Proceedings

of the

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada

Gijón (Asturias), Spain

June 14-18, 2021

SēMA
Sociedad Española
de Matemática Aplicada

Universidad de Oviedo

Editors:
Rafael Gallego, Mariano Mateos

Universidad de Oviedo
Servicio de Publicaciones de la Universidad de Oviedo
Campus de Humanidades. Edificio de Servicios. 33011 Oviedo (Asturias)
Tel. 985 10 95 03 Fax 985 10 95 07
http: www.uniovi.es/publicaciones
servipub@uniovi.es

ISBN: 978-84-18482-21-2

Foreword

It is with great pleasure that we present the Proceedings of the 26th Congress of Differential Equations and Applications / 16th Congress of Applied Mathematics (XXVI CEDYA / XVI CMA), the biennial congress of the Spanish Society of Applied Mathematics S$\vec{e}$MA, which is held in Gijón, Spain from June 14 to June 18, 2021.

In this volume we gather the short papers sent by some of the almost three hundred and twenty communications presented in the conference. Abstracts of all those communications can be found in the abstract book of the congress. Moreover, full papers by invited lecturers will shortly appear in a special issue of the S$\vec{e}$MA Journal.

The first CEDYA was celebrated in 1978 in Madrid, and the first joint CEDYA / CMA took place in Málaga in 1989. Our congress focuses on different fields of applied mathematics: Dynamical Systems and Ordinary Differential Equations, Partial Differential Equations, Numerical Analysis and Simulation, Numerical Linear Algebra, Optimal Control and Inverse Problems and Applications of Mathematics to Industry, Social Sciences, and Biology. Communications in other related topics such as Scientific Computation, Approximation Theory, Discrete Mathematics and Mathematical Education are also common.

For the last few editions, the congress has been structured in mini-symposia. In Gijón, we will have eighteen minis-symposia, proposed by different researchers and groups, and also five thematic sessions organized by the local organizing committee to distribute the individual contributions. We will also have a poster session and ten invited lectures. Among all the mini-symposia, we want to highlight the one dedicated to the memory of our colleague Francisco Javier "Pancho" Sayas, which gathers two plenary lectures, thirty-six talks, and more than forty invited people that have expressed their wish to pay tribute to his figure and work.

This edition has been deeply marked by the COVID-19 pandemic. First scheduled for June 2020, we had to postpone it one year, and move to a hybrid format. Roughly half of the participants attended the conference online, while the other half came to Gijón. Taking a normal conference and moving to a hybrid format in one year has meant a lot of efforts from all the parties involved. Not only did we, as organizing committee, see how much of the work already done had to be undone and redone in a different way, but also the administration staff, the scientific committee, the mini-symposia organizers, and many of the contributors had to work overtime for the change.

Just to name a few of the problems that all of us faced: some of the already accepted mini-symposia and contributed talks had to be withdrawn for different reasons (mainly because of the lack of flexibility of the funding agencies); it became quite clear since the very first moment that, no matter how well things evolved, it would be nearly impossible for most international participants to come to Gijón; reservations with the hotels and contracts with the suppliers had to be cancelled; and there was a lot of uncertainty, and even anxiety could be said, until we were able to confirm that the face-to-face part of the congress could take place as planned.

On the other hand, in the new open call for scientific proposals, we had a nice surprise: many people that would have not been able to participate in the original congress were sending new ideas for mini-symposia, individual contributions and posters. This meant that the total number of communications was about twenty percent greater than the original one, with most of the new contributions sent by students.

There were almost one hundred and twenty students registered for this CEDYA / CMA. The hybrid format allows students to participate at very low expense for their funding agencies, and this gives them the opportunity to attend different conferences and get more merits. But this, which can be seen as an advantage, makes it harder for them to obtain a full conference experience. Alfréd Rényi said: "a mathematician is a device for turning coffee into theorems". Experience has taught us that a congress is the best place for a mathematician to have a lot of coffee. And coffee cannot be served online.

In Gijón, June 4, 2021

The Local Organizing Committee from the Universidad de Oviedo

Scientific Committee

- Juan Luis Vázquez, Universidad Autónoma de Madrid
- María Paz Calvo, Universidad de Valladolid
- Laura Grigori, INRIA Paris
- José Antonio Langa, Universidad de Sevilla
- Mikel Lezaun, Euskal Herriko Unibersitatea
- Peter Monk, University of Delaware
- Ira Neitzel, Universität Bonn
- José Ángel Rodríguez, Universidad de Oviedo
- Fernando Terán, Universidad Carlos III de Madrid

Sponsors

- Sociedad Española de Matemática Aplicada
- Departamento de Matemáticas de la Universidad de Oviedo
- Escuela Politécnica de Ingeniería de Gijón
- Gijón Convention Bureau
- Ayuntamiento de Gijón

Local Organizing Committee from the Universidad de Oviedo

- Pedro Alonso Velázquez
- Rafael Gallego
- Mariano Mateos
- Omar Menéndez
- Virginia Selgas
- Marisa Serrano
- Jesús Suárez Pérez del Río

Contents

CONTENTS

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 8–14)

On numerical approximations to diffuse-interface tumor growth models

Daniel Acosta-Soba[1], **Francisco Guillén-González**[2], **J. Rafael Rodríguez-Galván**[3]
1. daniel.acosta@uca.es *Universidad de Cádiz, Spain*
2. guillen@us.es *Universidad de Sevilla, Spain*
3. rafael.rodriguez@uca.es *Universidad de Cádiz, Spain*

Abstract

This work is devoted to developing new numerical schemes for a tumor-nutrient PDE model. It is based on phase field equations for the tumor variable and a diffusive equation for the nutrient one, coupled by reaction terms and cross-diffusion terms. The model conserves the sum of tumor+nutrient and has a dissipative energy law.

We introduce two different time-discrete schemes: one is based on an Eyre-type decomposition of the energy and the other is an energy quadratization scheme. Both (continuous) Finite Elements and Discontinuous Galerkin are used for space discretization.

The schemes are compared analytically and computationally.

1. Introduction

In [3] this tumor-nutrient PDE model is proposed:

$$\partial_t u = \nabla \cdot (M_u \nabla \mu_u) + \delta P_0 u_+ (\mu_n - \mu_u) \qquad \text{in } \Omega \times (0,T], \tag{1.1a}$$

$$\mu_u = f'(u) - \varepsilon^2 \Delta u - \chi_0 n \qquad \text{in } \Omega \times (0,T], \tag{1.1b}$$

$$\partial_t n = \nabla \cdot (M_n \nabla \mu_n) - \delta P_0 u_+ (\mu_n - \mu_u) \qquad \text{in } \Omega \times (0,T], \tag{1.1c}$$

$$\mu_n = \frac{1}{\delta} n - \chi_0 u \qquad \text{in } \Omega \times (0,T], \tag{1.1d}$$

$$\nabla u \cdot \mathbf{n} = M_u \nabla \mu_u \cdot \mathbf{n} = M_n \nabla n \cdot \mathbf{n} = 0 \qquad \text{on } \partial\Omega \times (0,T], \tag{1.1e}$$

$$u(x,0) = u_0, \quad n(x,0) = n_0 \qquad \text{in } \Omega, \tag{1.1f}$$

where u is the tumor cell volume fraction and n is the nutrient density. Here M_u, M_n are nonnegative mobility functions, $\varepsilon, \delta, \chi_0 > 0$ and $f(u) = \Gamma u^2 (1-u)^2$, $\Gamma > 0$. Then, $f(u)$ is a polynomial double-well potential with stable values 0 and 1.

This model is conservative in the sense that $\partial_t \left(\int_\Omega u + n \right) = 0$ and the following energy is dissipated:

$$E(u,n) := \frac{\varepsilon^2}{2} \int_\Omega |\nabla u|^2 + \int_\Omega f(u) - \chi_0 \int_\Omega un + \frac{1}{2\delta} \int_\Omega n^2. \tag{1.2}$$

Specifically, the following energy law holds:

$$\frac{d}{dt} E = - \int_\Omega M_u |\nabla \mu_u|^2 - \int_\Omega M_n |\nabla \mu_n|^2 - \delta P_0 \int_\Omega u_+ (\mu_n - \mu_u)^2.$$

Well-posedness of this model is studied in [1, 2] for the case $\chi_0 = 0$ while, to the best knowledge of the authors, it is still an open problem for the cross-difussion case ($\chi_0 \neq 0$).

Also in [3], an Eyre-type (convex-splitting) time-discrete numerical scheme is studied which conserves the mass and has non-increasing discrete energy, giving several numerical examples using (continuous) Finite Elements (FE) as space discretization.

In this work, we introduce a slight modification of the nonlinear time semidiscretization shown in [3] which will lead to a first order linear scheme. Besides, we propose another semidiscrete-time scheme based on Invariant Energy Quadratization (IEQ). Furthermore, we extend the previous FE scheme to a Discontinuous Galerkin space discretization (DG). Finally, we compare the results both theoretically and computationally.

2. Discrete schemes

Troughout this section we show the different aforementioned time-semidiscrete and space-semidiscrete schemes. In particular, we focus on the theoretical properties of the Eyre-FE and IEQ-DG fully-discrete schemes. The proofs of the results will appear together with a numerical comparison of the other different time-space combinations in a forthcoming paper.

2.1. Fully-discrete Eyre-FE scheme

The Eyre-type scheme consist of splitting the energy functional (1.2) into two terms

$$E(u,n) = E_i(u,n) + E_e(u,n),$$

where $E_i(u,n)$ is a convex term (that we will treat implicitly) and $E_e(u,n)$ is a non-convex term (which we will treat explictly) so as to avoid the nonlineaity of the model (1.1a)–(1.1f).

In [3] the following splitting is considered

$$\begin{aligned} E_i(u,n) &:= \frac{\varepsilon^2}{2}\int_\Omega |\nabla u|^2 + \frac{3\Gamma}{2}\int_\Omega u^2 - \chi_0\int_\Omega un + \left(\frac{\alpha}{2}+\frac{1}{2\delta}\right)\int_\Omega n^2,\\ E_e(u,n) &:= \Gamma\int_\Omega \left(u^4 - 2u^3 - \frac{1}{2}u^2\right) - \frac{\alpha}{2}\int_\Omega n^2, \end{aligned}$$

where $\alpha \in \mathbb{R}$ is a stabilization parameter which must satisfy that $\alpha > \frac{\chi_0^2}{3\Gamma}$ in order to make $E_i(\cdot,\cdot)$ convex with respect to u and n independently.

Regarding the previous splitting, we propose the following linear scheme using a FE space-semidiscretization associated with a triangulation family $\mathcal{T}_h$ of Ω: find, for each $m \in \mathbb{N}\cup\{0\}$, $u^{m+1}, n^{m+1}, \mu_u^{m+1} \in \mathbb{P}_k^{\mathrm{cont}}(\mathcal{T}_h)$ such that for every $\overline{\mu}_u, \overline{u}, \overline{\mu}_n \in \mathbb{P}_k^{\mathrm{cont}}(\mathcal{T}_h)$,

$$\left(\delta_t u^{m+1}, \overline{\mu}_u\right)_{L^2(\Omega)} = -\left(M_u(u^m)\nabla\mu_u^{m+1}, \nabla\overline{\mu}_u\right)_{(L^2(\Omega))^d} + \delta P_0\left(u_+^m(\mu_n^{m+1}-\mu_u^{m+1}), \overline{\mu}_u\right)_{L^2(\Omega)}, \tag{2.1a}$$

$$\begin{aligned}\left(\mu_u^{m+1}, \overline{u}\right)_{L^2(\Omega)} &= \varepsilon^2\left(\nabla u^{m+1}, \nabla\overline{u}\right)_{(L^2(\Omega))^d} + 3\Gamma\left(u^{m+1}, \overline{u}\right)_{L^2(\Omega)}\\ &\quad + \Gamma\left(4(u^m)^3 - 6(u^m)^2 - u^m, \overline{u}\right)_{L^2(\Omega)} - \chi_0\left(n^{m+1}, \overline{u}\right)_{L^2(\Omega)},\end{aligned} \tag{2.1b}$$

$$\left(\delta_t n^{m+1}, \overline{\mu}_n\right)_{L^2(\Omega)} = -\left(M_n(n^m)\nabla\mu_n^{m+1}, \nabla\overline{\mu}_n\right)_{(L^2(\Omega))^d} - \delta P_0\left(u_+^m(\mu_n^{n+1}-\mu_u^{m+1}), \overline{\mu}_n\right)_{L^2(\Omega)}, \tag{2.1c}$$

where we denote $\delta_t u^{m+1} = \frac{u^{m+1}-u^m}{\Delta t}$ and

$$\mu_n^{m+1} = \frac{1}{\delta}n^{m+1} + \alpha\left(n^{m+1}-n^m\right) - \chi_0 u^{m+1}. \tag{2.2}$$

Remark 2.1 The previous time-semidiscrete scheme (2.1) differs from the one proposed in [3] in the way we treat u_+ in the reaction terms. Whereas in [3] this term is approximated using a Crank-Nicolson scheme, we just treat this term explicitly in order to avoid the nonlinearity.

Despite using a lower-order time-discretization scheme than in [3] for one of the terms in (2.1a)–(2.1c), we can also afford first-order consistency in time. In particular, using the ideas in [3] we get the following result.

Theorem 2.2 *The scheme* (2.1a)–(2.1c) *has the following properties*:

1. *If there is a smooth enough solution u, μ_u, n of* (1.1a)–(1.1f)*, then the scheme has first-order consistency in time.*
2. *The mass is conserved in the following sense*: $\int_\Omega (u^{m+1}+n^{m+1}) = \int_\Omega (u^m + n^m)$ *for* $m \geq 0$.
3. *If* $u^i \in \left[\frac{1}{2}-\frac{1}{\sqrt{3}}, \frac{1}{2}+\frac{1}{\sqrt{3}}\right]$ *for* $i \in \{m, m+1\}$, $m \geq 0$, *then* $E(u^{m+1}, n^{m+1}) \leq E(u^m, n^m)$.

Remark 2.3 It is granted by the previous result that the energy of the solution of the scheme (2.1a)–(2.1c) decreases as long as the solution u belongs to the interval $\left[\frac{1}{2}-\frac{1}{\sqrt{3}}, \frac{1}{2}+\frac{1}{\sqrt{3}}\right] \approx [-0.08, 1.08]$, where $E_e(\cdot,\cdot)$ is concave. Nevertheless, the model (1.1a)–(1.1f) has no maximum principle so the solution may not be bounded by this range. Hence, it is not guaranteed that the energy always decreases with this scheme (2.1a)–(2.1c).

2.2. Fully-discrete IEQ-DG scheme

The following space discretization of the variational formulation of the previous model (1.1a)–(1.1f) using Symmetric Interior Penalty (SIP) Discontinuous Galerkin is introduced: find $u(\cdot,t), n(\cdot,t), \mu_u(\cdot,t) \in \mathbb{P}_k^{\text{disc}}(\mathcal{T}_h)$ for a.e. $t \in (0,T)$ such that

$$\int_\Omega \partial_t u \overline{\mu}_u = -a_h^{\text{sip}}(M_u(u);\mu_u,\overline{\mu}_u) + \delta P_0 \int_\Omega u_+(\mu_n-\mu_u)\overline{\mu}_u, \tag{2.3a}$$

$$\int_\Omega \mu_u \overline{u} = a_h^{\text{sip}}(\varepsilon^2;u,\overline{u}) + \int_\Omega f'(u)\overline{u} - \chi_0 \int_\Omega n\overline{u}, \tag{2.3b}$$

$$\int_\Omega \partial_t n \overline{\mu}_n = -a_h^{\text{sip}}(M_n(n);\mu_n,\overline{\mu}_n) - \delta P_0 \int_\Omega u_+(\mu_n-\mu_u)\overline{\mu}_n, \tag{2.3c}$$

$$\mu_n = \frac{1}{\delta} n - \chi_0 u, \tag{2.3d}$$

for every $\overline{\mu}_u, \overline{u}, \overline{\mu}_n \in \mathbb{P}_k^{\text{disc}}(\mathcal{T}_h)$, with $a_h^{\text{sip}}(\cdot;\cdot,\cdot)$ the SIP-blinear form defined for $v, w \in \mathbb{P}_k^{\text{disc}}(\mathcal{T}_h)$ as in [4], where this kind of IEQ-DG schemes are studied for the Cahn-Hilliard equations:

$$\begin{aligned} a_h^{\text{sip}}(M(a);v,w) \coloneqq & \sum_{K\in\mathcal{T}_h}\int_K M(a)\nabla_h v\cdot\nabla_h w - \sum_{e\in\mathcal{E}_h^i}\int_e M(\{\!\{a\}\!\})\left(\{\!\{\nabla_h v\cdot\mathbf{n}_e\}\!\}[\![w]\!] + \{\!\{\nabla_h w\cdot\mathbf{n}_e\}\!\}[\![v]\!]\right) \\ & + \sigma\sum_{e\in\mathcal{E}_h^i}\int_e \frac{1}{h_e} M(\{\!\{a\}\!\})[\![v]\!][\![w]\!]. \end{aligned}$$

Regarding the IEQ time-semidiscretization we have that, taking $B > 0$ and defining the artifical variable $U = \sqrt{f(u)+B}$ and $H(u) = \frac{f'(u)}{2\sqrt{f(u)+B}}$, then $f'(u) = 2H(u)U(u)$ and

$$\partial_t U = H(u)\partial_t u. \tag{2.4}$$

Now, we will approximate (2.4) in the $(m+1)$-th time iteration in two steps: first, we use the projection operator $\Pi_h\colon L^2(\Omega) \longrightarrow \mathbb{P}_k^{\text{disc}}(\mathcal{T}_h)$ to calculate $U_h^m = \Pi_h U^m$ and then we use the semidiscrete scheme

$$\frac{U^{m+1}-U_h^m}{\Delta t} = H(u^m)\frac{u^{m+1}-u^m}{\Delta t},$$

where $U^0 = \sqrt{f(u^0)+B}$.

Therefore, we propose the next fully discrete IEQ-DG scheme: find, for each $m \in \mathbb{N}\cup\{0\}$, $u^{m+1}, n^{m+1}, \mu_u^{m+1} \in \mathbb{P}_k^{\text{disc}}(\mathcal{T}_h)$ such that for every $\overline{\mu}_u, \overline{u}, \overline{\mu}_n \in \mathbb{P}_k^{\text{disc}}(\mathcal{T}_h)$,

$$\left(\delta_t u^{m+1}, \overline{\mu}_u\right) + a_h^{\text{sip}}(M_u(u^m);\mu_u^{m+1},\overline{\mu}_u) = \delta P_0\left(u_+^m(\mu_n^{m+1}-\mu_u^{m+1}),\overline{\mu}_u\right), \tag{2.5a}$$

$$\left(\mu_u^{m+1},\overline{u}\right) = a_h^{\text{sip}}(\varepsilon^2;u^{m+1},\overline{u}) + \left(2H(u^m)U^{m+1},\overline{u}\right) - \chi_0\left(n^m,\overline{u}\right), \tag{2.5b}$$

$$\left(\delta_t n^{m+1}, \overline{\mu}_n\right) + a_h^{\text{sip}}(M_n(n^m);\mu_n^{m+1},\overline{\mu}_n) = -\delta P_0\left(u_+^m(\mu_n^{m+1}-\mu_u^{m+1}),\overline{\mu}_n\right), \tag{2.5c}$$

where

$$\mu_n^{m+1} = \frac{1}{\delta} n^{m+1} - \chi_0 u^{m+1}, \tag{2.6a}$$

$$U^{m+1} = U_h^m + H(u^m)(u^{m+1}-u^m). \tag{2.6b}$$

In practice, in order to solve the previous system of equations (2.5a)–(2.5c) minimising the computational costs, in each time step we do the following:

1. We introduce the expressions (2.6a)–(2.6b) in (2.5a)–(2.5c) and we solve the system of equations.
2. The approximation U_h^{m+1} is obtained by projecting (2.6b) into $\mathbb{P}_k^{\text{disc}}(\mathcal{T}_h)$.

Theorem 2.4 *The scheme* (2.5a)–(2.5c) *has the following properties*:

1. *The mass is conserved*: $\int_\Omega (u^{m+1}+n^{m+1}) = \int_\Omega (u^m+n^m)$.

2. *The following energy law holds*:

$$\delta_t E_h(u^{m+1}, U_h^{m+1}, n^{m+1}) + a_h^{sip}(M_u(u^m); \mu_u^{m+1}, \mu_u^{m+1}) + a_h^{sip}(M_n(n^m); \mu_n^{m+1}, \mu_n^{m+1}) + \delta P_0 \int_\Omega u_+^m (\mu_n^{m+1} - \mu_u^{m+1})^2 \le 0$$

for the modified energy

$$E_h(u, U_h, n) = a_h^{sip}\left(\frac{\varepsilon^2}{2}; u, u\right) + \int_\Omega U_h^2 - \chi_0 \int_\Omega un + \frac{1}{2\delta}\int_\Omega n^2.$$

Remark 2.5 Observe that the EQ-DG scheme (2.5a)–(2.5c) is unconditionally energy-stable for a modified energy obtained by using the artificial variable U. Nonetheless, if the approximation is good enough (for small Δt and h) we have that $E_h(u^m, n^m) + B|\Omega| \approx E_h(u^m, U_h^m, n^m)$ where now

$$E_h(u, n) = a_h^{\text{sip}}\left(\frac{\varepsilon^2}{2}; u, u\right) + \int_\Omega f(u) - \chi_0 \int_\Omega un + \frac{1}{2\delta}\int_\Omega n^2,$$

is the natural discrete energy of the model in $\mathbb{P}_k^{\text{disc}}(T_h)$.

2.3. Numerical experiments

In this section we show several numerical experiments with the purpose of comparing both the Eyre-FE (2.1a)–(2.1a) and the IEQ-DG (2.5a)–(2.5c) schemes and studying some properties of the tumor model (1.1a)–(1.1f) by reproducing some results of the papers [3, 8].

Example 1. Circular tumor growth
We show the results we got with the two schemes for a initial small tumor (radius 0.1) located in the center of the domain $\Omega = [-1, 1]^2$, considering, at the beginning, the extracellular water to be completely nutrient-rich, $n_0 = 1$, in Ω.

We take the parameters $\varepsilon = 0.005$, $\delta = 0.01$, $P_0 = 0.1$, $\chi_0 = 0.05$, $\Gamma = 0.045$ y $\alpha = \frac{\chi_0^2}{3\Gamma} + 0.1$ as in [3] so as to reproduce its results. Likewise, we will take the mobility functions $M_u(u) = 200u^2$ and $M_n = \delta$. In the case of the IEQ-DG scheme we use $\sigma = 15$ and $B = 1$.

A time step and a mesh size $h \approx 0.04$ are used together with polynomials of order $k = 2$.

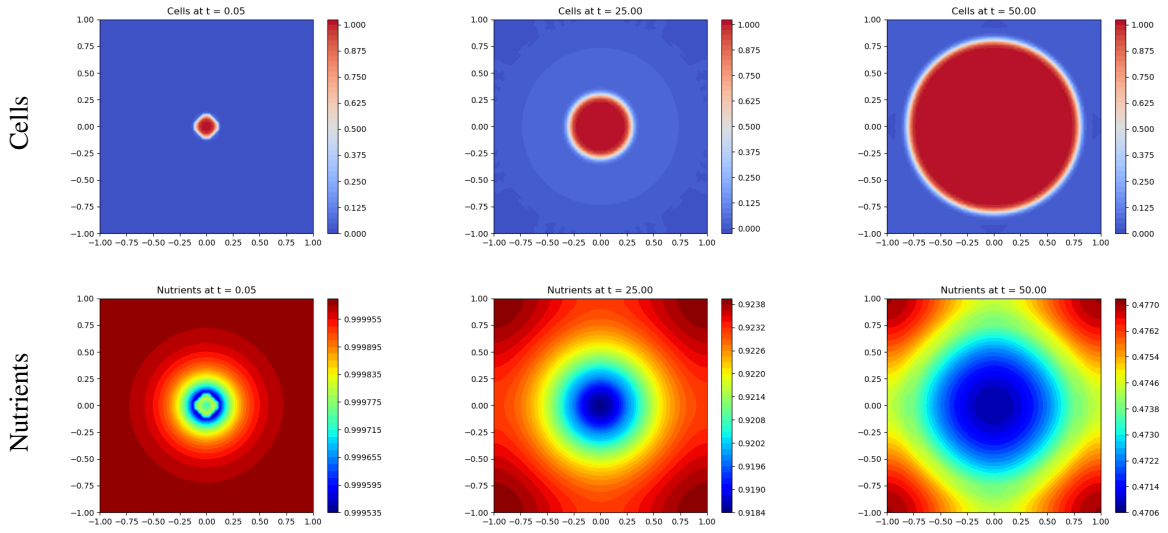


Fig. 1 Example 1. Cells and nutrients at different time steps ($\Delta t = 5 \cdot 10^{-2}$) with the **Eyre-FE** scheme (2.1).

Comparing Fig. 1 and Fig. 2, despite the qualitative behaviour of the schemes seem to be similar, the scheme IEQ-DG is much more unstable in time and a significantly smaller time step is needed to control the spurious oscillations over time. The time step used is $\Delta t = 0.05$ for Eyre-FE and $\Delta t = 0.002$ for IEQ-DG. This issue makes it very difficult to reach the final time $T = 50$ with the IEQ-DG scheme as it was easily done using the scheme

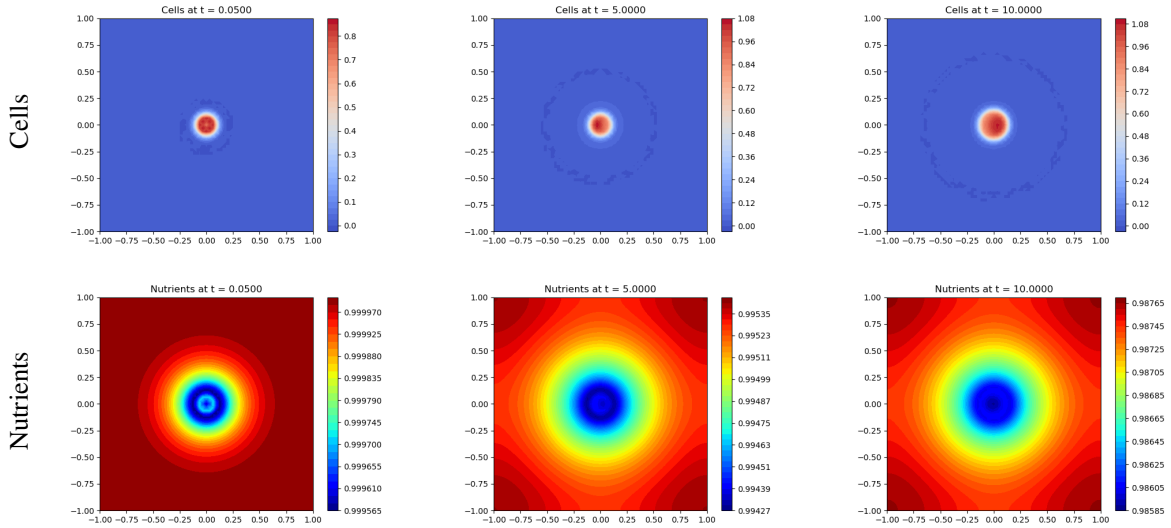


Fig. 2 Example 1. Cells and nutrients at different time steps ($\Delta t = 2 \cdot 10^{-3}$) with the **IEQ-DG** scheme (2.5).

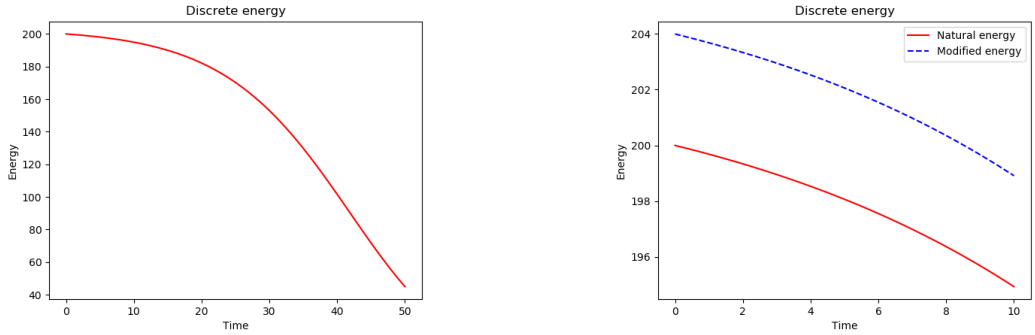


Fig. 3 Example 1. Discrete energy with the **Eyre-FE** scheme (2.1) on the left and with the **IEQ-DG** scheme (2.5) on the right.

Eyre-FE. The unstability of these well-known and widely studied IEQ time-semidiscretization technique has been spotted in several papers such as [5–7].

In both cases the mass is conserved and the energy decresases (Fig. 3). Notice that on the right of Fig. 3 we can observe that both the natural and the modified energies differ only in the constant $B|\Omega| = 4$.

Example 2. Elliptic tumor with nutrient source growth

Now, we slightly modify the model as it was done in [3] in order to increase the interaction between the tumor and the nutrients. To this aim, we guess that the diffusion of the nutrients is much faster than the growth of the tumor cells so we consider the elliptic version of the n-equation, taking $\partial_t n = 0$. Moreover, we change $\nabla n \cdot \mathbf{n} = 0$ to $n = 1$ on $\partial\Omega$.

$$\partial_t u = M_u \Delta \mu_u + \delta P_0 u_+ (\mu_n - \mu_u) \qquad \text{in } \Omega \times (0,T), \tag{2.7a}$$

$$\mu_u = F'(u) - \varepsilon^2 \Delta u - \chi_0 n \qquad \text{in } \Omega \times (0,T), \tag{2.7b}$$

$$0 = M_n \Delta \mu_n - \delta P_0 u_+ (\mu_n - \mu_u) \qquad \text{in } \Omega \times (0,T), \tag{2.7c}$$

$$\mu_n = \frac{1}{\delta} n - \chi_0 u \qquad \text{in } \Omega \times (0,T), \tag{2.7d}$$

$$\nabla u \cdot \mathbf{n} = \nabla \mu_u \cdot \mathbf{n} = 0 \qquad \text{on } \partial\Omega \times (0,T), \tag{2.7e}$$

$$n = 1 \qquad \text{on } \partial\Omega \times (0,T), \tag{2.7f}$$

$$u(x,0) = u_0, \quad n(x,0) = n_0 \qquad \text{in } \Omega, \tag{2.7g}$$

where $u_0, n_0 \in L^2(\Omega)$.

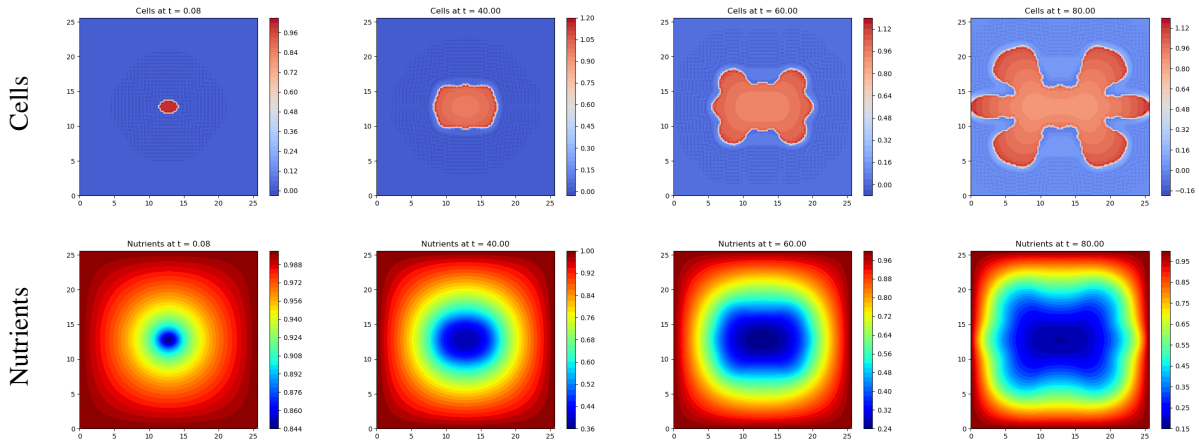


Fig. 4 Example 2. Cells and nutrients of the variant (2.7) with the **Eyre-FE** scheme (2.1).

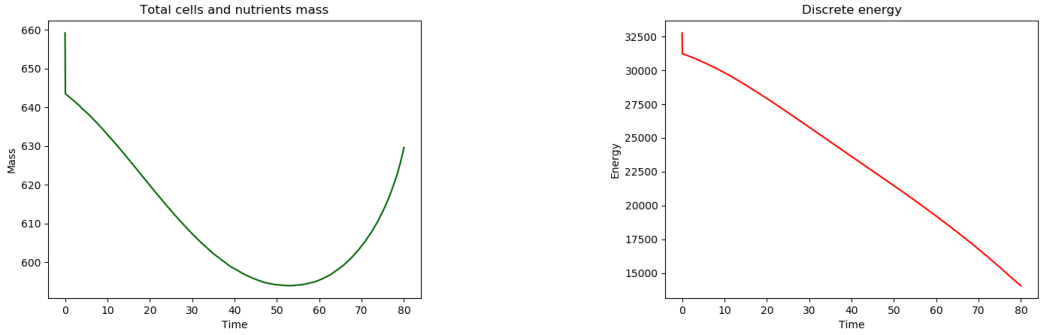


Fig. 5 Example 2. Mass and discrete energy of the variant (2.7) with the **Eyre-FE** scheme (2.1).

Consider the domain $\Omega = [0, 25.6]^2$. We take $u_0 = 1$ in the region $\left\{(x, y) \in \mathbb{R}^2 : \frac{(x-12.8)^2}{1.7} + \frac{(y-12.8)^2}{0.9} \leq 1\right\}$ and $n_0 = 1$ in Ω.

We keep the same parameters than in the example 1 as it is done in [3]. We take $\Delta t = 8 \cdot 10^{-2}$ and $h \approx 0.36$ with polynomials of order $k = 1$.

The Eyre-FE scheme (2.1a)–(2.1c) is used to obtain the solution that is shown in Fig. 4. It is remarkable to underline the evolution of shape of the tumor over time, forming buds towards the higher levels of nutrients as in [3].

Nontheless, the solution escapes from the range $[0, 1]$ which is due to the lack of maximum principle of the model (1.1) and, consequently, of the Eyre-FE scheme (2.1a)–(2.1c). Moreover, some oscillations are produced which may be due to the bad approximations of the cross-diffusion terms as $\chi_0 \neq 0$.

In this case, the energy and mass functions that we obtain with the modification of the model (2.7a)–(2.7g) are shown in the Fig. 5. Now, the energy may not be dissipative in general and neither is tumor + nutrient conserved.

Example 3. Aggregation of circular tumors

We show a simulation of the aggregation process of three circular tumors as it was done in [8]. To this aim, we consider the domain $\Omega = [-1, 1]^2$ and we guess that there are three tumors in the regions $\overline{B}((0.2, 0.2), 0.01)$, $\overline{B}((0.3, -0.5), 0.01)$ and $\overline{B}((-0.15, -0.15), 0.03)$ with the maximum concentration of nutrients in the extracellular water $n_0 = 1$.

This time we use the IEQ-DG scheme (2.5a)–(2.5c). We take the parameters $\varepsilon = 0.02$, $\delta = 0.01$, $P_0 = 100$, $\chi_0 = 0$, $\Gamma = 0.045$ and $B = 1$. We consider the constant mobility functions $M_u = 1$ and $M_n = \delta$. Moreover we use polynomials of order $k = 1$ with a penalization parameter $\sigma = 4$ and we take $\Delta t = 10^{-4}$ and $h \approx 0.02$.

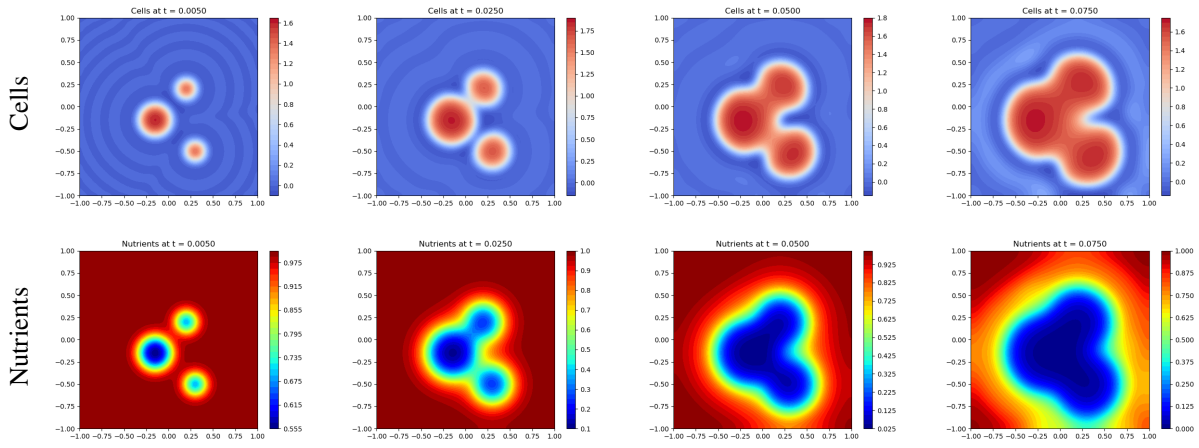


Fig. 6 Example 3. Cells and nutrients at different time steps with the **IEQ-DG** scheme (2.5).

Looking at the Fig. 6, the solution is not bounded in the interval $[0, 1]$ and some remarkable oscillations are produced as there is no maximum principle and we are not properly approximating the cross diffusion terms.

The mass is conserved and both the natural and the modified energies have the same decreasing behaviour (Fig. 7) differing only in the constant $B|\Omega| = 4$.

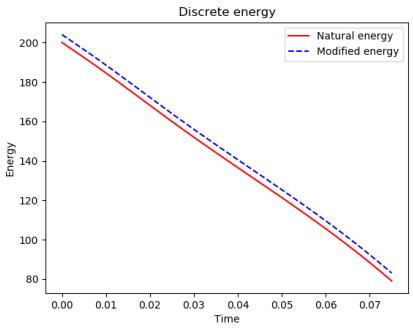


Fig. 7 Example 3. Discrete energy with the **IEQ-DG** scheme (2.5).

Acknowledgements

The first author has been supported by *UCA FPU contract UCA/REC14VPCT/2020 funded by Universidad de Cádiz*. The second and third authors have been supported by *Proyecto PGC2018-098308-B-I00, funded by FEDER/Ministerio de Ciencia e Innovación - Agencia Estatal de Investigación, Spain*.

References

[1] Pierluigi Colli, Gianni Gilardi, and Danielle Hilhorst. On a Cahn-Hilliard type phase field system related to tumor growth. *Discrete & Continuous Dynamical Systems - A*, 35(6):2423–2442, June 2015.

[2] Sergio Frigeri, M. Grasselli, and Elisabetta Rocca. On a diffuse interface model of tumor growth. *European Journal of Applied Mathematics*, 26(2):215–243, January 2015.

[3] Andrea Hawkins-Daarud, Kristoffer G. van der Zee, and J. Tinsley Oden. Numerical simulation of a thermodynamically consistent four-species tumor growth model. *International Journal for Numerical Methods in Biomedical Engineering*, 28(1):3–24, January 2012.

[4] Hailiang Liu and Peimeng Yin. Unconditionally energy stable discontinuous galerkin schemes for the cahn–hilliard equation. *Journal of Computational and Applied Mathematics*, 390:113375, July 2021.

[5] Zhengguang Liu. Efficient invariant energy quadratization and scalar auxiliary variable approaches without bounded below restriction for phase field models. *arXiv:1906.03621 [math]*, November 2019.

[6] Jie Shen, Jie Xu, and Jiang Yang. The scalar auxiliary variable (SAV) approach for gradient flows. *Journal of Computational Physics*, 353:407–416, January 2018.

[7] Jie Shen, Jie Xu, and Jiang Yang. A New Class of Efficient and Robust Energy Stable Schemes for Gradient Flows. *SIAM Review*, 61(3):474–506, January 2019.

[8] X. Wu, G. J. van Zwieten, and K. G. van der Zee. Stabilized second-order convex splitting schemes for Cahn-Hilliard models with application to diffuse-interface tumor-growth models. *International Journal for Numerical Methods in Biomedical Engineering*, 30(2):180–203, 2014.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 15–22)

An optimized sixth-order explicit RKN method to solve oscillating systems

Musa Ahmed Demba[1], Higinio Ramos[2], Poom Kumam[3], Wiboonsak Watthayu[4]

1. musdem2004@gmail.com *Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand.*
KMUTTFixed Point Research Laboratory, KMUTT-Fixed Point Theory and Applications Research Group, Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi (KMUTT), 126 Pracha-Uthit Road, Bang Mod, Thrung Khru, Bangkok 10140, Thailand.
Department of Mathematics, Faculty of Computing and Mathematical Sciences, Kano University of Science and Technology, Wudil, P.M.B 3244 Kano State, Nigeria
2. higra@usal.es *Department of Applied Mathematics, Faculty of Sciences, University of Salamanca, Salamanca 37008, Spain. Escuela Politécnica Superior, Avda. de Requejo, 33, 49022 Zamora, Spain*
3. poom.kum@kmutt.ac.th *Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand.*
KMUTTFixed Point Research Laboratory, KMUTT-Fixed Point Theory and Applications Research Group, Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi (KMUTT), 126 Pracha-Uthit Road, Bang Mod, Thrung Khru, Bangkok 10140, Thailand
4. wiboonsak.wat@kmutt.ac.th *Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand*

Abstract

Optimization of the sixth-order explicit Runge-Kutta-Nyström method with six stages derived by El-Mikkawy and Rahmo using the phase-fitted and amplification-fitted techniques with constant step-size is developed in this paper. The new method integrates exactly the common test: $y'' = -w^2 y$. The local truncation error of the new method is computed, showing that the order of convergence is maintained. The stability analysis is addressed, showing that the developed method has a periodicity interval. The numerical experiments demonstrate the high performance of the proposed scheme compared to other existing RKN codes with six stages and same order.

1. Introduction

This paper aim to effectively solve the special second-order initial-value problem of the form

$$y'' = f(x, y), \qquad y(x_0) = y_0, \qquad y'(x_0) = y'_0, \tag{1.1}$$

assuming that their solutions are oscillatory, where $y \in \mathfrak{R}^d$ and $f : \mathfrak{R} \times \mathfrak{R}^d \to \mathfrak{R}^d$ are sufficiently differentiable. In recent and past years, the search of new numerical algorithms to effectively solve (1.1) has brought the attention of many researchers due to the great role the problem played in so many areas of applied sciences. To solve (1.1) directly, the class of Runge-Kutta-Nyström (RKN) methods has been largely used. Regarding the effective use of these methods, some RKN methods of sixth-order with six stages have been developed in [6], [7], and [1]. A lot of adapted RKN methods have been developed, which are of less algebraic-order than the constructed method in this paper. To mention a few, we cite those in [2, 3, 9–11, 13]. Recently, Demba et al. [4, 5] derived two new explicit RKN methods trigonometrically adapted for solving the kind of problems in (1.1).
This work aims at the development of a new phase- and amplification-fitted sixth order explicit RKN method with six stages based on the sixth order method of the RKN6(4)6ER pair given in [1] for solving the problem in (1.1). The derived method solves exactly the test equation $y'' = -w^2 y$. The numerical experiments reveal the effectiveness of the developed method compared to standard RKN codes of sixth order with six stages.

The remaining part of this paper is organized in this way: the basic theory of explicit RKN methods, the definitions of phase-lag and amplification error, and the definitions regarding the stability analysis are addressed in Section 2. Section 3 is devoted to the construction of the new code, to determine the order and error analysis, and to bring some details about the periodicity interval of the derived code. Some numerical examples are presented in Section 4, showing the good performance of the proposed scheme. Comments on the obtained results are given in Section 5, and finally, Section 6 gives a conclusion.

2. Fundamental Concepts

2.1. Explicit Runge-Kutta-Nyström Methods

An explicit RKN method with r stages is generally expressed by the formulas:

$$y_{n+1} = y_n + hy'_n + h^2 \sum_{l=1}^{r} b_l f(x_n + c_l h, Y_l), \tag{2.1}$$

$$y'_{n+1} = y'_n + h \sum_{l=1}^{r} d_l f(x_n + c_l h, Y_l), \tag{2.2}$$

$$Y_l = y_n + c_l h y'_n + h^2 \sum_{j=1}^{l-1} a_{lj} f(x_n + c_j h, Y_j), \quad l = 1, \ldots, r, \tag{2.3}$$

where as usual, y_{n+1} and y'_{n+1} denote approximations for $y(x_{n+1})$ and $y'(x_{n+1})$, respectively, and the grid points on the integration interval $[x_0, x_N]$ are given by $x_j = x_0 + jh,\ j = 0, 1, \ldots, N$, with h a fixed step-size.
The above method may be formulated compactly using the Butcher array in the form

$$\begin{array}{c|c} c & A \\ \hline & b^T \\ & d^T \end{array}$$

being $A = (a_{ij})_{r\times r}$ a matrix of coefficients, $c = (c_1, c_2, \ldots, c_r)^T$ is the vector of stages, and $b = (b_1, b_2, \ldots, b_r)^T$, $d = (d_1, d_2, \ldots, d_r)^T$ are two vectors containing the remaining coefficients of the method. For short, this can be denoted as (c, A, b, d).

Definition 1 ([8]) *An explicit Runge-Kutta-Nyström method as given in equations* (2)–(4) *is said to have algebraic order k if at any grid point x_{n+1} it holds*

$$\begin{cases} y_{n+1} - y(x_n + h) = O(h^{k+1}), \\ y'_{n+1} - y'(x_n + h) = O(h^{k+1}). \end{cases} \tag{2.4}$$

2.2. Analysis of Phase-lag, amplification error and stability

Applying the RKN method in (2)–(4) to the test equation $y'' = -w^2 y$, the phase-lag, amplification error and the linear stability analysis are derived. In particular, letting $\tilde{h} = -(wh)^2$, the approximate solution provided by (2)–(4) verifies the recurrence equation:

$$L_{n+1} = E(\tilde{h}) L_n,$$

where

$$L_{n+1} = \begin{bmatrix} y_{n+1} \\ hy'_{n+1} \end{bmatrix},\ L_n = \begin{bmatrix} y_n \\ hy'_n \end{bmatrix},\ E(\tilde{h}) = \begin{bmatrix} 1 + \tilde{h} b^T N^{-1} e & wh(1 + \tilde{h} b^T N^{-1} c) \\ -whd^T N^{-1} e & 1 + \tilde{h} d^T N^{-1} c \end{bmatrix},$$

$N = I - \tilde{h}A$, with I the identity matrix of dimension six, $A = \left(a_{ij}\right)_{6\times 6}, b, c, d$ are the corresponding matrix and vectors collecting the coefficients, and $e = [1, 1, 1, 1, 1, 1]^T$.

For enough small values of $\mu = wh$, it can be assumed that the matrix $E(\tilde{h})$ possesses conjugate complex eigenvalues [15]. Under this assumption, an oscillatory numerical solution should be provided by the method. The oscillatory character depends on the eigenvalues of the stability matrix $E(\tilde{h})$. The characteristic equation of this matrix can be expressed as:

$$\lambda^2 - \lambda Tr(E(\tilde{h})) + Det(E(\tilde{h})) = 0. \tag{2.4}$$

Theorem 2 ([1]) *If we apply to the common test equation $y'' = -w^2 y$ the Runge-Kutta-Nyström scheme in* (2)–(4)*, we get the formula for calculating directly the phase-lag* (*or dispersion error*) $\Psi(\mu)$ *given by*:

$$\Psi(\mu) = \mu - cos^{-1}\left(\frac{Tr(E(\tilde{h}))}{2\sqrt{Det(E(\tilde{h}))}}\right). \tag{2.4}$$

If $\Psi(\mu) = O(\mu^{l+1})$, then it is said that the method has a phase-lag of order l. For an explicit RKN method, $Tr(E(\tilde{h})$ and $Det(E(\tilde{h})$ are polynomials in μ (in case of an implicit RKN method these would be rational functions).

Definition 3 *An explicit Runge-Kutta-Nyström method as given in equations* (2)–(4) *is said to be phase-fitted, if the phase-lag is zero.*

Definition 4 ([1]) *For the Runge-Kutta-Nyström method given in equations* (2)–(4)*, the value* $\beta(\mu) = 1 - \sqrt{Det(E(\tilde{h})}$ *is known as the amplification error* (*or dissipative error*)*. If* $\beta(\mu) = O(\mu^{s+1})$*, then it is said that the method has an amplification error of order s.*

Definition 5 *An explicit Runge-Kutta-Nyström method as given in equations* (2)–(4) *is said to be amplification-fitted if the amplification-error is zero.*

Definition 6 *An interval* $(-\tilde{h}_b, 0)$*,* $\tilde{h}_b \in \mathfrak{R}$ *is named as the primary interval of periodicity of the method in* (2)–(4)*, if* $\tilde{h}$ *is the highest value such that for all* $\tilde{h}_b \in (-\tilde{h}_b, 0)$*,* $|\lambda_{1,2}| = 1$ *and* $\lambda_1 \neq \lambda_2$*. Where* $\lambda_{1,2}$ *are the solutions of the equation in* (2.2)*.*

3. Development of the new scheme

In this section, we will obtain a sixth order explicit phase- and amplification-fitted RKN scheme based on the higher-order method in the RKN6(4)6ER embedded pair derived by El-Mikkawy and Rahmo in [7], which we named as RKN6-6ER. The coefficients of the sixth order RKN method in [7] are shown in Table 1 with the correct value of a_{54} as given in [1].

Tab. 1 The RKN6-6ER Method in [7]

0						
$\frac{1}{77}$	$\frac{1}{11858}$					
$\frac{1}{3}$	$-\frac{7189}{17118}$	$\frac{4070}{8559}$				
$\frac{2}{3}$	$\frac{4007}{2403}$	$-\frac{589655}{355644}$	$\frac{25217}{118548}$			
$\frac{13}{15}$	$-\frac{4477057}{843750}$	$\frac{13331783894}{2357015625}$	$-\frac{281996}{5203125}$	$\frac{563992}{7078125}$		
1	$\frac{17265}{2002}$	$-\frac{1886451746}{212088107}$	$\frac{22401}{31339}$	$\frac{2964}{127897}$	$\frac{178125}{5428423}$	
	$-\frac{341}{780}$	$\frac{386683451}{661053840}$	$\frac{2853}{11840}$	$\frac{267}{3020}$	$\frac{9375}{410176}$	0
	$-\frac{341}{780}$	$\frac{29774625727}{50240091840}$	$\frac{8559}{23680}$	$\frac{801}{3020}$	$\frac{140625}{820352}$	$\frac{847}{18240}$

In order to get the new adapted scheme, we equate to zero the phase-lag $\Psi(\mu)$ and the amplification error $\beta(\mu)$, and we get the system:

$$\begin{cases} \Psi(\mu) = 0 \\ \beta(\mu) = 0. \end{cases} \tag{3.0}$$

We solve this system considering the coefficients in Table 1 except two of them which are taking as unknowns. Specifically, we take b_5 and d_5 as unknowns. We obtain the following values:

$$
\begin{aligned}
b_5 \;=\; & \frac{2503125}{410176M}\Big(-1258632233707368303463680000000 + 52499468404370638714802508000 0\,\mu^2 \\
& +380278327832939254939061688 00\,\mu^4 - 42305110040020986855472545000\,\mu^6 + 63894963509037530795250171 00\,\mu^8 \\
& -396360945814751886526623990\,\mu^{10} + 12393674919826270714885995\,\mu^{12} - 16375738211195081948868 6\,\mu^{14} \\
& +4438802446260702785 20\,\mu^{16} - 3556135517458913619310080000\,\mu^2\cos(\mu) \\
& +796929595765552632521698560 0\,\mu^6\cos(\mu) - 1257180203210973608863296 00\,\mu^8\cos(\mu) \\
& -7426931555859094858069370880 0\,\mu^4\cos(\mu) + 1258632233707368303463680000000\cos(\mu)\Big),
\end{aligned}
\tag{3.1}
$$

$$
\begin{aligned}
d_5 \;=\; & \frac{625}{820352M}\Big(4882682690886773063720\,\mu^{20} - 1766435438191731348692196\,\mu^{18} + 1426712864988780120153495 60\,\mu^{16} \\
& -10126226143892166109616015370\,\mu^{14} + 4754939043963115273766323268 25\,\mu^{12} \\
& +5468819730508407827785271040 0\,\mu^{10}\cos(\mu) - 1039168019912554487955565244565 0\,\mu^{10} \\
& -1038190029658946246749232966400 0\,\nu^8\cos(\nu) + 1130867607580895732412988295865 00\,\mu^8 \\
& +2536902040490601057324033984000 00\,\mu^6\cos(\mu) - 3817218324598810210637764771950 00\,\mu^6 \\
& -1775893905546681693988359573660000\,\mu^4 - 6305505579734821871351779238400 00\,\mu^4\cos(\mu) \\
& +3199753041551405164628715874500000 0\,\mu^2 - 6721096127997346740496051200000 00\,\mu^2\cos(\nu) \\
& +7561233143997015083058057600000000 0\cos(\mu) - 75612331439970150830580576000000000\Big),
\end{aligned}
\tag{3.2}
$$

where

$$
\begin{aligned}
M \;=\; & \mu^2\Big(-28803310743425593080234375000000 + 48005517905709321800390625000 00\,\mu^2 \\
& +2409864727821008473951031250 00\,\mu^4 - 21157585474732123459365303750 0\,\mu^6 \\
& +2769337946941422457432279275 0\,\mu^8 - 1543565245575968927989765335\,\mu^{10} \\
& +5515885104849964144936935 0\,\mu^{12} - 8615785571703447482682 48\,\mu^{14} + 2441341345443386531860\,\mu^{16}\Big).
\end{aligned}
$$

The corresponding Taylor series expansions in powers of μ are given by

$$\begin{aligned}
b_5 &= \frac{9375}{410176} - \frac{261461}{93847723200}\mu^6 + \frac{20361401}{369525410100000}\mu^8 - \frac{177044709462626977}{866977960060782108000000 0}\mu^{10} \\
&+\frac{11347558575343312922557}{88756868661222568306500000000 0}\mu^{12} - \frac{10147779116018364843223853 9}{13668557773828275519201000000000000 0}\mu^{14} + \cdots \quad , \\
d_5 &= \frac{140625}{820352} - \frac{1}{213290280}\mu^6 - \frac{618923}{739050820200}\mu^8 - \frac{1251344791}{93120403345200000}\mu^{10} \\
&-\frac{190297638076116325219}{73964057217685473588750000 00}\mu^{12} + \frac{35276945432092739240316 79}{99407692900569276503280000000000 0}\mu^{14} + \cdots \quad . \qquad (3.3)
\end{aligned}$$

As expected, when $\mu \to 0$, the newly obtained coefficients b_5, d_5 become the coefficients of the counterpart scheme in the original method. The new adapted RKN scheme will be named as PFAFRKN6-6ER.

3.1. Order of Convergence

This section is devoted to present the local truncation error of the proposed method and to get the order of convergence. This is accomplished by using the usual tool of Taylor expansions. The local truncation errors (LTE) at the point x_{n+1} of the solution and the first derivative are given respectively by:

$$\begin{aligned}
LTE &= y_{n+1} - y(x_n + h), \\
LTE_{der} &= y'_{n+1} - y'(x_n + h). \qquad (3.3)
\end{aligned}$$

Proposition 7 *The corresponding LTEs of the formulas to provide the solution and the derivative with the new RKN method are, respectively*:

$$\begin{aligned}
LTE &= \frac{h^7}{213290280}(f_y)^2(f_x + f_y y') + O(h^8), \\
LTE_{der} &= \frac{h^7}{5040}(f_{xxxxxx} + 15(y')^4 f_{yyyyy} y'' + 60(y')^3 f_{xyyyy} y'' + 60y' f_{xxxyy} y'' + 90(y')^2 f_{xxyyy} y'' + 21 f_y f_{yxx} y'' \\
&+60y'' f_{xyy} f_x + 15y'' f_{yy} f_{xx} + 18(y'')^2 f_{yy} f_y + 90y' f_{xyyy}(y'')^2 + 45(y')^2 f_{yyyy}(y'')^2 + 33(y')^2 (f_{yy})^2 y'' \\
&+48y' f_{xy} f_{yxx} + 10 f_y f_{xy} f_x + 12(f_y)^2 y' f_{xy} + 60y' f_{xxyy} f_x + 60(y')^2 f_{xyyy} f_x + 20(y')^3 f_{yyyy} f_x \\
&+24 f_y y' f_{xxxy} + 30y' f_{xyy} f_{xx} + 15(y')^2 f_{yyy} f_{xx} + 6y' f_{yy} f_{xxx} + 78(y')^2 f_{xyy} f_{xy} + 66(y')^2 f_y f_{xxyy} \\
&+33(y')^2 f_{yy} f_{yxx} + 64(y')^3 f_y f_{xyyy} + 36(y')^3 f_{yyy} f_{xy} + 48(y')^3 f_{yy} f_{xyy} + 21(f_y)^2 (y')^2 f_{yy} \\
&+21(y')^4 f_y f_{yyyy} + 21(y')^4 f_{yyy} f_{yy} + 15(y'')^3 f_{yyy} + 45(y'')^2 f_{xxyy} + 15y'' f_{xxxxy} + 18y''(f_{xy})^2 \\
&+(f_y)^3 y'' + (y')^6 f_{yyyyyy} + 6(y')^5 f_{xyyyyy} + (f_y)^2 f_{xx} + 6 f_{xxx} f_{xy} + f_y f_{xxxx} + 20 f_x f_{xxxy} \\
&+6y' f_{xxxxxy} + 15 f_{yxx} f_{xx} + 15(y')^4 f_{xxyyyy} + 15(y')^2 f_{xxxxyy} + 20(y')^3 f_{xxxyyy} + 10 f_{yy}(f_x)^2 \\
&+81(y')^2 f_{yyy} f_y y'' + 60y' f_{yyy} f_x y'' + 102y' f_y f_{xyy} y'' + 66y' f_{yy} f_{xy} y'' + 30y' f_{yy} f_y f_x) + O(h^8), \qquad (3.4)
\end{aligned}$$

*from which we can infer that the PFAFRKN*6-6*ER method has order six.*

3.2. Periodicity interval of the new method

Using the Maple package, from the definition in (6), the following result can be readily obtained.

Proposition 8 *The newly derived method, PFAFRKN*6-6*ER, has* $(-39.11, 0)$ *as the primary interval of periodicity.*

4. Some Numerical Examples

To assess the performance of the new scheme, we have considered the following RKN codes of the same order and stages to get fair comparisons:

- PFAFRKN6-6ER: The constructed adapted RKN code developed here,
- RKN6-6ER: An explicit sixth-order six stage RKN method presented in [7],
- RKN6-6ER-PFAF: An optimized explicit sixth-order six stage RKN method derived by Anastassi and Kosti in [1],
- RKN6-6FM: An explicit sixth-order six stage RKN method developed by Dormand et al. in [6].

We will consider different oscillatory problems appeared in the literature to test the performance of the above methods:

Problem. (Non-linear System in [14])

$$\begin{aligned} y_1'' + w^2 y_1 &= \frac{2y_1 y_2 - \sin(2wx)}{({y_1}^2 + {y_2}^2)^{\frac{3}{2}}}, \quad y_1(0) = 1, y_1'(0) = 0, \\ y_2'' + w^2 y_2 &= \frac{{y_1}^2 - {y_2}^2 - \sin(2wx)}{({y_1}^2 + {y_2}^2)^{\frac{3}{2}}}, \quad y_2(0) = 0, y_2'(0) = w, \quad x \in [0, 4000], \end{aligned}$$

with exact solution given by

$$y_1(x) = \cos(wx), \quad y_2(x) = \sin(wx).$$

To use the adapted methods we have taken the parameter value $w = 5$.

To tests the accuracy of the considered methods we have taken the integration interval $[x_0, x_N] = [0, T]$, with step-length, $h = \frac{x_N - x_0}{N} = \frac{T}{N}$, where the end point takes different values, $T = 100, 1000, 4000$.

The numerical data is given in Tables 2, considering different step-sizes h. The tables contain the maximum absolute errors

$$Max\ Abs\ Err = \max_{n=1,2,\ldots,N} ||y(x_n) - y_n||.$$

Tab. 2 Numerical data corresponding to Problem

h	Methods	T = 100	T = 1000	T = 4000
0.05	PFAFRKN6-6ER	3.802533(-10)	2.155096(-9)	9.277232(-9)
	RKN6-6ER	4.282131(-8)	1.632977(-7)	1.632977(-7)
	RKN6-6ER-PFAF	1.899990(-8)	7.334019(-8)	7.667858(-8)
	RKN6-6FM	1.953175(-7)	7.542634(-7)	7.542634(-7)
0.075	PFAFRKN6-6ER	9.475666(-9)	3.697725(-8)	3.697725(-8)
	RKN6-6ER	7.249395(-7)	2.799656(-6)	2.804762(-6)
	RKN6-6ER-PFAF	1.214422(-5)	4.709903(-5)	4.713322(-5)
	RKN6-6FM	2.236286(-6)	8.631087(-6)	8.634346(-6)
0.1	PFAFRKN6-6ER	9.349917(-8)	3.600327(-7)	3.600327(-7)
	RKN6-6ER	5.491464(-6)	2.106615(-5)	2.106615(-5)
	RKN6-6ER-PFAF	1.223785(-3)	4.727541(-3)	4.727541(-3)
	RKN6-6FM	1.261951(-5)	4.880468(-5)	4.881228(-5)
0.125	PFAFRKN6-6ER	5.305980(-7)	2.048570(-6)	2.048570(-6)
	RKN6-6ER	2.598789(-5)	1.008581(-4)	1.008581(-4)
	RKN6-6ER-PFAF	4.319413(-2)	2.315899(-1)	2.324977(-1)
	RKN6-6FM	4.804257(-5)	1.875787(-4)	1.876650(-4)

In order to show even more the efficiency of the developed PFAFRKN6-6ER code, we present the efficiency curve for the considered problem for $T = 100$. In Figure 1 the logarithm of the maximum absolute global error versus the logarithm of the total number of function evaluations have been plotted. It can be observed the good behavior of the new code.

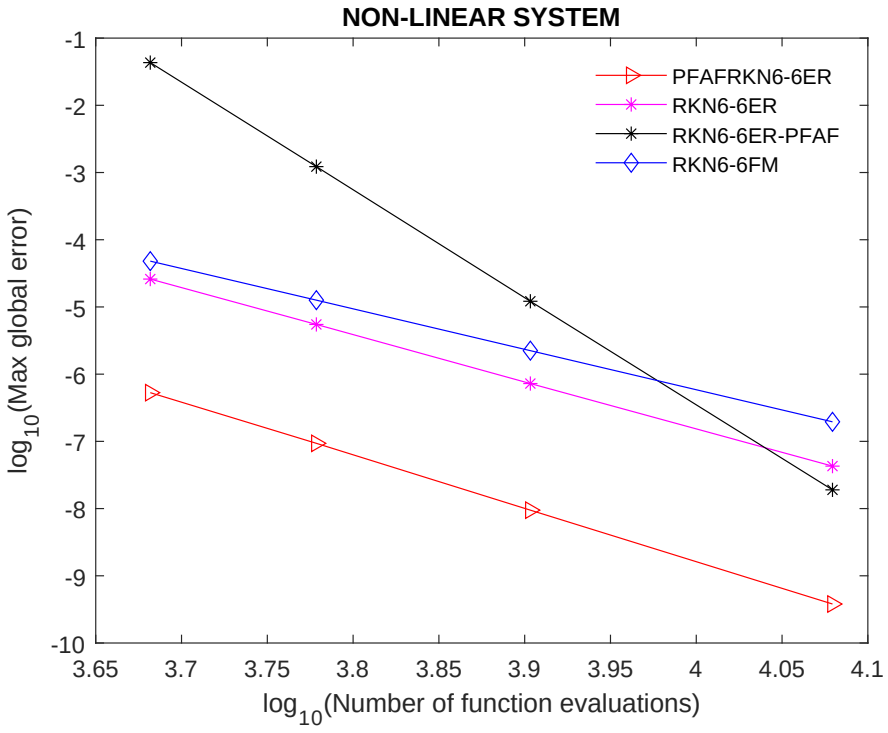


Fig. 1 Efficiency curves corresponding to the Problem

To further demonstrate the efficiency of the constructed PFAFRKN6-6ER code, we present in Figures 2 the logarithms of the maximum absolute global errors versus the CPU time used. It can be observed the good behavior of the new code.

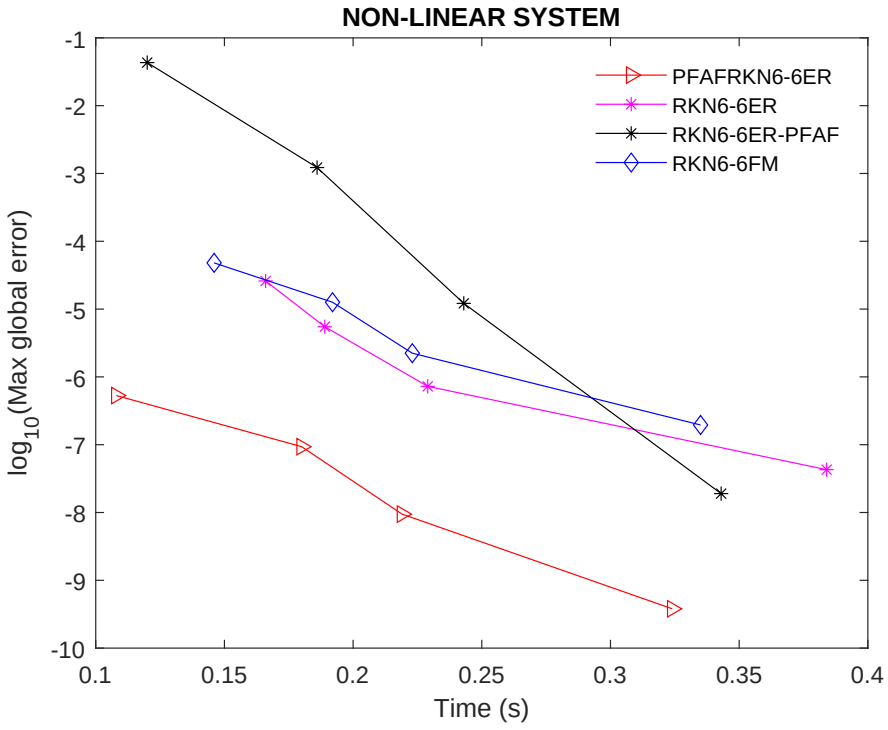


Fig. 2 Efficiency curves corresponding to Problem

5. Discussion

The new PFAFRKN6-6ER code gives minimum error norm, minimum number of function evaluation per steps, and minimum computational cost (Time(s)). Table 2 and Figures $1-2$ put an evidence that PFAFRKN6-6ER is a very efficient scheme. Therefore, we can say that PFAFRKN6-6ER is more appropriate for solving the type of problem in (1.1) than the other existing RKN methods of order 6 with six stages in the literature.

6. Conclusion

In this study, we have used the methodology for constructing the phase-fitted and amplification-fitted methods to develop an efficient new explicit phase- and amplification-fitted RKN code based on the RKN6-6ER method due to El-Mikkawy and Rahmo [7]. The new developed method has two variable coefficients depending on the parameter $\mu = wh$, which is usually known as the parameter frequency [12, 16]. We computed the local truncation error of the new method, confirming that the order of convergence of the underlying code is maintained. In addition, the periodicity interval of the new code has been obtained. The obtained numerical results clearly show that PFAFRKN6-6ER is more accurate and efficient than other sixth-order six-stage RKN codes in the literature.

Acknowledgements

The authors want to thank the Center of Excellence in Theoretical and Computational Science (TaCS-CoE), King Mongkut's University of Technology, Thonburi, Bangkok, Thailand for the financial support. The first author appreciates the support by the Petchra Pra Jom Klao PhD Research Scholarship from KMUTT with Grant No. 15/2562.

References

[1] Z. A. Anastassi and A. A. Kosti. A 6 (4) optimized embedded Runge–Kutta–Nyström pair for the numerical solution of periodic problems. *Journal of Computational and Applied Mathematics*, 275:311–320, 2015.

[2] R. D'Ambrosio, B. Paternoster, and G. Santomauro. Revised exponentially fitted Runge-Kutta-Nyström methods. *Applied Mathematics Letters*, 30:56–60, 2014.

[3] M. A. Demba, N. Senu, and F. Ismail. A four-stage third-order symplectic explicit trigonometrically-fitted Runge-Kutta-Nyström method for the numerical integration of oscillatory initial-value problems. *International Journal of Pure and Applied Mathematics*, 111(2):165–178, 2016.

[4] M. A. Demba, N. Senu, and F. Ismail. New explicit trigonometrically-fitted fourth-order and fifth-order Runge-Kutta-Nyström methods for periodic initial value problems. *International Journal of Pure and Applied Mathematics*, 109(2):207–222, 2016.

[5] M. A. Demba, N. Senu, and F. Ismail. A new fourth-order four stage explicit trigonometrically-fitted Runge-Kutta-Nyström method for solving periodic problems. In *AIP Conference Proceedings*, volume 1739, page 020041. AIP Publishing, 2016.

[6] J. R Dormand, M. E. A El-Mikxawy, and P. J Prince. Families of Runge-Kutta-Nyström formulas. In *IMA Journal of Numerical Analysis*, 1987.

[7] M. El-Mikkawy and E. Rahmo. A new optimized non-FSAL embedded Runge–Kutta–Nyström algorithm of orders 6 and 4 in six stages. *Applied mathematics and computation*, 145(1):33–43, 2003.

[8] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I, Nonstiff Problems, 2nd Edition*. Springer, Berlin, 2008.

[9] A. A Kosti, Z. A. Anastassi, and T. E. Simos. An optimized explicit Runge–Kutta–Nyström method for the numerical solution of orbital and related periodical initial value problems. *Computer Physics Communications*, 183(3):470–479, 2012.

[10] T. Monovasilis, Z. Kalogiratou, and T. E. Simos. Exponentially fitted symplectic Runge-Kutta-Nyström methods. *Appl. Math. Inf. Sci*, 7(1):81–85, 2013.

[11] K. W. Moo, N. Senu, F. Ismail, and M. Suleiman. New phase-fitted and amplification-fitted fourth-order and fifth-order Runge–Kutta–Nyström methods for oscillatory problems. In *Abstract and Applied Analysis*, volume 2013. Hindawi, 2013.

[12] H. Ramos and J. Vigo-Aguiar. On the frequency choice in trigonometrically fitted methods. *Applied Mathematics Letters*, 23(11):1378–1381, 2010.

[13] T. E. Simos. Exponentially-fitted Runge-Kutta-Nyström method for the numerical solution of initial-value problems with oscillating solutions. *Applied Mathematics Letters*, 15(2):217–225, 2002.

[14] H. Van de Vyver. A Runge–Kutta–Nyström pair for the numerical integration of perturbed oscillators. *Computer Physics Communications*, 167(2):129–142, 2005.

[15] P. J. Van der Houwen and B. P. Sommeijer. Diagonally implicit Runge–Kutta–Nyström methods for oscillatory problems. *SIAM Journal on Numerical Analysis*, 26(2):414–429, 1989.

[16] J. Vigo-Aguiar and H. Ramos. On the choice of the frequency in trigonometrically-fitted methods for periodic problems. *Journal of Computational and Applied Mathematics*, 277:94–105, 2015.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 23–30)

The propagation of smallness property and its utility in controllability problems

Jone Apraiz
jone.apraiz@ehu.eus
Universidad del País Vasco, Spain

Abstract

In this article, we would like to transmit the origins and the importance of the propagation of smallness property, delving into its origins and analyzing its utility in parabolic controllability problems. Generally speaking, the propagation of smallness property analyzes and tries to quantify the rate of growth of a function in one domain knowing its values in two other domains related to the first one. We will mention the first historical ideas and results related to this property (harmonic measure, two-constants theorem, Hadamard three-circles theorem), and, then, we will see how we have applied them in order to solve parabolic evolutions' interior and boundary controllability problems.

1. Introduction

In this presentation, we want to introduce the readers to the propagation of smallness property and some of its utilities. To achieve this purpose, we will try to show what the propagation of smallness property is, where it comes from, why it has been useful, and see some of its utilities in controllability problems.

We have divided our presentation into two parts. In the first part, entitled "History of propagation of smallness property", in section 2, we will define the harmonic measure (which could be thought of as one of the origins of the propagation of smallness property), review the origins of our property and sketch out the main results where it started to be analyzed: the two-constant theorem and the Hadamard three-circles theorem. In the second part, entitled "Some applications of the propagation of smallness property", in section 3, we will see the applications of the propagation of smallness property: the extension of the Hadamard three-circles theorem to higher dimensions and other domains, and the application to some specific controllability problems of parabolic evolution equations.

First of all, in order to start with our research and to give the reader an idea of the mathematical concept we are going to talk about, we will give a general definition of what the propagation of smallness property is. The inequality that we can see in the next definition shows what the results of propagation of smallness usually look like, but, as we will see throughout our article, it can have different variations.

Definition 1.1 (Propagation of smallness) Given three subsets of $\mathbb{R}^n$, E, B_1 and B_2, verifying $E \subset B_1 \subset B_2$ and a class of functions $\mathcal{A} \subset C(B_2)$, we say that E is a propagation of smallness set for $\mathcal{A}$ if, for any $u \in \mathcal{A}$, there exists $\alpha = \alpha(E, B_1, B_2) \in (0, 1)$ such that

$$||u||_{\infty,B_1} \leq ||u||^{\alpha}_{\infty,E} ||u||^{1-\alpha}_{\infty,B_2} . \tag{1.1}$$

Inequality (1.1) assures that, having some general boundedness conditions for a function on the smaller and bigger domains, we can establish estimates for the function on the intermediate domain. The name of this property, as we can predict from the above inequality, could come from the fact that the "information of the function" on the smaller domain is propagating or affecting over bigger domains in some sense.

There are some references that sometimes allude to this concept as "rate of growth result" or "quantification of the propagation of smallness" too.

2. History of the propagation of smallness property

We will start with the first part of our study by analyzing where the propagation of smallness property comes from. Looking back at the history of different areas of mathematics, we can appreciate that one of the concepts that started involving the analysis of the rate of growth of functions was the harmonic measure. Therefore, we will start by defining this measure and giving the historical context in which it was developed.

2.1. Harmonic measure

In this section, we will introduce the concept of harmonic measure and its importance for our main topic, the propagation of smallness property. The harmonic measure is a concept that can be considered or defined inside the theory of harmonic functions. But the harmonic measure is very useful in other areas, and it can also be defined using partial differential equations (using an integral that is a solution to the Dirichlet problem) or probabilistic theory (defining it as a boundary hitting distribution of Brownian motion). For the former, we can learn more about it in [3]; and, for the latter, we can look at these Shizuo Kakutani's references, [15] and [16].

If we locate the harmonic measure inside the theory of harmonic functions, we can say that it is connected to estimating the modulus of an analytic function inside a domain given certain bounds on the modulus of the function on the boundary of the domain. This is mainly what the propagation of smallness property pursues and the main reason why the harmonic measure is found in the origins of this property.

If we have to find when and how the harmonic measure was defined for the first time, although as in a lot of areas of mathematics this could be difficult to specify, in many references we look into, it seems that all of them point to Rolf Nevanlinna in the 1920s, see for example [29], [30] and [31]. In Nevanlinna's *Eindeutige Analytische Funktionen* book, [30], mainly in chapters I, II and III's section §1, we can follow the ideas and details of how he worked on and built the harmonic measure that we know nowadays. In this book, Nevanlinna starts by defining the arcs in the unit circle and analyzing conformal mappings. Then, he defines the harmonic measure and studies its properties in the same circle to later extend the same idea to any bounded region on the plane. He also established the principles or properties that the harmonic measure verifies.

Although it was R. Nevanlinna who gave the name to the harmonic measure, we can see in [25], that, some years earlier, Torsten Carleman and Henri Milloux started to use it trying to measure the growth of analytic functions or, in some sense, studying the propagation of smallness property.

In the following paragraphs, we will introduce the harmonic measure giving the definition and some of its main properties using the area of complex analysis or harmonic functions.

First of all, we will recall three definitions that we will use soon.

Definition 2.1 (Jordan curve) We say that a plane curve is a Jordan curve if it is simple and closed or if it is a homeomorphic image of the unit circle.

Definition 2.2 (Jordan domain) A Jordan domain is a domain in a plane bounded by a Jordan curve or a finite number of Jordan curves.

Definition 2.3 Let Ω and Ω' be two open subsets of $\mathbb{C}$. If $f : \Omega \to \Omega'$ is one-to-one, and f and f^{-1} are holomorphic, then, we say that f is a conformal mapping from Ω to Ω'.

Let us now give the exact definition of the harmonic measure that we will use in the next section, 2.2.

Definition 2.4 (Harmonic measure) Let $D \subset \mathbb{C}$ be a domain bounded by a finite number of Jordan curves, Γ. Let $\Gamma = \alpha \cup \beta$, $Int(\alpha) \cap Int(\beta) = \emptyset$, where α and β are finite sets of Jordan curves. We call the harmonic measure of α with respect to D and evaluated at the point z, to the harmonic function with boundary limit 1 at points of α and boundary limit 0 at points of β. It is expressed as $\omega(z, \alpha; D)$ and verifies

$$\lim_{z\to\alpha} \omega(z, \alpha; D) = 1 \quad \text{and} \quad \lim_{z\to\beta} \omega(z, \alpha; D) = 0.$$

The fact that $\omega(z, \alpha; D)$ is a harmonic function is the reason why this measure is called harmonic measure. The existence of the harmonic measure can be proved using the theory and results from/of [1], [7] and [32]. Whereas, the uniqueness of the harmonic function can be proved using the maximum modulus principle and the Lindelöf theorem that can be found in [24].

2.2. The origins of the propagation of smallness property

As we mentioned in the above section, although maybe other mathematicians in other moments were trying to research this property of propagation of smallness, it seems that the topic began to solidify when R. Nevanlinna started with his works on the harmonic measure in the 1920s. Nevertheless, we can also see in T. Carleman's works, for example in [8], that he was pursuing that idea too, trying to study properties of the harmonic measure and finding bounds for it. This fact is very significant because of the big influence that the later developed Carleman inequalities have had on many branches of mathematics where partial differential equations are involved. Actually, these inequalities are very commonly used, for example, in the control theory of partial differential equations.

We can see in [30, chapter IV, section § 2] the principle of monotoneity for the harmonic measure that Carleman proved. The method he used to prove it provides a means to estimate the harmonic measure from above and below. The authors Alexander Ostrowski and Stefan Emanuel Warschawski also employed the results and method used by Carleman, see [33] and [37].

There is another mathematician that was behind the beginnings of the propagation of smallness property, H. Milloux (for example, see [26]). In [30], we can also find the Carleman-Milloux problems that are related to the study of the harmonic measure and its boundedness.

Now, we will see and comment on the main results we know are involved in the origins of the propagation of smallness property.

We have rebuilt the next theorem a bit looking at [14] and [30] (chapter III, section §2), to deduce this version of the two-constants theorem.

Theorem 2.5 (Two-constants theorem) *Suppose that f is a bounded analytic function in a Jordan domain Ω such that $|f(z)| \leq M$ in Ω and $\lim_{z\to\zeta} \sup |f(z)| \leq m < M$ when $z \in \Omega$ and $\zeta \in E \subset \partial\Omega$. Then, for any $z \in \Omega$,*

$$\log |f(z)| \leq \omega(z, E; \Omega) \log m + (1 - \omega(z, E; \Omega)) \log M, \tag{2.1}$$

or similarly,

$$|f(z)| \leq m^{\omega(z,E;\Omega)} M^{1-\omega(z,E;\Omega)}. \tag{2.2}$$

Remark 2.6 In the previous theorem, it's the second inequality, (2.2), which mostly appears in the literature.

There are more versions of the two-constants theorem, for example, L. Ahlfors and S. G. Krantz, in the books *Complex Analysis* and *Geometric Funtcion Theory: Explorations in Complex Analysis*, respectively, [1] and [18], present a comparison tool for the harmonic measure called *the majorization principle*. Then, they prove the two-constants theorem using this principle.

We can also observe in [14] that the two-constants theorem can also be seen as a particular case of a n-constants theorem.

Moving forward with the history of the propagation of smallness, we will introduce the Hadamard three-circles theorem. Before we set forth the theorem, we will remember who discovered it. Jacques Hadamard published this result without a proof in 1896 in *Sur les Fonctions Entières*, [13]. Then, in 1935, it seems that he presented a proof in *Selecta: Jubilé Scientifique de M. Jacques Hadamard*, [12], but, meanwhile, other proofs were given in [5], [9] and [10].

The next version of this classic theorem has been obtained from [14].

Theorem 2.7 (Hadamard three-circles theorem) *Let $f(z)$ be an analytic function in $|z| < R$ and $M(r) = \max_{\theta} |f(r e^{i\theta})|$. If $0 < r_1 \leq r \leq r_2 < R$, then,*

$$\log M(r) \leq \frac{\log r_2 - \log r}{\log r_2 - \log r_1} \log M(r_1) + \frac{\log r - \log r_1}{\log r_2 - \log r_1} \log M(r_2).$$

Remark 2.8 The constant $\frac{\log r_2 - \log r}{\log r_2 - \log r_1}$ is sometimes called the Hadamard exponent and is denoted by $\beta_H\left(\frac{r_1}{r_2}, \frac{r}{r_2}\right)$, see for example [17].

Seeing what the theorem assures, we can realize where its name comes from. The result compares the logarithm of the modulus of a function in three circles of radii r_1, r_2 and r.

There is also a characterization of the Hadamard three-circles theorem using the logarithmically convex property: $M(r)$ is a logarithmically convex function with respect to the variable $\log r$.

3. Some applications of the propagation of smallness property

With this second part of our presentation, we would like to show the readers how the original ideas for the propagation of smallness property have been broadened to different results in higher dimensions and see one of the applications for parabolic evolution equations to controllability problems.

3.1. Three-spheres and three-regions theorems for complex-valued harmonic functions

In this section, we will see three propagation of smallness property results extended to higher dimensions. The different three-spheres type theorems have become very useful and common in many areas of mathematical analysis.

If we have to go back to the origins of the purpose of mathematicians to extend the Hadamard three-circles propagation of smallness result to higher dimensions, one of the mathematicians we will find is Evgenii Landis. He proved a three-spheres theorem for harmonic functions in $\mathbb{R}^n$ and also a more general one for solutions of second order elliptic equations. He published them in 1963 in *Some Problems of the Qualitative Theory of Second Order Elliptic Equation*, [20].

There are other mathematicians that could be the forefathers of the expansion of the three-spheres theorem to higher dimensions and to other equations: Nicola Garofalo and Fang-Hua Lin, Jacob Korevaar and Jan L. H. Meyers, Raymond Brummelhuis or Igor Kukavica (see [11], [17], [6] and [19]).

We have chosen J. Korevaar's and J. L. H. Meyers' results to comment about here because of their similarity to the Hadamard three-circles theorem we have presented in section 2.2. Another reason why we have also chosen their results, is because we have used analogous techniques to theirs to prove the null controllability problems we will see later (theorems 3.7 and 3.8).

In 1994, J. Korevaar and J. L. H. Meyers published the article *Logarithmic Convexity for Supremum Norms of Harmonic Functions*, [17], where they mainly proved that the L^2-version of the three-spheres theorem was valid for harmonic functions in $\mathbb{R}^n$ with the same coefficients as the logarithm terms that appear in the Hadamard theorem. We will see below the two main theorems J. Korevaar and J. L. H. Meyers proved in [17].

We will begin with the first theorem J. Korevaar and J. L. H. Meyers proved in [17], they named it the "three-balls theorem".

Theorem 3.1 (Three-balls theorem) *Suppose $0 < \rho < r < R$ and $n \geq 2$. Then, there exists a constant $\alpha \in (0,1)$, depending only on ρ/R, r/R and n, such that for all complex-valued harmonic functions u on the ball $B(0,R)$ in $\mathbb{R}^n$,*

$$||u||_r \leq ||u||_\rho^\alpha ||u||_R^{1-\alpha},$$

where $||u||_t = \sup |u(x)|$ on the ball $B(0,t)$.

The similarity of this theorem to the Hadamard three-circle theorem is clear. Indeed, in the article [17], they specifically mention Hadamard's theorem as a special case of this.

We will briefly sketch out how J. Korevaar and J. L. H. Meyers proved the three-balls theorem:

- If u is harmonic on the closed ball $\overline{B}(0,R)$ in $\mathbb{R}^n$, then, it can be expressed by the Poisson integral.
- A harmonic function $u(x)$ can also be represented by Laplace series, an orthogonal decomposition using spherical harmonics (see [35] for more details).
- Using Hadamard three-circles theorem and the Laplace series for $u(x)$, it can be proved that $\log ||u||_{r,2}$ is a convex function of $\log r$.
- The maximum principle for harmonic functions gives us a corresponding arithmetic inequality for the sup or uniform norm $||u||_r$ on spheres $S(0,r)$.
- Change scale in the theorem, $R = 1$ so that $0 < \rho < r < 1$.
- We may assume that u is bounded and that $||u||_1 = \sup_B |u| = 1$, because if u is not bounded on $B = B(0,1)$ or if $u \equiv 0$, there is nothing to prove.
- Setting $u_\tau(x) = u(\tau x)$ with $\tau \to 1$, we have that $u(x) = \lim u_\tau(x)$ on B, uniformly on compact subsets. Thus, $||u_\tau||_s \to ||u||_s$ if $s < 1$ while $||u_\tau||_1 = ||u||_\tau \leq ||u||_1$. Hence, it is enough to prove the three-balls theorem for functions (such as u_τ) which are harmonic on the closed unit ball.
- Assuming some nomralization conditions and setting

$$m(\epsilon) = m(\epsilon, \rho, r, n) \overset{def}{=} \sup_{u \in H_\epsilon} ||u||_r = \sup_{u \in H_\epsilon} u(re_1),$$

it follows from the preceding that the three-balls theorem is equivalent to the next proposition.

Proposition 3.2 *There is a constant $\alpha \in (0,1)$ (depending on ρ, r and n) such that*

$$m(\epsilon) \leq \epsilon^\alpha, \quad 0 < \epsilon \leq 1. \tag{3.1}$$

Remark 3.3 Some authors have obtained three-balls type theorems with an additional constant $C = C(\rho/R, r/R, n)$,

$$||u||_r \leq C||u||_\rho^\alpha ||u||_R^{1-\alpha}. \tag{3.2}$$

The first author that proved this kind of inequality could have been E. M. Landis in [21], 1963. He actually proved an inequality (3.2) for solutions of second order linear elliptic partial differential equations. In 1995, R. Brummelhuis, [6], removed the constant C in (3.2) for the partial differential equations case. On the other hand, E. D. Solomentsev obtained a three-spheres theorem for harmonic functions in which the right-hand side of inequality (3.2) also involved the normal derivative of u. This was in 1966, [34].

There is an extension of the three-balls theorem, the three-regions theorem. It was proved by Errett Albert Bishop, [4], in 1963 using the standard covering argument.

Theorem 3.4 (Three-regions theorem) *Let Ω be a connected domain in $\mathbb{R}^n$, $n \geq 2$, $\Omega_0 \subset \Omega$ a nonempty open subset and $E \subset \Omega$ a compact subset (which may be just a point). Then, there is a constant $\alpha \in (0, 1]$ depending only on E, Ω_0 and Ω such that for all complex-valued harmonic functions u on Ω,*

$$\sup_E |u| \leq (\sup_{\Omega_0} |u|)^\alpha (\sup_\Omega |u|)^{1-\alpha}.$$

Now, we will see the second theorem J. Korevaar and J. L. H. Meyers proved in [17], a propagation of smallness property for arbitrary harmonic functions in $\mathbb{R}^n$ with $n \geq 2$. We can observe that the sets where they applied the propagation of smallness property are not concentric.

Theorem 3.5 *Let Ω be a domain in $\mathbb{R}^n$, $\Omega_0 \subset \Omega$ a nonempty subdomain and $E \subset \Omega$ a nonempty compact subset. Then, there is a constant $\alpha = \alpha(E, \Omega_0, \Omega) \in (0, 1]$ such that for all complex-valued harmonic functions u on Ω,*

$$||u||_E \leq ||u||_{\Omega_0}^\alpha ||u||_\Omega^{1-\alpha},$$

where $||u||_A = \sup_{x \in A} |u(x)|$.

Proof Let Ω, Ω_0 and E be subsets of $\mathbb{R}^n$ as in theorem 3.5 and let u be harmonic on Ω. As we want to obtain a constant $\alpha \in (0, 1)$, we may assume that u is bounded and that $||u||_\Omega \leq 1$. We can also assume that Ω is not all of $\mathbb{R}^n$. Otherwise, u would be a constant and there would be nothing to prove.

We will now enclose the compact set E in a finite union $\Omega_p = \Omega_0 \cup B_1 \cup \ldots \cup B_p$, where $\overline{B}_1 \subset \Omega$ is a ball with center in Ω_0 and, in general, B_k ($\overline{B}_k \subset \Omega$) is a ball with center in $\Omega_{k-1} = \Omega_0 \cup B_1 \cup \ldots \cup B_{k-1}$. In order to explain the idea of the proof, we will analyze the case of B_1 and $\Omega = \Omega_0 \cup B_1$. Let V_1 and W_1 be balls concentric with B_1 such that V_1 is maximal in Ω_0 and W_1 is maximal in Ω. Then, by the three-balls theorem, theorem 3.1, there is a constant $\alpha_1 \in (0, 1)$, depending only on the radii of B_1, V_1 and W_1 and on n, such that

$$||u||_{B_1} \leq ||u||_{V_1}^{\alpha_1} ||u||_{W_1}^{1-\alpha_1} \leq ||u||_{\Omega_0}^{\alpha_1}.$$

Since $||u||_{\Omega_0} \leq 1$, we have also that $||u||_{\Omega_0} \leq ||u||_{\Omega_0}^{\alpha_1}$, and therefore, $||u||_{\Omega_1} \leq ||u||_{\Omega_0}^{\alpha_1}$. In the next step, we can similarly prove that

$$||u||_{\Omega_2} \leq ||u||_{\Omega_1}^{\alpha_2} \leq ||u||_{\Omega_0}^{\alpha_1 \alpha_2}.$$

Thus, following the same process, we can obtain

$$||u||_E \leq ||u||_{\Omega_p} \leq ||u||_{\Omega_0}^{\alpha_1 \ldots \alpha_p},$$

where $\alpha_k \in (0, 1)$ are constants depending only on the geometry. Now, putting things together, we have obtained what we wanted to prove,

$$||u||_E \leq ||u||_{\Omega_0}^\alpha ||u||_\Omega^{1-\alpha}.$$

□

We have given the above proof in order to show the *covering by balls' argument* in a simple way, which is the argument they use in the proof of the above theorem.

Remark 3.6 We can observe that, depending on the topic or area they are going to be used in, the hypothesis of three-balls or three-spheres type theorems can change (especially the equation that the function u verifies or the type of the domain we will apply the propagation of smallness to).

3.2. Applications to controllability problems

In this section, we will see one of the propagation of smallness property's utilities in parabolic evolutions' interior and boundary controllability problems. Following similar results and tools as we have seen in the previous sections, in the article *Null-Control and Measurable Sets*, [2], we showed that there are some propagation of smallness properties (two-constants or three-spheres type inequalities) that can drive us to prove the null controllability problems we will present soon for the interior case (theorem 3.7) and for the boundary case (theorem 3.8) of a domain.

The null-controllability problems we present below, are for some time-independent parabolic evolutions with controls acting over measurable sets $\mathcal{D} = \omega \times [0,T]$ and $\mathcal{J} = \gamma \times [0,T]$ with positive measure.

Here, we focus on the heat equation over a smooth and bounded domain Ω in $\mathbb{R}^n$ and for a time interval $(0,T)$, for a distributed control f in the interior case, or for a boundary control h in the boundary case. We set forth the null-controllability for the problems

$$\begin{cases} \triangle u - \partial_t u = f(x,t)\chi_\omega(x), & \text{in } \Omega \times (0,T), \\ u = 0, & \text{on } \partial\Omega \times [0,T], \\ u(0) = u_0, & \text{in } \Omega, \end{cases} \tag{3.3}$$

and

$$\begin{cases} \triangle u - \partial_t u = 0, & \text{in } \Omega \times (0,T), \\ u = h(x,t)\chi_\gamma(x), & \text{on } \partial\Omega \times [0,T], \\ u(0) = u_0, & \text{in } \Omega, \end{cases} \tag{3.4}$$

where $\omega \subset \Omega$ is an interior control region and $\gamma \subset \partial\Omega$ is a boundary control region. In the next theorems, we give a formal account of the null controllability when the control regions ω and γ are measurable sets with positive measure.

Theorem 3.7 *Let $n \geq 2$. Then, $\triangle - \partial_t$ is null-controllable at all positive times with distributed controls acting over a measurable set $\omega \subset \Omega$ with positive Lebesgue measure when*

$$\triangle = \nabla \cdot (\mathbf{A}(x)\nabla \cdot) + V(x),$$

is a self-adjoint elliptic operator, the coefficients matrix $\mathbf{A}$ is smooth in $\overline{\Omega}$, V is bounded in Ω and both are real-analytic in an open neighborhood of ω. The same holds when $n = 1$,

$$\triangle = \frac{1}{\rho(x)}\left[\partial_x\left(a(x)\partial_x\right) + b(x)\partial_x + c(x)\right]$$

and a, b, c and ρ are measurable functions in $\Omega = (0,1)$.

Theorem 3.8 *Let $n \geq 2$. Then, $\triangle - \partial_t$ is null-controllable at all times $T > 0$ with boundary controls acting over a measurable set $\gamma \subset \partial\Omega$ with positive surface measure when*

$$\triangle = \nabla \cdot (\mathbf{A}(x)\nabla \cdot) + V(x)$$

is a self-adjoint elliptic operator, the coefficients matrix $\mathbf{A}$ is smooth in $\overline{\Omega}$, V is bounded in Ω and $\mathbf{A}$, V and $\partial\Omega$ are real-analytic in an open neighborhood of γ in $\overline{\Omega}$.

The results in theorems 3.7 and 3.8 follow from:

- A straightforward application of the linear construction of the control function for the systems (3.3) and (3.4) developed in [22] by Gilles Lebeau and Luc Robbiano.
- For theorem 3.7, a finite number of applications of the following (theorem 3.9) propagation of smallness estimate from measurable sets established in [36] by Sergio Vessella (see also [27] and [28] for other close results).
- A suitable covering argument for the domain Ω (in theorem 3.7) and $\partial\Omega$ (in theorem 3.8).
- For theorem 3.8, a successive iteration of a finite number of applications of the *three-spheres type inequalities* associated with the obvious extension of theorem 3.9 for real analytic functions defined over real analytic hypersurfaces in $\mathbb{R}^{n+1}$

Theorem 3.9 *Assume that $f : B_{2R} \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ is a real-analytic function verifying*

$$|\partial^\alpha f(x)| \leq \frac{M|\alpha|!}{(\rho R)^{|\alpha|}}\ ,\ \textit{when}\ \ x \in B_{2R},\ \alpha \in \mathbb{N}^n, \tag{3.5}$$

for some $M > 0$, $0 < \rho \leq 1$ and $E \subset B_{\frac{R}{2}}$ is a measurable set with positive measure. Then, there are positive constants $N = N(\rho, |E|/|B_R|)$ and $\theta = \theta(\rho, |E|/|B_R|)$ such that

$$\|f\|_{L^\infty(B_R)} \leq N \left(\fint_E |f|\, dx \right)^\theta M^{1-\theta}.$$

Acknowledgements

The author was supported by the Spanish Government's Ministry of Science, Innovation and Universities' grant, PGC2018-094522-B-I00, and the Basque Government's grant, IT12247-19.

References

[1] Lars V. Ahlfors. *Complex Analysis*. McGraw-Hill, New York, 3rd ed., 1979.

[2] Jone Apraiz, Luis Escauriaza. Null-Control and Measurable Sets. *ESAIM: Control, Optimisation and Calculus of Variations,* Tome. 19, no. 1, p. 239-254, 2013.

[3] Christopher J. Bishop. Harmonic measure: algorithms and applications. *Proc. Int. Cong. of Math, Rio de Janeiro*, Vol. 2, (2017) 1507–1534.

[4] Errett A. Bishop. Holomorphic completions, analytic continuation, and the interpolation of semi-norms. *Ann. of Math.* 78, 468–500, 1963.

[5] Otto Blumenthal. Über ganze transzendente Funktionen. *Jber. Deutschen Math. Verein.* 16, 97–109, 1907.

[6] Raymond Brummelhuis. Three spheres theorem for second order elliptic equations. *Journal de Analyse Mathématique*, 65, 179–180, 1995.

[7] Constantin Carathéodory. Über die gegenseitige Beziehung der Ränder bei der konformen Abbildungen des Inneren einer Jordanschen Kurve auf einen Kreis. *Math. Ann.* 73, 305–320, 1913.

[8] Torsten Carleman. Sur les fonctions inverses des fonctions entieres. *Ark. Mat. Astron. Fys.* 15, no. 10, 1921.

[9] Georg Faber. Sur le mode de croissance des fonctions entières. *Bull. Soc. Math. France*, 35, 213–232, 1907.

[10] Georg Faber. Über das Anwachsen analytischer Funktionen. *Math. Ann.* 63, 549–551, 1907.

[11] Nicola Garofalo, Fang-Hua Lin. Monotonicity properties of variational integrals, A_p weights and unique continuation. *Indiana Univ. Math. J.* 35, 245–268, 1986.

[12] Jacques Hadamard. *Selecta: Jubilé Scientifique de M. Jacques Hadamard.* Paris, 1935.

[13] Jacques Hadamard. Sur les fonctions entières. *Bull. Soc. Math*. France, 24, 186-187, 1896.

[14] Einar Hille. *Analytic function theory.* Chelsea Publishing Company, New York, 2nd ed. 1977.

[15] Shizuo Kakutani. On Brownian motion in n-space. *Proc. Imp. Acad. Tokyo*, 20, (9), 648–652, 1944.

[16] Shizuo Kakutani. Two-dimensional Brownian Motion and Harmonic Functions. *Proceedings of the Imperial Academy*, 20, 706–714, 1944.

[17] Jacob Korevaar, Jan L. H. Meyers. Logarithmic convexity for supremum norms of harmonic functions. *Bull. London Math. Soc.* 26, no. 4, 353–362, 1994.

[18] Steve G. Krantz. *Geometric Funtcion Theory, Explorations in Complex Analysis.* Birkhäuser, Springer, 2006.

[19] Igor Kukavica. Nodal volumes for eigenfunction of analytic regular elliptic problems. *J. Anal. Math.* 67, 269–204, 1995.

[20] Evgenii M. Landis. Some problems of the qualitative theory of second order elliptic equation. *Russian Math Surveys*, 18, 1–62, 1963.

[21] Evgenii M. Landis. A three-spheres theorem. *Dokl. Akad. Nauk SSSR* 148, 277–279, 1963 (Russian). English translation in *Soviet Math. Doklady* 4, no. 1, 76–78, 1963.

[22] Gilles Lebeau, Luc Robbiano. Contrôle exact de l'equation de la chaleur. *Commun. Part. Diff. Eq.* **20**, 1, 335-356, 1995.

[23] Gilles Lebeau, Enrique Zuazua. Null controllability of a system of linear thermoelasticity. *Arch. Ration. Mech. An.* **141** (4), 297–329, 1998.

[24] Ernst Leonard Lindelöf. Quelque remarques sur la croissance de la fonction $\zeta(s)$. *Bull. Sci. Math.* 32, 341-356, 1908.

[25] Donald E. Marshall, Carl Sundberg. Harmonic measure and radial projection. *Trans. Am. Math. Soc.* 316, 1, 81–95, 1989.

[26] Henri Milloux. Le theoreme de M. Picard, suites de fonctions holomorphes, fonctions meromorphes et fonctions entieres. *J. Math. pures appl.* 9, 3, 1924.

[27] Nikolai S. Nadirashvili. A generalization of Hadamard's three circles theorem. *Moscow Univ. Math. Bull.* **31**, 3, 30–32, 1976.

[28] Nikolai S. Nadirashvili. Estimation of the solutions of elliptic equations with analytic coefficients which are bounded on some set. *Vestnik Moskov. Univ. Ser. I Mat. Mekh.* **102**, 2, 42-46, 1979.

[29] Rolf Nevanlinna. Das harmonische Mass von Punktmengen und seine Anwendung in der Funktionentheorie. *Comptes rendus du huitème congrès des mathématiciens scandinaves*, Stockholm, 116–133, 1934..

[30] Rolf Nevanlinna. *Eindeutige analytische Funktionen. Grundlehren Math. Wiss. Springer-Verlag, Berlin, 46, 1936. English translation of second edition,* Analytic functions. Grundlehren Math. Wiss. Springer-Verlag, Berlin, 162, 1970.

[31] Frithiof Nevanlinna, Rolf Nevanlinna. Ueber die Eigenschaften einer analytischen Funktion in der Umgebung einer singul´aren Stelle oder Linie. *Acta Soc. Sci. Fennica*, 50, 5, 1–46, 1922.

[32] William F. Osgood, E. H. Taylor. Conformal transformations on the boundaries of their regions of definition. *Trans. Amer. Math. Soc.* 14, No. 2, 277–298, 1913.

[33] Alexander Ostrowski. Über quasianalytische Funktionen und Bestimmtheit asymptotischer Entwicklungen. *Acta math.* 53, 1929.

[34] E. D. Solomentsev. A three-spheres theorem for harmonic functions. *Akad. Nauk Armyan. SSR Dokl.* 42, 274–278, 1966 (Russian).

[35] Elias M. Stein, Guido Weiss. *Introduction to Fourier Analysis on Euclidean Spaces.* Princeton University Press, Princeton, N.J., 1971.

[36] Sergio Vessella. A continuous dependence result in the analytic continuation problem. *Forum Math.* **11**, 6, 695-703, 1999.

[37] Stefan E. Warschawski. Über das Randverhalten der Ableitung der Abbildungsfunktion bei konformer Abbildung. *Math. Z.* 35, 1932.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 31–38)

Theoretical and numerical results for some inverse problems for PDEs

Jone Apraiz[1], **Anna Doubova**[2], **Enrique Fernández-Cara**[3], **Masahiro Yamamoto**[4]
1. jone.apraiz@ehu.eus *Universidad del País Vasco, Spain*
2. doubova@us.es *Universidad de Sevilla, Spain*
3. cara@us.es *Universidad de Sevilla, Spain*
4. myama@next.odn.ne.jp *University of Tokyo, Japan*

Abstract

We consider geometric inverse problems concerning the one-dimensional Burgers equation and some related nonlinear systems (involving heat effects and variable density). In these problems, the goal is to find the size of the spatial interval from some appropriate boundary observations of the solution. Depending on the properties of the initial and boundary data, we prove uniqueness and non-uniqueness results. On the other hand, we also solve these inverse problems numerically and compute approximations of the interval sizes.

1. Introduction

In this work we will deal with inverse problems for some nonlinear time-dependent PDEs in one spatial dimension. The analysis and solution of inverse problems of many kinds has recently increased a lot because of their relevance in many applications: elastography and medical imaging, seismology, potential theory, ion transport problems or chromatography, finances, etc.; see for instance [5]. The variety of inverse problems is huge in comparison with their direct analogs and many inverse problems coming from very classical and basic direct problems wait for theoretical and numerical research (see for example [3] and [8]) where many theoretical and numerical aspects of inverse problems for partial differential equations are depicted.

In this paper, we consider problems related to the identification of the size of the spatial interval where a time-dependent governing equation must be satisfied. We will focus on the Burgers equation and some variants, satisfied for $(x,t) \in (0,\ell) \times (0,T)$. We will assume that the equation is complemented with boundary and initial conditions corresponding to known data, respectively for $x \in \{0,\ell\}$ and $t = 0$. Then, we will try to determine the width ℓ of the spatial x-interval from some extra information, for instance given at $x = 0$. The main goals will be to establish uniqueness and to compute approximations of the solutions to the inverse problems. Related questions have been analyzed recently for the linear heat and wave equations in [1].

The plan is the following. In Section 2, we consider the viscous Burgers equation under several different circumstances. Section 3 and Section 4 respectively deal with the Burgers equation coupled to a heat equation and the variable density Burgers system. Finally, we present the results of some numerical experiments in Section 5.

2. Some positive and negative results for the viscous Burgers equation

Let us consider the following system for the Burgers equation:

$$\begin{cases} u_t - u_{xx} + uu_x = 0, & 0 < x < \ell,\ 0 < t < T, \\ u(0,t) = \eta(t),\ \ u(\ell,t) = 0, & 0 < t < T, \\ u(x,0) = u_0(x), & 0 < x < \ell. \end{cases} \tag{2.1}$$

We will analyze the following question:

Uniqueness for Burgers equation: *Let u^ℓ and u^L be the solutions to* (2.1) *associated to the spatial intervals* $(0,\ell)$ *and* $(0,L)$*, respectively. Assume that the corresponding observations $u^\ell_x(0,\cdot)$ and $u^L_x(0,\cdot)$ coincide, that is,*

$$u^\ell_x(0,t) = u^L_x(0,t) \ \ in \ \ (0,T).$$

Then, do we have $\ell = L$?

In the sequel, we will provide some positive and negative answers to this question, depending on the kind of imposed boundary or initial data.

 ISBN 978-84-18482-21-2

2.1. The simplest cases: zero initial and/or boundary data

Case I: $\eta \not\equiv 0$ **and** $u_0 \equiv 0$. If $u_0 \equiv 0$, we get uniqueness:

Theorem 2.1 *Assume that* $0 < \ell \leq L < T$, $\eta \in L^\infty(0,T)$ *satisfies* $\eta \not\equiv 0$ *and* $u_0 \equiv 0$. *Let* u^ℓ *and* u^L *be the solutions to* (2.1) *respectively corresponding to* ℓ *and* L *and let us assume that*

$$|u_x^\ell(x,t)| \leq M \ \text{ in } \ (0,\ell)\times(T_0,T), \ \ |u_x^L(x,t)| \leq M \ \text{ in } \ (0,L)\times(T_0,T),$$

where $M > 0$ *and* $u_x^\ell(0,t) = u_x^L(0,t)$ *in* $(0,T)$. *Then,* $\ell = L$.

The proof is standard and it is based on a suitable unique continuation property. For the details, see [2].

Case II: $\eta \equiv 0$ **and** $u_0 \not\equiv 0$. In this case, we can show that, as in the case of the linear heat equation (see [1]), non-uniqueness holds in general. A counter-example to uniqueness can be found (see [2]).

2.2. Results where $\eta(t) \not\equiv 0$ and $u_0(x) \not\equiv 0$

In this case we have the following result, which proof can be found in [2].

Theorem 2.2 *Assume that* $0 < \ell \leq L \leq L_*$, $0 < T_0 < T$,

$$u_x^\ell(0,t) = u_x^L(0,t) \ \text{ in } \ (0,T), \ \ \|u_0\|_{L^2(0,L)} \leq M_0,$$
$$|u_x^\ell(x,t)| \leq M \ \text{ in } \ (0,\ell)\times(T_0,T) \ \text{ and } \ |u_x^L(x,t)| \leq M \ \text{ in } \ (0,L)\times(T_0,T),$$

where L_*, M_0 *and* M *are some positive constants. There exists* δ_0 (*only depending on* L_*, T_0, T, M_0 *and* M) *such that, if*

$$\int_{T_0}^{T} |\eta(t)|^2\, dt \geq \delta_0, \tag{2.2}$$

then one necessarily has $\ell = L$.

3. The Burgers equation with heat effects

Given $k \in \mathbb{R}$, the system is now

$$\begin{cases} u_t - u_{xx} + uu_x = k\theta, & 0 < x < \ell,\ t > 0, \\ \theta_t - \theta_{xx} + u\theta_x = 0, & 0 < x < \ell,\ t > 0, \\ u(0,t) = \eta(t), \ \ u(\ell,t) = 0, & t > 0, \\ \theta(0,t) = \lambda(t), \ \ \theta(\ell,t) = 0, & t > 0, \\ u(x,0) = u_0(x), \ \ \theta(x,0) = \theta_0(x), & 0 < x < \ell. \end{cases} \tag{3.1}$$

The uniqueness property we will analyze here in the following:

Uniqueness for Burgers equation with heat effects: *Let* (u^ℓ, θ^ℓ) *and* (u^L, θ^L) *be the solutions to* (3.1) *associated to the spatial intervals* $(0,\ell)$ *and* $(0,L)$, *respectively. Assume that the corresponding observations* $(u_x^\ell(0,\cdot), \theta_x^\ell(0,\cdot))$ *and* $(u_x^L(0,\cdot), \theta_x^L(0,\cdot))$ *coincide, that is,*

$$u_x^\ell(0,t) = u_x^L(0,t) \ \text{ and } \ \theta_x^\ell(0,t) = \theta_x^L(0,t) \ \text{ in } \ (0,T).$$

Then, do we have $\ell = L$*?*

If $(u_0, \theta_0) \equiv (0,0)$, we have again uniqueness:

Theorem 3.1 *Assume that* $0 < \ell \leq L < T$, $\eta, \lambda \in L^\infty(0,T)$ *satisfy* $(\eta,\lambda) \not\equiv (0,0)$ *and* $(u_0,\theta_0) \equiv (0,0)$. *Let* (u^ℓ,θ^ℓ) *and* (u^L,θ^L) *be the solutions to* (3.1) *respectively corresponding to* ℓ *and* L *and let us assume that* $|u_x^\ell(x,t)| \leq M$ *in* $(0,\ell)\times(T_0,T)$, $|u_x^L(x,t)| \leq M$ *in* $(0,L)\times(T_0,T)$, $u_x^\ell(0,\cdot) = u_x^L(0,\cdot)$ *and* $\theta_x^\ell(0,\cdot) = \theta_x^L(0,\cdot)$. *Then,* $\ell = L$.

The proof is very similar to the proof of Theorem 2.1 (see [2]).

On the other hand, it is obvious that the solution to (2.1) is a particular solution to (3.1), corresponding to $\theta_0 \equiv 0$ and $\lambda \equiv 0$. Consequently, the counter-example mentioned before is also a counter-example to uniqueness in this case when we allow u_0 to be nonzero.

To our knowledge, it is unknown if a counter-example to uniqueness can also be found with $\theta_0 \not\equiv 0$.

As before, we can deduce a uniqueness result for (3.1) for large η. More precisely, the following holds:

Theorem 3.2 *Assume that* $0 < \ell \leq L \leq L_*$, $0 < T_0 < T$,

$$u_x^\ell(0,t) = u_x^L(0,t) \ \text{ and } \ \theta_x^\ell(0,t) = \theta_x^L(0,t) \ \text{ in } \ (0,T), \ \ \|(u_0,\theta_0)\|_{L^2(0,L)} \leq M_0,$$
$$|u_x^\ell(x,t)| \leq M \ \text{ in } \ (0,\ell)\times(T_0,T) \ \text{ and } \ |u_x^L(x,t)| \leq M \ \text{ in } \ (0,L)\times(T_0,T).$$

There exists δ_1 (*only depending on* L_*, T_0, T, M_0 *and* M) *such that, if*

$$\int_{T_0}^{T} |\eta(t)|^2\,dt \geq \delta_1, \tag{3.2}$$

then one necessarily has $\ell = L$.

The proof is similar to the proof of Theorem 2.2 (see [2]).

4. The case of the variable density Burgers equation

This is more interesting, but also more difficult. We consider a non-homogeneous (or variable density) one-dimensional fluid, modeled as follows:

$$\begin{cases} \rho(u_t + uu_x) - u_{xx} = 0, & 0 < x < \ell,\ t > 0,\\ \rho_t + u\rho_x = 0, & 0 < x < \ell,\ t > 0,\\ u(0,t) = \overline{u}(t),\ \ u(\ell,t) = 0, & t > 0,\\ \rho(0,t) = \overline{\rho}(t), & t \in \mathbb{R}_+ \cap \{t : \overline{u}(t) > 0\},\\ u(x,0) = u_0(x),\ \ \rho(x,0) = \rho_0(x), & 0 < x < \ell. \end{cases} \tag{4.1}$$

Of course, this can be viewed as a toy model for the variable density Navier-Stokes system. The corresponding inverse problem is the following:

This is the uniqueness question we are interested in:

Uniqueness for variable density Burgers equation: *Let* (u^ℓ, ρ^ℓ) *and* (u^L, ρ^L) *be the solutions to* (4.1) *respectively associated to* $(0,\ell)$ *and* $(0,L)$. *Assume that the corresponding* $(u_x^\ell(0,\cdot), \rho^\ell(0,\cdot))$ *and* $(u_x^L(0,\cdot), \rho^L(0,\cdot))$ *coincide. Then, do we have* $\ell = L$*?*

4.1. A result for zero initial data

When the initial data vanish, we have a positive uniqueness result for this problem:

Theorem 4.1 *Assume that* $0 < \ell \leq L$, $T > 0$ *and* (u_0, ρ_0) *and* $(\overline{u}, \overline{\rho})$ *satisfy*

$$\begin{cases} \overline{u}, \overline{\rho} \in L^\infty(0,T),\ \ \overline{u} \not\equiv 0,\ \ \overline{\rho} \geq 0,\\ u_0 \equiv 0,\ \ \rho_0 \in L^\infty(0,L),\ \ \rho_0 \geq a_0 > 0. \end{cases}$$

Let (u^ℓ, ρ^ℓ) *and* (u^L, ρ^L) *be the solutions to* (4.1) *for* $0 < t < T$ *respectively corresponding to* ℓ *and* L *and let us assume that* $|u_t^\ell| + |u_x^\ell| + |\rho_x^\ell| \leq M$ *and* $|u_t^L| + |u_x^L| + |\rho_x^L| \leq M$ *respectively in* $(0,\ell)\times(T_0,T)$ *and* $(0,L)\times(T_0,T)$ *and* $u_x^\ell(0,\cdot) = u_x^L(0,\cdot)$. *Then,* $\ell = L$.

For the proof, we use a unique continuation property satisfied for the solutions to systems of the form

$$\begin{cases} a(x,t)v_t - v_{xx} + b(x,t)v_x + c(x,t)v + d(x,t)p = 0, & (x,t) \in Q,\\ p_t + m(x,t)p_x + r(x,t)v = 0, & (x,t) \in Q, \end{cases} \tag{4.2}$$

where we assume that $Q := (0,\ell)\times(0,T)$,

$$a, b, c, d, m, r \in C^0(\overline{Q}) \ \text{ and } \ a \geq a_0 > 0 \ \text{ in } \ Q. \tag{4.3}$$

More precisely, the following holds:

Proposition 4.2 *Assume that* (4.3) *is satisfied and* (v,p) *solves* (4.2), *with* $v, v_x, v_{xx}, p, p_x \in C^0(\overline{Q})$. *Also, assume*

$$\begin{cases} v(0,t) = 0,\ \ v_x(0,t) = 0,\ \ p(0,t) = 0, & 0 < t < T,\\ v(x,0) = 0,\ \ p(x,0) = 0, & 0 < x < \ell. \end{cases} \tag{4.4}$$

Then, one has $v \equiv 0$ *and* $p \equiv 0$.

The proof of this Proposition relies on appropriate local Carleman inequalities for the solutions to (4.2) and is postponed to Section 4.2.

Proof of Theorem 4.1: Note that $u^\ell \in L^\infty((0,\ell)\times(0,T))$ and $u^L \in L^\infty((0,L)\times(0,T))$. If we set $v := u^\ell - u^L$ and $p := \rho^\ell - \rho^L$, one has

$$\begin{cases} \rho^\ell v_t - v_{xx} + \rho^\ell u^\ell v_x + \rho^\ell u^\ell_x v - (u^L_t + u^L u^L_x)p = 0, & 0 < x < \ell,\ t > 0, \\ p_t + u^\ell p_x + \rho^L_x v = 0, & 0 < x < \ell,\ t > 0, \\ v(0,t) = 0,\ \ v_x(0,t) = 0,\ \ p(0,t) = 0, & t > 0, \\ v(x,0) = 0,\ \ p(x,0) = 0, & 0 < x < \ell. \end{cases}$$

Consequently, v and p satisfies (4.2) with $a = \rho^\ell$, $b = \rho^\ell u^\ell$, $c = \rho^\ell u^\ell_x$, $d = -(u^L_t + u^L u^L_x)$, $m = u^\ell$ and $r = \rho^L_x$ and (4.4).

In view of Proposition 4.2, one has $v = 0$ and $p = 0$ in $(0,\ell)\times(0,T)$. This yields $u^L(x,t) = 0$ in $(\ell, L)\times(0,T)$. Since the PDEs satisfied by u^L and ρ^L also possess the unique continuation property, we find that $u^L \equiv 0$, which is impossible. □

It would be interesting to find nonzero initial data (u_0, ρ_0) such that uniqueness fails. On the other hand, it would also be interesting to prove a result similar to Theorem 3.2 asserting that, if the boundary data are large enough (with respect to the other data in the system), uniqueness is satisfied. However, to our knowledge this is unknown.

A still more complex situation is found when we deal with a variable density fluid where thermal effects are relevant. For example, we can consider the variable density Boussinesq-like system

$$\begin{cases} \rho(u_t + uu_x) - u_{xx} = \theta, & 0 < x < \ell,\ t > 0, \\ \rho(\theta_t + u\theta_x) - \theta_{xx} = u_x^2, & 0 < x < \ell,\ t > 0, \\ \rho_t + u\rho_x = 0, & 0 < x < \ell,\ t > 0, \\ u(0,t) = \overline{u}(t),\ \ u(\ell,t) = 0, & t > 0, \\ \rho(0,t) = \overline{\rho}(t), & t \in \mathbb{R}_+ \cap \{t : \overline{u}(t) > 0\}, \\ \theta_x(0,t) = \theta_x(\ell,t) = 0, & t > 0, \\ \rho(x,0) = \rho_0(x),\ \ u(x,0) = u_0(x),\ \ \theta(x,0) = \theta_0(x), & 0 < x < \ell. \end{cases} \tag{4.5}$$

This is the related inverse problem: (u_0, θ_0, ρ_0) and $(\overline{u}, \overline{\rho})$ are given and the additional observations $\beta := u_x|_{x=0}$ and $\zeta := \theta|_{x=0}$ are known for $t \in (0,T)$ and we try to find ℓ.

A result similar to Theorem 4.1 can also be established in this case. The details are left to the reader.

4.2. Proof of Proposition 4.2

The proof of Proposition 4.2 can be obtained by combining two Carleman inequalities that can be deduced for the solutions to the first and the second equation in (4.2). The main steps are the following: to choose a suitable weight function (the same in both inequalities); to argue as in [9] and [7], then, to sum and eliminate all undesirable terms in the right hand side.

Step 1: Let us first recall some known Carleman estimates for the solutions to equations like in (4.2).

Thus, assume that a, b and c are as in Proposition 4.2 and set $Lv := av_t - v_{xx} + bv_x + cv$ for any suitable v. For any $\lambda > 0$, $\beta > 0$, $x_0 > \ell$, $\delta > 0$ and $T > 0$ (to be definitively fixed below), we take

$$\varphi(x,t) := e^{\lambda\psi(x,t)}, \ \text{ with } \ \psi(x,t) := |x - x_0|^2 - \frac{2\delta\beta}{T}|t - T/2|. \tag{4.6}$$

Note that φ can be used in the proof of the Carleman inequality in Theorem 2.1 in [9, Ch. 4]. Consequently, the following holds:

Theorem 4.3 *There exists $\lambda_0 > 0$ with the following property: for any $\lambda \geq \lambda_0$, there exist constants $s_0 = s_0(\lambda) > 0$ and $C_0 = C_0(\lambda)$ such that*

$$\begin{aligned} &\iint_Q \left(\frac{1}{s\varphi}(|v_t|^2 + |v_{xx}|^2) + s\lambda^2\varphi|v_x|^2 + s^3\lambda^4\varphi^3|v|^2\right) e^{2s\varphi}\,dx\,dt \\ &\quad \leq C_0\left(\iint_Q |Lv|^2 e^{2s\varphi}\,dx\,dt + \int_0^T \left(s^3\lambda^3\varphi^3|v|^2 + s\lambda\varphi|v_x|^2 + |v_t|^2\right)e^{2s\varphi}\,dt\Big|_{x=0,\ell}\right. \\ &\quad \left. + s^2\lambda^2 e^{C_0\lambda}\int_0^\ell \left(|v|^2 + |v_x|^2\right)e^{2s\varphi}\,dx\Big|_{t=0,T}\right) \end{aligned} \tag{4.7}$$

for all $s \geq s_0$ and any $v \in H^{2,1}(Q)$.

Now, let m be as in (4.3) and let us set $B := \varphi_t + m\varphi_x$ and $Ep := p_t + mp_x$ for any p. We can also adapt the proof of the Carleman estimate for transport equations in Proposition 2.1 in [9, Ch. 3] and deduce the following:

Theorem 4.4 *Assume that* $\min_{(x,t)\in\overline{Q}} |B(x,t)| \geq B_0 > 0$. *Then, there exist constants* $s_0 > 0$ *and* $C > 0$ *such that*

$$s^2 \iint_Q \Big|p|^2 e^{2s\varphi}\, dx\, dt \leq C \iint_Q |Ep|^2 e^{2s\varphi}\, dx\, dt + s \int_0^T mB|p|^2 e^{2s\varphi}\, dt\Big|_{x=0}^{x=\ell} + s \int_0^\ell B|p|^2 e^{2s\varphi}\, dx\Big|_{t=0}^{t=T} \tag{4.8}$$

for all $s \geq s_0$ *and any* $p \in H^1(Q)$.

Step 2: Let us assume that $t_0 \in (0,T)$ and $\delta > 0$ is such that $0 < t_0 - \delta < t_0 + \delta < T$ and let us set

$$Q_\delta := (0,\ell) \times (t_0 - \delta, t_0 + \delta).$$

Let us introduce the new variable $\widetilde{t}$ with $\widetilde{t} = t_0 - \delta + \dfrac{2\delta}{T} t$ and the new function $\widetilde{\varphi}$ with

$$\widetilde{\varphi}(x,\widetilde{t}) := e^{\lambda\widetilde{\psi}(x,\widetilde{t})} \text{ and } \widetilde{\psi}(x,\widetilde{t}) := \psi(x,t) \equiv |x - x_0|^2 - \beta|\widetilde{t} - t_0|.$$

Then, (4.7) can be rewritten as an estimate in Q_δ. If we denote $\widetilde{t}$ (resp. $\widetilde{\varphi}$) again by t (resp. φ), the following is found:

$$\iint_{Q_\delta} \Big(\frac{1}{s\varphi}(|v_t|^2 + |v_{xx}|^2) + s\lambda^2\varphi|v_x|^2 + s^3\lambda^4\varphi^3|v|^2\Big)e^{2s\varphi}\, dx\, dt \leq C\left(\iint_Q |p|^2 e^{2s\varphi}\, dx\, dt + K_1 + K_2\right), \tag{4.9}$$

where

$$\begin{aligned} K_1 &:= \int_{t_0-\delta}^{t_0+\delta} \Big(s^3\lambda^3\varphi^3|v|^2 + s\lambda\varphi|v_x|^2 + |v_t|^2\Big)e^{2s\varphi}\, dt\Big|_{x=0,\ell} \\ &\leq Cs^3\lambda^3 e^{C\lambda} \int_{t_0-\delta}^{t_0+\delta} \Big(|v(0,t)|^2 + |v_x(0,t)|^2 + |v_t(0,t)|^2\Big)e^{2s\varphi(0,t)}\, dt \\ &\quad + Cs^3\lambda^3 e^{C\lambda} M^2 \int_{t_0-\delta}^{t_0+\delta} e^{2s\varphi(\ell,t)}\, dt \end{aligned} \tag{4.10}$$

and

$$K_2 := Cs^2\lambda^2 e^{C\lambda} \int_0^\ell \Big(|v|^2 + |v_x|^2\Big)e^{2s\varphi}\, dx\Big|_{t=t_0-\delta,t_0+\delta} \leq Cs^2\lambda^2 e^{C\lambda} M^2 e^{2se^{\lambda(|x_0|^2-\beta\delta)}}. \tag{4.11}$$

On the other hand, (4.8) written in Q_δ and applied to the second equation of (4.2) gives:

$$s^2 \iint_{Q_\delta} |p|^2 e^{2s\varphi}\, dx\, dt \leq C \iint_{Q_\delta} |v|^2 e^{2s\varphi}\, dx\, dt + s \int_{t_0-\delta}^{t_0+\delta} mB|p|^2 e^{2s\varphi}\Big|_{x=0}^{x=\ell} dt + s \int_0^\ell B|p|^2 e^{2s\varphi}\, dx\Big|_{t=t_0-\delta}^{t=t_0+\delta}$$

and we find that

$$s^2 \iint_{Q_\delta} \Big|p|^2 e^{2s\varphi}\, dx\, dt \leq C \iint_{Q_\delta} |v|^2 e^{2s\varphi}\, dx\, dt + R_1 + R_2, \tag{4.12}$$

where

$$\begin{aligned} R_1 &:= Cse^{C\lambda} M^2 \int_{t_0-\delta}^{t_0+\delta} |p|^2 e^{2s\varphi}\, dt\Big|_{x=0}^{x=\ell} \\ &\leq Cse^{C\lambda} M^2 \int_{t_0-\delta}^{t_0+\delta} |p(0,t)|^2 e^{2s\varphi(0,t)}\, dt + Cse^{C\lambda} M^4 \int_{t_0-\delta}^{t_0+\delta} e^{2s\varphi(\ell,t)}\, dt \end{aligned} \tag{4.13}$$

and

$$R_2 := Cse^{C\lambda} M \int_0^\ell |p|^2 e^{2s\varphi}\, dx\Big|_{t=t_0-\delta}^{t=t_0+\delta} \leq Cse^{C\lambda} M^3 e^{2se^{\lambda(|x_0|^2-\beta\delta)}}. \tag{4.14}$$

In (4.10), (4.11), (4.13) and (4.14), we have used that $|v| + |v_x| + |v_t| + |p| \leq M$ in $\overline{Q}$. It is not restrictive to assume that $M \geq 1$.

Step 3: After adding (4.9) and (4.12), taking into account the estimates of the K_i and R_i and assuming that s and λ are sufficiently large, we find:

$$\begin{aligned} &\iint_{Q_\delta} \Big(\frac{1}{s\varphi}(|v_t|^2 + |v_{xx}|^2) + s\lambda^2\varphi|v_x|^2 + s^3\lambda^4\varphi^3|v|^2\Big)e^{2s\varphi}\, dx\, dt + s^2 \iint_{Q_\delta} \Big|p|^2 e^{2s\varphi}\, dx\, dt \\ &\quad \leq Cs^3\lambda^3 e^{C\lambda} M^2 \int_{t_0-\delta}^{t_0+\delta} \Big(|v(0,t)|^2 + |v_x(0,t)|^2 + |v_t(0,t)|^2 + |p(0,t)|^2\Big)e^{2s\varphi(0,t)}\, dt \\ &\quad\quad + Cs^3\lambda^3 e^{C\lambda} M^4 \int_{t_0-\delta}^{t_0+\delta} e^{2s\varphi(\ell,t)}\, dt + Cs^2\lambda^2 e^{C\lambda} M^3 e^{2se^{\lambda(|x_0|^2-\beta\delta)}}. \end{aligned} \tag{4.15}$$

Now, we argue as follows:

- First, we fix $\lambda > 0$ such that (4.15) holds and choose x_0, t_0 and δ as before and $\varepsilon \in (0, \ell)$.
- Then, we take $\beta > 0$ large enough and such that $\beta\delta/2 > \ell x_0 + \ell^2$.
- Finally, we choose $\kappa \in (0, \delta/2)$ such that $\beta\kappa < 2\varepsilon(x_0 - \ell) + \varepsilon^2$.

With these constants ε and κ, one has

$$|x-x_0|^2-\beta|t-t_0| \geq \mu := |x_0-\ell+\varepsilon|^2-\beta\kappa > \max(|x_0-\ell|^2, |x_0|^2-\beta\delta) \tag{4.16}$$

for all $(x,t) \in (0, \ell-\varepsilon) \times (t_0-\kappa, t_0+\kappa)$. Taking into account (4.4), we deduce from (4.15) that

$$\begin{aligned} &\iint_{(0,\ell-\varepsilon)\times(t_0-\kappa,t_0+\kappa)} \left(s\lambda^4 e^{3\lambda\mu}|v|^2 + |p|^2\right) dx\, dt \\ &\qquad \leq 2\delta C s\lambda^3 e^{C\lambda} M^4 e^{2s(e^{\lambda|x_0-\ell|^2}-e^{\lambda\mu})} + Cs\lambda^2 e^{C\lambda} M^3 e^{2s(e^{\lambda(|x_0|^2-\beta\delta)}-e^{\lambda\mu})} \\ &\qquad \leq C_* s \left(e^{2s(e^{\lambda|x_0-\ell|^2}-e^{\lambda\mu})} + e^{2s(e^{\lambda(|x_0|^2-\beta\delta)}-e^{\lambda\mu})}\right), \end{aligned} \tag{4.17}$$

where C_* depends on M, δ and λ but is independent of s. But, in view of (4.16), this right hand side goes to zero as $s \to +\infty$. Consequently, $v(x,t) = 0$ and $p(x,t) = 0$ in $(0, \ell-\varepsilon) \times (t_0-\kappa, t_0+\kappa)$.

Since ε and κ are arbitrarily small and t_0 is arbitrary in $(0,T)$, $v \equiv 0$ and $p \equiv 0$ and the proof is achieved.

5. Some numerical results

We will present here some numerical results for the inverse problems for the Burgers equation.

Reformulation of inverse problem: *Given $T > 0$, $\eta = \eta(t)$, $u_0 = u_0(x)$ and $\beta = \beta(t)$, find $\ell \in (\ell_0, \ell_1)$ such that*

$$J_1(\ell) \leq J_1(\ell') \quad \forall\, \ell' \in (\ell_0, \ell_1), \tag{5.1}$$

where J is given by

$$J_1(\ell) := \frac{1}{2}\int_0^T |\beta(t) - u_x^\ell(0,t)|^2\, dt. \tag{5.2}$$

Here, u^ℓ is the state, i.e. the solution to (2.1) *corresponding to the unknown length ℓ.*

Three different situations will be analyzed for the Burgers equation. In the first two cases, we will check that uniqueness holds: zero initial data and nonzero initial data and sufficiently large η. In the third case we will consider a non-uniqueness situation corresponding to some nonzero initial data and "small" η and we will study the behavior of the numerical algorithm. To this purpose, we will implement the `fmincon` function from the MatLab Optimization Toolbox using the `active-set` minimization algorithm.

Case 1.1: Burgers equation with $u_0 = 0$ and $\eta \neq 0$. We take $T = 5$, $\eta(t) = 5\sin^3 t$ in $(0,T)$ and $u_0(x) \equiv 0$. Starting from $L_i = 3$, our goal is to recover the desired value of the length $L_d = 2$.

The results of this numerical experiments can be seen in Table 1, where the effect of random noise on the target are shown. The computed length is denoted by L_c. The corresponding solution to (5.1)–(5.2) is displayed in Figure 1. The evolution of the iterates and the cost in the minimization process in the absence of random noise appear in Figures 2 and 3, respectively.

% noise	Cost	Iterates	Computed L_c
1%	1.e-3	12	1.997140631
0.1%	1.e-5	15	1.999169558
0.01%	1.e-7	11	1.999912907
0.001%	1.e-9	10	2.000021375
0%	1.e-17	9	1.999999985

Tab. 1 Case $u_0 = 0$ and $\eta \neq 0$. Results with random noise in the target ($L_d = 2$).

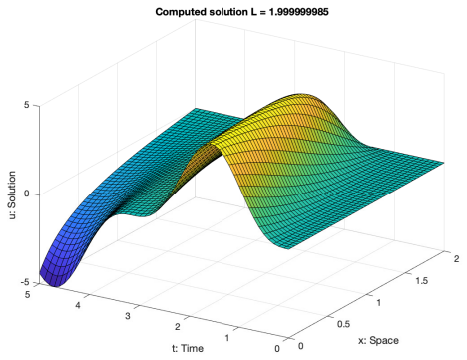


Fig. 1 Case $u_0 = 0$ and $\eta \neq 0$. The computed solution.

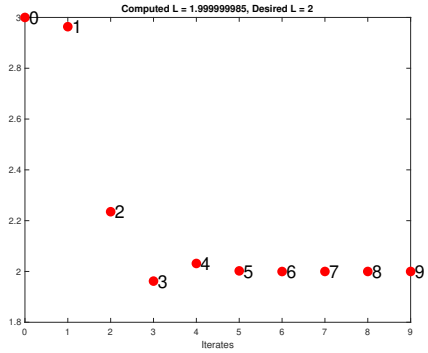


Fig. 2 Case $u_0 = 0$ and $\eta \neq 0$. The iterates.

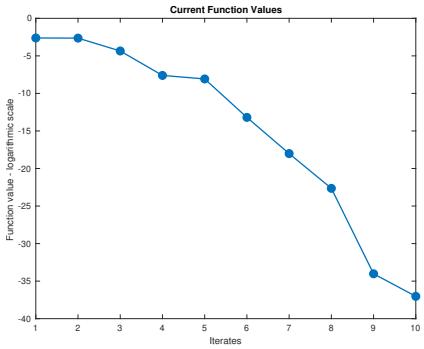


Fig. 3 Case $u_0 = 0$ and $\eta \neq 0$. Evolution of the cost.

Case 1.2: Burgers equation with $u_0 \neq 0$ and large η. We take $T = 5$, $\eta(t) = 5\sin(t)^3$ in $(0,T)$ and $u_0(x) \equiv 3x(2-x)$. Now, starting from $L_i = 2.4$, the target value that we want to recover is $L_d = 2$.

The results of the numerical implementation are shown in Table 2, where again random noise was incorporated. The contents of Figures 4, 5 and 6 are similar to those above.

% noise	Cost	Iterates	Computed L_c
1%	1.e-2	6	2.032815856
0.1%	1.e-5	11	2.012510004
0.01%	1.e-5	9	1.985859861
0.001%	1.e-6	9	1.994836103
0%	1.e-6	9	1.997637334

Tab. 2 Case $u_0 \neq 0$ and large η. Results with random noise in the target ($L_d = 2$).

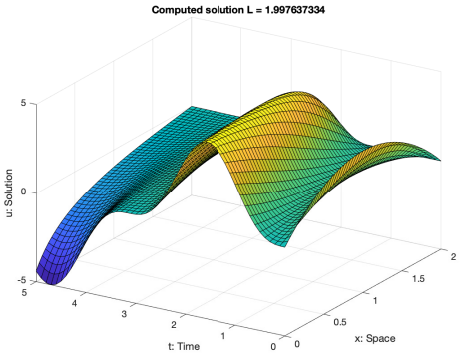


Fig. 4 Case $u_0 \neq 0$ and large η. The computed solution.

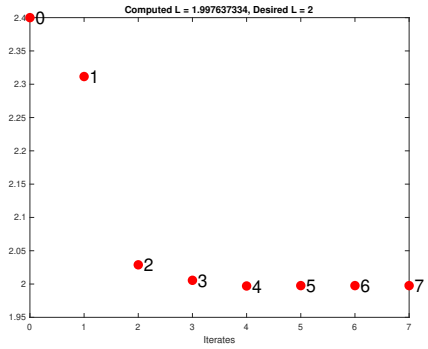


Fig. 5 Case $u_0 \neq 0$ and large η. The iterates.

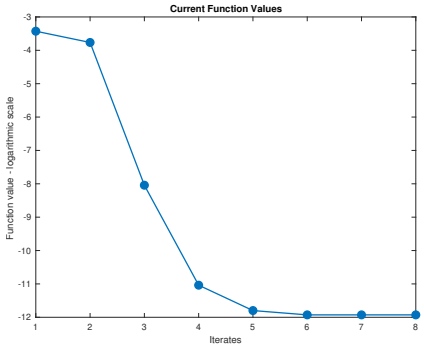


Fig. 6 Case $u_0 \neq 0$ and large η. The cost.

Case 1.3: Burgers equation with $u_0 \neq 0$ and "small" η.

Here, we deal with a non-uniqueness situation. Our aim is to investigate the behavior of the algorithm in a situation of this kind.

We take $T = 6$, $\eta = 0$ in $(0,T)$ and $u_0(x) \equiv \pi\sin(\pi x/2)/(2+\cos(\pi x/2))$. Note that we have $u_0(x) \equiv \sin(3\pi x/L_d^1)/(2+\cos(3\pi x/L_d^1)) \equiv \sin(3\pi x/L_d^2)/(2+\cos(3\pi x/L_d^2))$, with $L_d^1 = 6$ and $L_d^2 = 4$; consequently, this initial data can be used to prove non-uniqueness. We will consider the following experiments:

- First, we start from $L_i = 5.6$. The computed value is $L_c^1 = 5.998083259$ and the associated cost is $J(L_c^1) < 10^{-8}$.
- Then, we start from $L_i = 4.6$. The computed value is $L_c^2 = 4.000601673$ and the associated cost is again $J(L_c^2) < 10^{-9}$.

The corresponding computed boundary observations are displayed in Figures 9 and 10, respectively. Thus, we confirm that these identical observations correspond, as we already knew, two different solutions.

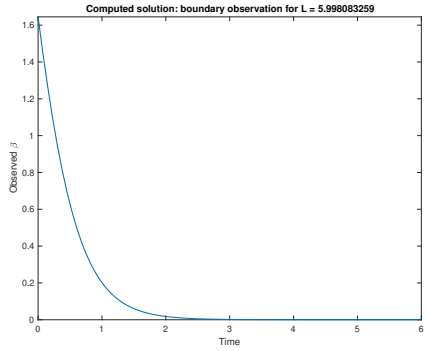


Fig. 7 Case $\eta = 0, u_0(x) \neq 0$. The computed boundary observation $u_x(0,\cdot)$ for $L_c^1 = 5.996562049$.

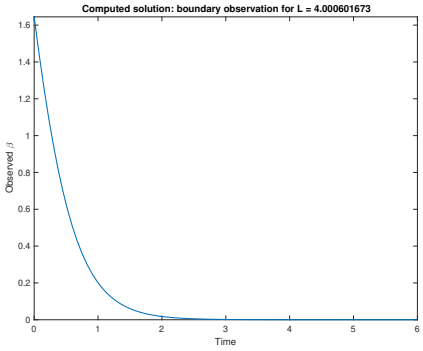


Fig. 8 Case $\eta = 0$, $u_0(x) \neq$. The computed boundary observation $u_x(0,\cdot)$ for $L_c^2 = 4.007345905$

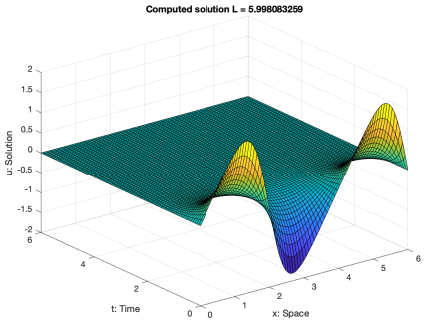


Fig. 9 Case $\eta = 0$, $u_0(x) \neq 0$. The computed solution corresponding to $L_c^1 = 5.998083259$.

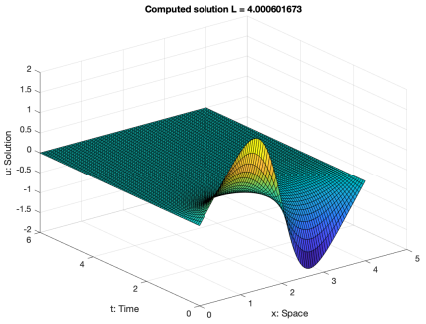


Fig. 10 Case $\eta = 0$, $u_0(x) \neq 0$. The computed solution corresponding to $L_c^2 = 4.000601673$.

Acknowledgements

The first author was supported by the Spanish Government's Ministry of Science, Innovation and Universities (MICINN), under grant PGC2018-094522-B-I00 and the Basque Government, under grant IT12247-19. The second and third authors were partially supported by MICINN, under grant MTM2016-76690-P.

The fourth author was supported by Grant-in-Aid for Scientific Research (S) 15H05740 of Japan Society for the Promotion of Science and by The National Natural Science Foundation of China (no. 11771270, 91730303). This work was prepared with the support of the "RUDN University Program 5-100".

References

[1] J. Apraiz, J. Cheng, A. Doubova, E. Fernández-Cara and M. Yamamoto, *Some Results Concerning Inverse Problems for One-Dimensional in Space PDEs*, to appear (2021).

[2] J. Apraiz, A. Doubova, E. Fernández-Cara and M. Yamamoto, *Some Inverse Problems for the Burgers Equation and Related Systems*, submitted.

[3] M. Bellassoued, M. Yamamoto, *Carleman estimates and applications to inverse problems for hyperbolic systems*, Springer Monographs in Mathematics. Springer, Tokyo, 2017.

[4] C. Fabre and G. Lebeau, *Prolongement unique des solutions de l'equation de Stokes*, Comm. Partial Differential Equations, **21**, no. 3 - 4, (1996), 573 – 596.

[5] M. Hanke, *A taste of inverse problems — basic theory and examples*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2017.

[6] A. Hasanov, V.G. Romanov, *Introduction to inverse problems for differential equations*, Springer, Cham, 2017.

[7] X. Huang, O. Yu. Imanuvilov and M. Yamamoto. Stability for inverse source problems by Carleman estimates. *Inverse Problems*, Vol. 36, no.12, (2020), 125006.

[8] V. Isakov. Inverse Problems for Partial Differential Equations. *Springer, New York*, 2006.

[9] M. Yamamoto. Lectures on inverse problems. *Università Roma 2 "Tor Vergata"*, 2021.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 39–43)

Pricing TARN options with a stochastic local volatility model

Iñigo Arregui [1,2], **Jonatan Ráfales** [1,2]
1. *Depto. de Matemáticas, Universidad de A Coruña, A Coruña, Spain*
2. *CITIC, A Coruña, Spain*

Abstract

Target Accumulation Redemption Notes (TARNs) are financial derivatives which give their holders the right to receive periodic coupons until the accumulated sum of those ones reaches an agreed target. In this work, we solve a partial differential equations (PDEs) model for pricing TARN options by implementing an alternating-direction implicit finite difference method (ADI method). We combine the numerical solution with a stochastic local volatility (SLV) technique and show the numerical results for a particular example.

1. Introduction

European and American options are the most well-known derivative products, and have been widely studied from the financial and mathematical points of view. For these options, jointly known as vanilla options, their price depends mainly on the value of the underlying asset.

On the opposite side, exotic options present different features which also have an effect on the price. For example, Bermudan options offer multiple exercise dates, the pay-off of Asian options depends on the average price of the underlying asset and the pay-off of barrier options depends on whether or not the price of the underlying asset reaches an agreed value during the option's lifetime. Other examples of exotic products are the target redemption products, whose notional amount increases until a certain target is reached [5]. In particular, the value of a Target Accumulation Redemption Note (TARN) depends on an accumulated amount: if the sum of coupons reaches an agreed target before the maturity date, the holder of the note receives a final payment, also known as knockout, and the contract terminates. These products are usually traded in foreign exchange (FX) markets.

2. Mathematical model

We propose a stochastic local volatility model based on Heston model [4]:

$$\begin{cases} dS_t = (r_d(t) - r_f(t))S_t\,dt + L(S_t,t)\sqrt{V_t}S_t\,dW_t^1 \\ dV_t = \kappa(\theta - V_t)\,dt + \lambda\sqrt{V_t}\,dW_t^2 \\ dW_t^1\,dW_t^2 = \rho\,dt\,, \end{cases} \tag{2.1}$$

where L is the leverage function, which represents the contribution of the local volatility and will be calibrated with the help of market data, S_t is the underlying asset, V_t is the stochastic variance, W_t^1 and W_t^2 are two Brownian motions, r_d and r_f are the domestic and foreign interest rates, respectively, and κ, θ, λ and ρ are the Heston parameters.

Let $r(t) = r_d(t) - r_f(t)$. Let us assume $2\kappa\theta \leq \lambda^2$ and $V_0 > 0$. Then, Feller condition states that $V_t > 0$ for every $t > 0$. Moreover, we will assume S_0, κ, θ and λ are strictly positive, $-1 < \rho < 1$ and the leverage function is positive and bounded. Under these hypotheses, it is proven [4] that there exists a unique solution of model (2.1) and there also exists a function $p := p(S,V,t)$, called transition probability function, solution of the Fokker-Planck (FP) equation:

$$\frac{\partial p}{\partial t} = -\frac{\partial}{\partial S}\big(r(t)Sp\big) - \frac{\partial}{\partial V}\big(\kappa(\theta - V)p\big) + \frac{1}{2}\frac{\partial^2}{\partial S^2}\big(L^2S^2Vp\big) + \frac{\partial^2}{\partial S\partial V}\big(\lambda\rho LSVp\big) + \frac{1}{2}\frac{\partial^2}{\partial V^2}\big(\lambda^2Vp\big)\,, \tag{2.2}$$

such that the leverage function can be written as [4]:

$$L(S,t) = \sigma_{LV}(S,t)\sqrt{\frac{\int_{\mathbb{R}} p(S,V,t)\,dV}{\int_{\mathbb{R}} V\,p(S,V,t)\,dV}}\,. \tag{2.3}$$

Let $S(t)$ be the FX rate at time t, t_0 the actual date and $t_1, t_2, \ldots, t_K$ the so-called fixing dates. Moreover, let E be the strike, U the target accrual level and $A(t)$ the accumulated amount at time t [3]. On each fixing date, t_k, there is a cash flow payment:

$$\widetilde{C}_k \equiv \beta(S(t_k) - E) \times 1_{\beta\times S(t_k)\geq\beta\times E}\,,$$

where β is a strategy on foreign exchange ($\beta = 1$ for a call option and $\beta = -1$ for a put option), until the accumulated amount A reaches the target U. Let $t_{\widetilde{K}}$ be the first fixing date when the target is breached:

$$\widetilde{K} = \min_{k=1,\dots,K} \{k : A(t_k) \geq U\}$$

and let $\widetilde{K} = K$ if the target is not breached, the payment can be written as:

$$C_k(S,A) = \begin{cases} \widetilde{C}_k \times \left(1_{A(t_{k-1})+\widetilde{C}_k<U} + W_k \times 1_{A(t_{k-1})+\widetilde{C}_k \geq U}\right), & \text{if } t_k \leq t_{\widetilde{K}}, \\ 0, & \text{if } t_k > t_{\widetilde{K}}, \end{cases} \tag{2.4}$$

where $A(t_{k-1})$ is the accumulated amount immediately after t_{k-1}. This magnitude is given by a piecewise constant function:

$$A(t) = \begin{cases} A(t_{k-1}), & \text{if } t_{k-1} \leq t < t_k, \\ A(t_{k-1}) + C_k(S(t_k), A(t_{k-1})), & \text{if } t = t_k. \end{cases}$$

Moreover, W_k is a weight that depends on the knockout when the target is breached. We will consider three types of knockout, such that the weight can be written as:

$$W_k = \begin{cases} 1, & \text{in the case of full gain,} \\ \dfrac{U - A(t_{k-1})}{\beta \times (S(t_k) - E)}, & \text{in the case of part gain,} \\ 0, & \text{in the case of no gain.} \end{cases}$$

Finally, let $u := u(S,V,t,A)$ be the value of a TARN option, the following SLV option pricing PDE:

$$\frac{\partial u}{\partial t} + r(t)S\frac{\partial u}{\partial S} + (\kappa(\theta - V))\frac{\partial u}{\partial V} + \frac{1}{2}L^2VS^2\frac{\partial^2 u}{\partial S^2} + \rho L\lambda SV\frac{\partial^2 u}{\partial S \partial V} + \frac{1}{2}\lambda^2 V\frac{\partial^2 u}{\partial V^2} - r_d(t)u = 0$$

is valid between fixing dates. Furthermore, for each fixing date we can pose:

$$u\big(S,V,t_k^-,A(t_k^-)\big) = u\big(S,V,t_k,A(t_k^-) + C_k(S,A(t_k^-))\big) + C_k\big(S,A(t_k^-)\big),$$

where C_k is given by (2.4) and t_k^- is the time infinitesimally before t_k.

2.1. Numerical methods

We compute the optimal Heston parameters $(\kappa,\theta,\lambda,\rho)$ for the calibration of the SLV model. For each maturity, the COS method is developed to price European options under the Heston model, $[w^{\text{Hes}}]_i^m := w^{\text{Hes}}(E_i,T^m)$ and the Levenberg-Marquardt non-linear least squares algorithm is applied to find the optimal parameters by minimizing

$$\min_{\kappa,\theta,\lambda,\rho} \sum_{i=1}^{N_E} \left([w^{\text{Hes}}(\kappa,\theta,\lambda,\rho)]_i^m - w_i^m\right)^2,$$

where $w_i^m := w(E_i,T^m)$ are the market data for different strikes and maturities.

As Feller condition does not always hold in real markets, we propose a logarithmic change of variable [2]. Thus, let S_0 and V_0 be the initial values of the underlying and the variance, and let

$$X_t = \log(S_t/S_0), \qquad Z_t = \log(V_t/V_0).$$

In the new domain $(-\infty,\infty)\times(-\infty,\infty)$, model (2.1) is rewritten as:

$$\begin{cases} dX_t = \left(r(t) - \frac{1}{2}L(X_t,t)^2V_0e^{Z_t}\right)dt + L(X_t,t)\sqrt{V_0e^{Z_t}}\,dW_t^1 \\ dZ_t = \left((\kappa\theta - \frac{1}{2}\lambda^2)\frac{1}{V_0e^{Z_t}} - \kappa\right)dt + \lambda\frac{1}{\sqrt{V_0e^{Z_t}}}\,dW_t^2 \\ dW_t^1\,dW_t^2 = \rho dt. \end{cases}$$

A previous step to apply numerical methods is the truncation of the new unbounded domain to a bounded one. Thus, we consider the fixed domain $\Omega = (X_{\min},X_{\max})\times(Z_{\min},Z_{\max})$, which is finer around $(X_0 = 0, Z_0 = 0)$, and pose the FP equation (2.2):

$$\begin{aligned} \frac{\partial p}{\partial t} = &-\frac{\partial}{\partial X}\left((r(t) - \frac{1}{2}L^2V_0e^Z)p\right) - \frac{\partial}{\partial Z}\left(((\kappa\theta - \frac{1}{2}\lambda^2)\frac{1}{V_0e^Z} - \kappa)p\right) \\ &+ \frac{1}{2}\frac{\partial^2}{\partial X^2}\left(L^2V_0e^Z p\right) + \frac{\partial^2}{\partial X\partial Z}\left(\lambda\rho Lp\right) + \frac{1}{2}\frac{\partial^2}{\partial Z^2}\left(\lambda^2\frac{1}{V_0e^Z}p\right), \end{aligned} \tag{2.5}$$

with the initial condition:

$$p(X, Z, 0) = \delta(X)\delta(Z), \tag{2.6}$$

where δ is the Dirac function. Additionally, we impose the boundary conditions:

$$\frac{\partial^2 p}{\partial X^2}(X_{\min}, Z, t) = 0, \quad \frac{\partial^2 p}{\partial X^2}(X_{\max}, Z, t) = 0, \quad \frac{\partial^2 p}{\partial Z^2}(X, Z_{\min}, t) = 0, \quad \frac{\partial^2 p}{\partial Z^2}(X, Z_{\max}, t) = 0. \tag{2.7}$$

Moreover, the leverage function (2.3) is given by:

$$L(X, t) = \sigma_{LV}(X, t)\sqrt{\frac{\int_{\mathbb{R}^+} p(X, Z, t)dZ}{\int_{\mathbb{R}^+} V_0 e^Z p(X, Z, t)dZ}}. \tag{2.8}$$

As we can see in (2.5) and (2.8), the computing of the transition probability function needs the leverage function and reciprocally. Therefore, we propose a fixed point scheme to solve the problem at each time step, in which an alternating directions implicit (ADI) method is developed to solve the FP problem and the trapezoidal rule is applied to approximate the leverage function. The modified Douglas ADI scheme at each step can be written as:

$$\begin{aligned} A &= p^{n-1} + \Delta t^n \big[F_0(p^{n-1}, t^{n-1}) + F_1(p^{n-1}, t^{n-1}) + F_2(p^{n-1}, t^{n-1})\big], \\ B - \alpha\Delta t^n F_1(B, t^n) &= A - \alpha\Delta t^n F_1(p^{n-1}, t^{n-1}), \qquad (2.9) \\ C - \alpha\Delta t^n F_2(C, t^n) &= B - \alpha\Delta t^n F_2(p^{n-1}, t^{n-1}), \qquad (2.10) \\ p^n &= C, \end{aligned}$$

for $n = 1, 2, \ldots, N_T$, where

$$\begin{aligned} F_0(p, t) &= \frac{\partial^2}{\partial X \partial Z}\big(\lambda\rho L p\big), \\ F_1(p, t) &= -\frac{\partial}{\partial Z}\Big(\big((\kappa\theta - \frac{1}{2}\lambda^2)\frac{1}{V_0 e^Z} - \kappa\big)p\Big) + \frac{1}{2}\frac{\partial^2}{\partial Z^2}\big(\frac{\lambda^2}{V_0 e^Z}p\big), \\ F_2(p, t) &= -\frac{\partial}{\partial X}\big((r(t) - \frac{1}{2}L^2 V_0 e^Z)p\big) + \frac{1}{2}\frac{\partial^2}{\partial X^2}\big(L^2 V_0 e^Z p\big). \end{aligned}$$

In addition, a mixing fraction parameter, η, is applied to the volatility of the volatility, λ:

$$dV_t = \kappa(\theta - V_t)dt + \eta\lambda\sqrt{V_t}dW_t^2.$$

For each maturity, the ADI method is developed to price options under the SLV model, $[y^{\text{SLV}}]_i^m := y^{\text{SLV}}(E_i, T^m)$ and the golden section search algorithm is applied to determine the optimal parameter by minimizing

$$\min_{\eta} \sum_{i=1}^{N_E} \big([y^{\text{SLV}}(\eta)]_i^m - y_i^m\big)^2,$$

where $y_i^m := y(E_i, T^m)$ are the market data for different strikes and maturities.

Finally, we introduce the time-to-maturity variable ($\tau = T - t$) and deduce the TARN price PDE in terms of X and Z:

$$\begin{aligned} \frac{\partial u}{\partial \tau} &= \big(r(\tau) - \frac{1}{2}L^2 V_0 e^Z\big)\frac{\partial u}{\partial X} + \frac{1}{2}L^2 V_0 e^Z \frac{\partial^2 u}{\partial X^2} + \lambda\eta\rho L\frac{\partial^2 u}{\partial X \partial Z} \\ &\quad + \big((\kappa\theta - \frac{1}{2}\lambda^2\eta^2)\frac{1}{V_0 e^Z} - \kappa\big)\frac{\partial u}{\partial Z} + \frac{1}{2}\frac{\lambda^2\eta^2}{V_0 e^Z}\frac{\partial^2 u}{\partial Z^2} - r_d(\tau)u, \end{aligned} \tag{2.11}$$

with the initial condition:

$$u(X, Z, 0, A) = 0. \tag{2.12}$$

Additionaly, we assume the boundary conditions [4]:

$$\begin{aligned} &\frac{1}{S_0^2}e^{-2X}\big(\frac{\partial^2 u}{\partial X^2}(X_{\min}, Z, t) - \frac{\partial u}{\partial X}(X_{\min}, Z, t)\big) = 0, \quad \frac{1}{V_0^2}e^{-2Z}\big(\frac{\partial^2 u}{\partial Z^2}(X, Z_{\min}, t) - \frac{\partial u}{\partial Z}(X, Z_{\min}, t)\big) = 0, \\ &\frac{1}{S_0^2}e^{-2X}\big(\frac{\partial^2 u}{\partial X^2}(X_{\max}, Z, t) - \frac{\partial u}{\partial X}(X_{\max}, Z, t)\big) = 0, \quad \frac{1}{V_0^2}e^{-2Z}\big(\frac{\partial^2 u}{\partial Z^2}(X, Z_{\max}, t) - \frac{\partial u}{\partial Z}(X, Z_{\max}, t)\big) = 0. \end{aligned} \tag{2.13}$$

As for the numerical solution of the FP problem, we also propose the use of the ADI algorithm to solve (2.11 − 2.13), jointly with the jump condition for each fixing date τ_k:

$$u\big(X, Z, \tau_k^+, A(\tau_k^+)\big) = u\big(X, Z, \tau_k, A(\tau_k^+) + C_k(X, A(\tau_k^+))\big) + C_k\big(X, A(\tau_k^+)\big),$$

where τ_k^+ is the time infinitesimally after τ_k. Fig. 1 shows a sketch of the scheme.

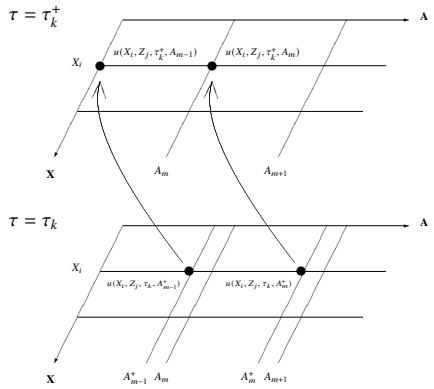


Fig. 1 Jump condition

2.2. Numerical results

We present a numerical test for the valuation of a TARN call option in the frame of foreign exchange. For this aim, we have considered the US dollar and British pound as domestic and foreign currencies, respectively, an initial underlying $S_0 = 1.320$, an initial variance $V_0 = 0.004$ and a strike $E = 1.283$. Furthermore, the domestic and foreign interest rates are shown in Tab. 1 of [1], the maturity period is $T = 12$ months and the fixing dates are taken every 30 days, thus $K = 12$.

Fig. 2 shows the market implied volatility (left) and the local volatility (right), which is computed by means of Dupire's formula. Thus, we apply the previous techniques to compute the Heston parameters, which are shown in Tab. 3 (left) of [1].

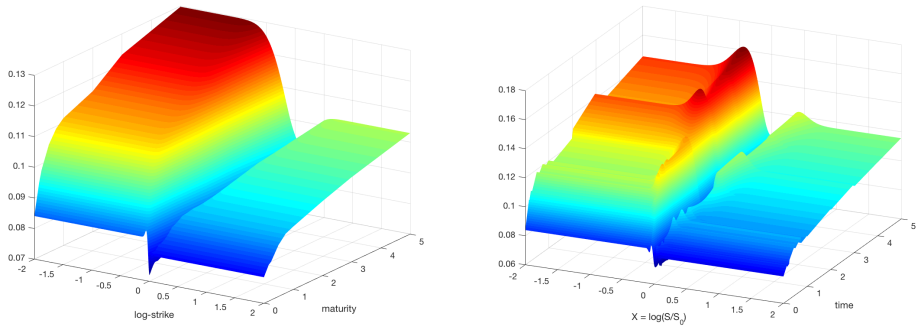


Fig. 2 The implied volatility (left) and the local volatility (right)

Next, we have approximated the (X, Z) domain with a mesh similar to the one plotted in Fig. 3, consisting of 400×100 nodes. As we have previously detailed, the mesh is finer around the point $(X_0 = 0, Z_0 = 0)$. We have also refined the mesh for values close to $Z_{\min}$ in order to minimize the errors arising from the fact that the Feller condition may be not accurate for these values of the volatility. Moreover, we have used 180 time steps and the parameter $\alpha = 0.5$ in the ADI method. Fig. 4 shows the computed solution of the FP problem at the maturity (left) and the leverage function (right).

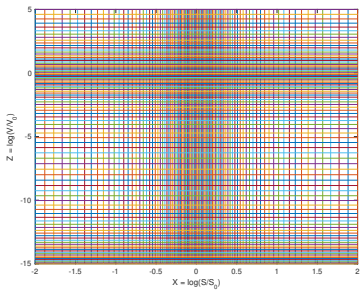


Fig. 3 Mesh

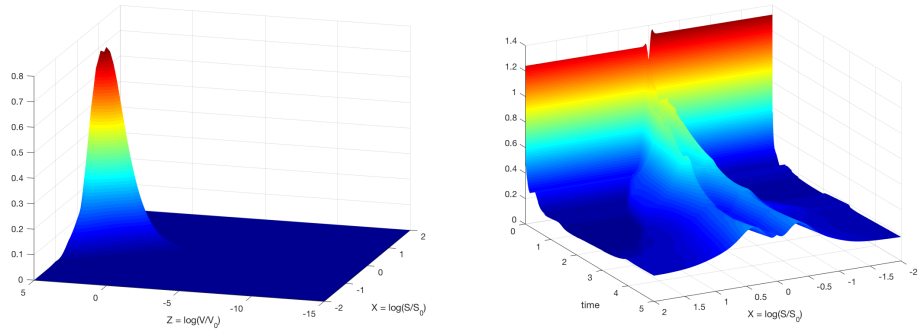


Fig. 4 Numerical solution of the FP equation (left) and the leverage function (right)

Finally, we calibrate the mixing fraction parameter η, which is shown in Tab. 3 (right) of [1], and show the numerical approximation of the TARN price for different knockouts in Fig. 5. More details and results are available in [1].

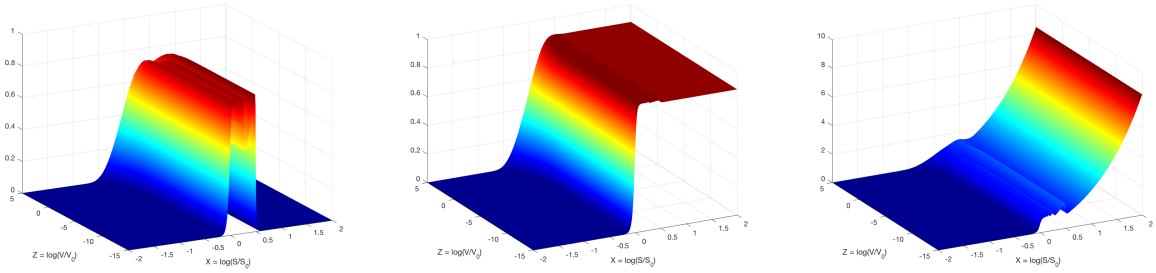


Fig. 5 TARN option price: no gain (left), part gain (center) and full gain (right) knockouts for $U = 0.90$.

2.3. Conclusions

We solve a partial differential equations model to price TARN options. We have improved previous results by introducing a SLV technique in order to better reflect the market volatilities (taking advantage of local volatility methods) on a path dependent derivative product (for which stochastic volatility methods are more convenient). This SLV approach can be extended to other kinds of exotic options.

An alternating directions method (ADI) is implemented and the volatility surfaces, transition probability function, leverage function and option price are computed. Therefore, we have a tool to valuate this kind of exotic options.

Acknowledgements

Authors are financially supported by Spanish Ministerio de Ciencia e Innovación (grant PID2019-108584RB-100). Authors also acknowledge the support received from the Centro de Investigación de Galicia (CITIC), funded by Xunta de Galicia and the European Union (European Regional Development Fund- Galicia 2014-2020 Program), by grant ED431G 2019/01.

References

[1] Iñigo Arregui and Jonatan Ráfales. A stochastic local volatility technique for TARN options. *International Journal of Computer Mathematics*, 97(5):1133–1149, 2020.

[2] Fang Fang and Cornelis W. Oosterlee. A novel pricing method for European options based on Fourier-Cosine series expansions. *SIAM Journal on Scientific Computing*, 31(2):826–848, 2009.

[3] Xiaolin Luo and Pavel V. Shevchenko. Pricing TARNs using a finite difference method. *The Journal of Derivatives*, 23(1):62–72, 2015.

[4] Yu Tian. *The hybrid stochastic-local volatility model with applications in pricing FX options*. PhD thesis, Monash University, 2013.

[5] Uwe Wystup. *FX options and Structured Products*. John Wiley & Sons, Chichester, 2017.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 44–50)

XVA for American options with two stochastic factors: modelling, mathematical analysis and numerical methods

I. Arregui[1] and B. Salvador[1] and D. Ševčovič[2] and C. Vázquez[1]
1. Universidade da Coruña, Spain
2. Comenius University, Bratislava, Slovakia

Abstract

In this work, we derive new linear and nonlinear partial differential equations (PDEs) models for pricing American options and total value adjustment in the presence of counterparty risk. Moreover, stochastic spreads are considered, which increases the dimension of the problem.

1. Introduction

Counterparty risk can be understood as the risk to each party of a contract from a future situation in which one of the counterparties cannot live up its contractual obligations. Since the last financial, crisis when several institutions went bankrupt, a relevant effort in quantitative finance research concerns to the consideration of counterparty risk in financial contracts, specially in the pricing of derivatives As a consequence, different adjustments on the value of the derivative without counterparty risk (hereafter referred as risk–free derivative) are being included in the derivative pricing. For example, the credit value adjustment (CVA) refers to the variation on the price of a contract due to the possibility of default of one (or both) of the counterparties. Adjustments on debit (DVA) and funding (FVA) are also important issues included in the so called total value adjustment (XVA). The XVA incorporates the sum of all the adjustments related to counterparty risk.

In a previous work [2], European and American options have been priced considering corunterparty risk. In such model, constant intensities of default for both counterparties have been assumed. So that a model depending on just one underlying stochastic factor (the underlying asset) is posed to price XVA. However, the intensity of default is not always constant, then stochastic intensities of default has been assumed in [3] as a result a model depending on two stochastic factors (the asset price and the spread from the investor) was deduced to price European options. In the current work, as we have done in [3], we consider that only the investor is defaultable and presents a stochastic intensity of default. Moreover, similar hypotheses as in the European options model introduced in [3] are assumed. Then, we extend the models introduced in [2], [3] to price the American options considering counterparty risk and compute the associated total value adjustment when stochastic intensity of default is assumed. So, we deduce a two dimensional PDE model for the American risky derivative value with stochastic intensity of deafult. The plan of the chapter is the following. In Section 1 we pose the complementarity problems deduced from the hedging arguments. In Section 2 we present the mathematical analysis of the previous problems. Section 3 presents the numerical methods and Section 4 shows some illustrative numerical results.

2. Mathematical model

In this section, we deduce the models for American options considering counterparty risk. With this aim, we consider self–financing portfolio and non–arbitrage scenarios. Moreover, we assume an investor as a risky counterparty and consider that the issuer's intensity of default is null. Thus, the underlying asset price S, and the short term CDS spread of the investor h are modelled by the following system of stochastic differential equations:

$$\begin{aligned} dS_t &= (r(t) - q(t))\, S_t\, dt + \sigma^S(t)\, S_t\, dW_t^S\,, \\ dh_t &= (\mu^h(t) - M^h(t)\sigma^h(t))\, dt + \sigma^h(t)\, dW_t^h\,, \end{aligned}$$

where $(r(t) - q(t))$ and $(\mu^h(t) - M^h(t)\sigma^h(t))$ are the (respective) drifts of the processes. Moreover, $r(t)$ denotes the risk-free interest rate, $q(t)$ is the asset dividend yield rate, $M^h(t)$ is the market price of investor's credit risk, $\sigma^S(t,S)$ and $\sigma^h(t,h)$ are the volatility functions, and W_t^S and W_t^h are two correlated Wiener processes (i.e., $\rho\, dt = dW_t^S\, dW_t^h$) so that ρ is the instantaneous correlation between S_t and h_t.

Thus, we consider a derivative trade between a hedger and an investor, where only the investor has probability of default. The risky derivative value from the point of view of the investor, at time t, is denoted by $\widehat{V}(t, S_t, h_t, J_t^I)$, and depends on the spot value of the asset (S_t), on the spread of the investor (h_t) and on the investor's default state at time t (J_t^I). Remind that $J_t^I = 1$ in case of default before or at time t, otherwise $J_t^I = 0$. The risk–free

American option value, corresponding to the same contract between two free–bankruptcy counterparties, is denoted by $\widehat{V}(t,S_t)$ and does not include any counterparty risk adjustment, whereas the risky derivative price $\widehat{V}_t$ includes total value adjustment.

The risky derivative price in case of the investor makes default is given by:

$$\widehat{V}(t,S_t,h_t,1) = RM^+(t,S_t,h_t) + M^-(t,S_t,h_t), \tag{2.1}$$

where $M(t,S_t,h_t)$ denotes the mark–to–market price, $M^+ = \max(M,0)$ and $M^- = \min(0,M)$. In terms of the mark-to-market condition (2.1), we introduce $\Delta\widehat{V}$ as the variation of $\widehat{V}$ at default, which is given by:

$$\Delta\widehat{V}_t = RM_t^+ + M_t^- - \widehat{V}_t\,, \tag{2.2}$$

where $M_t = M(t,S_t,h_t)$. As it is usually assumed in the literature [4], and as we did in [2] and [3], we consider two possibilities for the mark–to–market value: either the risk–free value, either the derivative value including counterparty risk.

The hedger will trade with different financial instruments to hedge the market risk, the spread risk and the investor's default risk. Thus, in order to derive the value of American options with counterparty risk, we consider the same self–financing portfolio built for European options in [3], Π_t, which is designed to hedge all underlying risk factors:

$$\Pi_t = \alpha(t)H(t) + \beta(t) + \gamma(t)\mathrm{CDS}(t,T) + \varepsilon(t)\mathrm{CDS}(t,t+dt) + \Omega(t)B(t,t+dt)\,. \tag{2.3}$$

Furthermore, in order to avoid arbitrage opportunities we introduce the following hedging inequality:

$$d\widehat{V}_t \le d\Pi_t\,. \tag{2.4}$$

Next, by applying Itô's Lemma for jump diffusion processes, we obtain the variation $d\widehat{V}_t$ of the derivative value $\widehat{V}_t$. Thus, replacing the change of the portfolio and the change of the derivative in (2.4), the hedging equation is transformed into:

$$\begin{aligned}
&\frac{\partial\widehat{V}}{\partial t} + \frac{1}{2}(\sigma^S)^2S^2\frac{\partial^2\widehat{V}}{\partial S^2} + \frac{1}{2}(\sigma^h)^2\frac{\partial^2\widehat{V}}{\partial h^2} + \rho\sigma^S\sigma^h S\frac{\partial^2\widehat{V}}{\partial S\partial h}\\
&\quad\le \frac{\partial\widehat{V}/\partial S}{\partial H/\partial S}\left(cH - (r-q)S\frac{\partial H}{\partial S}\right) + \frac{\partial\widehat{V}/\partial S}{\partial H/\partial S}(-fH)\\
&\quad+ \frac{\partial\widehat{V}/\partial h}{\partial\mathrm{CDS}(t,T)/\partial h}\left(-\frac{h}{1-R}\Delta\mathrm{CDS}(t,T) - \left(\mu^h - M\sigma^h\right)\frac{\partial\mathrm{CDS}(t,T)}{\partial h}\right)\\
&\quad+ \left(\frac{\partial\widehat{V}/\partial h}{\partial\mathrm{CDS}(t,T)/\partial h}\frac{\Delta\mathrm{CDS}(t,T)}{1-R} - \frac{\Delta\widehat{V}}{1-R}\right)h + f\widehat{V}\,,
\end{aligned} \tag{2.5}$$

in $[0,T)\times(0,\infty)\times(0,\infty)$. Then, the American option value when considering counterparty risk is modelled by the following complementarity problem:

$$\begin{cases}
\mathcal{L}(\widehat{V}) = \dfrac{\partial\widehat{V}}{\partial t} + \tilde{\mathcal{L}}_{Sh}\widehat{V} + \dfrac{\Delta\widehat{V}}{1-R}h - f\widehat{V} \le 0\\
\widehat{V}(t,S,h) \ge G(S)\\
\mathcal{L}(\widehat{V})(\widehat{V}-G) = 0\\
\widehat{V}(T,S,h) = G(S)\,,
\end{cases} \tag{2.6}$$

where $G(S)$ represents the option payoff and the differential operator $\widetilde{\mathcal{L}}_{Sh}$ is

$$\mathcal{L}_{Sh}V \equiv \frac{1}{2}(\sigma^S)^2S^2\frac{\partial^2V}{\partial S^2} + \frac{1}{2}(\sigma^h)^2\frac{\partial^2V}{\partial h^2} + \rho\sigma^S\sigma^h S\frac{\partial^2V}{\partial h\partial S} + (r-q)S\frac{\partial V}{\partial S} - \frac{\kappa}{1-R}h\frac{\partial V}{\partial h}\,.$$

According to the mark-to–market choices, two alternative linear complementarity problems are deduced:

- If $M = \widehat{V}$, we deduce the nonlinear complementarity problem:

$$\begin{cases}
\mathcal{L}_1(\widehat{V}) = \dfrac{\partial\widehat{V}}{\partial t} + \mathcal{L}_{Sh}\widehat{V} - f\widehat{V} - h\widehat{V}^+ \le 0, \quad \text{in } [0,T)\times(0,\infty)\times(0,\infty)\\
\widehat{V}(t,S,h) \ge G(S)\\
\mathcal{L}_1(\widehat{V})(\widehat{V}-G) = 0\\
\widehat{V}(T,S,h) = G(S)\,.
\end{cases} \tag{2.7}$$

- If $M = V$, the following linear complementarity problem is derived:

$$\begin{cases} \mathcal{L}_2(\widehat{V}) = \dfrac{\partial \widehat{V}}{\partial t} + \mathcal{L}_{Sh}\widehat{V} - \left(\dfrac{h}{1-R} + f\right)\widehat{V} \\ \qquad\qquad -((1-R)V^+ - V)\dfrac{h}{1-R} \leq 0, \quad \text{in } [0,T)\times(0,\infty)\times(0,\infty) \\ \widehat{V}(t,S,h) \geq G(S) \\ \mathcal{L}_2(\widehat{V})(\widehat{V} - G) = 0 \\ \widehat{V}(T,S,h) = G(S)\,. \end{cases} \tag{2.8}$$

Moreover, to compute the XVA value, we consider that $\widehat{V} = V + U$ where U denotes the XVA, then the adjustments can be obtained as the difference of the risky derivative value, $\widehat{V}$, and the risk–free derivative value, V, which is the solution of the classical Black-Scholes American problem:

$$\begin{cases} \mathcal{L}_3(V) = \dfrac{\partial V}{\partial t} + \mathcal{L}_S V - fV \leq 0\,, \quad \text{in } [0,T)\times(0,\infty) \\ V(t,S) \geq G(S) \\ \mathcal{L}_3(V)(V - G) = 0 \\ V(T,S) = G(S)\,, \end{cases} \tag{2.9}$$

where the operator $\mathcal{L}_S$ is given by

$$\mathcal{L}_S V \equiv \frac{(\sigma^S)^2}{2} S^2 \frac{\partial^2 V}{\partial S^2} + (r-q)S\frac{\partial V}{\partial S}\,.$$

In order to numerically solve problems (2.7) and (2.8) by a finite element method, we proceed to localize the problems on a bounded domain. For this purpose, let us consider $\Omega = (0, S_\infty)\times(0, h_\infty)$ for large enough values of S_∞ and h_∞, so that their choice does not affect the solution in the domain of financial interest. We need to impose appropiate boundary conditions on the risky derivative value problem in the bounded domain. For this purpose, we consider the same boundary conditions than for V and $\widehat{V}$ as in the case of European options in [3]. Then, at $S = 0$ and $S = S_\infty$, the derivative value is given by:

$$\begin{cases} \widehat{V}(t,0,h) = V(t,0) = V_0(t)\,, \\ \widehat{V}(t,S_\infty,h) = V(t,S_\infty) = V_\infty(t)\,, \end{cases}$$

where the values of $V_0(t)$ and $V_\infty(t)$ are respectively given by:

$$V_0(t) = \begin{cases} 0\,, & \text{for a call option,} \\ K\exp(-f(T-t))\,, & \text{for a put option,} \end{cases} \tag{2.10}$$

$$V_\infty(t) = \begin{cases} S_\infty - K\,, & \text{for a call option,} \\ 0\,, & \text{for a put option.} \end{cases} \tag{2.11}$$

In the next section, the existence and uniqueness of solution of problem (2.7) are studied. For this purpose, we introduce the problem which models the XVA in order to obtain a problem with homogeneous boundary conditions. Then, we split up the risky derivative value, $\widehat{V}$, as the sum of the XVA, U, plus the total value adjustment, V, i.e. $\widehat{V} = V + U$. Introducing this breakdown in (2.7), the following nonlinear complementarity problem is deduced:

$$\begin{cases} \mathcal{L}_t(U) = \dfrac{\partial U}{\partial t} + \mathcal{L}_{Sh}U - fU - h(U+V)^+ \leq -\dfrac{\partial V}{\partial t} - \mathcal{L}_S V + fV\,, \qquad t\in[0,T)\,, \quad (S,h)\in\Omega \\ U(t,S,h) \geq G(S) - V(t,S) \\ \left[\mathcal{L}_t(U) - \left(-\dfrac{\partial V}{\partial t} - \mathcal{L}_S V + fV\right)\right]\left[U - (G(S) - V(t,S))\right] = 0 \\ U(T,S,h) = 0 \\ U(t,0,h) = 0 \\ U(t,S_\infty,h) = 0 \\ U(t,S,0) = 0 \\ (A\nabla U\cdot\vec{n})(\tau,S,h_\infty) = 0\,. \end{cases} \tag{2.12}$$

For the linear problem (2.8), the same boundary conditions are considered.

3. Mathematical analysis

In this section we prove the existence and uniqueness of solution for the XVA problem (2.12) for a given function V. Then, taking into account the existence and uniqueness of solution V for the classical Black-Scholes problem, we obtain the existence and uniqueness of solution for problem (2.7). Introducing the time to maturity variable, $\tau = T - t$, as well as the new variables and unknown:

$$x = \ln \frac{S}{K}, \qquad u(\tau, x, h) = U(t, S, h), \qquad v(\tau, x) = V(t, S)\,.$$

we pose the nonlinear complementarity problem (2.12) as follows:

$$\begin{cases} \mathcal{L}_\tau(u) = \dfrac{\partial u}{\partial \tau} + \mathcal{A}u - \Phi(\tau, u) \geq \ell\,, \quad (x,h) \in \widehat{\Omega}, \quad \tau \in (0,T] \\ u \geq \psi \\ [\mathcal{L}_\tau(u) - \ell]\ [u - \psi] = 0 \\ u(0,S,h) = 0 \\ u(\tau, x_0, h) = 0 \\ u(\tau, x_\infty, h) = 0 \\ u(\tau, x, 0) = 0 \\ (\widehat{A}\nabla u \cdot \vec{n})(\tau, x, h_\infty) = 0\,, \end{cases} \tag{3.1}$$

Theorem 3.1 *The following statements are satisfied*:

1. *The continuous operator $\mathcal{A}$ satisfies Gårding's inequality, i.e.*:

$$(\mathcal{A}z, z) \geq \omega_1 \|z\|^2_{H^1_\Gamma(\widehat{\Omega})} - \omega_2 \|z\|^2_{L^2(\widehat{\Omega})}\,, \quad \forall z \in H^1_\Gamma(\widehat{\Omega})\,, \tag{3.2}$$

   with $\omega_1 > 0$ and $\omega_2 \in \mathbb{R}$.

2. $\ell \in L^2(0,T; L^2(\widehat{\Omega})) \subset L^2(0,T; W^*)$.

3. *Let $D(\phi) = \left\{ z \in H^1_\Gamma(\widehat{\Omega}) \,/\, \phi(z) < \infty \right\}$ and $u_0 = u(0,x,h)$. Then, $u_0 \in \overline{D(\phi)}$.*

4. *$\Phi(\tau, \varphi)$ is Lipschitz continuous on variable φ, i.e.*

$$\|\Phi(\tau, \varphi_1) - \Phi(\tau, \varphi_2)\|_{L^2(\widehat{\Omega})} \leq L_G \|\varphi_1 - \varphi_2\|_{H^1_\Gamma(\widehat{\Omega})}\,.$$

Therefore, the nonlinear variational inequality (3.1) *has a unique solution $u \in L^2(0,T; H^1_\Gamma(\widehat{\Omega})) \cap C([0,T]; L^2(\widehat{\Omega}))$; in particular $u \in W^{1,2}(0,T; L^2(\widehat{\Omega}))$ and satisfies*

$$\|u\|_{W^{1,2}(0,T;L^2(\widehat{\Omega}))} \leq C_1 \left(1 + \|u_0\|_{L^2(\widehat{\Omega})} + \|\ell\|_{L^2(0,T;H^1_\Gamma(\widehat{\Omega}))}\right)\,. \tag{3.3}$$

4. Numerical simulation

The numerical approximation is mainly based on finite element methods combined with the method of characteristics. Moreover, a fixed point scheme is implemented for the nonlinear complementarity problem.

4.1. The method of characteristics

More precisely, taking into account the advective term, the risky derivative problem is approximated by

$$\begin{cases} \mathcal{L}_1^n(\widehat{V}^{n+1}) = \dfrac{\widehat{V}^{n+1} - \widehat{V}^n \circ \chi^n}{\Delta\tau^n} - \operatorname{div}(A\nabla\widehat{V}^{n+1}) + f\widehat{V}^{n+1} + h(\widehat{V}^{n+1})^+ \geq 0\,, \\ \widehat{V}^0(S,h) = 0\,, \\ \widehat{V}^{n+1}(S,h) \geq G(S)\,, \\ \mathcal{L}_1^n(\widehat{V}^{n+1})(\widehat{V}^{n+1} - G) = 0\,, \end{cases} \tag{4.1}$$

for $n = 0, 1, 2 \ldots, N_T - 1$, where $\widehat{V}^n(\cdot) \approx \widehat{V}(\tau^n, \cdot)$ and $\chi^n = \chi(\tau^n) = \chi((S,h), \tau^{n+1}; \tau^n)$ represents the characteristic curve passing through point (S,h) at time τ^{n+1}. Then function χ is the solution of the final value ODE problem:

$$\begin{cases} \dfrac{d\chi_1}{d\tau} = \left(\left(\sigma^S\right)^2 - (r-q)\right)\chi_1, \\ \chi_1(\tau^{n+1}) = S, \end{cases} \qquad \begin{cases} \dfrac{d\chi_2}{d\tau} = \dfrac{\rho\sigma^S\sigma^h}{2} + \dfrac{\kappa}{1-R}\chi_2, \\ \chi_2(\tau^{n+1}) = h\,, \end{cases} \tag{4.2}$$

The components of χ^n can thus be deduced and are given by:

$$\chi_1^n = S \exp\left(-((\sigma^S)^2 - r + q)(\tau^{n+1} - \tau^n)\right),$$
$$\chi_2^n = -\frac{(1-R)\sigma^S\sigma^h\rho}{2\kappa} + \left(h + \frac{(1-R)\sigma^S\sigma^h\rho}{2\kappa}\right)\exp\left(\frac{-\kappa}{1-R}(\tau^{n+1} - \tau^n)\right).$$

4.2. Fixed point scheme

In order to solve the nonlinearity of problem (4.1), a fixed point scheme is proposed at each iteration of the characteristics method. Thus, the global scheme is shown in Algorithm 1.

Algorithm 1

Let $N_T > 1$, $n = 0$, $\varepsilon > 0$ and $\widehat{V}^0$ given
For $n = 1, 2, \ldots, N_T - 1$:

1. Let $\widehat{V}^{n+1,0} = \widehat{V}^n$, $k = 0$, $e = \varepsilon + 1$
2. For $k = 0, 1, \ldots$
   - Search $\widehat{V}^{n+1,k+1}$ solution of:

$$\begin{aligned}
&(1 + \Delta\tau^n f)\,\widehat{V}^{n+1,k+1} - \Delta\tau^n \operatorname{div}(A\nabla\widehat{V}^{n+1,k+1}) \\
&\qquad\qquad \geq \widehat{V}^n \circ \chi^n - \Delta\tau^n\, h\,(\widehat{V}^{n+1,k})^+ \\
&\widehat{V}^{n+1,k+1}(S,h) \geq G(S) \\
&\mathcal{L}_1^n(\widehat{V}^{n+1,k+1})(\widehat{V}^{n+1,k+1} - G) = 0
\end{aligned} \tag{4.3}$$

   - Compute the relative error $e = \dfrac{\|\widehat{V}^{n+1,k+1} - \widehat{V}^{n+1,k}\|}{\|\widehat{V}^{n+1,k+1}\|}$

   until $e < \varepsilon$.

4.3. Finite element method

For the spatial discretization of (4.3) a triangular mesh of Ω and the associated finite element space of piecewise linear Lagrange polynomials are considered. For fixed natural numbers $N_S > 0$ and $N_h > 0$, we consider a uniform mesh of the computational domain Ω, the nodes of which are (S_i, h_j), with $S_i = i\Delta S$ $(i = 0, \ldots, N_S + 1)$ and $h_j = j\Delta h$ $(j = 0, \ldots, N_h + 1)$, where $\Delta S = S_\infty/(N_S + 1)$ and $\Delta h = h_\infty/(N_h + 1)$ denote the constant mesh steps in each coordinate. Associated to this uniform mesh, a piecewise linear Lagrange finite element discretization is considered. More precisely, we introduce the finite element spaces

$$W_h = \{\varphi_h \in C(\Omega) \,/\, \widetilde{\varphi}|_{T_j} \in \mathcal{P}_1\,,\ \forall T_j \in \mathcal{T}\}\,,$$
$$\mathcal{K}_h = \{\varphi_h \in W_h \,/\, \varphi_h = \widehat{V} \text{ on } \Gamma_1^{*,+} \cup \Gamma_2^{*,-} \text{and } \varphi_h \geq G(S)\}\,,$$

in order to find $\widehat{V}_h^{n+1,k+1} \in \mathcal{K}_h$ satisfying the boundary conditions and such that:

$$\begin{aligned}
&\int_\Omega (1 + \Delta\tau^n f)\widehat{V}_h^{n+1,k+1}\big(\varphi_h - \widehat{V}_h^{n+1,k+1}\big)\,dS\,dh \\
&\quad + \Delta\tau^n \int_\Omega A\nabla\widehat{V}_h^{n+1,k+1}\,\nabla\big(\varphi_h - \widehat{V}_h^{n+1,k+1}\big)\,dS\,dh \\
&\quad - \Delta\tau^n \int_{\Gamma_2^{*,+}} (A\nabla V_h^{n+1,k+1}, n)(\varphi_h - \widehat{V}_h^{n+1,k+1})\partial\gamma \\
&\quad \geq \int_\Omega \big(\widehat{V}_h^n \circ \chi^n\big)\big(\varphi_h - \widehat{V}_h^{n+1,k+1}\big)\,dS\,dh - \Delta\tau^n \int_\Omega h\big(\widehat{V}_h^{n+1,k}\big)^+\big(\varphi_h - \widehat{V}_h^{n+1,k+1}\big)\,dS\,dh\,,
\end{aligned}$$

for all $\varphi_h \in \mathcal{K}_h$. Quadrature formula based on the midpoints of the edges of the triangles has been used to obtain the coefficients of the matrix and the right hand side vector which define the linear system associated to the discretized problem.

After the time discretization with the method of characteristics and the spatial discretization with finite elements, the fully discretized problem can be written in the form:

$$\begin{cases} A_h \widehat{V}_h^{n+1,k+1} \geq b_h^{n+1,k+1}, \\ \widehat{V}_h^{n+1,k+1} \geq \Psi_h, \\ (A_h \widehat{V}_h^{n+1,k+1} - b_h^{n+1,k+1})(\widehat{V}_h^{n+1,k+1} - \Psi_h) = 0, \end{cases} \tag{4.4}$$

where Ψ_h denotes the discretized exercise value, $G(S)$, which also coincides with the value at maturity.

In order to solve problem (4.4), the augmented Lagrangian active set (ALAS) algorithm is employed.

5. Numerical results

Finally, in order to show the relevance of incorporating counterparty risk pricing derivatives we show some numerical results to understand the behaviour of the total value adjustment for American options. We focus on an American put option sold by the investor. The maturity time is $T = 0.5$ years and is discretized with $N_T = 700$ time steps. Firstly, we plot the risky and risk-free derivative value and the XVA. Moreover, we present the exercise region for both derivative value in order to show how affects the counterparty risk in the early exercise.

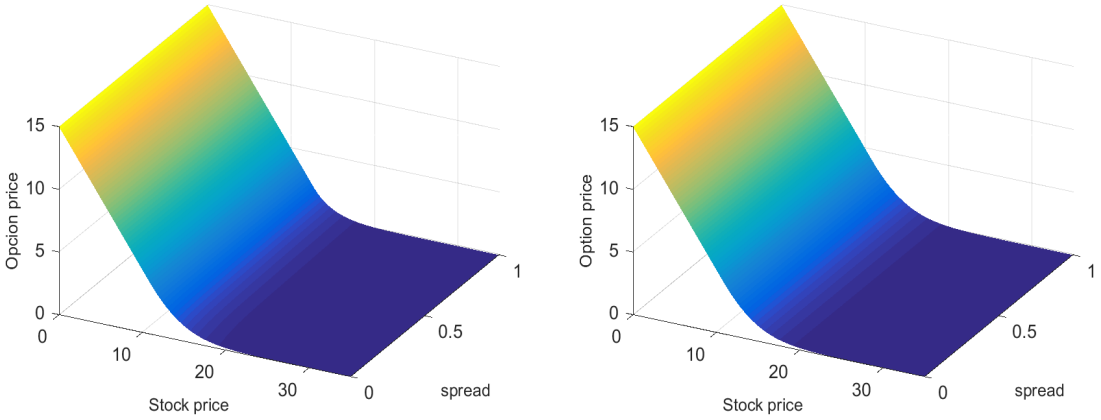


Fig. 1 American put option value risky valur (left), risk–free value (right)

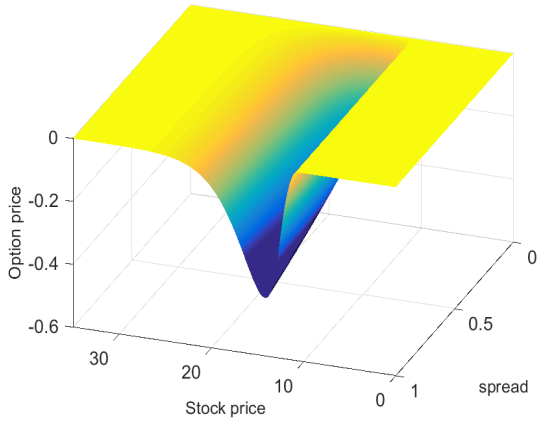


Fig. 2 Total value adjustment

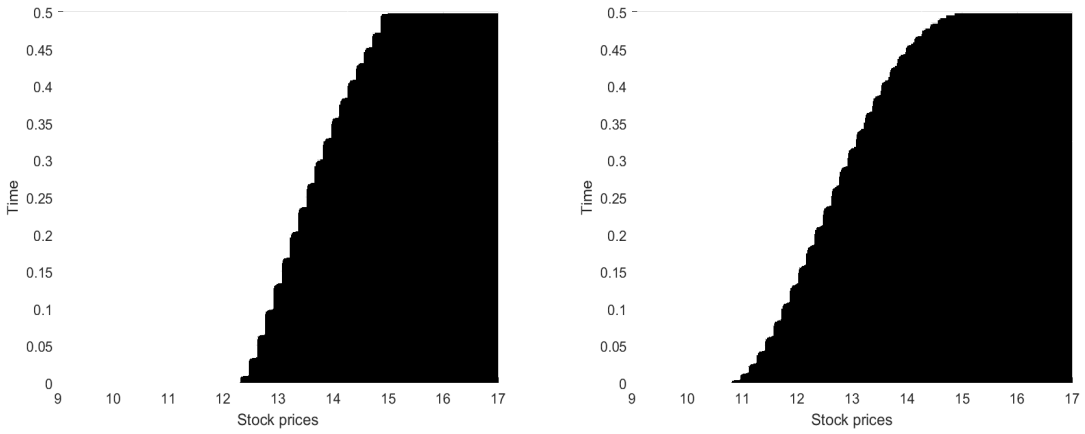


Fig. 3 Exercise regions (white) risky value (left) risk–free value (right)

References

[1] I. Arregui and B. Salvador and C. Vázquez A Monte Carlo approach to American options pricing including counterparty risk. International Journal of Computer Mathematics, 96(11): 2157–2176, 2019.

[2] I. Arregui and B. Salvador and C. Vázquez PDE models and numerical methods for total value adjustment in European and American options with counterparty risk. Applied Mathematics and Computation, 308: 31–53, 2017.

[3] I. Arregui and B. Salvador and D. Ševčovič and C. Vázquez Total value adjustment for European options with two stochastic factors. Mathematical model, analysis and numerical simulation. Computers & Mathematics with Applications, 76: 725–740, 2018.

[4] C. Burgard and M. Kjaer PDE representations of derivatives with bilateral counterparty risk and funding costs Journal of Credit Risk, 7(3): 1–19, 2011.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 51–57)

A numerical method to solve Maxwell's equations in 3D singular geometry

Franck Assous[1], **Irina Raichik**[2]
1. *Ariel University, Israel*
2. *Bar-Ilan University, Israel*

Abstract

We propose a new method to solve the 3D Maxwell equations in axisymmetric singular domains, containing reentrant corner or edges. By doing a Fourier analysis, one arrives to a sequence of singular problems set in 2D domains, and 3D solutions are computed by solving 2D problems, depending on a Fourier mode k. For each k, the solution is decomposed into a regular and a singular part. The regular part is computed with a finite element method. The singular part belongs to a finite-dimensional subspace and is computed by an appropriate numerical approach, *only* for the modes $k = 0, \pm 1, 2$. The total the solution is then reconstructed, based on a non stationary variational formulation. Numerical examples will be shown.

1. Introduction

This article is part of the efforts made in the framework of non-smooth problems, i.e. problems set in non convex curvilinear polyhedra: such domains containing reentrant edges, they generate singularities in Maxwell's equations solutions. From a more intuitive point of view, the term singularities means that such geometrical features can generate, in their vicinity, very strong electromagnetic fields, that have to be carefully handled and are often difficult to compute. Moreover, as shown in [2], the impossibility of correctly handling these singularities may have drastic consequences on the phenomenon one wants to model.

In this context, many methods have been proposed to compute the solution to the Maxwell equations. We can mention the edge finite element method, introduced by Nédélec [9], that has demonstrated efficiency for the static and eigenvalue problems. More recently, discontinuous Galerkin method has been introduced [8] and have been extensively studied since then. In [5], Brenner et al. have also proposed an adaptive finite element method that works in dimension two.

Nevertheless, it is interesting for some applications to have a *continuous approximation* of the solutions, that can capture both the curl and the divergence of the electromagnetic fields, for instance when coupling the Maxwell equations in other equations, like the Vlasov one, see [3]. But the latter works only in convex (curvilinear) polyhedra.

In this paper, we consider three-dimensional axisymmetric domains with non axisymmetric data. Due to the axisymmetric assumption, the singular computational domain can be reduced to a subset of $\mathbb{R}^2$. However, the data being *arbitrary*, i.e. not necessarily axisymmetric, the electromagnetic field and other vector quantities still belong to $\mathbb{R}^3$. Hence, we take advantage that the domain is transformed into a two-dimensional one, and based on a Fourier analysis in the third dimension, one arrives to a sequence of singular problems set in a 2D singular domain. We then derive a variational formulation from which we propose a finite element method to solve the problem and numerically compute the solution.

2. Setting of the problem

We consider an axisymmetric bounded and simply connected Lipschitz domain Ω in $\mathbb{R}^3$, with a boundary Γ, $\mathbf{n}$ being the unit outward normal to Γ. We denote by c and ε_0 the speed of light and the dielectric permittivity respectively.

Hence, the evolution of a time-dependent electromagnetic field $\mathbf{E}(\mathbf{x},t), \mathbf{B}(\mathbf{x},t)$ propagating in vacuum is governed by Maxwell's equations[1]:

$$\frac{\partial \mathbf{E}}{\partial t} - c^2 \mathbf{curl}\ \mathbf{B} = -\frac{1}{\varepsilon_0}\mathbf{J}, \tag{2.1}$$

$$\frac{\partial \mathbf{B}}{\partial t} + \mathbf{curl}\ \mathbf{E} = 0, \tag{2.2}$$

$$\mathrm{div}\ \mathbf{E} = \frac{\rho}{\varepsilon_0}, \tag{2.3}$$

$$\mathrm{div}\ \mathbf{B} = 0\,, \tag{2.4}$$

where $\rho(\mathbf{x},t)$ and $\mathbf{J}(\mathbf{x},t)$ are the charge and current densities, that depend on the space variable $\mathbf{x}$ and on the time variable t. These equations are supplemented with perfect conductor boundary conditions, and homogeneous initial conditions at initial time $t = 0$.

We assume now that the domain Ω is axisymmetric, limited by the surface of revolution Γ, and we denote by ω and γ_b their intersections with a meridian half-plane. The boundary $\partial\omega := \gamma$ corresponds to $\gamma_a \cup \gamma_b$, where either $\gamma_a = \emptyset$ when γ_b is a closed contour (i.e. Ω does not contain the axis), or γ_a is the segment of the axis lying between the extremities of γ_b, see Fig.1. The natural coordinates for this domain are the cylindrical coordinates (r,θ,z), with the basis vectors $(\mathbf{e}_r, \mathbf{e}_\theta, \mathbf{e}_z)$. A meridian half-plane is defined by the equation θ =constant, and (r,z) are Cartesian coordinates in this half-plane.

However, even if we assumed symmetry of revolution for the domain Ω, we *do not assumed* such a symmetry for the data. Consequently, the problem can not be reduced to a two-dimensional one by assuming that derivative with respect to the azimuthal variable θ vanishes, i.e. $\partial/\partial\theta = 0$, as made for example in [1]: we have to continue to deal with a three-dimensional problem.

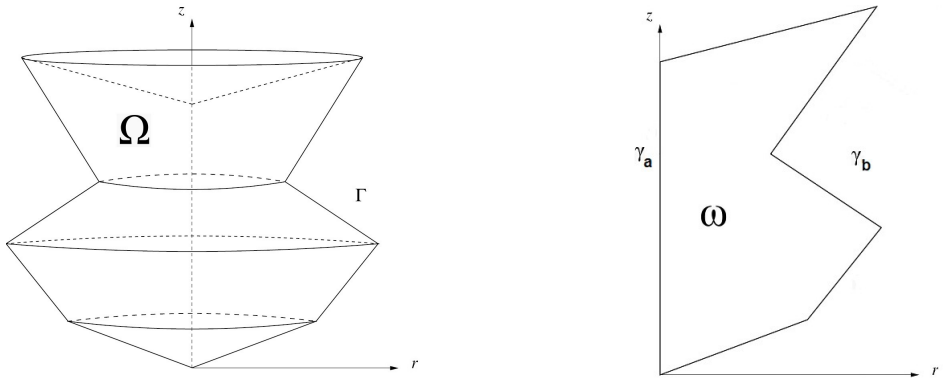


Fig. 1 Example of 3D domain Ω, and its corresponding 2D intersection with meridian half-plane ω.

Following [3], it is more efficient if one wishes to use nodal finite element methods, for instance for charge particle simulations as in the context of Vlasov-Maxwell computations, to eliminate the magnetic field $\mathbf{B}$ (respectively the electric field $\mathbf{E}$) from Eqs. (2.1-2.4). Hence, Maxwell's equations reduce to two second-order wave equations for each field separately:

$$\frac{\partial^2 \mathbf{E}}{\partial t^2} + c^2 \mathbf{curl}\,\mathbf{curl}\,\mathbf{E} = -\frac{1}{\varepsilon_0}\frac{\partial \mathbf{J}}{\partial t},$$

$$\frac{\partial^2 \mathbf{B}}{\partial t^2} + c^2 \mathbf{curl}\,\mathbf{curl}\,\mathbf{B} = \frac{1}{\varepsilon_0}\mathbf{curl}\,\mathbf{J}\,,$$

the constraints equations, namely divergence and boundary conditions, still holding.

3. Two-dimensional space reduction

As the data we consider are not axisymmetric, one can not perform $\partial/\partial\theta = 0$. However, one can use the cylindrical symmetry of the domain Ω to characterize the quantities defined on it, through their Fourier series in θ, the coefficients of which being functions defined on ω.

[1]In the text, names of function spaces of scalar fields usually begin by an italic letter, whereas they begin by a bold letter for spaces of vector fields.

Hence, we will consider, for a given vector field $\mathbf{w}(r,\theta,z)$

$$\mathbf{w}\,(r,\theta,z) = \frac{1}{\sqrt{2\pi}} \sum_{k\in\mathbb{Z}} \mathbf{w}^k\,(r,z)\,e^{ik\theta}\,,$$

and the following weighted Lebesgue space

$$L^2_r(\omega) := \left\{ w \text{ measurable on } \omega : \iint_\omega |w(r,z)|^2\, r dr dz < \infty \right\},$$

that will be the space of Fourier coefficients (at all modes) of functions[2] in $\mathbf{L}^2\,(\Omega)$.

At the same time, let us also define the space of relevant Fourier coefficients for the electromagnetic fields. It is easy to check that, for $\mathbf{w} \in \mathbf{H}\,(\mathrm{div}\,;\Omega)$, resp. $\mathbf{H}\,(\mathbf{curl}\,;\Omega)$, one has

$$\mathrm{div}\,\mathbf{w} = \frac{1}{\sqrt{2\pi}} \sum_{k\in\mathbb{Z}} \mathrm{div}_{\,k}\mathbf{w}^k e^{ik\theta} \;\text{ resp. }\; \mathbf{curl}\,\mathbf{w} = \frac{1}{\sqrt{2\pi}} \sum_{k\in\mathbb{Z}} \mathbf{curl}_{\,k}\mathbf{w}^k e^{ik\theta},$$

where the operators for each mode k are defined as

$$\mathrm{div}_{\,k}\mathbf{w} := \frac{1}{r}\frac{\partial\,(rw_r)}{\partial r} + \frac{ik}{r}w_\theta + \frac{\partial w_z}{\partial z}; \qquad (\mathbf{curl}_{\,k}\mathbf{w})_r := \frac{ik}{r}w_z - \frac{\partial w_\theta}{\partial z};$$

$$(\mathbf{curl}_{\,k}\mathbf{w})_\theta := \frac{\partial w_r}{\partial z} - \frac{\partial w_z}{\partial r}; \qquad (\mathbf{curl}_{\,k}\mathbf{w})_z := \frac{1}{r}\left(\frac{\partial\,(rw_\theta)}{\partial r} - ikw_r\right).$$

The regularity of $\mathbf{w}$ only depends on the regularity of its Fourier components $\mathbf{w}^k$, for $k \in \mathbb{Z}$. Let us now introduce the spaces for the **curl** and div operators

$$\mathbf{H}_0\,(\mathbf{curl}\,;\Omega) = \{\mathbf{v} \in \mathbf{H}\,(\mathbf{curl}\,;\Omega) : \mathbf{v}\times\mathbf{n}|_\Gamma = 0\} \;\text{ and }\; \mathbf{H}_0\,(\mathrm{div}\,;\Omega) = \{\mathbf{v} \in \mathbf{H}\,(\mathrm{div}\,;\Omega) : \mathbf{v}\cdot\mathbf{n}|_\Gamma = 0\}\,.$$

Hence, electric and magnetic field naturally belongs to the spaces

$$\mathbf{X}\,(\Omega) = \mathbf{H}_0\,(\mathbf{curl}\,;\Omega) \cap \mathbf{H}\,(\mathrm{div}\,\mathbf{v};\Omega) \;\text{ and }\; \mathbf{Y}\,(\Omega) = \mathbf{H}\,(\mathbf{curl}\,;\Omega) \cap \mathbf{H}_0\,(\mathrm{div}\,\mathbf{v};\Omega)$$

As a consequence, a function $\mathbf{v}$ belongs to $\mathbf{X}(\Omega)$ if and only if, for all $k \in \mathbb{Z}$, its Fourier coefficients $\mathbf{v}^k$ belong to the space $\mathbf{X}_{(k)}(\omega)$ defined by

$$\mathbf{X}_{(k)}(\omega) = \{\mathbf{v}^k \in \mathbf{L}^2_r(\omega),\; \mathbf{curl}_{\,k}\mathbf{v}^k \in \mathbf{L}^2_r(\omega)\,, \mathrm{div}_{\,k}\mathbf{v}^k \in \mathbf{L}^2_r(\omega)\,, \mathbf{v}^k\times\mathbf{n}_|\gamma_b = 0\}\,.$$

In a similar way, one introduces the space $\mathbf{Y}_{(k)}(\omega)$ for the Fourier coefficients of elements of $\mathbf{Y}(\Omega)$, namely

$$\mathbf{Y}_{(k)}(\omega) = \{\mathbf{v}^k \in \mathbf{L}^2_r(\omega),\; \mathbf{curl}_{\,k}\mathbf{v}^k \in \mathbf{L}^2_r(\omega)\,, \mathrm{div}_{\,k}\mathbf{v}^k \in \mathbf{L}^2_r(\omega)\,, \mathbf{v}^k\cdot\mathbf{n}_|\gamma_b = 0\}$$

A useful property concerning these spaces (see [6]) is that $\mathbf{X}_{(k)}(\omega)$ and $\mathbf{Y}_{(k)}(\omega)$ are independent of k, for $|k| \geq 2$. This allows us to compute the singular subspaces only for the modes $|k| \leq 2$, while the modes ± 2 will be use to compute all the higher modes $|k| > 2$.

Applying the dimension reduction, and using the linearity of the Maxwell equations together with the orthogonality of the different Fourier modes, we can reduce the 3D equations to a series of 2D formulations solved by the Fourier coefficients $(\mathbf{E}_k,\, \mathbf{B}_k)$, for each mode k, where the operators $\mathbf{curl}_{\,k}$ and $\mathrm{div}_{\,k}$ are involved. Let us introduce the operator $a_k(\cdot,\cdot)$ defined by

$$a_k(\mathbf{u},\mathbf{v}) = (\mathbf{curl}_{\,k}\mathbf{u}, \mathbf{curl}_{\,k}\mathbf{v}) + (\mathrm{div}_{\,k}\mathbf{u}, \mathrm{div}_{\,k}\mathbf{v})\,. \tag{3.1}$$

We get that each mode $\mathbf{E}^k$ is solution to the following variational formulation:
find $\boldsymbol{E}^k(t) \in \boldsymbol{X}_{(k)}(\omega)$ such that, for all $\boldsymbol{F} \in \boldsymbol{X}_{(k)}(\omega)$:

$$\frac{d^2}{dt^2}\left(\mathbf{E}^k(t),\mathbf{F}\right) + c^2\, a_k\left(\mathbf{E}^k(t),\mathbf{F}\right) = -\frac{1}{\varepsilon_0}\left(\partial_t\mathbf{J}^k,\mathbf{F}\right) + \frac{1}{\varepsilon_0}\left(\rho^k, \mathrm{div}_{\,k}\mathbf{F}\right), \tag{3.2}$$

[2]In the text, we shall also use the standard spaces and norms

where ρ^k and $\mathbf{J}^k$ denote the Fourier coefficients of the charge and current density ρ and $\mathbf{J}$ respectively, that depend (in space) only on (r, z).

In the same way one gets that the Fourier coefficients $\mathbf{B}_k(t)$ verify the variational formulation, for each mode k:
find $\boldsymbol{B}^k(t) \in \boldsymbol{Y}_{(k)}(\omega)$ *such that, for all* $\boldsymbol{C} \in \boldsymbol{Y}_{(k)}(\omega)$ *:*

$$\frac{d^2}{dt^2}\left(\mathbf{B}^k(t), \mathbf{C}\right) + c^2\, a_k\left(\mathbf{B}^k(t), \mathbf{C}\right) = \frac{1}{\varepsilon_0}\left(\mathbf{curl}_k \mathbf{J}^k, \mathbf{C}\right). \tag{3.3}$$

4. Decomposition in regular and singular parts

Due to the geometrical reduction, the geometrical singularities remain in the two-dimensional domain ω (see Figure 1). We briefly recall here some useful results helpful to understand the construction of the numerical method (see [2], [6] for details). As a first step, we introduce, for each Fourier mode k, the regular subspaces $\mathbf{X}^R_{(k)}$ and $\mathbf{Y}^R_{(k)}$, defined by:

$$\mathbf{X}^R_{(k)} := \mathbf{X}_{(k)} \cap \mathbf{H}^1_r(\omega)\,, \text{ and } \mathbf{Y}^R_{(k)} := \mathbf{Y}_{(k)} \cap \mathbf{H}^1_r(\omega)\,.$$

These subspaces are regular, in the sense that they coincide to the spaces of solutions in the case of a regular domain. Using now that $\mathbf{X}^R_{(k)}$ and $\mathbf{Y}^R_{(k)}$ are closed subspaces of $\mathbf{X}_{(k)}$ and $\mathbf{Y}_{(k)}$ respectively, we deduce the following decomposition

$$\mathbf{X}_{(k)} = \mathbf{X}^R_{(k)} \oplus \mathbf{X}^S_{(k)} \qquad \text{and} \qquad \mathbf{Y}_{(k)} = \mathbf{Y}^R_{(k)} \oplus \mathbf{Y}^S_{(k)}\,,$$

where $\mathbf{X}^S_{(k)}$ and $\mathbf{Y}^S_{(k)}$ are singular subspaces, equal to $\{0\}$ if the domain Ω (or equivalently ω) is regular.

The second step is to characterize these singular spaces, that have been proved to be finite dimensional. We have

Theorem 4.1 *The singular spaces* $X^S_{(k)}$ *and* $Y^S_{(k)}$ *are of finite dimension, namely*

- *For* $k = 0$

$$\dim Y^S_{(k)} := N_B = \textit{ number of reentrant edges,}$$
$$\dim X^S_{(k)} := N_E = N_B + \textit{ number of conical points with vertex angle } > \frac{\pi}{\beta},\ (\beta \simeq 1.3731)$$

- *For* $k \neq 0$

$$\dim Y^S_{(k)} := N_B = \dim X^S_{(k)} := N_E = \textit{ number of reentrant edges.}$$

From these properties, one can decompose, for each mode k, the electromagnetic field $(\mathbf{E}^k, \mathbf{B}^k)$ into a regular and a singular part, namely

$$(\mathbf{E}^k(t), \mathbf{B}^k(t)) = (\mathbf{E}^k_R(t), \mathbf{B}^k_R(t)) + (\mathbf{E}^k_S(t), \mathbf{B}^k_S(t))\,. \tag{4.1}$$

Moreover, since the singular spaces are of finite dimension, one can introduce their respective basis $(\mathbf{x}^k_{S,j})_{j=1,N_E}$ and $(\mathbf{y}^k_{S,j})_{j=1,N_B}$ for a given Fourier mode k. Using now that these basis are time independent, one can express the singular parts $\mathbf{E}^k_S(t)$ and $\mathbf{B}^k_S(t)$ as

$$\mathbf{E}^k_S(t) = \sum_{j=1}^{N_E} \mu^k_{E,j}(t)\mathbf{x}^k_{S,j} \qquad \text{and} \qquad \mathbf{B}^k_S(t) = \sum_{j=1}^{N_B} \mu^k_{B,j}(t)\mathbf{y}^k_{S,j}\,,$$

where $\mu^k_{E,j}(t)$ and $\mu^k_{B,j}(t)$ are smooth functions in time (at least continuous). As a consequence, the decomposition (4.1) of the electromagnetic, that will be useful for the numerical method, can be finally expressed, for each k,

$$\mathbf{E}^k(t) = \mathbf{E}^k_R(t) \oplus \sum_{j=1}^{j=N_E} \mu^k_{E,j}(t)\mathbf{x}^k_{S,j}, \qquad \mathbf{B}^k(t) = \mathbf{B}^k_R(t) \oplus \sum_{j=1}^{j=N_B} \mu^k_{B,j}(t)\mathbf{y}^k_{S,j}\,. \tag{4.2}$$

From a numerical point of view, as explained above, it is sufficient to compute them only for $k = -1, 0, 1, 2$. As these basis are not time-dependent, the computations will be carried out only once as an initialization procedure. This has been previously presented in [4] where details can be found.

5. Solving the time-dependent problem

In this section, we present the case of the magnetic field formulation. The electric field formulation is similar and can be derived in the same way. Therefore, we consider the variational formulation (3.3), in which we substitute the decomposition of the magnetic field (4.2) in regular and singular parts. Using that the singular basis $\mathbf{y}_{S,j}^k$ are time-independent, and denoting by $''$ the second derivative in time, we get

$$\frac{d^2}{dt^2}\left(\mathbf{B}_R^k(t),\mathbf{C}\right)+\sum_{j=1}^{N_B}(\mu_{B,j}^k)''\left(\mathbf{y}_{S,j}^k,\mathbf{C}\right)+c^2\,a_k\left(\mathbf{B}_R^k(t),\mathbf{C}\right)+c^2\sum_{j=1}^{N_B}\mu_{B,j}^k(t)\,a_k\left(\mathbf{y}_{S,j}^k,\mathbf{C}\right)$$
$$=\frac{1}{\varepsilon_0}\left(\mathbf{curl}_{\,k}\mathbf{J}^k,\mathbf{C}\right),\qquad\forall\mathbf{C}\in\mathbf{Y}_{(k)}^R(\omega)\,. \tag{5.1}$$

In addition, we add to the space of test functions $\mathbf{Y}_R(\omega)$ the fonctions $(\mathbf{y}_{S,j}^k)_{j=1,N_B}$. This yields the N_B additional equations

$$\frac{d^2}{dt^2}\left(\mathbf{B}_R^k(t),\mathbf{y}_{S,i}^k\right)+\sum_{j=1}^{N_B}(\mu_{B,j}^k)''\left(\mathbf{y}_{S,j}^k,\mathbf{y}_{S,i}^k\right)+c^2\,a_k\left(\mathbf{B}_R^k(t),\mathbf{y}_{S,i}^k\right)+c^2\sum_{j=1}^{N_B}\mu_{B,j}^k(t)\,a_k\left(\mathbf{y}_{S,j}^k,\mathbf{y}_{S,i}^k\right)$$
$$=\frac{1}{\varepsilon_0}\left(\mathbf{curl}_{\,k}\mathbf{J}^k,\mathbf{y}_{S,i}^k\right),\quad\forall\mathbf{y}_{S,i}^k\in\mathbf{Y}_{(k)}^S(\omega),1\le i\le N_B\,.$$

Moreover, using the orthogonality for each k of $\mathbf{Y}_{(k)}^R$ and $\mathbf{Y}_{(k)}^S$ with respect to the equivalent scalar product $a_k(\cdot,\cdot)$ defined by (3.1), we can eliminate the corresponding terms in the formulations above. This variational formulation is finally expresses as

Find $(\boldsymbol{B}_R^k,\boldsymbol{\mu}_B^k)\in\boldsymbol{Y}_{(k)}^R\times\mathbb{R}^{N_B}$ *such that*

$$\begin{cases}\left(\dfrac{\partial^2\mathbf{B}_R^k(t)}{\partial t^2},\mathbf{C}\right)+\displaystyle\sum_{j=1}^{N_B}(\mu_{B,j}^k)''\left(\mathbf{y}_{S,j}^k,\mathbf{C}\right)+c^2\,a_k\left(\mathbf{B}_R^k(t),\mathbf{C}\right)\\ \qquad\qquad=\dfrac{1}{\varepsilon_0}\left(\mathbf{curl}_{\,k}\mathbf{J}^k,\mathbf{C}\right),\quad\forall\mathbf{C}\in\mathbf{Y}_{(k)}^R(\omega)\,,\\ \left(\dfrac{\partial^2\mathbf{B}_R^k(t)}{\partial t^2},\mathbf{y}_{S,i}^k\right)+\displaystyle\sum_{j=1}^{N_B}(\mu_{B,j}^k)''\left(\mathbf{y}_{S,j}^k,\mathbf{y}_{S,i}^k\right)+c^2\sum_{j=1}^{N_B}\mu_{B,j}^k(t)\,a_k\left(\mathbf{y}_{S,j}^k,\mathbf{y}_{S,i}^k\right)\\ \qquad\qquad=\dfrac{1}{\varepsilon_0}\left(\mathbf{curl}_{\,k}\mathbf{J}^k,\mathbf{y}_{S,i}^k\right),\quad\forall\mathbf{y}_{S,i}^k\in\mathbf{Y}_{(k)}^S(\omega).\end{cases} \tag{5.2}$$

From a computational point of view, it is worth to rewrite the bilinear form $a_k(\cdot,\cdot)$ involved above, depending on the values of k. Performing a simple integration by parts shows that

$$\begin{aligned}a_k(\mathbf{u},\mathbf{v}) &= a_0(\mathbf{u}_m,\mathbf{v}_m)+k^2\left(\frac{\mathbf{u}_m}{r},\frac{\mathbf{v}_m}{r}\right)+(\mathbf{curl}\,u_\theta,\mathbf{curl}\,v_\theta)+k^2\left(\frac{u_\theta}{r},\frac{v_\theta}{r}\right)\\ &+ \imath kB(\mathbf{u},\mathbf{v}) \quad+\imath kC(\mathbf{u},\mathbf{v})\,,\end{aligned}$$

where $a_0(\cdot,\cdot)$ denotes the operator $a_k(\cdot,\cdot)$ for $k=0$ (namely in the "full" axisymmetric case), $\mathbf{u}_m:=(u_r,u_z)$ and the vector **curl** of a scalar field w is defined by

$$\mathbf{curl}\,w:=-\partial_z w\mathbf{e}_r+r^{-1}\partial_r(rw)\,\mathbf{e}_z\,.$$

In addition, the two bilinear forms $B(\mathbf{u},\mathbf{v})$ and $C(\mathbf{u},\mathbf{v})$ are defined by

$$B(\mathbf{u},\mathbf{v}):=\int_{\gamma_b}(\mathbf{u}_m\cdot\mathbf{n})\,\bar{v}_\theta-\mathbf{u}_\theta\,(\bar{\mathbf{v}}_m\cdot\mathbf{n})\,d\gamma\,,$$

and

$$C(\mathbf{u},\mathbf{v}):=\int\int_\omega 2\,(u_\theta\bar{v}_r-u_r\bar{v}_\theta)\,\frac{d\omega}{r}\,.$$

Note that the term $B(\mathbf{u},\mathbf{v})$ is vanishes as soon $\mathbf{u}\cdot\mathbf{n}=\mathbf{v}\cdot\mathbf{n}=0$, that is the case for the magnetic field, due to the perfect conductor boundary condition. The same is true if $\mathbf{u}\times\mathbf{n}=\mathbf{v}\times\mathbf{n}=0$, that is the case for the electric field. In addition, the term $C(\mathbf{u},\mathbf{v})$ is not singular despite the presence of $1/r$ in the integral. Indeed, only on the boundary γ_a one may have $r=0$, but $u_\theta=v_\theta=0$ (that is in practice B_θ^k or E_θ^k for the electric case) due the symmetry

condition on the axis γ_a.

Starting from this variational formulation, we are now ready to derive a finite element approximation. Let $\mathbf{Y}^{R,h}_{(k)} \subset \mathbf{Y}^{R}_{(k)}$ be the space of discretized test functions of dimension N_h. We actually used the P_2 finite element, and denote by $\mathcal{T}_h$ the mesh of ω made of triangles K_h. Then, the approximation space for the vector fields is made of functions which are component-wise P_2-conforming on the triangulation.

Let now $\mathbf{B}^{k,h}(t) = \mathbf{B}^{k,h}_R(t) + \sum_{j=1}^{N_B} \mu^{k,h}_{B,j}(t)\, \mathbf{y}^{k,h}_{S,j}$ be the discrete solution. After discretization in space, the semi-discretized variational formulation is written (with the addition of the index ^{h}) in the same way as (5.2). It can be expressed equivalently as a linear system:

$$\frac{d^2}{dt^2}\mathbb{M}_{rr} B^k_R \;+\; \mathbb{M}^k_{rs}\mu^k_B{}'' \;+\; c^2\mathbb{K}^k_{rr} B^k_R \;=\; \frac{1}{\varepsilon_0}\mathbb{R}^k_{rr} J^k \;, \tag{5.3}$$

$$\frac{d^2}{dt^2}\mathbb{M}^k_{sr} B^k_R \;+\; \mathbb{M}^k_{ss}\mu^k_B{}'' \;+\; c^2\mathbb{K}^k_{ss}\mu^k_B{}'' \;=\; \frac{1}{\varepsilon_0}\mathbb{R}^k_{sr} J^k \;, \tag{5.4}$$

where $\mathbb{M}_{rr}$ denotes the mass matrix that does not depend to the Fourier mode k, $\mathbb{M}^k_{rs}$ is a (N_h, N_B) rectangular matrix coming from the integral over ω of the product of the N_B singular functions $\mathbf{y}^{k,h}_{S,j}$ by the basis functions of $\mathbf{Y}^{R,h}_{(k)}$, $\mathbb{M}^k_{sr}$ being its transpose. Similarly, the matrix $\mathbb{K}^k_{rr}$ is associated to the term $a_k\left(\mathbf{B}^k_R(t), \mathbf{C}\right)\mathbb{R}^k_{rr}$ coming from the source term with $\mathbf{curl}_k \mathbf{J}^k$, and μ^k_B standing for the vector of $\mathbb{R}^{N_B}$ of entries $(\mu^k_{B,j})$. Finally, $\mathbb{M}^k_{ss}$ and $\mathbb{K}^k_{ss}$ are the "singular" mass and rigidity matrices of dimension (N_B, N_B), associated to the term $\left(\mathbf{y}^k_{S,j}, \mathbf{y}^k_{S,i}\right)$ and $a_k\left(\mathbf{y}^k_{S,j}, \mathbf{y}^k_{S,i}\right)$ respectively. For these singular matrices, the computation must be carried out precisely in the neighborhood of the singularities by using a quadrature formula of high order.

We then perform a time discretization involving a second-order explicit (leap-frog) scheme. Here the notation X^n (resp. X^{n+1}) stands for a variable X at time $t^n = n\Delta t$ (resp. $t^{n+1} = (n+1)\Delta t$), where Δt is the time-step. F^n, G^n, H^n is the set of quantities known at time t^n for each equation of the scheme (5.3)-(5.4), which can be rewritten as

$$\mathbb{M}_{rr} B^{k,n+1}_R + \mathbb{M}^k_{rs}\mu^{k,n+1}_B = F^{k,n} \;, \tag{5.5}$$

$$\mathbb{M}^k_{sr} B^{k,n+1}_R + \mathbb{M}^k_{ss}\mu^{k,n+1}_B = G^{k,n} \;. \tag{5.6}$$

To solve this linear system, a convenient way is to decouple $\mu^{k,n+1}_B$ and the unknown $B^{k,n+1}_R$as proposed in [2] for a two-dimensional Cartesian Maxwell system of equations, in the case of $N_B = 1$. The method developed here is more general, since it is also adapted to a domain with $N_B \geq 1$. For this purpose, we simply substitute (5.5)$-\mathbb{M}^k_{rs}(\mathbb{M}^k_{ss})^{-1}$(5.6) to obtain a system where $\mu^{k,n+1}_B$ does no appear anymore. It remains now to invert this system to compute $B^{k,n+1}_R$, and then, at the corresponding time, the value $\mu^{k,n+1}_B$ by solving (5.6).

Compared to the system one would obtained in a regular domain, the additional effort is essentially the computation of the matrix $(\mathbb{M}^k_{ss})^{-1}$. $\mathbb{M}^k_{ss}$ being a symmetric definite positive matrix (by construction) of dimension (N_B, N_B), i.e. a few units (and often $N_B = 1$), $(\mathbb{M}^k_{ss})^{-1}$ is very easy to compute once and for all, for any mode $k, |k| \leq 2$.

6. Numerical results

We present here numerical results to illustrate the method. For the sake of simplicity, we restrict ourselves to a domain with only one singular point. Hence, we will consider a 3-D top hat domain Ω with a reentrant circular edge, that corresponds, for a given θ, to an L-shaped 2-D domain ω with a reentrant corner. We introduce an unstructured mesh of ω made up of triangles, with no particular refinement near the reentrant corner. The variational formulations are approximated by a finite element method with FreeFem++ package [7]. The singular basis being computed as described in [4], we focus here on the computation of the time-dependent solutions. As in the previous section, we will concentrate on the magnetic case.

In addition, we assume that a perfectly conducting boundary condition is imposed on ω, and we want to numerically compute $\mathbf{B}^k(t) = \mathbf{B}^k_R(t) + \mu^k_B(t)\mathbf{y}^k_S$, assuming that singular basis $\mathbf{y}^k_S$ was already computed. More precisely, we are interested in computing the magnetic field $\mathbf{B}^k(t)$ created by a current loop, with initial conditions set to zero, and a current defined by $\mathbf{J}(t) = 10\sin(\lambda t)\mathbf{e}_\theta$, with a frequency $\lambda/2\pi = 2.5$GHz. The support of this current is a little disc centered around the middle of the domain. This current generates a wave that propagates circularly around the

current source. Physically, as long as the wave has not reached the reentrant corner, the field is smooth.

Let t_I be the impact time of the wave on the reentrant corner. Then, if one writes $\mathbf{B}^k(t) = \mathbf{B}^k_R(t) + \mu^k_B(t)\mathbf{y}^k_S$, $\mu^k_B(t) = 0$ for all t lower than t_I, and $\mathbf{B}^k(t)$ and $\mathbf{B}^k_R(t)$ coincide. On the other hand, for $t > t_I$, $\mu^k_B(t) \neq 0$ (and so $\mu^k_B(t)\mathbf{y}^k_S$ is) and the total field differs from its regular part.

This behavior is illustrated, for $k = 1$, on Figures 2 and 3. Similar results are obtained for other values of k.

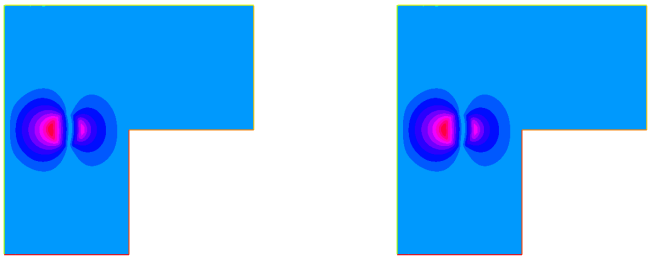


Fig. 2 $\mathbf{B}^1(t_1)$ and $\mathbf{B}^1_R(t_1)$, for $t_1 < t_I$ (case $k = 1$), z-component.

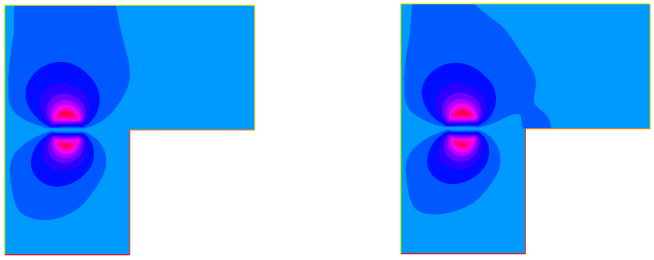


Fig. 3 $\mathbf{B}^1(t_2)$ and $\mathbf{B}^1_R(t_2)$, for $t_2 > t_I$ (case $k = 1$), r-component, in 2D and 3D view.

References

[1] F. Assous, P. Ciarlet, Jr., S. Labrunie, J. Segré, Numerical solution to the time-dependent Maxwell equations in axisymmetric singular domains:The Singular Complement Method, *J. Comput. Phys.*, 191, 147-176 (2003).

[2] F. Assous, P. Ciarlet, Jr., J. Segré, Numerical solution to the time-dependent Maxwell equations in two-dimensional singular domain: The Singular Complement Method, *J. Comput. Phys.*, 161, 218-249 (2000).

[3] F. Assous, P. Degond, E. Heintzé, P.A. Raviart, J. Segré, On a finite element method for solving the three-dimensional Maxwell equations, *J. Comput. Phys.*, 109, 222-237 (1993).

[4] F. Assous, I. Raichik, Numerical solution to the 3D Static Maxwell equations in axisymmetric singular domains with arbitrary data, *Comput. Meths. Applied Maths.*, 20 (3), 419-435 (2020).

[5] S.C. Brenner, J. Gedicke, L.-Y. Sung, An adaptive P1 finite element method for two-dimensional Maxwell's equations, *Journal of Scientific Computing*, 55, 738-754 (2013).

[6] P. Ciarlet Jr., S. Labrunie, Numerical solution of Maxwell's equations in axisymmetric domains with the Fourier singular complement method, *Diff. Eq. & Applic.* 3-1, 113-155 (2011).

[7] F. Hecht, New development in FreeFem++. *J. Numer. Math.* 20 (3-4): 251–265 (2012).

[8] J. S. Hestaven, T. Warburton, Nodal discontinuous Galerkin methods, Texts in Applied Mathematics 54, Springer (2008).

[9] J-C. Nédelec, Mixed finite elements in $\mathbb{R}^3$, *Numer. Math.*, 35, 315–341 (1980).

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 58–64)

Analysis of a SEIRS metapopulation model with fast migration

Pilar Atienza, Luis Sanz-Lorenzo
Universidad Politécnica de Madrid, Spain

Abstract

Metapopulation models for the study of a infectious disease in a population with space structure involve a large number of equations. Therefore the mathematical analysis of these models yields only partial results. We propose a model in which, as it is often the case in practical situations, the time scale of the transport of individuals is much master that that of the disease. Then we make use of approximate reduction techniques in order to reduce the system's dimension, and carry out a thorough analysis of the reduced model. In particular we characterize the number and stability of equilibria, provide conditions for the disease to become endemic (resp. die out) and show that certain counter-intuitive behaviors can arise.

1. Introduction

Classical deterministic epidemic models assume an homogeneous spatial distribution of individuals. However, travels of individuals between different regions have proved to have a great influence on the spatial spread of diseases. Therefore, given the characteristics of current society, in which most humans live in cities and travel along defined routes, it seems reasonable to include spatial variation into epidemic models.

The spatial spread of infectious diseases is a complex phenomenon to model. The usual approach in the literature is to use the so called metapopulation models, in which the population is distributed into discrete spatial sites, called patches, amongst which they may migrate. This movement of individuals is captured by a directed graph, in which the vertices represent the geographical regions and the arcs represent the connections between them.

Epidemic metapopulation models have been formulated and discussed in the literature for different diseases, see for example [2–4], yielding systems that consist of a large number of ordinary differential equations. This complexity greatly limits the analytical study that can be carried out, and only partial results have been obtained for these models. In particular, in [2], a SEIRS model with spatial distribution is formulated and studied. The expression of the basic reproduction number of the model is derived, but the existence and stability of endemic equilibria is only considered numerically, as an analytical approach seems unfeasible.

Furthermore, in many practical situations it can be assumed that travel of individuals between patches is much faster than the dynamics of the corresponding disease. For example, in the case of human diseases travel between different cities can be done in the span of a few hours, while the development of an infectious disease may take days or even weeks. This fact justifies the use of two different time scales to formulate the disease dynamics and the movement of individuals between sites.

One can make use of the existence of two different time scales in order to obtain a reduced approximation of the model. Indeed, in [7,8] approximate aggregation techniques are presented for the study of complex population dynamics in which two time scales are considered. Loosely speaking, this method consists in taking the fast process in the original model to its equilibrium value, yielding a system whose dimension can be reduced. These techniques allow us to obtain an approximated simplified model that, under the assumption that migration of individuals is sufficiently faster than the disease dynamics, behave qualitatively similar to the original model, while its dimension and, hence, complexity is highly reduced.

The objective of this work is to formulate a SEIRS metapopulation model, derive its approximate reduced version by means of aggregation techniques and carry out an exhaustive study of this simplified system. We aim to obtain stronger results than those already found in the literature for the case of metapopulation models without a time-scale approach.

In Section 2 the model is formulated, considering a metapopulation of r patches in which the local dynamics of the disease in each site is of SEIRS type, whereas migration of individuals among sites is linear. This model is described by a system of $4r$ differential equations. Next, under the assumption that migration of individuals between regions is fast with respect to the disease dynamics, we make use of approximate aggregation techniques in order to reduce the system's dimension. The resulting system consists of only 4 differential equations.

Section 3 carries out a thorough analytical study of the approximated model. The basic reproduction number $\mathcal{R}_0$ of the model is obtained by the next generation matrix method [1,6]. The analysis carried out shows that the model can only have two behaviours: if $\mathcal{R}_0 < 1$ the disease will die out in every region, whereas if $\mathcal{R}_0 > 1$ the disease will be globally endemic.

Moreover, the reduction on the systems dimension has allowed us to obtain results that can not be found in the literature on epidemic metapopulation models without a time-scale approach. In particular, some conclusions are given regarding the existence of endemic equilibria in relation to the value of the basic reproduction number.

Finally, in Section 4 we study the influence that the parameters of migration have on the behaviour of the disease. The analysis shows that the dynamics the disease would have in the isolated patches can vary under the existence of migration of individuals. In particular, under certain conditions a counter-intuitive scenario can arise: let us assume we have r separated sites and in all of them the disease dies out (resp. is endemic). Then if migration of individuals between them is allowed, then for certain values of the parameters regarding migration, the behaviour of the disease may change, i.e. it might become endemic (resp. die out) globally.

The system formulation and results are also particularized for a SIRS model.

2. Formulation and reduction of the model

We consider a population spread out among r different patches amongst which they can migrate, and affected by a disease. We assume that the local dynamics of the disease in each site follows a classical SEIRS model, whereas migration of individuals among sites is linear, see Figure 1.

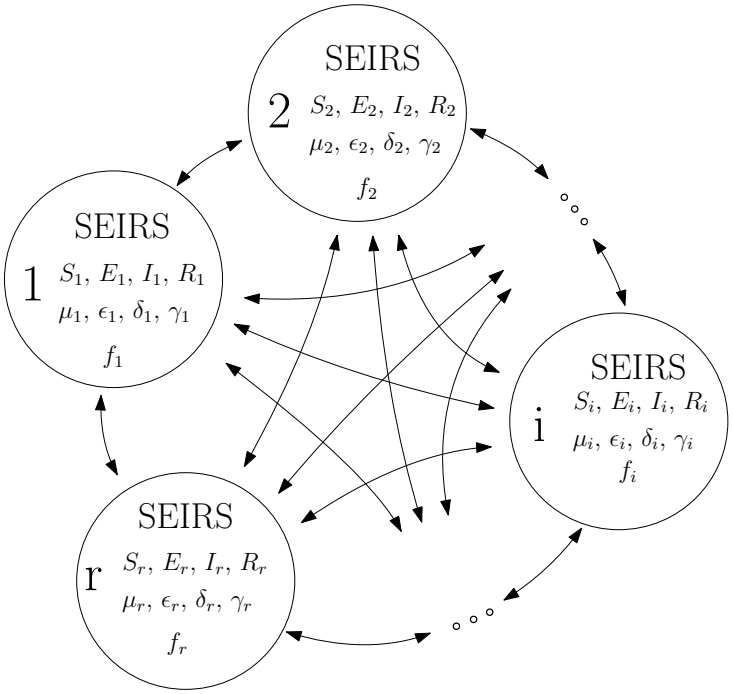


Fig. 1 SEIRS metapopulation model consisting of r interconnected patches.

This setting has been explored in several works [2–4] but the large dimension of the resulting model makes its analytical study very difficult. In our approach we make use of the fact that in many practical situations the time scale of migration is much faster than that of the epidemic process to formulate a two-time scale model.

The "slow time" variable, t, is used to describe the dynamics of the disease, whereas we denote the "fast time "variable as τ. Derivatives of a function g with respect this two different time variables are denoted as $\frac{dg}{dt} := \dot{g}$ and $\frac{dg}{d\tau} := g'$ respectively.

In order to formulate the model, we first consider both phenomena separately, and then combine them into the resulting complete system.

2.1. Local dynamics in each patch

Let S_i, E_i, I_i, and R_i denote the number of susceptible, exposed (infected but can not transmit the disease), infectious (infected and contagious) and recovered individuals who are present in patch i, $i = 1, ..., r$, at time t, respectively. The equations of a classical SEIRS model for patch i are:

$$\begin{aligned}
\dot{S}_i &= -f_i(S_i, E_i, I_i, R_i) + \mu_i N_i - \mu_i S_i + \delta_i R_i \\
\dot{E}_i &= +f_i(S_i, E_i, I_i, R_i) - (\epsilon_i + \mu_i) E_i \\
\dot{I}_i &= \epsilon_i E_i - (\gamma_i + \mu_i) I_i \\
\dot{R}_i &= \gamma_i I_i - (\delta_i + \mu_i) R_i
\end{aligned} \tag{2.1}$$

where the subscript i denotes the patch number each epidemiological variable refers to and the time derivatives are with respect to the slow time t. We denote by f_i the incidence function, over which for the moment we do not impose any condition, μ_i denotes the birth and the death rate of individuals (as it is the case in most models, both are assumed equal so that the total population in each patch remains constant), $1/\epsilon_i$ is the average latency time, $1/\gamma_i$ is the average recovery time for infectious individuals and $1/\delta_i$ is the average time of the immunity period for recovered individuals.

2.2. Migration model

Let us denote $\mathbf{X}(t) := (S_1, ..., S_r, E_1, ..., E_r, I_1, ..., I_r, R_1, ..., R_r)$,

For each $Y \in \{S, E, I, R\}$, let $\mathbf{Y}(t) := (Y_1, ..., Y_r)$ and let

$$Y := \sum_{i=1}^{r} Y_i \tag{2.2}$$

denote the total population in each of the epidemiological classes. We refer to S, E, I and R as "global variables".

We assume that migration between sites is a linear process. The migration rate (with respect to the "fast time") of individuals of each class $\alpha \in \{S, E, I, R\}$ from patch j to i, is denoted by m_{ij}^{α}, $i, j = 1, ..., r$. Therefore, the migration matrix for class α is

$$M^{\alpha} := \begin{pmatrix} -\sum\limits_{i=1, i\neq 1}^{r} m_{i1}^{\alpha} & m_{12}^{\alpha} & \cdots & m_{1r}^{\alpha} \\ m_{21}^{\alpha} & -\sum\limits_{i=1, i\neq 2}^{r} m_{i2}^{\alpha} & \cdots & m_{2r}^{\alpha} \\ \vdots & \vdots & \ddots & \vdots \\ m_{r1}^{\alpha} & m_{r2}^{\alpha} & \cdots & -\sum\limits_{i=1, i\neq r}^{r} m_{ir}^{\alpha} \end{pmatrix},$$

so that it is a Metzler matrix. The matrix characterizing migration for the whole population is then $M = \operatorname{diag}\left(M^S, M^E, M^I, M^R\right)$ and therefore migration dynamics is described by the equation $\mathbf{X}' = M\mathbf{X}$.

We assume that an individual initially present in any patch can (directly or indirectly) travel to any other patch. This assumption results in the fact that that digraph corresponding to the migration is strongly connected and therefore each migration matrix M^{α} is irreducible.

The following result characterizes the asymptotic behavior of migration:

Theorem 2.1 *For each $\alpha \in \{S, E, I, R\}$, let M^{α} be irreducible. Let $\mathbf{v}^{\alpha} > 0$ be the right eigenvector of matrix M^{α} associated to eigenvalue $\lambda = 0$ and normalized so that the sum of its components is 1. Then, if the initial condition $\mathbf{X}(0)$ contains at least an individual in each of the epidemiological clases $\{S, E, I, R\}$, then the dynamics of migration for the whole population tends to the equilibrium*

$$\mathbf{X}_e = \begin{pmatrix} \mathbf{v}^S S \\ \mathbf{v}^E E \\ \mathbf{v}^I I \\ \mathbf{v}^R R \end{pmatrix}, \tag{2.3}$$

where S, E, I and R are given by (2.2).

2.3. Complete model with two time scales

We proceed with the formulation of the complete model, that takes into account the joint effect of the local disease dynamics in each site and inter-site migration. In order to take into account the existence of two different time scales, we define the ratio of characteristic times $\varepsilon := t/\tau$, which under our hypotheses is a small positive number. Putting together the equations regarding the disease dynamics in each site (2.1) with those corresponding to migration, we obtain the following model:

$$\begin{aligned} \mathbf{S}' &= \varepsilon \left[-\mathbf{f}(\mathbf{X}) + D_{\mu}\mathbf{N} - D_{\mu}\mathbf{S} + D_{\delta}\mathbf{R}\right] + M^S\mathbf{S} \\ \mathbf{E}' &= \varepsilon \left[+\mathbf{f}(\mathbf{X}) - \left(D_{\epsilon} + D_{\mu}\right)\mathbf{E}\right] + M^E\mathbf{E} \\ \mathbf{I}' &= \varepsilon \left[D_{\epsilon}\mathbf{E} - \left(D_{\mu} + D_{\gamma}\right)\mathbf{I}\right] + M^I\mathbf{I} \\ \mathbf{R}' &= \varepsilon \left[D_{\gamma}\mathbf{I} - \left(D_{\delta} + D_{\mu}\right)\mathbf{R}\right] + M^R\mathbf{R}, \end{aligned} \tag{2.4}$$

where $\mathbf{f}(\mathbf{X}) := \begin{pmatrix} f_1(\mathbf{X}_1) \\ \vdots \\ f_r(\mathbf{X}_r) \end{pmatrix}$ and $D_{\sigma} := \operatorname{diag}(\sigma_1, ..., \sigma_r)$ for $\sigma \in \{\mu, \delta, \epsilon, \gamma\}$.

2.4. Reduction of the model

In order to carry out the analysis of model (2.4), we make use of the existence of different time scales and apply approximate aggregation techniques (see [7, 8] for a survey and a collection of the main results) in order to reduce its dimension. The reduction procedure consists on replacing the population vector with its equilibrium value for the migration process, i.e., replacing $\mathbf{X}(t)$ with $\mathbf{X}_e$ given by (2.3) and then summing the equations corresponding to each infectious class, in such a way that we obtain an autonomous model in the global variables S, E, I and R.

The resulting reduced or aggregated model is:

$$\begin{aligned}
\dot{S} =& - f(S,E,I,R) + \mu^I I + \mu^E E + \left(\mu^R + \delta\right) R \\
\dot{E} =& + f(S,E,I,R) - \left(\epsilon + \mu^E\right) E \\
\dot{I} =& \epsilon E - \left(\mu^I + \gamma\right) I \\
\dot{R} =& \gamma I - \left(\delta + \mu^R\right) R.
\end{aligned} \tag{2.5}$$

where

$$\begin{aligned}
f(S,E,I,R) &:= \sum_{i=1}^{r} f_i\left(v_i^S S, v_i^E E, v_i^I I, v_i^R R\right) \\
\mu^S &:= \sum_{i=1}^{r} \mu_i v_i^S \in \mathbb{R}\,,\ \mu^E := \sum_{i=1}^{r} \mu_i v_i^E \in \mathbb{R} \\
\mu^R &:= \sum_{i=1}^{r} \mu_i v_i^R \in \mathbb{R}\,,\ \mu^I := \sum_{i=1}^{r} \mu_i v_i^I \in \mathbb{R} \\
\delta &:= \sum_{i=1}^{r} \delta_i v_i^R \in \mathbb{R}\,,\ \gamma := \sum_{i=1}^{r} \gamma_i v_i^I \in \mathbb{R} \\
\epsilon &:= \sum_{i=1}^{r} \epsilon_i v_i^E \in \mathbb{R}
\end{aligned}$$

Note that the reduced model consists of only 4 differential equations, in contrast to the $4r$ equations that constitute the original metapopulation model (2.4).

The results in the field of approximate aggregation techniques allow us to claim that, loosely speaking, if the reduced model has an attractor which is locally structurally stable (in particular this holds for a hyperbolic equilibrium) and ϵ is small enough, i.e., if the separation of time scales between the disease dynamics and migration is large enough, then the original model also has a corresponding attractor which is $O(\epsilon)$-close and has the same stability properties. Therefore, the analysis of the reduced system provides relevant qualitative information about the behavior of the original model.

3. Analysis of the reduced model

In contrast with the original model (2.4), model (2.5) is amenable to an analytical study. In order to do so we will consider, as most epidemiological models do, the case of standard incidence in each patch, i.e., $f_i(S_i, E_i, I_i, R_i) = \beta_i \frac{I_i S_i}{N_i}$, and so we have

$$f(S,E,I,R) = \sum_{i=1}^{r} \beta_i \frac{v_i^S v_i^I}{v_i^S S + v_i^E E + v_i^I I + v_i^R R} IS.$$

It must be taken into account that the total population size ($N := S + E + I + R$) remains constant, so the analysis of the model can be reduced to 3 equations, for example those of E, I, R.

It is straightforward to show that for all non-negative initial conditions system (2.5) has an unique solution for all times and the solution remains non-negative. The study will be reduced to region

$$K := \{(E,I,R) : E \geq 0, I \geq 0, R \geq 0, E + I + R \leq N\}, \tag{3.1}$$

as only solutions inside that region have an epidemiological interpretation. It can be shown taht that K is positively invariant, i.e. solutions lie in K for all positive times given initial conditions inside that region. Axis $\{E = 0\} \cap \{I = 0\}$ is also positively invariant.

	Number of EE	Stability of DFE	Stability of EE	Case number
$\mathcal{R}_0 < 1$	0	GAS	-	1
$\mathcal{R}_0 > 1$	1	US	GAS (*)	2

Tab. 1 Summary of possible scenarios of aggregated model

(*) Not analytically proven.

The basic reproduction number of the model can be calculated by using the next generation matrix approach [1,6], yielding

$$\mathcal{R}_0 = \frac{\epsilon \sum_{i=1}^{r} \beta_i v_i^I}{\left(\epsilon + \mu^E\right)\left(\mu^I + \gamma\right)}. \tag{3.2}$$

Next, we study the stability of the disease free equilibrium (DFE) and the existence of endemic equilibria (EE) in relation to the value of $\mathcal{R}_0$.

Let $\bar{S} := S/N$, $\bar{E} := E/N$, $\bar{I} := I/N$, and $\bar{R} := R/N$ denote the proportions of individuals in each class. We define

$$g(\bar{I}) := \left(1 - \frac{\bar{I}}{H}\right) \sum_{i=1}^{r} \beta_i \frac{v_i^S v_i^I}{v_i^S + \left(v_i^I + b v_i^E + a v_i^R - \frac{v_i^S}{H}\right)\bar{I}}, \tag{3.3}$$

where

$$\sigma := \frac{\epsilon}{\left(\epsilon + \mu^E\right)\left(\gamma + \mu^I\right)}, \tag{3.4}$$

$$H := \frac{\epsilon\left(\delta + \mu^R\right)}{\epsilon\gamma + \left(\delta + \mu^R\right)\left(\epsilon + \gamma + \mu^I\right)}, \tag{3.5}$$

$$a := \frac{\gamma}{\delta + \mu^R}, \tag{3.6}$$

$$b := \frac{\gamma + \mu^I}{\epsilon}. \tag{3.7}$$

Then we have following result:

Theorem 3.1 *Let us consider the reduced model* (2.5)*, for which the basic reproduction number is given by* (3.2)*. Then*:

1. *If* $\mathcal{R}_0 < 1$ *the DFE is hyperbolic and globally asymptotically stable* (*GAS*) *in region K, so the disease will die out asymptotically for any initial condition. No endemic equilibrium exists in this case.*

2. *If* $\mathcal{R}_0 > 1$ *the DFE is hyperbolic and unstable* (*US*)*. There exists a unique endemic equilibrium with the form* $(\bar{E}_e, \bar{I}_e, \bar{R}_e)$*, where* $\bar{I}_e$ *is the only solution to the equation*

$$g(\bar{I}) = \frac{1}{\sigma} \tag{3.8}$$

in the interval $\bar{I}_e \in (0, H)$*, and* $\overline{E}_e = b\bar{I}_e$*,* $\overline{R}_e = a\bar{I}_e$*, where* $g(\bar{I})$*,* σ*, H, a and b are given by* (3.3-3.7)*.*

Furthermore, although we have not been able to showed it analytically, if $\mathcal{R}_0 > 1$ simulations suggest that the EE of the model is GAS in K, and so the disease will be endemic for any initial condition with infected individuals.

Table 1 summarizes the two possible scenarios. For the particular case of a two region metapopulation (i.e. $r = 2$), Figures 2 (a) and (b) show the three dimensional EIR phase diagrams for the two cases $\mathcal{R}_0 < 1$ and $\mathcal{R}_0 > 1$.

3.1. Particular case of a SIRS model

The classical SIRS model has the same structure than the SEIRS model, but without latency period, so the exposed class is not considered. Expressions for a SIRS model can be obtained from those of the SEIRS by letting $\epsilon \to \infty$. Therefore, the reduced model is

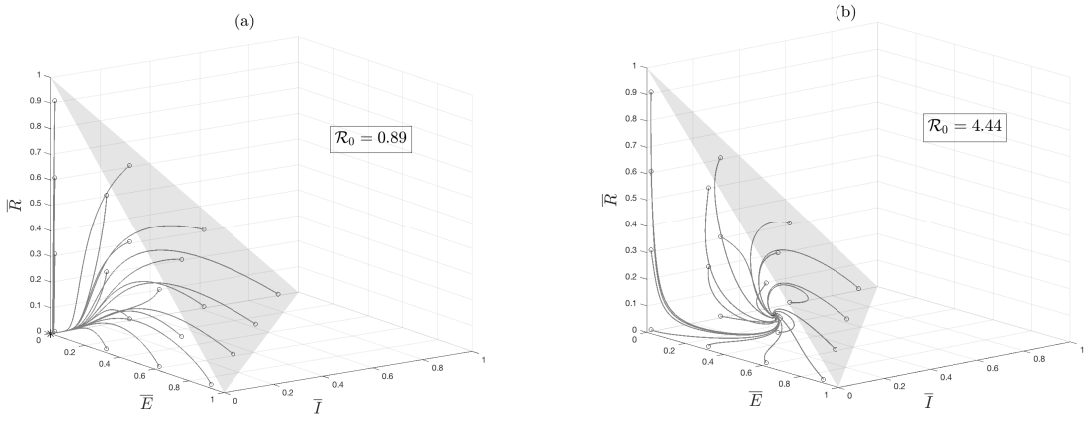


Fig. 2 (a) Case 1: DFE GAS. (b) Case 2: EE GAS.

$$\begin{aligned}
\dot{S} &= -f(S,E,I,R) + \mu^I I + \left(\mu^R + \delta\right) R \\
\dot{I} &= +f(S,E,I,R) - \left(\mu^I + \gamma\right) I \\
\dot{R} &= \gamma I - \left(\delta + \mu^R\right) R.
\end{aligned} \tag{3.9}$$

The basic reproduction number for this model is

$$\mathcal{R}_0 = \frac{\sum_{i=1}^{r} \beta_i v_i^I}{\left(\mu^I + \gamma\right)}, \tag{3.10}$$

and results of Theorem 3.1 apply.

4. Discussion: Influence of migration in the dynamics of the disease

One of the main purposes of this work is to evaluate how migration of individuals between sites can affect the behaviour of the disease. Namely, we aim to explore the possibility that an endemic behaviour of the disease on the isolated sites might be modified by the existence of migration between them.

First of all, it is important to observe that, according to the model, the disease will have the same behaviour in all interconnected regions, i.e., either the disease will be endemic in all of the sites, or it will die out globally. In the first case ($\mathcal{R}_0 > 1$), the asymptotic distribution of individuals of each class α between the different patches can be obtained by the eigenvector $\mathbf{v}^\alpha$, that, as stated in Theorem 2.3, provides the equilibrium distribution for migration of those individuals.

In addition, we want to study the possibility that, provided that the behaviour of the disease is the same in all patches separately, the migration of individuals amongst the sites could alter this behavior. Intuitively, we might consider that this can not happen, but in fact it can be proven that this counter-intuitive scenario may arise. That is, even when the basic reproduction number in every patch separately is $\mathcal{R}_0^i < 1$ (resp. $\mathcal{R}_0^i > 1$), the global basic reproduction number of the aggregated model could be $\mathcal{R}_0 > 1$ (resp. $\mathcal{R}_0 < 1$).

It is important to remark that this counter-intuitive situation is only possible under certain circumstances. In particular, it can not happen if the value of parameters μ_i, γ_i and ε_i is the same for all patches i, i.e., if the birth-death parameters and those regarding latency time, recovery time and incubation time are the same in all sites.

For the particular case of a two region metapopulation, Figure 3 shows an example of this situation, in which the disease dies out in both patches 1 and 2 if they are separated, but it becomes endemic if individuals can travel between the sites. For this case ($r = 2$), this situation arises if and only if the following conditions are satisfied:

$$\begin{aligned}
\max\left\{\frac{\epsilon_1}{\epsilon_1 + \mu_1}, \frac{\gamma_1 + \mu_1}{\beta_1}\right\} &< \frac{\epsilon_2}{\epsilon_2 + \mu_2} \\
\frac{\beta_2}{\beta_1} &< \frac{\gamma_2 + \mu_2}{\gamma_1 + \mu_1}.
\end{aligned}$$

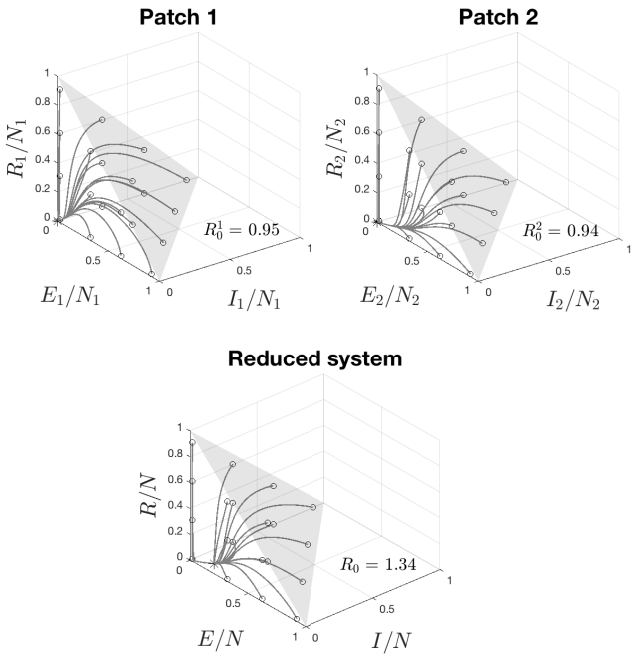


Fig. 3 Counter-intuitive scenario for a two region metapopulation

In the particular case of a SIRS model it can be proven that the counter-intuitive situation presented above cannot occur, i.e., if $\mathcal{R}_0^i < 1$ (resp. $\mathcal{R}_0^i > 1$) for all $i = 1, ..., r$, the global basic reproduction number of the reduced model is $\mathcal{R}_0 > 1$ (resp. $\mathcal{R}_0 < 1$).

In conclusion, our analysis shows that migration of individuals between different sites has a great influence on the behaviour of the disease, and we can quantify how the values of the parameters affecting migration affect this behavior.

Further study will be carried out in this field. In the first place, we want to generalize the study to other forms of incidence functions (amongst them mass action incidence) and analyse the more realistic situation in which the recruitment of individuals is not linear and, correspondingly, the local population in each patch is not necessarily constant.

Additionally, we will extend this analysis of models with time scales to SLIAR models [5], a family of models that can be used for the study of epidemics such as COVID-19.

References

[1] Antoine Perasso. An introduction to the basic reproduction number in mathematical epidemiology. *SESAIM: Proceedings and Surveys*, 62:123–138, 2018.

[2] Julien Arino. Diseases in metapopulations. *Series in contemporary applied mathematics*, 11:65-123, 2009.

[3] Julien Arino and P Van den Driessche. A multi-city epidemic model. *Mathematical Population Studies*, 10(3):175-193, 2003.

[4] Julien Arino and P Van den Driessche. Disease spread in metapopulations. *Nonlinear Dynamics and Evolution Equations, Fields Inst. Commun*, 48:1-13, 2006.

[5] Julien Arino, Fred Brauer, Pauline van den Driessche, James Watmough, and Jianhong Wu. Simple models for containment of a pandemic. *Journal of the Royal Society Interface*, 3(8):453–457, 2006.

[6] Pauline Van den Driessche and James Watmough. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Mathematical biosciences*, 180(1-2):29–48, 2002.

[7] Pierre Auger, R Bravo De La Parra, J-C Poggiale, E Sánchez, and TriNguyen-Huu. Aggregation of variables and applications to population dynamics. *Structured population models in biology and epidemiology*, pages 209–263. Springer, 2008.

[8] Pierre Auger, R Bravo de La Parra, Jean-Christophe Poggiale, E Sánchez,and L Sanz. Aggregation methods in dynamical systems and applications in population and community dynamics. *Physics of Life Reviews*, 5(2):79–105,2008.

[9] Zhihua Liu, Pierre Magal, Ousmane Seydi, and Glenn Webb. A model to predict COVID-19 epidemics with applications to South Korea, Italy, and Spain. *SIAM*, News, 1:1–2, 2020.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 65–72)

Goal-oriented adaptive finite element methods with optimal computational complexity

Roland Becker[1], **Gregor Gantner**[2], **Michael Innerberger**[3], **Dirk Praetorius**[3]
1. *Université de Pau et des Pays de l'Adour, France*
2. *University of Amsterdam, Korteweg-de Vries Institute for Mathematics, The Netherlands*
3. *TU Wien, Institute of Analysis and Scientific Computing, Austria*

Abstract

We consider a linear symmetric and elliptic partial differential equation (PDE) and a linear goal functional. We design a goal-oriented adaptive finite element method (GOAFEM), which steers the adaptive mesh-refinement as well as the approximate solution of the arising linear systems by means of a contractive iterative solver like the optimally preconditioned conjugate gradient method (PCG). We prove linear convergence of the proposed adaptive algorithm with optimal algebraic rates with respect to the number of degrees of freedom as well as the computational cost.

1. Our interest in a nutshell

Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain, $d \geq 2$. For given $f, \boldsymbol{f} \in L^2(\Omega)$ with $f(x) \in \mathbb{R}$ and $\boldsymbol{f}(x) \in \mathbb{R}^d$, we consider the linear symmetric and elliptic PDE

$$-\operatorname{div} \boldsymbol{A}\nabla u^\star = f + \operatorname{div} \boldsymbol{f} \quad \text{in } \Omega, \tag{1.1a}$$
$$u^\star = 0 \quad \text{on } \Gamma := \partial\Omega, \tag{1.1b}$$

where the right-hand side is understood in the distributional sense. As usual, we assume that $\boldsymbol{A} \in L^\infty(\Omega)$, where $\boldsymbol{A}(x) \in \mathbb{R}^{d\times d}_{\mathrm{sym}}$ is symmetric and uniformly positive definite so that the Lax–Milgram lemma proves existence and uniqueness of the so-called *primal solution* $u^\star \in \mathbb{V} := H^1_0(\Omega)$ to

$$a(u^\star, v) := \int_\Omega (\boldsymbol{A}\nabla u^\star)\cdot\nabla v \,\mathrm{d}x = \int_\Omega fv - \boldsymbol{f}\cdot\nabla v \,\mathrm{d}x =: F(v) \quad \text{for all } v \in \mathbb{V}. \tag{1.2}$$

Given $g, \boldsymbol{g} \in L^2(\Omega)$ with $g(x) \in \mathbb{R}$ and $\boldsymbol{g}(x) \in \mathbb{R}^d$, we aim to approximate the linear goal quantity $G(u^\star)$, where

$$G(v) = \int_\Omega gv - \boldsymbol{g}\cdot\nabla v \,\mathrm{d}x \quad \text{for all } v \in \mathbb{V}. \tag{1.3}$$

Then, the Lax–Milgram lemma also proves the existence and uniqueness of the so-called *dual solution* $z^\star \in \mathbb{V}$ to

$$a(v, z^\star) = G(v) \quad \text{for all } v \in \mathbb{V}. \tag{1.4}$$

Following [1,7,10], our numerical approach approximates $u^\star \approx u_\ell$ and $z^\star \approx z_\ell$ by discrete functions $u_\ell, z_\ell \in \mathbb{V}(\mathcal{T}_\ell)$ from the same conforming FEM space $\mathbb{V}(\mathcal{T}_\ell) \subseteq \mathbb{V}$. However, unlike [1, 7, 10], we do not assume that u_ℓ and z_ℓ are the exact Galerkin approximations of $u^\star$ and $z^\star$, respectively. Then, the discrete goal quantity

$$G_\ell(u_\ell, z_\ell) := G(u_\ell) + [F(z_\ell) - a(u_\ell, z_\ell)] \tag{1.5}$$

leads to the error estimate

$$G(u^\star) - G_\ell(u_\ell, z_\ell) = G(u^\star - u_\ell) - [F(z_\ell) - a(u_\ell, z_\ell)] = a(u^\star - u_\ell, z^\star - z_\ell) \leq |\!|\!| u^\star - u_\ell |\!|\!|\, |\!|\!| z^\star - z_\ell |\!|\!|, \tag{1.6}$$

where $|\!|\!| \cdot |\!|\!|$ is the $a(\cdot,\cdot)$-induced energy norm $|\!|\!| v |\!|\!|^2 := a(v, v)$.

Let $\|u^\star\|_{\mathbb{A}_s} + \|z^\star\|_{\mathbb{A}_t} < \infty$ denote that, starting from an initial mesh $\mathcal{T}_0$, the primal solution $u^\star$ can be approximated at algebraic rate $s > 0$ and the dual solution $z^\star$ can be approximated at algebraic rate $t > 0$ with respect to the number $\#\mathcal{T}_\ell$ of elements (see Section 3.2 for the precise definition). Due to estimate (1.6), a rate-optimal GOAFEM will then aim to generate a sequence $(\mathcal{T}_\ell)_{\ell\in\mathbb{N}_0}$ of meshes such that

$$\forall s, t > 0 \quad \Big[\|u^\star\|_{\mathbb{A}_s} + \|z^\star\|_{\mathbb{A}_t} < \infty \quad \Longrightarrow \quad \sup_{\ell\in\mathbb{N}_0} [\#\mathcal{T}_\ell]^{s+t}\,[G(u^\star) - G_\ell(u_\ell, z_\ell)] < \infty\Big], \tag{1.7}$$

i.e., the goal error decays with rate $s+t$ with respect to the number of elements (which usually is proportional to the number of the degrees of freedom of $\mathbb{V}(\mathcal{T}_\ell)$). Moreover, the computation of $u_\ell = u_\ell^m$ and $z_\ell = z_\ell^n$ will usually require $m = \underline{m}(\ell) \in \mathbb{N}$ solver steps of an iterative solver on the mesh $\mathcal{T}_\ell$ to compute u_ℓ and $n = \underline{n}(\ell) \in \mathbb{N}$ solver steps to compute z_ℓ. If each solver step is of linear cost, then a cost-optimal GOAFEM will even aim to generate a sequence $(\mathcal{T}_\ell)_{\ell\in\mathbb{N}_0}$ of meshes such that

$$\forall s,t>0 \quad \Big[\|u^\star\|_{\mathbb{A}_s} + \|z^\star\|_{\mathbb{A}_t} < \infty \quad\Longrightarrow\quad \sup_{\ell\in\mathbb{N}_0}\Big(\sum_{\ell'=0}^{\ell}[\underline{m}(\ell')+\underline{n}(\ell')]\,\#\mathcal{T}_{\ell'}\Big)^{s+t}[G(u^\star)-G_\ell(u_\ell,z_\ell)]<\infty\Big], \tag{1.8}$$

where we note that the computation of u_ℓ and z_ℓ depends on the full adaptive history. In explicit terms, (1.8) states that the goal error decays with rate $s+t$ with respect to the computational cost (and hence usually also with respect to the computational time). Clearly, cost-optimality (1.8) implies rate-optimality (1.7).

In our contribution, we report on our recent preprint [2], and present and discuss an adaptive strategy which, based on standard residual error estimators and an optimally preconditioned CG method, steers the local mesh-refinement as well as the termination of the iterative solver and is proven to be cost-optimal in the sense of (1.8).

2. Mathematical prerequisites

2.1. Mesh-refinement

Throughout, we employ newest vertex bisection for refining conforming simplicial triangulations [3, 4, 13]. For each conforming triangulation $\mathcal{T}_H$ and marked elements $\mathcal{M}_H \subseteq \mathcal{T}_H$, let $\mathcal{T}_h := \texttt{refine}(\mathcal{T}_H, \mathcal{M}_H)$ be the coarsest conforming triangulation, where all $T \in \mathcal{M}_H$ have been bisected. We write $\mathcal{T}_h \in \mathbb{T}(\mathcal{T}_H)$, if $\mathcal{T}_h$ results from $\mathcal{T}_H$ by finitely many steps of refinement.

For the later adaptive algorithm, let $\mathcal{T}_0$ be a given conforming triangulation of Ω. Let $\mathbb{T} := \mathbb{T}(\mathcal{T}_0)$ denote the set of all admissible triangulations and $\mathbb{T}(N) := \big\{\mathcal{T}_H \in \mathbb{T} : \#\mathcal{T}_H - \#\mathcal{T}_0 \le N\big\}$ for all $N \in \mathbb{N}_0$.

2.2. Discrete spaces and exact discrete solutions

For a conforming triangulation $\mathcal{T}_H$ of Ω and a polynomial degree $p \ge 1$, let

$$\mathbb{V}(\mathcal{T}_H) := \big\{v_H \in \mathbb{V} : \forall T \in \mathcal{T}_H \quad v_H|_T \text{ is a polynomial of degree} \le p\big\}. \tag{2.1}$$

Let $u_H^\star, z_H^\star \in \mathbb{V}(\mathcal{T}_H)$ be the exact Galerkin approximations of the primal and dual solution, i.e.,

$$a(u_H^\star, v_H) = F(v_H) \quad\text{and}\quad a(v_H, z_H^\star) = G(v_H) \quad \text{for all } v_H \in \mathbb{V}(\mathcal{T}_H), \tag{2.2}$$

where existence and uniqueness of $u_H^\star, z_H^\star \in \mathbb{V}(\mathcal{T}_H)$ follow again from the Lax–Milgram lemma.

2.3. Residual a-posteriori error estimator

To measure the discretization errors, we employ standard residual error estimators: For each $w_H \in \mathbb{V}(\mathcal{T}_H)$, define

$$\eta_H(w_H) := \eta_H(\mathcal{T}_H, w_H) \quad\text{with}\quad \eta_H(\mathcal{U}_H, w_H) := \Big(\sum_{T\in\mathcal{U}_H}\eta_H(T,w_H)^2\Big)^{1/2} \quad \text{for all } \mathcal{U}_H \subseteq \mathcal{T}_H, \tag{2.3a}$$

where the local contributions read

$$\eta_H(T,w_H)^2 := h_T^2\,\|f+\operatorname{div}\boldsymbol{f}+\operatorname{div}(\boldsymbol{A}\nabla w_H)\|_{L^2(T)}^2 + h_T\,\|[\![(\boldsymbol{A}\nabla w_H+\boldsymbol{f})\cdot\boldsymbol{n}]\!]\|_{L^2(\partial T\cap\Omega)}^2 \quad \text{for all } T\in\mathcal{T}_H. \tag{2.3b}$$

Here, $[\![(\cdot)\cdot\boldsymbol{n}]\!]$ denotes the normal jump across interior facets of the triangulation and $h_T := |T|^{1/d}$ denotes the local element size. We note that η_H requires additional regularity of $\boldsymbol{A}$ and $\boldsymbol{f}$, namely that $\boldsymbol{A}|_T \in W^{1,\infty}(T)$ and $\boldsymbol{f}|_T \in H(\operatorname{div};T)$ with $\boldsymbol{f}\cdot\boldsymbol{n}|_{\partial T} \in L^2(\partial T)$ for all $T\in\mathcal{T}_H$. With the same requirements on $\boldsymbol{g}$, define

$$\zeta_H(w_H) := \zeta_H(\mathcal{T}_H, w_H) \quad\text{with}\quad \zeta_H(\mathcal{U}_H, w_H) := \Big(\sum_{T\in\mathcal{U}_H}\zeta_H(T,w_H)^2\Big)^{1/2} \quad \text{for all } \mathcal{U}_H \subseteq \mathcal{T}_H, \tag{2.4a}$$

where the local contributions read

$$\zeta_H(T,w_H)^2 := h_T^2\,\|g+\operatorname{div}\boldsymbol{g}+\operatorname{div}(\boldsymbol{A}\nabla w_H)\|_{L^2(T)}^2 + h_T\,\|[\![(\boldsymbol{A}\nabla w_H+\boldsymbol{g})\cdot\boldsymbol{n}]\!]\|_{L^2(\partial T\cap\Omega)}^2 \quad \text{for all } T\in\mathcal{T}_H. \tag{2.4b}$$

It is well-known [3, 4] that the residual error estimator has the following four properties: For all triangulations $\mathcal{T}_H \in \mathbb{T}$, refinements $\mathcal{T}_h \in \mathbb{T}(\mathcal{T}_H)$, non-refined elements $\mathcal{U}_H \subseteq \mathcal{T}_H \cap \mathcal{T}_h$, and discrete functions $w_H \in \mathbb{V}(\mathcal{T}_H)$ and $w_h \in \mathbb{V}(\mathcal{T}_h)$, there holds the following:

(A1) stability: $|\eta_h(\mathcal{U}_H, w_h) - \eta_H(\mathcal{U}_H, w_H)| \le C_{\mathrm{stab}} \, |\!|\!| w_h - w_H |\!|\!|$, $|\zeta_h(\mathcal{U}_H, w_h) - \zeta_H(\mathcal{U}_H, w_H)| \le C_{\mathrm{stab}} \, |\!|\!| w_h - w_H |\!|\!|$;

(A2) reduction: $\eta_h(\mathcal{T}_h \backslash \mathcal{T}_H, w_H) \le q_{\mathrm{red}} \, \eta_H(\mathcal{T}_H \backslash \mathcal{T}_h, w_H)$, $\zeta_h(\mathcal{T}_h \backslash \mathcal{T}_H, w_H) \le q_{\mathrm{red}} \, \zeta_H(\mathcal{T}_H \backslash \mathcal{T}_h, w_H)$;

(A3) reliability: $|\!|\!| u^\star - u_H^\star |\!|\!| \le C_{\mathrm{rel}} \, \eta_H(u_H^\star)$, $|\!|\!| z^\star - z_H^\star |\!|\!| \le C_{\mathrm{rel}} \, \zeta_H(z_H^\star)$;

(A4) discrete reliability: $|\!|\!| u_h^\star - u_H^\star |\!|\!| \le C_{\mathrm{drel}} \, \eta_H(\mathcal{T}_H \backslash \mathcal{T}_h, u_H^\star)$, $|\!|\!| z_h^\star - z_H^\star |\!|\!| \le C_{\mathrm{drel}} \, \zeta_H(\mathcal{T}_H \backslash \mathcal{T}_h, z_H^\star)$.

Here, $0 < q_{\mathrm{red}} := 2^{-1/(2d)} < 1$ and the constants $C_{\mathrm{stab}}, C_{\mathrm{drel}} > 0$ depend only on the initial triangulation $\mathcal{T}_0$, the diffusion coefficient $\boldsymbol{A}$, and the polynomial degree $p \in \mathbb{N}$, while $C_{\mathrm{rel}} > 0$ is independent of p. Moreover, we note that, in the present setting, discrete reliability (A4) implies reliability (A3).

2.4. Contractive iterative solver and approximate discrete solutions

In our numerical experiments, we employ a preconditioned CG method [9] with optimal multilevel additive Schwarz preconditioner from [5,12] to approximately solve the discrete formulations (2.2). Since the condition numbers of the arising preconditioned linear systems are uniformly bounded, it is well-known that the preconditioned CG method leads to a uniform contraction for the solver. In explicit terms, given arbitrary initial guesses $u_H^0, z_H^0 \in \mathbb{V}(\mathcal{T}_H)$, the solver creates sequences $(u_H^m)_{m \in \mathbb{N}_0}$ and $(z_H^n)_{n \in \mathbb{N}_0}$ such that

$$|\!|\!| u_H^\star - u_H^m |\!|\!| \le q_{\mathrm{ctr}} \, |\!|\!| u_H^\star - u_H^{m-1} |\!|\!| \quad \text{and} \quad |\!|\!| z_H^\star - z_H^n |\!|\!| \le q_{\mathrm{ctr}} \, |\!|\!| z_H^\star - z_H^{n-1} |\!|\!| \quad \text{for all } m, n \in \mathbb{N}, \tag{2.5}$$

where the contraction constant $0 < q_{\mathrm{ctr}} < 1$ is independent of $\mathcal{T}_H$ and the initial guesses, but only depends on $\mathcal{T}_0$; see [12]. It follows from the triangle inequality that, for instance,

$$|\!|\!| u_H^\star - u_H^m |\!|\!| \le q_{\mathrm{ctr}} \, |\!|\!| u_H^\star - u_H^{m-1} |\!|\!| \le \frac{q_{\mathrm{ctr}}}{1 - q_{\mathrm{ctr}}} \, |\!|\!| u_H^m - u_H^{m-1} |\!|\!| \quad \text{for all } m \in \mathbb{N}. \tag{2.6}$$

Together with stability (A1) and reliability (A3), this leads to

$$\begin{aligned} |\!|\!| u^\star - u_H^m |\!|\!| \le |\!|\!| u^\star - u_H^\star |\!|\!| + |\!|\!| u_H^\star - u_H^m |\!|\!| &\le C_{\mathrm{rel}} \, \eta_H(u_H^\star) + |\!|\!| u_H^\star - u_H^m |\!|\!| \\ &\le C_{\mathrm{rel}} \, \eta_H(u_H^m) + (1 + C_{\mathrm{stab}}) \, |\!|\!| u_H^\star - u_H^m |\!|\!| \\ &\le C_{\mathrm{rel}} \, \eta_H(u_H^m) + (1 + C_{\mathrm{stab}}) \, \frac{q_{\mathrm{ctr}}}{1 - q_{\mathrm{ctr}}} \, |\!|\!| u_H^m - u_H^{m-1} |\!|\!|. \end{aligned} \tag{2.7}$$

We note that this estimate provides reliable a-posteriori error control for any approximate discrete solution u_H^m (with at least $m \ge 1$). On the one hand, this motivates to terminate the solver for some $m \ge 1$ as soon as

$$|\!|\!| u_H^m - u_H^{m-1} |\!|\!| \le \lambda_{\mathrm{ctr}} \, \eta_H(u_H^m)$$

for some user-prescribed parameter $\lambda_{\mathrm{ctr}} > 0$. On the other hand, this also motivates to employ nested iteration $u_h^0 := u_H^m$ if $\mathcal{T}_h = \texttt{refine}(\mathcal{T}_H, \mathcal{M}_H)$ and the iterative solver on $\mathcal{T}_H$ is terminated for u_H^m.

The same considerations apply to the iterative approximation of the discrete dual solution $z_H^\star \approx z_H^n$, i.e.,

$$|\!|\!| z^\star - z_H^n |\!|\!| \le C_{\mathrm{rel}} \, \eta_H(z_H^n) + (1 + C_{\mathrm{stab}}) \, \frac{q_{\mathrm{ctr}}}{1 - q_{\mathrm{ctr}}} \, |\!|\!| z_H^n - z_H^{n-1} |\!|\!|$$

and the solver is terminated for some $n \ge 1$ as soon as

$$|\!|\!| z_H^n - z_H^{n-1} |\!|\!| \le \lambda_{\mathrm{ctr}} \, \zeta_H(z_H^n).$$

2.5. A-posteriori control of the goal error

Recall the discrete goal quantity $G_H(u_H^m, z_H^n) = G(u_H^m) + [F(z_H^n) - a(u_H^m, z_H^n)]$ from the introduction. Combining (1.6) and (2.7), we see that

$$|G(u^\star) - G_H(u_H^m, z_H^n)| \le C'_{\mathrm{rel}} \big[\eta_H(u_H^m) + |\!|\!| u_H^\star - u_H^m |\!|\!| \big] \big[\zeta_H(z_H^n) + |\!|\!| z_H^\star - z_H^n |\!|\!| \big] \tag{2.8}$$

as well as

$$|G(u^\star) - G_H(u_H^m, z_H^n)| \le C''_{\mathrm{rel}} \big[\eta_H(u_H^m) + |\!|\!| u_H^m - u_H^{m-1} |\!|\!| \big] \big[\zeta_H(z_H^n) + |\!|\!| z_H^n - z_H^{n-1} |\!|\!| \big], \tag{2.9}$$

where $C'_{\mathrm{rel}} = \max\{C_{\mathrm{rel}}, \, 1 + C_{\mathrm{stab}}\}^2$ and $C''_{\mathrm{rel}} = \max\{C_{\mathrm{rel}}, \, (1 + C_{\mathrm{stab}}) \, \frac{q_{\mathrm{ctr}}}{1 - q_{\mathrm{ctr}}}\}^2$ for any approximate discrete solutions u_H^m and z_H^n with $m, n \ge 1$. We will use (2.8) to formulate linear convergence of the adaptive algorithm. Moreover, one can use (2.9) for a-posteriori error control of the goal error.

3. Cost-optimal adaptive algorithm

3.1. Adaptive algorithm

Algorithm 3.1 *Let $u_0^0, z_0^0 \in \mathbb{V}(\mathcal{T}_0)$ be initial guesses. Let $0 < \theta \le 1$ as well as $\lambda_{\mathrm{ctr}} > 0$ be fixed marking parameters. For all $\ell = 0, 1, 2, \dots$, perform the following steps* (i)–(vi):

(i) *Employ* (*at least one step of*) *the iterative solver to compute iterates $u_\ell^1, \dots, u_\ell^m$ and $z_\ell^1, \dots, z_\ell^n$ together with the corresponding refinement indicators $\eta_\ell(T, u_\ell^k)$ and $\zeta_\ell(T, z_\ell^k)$ for all $T \in \mathcal{T}_\ell$, until*

$$|||u_\ell^m - u_\ell^{m-1}||| \le \lambda_{\mathrm{ctr}}\, \eta_\ell(u_\ell^m) \quad \text{and} \quad |||z_\ell^n - z_\ell^{n-1}||| \le \lambda_{\mathrm{ctr}}\, \zeta_\ell(z_\ell^n). \tag{3.1}$$

(ii) *Define $\underline{m}(\ell) := m$, $\underline{n}(\ell) := n$, and $\underline{k}(\ell) := \max\{\underline{m}(\ell), \underline{n}(\ell)\}$.*

(iii) *If $\eta_\ell(u_\ell^m) = 0$ or $\zeta_\ell(z_\ell^n) = 0$, then define $\underline{\ell} := \ell$ and terminate.*

(iv) *Otherwise construct a set $\mathcal{M}_\ell \subseteq \mathcal{T}_\ell$ such that*

$$2\theta\, \eta_\ell(u_\ell^m)^2 \zeta_\ell(z_\ell^n)^2 \le \eta_\ell(\mathcal{M}_\ell, u_\ell^m)^2 \zeta_\ell(z_\ell^n)^2 + \eta_\ell(u_\ell^m)^2 \zeta_\ell(\mathcal{M}_\ell, z_\ell^n)^2. \tag{3.2}$$

(v) *Generate $\mathcal{T}_{\ell+1} := \mathtt{refine}(\mathcal{T}_\ell, \mathcal{M}_\ell)$.*

(vi) *Define the initial guesses $u_{\ell+1}^0 := u_\ell^m$ and $z_{\ell+1}^0 := z_\ell^n$.*

Remark 3.2 Given some $0 < \vartheta \le 1$, possible realizations of the marking strategy (3.2) are the following, which are all based on the Dörfler marking strategy proposed in [6]:
• The strategy proposed in [1] defines $\rho_\ell(T, u_\ell^m, z_\ell^n)^2 := \eta_\ell(T, u_\ell^m)^2 \zeta_\ell(z_\ell^n)^2 + \eta_\ell(u_\ell^m)^2 \zeta_\ell(T, z_\ell^n)^2$ for all $T \in \mathcal{T}_\ell$ and then determines $\mathcal{M}_\ell \subseteq \mathcal{T}_\ell$ such that

$$\vartheta\, \rho_\ell(u_\ell^m, z_\ell^n) \le \rho_\ell(\mathcal{M}_\ell, u_\ell^m, z_\ell^n), \tag{3.3}$$

where we employ the same abbreviated notation as in (2.3)–(2.4). This guarantees (3.2) with $\theta = \vartheta^2$.
• The strategy proposed in [10] determines $\mathcal{M}_\ell^{u+}, \mathcal{M}_\ell^{z+} \subseteq \mathcal{T}_\ell$ such that

$$\vartheta\, \eta_\ell(u_\ell^m) \le \eta_\ell(\mathcal{M}_\ell^{u+}, u_\ell^m) \quad \text{and} \quad \vartheta\, \zeta_\ell(z_\ell^n) \le \zeta_\ell(\mathcal{M}_\ell^{z+}, z_\ell^n) \tag{3.4}$$

and then picks $\mathcal{M}_\ell \in \{\mathcal{M}_\ell^{u+}, \mathcal{M}_\ell^{z+}\}$ with $\#\mathcal{M}_\ell = \min\{\#\mathcal{M}_\ell^{u+}, \#\mathcal{M}_\ell^{z+}\}$. This guarantees (3.2) with $\theta = \vartheta^2/2$.
• The strategy proposed in [7] determines $\mathcal{M}_\ell^{u+}, \mathcal{M}_\ell^{z+} \subseteq \mathcal{T}_\ell$ as in (3.4), but chooses $\mathcal{M}_\ell := \mathcal{M}_\ell^u \cup \mathcal{M}_\ell^z$, where $\mathcal{M}_\ell^u \subseteq \mathcal{M}_\ell^{u+}$ and $\mathcal{M}_\ell^z \subseteq \mathcal{M}_\ell^{z+}$ satisfy $\#\mathcal{M}_\ell^u = \#\mathcal{M}_\ell^z = \min\{\#\mathcal{M}_\ell^{u+}, \#\mathcal{M}_\ell^{z+}\}$. Again, this guarantees (3.2) with $\theta = \vartheta^2/2$. ∎

Remark 3.3 We note that, for fixed ℓ, the potential cost of one step of Algorithm 3.1 amounts to $\mathcal{O}(\underline{k}(\ell)\, \#\mathcal{T}_\ell)$:
• Performing one solver step of the preconditioned CG method in Algorithm 3.1(i) (for both u_ℓ^k and z_ℓ^k) with the optimal multilevel additive Schwarz preconditioner from [5, 12] is of computational cost $\mathcal{O}(\#\mathcal{T}_\ell)$. Computing the local contributions $\eta_\ell(T, u_\ell^k)$ and $\zeta_\ell(T, z_\ell^k)$ for all $T \in \mathcal{T}_\ell$ amounts to a computational cost $\mathcal{O}(\#\mathcal{T}_\ell)$.
• According to [11, 13], the marking strategies of Remark 3.2 allow to construct a set $\mathcal{M}_\ell$ in Algorithm 3.1(iv) satisfying (3.2) at computational cost $\mathcal{O}(\#\mathcal{T}_\ell)$, and the related sets (i.e., $\mathcal{M}_\ell$ in (3.3) resp. $\mathcal{M}_\ell^{u+}$ and $\mathcal{M}_\ell^{z+}$ in (3.4)) have even minimal cardinality [11] resp. up to a factor 2 minimal cardinality [13].
• Finally, also the local mesh-refinement in Algorithm 3.1(v) can be done at linear cost $\mathcal{O}(\#\mathcal{T}_\ell)$. ∎

3.2. Approximation class and rate-optimal convergence

We follow the seminal works [4, 13] and the abstract framework [3] to formulate a possible algebraic convergence rate for the approximation of the primal and dual problem. For any $s, t > 0$, let

$$\|u^\star\|_{\mathbb{A}_s} := \sup_{N \in \mathbb{N}_0} (N+1)^s \min_{\mathcal{T}_{\mathrm{opt}} \in \mathbb{T}(N)} \eta_{\mathrm{opt}}(u_{\mathrm{opt}}^\star) \in \mathbb{R}_{\ge 0} \cup \{\infty\}, \tag{3.5a}$$

$$\|z^\star\|_{\mathbb{A}_t} := \sup_{N \in \mathbb{N}_0} (N+1)^t \min_{\mathcal{T}_{\mathrm{opt}} \in \mathbb{T}(N)} \zeta_{\mathrm{opt}}(z_{\mathrm{opt}}^\star) \in \mathbb{R}_{\ge 0} \cup \{\infty\}, \tag{3.5b}$$

where $u_{\mathrm{opt}}^\star, z_{\mathrm{opt}}^\star \in \mathbb{V}(\mathcal{T}_{\mathrm{opt}})$ are the exact discrete solutions and $\eta_{\mathrm{opt}}(u_{\mathrm{opt}}^\star), \zeta_{\mathrm{opt}}(z_{\mathrm{opt}}^\star)$ are the corresponding error estimators. It is shown in [3] that the approximabilities (3.5) can equivalently be defined through the minimal total error (i.e., energy error plus data oscillations) instead of the error estimator.

By definition, $\|u^\star\|_{\mathbb{A}_s} + \|z^\star\|_{\mathbb{A}_t} < \infty$ yields that $\eta_{\mathrm{opt}}(u_{\mathrm{opt}}^\star) = \mathcal{O}((\#\mathcal{T}_{\mathrm{opt}})^{-s})$ and $\zeta_{\mathrm{opt}}(z_{\mathrm{opt}}^\star) = \mathcal{O}((\#\mathcal{T}_{\mathrm{opt}})^{-t})$ along respective sequences of optimal triangulations. In particular, one can thus hope for $|G(u^\star) - G_{\mathrm{opt}}(u_{\mathrm{opt}}^\star, z_{\mathrm{opt}}^\star)| \le C_{\mathrm{rel}}^2\, \eta_{\mathrm{opt}}(u_{\mathrm{opt}}^\star)\zeta_{\mathrm{opt}}(z_{\mathrm{opt}}^\star) = \mathcal{O}((\#\mathcal{T}_{\mathrm{opt}})^{-(s+t)})$, where the first estimate follows from (1.6) and reliability (A3). This convergence behavior is indeed proved by our main results presented and discussed in Section 3.3 and confirmed by the numerical experiments below.

3.3. Main results of own preprint [2]

For each adaptive level ℓ, Algorithm 3.1 performs at least one solver step to compute u_ℓ^m as well as one solver step to compute z_ℓ^n. By definition, $\underline{m}(\ell) \geq 1$ is the solver step for which the discrete solution $u_\ell^{\underline{m}(\ell)}$ is accepted (to contribute to the set of marked elements $\mathcal{M}_\ell$). Analogously, $\underline{n}(\ell) \geq 1$ is the solver step for which the discrete solution $z_\ell^{\underline{n}(\ell)}$ is accepted (to contribute to $\mathcal{M}_\ell$). If the iterative solver for either the primal or the dual problem fails to converge for some level $\ell \in \mathbb{N}_0$, i.e., (3.1) cannot be achieved for finite m, or n, we define $\underline{m}(\ell) := \infty$, or $\underline{n}(\ell) := \infty$, respectively, and $\underline{\ell} := \ell$. To simplify the notation, we define

$$\begin{aligned} u_\ell^k &:= u_\ell^{\underline{m}(\ell)} \quad \text{for all } k \in \mathbb{N} \text{ with } \underline{m}(\ell) < k \leq \underline{k}(\ell), \\ z_\ell^k &:= z_\ell^{\underline{n}(\ell)} \quad \text{for all } k \in \mathbb{N} \text{ with } \underline{n}(\ell) < k \leq \underline{k}(\ell). \end{aligned} \tag{3.6}$$

If Algorithm 3.1 does not terminate in step (iii) for some $\ell \in \mathbb{N}$, e.g., if neither the primal nor the dual solution can be resolved by a finite dimensional subspace of $H_0^1(\Omega)$, we define $\underline{\ell} := \infty$. To thoroughly formulate the convergence of Algorithm 3.1, we define the set

$$\mathcal{Q} := \big\{(\ell, k) \in \mathbb{N}_0^2 : \ell \leq \underline{\ell} \text{ and } 1 \leq k \leq \underline{k}(\ell)\big\} \quad \text{as well as} \quad |(\ell, k)| := k + \sum_{j=0}^{\ell-1} \underline{k}(j). \tag{3.7}$$

Note that $|(\ell, k)|$ is proportional to the overall number of solver steps needed to compute u_ℓ^k and z_ℓ^k. Moreover, $|(\cdot,\cdot)|$ also provides the natural ordering of $\mathcal{Q}$: If $(\ell, k), (\ell', k') \in \mathcal{Q}$ with $|(\ell', k')| \geq |(\ell, k)|$, then u_ℓ^k is computed earlier by Algorithm 3.1 than $u_{\ell'}^{k'}$.

Finally, recall from (2.8) the goal error estimate

$$(C'_{\mathrm{rel}})^{-1} |G(u^\star) - G_\ell(u_\ell^k, z_\ell^k)| \leq \big[\eta_\ell(u_\ell^k) + |\!|\!| u_\ell^\star - u_\ell^k |\!|\!|\big]\big[\zeta_\ell(z_\ell^k) + |\!|\!| z_\ell^\star - z_\ell^k |\!|\!|\big] =: \Lambda_\ell^k. \tag{3.8}$$

With these notations, there holds the following linear convergence result (stated in terms of the *quasi-error product* Λ_ℓ^k defined in (3.8)) for every choice of the adaptivity parameters. Clearly, this and the subsequent theorems are only of interest if the index set $\mathcal{Q}$ is countably infinite (i.e., Algorithm 3.1 does not terminate in step (iii)). However, we stress that the theorems hold for both cases $\underline{\ell} = \infty$ and $\underline{\ell} < \infty$ (which also covers the case of finite $\mathcal{Q}$).

Theorem 3.4 (full linear convergence) *For any $0 < \theta \leq 1$ and $\lambda_{\mathrm{ctr}} > 0$, Algorithm 3.1 guarantees that*

$$\Lambda_{\ell'}^{k'} \leq C_{\mathrm{lin}} q_{\mathrm{lin}}^{|(\ell',k')|-|(\ell,k)|}\, \Lambda_\ell^k \quad \textit{for all } (\ell, k), (\ell', k') \in \mathcal{Q} \textit{ with } |(\ell', k')| \geq |(\ell, k)|, \tag{3.9}$$

where the constants $C_{\mathrm{lin}} > 0$ and $0 < q_{\mathrm{lin}} < 1$ depend only on C_{stab}, C_{rel}, q_{ctr}, and the (arbitrary) adaptivity parameters θ and λ_{ctr}. ∎

As is observed in [2], full linear convergence is indeed the key argument to link rates with respect to the number of elements (1.7) and rates with respect to the overall computational cost (1.8): Recall that any step $(\ell, k) \in \mathcal{Q}$ of Algorithm 3.1 depends on the full history of preceding steps. Under realistic assumptions on the single steps of Algorithm 3.1 (see Remark 3.3), the total work spent to compute u_ℓ^k and z_ℓ^k is hence of order

$$\mathtt{work}(\ell, k) := \sum_{\substack{(\ell',k') \in \mathcal{Q} \\ |(\ell',k')| \leq |(\ell,k)|}} \#\mathcal{T}_{\ell'}. \tag{3.10}$$

With this understanding and interpretation, there holds the following result, which proves that the convergence rate $r > 0$ with respect to the number of elements will coincide with the convergence rate with respect to the overall computational cost (and hence with respect to the overall computational time). In particular, this proves that (1.7) and (1.8) are indeed equivalent in the present setting. We stress that, unlike the rough statement in the introduction, the theorem covers all iterates (u_ℓ^k, z_ℓ^k) for $(\ell, k) \in \mathcal{Q}$ instead of only the final iterates $(u_\ell^{\underline{m}(\ell)}, z_\ell^{\underline{n}(\ell)})$.

Theorem 3.5 (identification of convergence rates) *Let $r > 0$ and define $C_r := \sup_{(\ell,k)\in\mathcal{Q}} (\#\mathcal{T}_\ell - \#\mathcal{T}_0 + 1)^r \Lambda_\ell^k \in \mathbb{R}_{\geq 0} \cup \{\infty\}$. Then, it holds that*

$$C_r \leq \sup_{(\ell,k)\in\mathcal{Q}} (\#\mathcal{T}_\ell)^r \Lambda_\ell^k \leq \sup_{(\ell,k)\in\mathcal{Q}} \mathtt{work}(\ell, k)^r \Lambda_\ell^k \leq C_{\mathrm{rate}} C_r, \tag{3.11}$$

where $C_{\mathrm{rate}} > 0$ depends only on r, $\#\mathcal{T}_0$, and the constants C_{lin} and q_{lin} from Theorem 3.4. ∎

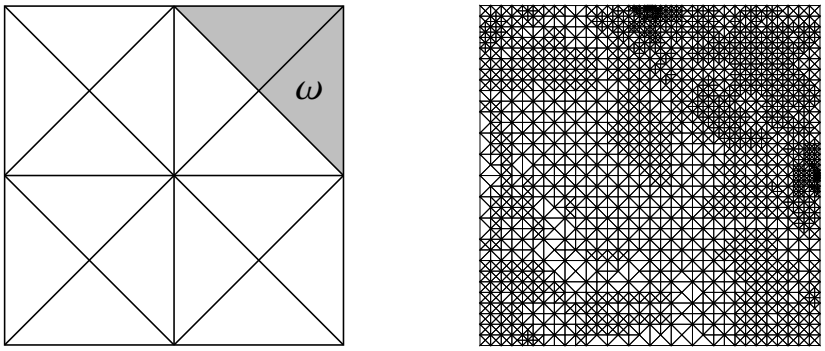


Fig. 1 Left: Initial mesh $\mathcal{T}_0$ for numerical experiments, where the set ω of (4.2) is highlighted in gray. Right: Mesh after 14 iterations of Algorithm 3.1 with $\#\mathcal{T}_{14} = 4.157$ elements.

Finally, this allows to formulate our main result, which states that Algorithm 3.1 leads to optimal convergence rates with respect to the number of elements and hence also with respect to the computational cost provided that the adaptivity parameters are sufficiently small. Indeed, this provides the formal statement of (1.7)–(1.8). In explicit terms, Algorithm 3.1 will lead to any possible rate $s+t$ with respect to the computational cost, if the primal problem allows for rate s and the dual problem allows for rate t with respect to the number of elements along respective sequences of optimal meshes.

Theorem 3.6 (optimal convergence) *Define* $\lambda_\star := \frac{1-q_{\mathrm{ctr}}}{q_{\mathrm{ctr}} C_{\mathrm{stab}}}$. *Let* $0 < \theta \le 1$ *and* $0 < \lambda_{\mathrm{ctr}} < \lambda_\star$ *be sufficiently small such that*

$$0 < \left(\frac{\sqrt{2\theta} + \lambda_{\mathrm{ctr}}/\lambda_\star}{1 - \lambda_{\mathrm{ctr}}/\lambda_\star} \right)^2 < \frac{1}{1 + C_{\mathrm{stab}}^2 C_{\mathrm{drel}}^2}. \tag{3.12}$$

Suppose that the set $\mathcal{M}_\ell$ *in Algorithm* 3.1(iv) *has up to some fixed constant* C_{mark} *minimal cardinality with* (3.2) (*e.g., by using one of the strategies from Remark* 3.2). *Let* $s, t > 0$ *with* $\|u^\star\|_{\mathbb{A}_s} + \|z^\star\|_{\mathbb{A}_t} < \infty$. *Then, there exists a constant* $C_{\mathrm{opt}} > 0$ *such that*

$$\sup_{(\ell,k)\in\mathcal{Q}} \mathtt{work}(\ell,k)^{s+t} \Lambda_\ell^k \le C_{\mathrm{opt}} \max\left\{ \|u^\star\|_{\mathbb{A}_s} \|z^\star\|_{\mathbb{A}_t}\,,\, \Lambda_0^0 \right\}. \tag{3.13}$$

The constant C_{opt} *depends only on* $C_{\mathrm{stab}}, C_{\mathrm{drel}}, q_{\mathrm{ctr}}, C_{\mathrm{mark}}, \theta, \lambda_{\mathrm{ctr}}, \#\mathcal{T}_0, s, t$ *and on fine properties of mesh-refinement by newest vertex bisection.* ∎

4. Numerical experiments

We underline our theoretical results by a numerical example. To this end, we consider the model problem

$$-\Delta u^\star = 2x_1(x_1-1) + 2x_2(x_2-1) \quad \text{in } \Omega := (0,1)^2, \qquad u^\star = 0 \quad \text{on } \Gamma := \partial\Omega \tag{4.1}$$

together with $\omega := \{x \in \Omega : x_1 + x_2 \ge 3/2\}$ and the goal functional

$$G(v) := \int_\omega \frac{\partial v}{\partial x_1}\,\mathrm{d}x, \quad \text{for all } v \in H_0^1(\Omega). \tag{4.2}$$

Hence, the given data read $f = 2x_1(x_1-1) + 2x_2(x_2-1)$, $\boldsymbol{f} = 0$, $g = 0$, and $\boldsymbol{g} = (-1,0)^\top \chi_\omega$, and the exact solution to (4.1) is $u^\star = x_1 x_2 (x_1-1)(x_2-1)$ with goal value $G(u^\star) = 11/960$.

We compute a numerical solution to (4.1)–(4.2) with Algorithm 3.1, where we set $u_0^0 = z_0^0 = 0$, $\theta = 0.25$, and, if not stated otherwise, $\lambda_{\mathrm{ctr}} = 4^{-6} \approx 2 \cdot 10^{-4}$. The initial triangulation $\mathcal{T}_0$ of Ω as well as a triangulation resulting from several steps of Algorithm 3.1 are shown in Figure 1. As marking strategy, we employ the one from [1], which is outlined in (3.3) in Remark 3.2.

In our experiments, we compare two iterative solvers: first, a non-preconditioned conjugate gradient method [9], which is denoted by CG; second, a preconditioned conjugate gradient method with optimal multilevel additive Schwarz preconditioner [12], which is denoted by PCG. We evaluate data obtained of Algorithm 3.1 as well as from a slight modification, where we do not employ nested iteration, but start from scratch at each mesh level ℓ, i.e., Algorithm 3.1(vi) is replaced by:

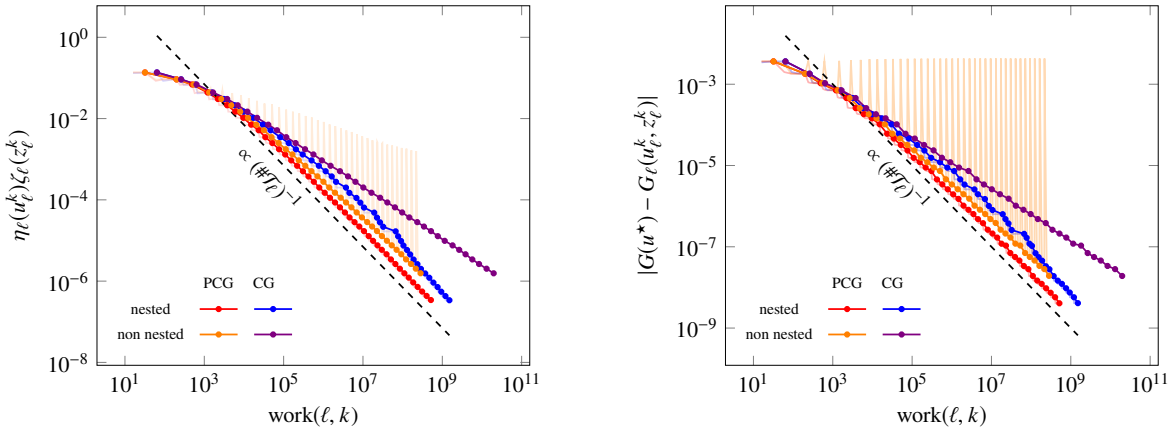


Fig. 2 Error measures over total computational cost. The markers represent the values for the final iterates $u_\ell^{m(\ell)}$ and $z_\ell^{n(\ell)}$ on each level. The transparent lines in the background correspond to the values for all iterates u_ℓ^k and z_ℓ^k. For CG without nested iteration, this line is not shown. Note that, for the data with nested iteration, the data for general iterates stays within close vicinity of that for the final ones.

(vi') Define the initial guesses $u_{\ell+1}^0 = z_{\ell+1}^0 = 0$.

In Figure 2, we report on the values of the estimator product $\eta_\ell(u_\ell^k)\zeta_\ell(z_\ell^k)$ and the goal error $|G(u^\star) - G_\ell(u_\ell^k, z_\ell^k)|$ over the overall computational cost work(ℓ, k). The results in Figure 2 are in accordance with Theorem 3.5, which states that, for the optimally preconditioned conjugate gradient method with nested iteration, the goal error decays with optimal rate over work(ℓ, k). For the non-preconditioned CG and also without nested iteration, Theorem 3.5 cannot make any statement. Indeed, our data suggest that in this case the rate is worse than that of PCG with nested iteration and hence not optimal.

From Figure 2, one also recognizes why nested iteration provides an advantage with respect to the computational cost. While for the algorithms with nested iteration the data for all iterates u_ℓ^k and z_ℓ^k (shown as transparent lines in the background) stays in close proximity to that of the final iterates $u_\ell^{m(\ell)}$ and $z_\ell^{n(\ell)}$, this is not the case for PCG without nested iteration, where the data for all iterates highly oscillate. Note that, in this case, the estimator product of the initial iterates $\eta_\ell(u_\ell^0)\zeta_\ell(z_\ell^0)$ goes down only because of the mesh-width weighting of the residual error estimates, while such an effect is not present for the goal error.

The cause for the difference in computational cost of PCG and CG becomes clear if one looks at the number of solver steps on each level that are necessary to reach the stopping criterion (3.1) for primal and dual solution, respectively. This is shown in Figure 3 (left). We see that, for PCG with nested iteration, the number of solver steps on each level is uniformly bounded, whereas for the other methods it further increases with increasing numbers of degrees of freedom. For the CG method with nested iteration, the number of necessary steps also seems relatively low. However, we stress that, in this case, there is no guaranteed upper bound, as is evident from the occasional peaks in the iteration count.

As a last observation, we give numerical evidence on the necessity of the definition of the discrete goal $G_\ell(u_\ell^k, z_\ell^k)$ from (1.5). To this end, we compare the algebraic rates of the quantities $|G(u^\star) - G_\ell(u_\ell^k, z_\ell^k)|$ (goal error), $|G(u^\star) - G(u_\ell^k)|$ (naïve goal error), $|F(z_\ell^k) - a(u_\ell^k, z_\ell^k)|$ (corrector term of discrete goal), and $\eta_\ell(u_\ell^k)\zeta_\ell(z_\ell^k)$ (estimator product, which is an upper bound of the goal error for the final iterates). These data are obtained by fitting the experimental data to a parametric model $p_1 \text{work}(\ell, k)^{-p_2}$ with $p_1, p_2 \geq 0$ with the MATLAB function `robustfit` applied to the natural logarithm of the data. The experimental convergence rates p_2 for different values of the solver parameter λ_{ctr} are shown in Figure 3 (right). It is evident that the rates of the corrector term are much lower than for the goal error and the estimator product. For large values of λ_{ctr}, the value of the corrector term is clearly large enough to disturb the rate of the naïve goal error, which makes it an unsuitable error quantity in this case.

For further experiments, we refer to the preprint [2].

Acknowledgements

The authors thankfully acknowledge support by the Austrian Science Fund (FWF) through the doctoral school *Dissipation and dispersion in nonlinear PDEs* (grant W1245), the SFB *Taming complexity in partial differential systems* (grant SFB F65), the stand-alone project *Computational nonlinear PDEs* (grant P33216), and the Erwin Schrödinger fellowship (grant J4379).

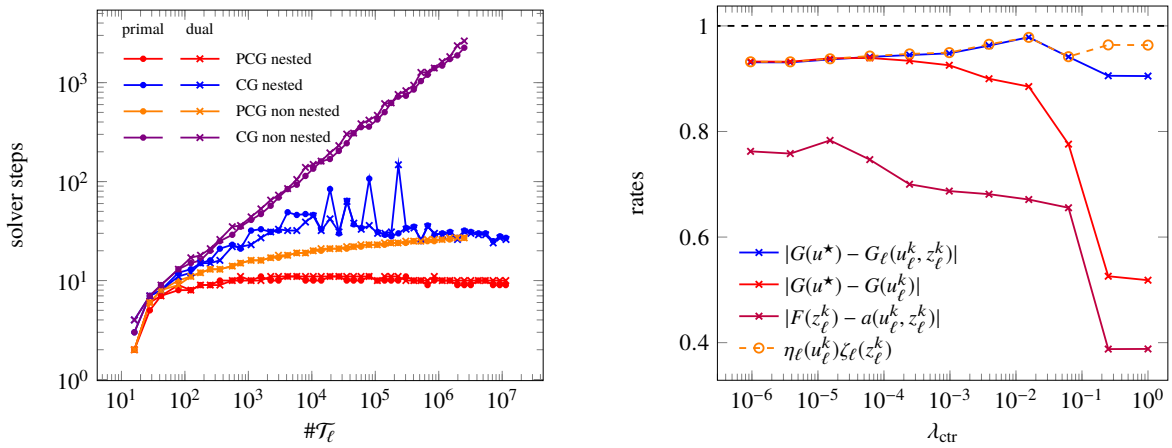


Fig. 3 Left: Number of steps of the iterative solver on each level ℓ of Algorithm 3.1. Right: Rates of various error measures for different values of λ_{ctr} in Algorithm 3.1 with PCG and nested iteration.

References

[1] Roland Becker, Elodie Estecahandy, David Trujillo: *Weighted marking for goal-oriented adaptive finite element methods*, SIAM J. Numer. Anal. 49 (2011), 2451–2469.

[2] Roland Becker, Gregor Gantner, Michael Innerberger, Dirk Praetorius: *Goal-oriented adaptive finite element methods with optimal computational complexity*, Preprint `arXiv:2101.11407`, 2021.

[3] Carsten Carstensen, Michael Feischl, Marcus Page, Dirk Praetorius, *Axioms of adaptivity*, Comput. Math. Appl. 67 (2014), 1195–1253.

[4] J. Manuel Cascon, Christian Kreuzer, Ricardo H. Nochetto, Kunibert G. Siebert: *Quasi-optimal convergence rate for an adaptive finite element method*, SIAM J. Numer. Anal. 46 (2008), 2524–2550.

[5] Long Chen, Ricardo H. Nochetto, Jinchao Xu: *Optimal multilevel methods for graded bisection grids*, Numer. Math. 120 (2012), 1–34.

[6] Willy Dörfler: *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal. 33 (1996), 1106–1124.

[7] Michael Feischl, Dirk Praetorius, Kristoffer G. van der Zee: *An abstract analysis of optimal goal-oriented adaptivity*, SIAM J. Numer. Anal. 54 (2016), 1423–1448.

[8] Michael B. Giles and Endre Süli: *Adjoint methods for PDEs: a posteriori error analysis and postprocessing by duality*, Acta Numer. 11 (2002), 145–236.

[9] Gene H. Golub and Charles F. Van Loan: *Matrix computations*, Johns Hopkins University Press, Baltimore, fourth edition, 2013.

[10] Mario S. Mommer, Rob Stevenson: *A goal-oriented adaptive finite element method with convergence rates*, SIAM J. Numer. Anal. 47 (2009), 861–886.

[11] Carl-Martin Pfeiler, Dirk Praetorius: *Dörfler marking with minimal cardinality is a linear complexity problem*, Math. Comp. 89 (2020), 2735–2752.

[12] Stefan Schimanko: *On optimality of adaptive FEM and BEM*, PhD thesis, TU Wien, Institute of Analysis and Scientific Computing, 2021.

[13] Rob Stevenson: *Optimality of a standard adaptive finite element method*, Found. Comput. Math. 7 (2007), 245–269.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 73–81)

On volume constraint problems related to the fractional Laplacian

J. C. Bellido[1] **and Alejandro Ortega**[2]
1. *Universidad de Castilla–La Mancha, Spain,* josecarlos.bellido@uclm.es
2. *Universidad Carlos III de Madrid, Spain,* alortega@math.uc3m.es

Abstract

In this work we study volume constraint problems involving the nonlocal operator $(-\Delta)^s_\delta$ depending upon a parameter $\delta > 0$ called horizon. We analyze the associated linear and spectral problems and the behavior of these volume constraint problems when $\delta \to 0^+$ and $\delta \to +\infty$. We prove spectral convergence to the classical Laplacian as $\delta \to 0^+$ under a suitable scaling and spectral convergence to the fractional Laplacian as $\delta \to +\infty$.

1. Introduction

We study volume constraint elliptic problems driven by a nonlocal operator closely related to the well-known fractional Laplace operator. In particular, given an open bounded domain $\Omega \subset \mathbb{R}^N$ with Lipschitz boundary and $\delta > 0$, a parameter called *horizon*, let us define the problem

$$\begin{cases} (-\Delta)^s_\delta u = f & \text{in} \quad \Omega, \\ u = 0 & \text{on} \quad \partial_\delta \Omega, \end{cases} \tag{P^s_δ}$$

where,

$$(-\Delta)^s_\delta u(x) = c_{N,s} P.V. \int_{B(x,\delta)} \frac{u(x) - u(y)}{|x-y|^{n+2s}} dy,$$

with $c_{N,s} = \frac{2^{2s} s \Gamma(\frac{N}{2}+s)}{\pi^{\frac{N}{2}} \Gamma(1-s)}$ a normalization constant and $\partial_\delta \Omega$ the nonlocal boundary given by

$$\partial_\delta \Omega = \{y \in \mathbb{R}^N \backslash \Omega : |x-y| < \delta \text{ for some } x \in \Omega\}.$$

Nonlocal and fractional elliptic problems have attracted a great attention in the mathematical community in the last decades, coming from fields as nonlocal diffusion [4, 13], statistical mechanics [2] and continuum mechanics, including peridynamics, [17, 20, 25]. Nonlocal variational problems are also important in the characterization of Sobolev spaces [9,18,21]. Interesting surveys on the fractional Laplacian and nonlocal elliptic problems are [19,22].

The operator $(-\Delta)^s_\delta$ is not new, and it has been addressed in different studies in the literature before. In view of the definition of $(-\Delta)^s_\delta$, it is clear that long-range interactions are neglected and only those exerted at distance smaller than $\delta > 0$ are taken into account, i.e., the horizon $\delta > 0$ represents the range of interactions. In this sense, the operator $(-\Delta)^s_\delta$, pertaining to the class of nonlocal elliptic operators, it is clearly inspired by peridynamics, where the elastic energy is computed through a double integral of a pairwise potential function, and it could actually be seen as a *peridynamic fractional Laplacian*. Peridynamics is a nonlocal continuum model for Solid Mechanics proposed by Silling, cf. [25]. The main difference with classical theory relies on the nonlocality, reflected in the fact that points separated by a positive distance exert a force upon each other. Since the use of gradients is avoided, peridynamics is a suitable framework for problems where discontinuities, such as fractures, appear naturally. In [16] a numerical study comparing $(-\Delta)^s_\delta$ with the fractional Laplacian, the spectral fractional Laplacian and the regional Laplacian is performed. In [1], the Fourier multiplier associated to $(-\Delta)^s_\delta$ is computed and, as a consequence, convergence of $(-\Delta)^s_\delta u(x)$ to $(-\Delta)u(x)$, for sufficiently smooth u, is obtained as $\delta \to 0^+$ or $s \to 1^-$. Also, $(-\Delta)^s_\delta$ was studied in [14] in connection with the fractional Laplacian, $(-\Delta)^s = (-\Delta)^s_\infty$, and with the motivation of computing numerical approximations. Notice that taking the limit as $\delta \to +\infty$ one recovers, at least formally, the usual nonlocal elliptic problem driven by the fractional Laplace operator with boundary condition on the complementary of the domain Ω.

In this work the limit properties of $(-\Delta)^s_\delta$, both as $\delta \to 0^+$ and as $\delta \to +\infty$, are addressed. In particular, by means of Γ-convergence techniques, we show, cf. [7], convergence of solutions and spectral stability, i.e., convergence of eigenvalues and eigenfunctions, to the classical Laplacian and to the fractional Laplacian as $\delta \to 0^+$ and as $\delta \to +\infty$ respectively. Therefore, the operator $(-\Delta)^s_\delta$ can be seen as an intermediate operator in between the local Laplacian and the fractional Laplacian.

The results for the case $\delta \to 0^+$ rely on a general Γ-convergence result from [6], while the results for the case $\delta \to +\infty$ are based on Γ-convergence properties of monotone sequences.

Closely related to our work is [3], where spectral stability as $\delta \to 0^+$ for certain nonlocal problems is shown without explicitly appealing to Γ-convergence. The advantage of the Γ-convergence approach is its adaptability to a nonlinear setting. Regarding this nonlinear setting, the spectral convergence of the fractional p-Laplacian to the classical p-Laplacian as $s \to 1^-$ is shown, by means of Γ-convergence techniques, in [11]. We extend the results of this work about spectral behavior to the nonlinear case dealing with the *peridynamic fractional p-Laplacian* in [8], where we obtain analogous results to those of [11] regarding the fractional p-Laplacian.

2. Preliminaries

Let $\Omega \subset \mathbb{R}^N$ be a regular bounded domain and consider the Sobolev space $H^s(\Omega) = \{v \in L^2(\Omega) : \|v\|_{H^s(\Omega)} < \infty\}$, where $\|v\|^2_{H^s(\Omega)} = \|v\|^2_{L^2(\Omega)} + |v|^2_{H^s(\Omega)}$ being $|\cdot|_{H^s(\Omega)}$ the Gagliardo semi-norm,

$$|v|^2_{H^s(\Omega)} = \int_\Omega \int_\Omega \frac{|v(x) - v(y)|^2}{|x-y|^{N+2s}} dydx.$$

Next, denoting by $\Omega^c = \mathbb{R}^N \backslash \Omega$, let us set the energy space $\mathcal{H}^s_0(\Omega) = \{v \in H^s(\mathbb{R}^N) : v = 0 \text{ on } \Omega^c\}$ endowed with the norm inherited from $H^s(\mathbb{R}^N)$. Let us note that, given $v \in \mathcal{H}^s_0(\Omega)$, although $v = 0$ on Ω^c, the norms $\|v\|_{H^s(\Omega)}$ and $\|v\|_{\mathcal{H}^s_0(\Omega)}$ are not the same. Indeed, denoting by $\mathcal{D} = \left(\mathbb{R}^N \times \mathbb{R}^N\right) \backslash \left(\Omega^c \times \Omega^c\right)$, we have the strict inclusion $\Omega \times \Omega \subsetneq \mathcal{D}$. Then, the norm $\|\cdot\|_{\mathcal{H}^s_0(\Omega)}$ takes into account the interaction between Ω and Ω^c, i.e.,

$$\|v\|^2_{\mathcal{H}^s_0(\Omega)} = \|v\|^2_{H^s(\mathbb{R}^N)} = \|v\|^2_{L^2(\Omega)} + \iint_{\mathcal{D}} \frac{|v(x) - v(y)|^2}{|x-y|^{N+2s}} dydx.$$

Therefore, the space $\mathcal{H}^s_0(\Omega)$ is the appropriate space to deal with homogeneous elliptic boundary value problems involving the fractional Laplace operator,

$$(-\Delta)^s_\infty u(x) = c_{N,s} P.V. \int_{\mathbb{R}^N} \frac{u(x) - u(y)}{|x-y|^{N+2s}} dy.$$

On the other hand, by the fractional Sobolev inequality, cf. [15, Th. 6.5], we can renormize the space $\mathcal{H}^s_0(\Omega)$ and consider it endowed with the norm

$$|||v|||^2_{\mathcal{H}^s_0} = \iint_{\mathcal{D}} \frac{|v(x) - v(y)|^2}{|x-y|^{N+2s}} dydx.$$

Next, given an horizon $\delta > 0$, let us define the (*nonlocally*) completed domain $\Omega_\delta = \Omega \cup \partial_\delta \Omega$, and the energy space $\mathbb{H}^s(\Omega_\delta) = \{v \in L^2(\Omega_\delta) : \|v\|_{\mathbb{H}^s(\Omega_\delta)} < \infty\}$ where $\|v\|^2_{\mathbb{H}^s(\Omega_\delta)} = \|v\|^2_{L^2(\Omega_\delta)} + |v|^2_{\mathbb{H}^s(\Omega_\delta)}$ with

$$|v|^2_{\mathbb{H}^s(\Omega_\delta)} = \int_{\Omega_\delta} \int_{\Omega_\delta \cap B(x,\delta)} \frac{|v(x) - v(y)|^2}{|x-y|^{N+2s}} dydx.$$

Note that, because of [5, Prop. 6.1], the spaces $\mathbb{H}^s(\Omega_\delta)$ and $H^s(\Omega_\delta)$ are isomorphic. In order to deal with the boundary value problem P^s_δ, we define the energy space $\mathbb{H}^{\delta,s}_0(\Omega) = \{v \in \mathbb{H}^s(\Omega_\delta) : v \equiv 0 \text{ on } \partial_\delta \Omega\}$ endowed with the norm inherited from $\mathbb{H}^s(\Omega_\delta)$. Let us notice that, given a function $v \in \mathbb{H}^{\delta,s}_0(\Omega)$, although we have $v = 0$ on $\partial_\delta \Omega = \Omega_\delta \backslash \Omega$, the norms $\|v\|_{\mathbb{H}^s(\Omega)}$ and $\|v\|_{\mathbb{H}^{\delta,s}_0(\Omega)}$ are not the same. Indeed, if $v = 0$ on $\partial_\delta \Omega$, since $\mathbb{H}^s(\Omega) = \{v \in L^2(\Omega) : \|v\|_{\mathbb{H}^s(\Omega)} < \infty\}$ with $\|v\|^2_{\mathbb{H}^s(\Omega)} = \|v\|^2_{L^2(\Omega)} + |v|^2_{H^s(\Omega)}$ and

$$\|v\|^2_{\mathbb{H}^{\delta,s}_0(\Omega)} = \|v\|^2_{L^2(\Omega)} + \iint_{\mathcal{D}_\delta} \frac{|v(x) - v(y)|^2}{|x-y|^{N+2s}} dydx,$$

with $\mathcal{D}_\delta = \Big(\Omega_\delta \times \big(\Omega_\delta \cap B(x,\delta)\big)\Big) \Big\backslash \Big(\partial_\delta \Omega \times \big(\partial_\delta \Omega \cap B(x,\delta)\big)\Big)$, we have the strict inclusion $\big(\Omega \times (\Omega \cap B(x,\delta))\big) \subsetneq \mathcal{D}_\delta$. Hence, the norm $\|\cdot\|_{\mathbb{H}^{\delta,s}_0(\Omega)}$ takes into account the interaction between Ω and $\partial_\delta \Omega$ in the sense that

$$|v|^2_{\mathbb{H}^s(\Omega_\delta)} = |v|^2_{H^s(\Omega)} + \int_{\partial_\delta \Omega} \int_{\Omega \cap B(x,\delta)} \frac{|v(y)|^2}{|x-y|^{N+2s}} dydx + \int_\Omega \int_{\partial_\delta \Omega \cap B(x,\delta)} \frac{|v(x)|^2}{|x-y|^{N+2s}} dydx.$$

Therefore, the space $\mathbb{H}^{\delta,s}_0(\Omega)$ is the appropriate space to deal with homogeneous elliptic boundary value problems involving the operator $(-\Delta)^s_\delta$. Moreover, comparing the norms $\|\cdot\|_{\mathcal{H}^s_0(\Omega)}$ and $\|\cdot\|_{\mathbb{H}^{\delta,s}_0(\Omega)}$ we observe that $\partial_\delta \Omega$

plays the role of Ω^c. Indeed, the sets Ω_δ and $\Omega_\delta \cap B(x,\delta)$ will lead to the complete space $\mathbb{R}^N$ for $\delta \to +\infty$, the set $\Omega \cap B(x,\delta)$ will eventually reach the set Ω for $\delta > 0$ big enough and the sets $\partial_\delta\Omega$ and $\partial_\delta\Omega \cap B(x,\delta)$ will reach Ω^c for $\delta \to +\infty$. In fact, $\mathcal{D}_{\delta_1} \subset \mathcal{D}_{\delta_2}$ for $\delta_1 < \delta_2$ and $\mathcal{D}_\delta \to \mathcal{D}$ as $\delta \to +\infty$.
Due to [5, Prop. 6.1] and [5, Lem. 6.2], we have $|v|_{\mathbb{H}^s(\Omega_\delta)} \le \|v\|_{\mathbb{H}^s(\Omega_\delta)} \le c|v|_{\mathbb{H}^s(\Omega_\delta)}$, for a positive constant $c \in \mathbb{R}$ and, then, we can renormize the space $\mathbb{H}_0^{\delta,s}(\Omega)$ and consider it endowed with the norm

$$|||v|||^2_{\mathbb{H}_0^{\delta,s}} = \int_{\Omega_\delta}\int_{\Omega_\delta\cap B(x,\delta)} \frac{|v(x)-v(y)|^2}{|x-y|^{N+2s}}dydx.$$

As a consequence, we have the following, cf. [7, Lem. 2] and [7, Lem. 4] respectively.

Lemma 2.1 *The space $\mathbb{H}_0^{\delta,s}(\Omega)$ is a Hilbert space endowed with norm $|||\cdot|||_{\mathbb{H}_0^{\delta,s}}$ induced by the scalar product*

$$\langle u,v\rangle_{\mathbb{H}_0^{\delta,s}} = \int_{\Omega_\delta}\int_{\Omega_\delta\cap B(x,\delta)} \frac{(u(x)-u(y))(v(x)-v(y))}{|x-y|^{N+2s}}dydx.$$

Analyze convergence phenomena for $\delta \to +\infty$ will require to study the relation between $\mathcal{H}_0^s(\Omega)$ and $\mathbb{H}_0^{\delta,s}(\Omega)$.

Lemma 2.2 *For any $\delta > 0$, the spaces $\mathbb{H}_0^{\delta,s}(\Omega)$ and $\mathcal{H}_0^s(\Omega)$ are isomorphic. In particular, there exists a constant $C = C(\delta) > 1$ such that $C(\delta) \to 1$ as $\delta \to +\infty$ and*

$$|||\cdot|||^2_{\mathbb{H}_0^{\delta,s}} \le |||\cdot|||^2_{\mathcal{H}_0^s} \le C(\delta)|||\cdot|||^2_{\mathbb{H}_0^{\delta,s}} \quad \text{for all } \delta > 0.$$

Now we make precise the definition of weak solution of problem P_δ^s.

Definition 2.3 We say that $u \in \mathbb{H}_0^{\delta,s}(\Omega)$ is a weak solution to problem P_δ^s if, for all $v \in \mathbb{H}_0^{\delta,s}(\Omega)$,

$$\frac{c_{N,s}}{2}\langle u,v\rangle_{\mathbb{H}_0^{\delta,s}} = \langle f,v\rangle_{L^2(\Omega)}.$$

3. Main Results

We present now the main results of the work. To that end, let us set $\kappa(N,s) = \frac{4N(1-s)}{\sigma_{N-1}c_{N,s}}$ with σ_{N-1} the surface of the unitary sphere $\mathbb{S}^{N-1}$ and $\partial_\infty\Omega = \mathbb{R}^N\setminus\Omega$ and consider the following problems,

$$RP_\delta^s \equiv \begin{cases} (-\Delta)_\delta^s u = \frac{\delta^{2(1-s)}}{\kappa(N,s)}f & \text{in } \Omega,\\ u = 0 & \text{on } \partial_\delta\Omega,\end{cases} \qquad P_0^1 \equiv \begin{cases} (-\Delta)u = f & \text{in } \Omega,\\ u = 0 & \text{on } \partial\Omega,\end{cases} \quad \text{and} \quad P_\infty^s \equiv \begin{cases} (-\Delta)_\infty^s u = f & \text{in } \Omega,\\ u = 0 & \text{on } \partial_\infty\Omega,\end{cases}$$

Our main results regarding the linear problems are the following, cf. [7, Th. 2] and [7, Th. 3].

Theorem 3.1 *Let $u^{\delta,s}$ and $u^{0,1}$ be the solutions of RP_δ^s and P_0^1 respectively. Then, up to a subsequence,*

$$u^{\delta,s} \to u^{0,1} \text{ in } L^2(\Omega) \quad \text{as } \delta \to 0^+.$$

Theorem 3.2 *Let $u^{\delta,s}$ and $u^{\infty,s}$ be the solutions of P_δ^s and P_∞^s respectively. Then, up to a subsequence,*

$$u^{\delta,s} \to u^{\infty,s} \text{ in } L^2(\Omega) \quad \text{as } \delta \to +\infty.$$

We continue with existence and stability issues for the eigenvalue problem

$$\begin{cases} (-\Delta)_\delta^s\varphi = \lambda\varphi & \text{in} \quad \Omega,\\ \varphi = 0 & \text{on} \quad \partial_\delta\Omega.\end{cases} \tag{EP_δ^s}$$

Using [5, Prop. 6.1, Lem. 6.2] and following [24, Prop. 9], we prove the following, cf. [7, Prop. 2].

Proposition 3.3 *Let $\delta > 0$, $s \in (0,1)$, $N > 2s$ and $\Omega \subset \mathbb{R}^N$ an open bounded set with Lipschitz boundary. Then, the following hold*:

1. *Problem EP^s_δ has a first positive eigenvalue that can be characterized as*

$$\lambda_1^{\delta,s} = \min_{\substack{u\in\mathbb{H}_0^{\delta,s}(\Omega)\\ \|u\|_{L^2(\Omega)}=1}} \frac{c_{N,s}}{2}|||u|||^2_{\mathbb{H}_0^{\delta,s}} = \frac{c_{N,s}}{2}\left\|\varphi_1^{\delta,s}\right\|^2_{\mathbb{H}_0^{\delta,s}},$$

where $\varphi_1^{\delta,s}\in\mathbb{H}_0^{\delta,s}(\Omega)$, is a nonnegative eigenfunction. In addition, the first eigenvalue $\lambda_1^{\delta,s}$ is simple.

2. *The eigenvalues of EP^s_δ are a countable set $\{\lambda_k^{\delta,s}\}_{k\in\mathbb{N}}$ satisfying*

$$0<\lambda_1^{\delta,s}<\lambda_2^{\delta,s}\le\ldots\le\lambda_k^{\delta,s}\le\ldots \quad \text{and} \quad \lambda_k^{\delta,s}\to+\infty \qquad \text{as } k\to+\infty.$$

Furthermore, for any $k\in\mathbb{N},\ k\ge 2$ the eigenvalues can be characterized as

$$\lambda_k^{\delta,s} = \min_{\substack{u\in\mathbb{P}_k^\delta\\ \|u\|_{L^2(\Omega)}=1}} \frac{c_{N,s}}{2}|||u|||^2_{\mathbb{H}_0^{\delta,s}} = \frac{c_{N,s}}{2}\left\|\varphi_k^{\delta,s}\right\|^2_{\mathbb{H}_0^{\delta,s}}.$$

where $\mathbb{P}_k^\delta=\{u\in\mathbb{H}_0^{\delta,s}(\Omega):\langle u,\varphi_j^{\delta,s}\rangle_{\mathbb{H}_0^{\delta,s}}=0,\ j=1,\ldots,k-1\}$ and an eigenfunction $\varphi_k^{\delta,s}\in\mathbb{P}_k^\delta$.

3. *The set of eigenfunctions $\{\varphi_k^{\delta,s}\}_{k\in\mathbb{N}}$ is an orthogonal basis of $\mathbb{H}_0^{\delta,s}(\Omega)$ and an orthonormal basis of $L^2(\Omega)$.*

4. *For any $k\in\mathbb{N}$, the eigenvalue $\lambda_k^{\delta,s}$ has finite multiplicity, $1\le m_k^{\delta,s}<\infty$ for all $k\in\mathbb{N}$.*

Moreover, arguing as in [23, Prop. 4], we also deduce the following, cf. [7, Lem. 5].

Lemma 3.4 *Let $\varphi_k^{\delta,s}\in\mathbb{H}_0^{\delta,s}(\Omega)$ be an eigenfunction of EP^s_δ, then $\varphi_k^{\delta,s}\in L^\infty(\Omega)$ for any $k\in\mathbb{N}$.*

Finally, we present the main results about the behavior of EP^s_δ when $\delta\to 0^+$ and $\delta\to+\infty$. To that end, let us consider the eigenvalue problems,

$$EP_0^1\equiv\begin{cases}(-\Delta)\varphi=\lambda\varphi & \text{in}\quad\Omega,\\ \varphi=0 & \text{on}\quad\partial\Omega,\end{cases}\qquad\text{and}\qquad EP_\infty^s\equiv\begin{cases}(-\Delta)^s_\infty\varphi=\lambda\varphi & \text{in}\quad\Omega,\\ \varphi=0 & \text{on}\quad\mathbb{R}^N\backslash\Omega.\end{cases}$$

It is well known cf. [12] that the problem EP_0^1 has a countable set of eigenvalues that we denote by $\{\lambda_k^{0,1}\}_{k\in\mathbb{N}}$ and such that

$$0<\lambda_1^{0,1}<\lambda_2^{0,1}\le\ldots\le\lambda_k^{0,1}\le\ldots \quad\text{and}\quad \lambda_k^{0,1}\to+\infty \text{ as } k\to+\infty.$$

Denoting by $m_k^{0,1}$ the multiplicity of the eigenvalue $\lambda_k^{0,1}$, we have $1\le m_k^{0,1}<\infty$ for all $k\in\mathbb{N}$. Moreover, there exists a countable set of eigenfunctions $\{\varphi_k^{0,1}\}_{k\in\mathbb{N}}$ that is an orthogonal basis of $H_0^1(\Omega)$ and an orthonormal basis of $L^2(\Omega)$. The first eigenvalue is simple and $\varphi_1^{0,1}>0$ in Ω.

Concerning the fractional eigenvalue problem, Servadei and Valdinoci proved, cf. [24], that EP^s_∞ has a countable set of eigenvalues that we denote by $\{\lambda_k^{\infty,s}\}_{k\in\mathbb{N}}$ and such that

$$0<\lambda_1^{\infty,s}<\lambda_2^{\infty,s}\le\ldots\le\lambda_k^{\infty,s}\le\ldots, \quad\text{and}\quad \lambda_k^{\infty,s}\to+\infty \text{ as } k\to+\infty.$$

Denoting by $m_k^{\infty,s}$ the multiplicity of the eigenvalue $\lambda_k^{\infty,s}$, we have $1\le m_k^{\infty,s}<\infty$ for all $k\in\mathbb{N}$. Moreover, there exists a countable set of eigenfunctions $\{\varphi_k^{\infty,s}\}_{k\in\mathbb{N}}$ that is an orthogonal basis of

$$\mathbb{H}_0^{\infty,s}(\Omega)=\left\{v\in L^2(\Omega):\iint_{\mathcal{D}}\frac{|v(x)-v(y)|^2}{|x-y|^{N+2s}}dydx<\infty,\ v=0 \text{ a.e. on } \mathbb{R}^N\backslash\Omega\right\},$$

and an orthonormal basis of $L^2(\Omega)$. The first eigenvalue is also simple and $\varphi_1^{\infty,s}\ge 0$ in Ω.

We relate the eigenvalues and eigenfunctions of EP^s_δ to those of the eigenvalue problems EP_0^1 and EP^s_∞ through the following results, cf. [7, Th. 4] and [7, Th. 5].

Theorem 3.5 *Let $\{(\lambda_k^{\delta,s}, \varphi_k^{\delta,s})\}_{k\in\mathbb{N}}$ be the set of eigenvalues and eigenfunctions of $(-\Delta)^s_\delta$ with homogeneous Dirichlet boundary condition on $\partial_\delta\Omega$ and let $\{(\lambda_k^{0,1}, \varphi_k^{0,1})\}_{k\in\mathbb{N}}$ be the set of eigenvalues of $(-\Delta)$ with homogeneous Dirichlet boundary condition on $\partial\Omega$. Then,*

$$\kappa(N,s)\frac{\lambda_k^{\delta,s}}{\delta^{2(1-s)}} \to \lambda_k^{0,1} \quad \text{as } \delta \to 0^+,$$

for $\kappa(N,s) = \frac{4N(1-s)}{\sigma_{N-1}c_{N,s}}$. Moreover, there exists a subsequence (that we do not relabel) such that, for every $k \in \mathbb{N}$,

$$\varphi_k^{\delta,s} \to \varphi_k^{0,1} \text{ in } L^2(\Omega) \quad \text{as } \delta \to 0^+.$$

As a consequence, $m_k^{\delta,s} \to m_k^{0,1}$ as $\delta \to 0^+$, for any $k \geq 1$.

As Theorems 3.1 and 3.5 show, even though the fractionality parameter s keeps fixed, the local problem driven by $(-\Delta)$ is recovered, under the appropriate rescaling, as $\delta \to 0^+$.

Theorem 3.6 *Let $\{(\lambda_k^{\delta,s}, \varphi_k^{\delta,s})\}_{k\in\mathbb{N}}$ be the set of eigenvalues and eigenfunctions of $(-\Delta)^s_\delta$ with homogeneous Dirichlet boundary condition on $\partial_\delta\Omega$ and let $\{(\lambda_k^{\infty,s}, \varphi_k^{\infty,s})\}_{k\in\mathbb{N}}$ be the set of eigenvalues of $(-\Delta)^s_\infty$ with homogeneous Dirichlet boundary condition on $\mathbb{R}^N\setminus\Omega$. Then,*

$$\lambda_k^{\delta,s} \to \lambda_k^{\infty,s} \quad \text{as } \delta \to +\infty,$$

and there exists a subsequence (that we do not relabeled) such that for every $k \in \mathbb{N}$,

$$\varphi_k^{\delta,s} \to \varphi_k^{\infty,s} \text{ in } L^2(\Omega) \quad \text{as } \delta \to +\infty.$$

As a consequence, $m_k^{\delta,s} \to m_k^{\infty,s}$ as $\delta \to \infty$, for any $k \geq 1$.

3.1. Γ-convergence

This section includes some results about Γ-convergence that play a crucial role in the proof of the above stated results. The limit process in the sense of Γ-convergence, denoted by $\xrightarrow{\Gamma}$, is the right concept of limit for variational problems since, together with equicoercivity or compactness, it implies that minimizers of I_δ converge to minimizers of I as well as their energies. A nice account on Γ-convergence is provided in [10]. We present now a result about Γ-convergence of functionals proved in [6] that is the core of the proof of Theorem 3.1 and Theorem 3.5. Let us consider a functional of the form

$$I(u) = \int_\Omega \int_{\Omega\cap B(x,\delta)} \omega(x-y, u(x)-u(y))dydx,$$

for a *potential function* $\omega(x,y) : \mathbb{R}^N \times \mathbb{R} \mapsto \mathbb{R}$ verifying that, for some $\beta \in \mathbb{R}$, the following limit exists,

$$\omega^\circ(x,y) = \lim_{t\to 0^+} \frac{1}{t^\beta}\omega(tx,ty).$$

Let $\overline{\omega}^c : \mathbb{R}^N \mapsto \mathbb{R}$ be the *limit density convexification* of $\overline{\omega}$ defined as $\overline{\omega}^c = \sup\{v : \ v \leq \overline{\omega} \text{ and } v \text{ convex}\}$, where $\overline{\omega} : \mathbb{R}^N \mapsto \mathbb{R}$ is the *limit density* of ω,

$$\overline{\omega}(F) = \int_{\mathbb{S}^{N-1}} \omega^\circ(z, Fz)d\sigma(z).$$

Under the hypotheses stated below, given the sequence of rescaled functionals

$$I_\delta(u) = \frac{N+\beta}{\delta^{N+\beta}} \int_\Omega \int_{\Omega\cap B(x,\delta)} \omega(x-y, u(x)-u(y))dydx,$$

we have,

$$I_\delta(u) \xrightarrow{\Gamma} I_0(u) = \int_\Omega \overline{\omega}^c(\nabla u)dx.$$

In particular, the above Γ-convergence is ensured by the next result, that also provides the compactness of uniformly bounded energy sequences. Let us set $\tilde{\Omega} = \{z = x - y \ : \ x, \ y \in \Omega\}$ and $\mathcal{A}_\delta = \{v \in L^p(\Omega) : v = 0 \text{ on } \partial_\delta\Omega\}$.

Theorem 3.7 ([6, *Th.* 1]) *Let $\Omega \subset \mathbb{R}^N$ be a bounded domain with Lipschitz boundary and $\omega : \tilde{\Omega} \times \mathbb{R} \mapsto \mathbb{R}$ satisfying the hypotheses (H1)-(H2) below. Then, the following holds*:

a) *Compactness: For each $\delta > 0$, let $u_\delta \in \mathcal{A}_\delta$ such that $\sup_\delta I_\delta(u_\delta) < +\infty$. Then, there exist $u \in W_0^{1,p}(\Omega)$ such that, for a subsequence, $u_\delta \to u$ strong in $L^p(\Omega)$ as $\delta \to 0^+$.*

b) *Γ-liminf inequality: For each $\delta > 0$ let $u_\delta \in \mathcal{A}_\delta$ and $u \in W_0^{1,p}(\Omega)$ such that $u_\delta \to u$ strong in $L^p(\Omega)$ as $\delta \to 0^+$. Then, $I_0(u) \leq \liminf_{\delta \to 0^+} I_\delta(u_\delta)$.*

c) *Γ-limsup inequality: For each $\delta > 0$ and $u \in W_0^{1,p}(\Omega)$ there exist $u_\delta \in \mathcal{A}_\delta$, called recovery sequence, such that $u_\delta \to u$ strong in $L^p(\Omega)$ as $\delta \to 0^+$ and $\limsup_{\delta \to 0^+} I_\delta(u_\delta) \leq I_0(u)$.*

For a general potential function $\omega(x, y)$ the hypotheses of Theorem 3.7 are quite involved but, if $\omega(x, y) = f(x)g(y)$ for f a Lebesgue measurable function and g a Borel measurable and convex function, the necessary hypotheses are:

H1) There exists constants $c_0, c_1 > 0$ and $h \in L^1(\mathbb{S}^{N-1})$ with $h \geq 0$ such that, for some $1 < p < +\infty$ and $0 \leq \alpha < N + p$,

$$c_0 \frac{|y|^p}{|x|^\alpha} \leq f(x)g(y) \leq c_1 h\left(\frac{x}{|x|}\right) \frac{|y|^p}{|x|^\alpha} \quad \text{for } x \in \widetilde{\Omega},\ y \in \mathbb{R}.$$

H2) The functions $f^\circ(x) = \lim_{t \to 0^+} t^\alpha f(tx)$ and $g^\circ(y) = \lim_{t \to 0^+} \frac{1}{t^p} g(ty)$, are continuous and, for each compact $K \subset \mathbb{R}$,

$$\lim_{t \to 0^+} \sup_{x \in \mathbb{S}^{N-1}} |t^\alpha f(tx) - f^\circ(x)| = 0 \quad \text{and} \quad \lim_{t \to 0^+} \sup_{K \subset \mathbb{R}} |\frac{1}{t^p} g(ty) - g^\circ(y)| = 0.$$

The following is a straightforward consequence, cf. [10], of the Γ-convergence and the compactness provided by Theorem 3.7. Notice that under previous hypothesis existence of minimizers for I_δ is guaranteed, cf. [5].

Corollary 3.8 *In the conditions of Theorem* 3.7*, let $u_\delta \in \mathbb{H}_0^{\delta,s}(\Omega)$ be a minimizer of I_δ, for any $\delta > 0$. Then, there exists $u_0 \in H_0^1(\Omega)$ a minimizer of I_0 such that, up to a subsequence,*

$$u_\delta \to u_0 \text{ strong in } L^2(\Omega) \quad \text{as } \delta \to 0^+ \quad \text{and} \quad I_\delta(u_\delta) \to I_0(u_0) \quad \text{as } \delta \to 0^+.$$

3.2. Taking the horizon $\delta \to 0^+$

One of the main steps to prove Theorem 3.1 and Theorem 3.5 is the following result, cf. [7, Lem. 6], concerning the Γ-convergence of the energy functional defining the eigenvalues. Among other things, it shows that, up to the appropriate scaling, all the functionals $I_{\delta,s}$ will Γ-converge to the same Γ-limit independently of s.

Lemma 3.9 *Let us consider the scaled functional*

$$I_{\delta,s}(u) = \frac{2(1-s)}{\delta^{2(1-s)}} \int_{\Omega_\delta} \int_{\Omega_\delta \cap B(x,\delta)} \frac{|u(x) - u(y)|^2}{|x - y|^{N+2s}} dydx,$$

defined on $\mathbb{H}_0^{\delta,s}(\Omega)$. Then, the Γ-limit of $I_{\delta,s}(u)$ as $\delta \to 0^+$ is given by

$$I_0(u) = \frac{\sigma_{N-1}}{N} \int_\Omega |\nabla u(x)|^2 dx.$$

The proof of Theorem 3.1 follows from Lemma 3.9 and Corollary 3.8. To prove Theorem 3.5, we use Lemma 3.9 and Corollary 3.8, from where we get the convergence of the first eigenvalue under the appropriate scaling, namely,

$$\kappa(N, s) \frac{\lambda_1^{\delta,s}}{\delta^{2(1-s)}} \to \lambda_1^{0,1} \quad \text{as } \delta \to 0^+.$$

Moreover, we also get $\varphi_1^{\delta,s} \to \varphi_1^{0,1}$ strong in $L^2(\Omega)$ as $\delta \to 0^+$. Next, we construct a recovery sequence by projecting appropriately on the second eigenspace $\mathbb{P}_2^\delta = \{u \in \mathbb{H}_0^{\delta,s}(\Omega) : \langle u, \varphi_1^{\delta,s} \rangle_{\mathbb{H}_0^{\delta,s}} = 0\}$. Thanks to the strong $L^2(\Omega)$-convergence of the first eigenfunction the convergence of the second eigenvalue and the second eigenfunction follows. To conclude we argue inductively.

In order to clarify why the scaling in Γ-convergence result is natural in our context, it is of interest to deduce from the classical *localization* result of Bourgain, Brezis and Mironescu the upper bound

$$\lim_{\delta\to 0^+} \frac{\kappa(N,s)}{\delta^{2(1-s)}}\lambda_1^{\delta,s} \leq \lambda_1^{0,1}.$$

Let $\{\rho_n(x)\}_{n\in\mathbb{N}}$ be a sequence of radial mollifiers, i.e.,

$$\rho_n(x) = \rho_n(|x|),\ \rho_n(x) \geq 0 \text{ and } \int \rho_n(x)dx = 1 \quad \text{and satisfying} \quad \lim_{n\to\infty}\int_\varepsilon^\infty \rho_n(r)r^{N-1} = 0\ \forall \varepsilon > 0.$$

Theorem 3.10 ([9, *Th.* 2]) *Assume $u \in L^p(\Omega)$, $1 < p < \infty$. Then, for a constant $C = C(N,p) > 0$, we have*

$$\lim_{n\to\infty}\int_\Omega\int_\Omega \frac{|u(x)-u(y)|^p}{|x-y|^p}\rho_n(x-y)dydx = C\int_\Omega |\nabla u|^p dx.$$

with the convention that $\int_\Omega |\nabla u|^p dx = \infty$ if $u \notin W^{1,p}(\Omega)$.

Since $H_0^1(\Omega) \subset \mathbb{H}_0^{\delta,s}(\Omega)$ for all $\delta > 0$, we have

$$\lambda_1^{\delta,s} = \min_{\substack{u\in\mathbb{H}_0^{\delta,s}(\Omega)\\ \|u\|_{L^2(\Omega)}=1}} \frac{c_{N,s}}{2}\int_{\Omega_\delta}\int_{\Omega_\delta\cap B(x,\delta)} \frac{|u(x)-u(y)|^2}{|x-y|^{N+2s}}dydx \leq \frac{c_{N,s}}{2}\int_{\Omega_\delta}\int_{\Omega_\delta\cap B(x,\delta)} \frac{|\psi_1^\delta(x)-\psi_1^\delta(y)|^2}{|x-y|^{N+2s}}dydx,$$

being ψ_1^δ the first eigenfunction $\varphi_1^{0,1}$ of the Laplace operator ($L^2(\Omega)$-normalized) extended by zero on $\partial_\delta\Omega$. In order to apply Theorem 3.10, let us rewrite the above inequality as

$$\lambda_1^{\delta,s} \leq \int_{\Omega_\delta}\int_{\Omega_\delta} \frac{|\psi_1^\delta(x)-\psi_1^\delta(y)|^2}{|x-y|^2}\rho_\delta(|x-y|)dydx,$$

with $\rho_\delta(z) = \dfrac{c_{N,s}}{2}\dfrac{\chi_{B(0,\delta)}(|z|)}{|z|^{N+2(s-1)}}$ and χ_A the characteristic function of the set A. Since

$$\int \rho_\delta(z)dz = \frac{\sigma_{N-1}c_{N,s}}{4(1-s)}\delta^{2(1-s)},$$

the sequence of radial mollifiers $\overline{\rho}_\delta(z) = \dfrac{4(1-s)}{\sigma_{N-1}}\dfrac{1}{\delta^{2(1-s)}}\dfrac{\chi_{B(0,\delta)}(|z|)}{|z|^{N+2(s-1)}}$ satisfy the hypotheses of Theorem 3.10. Then, because of Theorem 3.10, we conclude

$$\lim_{\delta\to 0^+}\frac{4(1-s)}{\sigma_{N-1}c_{N,s}}\frac{\lambda_1^{\delta,s}}{\delta^{2(1-s)}} \leq \lim_{\delta\to 0^+}\int_{\Omega_\delta}\int_{\Omega_\delta}\frac{|\psi_1^\delta(x)-\psi_1^\delta(y)|^2}{|x-y|^2}\overline{\rho}_\delta(|x-y|)dydx = C\int_\Omega |\nabla\varphi_1^{0,1}|^2dx,$$

since $\psi_1 = 0$ on $\partial_{\delta_0}\Omega$ and $\psi_1 = \varphi_1^{0,1}$ in Ω. Since for $p = 2$ the constant $C = C(N,p)$ appearing in Theorem 3.10 takes the value $C(N,2) = \frac{1}{N}$, taking in mind that $\|\varphi_1^{0,1}\|_{L^2(\Omega)} = 1$, the desired bound follows.

3.3. Taking the horizon $\delta \to +\infty$

Because of the definition of the operator $(-\Delta)_\delta^s$, as a restriction of the fractional Laplacian, it is plausible that if we take $\delta \to +\infty$ one recovers the definition of the standard fractional Laplacian, namely,

$$\lim_{\delta\to+\infty}(-\Delta)_\delta^s u(x) = c_{N,s}P.V.\int_{\mathbb{R}^N}\frac{u(x)-u(y)}{|x-y|^{N+2s}}dy.$$

A result in this line was given in [14, Th. 3.1], where it is showed the explicit convergence rate

$$\|u^{\delta,s} - u^{\infty,s}\|_{\mathbb{H}_0^{\delta,s}} \leq \frac{c}{(\delta-I)^{2s}}\|u^{\infty,s}\|_{L^2(\Omega)},$$

being $u^{\delta,s}$ and $u^{\infty,s}$ the solutions of P_δ^s and P_∞^s respectively, $c > 0$ is a constant independent of δ and $I = I(\Omega)$ a constant depending on the diameter of Ω. This is an important result from the point of view of the numerical

approximation of problems involving the fractional Laplacian but its proof strongly relies on the linearity of the problem P^s_δ. Instead, the proof of Theorem 3.2 and Theorem 3.6 are based on a general result about Γ-convergence that works for both the linear and nonlinear setting. We exploit this advantage to address the p -fractional Laplacian case, cf. [8], and extend the results of this work to the nonlinear setting.

The following Γ-convergence result, cf. [7, Lem. 7], is in the core of the proofs of Theorems 3.2 and Theorem 3.6. This result is analogous to Lemma 3.9 in relation to the proofs of Theorem 3.1 and Theorem 3.5.

Lemma 3.11 *Let us consider the functional*

$$\mathcal{E}_{\delta,s}(u) = \int_{\Omega_\delta} \int_{\Omega_\delta \cap B(x,\delta)} \frac{|u(x)-u(y)|^2}{|x-y|^{N+2s}} dydx.$$

defined on $\mathbb{H}_0^{\delta,s}(\Omega)$. *Then, the* Γ*-limit of* $\mathcal{E}_{\delta,s}(u)$ *is given by*

$$\mathcal{E}_{\infty,s}(u) = \int_{\mathbb{R}^N} \int_{\mathbb{R}^N} \frac{|u(x)-u(y)|^2}{|x-y|^{N+2s}} dydx \quad \text{as } \delta \to +\infty.$$

The above Lemma is an easy consequence of the monotonicity in $\delta > 0$ of the sequence of functionals $\mathcal{E}_{\delta,s}(u)$ and Γ-convergence properties. Indeed, since the sequence of functionals $\mathcal{E}_{\delta,s}(u)$ with $\delta \to +\infty$ is a monotone increasing sequence and functionals $\mathcal{E}_{\delta,s}$ are lower semicontinuous, cf. [5], because of [10, Remark 1.40], we conclude $\mathcal{E}_{\delta,s}(u) \overset{\Gamma}{\to} \mathcal{E}_{\infty,s}(u)$ as $\delta \to +\infty$.

The proof of Theorem 3.2 follows from Lemma 3.11 combined with the monotonicity in $\delta > 0$ together with Lemma 2.2, the compact embedding of $\mathcal{H}_0^s(\Omega)$ into $L^2(\Omega)$, cf. [15, Cor. 7.2], and the fact that Γ-convergence implies the convergence of the minimizers.

The proof of Theorem 3.6 follows by combining Lemma 3.11 and Lemma 2.2 with similar arguments to those used in Theorem 3.5.

Acknowledgements

This work was carried out while the second author was a postdoctoral researcher at Universidad de Castilla–La Mancha in Ciudad Real, funded by project MTM2017-83740-P of the *Agencia Estatal de Investigación, Ministerio de Ciencia e Innovación* (Spain).

References

[1] B. Alali and N. Albin, *Fourier multipliers for nonlocal Laplace operators*, Applicable Analysis, **0** (2019), pp. 1–21.

[2] G. Alberti and G. Bellettini, *A non-local anisotropic model for phase transitions: asymptotic behaviour of rescaled energies*, European J. Appl. Math., **9**, pp. 261–284, 1998.

[3] F. Andrés and J. Muñoz, *A type of nonlocal elliptic problem: existence and approximation through a Galerkin-Fourier method*, SIAM J. Math. Anal., **47**, pp. 498–525, 2015.

[4] F. Andreu-Vaillo, J. M. Mazón, J. D. Rossi and J. J. Toledo-Melero, *Nonlocal diffusion problems*, vol. **165** of Mathematical Surveys and Monographs, American Mathematical Society, Providence, RI; Real Sociedad Matemática Española, Madrid, 2010.

[5] J. C. Bellido and C. Mora-Corral, *Existence for nonlocal variational problems in peridynamics*, SIAM J. Math. Anal., **46**, pp. 890–916, 2014.

[6] J. C. Bellido, C. Mora-Corral and P. Pedregal, *Hyperelasticity as a* Γ*-limit of peridynamics when the horizon goes to zero*, Calc. Var. Partial Differential Equations, **54**, pp. 1643–1670, 2015.

[7] J.C. Bellido and A. Ortega, *A restricted nonlocal operator bridging together the Laplacian and the fractional Laplacian*, Calc. Var. **60**, 71, 2021.

[8] J.C. Bellido and A. Ortega, *Spectral Stability for the Peridynamic Fractional p-Laplacian*, Appl. Math. Optim., 2021.

[9] J. Bourgain, H. Brezis and P. Mironescu, *Another look at Sobolev spaces*, in Optimal control and partial differential equations, IOS, Amsterdam, pp. 439–455, 2001.

[10] A. Braides, Γ*-convergence for beginners*, vol. **22** of Oxford Lecture Series in Mathematics and its Applications, Oxford University Press, Oxford, 2002.

[11] L. Brasco, E. Parini and M. Squassina, *Stability of variational eigenvalues for the fractional* p*-Laplacian*, Discrete Contin. Dyn. Syst., **36**, pp. 1813–1845, 2016.

[12] H. Brezis, *Functional analysis, Sobolev spaces and partial differential equations*, Universitext, Springer, New York, 2011.

[13] C. Bucur and E. Valdinoci, *Nonlocal diffusion and applications*, vol. **20** of Lecture Notes of the Unione Matematica Italiana, Springer, [Cham]; Unione Matematica Italiana, Bologna, 2016.

[14] M. D'Elia and M. Gunzburger, *The fractional Laplacian operator on bounded domains as a special case of the nonlocal diffusion operator*, Comput. Math. Appl., **66**, pp. 1245–1260, 2013.

[15] E. Di Nezza, G. Palatucci and E. Valdinoci, *Hitchhiker's guide to the fractional Sobolev spaces*, Bull. Sci. Math., **136**, pp. 521–573, 2012.

[16] S. Duo, H. Wang and Y. Zhang, *A comparative study on nonlocal diffusion operators related to the fractional Laplacian*, Discrete Contin. Dyn. Syst. Ser. B, **24**, pp. 231–256, 2019.

[17] A. Evgrafov and J. C. Bellido, *From non-local Eringen's model to fractional elasticity*, Math. Mech. Solids, **24**, pp. 1935–1953, 2019.

[18] G. Leoni and D. Spector, *Characterization of Sobolev and BV spaces*, J. Funct. Anal., **261**, pp. 2926–2958, 2011.

[19] A. Lischke, G. Pang, M. Gulian and et al., *What is the fractional Laplacian? A comparative review with new results*, J. Comput. Phys., **404**, 109009, 62 pp., 2020.

[20] T. Mengesha and Q. Du, *Characterization of function spaces of vector fields and an application in nonlinear peridynamics*, Nonlinear Anal., **140**, pp. 82–111, 2016.

[21] A. C. Ponce, *A new approach to Sobolev spaces and connections to* Γ*-convergence*, Calc. Var. Partial Differential Equations, **19**, pp. 229–255, 2004.

[22] X. Ros-Oton, *Nonlocal elliptic equations in bounded domains: a survey*, Publ. Mat., **60**, pp. 3–26, 2016.

[23] R. Servadei and E. Valdinoci, *A Brezis-Nirenberg result for non-local critical equations in low dimension*, Communications on Pure & Applied Analysis, **12**, pp. 2445–2464, 2013.

[24] R. Servadei and E. Valdinoci, *Variational methods for non-local operators of elliptic type*, Discrete Contin. Dyn. Syst., **33**, pp. 2105–2137, 2013.

[25] S. A. Silling, *Reformulation of elasticity theory for discontinuities and long-range forces*, J. Mech. Phys. Solids, **48**, pp. 175–209, 2000.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 82–89)

A semi-implicit Lagrange-projection-type finite volume scheme exactly well-balanced for 1D shallow-water system

C. Caballero-Cárdenas[1], M. J. Castro[1], T. Morales de Luna[2], M. L. Muñoz-Ruiz[1]
1. *Universidad de Málaga, Spain*
2. *Universidad de Córdoba, Spain*

Abstract

In this work we present a numerical approximation of the shallow water equations based on a Lagrange-projection-type finite volume scheme. For the Lagrangian step we propose two different versions of the scheme, two explicit and other two semi-implicit that ensure first and second order of accuracy. The projection on the Eulerian coordinates will always be done explicitly, preserving the total order of the scheme. Special care is done for ensuring the well-balanced properties of the scheme. Several numerical experiments are included in order to illustrate the good behavior of the proposed schemes.

1. Introduction

Let us consider the shallow water equations (SWE), given by

$$\begin{cases} \partial_t h + \partial_x(hu) = 0, \\ \partial_t(hu) + \partial_x\left(hu^2 + g\dfrac{h^2}{2}\right) = -gh\partial_x z, \end{cases} \tag{1.1}$$

where $z(x)$ denotes a given smooth topography and $g > 0$ is the gravity constant. The primitive variables are the water depth $h \geq 0$ and its velocity u, which both depend on the space and time variables, respectively, $x \in \mathbb{R}$ and $t \in [0, \infty)$. We assume that the initial water depth $h(x, 0) = h_0(x)$ and velocity $u(x, 0) = u_0(x)$ are given.

The use of Lagrangian coordinates allows to describe the flow by following the fluid motion. With this in mind, for any given "fluid particle", ξ, we consider the characteristic curves

$$\begin{cases} \dfrac{\partial x}{\partial t}(\xi, t) = u(x(\xi, t), t), \\ x(\xi, 0) = \xi, \end{cases}$$

and given any function $(x, t) \mapsto \mathbf{U}(x, t)$ in Eulerian coordinates, we denote by

$$\overline{\mathbf{U}}(\xi, t) = \mathbf{U}(x(\xi, t), t)$$

its counterpart in Lagrangian coordinates.

Moreover, we define

$$L(\xi, t) = \frac{\partial x}{\partial \xi}(\xi, t),$$

which implies that $\partial_t L(\xi, t) = \partial_\xi \overline{u}(\xi, t)$.

Since system (1.1) can be written for smooth solutions as

$$\begin{cases} \partial_t h + u\partial_x h + h\partial_x u = 0, \\ \partial_t(hu) + u\partial_x(hu) + hu\partial_x u + \partial_x\left(g\dfrac{h^2}{2}\right) = -gh\partial_x z, \end{cases}$$

after multiplying both equations by $L(\xi, t)$ and setting $p = gh^2/2$ we obtain

$$\begin{cases} \partial_t(L\overline{h}) = 0, \\ \partial_t(L\overline{hu}) + \partial_\xi \overline{p} + g\overline{h}\partial_\xi \overline{z} = 0. \end{cases} \tag{1.2}$$

The Lagrangian-projection scheme can be interpreted as a two-step algorithm consisting in first solving the system in Lagrangian coordinates (1.2), which is known as the Lagrangian step, and then projecting the results in Eulerian coordinates, which is known as the Projection step. See [9] for more details. This strategy allows us to decouple the acoustic and transport phenomena and to design a natural implicit-explicit and large time steps could be considered with a CFL restriction based on the slower transport waves and not on the acoustic ones. We address the reader to [5–8] for further details. In this work we follow the strategy described in [3, 10] for the definition of the LP scheme and [4] to ensure its well-balanced character. Concerning the well-balanced property, we refer the reader to [1] for a different approach, and to [2, 4] and the references therein, for a review on this topic.

2. The Lagrange-projection numerical algorithm

Space and time will be discretized using a space step Δx and a time step Δt in a set of cells $[x_{i-1/2}, x_{i+1/2}]$ and instants $t^n = n\Delta t$, for $\in \mathbb{Z}$, $n \in \mathbb{N}$. We consider Δx and Δt constants for simplicity. We define $x_{i+1/2} = i\Delta x$ and $x_i = (x_{i-1/2} + x_{i+1/2})/2$, the cell interfaces and the cell centers, respectively. We consider for the variable ξ the same space discretization as for x, that is, $\Delta\xi = \Delta x$, $\xi_{i+1/2} = x_{i+1/2}$ and $\xi_i = x_i$ for all $i \in \mathbb{Z}$.

Let $\mathbf{U} = (h, hu)^t$. For a given initial condition $x \mapsto \mathbf{U}^0(x)$, we will consider a discrete initial condition $\mathbf{U}_i^0$, which approximates $\dfrac{1}{\Delta x}\displaystyle\int_{x_{i-1/2}}^{x_{i+1/2}} \mathbf{U}^0(x)dx$, for $i \in \mathbb{Z}$. The proposed algorithm aims at computing an approximation $\mathbf{U}_i^n$ of

$$\mathbf{U}_i^n \approx \frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} \mathbf{U}(x, t^n)dx,$$

where $x \mapsto \mathbf{U}(x, t^n)$ is the exact solution of (1.1) at time t^n, $n \in \mathbb{N}$. Given the sequence $\{\mathbf{U}_i^n\}_i$, it is a matter of defining the sequence $\{\mathbf{U}_i^{n+1}\}_i$, $n \in \mathbb{N}$, since $\{\mathbf{U}_i^0\}_i$ is assumed to be known.

Using these notation, the overall Lagrange-projection algorithm can be described as follows: for a given discrete state $\mathbf{U}_i^n = (h, hu)_i^n, i \in \mathbb{Z}$, that describes the system at instant t^n, the computation of the approximation $\mathbf{U}_i^{n+1} = (h, hu)_i^{n+1}$ at the next time level is a two-step process defined by

1. update $\mathbf{U}_i^n$ to $\overline{\mathbf{U}}_i^{n+1}$ by approximating the solution of (1.2);
2. update $\overline{\mathbf{U}}_i^{n+1}$ to $\mathbf{U}_i^{n+1}$ by projecting the solution back to the Eulerian coordinates.

3. The Lagrangian step

Concerning the Lagrangian step, we will approximate the solution of (1.2) in two ways: explicitly and implicitly.

For simplicity, from now on we will focus on the flat topography case, which corresponds to $\partial_x z = 0$. The non-flat case can be carried out by adapting the schemes presented here as proposed in [3, 4, 10]. A detailed description of the well-balanced schemes will be done during the presentation.

Following [3, 8], we will now consider a relaxation approach of the Lagrangian formulation:

$$\begin{cases} \partial_t \tau - \partial_m u = 0, \\ \partial_t u + \partial_m \Pi = \lambda(p - \Pi), \\ \partial_t \Pi + a^2 \partial_m u = 0, \end{cases} \tag{3.1}$$

where $\tau = 1/h$, $\tau\partial_x = \partial_m$, a is a constant satisfying the subcharacteristic condition $a > h\sqrt{gh}$ and $\lambda \to \infty$.

From a numerical point of view, the strategy consists in first solving (3.1) with $\lambda = 0$, that is,

$$\begin{cases} \partial_t \tau - \partial_m u = 0, \\ \partial_t u + \partial_m \Pi = 0, \\ \partial_t \Pi + a^2 \partial_m u = 0, \end{cases} \tag{3.2}$$

and then setting $\Pi = p(\tau)$ before going to the following iteration.

Defining the two new variables $\overrightarrow{w} = \Pi + au$ and $\overleftarrow{w} = \Pi - au$, system (3.2) can be written as

$$\begin{cases} \partial_t \tau - \partial_m u = 0, \\ \partial_t \overrightarrow{w} + a\partial_m \overrightarrow{w} = 0, \\ \partial_t \overleftarrow{w} - a\partial_m \overleftarrow{w} = 0. \end{cases} \tag{3.3}$$

In what follows, we present a first order fully explicit and semi-implicit scheme and their extensions to second order. The second order extension is performed using a second order reconstruction operator combined with a suitable second order time integration.

3.1. Explicit Lagrangian schemes

First order explicit Lagrangian scheme

The Lagrangian step for the first order explicit scheme can be written as (see [3] for details)

$$\begin{cases} (L\overline{h})_i^{n+1} = h_i^n, \\ (L\overline{hu})_i^{n+1} = (hu)_i^n - \dfrac{\Delta t}{\Delta x}(\pi_{i+1/2}^* - \pi_{i-1/2}^*), \end{cases}$$

where

$$L_i^{n+1} = 1 + \frac{\Delta t}{\Delta x}(u^*_{i+1/2} - u^*_{i-1/2})$$

and

$$\pi^*_{i+1/2} = \frac{\overrightarrow{w}_i^n + \overleftarrow{w}_{i+1}^n}{2} = \frac{\pi_i^n + \pi_{i+1}^n}{2} - \frac{a}{2}(u_{i+1}^n - u_i^n),$$
$$u^*_{i+1/2} = \frac{\overrightarrow{w}_i^n - \overleftarrow{w}_{i+1}^n}{2a} = \frac{u_i^n + u_{i+1}^n}{2} - \frac{1}{2a}(\pi_{i+1}^n - \pi_i^n).$$

Second order explicit Lagrangian scheme
The second order explicit scheme, after using a MUSCL-Hancock reconstruction, can be described again as

$$\begin{cases} (L\overline{h})_i^{n+1} = h_i^n, \\ (L\overline{hu})_i^{n+1} = (hu)_i^n - \dfrac{\Delta t}{\Delta x}(\pi^*_{i+1/2} - \pi^*_{i-1/2}), \end{cases}$$

with

$$L_i^{n+1} = 1 + \frac{\Delta t}{\Delta x}(u^*_{i+1/2} - u^*_{i-1/2}),$$

where now

$$\pi^*_{i+1/2} = \frac{\overrightarrow{w}^{n+1/2}_{i+1/2-} + \overleftarrow{w}^{n+1/2}_{i+1/2+}}{2},$$
$$u^*_{i+1/2} = \frac{\overrightarrow{w}^{n+1/2}_{i+1/2-} - \overleftarrow{w}^{n+1/2}_{i+1/2+}}{2a},$$

and

$$\overrightarrow{w}^{n+1/2}_{i+1/2-} = \overrightarrow{w}_i^n + \frac{\Delta x}{2}\delta\overrightarrow{w}_i^n - \frac{a\Delta t}{2h_i^n}\delta\overrightarrow{w}_i^n,$$
$$\overleftarrow{w}^{n+1/2}_{i+1/2+} = \overleftarrow{w}_{i+1}^n - \frac{\Delta x}{2}\delta\overleftarrow{w}_{i+1}^n + \frac{a\Delta t}{2h_{i+1}^n}\delta\overleftarrow{w}_{i+1}^n.$$

In the previous expression, $\delta\overrightarrow{w}_i^n$ and $\delta\overleftarrow{w}_{i+1}^n$ are approximations of the space derivative of $\overrightarrow{w}(x_i, t^n)$ and $\overleftarrow{w}(x_{i+1}, t^n)$ respectively, that are computed by means of a limiter that avoids the appearance of spurious oscillations in the presence of discontinuities. In this work we use:

$$\delta\overrightarrow{w}_i^n = \phi_L(d_{i,l}, d_{i,r})d_{i,l} + \phi_R(d_{i,l}, d_{i,r})d_{i,r} \tag{3.4}$$

where

$$d_{i,l} = \frac{\overrightarrow{w}_i^n - \overrightarrow{w}_{i-1}^n}{\Delta x},\ d_{i,r} = \frac{\overrightarrow{w}_{i+1}^n - \overrightarrow{w}_i^n}{\Delta x}.$$

and

$$\phi_L(a,b) = \begin{cases} \dfrac{|b|}{|a|+|b|}, \text{ if } |a|+|b| > 0, \\ 0, \text{ otherwise.} \end{cases} \quad \text{and } \phi_R(a,b) = \begin{cases} \dfrac{|a|}{|a|+|b|}, \text{ if } |a|+|b| > 0, \\ 0 \text{ otherwise.} \end{cases}$$

3.2. Implicit Lagrangian schemes

First order implicit Lagrangian scheme
In this section we present the first order implicit scheme for the Lagrangian step. Note, that formally, the explicit and the implicit schemes have a similar expression, but now, $\pi^*_{i+1/2}$ and $u^*_{i+1/2}$ have to be evaluated at time $t = t^{n+1}$:

$$\begin{cases} (L\overline{h})_i^{n+1} = h_i^n, \\ (L\overline{hu})_i^{n+1} = (hu)_i^n - \dfrac{\Delta t}{\Delta x}(\pi^*_{i+1/2} - \pi^*_{i-1/2}), \end{cases}$$

with

$$L_i^{n+1} = 1 + \frac{\Delta t}{\Delta x}(u^*_{i+1/2} - u^*_{i-1/2}),$$

where now

$$\pi^*_{i+1/2} = \frac{\overrightarrow{w}^{n+1}_i + \overleftarrow{w}^{n+1}_{i+1}}{2},$$
$$u^*_{i+1/2} = \frac{\overrightarrow{w}^{n+1}_i - \overleftarrow{w}^{n+1}_{i+1}}{2a},$$

and $\overrightarrow{w}^{n+1}_i$, $\overleftarrow{w}^{n+1}_i$ are the solutions of

$$\overrightarrow{w}^{n+1}_i = \overrightarrow{w}^{n}_i - \frac{a\Delta t}{h^n_i \Delta x}(\overrightarrow{w}^{n+1}_i - \overrightarrow{w}^{n+1}_{i-1}), \tag{3.5}$$

$$\overleftarrow{w}^{n+1}_i = \overleftarrow{w}^{n}_i + \frac{a\Delta t}{h^n_i \Delta x}(\overleftarrow{w}^{n+1}_{i+1} - \overleftarrow{w}^{n+1}_i), \tag{3.6}$$

Note that due to the special form of (3.5) and (3.6), $\overrightarrow{w}^{n+1}_i$ and $\overleftarrow{w}^{n+1}_i$ can be computed in a very simple way.

Second order implicit Lagrangian scheme
The second order implicit scheme is defined combining the second order Adams-Moulton scheme for the time integration, and a second order reconstruction procedure. The resulting scheme reads as follows

$$\begin{cases} (L\overline{h})^{n+1}_i = h^n_i, \\ (L\overline{hu})^{n+1}_i = (hu)^n_i - \dfrac{\Delta t}{2\Delta x}\left(\pi^{*,n}_{i+1/2} - \pi^{*,n}_{i-1/2} + \pi^{*,n+1}_{i+1/2} - \pi^{*,n+1}_{i-1/2}\right), \end{cases}$$

with

$$L^{n+1}_i = 1 + \frac{\Delta t}{2\Delta x}\left(u^{*,n}_{i+1/2} - u^{*,n}_{i-1/2} + u^{*,n+1}_{i+1/2} - u^{*,n+1}_{i-1/2}\right),$$

where

$$\pi^{*,\#}_{i+1/2} = \frac{\overrightarrow{w}^{\#}_{i+1/2-} + \overleftarrow{w}^{\#}_{i+1/2+}}{2},$$
$$u^{*,\#}_{i+1/2} = \frac{\overrightarrow{w}^{\#}_{i+1/2-} - \overleftarrow{w}^{\#}_{i+1/2+}}{2a},$$

where # stands for n or $n+1$. The space reconstruction is performed as follows

$$\overrightarrow{w}^{\#}_{i+1/2\pm} = \overrightarrow{w}^{\#}_i \mp \frac{\Delta x}{2}\left(\phi_L(d^n_{i,l}, d^n_{i,r})d^{\#}_{i,l} + \phi_R(d^n_{i,l}, d^n_{i,r})d^{\#}_{i,r}\right).$$

We can define $\overleftarrow{w}^{\#}_{i+1/2\pm}$ in a similar way. Note that the limiters are always evaluated at time t^n, therefore, the systems that define $\overrightarrow{w}^{n+1}_i$ and $\overleftarrow{w}^{n+1}_i$, respectively, remain linear and have the following form:

$$\overrightarrow{w}^{n+1}_i = \overrightarrow{w}^{n}_i - \frac{a\Delta t}{2h^n_i \Delta x}\left(\overrightarrow{w}^{n}_{i+1/2-} - \overrightarrow{w}^{n}_{i-1/2-} + \overrightarrow{w}^{n+1}_{i+1/2-} - \overrightarrow{w}^{n+1}_{i-1/2-}\right),$$
$$\overleftarrow{w}^{n+1}_i = \overleftarrow{w}^{n}_i + \frac{a\Delta t}{2h^n_i \Delta x}\left(\overleftarrow{w}^{n}_{i+1/2+} - \overleftarrow{w}^{n}_{i-1/2+} + \overleftarrow{w}^{n+1}_{i+1/2+} - \overleftarrow{w}^{n+1}_{i-1/2+}\right).$$

4. The projection step

Once the system in Lagrangian coordinates is solved, the result has to be projected in Eulerian coordinates. This step will always be done explicitly.

For doing the projection of $L\overline{U}(\xi, t)$ on the Eulerian cells $(x_{i-1/2}, x_{i+1/2})$, we need to compute

$$\mathbf{U}_i(t) = \frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} \mathbf{U}(x,t)dx.$$

Given $t \geq 0$ we define $\hat{\xi}_{i+1/2}(t)$ such that

$$x(\hat{\xi}_{i+1/2}(t), t) = x_{i+1/2}.$$

Thus, for any time $T \geq 0$, $\hat{\xi}_{i+1/2}(T)$ corresponds to the origin of the characteristic $x(\hat{\xi}_{i+1/2}, t)$ such that at time $t = T$ coincides with $x_{i+1/2}$.

Using this notation, we can write

$$\mathbf{U}_i(t) = \frac{1}{\Delta x} \int_{x(\hat{\xi}_{i-1/2}(t),t)}^{x(\hat{\xi}_{i+1/2}(t),t)} \mathbf{U}(x,t)dx = \frac{1}{\Delta x} \int_{\hat{\xi}_{i-1/2}(t)}^{\hat{\xi}_{i+1/2}(t)} L(\xi,t)\overline{\mathbf{U}}(\xi,t)d\xi,$$

and we can split the integral as follows

$$\mathbf{U}_i(t) = \frac{1}{\Delta x} \int_{\hat{\xi}_{i-1/2}(t)}^{\xi_{i-1/2}(t)} L(\xi,t)\overline{\mathbf{U}}(\xi,t)d\xi + \frac{1}{\Delta x} \int_{\xi_{i-1/2}(t)}^{\xi_{i+1/2}(t)} L(\xi,t)\overline{\mathbf{U}}(\xi,t)d\xi + \frac{1}{\Delta x} \int_{\xi_{i+1/2}(t)}^{\hat{\xi}_{i+1/2}(t)} L(\xi,t)\overline{\mathbf{U}}(\xi,t)d\xi.$$

Note that the middle integral in the right-hand-side equals $(L\overline{\mathbf{U}})_i(t)$, which is known from the Lagrangian step. Therefore,

$$\mathbf{U}_i^{n+1} = (L\overline{\mathbf{U}})_i^{n+1} + \frac{1}{\Delta x} \int_{\hat{\xi}_{i-1/2}}^{\xi_{i-1/2}} L(\xi,t^{n+1})\overline{\mathbf{U}}(\xi,t^{n+1})d\xi + \frac{1}{\Delta x} \int_{\xi_{i+1/2}}^{\hat{\xi}_{i+1/2}} L(\xi,t^{n+1})\overline{\mathbf{U}}(\xi,t^{n+1})d\xi. \tag{4.1}$$

It remains to evaluate the other two integrals. We will now present the corresponding first and second order numerical schemes for doing this.

4.1. First order projection scheme

The previous integrals can be approximated in the following way:

$$\frac{1}{\Delta x} \int_{\xi_{i+1/2}}^{\hat{\xi}_{i+1/2}} L(\xi,t^{n+1})\overline{\mathbf{U}}(\xi,t^{n+1})d\xi = \frac{\hat{\xi}_{i+1/2} - \xi_{i+1/2}}{\Delta x}(L\overline{\mathbf{U}})_{i+1/2}^{n+1},$$

where

$$(L\overline{\mathbf{U}})_{i+1/2}^{n+1} = \begin{cases} (L\overline{\mathbf{U}})_i^{n+1} \text{ for } \xi_{i+1/2} > \hat{\xi}_{i+1/2}, \\ (L\overline{\mathbf{U}})_{i+1}^{n+1} \text{ for } \xi_{i+1/2} \leq \hat{\xi}_{i+1/2}. \end{cases}$$

Moreover, since for sufficiently small Δt we can use the approximation

$$\hat{\xi}_{i+1/2} = x_{i+1/2} - \Delta t u_{i+1/2}^*,$$

then from (4.1) we obtain

$$\mathbf{U}_i^{n+1} = (L\overline{\mathbf{U}})_i^{n+1} - \frac{\Delta t}{\Delta x}\left(u_{i+1/2}^*(L\overline{\mathbf{U}})_{i+1/2}^{n+1} - u_{i-1/2}^*(L\overline{\mathbf{U}})_{i-1/2}^{n+1}\right).$$

4.2. Second order projection scheme

In order to obtain a second order approximation of the integrals in (4.1), we will consider a linear reconstruction of the cell averages of $(L\overline{\mathbf{U}})_i^{n+1}$ and the velocities $u_{i+1/2}^{*,n+1}$ that are continuously defined at the intercells.

It can be seen that we can write (4.1) again as

$$\mathbf{U}_i^{n+1} = (L\overline{\mathbf{U}})_i^{n+1} - \frac{\Delta t}{\Delta x}\left(u_{i+1/2}^*(L\overline{\mathbf{U}})_{i+1/2}^{n+1} - u_{i-1/2}^*(L\overline{\mathbf{U}})_{i-1/2}^{n+1}\right),$$

where

$$(L\overline{\mathbf{U}})_{i+1/2}^{n+1} = \begin{cases} (L\overline{\mathbf{U}})_{i+1/2-}^{n+1} \text{ for } u_{i+1/2}^* > 0, \\ (L\overline{\mathbf{U}})_{i+1/2+}^{n+1} \text{ for } u_{i+1/2}^* \leq 0, \end{cases}$$

and

$$\left(L\overline{\mathbf{U}}\right)_{i+1/2-}^{n+1} = \frac{1}{L_i^{n+1}}\left(\left(L\overline{\mathbf{U}}\right)_i^{n+1} + \frac{1}{2}\left(\delta L\overline{\mathbf{U}}\right)_i^{n+1}\left(\Delta x - \frac{\Delta t}{L_i^{n+1}}u_{i+1/2}^{*,n+1}\right)\right),$$

$$\left(L\overline{\mathbf{U}}\right)_{i+1/2+}^{n+1} = \frac{1}{L_{i+1}^{n+1}}\left(\left(L\overline{\mathbf{U}}\right)_{i+1}^{n+1} + \frac{1}{2}\left(\delta L\overline{\mathbf{U}}\right)_{i+1}^{n+1}\left(-\Delta x - \frac{\Delta t}{L_{i+1}^{n+1}}u_{i+1/2}^{*,n+1}\right)\right).$$

In the previous expressions, $\left(\delta L\overline{\mathbf{U}}\right)_i^{n+1}$ and $\left(\delta L\overline{\mathbf{U}}\right)_{i+1}^{n+1}$ are approximations of the derivatives of $L\overline{\mathbf{U}}$ at time t^{n+1} at x_i and x_{i+1}, respectively, that are computed using (3.4).

$(L\overline{\mathbf{U}})_{i-1/2}^{n+1}$ is defined in a similar way.

Now, the previous numerical schemes could be extended to the general case of non-flat topography. For that we follow [3, 4] to ensure the well-balanced properties of the schemes. More details will be shown during the presentation.

5. Numerical experiments

To be consistent with the numerical scheme presented previously, we only consider here a simple test case with flat bottom topography.

The aim of this test if to check that the numerical schemes that have been proposed achieves the expected order of accuracy. Let us consider a flat topography in the interval $[-5, 5]$, with initial zero velocity ($u = 0$) and an initial water depth with a small perturbation given by

$$h(x, 0) = 1 + 0.1 \exp\big(-x^2\big).$$

Figures 1 and 2 correspond to this initial condition.

Periodic boundary conditions are considered.

Tables 1 and 2 show the errors corresponding to the different methods. Notice that the errors for the explicit and the semi-implicit schemes decrease as the number of cells increase at the expected rate.

	Explicit scheme - Order 1				Explicit scheme - Order 2			
No. of cells	h		q		h		q	
	Error	Order	Error	Order	Error	Order	Error	Order
25	6.09E-2	-	1.60E-1	-	3.69E-2	-	1.18E-1	-
50	3.70E-2	0.72	9.97E-2	0.68	1.21E-2	1.61	3.86E-2	1.61
100	2.07E-2	0.84	5.63E-2	0.82	3.16E-3	1.93	1.03E-2	1.90
200	1.07E-2	0.95	2.95E-2	0.93	7.80E-4	2.02	2.58E-3	2.00
400	5.20E-3	1.05	1.44E-2	1.04	1.89E-4	2.05	6.27E-4	2.04

Tab. 1 Errors in L^1 norm and convergence rates for the explicit LP schemes of order 1 and 2.

	Implicit scheme - Order 1				Implicit scheme - Order 2			
No. of cells	h		q		h		q	
	Error	Order	Error	Order	Error	Order	Error	Order
25	7.99E-2	-	2.03E-1	-	3.80E-2	-	1.19E-1	-
50	5.13E-2	0.64	1.35E-1	0.59	1.27E-2	1.58	3.99E-2	1.57
100	3.05E-2	0.75	8.19E-2	0.73	3.35E-3	1.93	1.07E-2	1.90
200	1.66E-2	0.88	4.52E-2	0.86	8.00E-4	2.07	2.55E-3	2.07
400	8.19E-3	1.02	2.26E-2	1.00	1.85E-4	2.12	5.89E-4	2.12

Tab. 2 Errors in L^1 norm and convergence rates for the semi-implicit LP schemes of order 1 and 2.

We now consider a uniform mesh composed by 400 cells. The solution at time $t = 0.5$ for variables h and u using the order 1 and order 2 explicit schemes is shown in Figures 3 and 4. The same for the semi-implicit schemes can be seen in 5 and 6 with CFL value 0.5 and in Figures 7 and 8 with CFL value 2. Observe that the semi-implicit second order scheme with CFL=2 behaves similar to the explicit scheme with CFL=0.5.

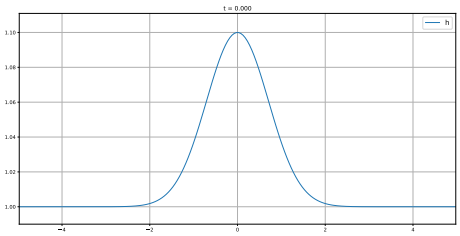


Fig. 1 Initial condition for the variable h.

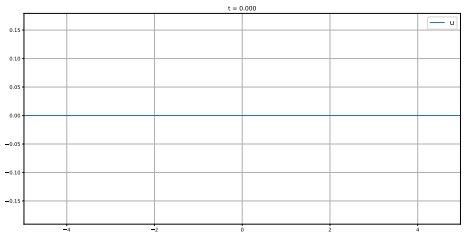


Fig. 2 Initial condition for the variable u.

Acknowledgements

This research has been partially supported by the Spanish Government and FEDER through the coordinated Research project RTI2018-096064-B-C1 and RTI2018-096064-B-C2, the Junta de Andalucía research project P18-RT-3163 and the Junta de Andalucía-FEDER-University of Málaga Research project UMA18-FEDERJA-161. C. Caballero-Cárdenas is also supported by a Grant from "Ministerio de Ciencia e Innovación", Spain (FPI2019/087773).

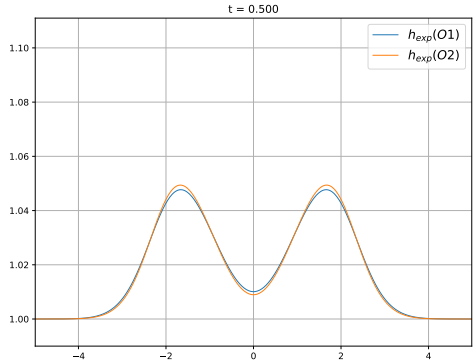


Fig. 3 Solution at time $t = 0.5$ for the variable h using the explicit schemes of order 1 and 2.

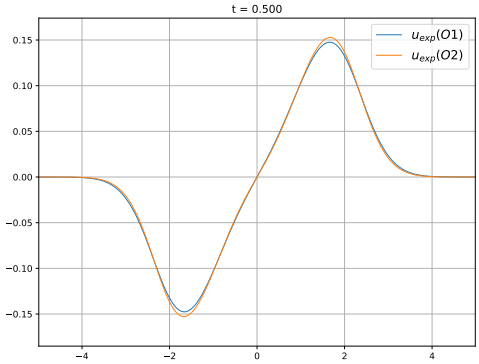


Fig. 4 Solution at time $t = 0.5$ for the variable u using the explicit schemes of order 1 and 2.

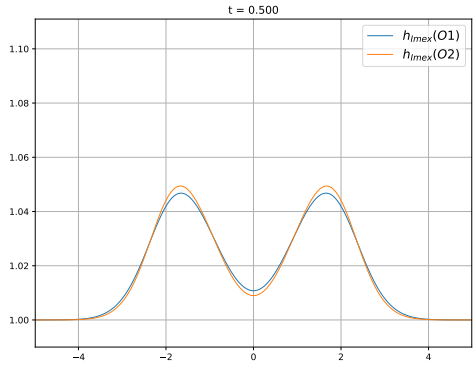


Fig. 5 Solution at time $t = 0.5$ for the variable h using the semi-implicit schemes of order 1 and 2 with CFL 0.5.

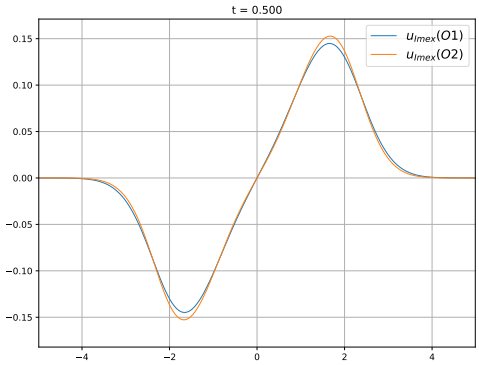


Fig. 6 Solution at time $t = 0.5$ for the variable u using the semi-implicit schemes of order 1 and 2 with CFL 0.5.

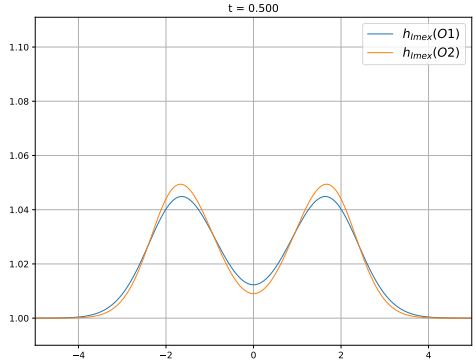


Fig. 7 Solution at time $t = 0.5$ for the variable h using the semi-implicit schemes of order 1 and 2 with CFL 2.

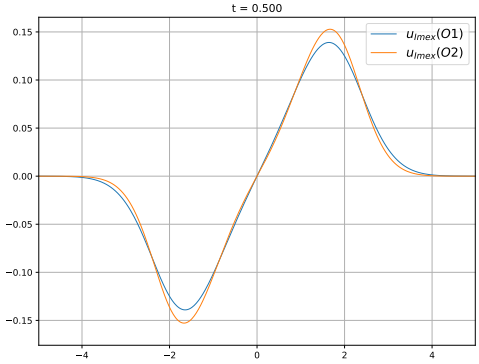


Fig. 8 Solution at time $t = 0.5$ for the variable u using the semi-implicit schemes of order 1 and 2 with CFL 2.

References

[1] Georgij Bispen, Koottungal Revi Arun, Maria Lukacova-Medvidova, and Sebastian Noelle. IMEX Large Time Step Finite Volume Methods for Low Froude Number Shallow Water Flows. *Communications in Computational Physics*, 16:307–347, 2014.

[2] François Bouchut. *Nonlinear stability of finite volume methods for hyperbolic conservation laws and well-balanced schemes for sources*. Frontiers in Mathematics. Birkäuser Verl., 2004.

[3] Manuel J. Castro, Christophe Chalons, and Tomás Morales de Luna. A Fully Well-Balanced Lagrange–Projection-Type Scheme for the Shallow-Water Equations. *SIAM Journal on Numerical Analysis*, 56(5):3071–3098, 2018.

[4] Manuel J. Castro and C. Parés. Well-Balanced High-Order Finite Volume Methods for Systems of Balance Laws. *Journal of Scientific Computing*, 82, 2020.

[5] Christophe Chalons, Mathieu Girardin, and Samuel Kokh. Large Time Step and Asymptotic Preserving Numerical Schemes for the Gas Dynamics Equations with Source Terms. *SIAM Journal on Scientific Computing*, 35:A2874–A2902, 2013.

[6] Christophe Chalons, Mathieu Girardin, and Samuel Kokh. Operator-splitting based AP schemes for the 1D and 2D gas dynamics equations with stiff sources. In *Hyperbolic Problems: Theory, Numerics, Applications*, volume 8 of *AIMS on Applied Mathematics*, pages 607–614, Springfield, MO, 2014.

[7] Christophe Chalons, Mathieu Girardin, and Samuel Kokh. An All-Regime Lagrange-Projection Like Scheme for the Gas Dynamics Equations on Unstructured Meshes. *Communications in Computational Physics*, 20:188–233, 2016.

[8] Christophe Chalons, Pierre Kestener, Samuel Kokh, and Maxime Stauffert. A large time-step and well-balanced Lagrange-Projection type scheme for the shallow-water equations. *Communications in Mathematical Sciences*, 15:765–788, 2017.

[9] Edwige Godlewski and Pierre-Arnaud Raviart. *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, volume 118 of *Appl. Math. Sci.* Springer, 1996.

[10] Tomas Morales De Luna, Manuel J. Castro, and Christophe Chalons. High-order fully well-balanced Lagrange-Projection scheme for shallow water. *Communications in Mathematical Sciences*, 18:781–807, 2020.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 90–96)

SEIRD model with nonlocal diffusion

Alfredo N. Calvo Pereira
Universidad de Salamanca
alcalvo@usal.es

Abstract

Some non-local diffussion processes may be described by an integral operator such as $\int K(x,y)(u(y)-u(x))dy$. In the compartmental models in epidemilogy, the diffusion terms may be replaced by an integral term, similar to above, to describe the dispersion of population. Classical numerical approximations, like finite element methods or finite difference methods, lead to systems of equations with dense matrices. Wavelets are a kind of functions which have some properties that make them a very useful tool for discretizing the convolution operators getting sparse matrices, thus saving computing time and storage memory.

1. Introduction

From the classical theoretical works on epidemic models by Kermack and McKendrick (1927), a big effort has been made to apply mathematics to the spread and control of infectious diseases (see, for instance, [5]). In particular, the called compartmental models distribute population in compartments and explain the mechanism through which individuals move from one compartment to another.

The spatial spread of the disease is incorporated in the models through diffusion terms that describe the local movement of population. The nonlocal diffusion models are more realistic, but when we discretize these models with classical methods as finite element or difference methods led to algebraic systems with large and dense matrices. We can circunvect this problem using a specific kind of functions, wavelets, to discretize the integral operators. Due to some of the properties of these functions, if the integral operators show certain properties, most of the matrix elements that we get from the discretization process are very close to zero and may be neglected, with the consequent saving in time and computer storage.

2. Multiresolution Analysis and wavelets

We give a brief review of Multiresolution Analysis, which is the starting point for constructing wavelets, and we show some relevant properties that explain why wavelets play a definitive role in the discretization of some kind of problems. For details, see, for instance, [2, 3].

Let us consider a sequence of spaces $\{V_j\}_{j\in\mathbb{Z}}$, that we will use to approximate functions with a level of resolution j, so we need a basis of functions to generate those spaces.

Definition 2.1 A sequence of spaces $\{V_j\}_{j\in\mathbb{Z}}$ of spaces $V_j \subset L_2(\mathbb{R})$ is called a Multiresolution Analysis, if

a) The spaces are nested, i.e. $V_j \subset V_{j+1}$, so the information contained in some level of resolution is also contained in the finer.

b) Any function may be approximated with arbitrary precission, i.e. $\lim_{j\to\infty} \|f - P_j f\|_{L^2} = 0$

c) The only common function is 0, i.e. $\lim_{j\to-\infty} \|P_j f\|_{L^2} = 0$

d) The spaces arise by scaling with factor 2: $f(\cdot) \in S_j \Leftrightarrow f(2\cdot) \in S_{j+1} \Leftrightarrow f(2^{-j}\cdot) \in S_0$

e) The spaces are shift-invariant: $f(\cdot) \in S_0 \Leftrightarrow f(\cdot - k) \in S_0$

f) $\exists\, \varphi$ s.t. $\Phi_j := \{\varphi_{j,k} : k \in \mathbb{Z}\}$ is a uniformly stable basis for V_j, $j \in \mathbb{Z}$, being $\varphi_{j,k}$ the dilated and traslated version of some function φ to be determined:

$$\varphi_{j,k}(x) = 2^{j/2}\varphi(2^j x - k) \tag{2.1}$$

From the condition a) and f) in definition 2.1, the function that is going to be the basis for the approximation space V_j have to verify the following recursion formula

$$\varphi(x) = \sum_k a_k \varphi(2x - k) \tag{2.2}$$

The function $\varphi(x)$ is called a scaling function or refinement function and its properties are determined by the refinement coefficientes $a_k \in \mathbb{R}, k \in \mathbb{Z}$. The approximation of functions is carried out projecting the function onto the space V_j, such that

$$P_j : L_2(\mathbb{R}) \to V_j, \quad P_j f = \sum_k c_{j,k} \varphi_{jk}(x). \tag{2.3}$$

The difference between two consecutive levels of resolution is encoded by functions $\psi_{j,k}(x)$, which *complete* the information in V_j to achive the information in V_{j+1}. These functions are called wavelets and the space W_j generated by

$$\psi_{j,k} = 2^{j/2} \psi(2^j x - k) \tag{2.4}$$

is the complement of V_j in V_{j+1}, i.e. $V_{j+1} = V_j \oplus W_j$. As $W_0 \subset V_1$, wavelets also verify a recursion formula similar to (2.2),

$$\psi(x) = \sum_k b_k \varphi(2x - k) \tag{2.5}$$

where the coefficients b_k are determined by the coefficients a_k.

So, we have spaces V_j and spaces W_j such that

$$V_j = \text{span } \{\varphi_{j,\lambda}\}, \quad W_j = \text{span } \{\psi_{j,\lambda}\}, \tag{2.6}$$

and we can approximate f in V_J or we can carry out an initial approximation V_{J_0} and add *details* to achive J, using spaces W_j, projecting the function on

$$V_J = V_{J-1} \oplus W_{J-1} = V_{J-2} \oplus W_{J-2} \oplus W_{J-1} = \cdots = V_{J_0} \oplus \bigoplus_{j=J_0}^{J-1} W_j \tag{2.7}$$

therefore we approximate f with a coarse level J_0 and we add finer resolution j:

$$f_J = \sum_\lambda c_\lambda^{J_0} \varphi_{J_0,k} + \sum_{j=J_0}^{J-1} \sum_\lambda d_k^j \psi_{j,k}. \tag{2.8}$$

The next two results (proofs and details can be seen in [2, 3] and references there) about wavelets are the key for compression and for the potential of wavelets in numerical calculations:

Proposition 2.2 *If the scaling functions reproduce exactly a polynomial of order d, the associated wavelets have d vanishing moments*

$$M_r(\psi_{j,k}) = \int_\Omega x^r \psi_{j,k}(x) dx = 0 \quad 0 \le r < d \tag{2.9}$$

Proposition 2.3 *If ψ have d vanishing moments, then*

$$|d_k^j| \lesssim 2^{-js} \|f\|_{s;supp\ \psi_{j,k}} \tag{2.10}$$

where d_k^j are the coefficients in (2.8). As a consequence of (2.10), many of the coefficientes of wavelets are small in the case of functions that are *regular enough* and, therefore, may be neglected.

3. Nonlocal models

The classical approach to diffusion problems implies local effects; nonlocal evolution equations of the form

$$\partial_t u(x,t) = \int_\Omega K(x.y) u(y,t) dy - u(x,t) \tag{3.1}$$

take into account long range effects and have been used to model nonlocal diffusion processes, replacing the local term

$$\partial_t u(x,t) = \nabla(\nu \nabla u(x,t)) \tag{3.2}$$

by its nonlocal version (3.1), [1].
Discretization in time of (3.1) leads to

$$\frac{u^{n+1}(x) - u^n(x)}{\delta t} = \int_\Omega K(x-y) u^{n+1}(y) - u^{n+1}(x) \tag{3.3}$$

Let us consider the variational formulation of (3.3), writing functions $u^n(x)$ in wavelet basis

$$u^n(x) = \sum_\lambda d^n_\lambda \psi_\lambda(x) \tag{3.4}$$

and using as test function $\psi_\mu(x)$ (to simplify notation, we don'n distinguish here between scaling functions and wavelets, and include the level of resolution j in the index λ, which should be (j,λ))

$$\begin{aligned} &\sum_\lambda d^{n+1} \int_\Omega \psi_\lambda(x)\psi_\mu(x)dx - \sum_\lambda d^n \int_\Omega \psi_\lambda(x)\psi_\mu(x)dx \\ &= \delta t \sum_\lambda d^{n+1} \int_\Omega\int_\Omega \psi_\lambda(y)\psi_\mu(x)dxdy - \delta t \sum_\lambda d^{n+1} \int_\Omega \psi_\lambda(x)\psi_\mu(x)dx \end{aligned} \tag{3.5}$$

and in matrix form

$$[(1+\delta t)\mathbf{B} - \delta \mathbf{K}]\, d^{n+1} = \mathbf{B} d^n \tag{3.6}$$

where the matrix elements for $\mathbf{B}$ and $\mathbf{K}$ are respectively

$$B_{\lambda\mu} = \int_\Omega \psi_\lambda(x)\psi_\mu(x)dx \tag{3.7}$$

$$K_{\lambda\mu} = \int_\Omega\int_\Omega K(x-y)\psi_\lambda(y)\psi_\mu(x)dxdy \tag{3.8}$$

If the scaling functions (and so wavelets) have compact support, most of elements in (3.7) are equal to zero, whereas the matrix $\mathbf{K}$ is a dense matrix, what involves a long computing time and storage. Anyway, if the distributional kernel verifies that

$$|\partial_x^\alpha \partial_y^\beta K(x,y)| \lesssim \mathrm{dist}(x,y)^{n+\alpha+\beta} \tag{3.9}$$

it is shown that many of the $|K_{\lambda,\mu}|$ in (3.8) are neglected and the matrix $\mathbf{K}$ has a sparse pattern similar to the local operators

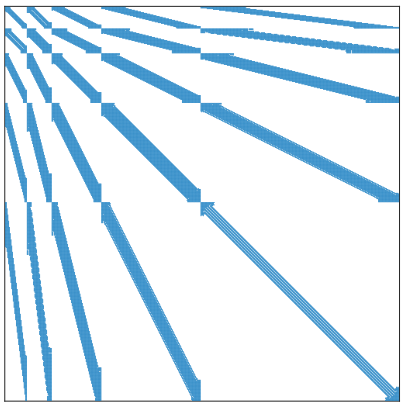


Fig. 1 Sparse pattern for non local operator

This fact, together with (2.10), makes wavelets a powerfull tool to design numerical algorithms to solve integro-differential equations wich come from the nonlocal diffusion models.

4. Epidemiological models

The starting point to describe transmission of diseases are the compartmental models by Kermack and McKendrick, in the final 20's and early 30's. These and later models share the idea of compartmental models, in which the population is distributed in compartments and may change between them under some conditions. In [6,7] a SEIRD (susceptible, exposed, infected, recovery and deceased) model is proposed, where the spatial spread of the infection is included trough a local diffusion term

$$\nabla \cdot (\nu_u \nabla u) \tag{4.1}$$

where u represents the population in any of the compartments before and ν_u is the diffusion coefficient for that population group. For a more realistic description of dispersion of the different groups of populations we can replace (4.1) by its non local version, as in (3.1)

$$\int_\Omega K(x-y)\,(u(y)-u(x))\,dy \tag{4.2}$$

where $K(x-y)$ represents the probability for the population to move from a location y to a location x. If the integral kernel $K(x-y)$ is normalized we can rewrite (4.2) by

$$\int_\Omega K(x-y)u(y)dy - u(x) = K*u - u \tag{4.3}$$

where $K*u$ is a convolution product. So, the equations for the nonlocal SEIRD model are

$$\partial_t S = K*S - S - \beta_i SI - \beta_e SE \tag{4.4}$$
$$\partial_t E = K*E - E + \beta_i SI + \beta_e SE - \sigma E - \phi_e E \tag{4.5}$$
$$\partial_t I = K*I - I + \sigma E - \phi_r I - \phi_d I \tag{4.6}$$
$$\partial_t R = K*R - R + \phi_e E + \phi_r I \tag{4.7}$$
$$\partial_t D = \phi_d I \tag{4.8}$$

where the $\{\beta_e, \beta_i\}$ are the contact rates of susceptible with exposed and infected, σ is the rate of exposed people who develops synthoms, $\{\phi_e, \phi_r\}$ are the rates of recovery for exposed and infected and ϕ_d is the rate of deceased people. The discretization in time leads to

$$\frac{S^{n+1}(x)-S^n(x)}{\delta t} = \int K(x-y)S^{n+1}(y)dy - S^{n+1}(x) - (\beta_i I^n(x) + \beta_e E^n(x))\,S^{n+1}(x) \tag{4.9}$$

$$\frac{E^{n+1}(x)-E^n(x)}{\delta t} = \int K(x-y)E^{n+1}(y)dy - E^{n+1}(x) + (\beta_i I^n(x) + \beta_e E^n(x))\,S^{n+1}(x) - (\sigma+\phi_e)E^{n+1}(x) \tag{4.10}$$

$$\frac{I^{n+1}(x)-I^n(x)}{\delta t} = \int K(x-y)I^{n+1}(y)dy - I^{n+1}(x) + \sigma E^{n+1}(x) - (\phi_r+\phi_d)I^{n+1}(x) \tag{4.11}$$

$$\frac{R^{n+1}(x)-R^n(x)}{\delta t} = \int K(x-y)R^{n+1}(y)dy - R^{n+1}(x) + \phi_e E^{n+1}(x) + \phi_r I^{n+1}(x) \tag{4.12}$$

$$\frac{D^{n+1}(x)-D^n(x)}{\delta t} = \phi_d I^{n+1}(x) \tag{4.13}$$

Let us consider the weak formulation for (4.9)-(4.12) (once we know I we get D from (4.13)) and write the functions $\{S^n, E^n, I^n, R^n\}$ in terms of wavelet basis,

$$S^n(x) = \sum_\lambda s^n_\lambda \psi_\lambda(x), \quad E^n(x) = \sum_\lambda e^n_\lambda \psi_\lambda(x), \tag{4.14}$$
$$I^n(x) = \sum_\lambda i^n_\lambda \psi_\lambda(x), \quad R^n(x) = \sum_\lambda r^n_\lambda \psi_\lambda(x) \tag{4.15}$$

we get

$$[(1+\delta t)\,\mathbf{B} - \delta t\mathbf{K}]\,s^{n+1} = \mathbf{B}s^n - \delta t\mathbf{F} \tag{4.16}$$
$$[(1+\delta t(1+\sigma+\phi_e)\,\mathbf{B} - \delta t\mathbf{K}]\,e^{n+1} = \mathbf{B}e^n + \delta t\mathbf{F} \tag{4.17}$$
$$[(1+\delta t(1+\phi_r+\phi_d)\,\mathbf{B} - \delta t\mathbf{K}]\,i^{n+1} = \mathbf{B}i^n + \delta t\sigma\mathbf{B}e^{n+1} \tag{4.18}$$
$$[(1+\delta t)\,\mathbf{B} - \delta t\mathbf{K}]\,r^{n+1} = \mathbf{B}r^n + \delta t\mathbf{B}(\phi_e e^{n+1} + \phi_r i^{n+1}) \tag{4.19}$$

where **F** in (4.16) and (4.17) is

$$\mathbf{F}_\mu = \int_\Omega p^{n+1}\psi_\mu(x)dx \tag{4.20}$$

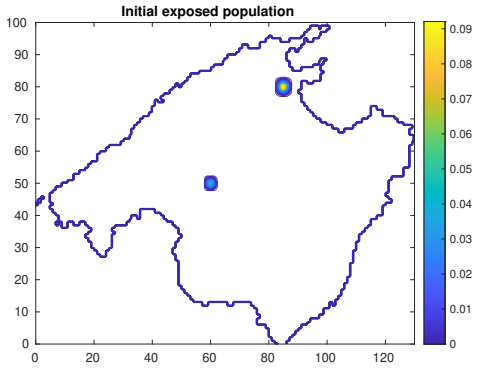
for the coefficients p^{n+1} for the expansion of

$$(\beta_i I^n(x) + \beta_e E^n(x))\, S^{n+1}(x) = \sum_\lambda p^{n+1}\psi_\lambda(x) \tag{4.21}$$

(this last term is treated with an implicit scheme for (4.16) and implicit in (4.17)).

5. Numerical test

To test the numerical algorithms, we have solved the nonlocal SEIRD model in Mallorca Island (Spain), to compare with the local model solved in [4]. We have considered two focus of asynthomatic people concentrated in a short distance, modeled by gaussian functions, as in the figure below,



and the dispersal kernel

$$K(x.y) = e^{-\|x-y\|^2/d^2}, \tag{5.1}$$

being d a kind of mean distance that individuals from different compartments *travel*.

Figures below show the evolution of infected people in different moments, comparing the local SEIRD model with the nonlocal with different values of d; as it is seen, longer is d, faster the epidemic spreads.

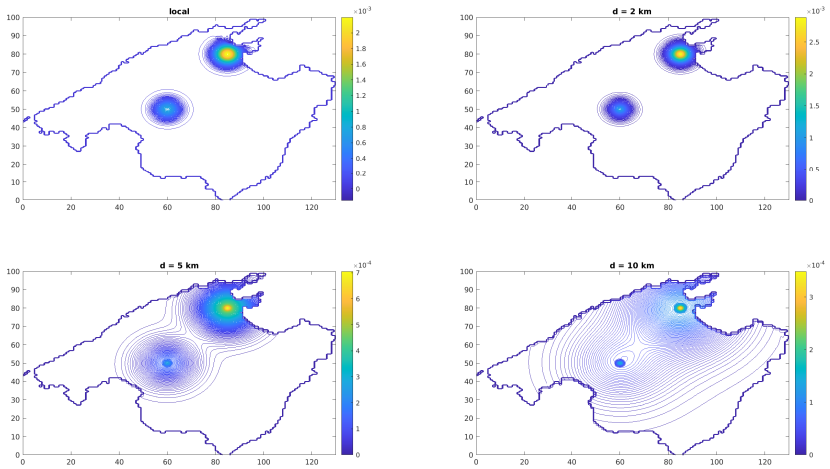


Fig. 2 infected population after 5 days

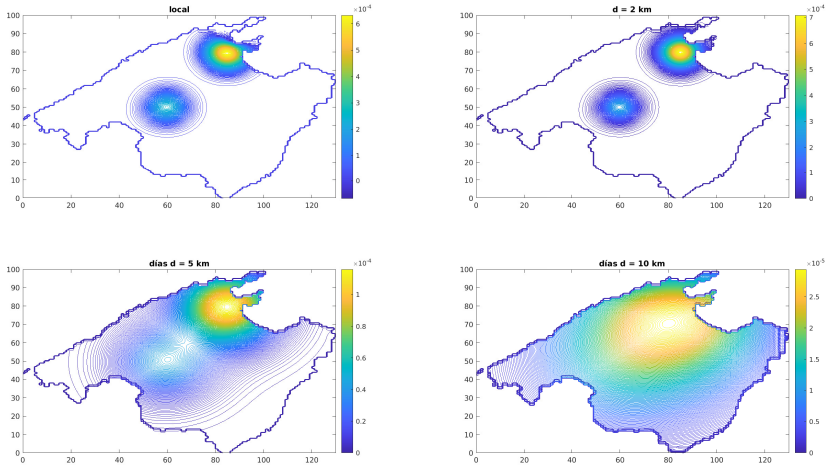


Fig. 3 infected population after 15 days

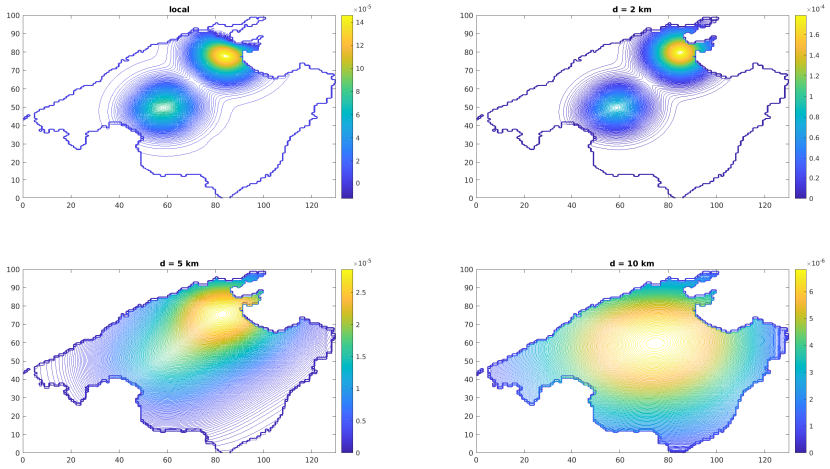


Fig. 4 infected population after 30 days

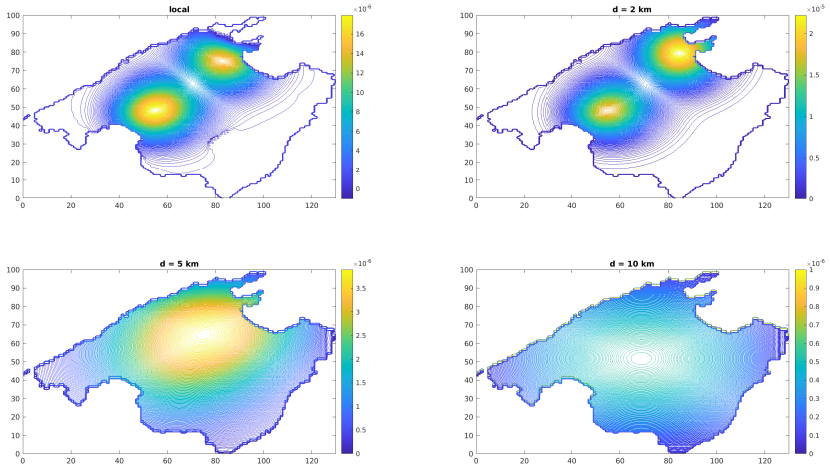


Fig. 5 infected population after 60 days

References

[1] Andreu Vaillo, F., Mazón, J., Rossi, J., Toledo, J. *Nonlocal Diffusion Problems* American Mathematicas Society. Mathematical Surveys and Monographs, VOL 165, 2010.

[2] Cohen, A. *Wavelet Methods in Numerical Analysis*. *Handbook of Numerical Analysis VOL VII*. P. Ciarlet and J. Lions, eds. Elsevier North-Holland, 2000.

[3] Urban, K. *Wavelet Methods for Elliptic Partial Differemtial Equations*. Oxford University Press, 2009

[4] Ferragut, L. Private comunication.

[5] Murray, J.D. *Mathematical biology II: spatial models and biomedical application* 3rd edn. Springer Verlag, Berlin.

[6] Viguerie A. et al. *Simulating the spread of COVID-19 via spatially-resolved susceptible-exposed- infected-recovered-deceased (SEIRD) model with heterogeneous diffusion.* Appl Math Lett 111

[7] Viguerie A. et al. *Diffusion–reaction compartmental models formulated in a continuum mechanics framework: application to COVID-19, mathematical analysis, and numerical study*. Computational Mechanics 66(4):1-22

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 97–104)

Two-sided methods for the nonlinear eigenvalue problem

Carmen Campos[1], **Jose E. Roman**[1]
Universitat Politècnica de València, Spain

Abstract

We discuss solvers for the general nonlinear eigenvalue problem that are able to compute both left and right eigenvectors. A possible application is the approximation of the resolvent of a matrix-valued function. Our focus is on large-scale problems, in the context of SLEPc, the Scalable Library for Eigenvalue Problem Computations. We present two-sided variants of the NLEIGS and SLP methods. For the latter, we have implemented a non-equivalence deflation scheme. The accuracy and performance of the methods are analyzed for several problems coming from real applications.

1. Introduction

Let the matrix-valued function $T : \Omega \to \mathbb{C}^{n\times n}$ be defined in the open set $\Omega \subseteq \mathbb{C}$. The nonlinear eigenvalue problem (NEP) is expressed as

$$T(\lambda)x = 0, \qquad x \neq 0, \tag{1.1}$$

where we are interested in computing eigenvalues $\lambda \in \mathbb{C}$ and (right) eigenvectors $x \in \mathbb{C}^n$ that satisfy (1.1). In the following, we assume that T is holomorphic in Ω and regular, that is, $\det T(z)$ is not identically zero. If T is a matrix polynomial, then the above problem is the polynomial eigenproblem, which includes the standard and generalized eigenvalue problem if the polynomial has degree one. Here we are interested in the non-polynomial case, where T is for instance a rational matrix or, more generally, a matrix whose entries are nonlinear functions of the parameter λ. Example applications where this type of problem appears are the simulation of photonic crystals [3], or the analysis of scattering resonances in metal-dielectric nano-structures [1]. The nonlinear eigenvalue problem has been addressed by many numerical linear algebra researchers in recent times, see the survey [5]. The SLEPc library (Scalable Library for Eigenvalue Problem Computations [7, 10]) implements several methods to solve the NEP, with details provided in our previous work [2].

In this work we focus on the case where, in addition to right eigenvectors x, it is also necessary to compute left eigenvectors y, that satisfy

$$y^*T(\lambda) = 0^*, \qquad y \neq 0, \tag{1.2}$$

where $(\cdot)^*$ denotes the conjugate transpose of a vector or matrix. Left eigenvectors can be obtained as right eigenvectors of T^*, but we are interested in methods that compute both x and y simultaneously, which are referred to as two-sided methods. In particular, we consider a two-sided variant of NLEIGS, already described in [2], and a two-sided version of SLP (Successive Linear Problems). The main contribution of this work is to show how to handle deflation when computing several eigenvalues with two-sided SLP.

The remaining material is organized as follows. Section 2 provides the necessary mathematical background for the NEP. Sections 3 and 4 describe the details of the NLEIGS and SLP methods, respectively. Some computational results are given in section 5, and then section 6 wraps up with some conclusions.

2. Nonlinear eigenvalue problems

The two-sided solution of the NEP consists in eigen-triples $\{(x_i, y_i, \lambda_i)\}_{i=1}^k$ satisfying both (1.1) and (1.2). We are interested mainly in large-scale problems, where the problem size n is large, so we can afford to compute only a relatively small number of eigen-triples. On the other hand, as opposed to the case of linear or polynomial eigenproblems, in the NEP one cannot normally aspire to compute the "full spectrum" as the solution structure of the NEP is more diverse (a NEP may have no solution at all, finitely many solutions, or infinitely many solutions [5]). Practical applications typically need to compute all eigenvalues inside a given region of the complex plane, or a few eigenvalues closest to a given target value.

The function T can always be expressed in *split form*

$$T(z) = \sum_{i=1}^{\ell} A_i f_i(z), \tag{2.1}$$

 ISBN 978-84-18482-21-2

for constant coefficient matrices $A_1, \ldots, A_\ell \in \mathbb{C}^{n\times n}$ and scalar holomorphic functions $f_1, \ldots, f_\ell : \Omega \to \mathbb{C}$, with $\ell \le n^2$. In many applications, T is naturally given in this form, usually with a small number of terms, $\ell \ll n^2$, and this often simplifies the implementation of application codes [2]. In addition, algorithms may sometimes exploit the structure of (2.1).

If the problem is expressed in split form, the accuracy of a computed eigen-triple $(\tilde{x}, \tilde{y}, \tilde{\lambda})$ can be assessed by means of the backward error

$$\eta(\tilde{x}, \tilde{y}, \tilde{\lambda}) = \max\left\{\frac{\|T(\tilde{\lambda})\tilde{x}\|}{f(\tilde{\lambda})\|\tilde{x}\|}, \frac{\|T^*(\tilde{\lambda})\tilde{y}\|}{f(\tilde{\lambda})\|\tilde{y}\|}\right\}, \qquad \text{with} \quad f(\tilde{\lambda}) = \sum_{i=1}^{\ell} |f_i(\tilde{\lambda})|\|A_i\|, \tag{2.2}$$

which involves the scaled residuals for both left and right eigenvectors. If the backward error (2.2) is small then the approximate eigen-triple $(\tilde{x}, \tilde{y}, \tilde{\lambda})$ is the exact eigen-triple of a nearby problem.

2.1. Resolvent

We now mention one relevant type of application where computing left eigenvectors of the NEP is required: it is the case where the (approximate) resolvent T^{-1} needs to be applied to a set of vectors.

Analogously to the Jordan form in linear eigenproblems, for the NEP there exists the Smith form, which is a factorization that reveals the structure of eigenvalues (including partial multiplicities) together with (generalized) right and left eigenvectors. The Smith form can be employed to operate with the resolvent. In particular, Keldysh's theorem provides an explicit formula for the resolvent. In case that all eigenvalues λ_i are semisimple, the resolvent can be written as

$$T^{-1}(z) = \sum_{i=1}^{k} (z - \lambda_i)^{-1} x_i y_i^* + R(z), \tag{2.3}$$

where k is the number of eigenvalues λ_i contained in Ω, x_i and y_i are the corresponding right and left eigenvectors, respectively, normalized so that $y_i^* T'(\lambda_i) x_i = 1$, and $R(z)$ is some function holomorphic in Ω. SLEPc provides functionality to efficiently apply the first part of (2.3) to a vector, once the eigen-triples have been computed with the methods described next.

3. Rational interpolation approach

A possible approach to solve (1.1) is to build a replacement of $T(\cdot)$ via interpolation that can then be linearized. A matrix polynomial is the most obvious candidate, but it is effective only in some cases, e.g., when the interpolation points are sufficiently far away from the singularities of $T(\cdot)$. A better strategy is to solve $R_d(\lambda)x = 0$ as a surrogate of (1.1), where $R_d(\cdot)$ is a rational matrix that interpolates $T(\cdot)$. The NLEIGS method [6] takes this route, with an R_d such that it is easy to build a linearization. In this work, we use the NLEIGS solver that we implemented in SLEPc. The details, including the two-sided version, are already described in [2], but we summarize them here for completeness.

3.1. Overview of NLEIGS

We want to find eigenvalues located inside a compact target set $\Sigma \subset \Omega$, in which $T(\cdot)$ is analytic, by approximating T with a rational matrix that interpolates it at nodes $\sigma_j \in \partial\Sigma$ (the boundary of Σ) and whose poles are chosen from the set of singularities of T, denoted by Ξ. This rational matrix is built as

$$R_d(\lambda) := \sum_{j=0}^{d} b_j(\lambda) D_j, \tag{3.1}$$

using the degree-graded rational Newton basis functions defined by the recurrence

$$b_0(\lambda) = 1, \quad b_j(\lambda) = \frac{\lambda - \sigma_{j-1}}{\beta_j(1 - \lambda/\xi_j)} b_{j-1}(\lambda), \quad j = 1, 2, \ldots \tag{3.2}$$

with nonzero poles at $\xi_j \in \Xi$. The β_j scaling factors are used to guarantee that $\max_{\lambda\in\partial\Sigma} |b_j(\lambda)| = 1$. In the case of pairwise distinct interpolation nodes σ_j, the coefficient matrices D_j of (3.1), called rational divided difference matrices, can be obtained from the interpolation conditions $R_j(\sigma_j) = T(\sigma_j)$, resulting in

$$D_0 = \beta_0 T(\sigma_0), \quad D_j = \frac{T(\sigma_j) - R_{j-1}(\sigma_j)}{b_j(\sigma_j)}, \quad j = 1, 2, \ldots. \tag{3.3}$$

If T is expressed in the split form (2.1), then the simpler form

$$D_j = \sum_{i=0}^{\ell} d_i^j A_i, \quad j \geq 0 \tag{3.4}$$

can be used, where d_i^j denotes the jth rational divided difference corresponding to the scalar function f_i. Additional details of how to compute these scalar divided differences in a numerically stable way can be found in [2]. In SLEPc, the degree d of the interpolant (3.1) is determined at run time, by evaluating the norms of the rational divided difference matrices as they are generated, until the relation $\|D_d\|/\|D_0\| < \texttt{tol}$ holds for a given tolerance.

The interpolation nodes and poles that determine $R_d(\lambda)$ are obtained as a sequence of Leja–Bagby points for (Σ, Ξ), see [6] for details. In SLEPc's implementation, the target set Σ is specified by the user by defining a region, whose boundary $\partial\Sigma$ is discretized automatically, and in the case of the singularity set Ξ, there is the possibility that the user provides a discretization of Ξ, or it can also be computed automatically (e.g., via the AAA method).

After building the rational approximation, a linearization is carried out to obtain a linear eigenvalue problem

$$\mathcal{A}y = \lambda\mathcal{B}y, \tag{3.5}$$

having the same eigenvalues λ as the rational eigenproblem $R_d(\lambda)x = 0$ and whose eigenvectors have the form

$$y = \begin{bmatrix} b_0(\lambda)x \\ \vdots \\ b_{d-1}(\lambda)x \end{bmatrix}. \tag{3.6}$$

The matrices of the linearization (3.5) can be simplified if the last pole is chosen as $\xi_d = \infty$ [6], resulting in

$$\mathcal{A} = \begin{bmatrix} D_0 & D_1 & \dots & D_{d-2} & (D_{d-1} - \frac{\sigma_{d-1}}{\beta_d}D_d) \\ \sigma_0 I & \beta_1 I & & & \\ & \ddots & \ddots & & \\ & & \ddots & \beta_{d-2}I & \\ & & & \sigma_{d-2}I & \beta_{d-1}I \end{bmatrix}, \quad \mathcal{B} = \begin{bmatrix} 0 & 0 & \dots & 0 & -\frac{D_d}{\beta_d} \\ I & \frac{\beta_1}{\xi_1}I & & & \\ & \ddots & \ddots & & \\ & & \ddots & \frac{\beta_{d-2}}{\xi_{d-2}}I & \\ & & & I & \frac{\beta_{d-1}}{\xi_{d-1}}I \end{bmatrix}. \tag{3.7}$$

Once the matrices $\mathcal{A}$, $\mathcal{B}$ of (3.7) have been built, the eigenproblem (3.5) must be solved, and we do this with the shift-and-invert Krylov-Schur method. This implies buiding a Krylov subspace associated with the matrix

$$\mathcal{S} = (\mathcal{A} - \sigma\mathcal{B})^{-1}\mathcal{B}, \tag{3.8}$$

where $\sigma \in \Sigma$ is the target value. To get an efficient algorithm, the action of $\mathcal{S}$ on a vector must be computed implicitly, without explicitly constructing the $d \cdot n$ matrices $\mathcal{A}$, $\mathcal{B}$ nor the inverse. This can be achieved by means of a block LU factorization of $(\mathcal{A} - \sigma\mathcal{B})$, as described next.

Suppose we want to compute the product $w = \mathcal{S}x$, by considering the d blocks of the vectors and deriving recurrence formulas for w^i, which denotes the ith block of vector $w = \text{vec}(w^0, \dots, w^{d-1})$. Consider the factorization $(\mathcal{A} - \sigma\mathcal{B})\Pi = U_\sigma L_\sigma$, where $\Pi = \left[\begin{smallmatrix} 0 & I_{(d-1)n} \\ I_n & 0 \end{smallmatrix}\right]$ is a permutation matrix, and L_σ, U_σ are block triangular matrices (lower and upper, respectively, see the expressions in [2]). Then $w = \mathcal{S}x = \Pi L_\sigma^{-1} U_\sigma^{-1}\mathcal{B}x$ can be computed by first solving the block upper triangular system $U_\sigma y = \mathcal{B}x$, with the recurrence

$$\begin{aligned} y^{d-1} &= \frac{1}{\sigma_{d-2} - \sigma}x^{d-2} + \frac{\beta_{d-1}}{(\sigma_{d-2} - \sigma)\xi_{d-1}}x^{d-1}, \\ y^j &= \frac{1}{\sigma_{j-1} - \sigma}x^{j-1} + \frac{\beta_j}{(\sigma_{j-1} - \sigma)\xi_j}x^j - \frac{\beta_j}{\sigma_{j-1} - \sigma}\left(1 - \frac{\sigma}{\xi_j}\right)y^{j+1}, \quad j = d-2, \dots, 1, \\ y^0 &= R_d(\sigma)^{-1}\left(-D_0y^1 - D_1y^2 - \cdots - D_{d-2}y^{d-1} - \frac{1}{\beta_d}D_dx^{d-1}\right), \end{aligned} \tag{3.9}$$

then the block lower triangular system $L_\sigma\tilde{w} = y$ with

$$\begin{aligned} \tilde{w}^0 &= b_{d-1}(\sigma)y^0, \\ \tilde{w}^j &= y^j + b_{j-1}(\sigma)y^0, \quad j = 1, \dots, d-1, \end{aligned} \tag{3.10}$$

and finally computing the solution vector by applying the permutation $w = \Pi\tilde{w}$. The most computationally expensive operation is building the sparse matrix $R_d(\sigma)$ and solving one linear system of equations with it.

SLEPc provides two different implementations of NLEIGS, one that operates with explicitly stored Krylov vectors (full basis) and another one with a compact representation (TOAR basis). The latter allows a significant reduction of both computational and storage cost [2], but cannot be used for left eigenvectors, so in the next subsection we assume a full basis.

3.2. Two-sided NLEIGS

We have implemented a two-sided variant of NLEIGS that allows computing both right and left eigenvectors, which is related to the two-sided Krylov–Schur of Zwaan and Hochstenbach [13]. The two-sided variant works with two Krylov bases, for the approximate right and left eigenspaces, respectively. The addition with respect to the previous subsection is that left basis vectors are obtained from matrix-vector products with the conjugate transpose, $\mathcal{S}^*$, so it is necessary to derive analogue recurrences that perform this operation by blocks. To obtain $w = \mathcal{S}^*x = \mathcal{B}^*U_\sigma^{-*}L_\sigma^{-*}\Pi^*x$ we proceed as follows. Let $y = L_\sigma^{-*}\Pi^*x$, then $L_\sigma^*y = \Pi^*x$ and we have

$$\begin{cases} y^i = x^{i-1}, \quad i = 1, \dots d-1, \\ y^0 = \bar{b}_{d-1}(\sigma)x^{d-1} + \displaystyle\sum_{i=1}^{d-1} \bar{b}_{i-1}(\sigma)y^i = \bar{b}_{d-1}(\sigma)x^{d-1} + \sum_{i=0}^{d-2} \bar{b}_i(\sigma)x^i, \end{cases} \tag{3.11}$$

where complex conjugation is denoted with a bar. For $z = U_\sigma^{-*}y$ the following relations hold,

$$\begin{aligned} R_d(\sigma)^*z^0 &= y^0, \\ D_0^*z^0 + (\bar{\sigma}_0 - \bar{\sigma})z^1 &= y^1, \\ D_1^*z^0 + \beta_1\left(1 - \frac{\bar{\sigma}}{\bar{\xi}_1}\right)z^1 + (\bar{\sigma}_1 - \bar{\sigma})z^2 &= y^2, \\ &\dots \\ D_{d-2}^*z^0 + \beta_{d-2}\left(1 - \frac{\bar{\sigma}}{\bar{\xi}_{d-2}}\right)z^{d-2} + (\bar{\sigma}_{d-2} - \bar{\sigma})z^{d-1} &= y^{d-1}, \end{aligned} \tag{3.12}$$

from which we obtain the recurrence

$$\begin{cases} z^0 = R_d(\sigma)^{-*}y^0, \\ z^1 = \dfrac{1}{\bar{\sigma}_0 - \bar{\sigma}}\left(y^1 - D_0^*z^0\right), \\ z^i = \dfrac{1}{\bar{\sigma}_{i-1} - \bar{\sigma}}\left(y^i - D_{i-1}^*z^0 - \beta_{i-1}\left(1 - \dfrac{\bar{\sigma}}{\bar{\xi}_{i-1}}\right)z^{i-1}\right), \quad i = 2, \dots, d-1. \end{cases} \tag{3.13}$$

Finally, $w = \mathcal{B}^*z$ is computed with

$$\begin{cases} w^0 = z^1, \\ w^i = \dfrac{\beta_i}{\bar{\xi}_i}z^i + z^{i+1}, \quad i = 1, \dots, d-2, \\ w^{d-1} = -\dfrac{1}{\beta_d}D_d^*z^0 + \dfrac{\beta_{d-1}}{\bar{\xi}_{d-1}}z^{d-1}. \end{cases} \tag{3.14}$$

As before, the most computationally expensive step is the application of $R_d(\sigma)^{-*}$. In case of using a direct method for the linear solves, the factorization needed for (3.9) can be reused here.

4. Newton-type approach

In this section we present the SLP method, which is one of the many Newton-type methods that we can find in the literature. In general, Newton-type methods for the NEP are appropriate whenever one eigenpair is required, but for several eigenpairs the complexity increases significantly due to the need of a deflation mechanism, and hence cannot compete with NLEIGS in terms of computational time. Another issue of these methods is that the initial guess may have a great influence on which eigensolution is found.

ALGORITHM 1: Successive Linear Problems (SLP)

Input: Initial eigenvalue approximation $\lambda^{(0)}$
Output: Computed eigenpair $(x^{(k)}, \lambda^{(k)})$
1 **for** *k=0,1,2,...* **do**
2 Evaluate $A = T(\lambda^{(k)}), \quad B = T'(\lambda^{(k)})$
3 **if** $k > 0$ and $\eta(x^{(k)}, \lambda^{(k)}) <$ `tol` **then** exit
4 Compute $(x^{(k+1)}, \mu^{(k)})$, the smallest magnitude eigenpair of $Ax = \mu Bx$
5 Update $\lambda^{(k+1)} = \lambda^{(k)} - \mu^{(k)}$
6 **end**

4.1. Overview of SLP

The Successive Linear Problems (SLP) method is a Newton-type iteration proposed by Ruhe [11]. It is very simple and susceptible of computing left eigenvectors as well. The method stems from a linearization using a first-order Taylor approximation, $T(\lambda + \mu) \approx T(\lambda) + \mu T'(\lambda)$, which suggests computing the correction μ as the smallest eigenvalue of the pencil $(T(\lambda), T'(\lambda))$, see Algorithm 1.

We implement step 4 of Algorithm 1 as computing the largest magnitude eigenvalue of $Cx = \theta x$, where $\theta = 1/\mu$ and $C = T(\lambda^{(k)})^{-1}T'(\lambda^{(k)})$, via a Krylov-Schur solver in which C is not built explicitly. The main drawback of SLP is that a factorization of $T(\lambda^{(k)})$ is required at each iteration, which is computationally expensive with respect to the rest of operations. In contrast, NLEIGS needs a single factorization throughout.

4.2. Two-sided SLP

In the two-sided variant, we must also compute approximations of left eigenvectors, by modifying step 4 of Algorithm 1. There are two approaches for this:

1. Use a two-sided Krylov-Schur method to compute the eigen-triple $(x^{(k+1)}, y^{(k+1)}, \theta^{(k)})$ corresponding to the smallest magnitude eigenvalue of $Ax = \mu Bx$. Note that when computed via C, left eigenvectors must be post-processed,
$$\tilde{y}^* C = \bar{\theta}\tilde{y}^*, \qquad y = B^{-*}\tilde{y}. \tag{4.1}$$
2. Run two independent Krylov-Schur solves for $Cx = \theta x$ and $Dy = \bar{\theta}y$, where $D = T(\lambda^{(k)})^{-*}T'(\lambda^{(k)})^*$.

We will use the second approach, not only because it avoids the post-processing (4.1), but mainly because the first approach cannot be used with the deflation scheme to be explained next.

When computing more than one eigenvalue, Newton-type solvers need some type of deflation to avoid reconverging to the same solution. In SLEPc [2] we implemented Effenberger deflation [4], that consists in building an extended nonlinear operator by padding $T(\lambda)$ with some rows and columns related to the eigenvalues and eigenvectors that we want to deflate. This formulation has no obvious extension to the case of needing left eigenvectors also. An alternative is the non-equivalence deflation technique [5,8], that aims at mapping the previously computed eigenvalues to infinity. We use this strategy here, since it is suitable for two-sided solvers.

Non-equivalence deflation in SLP Let $\{(x_i, \lambda_i)\}_{i=1}^k$ be the set of computed eigenpairs and let $\{y_i\}_{i=1}^k$ be vectors such that $y_i^* x_i = 1$, $i = 1, \ldots, k$. Then the operator

$$\widetilde{T}(\lambda) := T(\lambda) \prod_{i=1}^{k} \left(I_n - \frac{\lambda - \lambda_i - 1}{\lambda - \lambda_i} y_i x_i^* \right) \tag{4.2}$$

verifies that $\Lambda(\widetilde{T}) = (\Lambda(T) \setminus \{\lambda_i\}_{i=1}^k) \cup \{\infty\}$. Taking vectors Y bi-orthogonal to X (i.e., $Y^*X = I_k$), operator $\widetilde{T}$ can be expressed as

$$\widetilde{T}(\lambda) := T(\lambda)(I_n - YDX^*), \tag{4.3}$$

where $D = \mathrm{diag}(\alpha_i)$ and $\alpha_i = \frac{\lambda - \lambda_i - 1}{\lambda - \lambda_i}$, $i = 1, \ldots, k$. Note that once an eigenvector $\widetilde{x}_{k+1}$ of $\widetilde{T}(\lambda)$ has been computed, the corresponding eigenvector of the original problem $T(\lambda)$ can be recovered as $x_{k+1} = (I_n - YDX^*)\widetilde{x}_{k+1}$.

We take $\{y_i\}_{i=1}^k$ as the set of computed left eigenvectors properly bi-orthogonalized. After k eigen-triples $\{(x_i, y_i, \lambda_i)\}_{i=1}^k$ have been computed, the modified operators we will use with the SLP method are (4.3) and

$$\widetilde{T^*}(\lambda) := T^*(\lambda)(I_n - X\bar{D}Y^*). \tag{4.4}$$

Tab. 1 Summary of the two test problems employed for the computational experiments. The table shows the problem dimension and the region where eigenvalues are sought (used only by the NLEIGS solver). In the tests, *nev* eigenvalues are requested around the target value σ.

name	dim	region	*nev*	σ
gun	9956	see Fig. 1	4	$65000 + 500i$
dimer	1095073	$[-1, 20] \times [-2, 0]$	4	$2.7 - 0.25i$

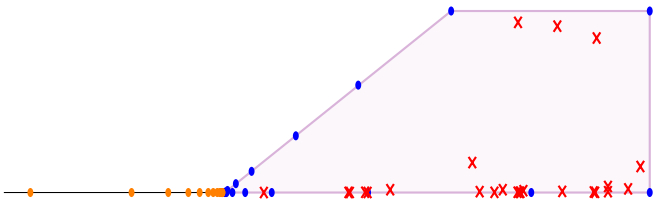


Fig. 1 Graphical representation of the gun problem. The shaded region is the target set Σ. The nodes (blue points) and poles (orange points) used in the NLEIGS solver are Leja-Bagby pairs, picked from the discretization of the boundary $\delta\Sigma$ and the singularity set Ξ, respectively. Red crosses are the eigenvalues lying inside the region.

In the last equation we have taken into account the fact that the columns of Y and X are, respectively, right and left eigenvectors of T^*. Matrix D is the same as in (4.3).

Now we give detailed expressions of the matrices we need in order to carry out the product of $(\widetilde{T}(\sigma))^{-1}\widetilde{T}'(\sigma)$ by a vector. Dropping the σ argument, these expressions become

$$\begin{aligned} \widetilde{T}^{-1} &= (I + YD(I - D)^{-1}X^*)T^{-1} \quad \text{and} \\ \widetilde{T}' &= T'(I_n - YDX^*) + T(-YD'X^*), \end{aligned} \tag{4.5}$$

where $D' = \mathrm{diag}(\alpha_i'(\sigma))$ and $\alpha_i'(\sigma) = \frac{1}{(\sigma-\lambda_i)^2}$. We have similar expressions for the operator $\widetilde{T^*}$:

$$\begin{aligned} (\widetilde{T^*})^{-1} &= (I + X\bar{D}(I - \bar{D})^{-1}Y^*)T^{-*} \quad \text{and} \\ (\widetilde{T^*})' &= (T^*)'(I_n - X\bar{D}Y^*) + T(-X\bar{D}'Y^*). \end{aligned} \tag{4.6}$$

To compute the inverses in (4.5) and (4.6) we have used the Woodbury matrix identity

$$(A + UV)^{-1} = A^{-1} - A^{-1}U(I + VA^{-1}U)^{-1}VA^{-1}, \tag{4.7}$$

particularized for the case $A = I$.

5. Computational results

We have carried out some computational tests to measure the accuracy and performance of the two-sided NLEIGS and SLP solvers. The runs are performed on Tirant III, a computer consisting of 336 computing nodes, each of them with two Intel Xeon SandyBridge E5-2670 processors (16 cores each) running at 2.6 GHz with 32 GB of memory, linked with an Infiniband network. We allocated 4 MPI processes per node at most. The results in this section correspond to SLEPc version 3.15, together with PETSc 3.15 and MUMPS 5.3.

All methods in SLEPc are implemented in parallel with a message-passing paradigm (MPI), so here we also include results on parallel performance. The details about parallelization are discussed in [2].

We have used two test problems for the computational experiments, which are summarized in Table 1. Both problems are expressed in the split form (2.1) and either the problem matrices or the f_i functions are complex, and so are the eigenvalues. The solvers have been run using complex arithmetic with double precision. The tolerance for the convergence criterion has been set to 10^{-8}. A more detailed description of the problems follows:

Fig. 2 Graphical representation of the dimer problem. The orange points are the 12 poles of the rational function. The shaded part is the region used in the NLEIGS solver, while the blue circle is the target value used in both NLEIGS and SLP. Red crosses are the 20 eigenvalues computed by NLEIGS (there are more eigenvalues inside the region).

Tab. 2 Results for both algorithms when solving the two test problems with 16 MPI processes. The table shows the number of iterations, the total number of required linear solves, the time used by the factorizations (PC), the total execution time, and the backward error.

test problem	solver	iter	lin. solves	PC	time	$\eta(x, y, \lambda)$
gun	NLEIGS	4	98	0.4	1.1	7×10^{-13}
	SLP	7	224	1.5	2.4	4×10^{-9}
dimer	NLEIGS	2	122	14.2	109.1	3×10^{-14}
	SLP	20	656	66.2	231.5	2×10^{-11}

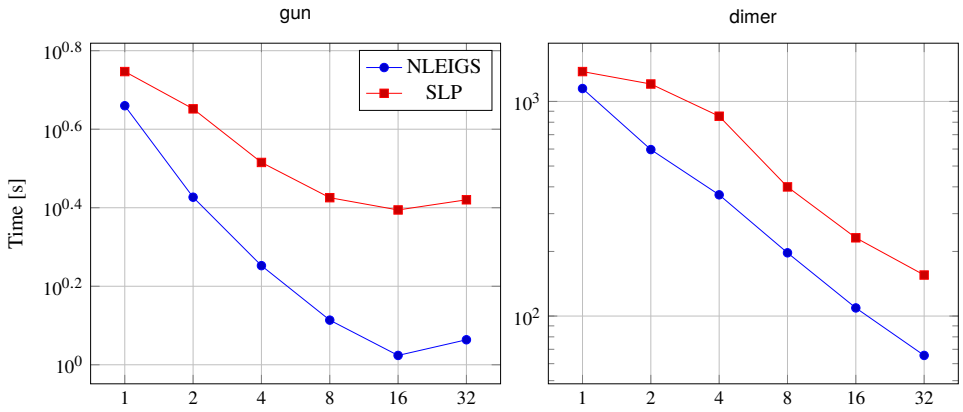


Fig. 3 Execution time (in seconds) with up to 32 nodes (128 MPI processes) for the problems gun (left) and dimer (right), solved with two-sided NLEIGS and two-sided SLP. The execution parameters are shown in Table 1.

- The gun problem models a radio-frequency gun cavity [9]. Here, $T(\lambda) = K - \lambda M + i\sqrt{\lambda - \kappa_1^2}W_1 + i\sqrt{\lambda - \kappa_2^2}W_2$, where K, M, W_1 and W_2 are real symmetric matrices and we set $\kappa_1 = 0$ and $\kappa_2 = 108.8774$. The nonlinearity comes from the square roots, and in this case $\Xi = (-\infty, \kappa_2^2)$. Figure 1 shows a picture of this problem.

- The dimer problem [1] corresponds to the analysis of scattering resonances (TM polarization) of a dimer nanostructure (two disks coated with gold). The partial differential equation is written as $\Delta u + \varepsilon(x, \omega)\omega^2 u = 0$, where ω is the complex frequency (eigenvalue). The relative permittivity $\varepsilon(x, \omega)$ is modeled as a rational function (Drude-Lorentz), so the nonlinear operator $T(\omega) = A_0 - \omega^2 A_1 - \omega^2 \varepsilon(x, \omega) A_2$ is rational. In the case of gold, the rational representation of ε has 12 poles, see Figure 2.

The accuracy of the computed eigen-triples is assessed with the backward error (2.2), using ∞-norms for practical computation of matrix norms. Table 2 shows the maximum value of this error estimate for all the eigen-triples computed by two-sided NLEIGS and two-sided SLP for the problems of Table 1. From the results, we can note the following:

- NLEIGS may return more eigenvalues than requested. This depends on the size of the Krylov subspace, which by default is $\max\{2 \cdot nev, nev + 15\}$. The number of NLEIGS iterations in Table 2 refers to *restarts*.

- All eigen-triples returned by both solvers have an accuracy below the requested tolerance. In the case of NLEIGS, the accuracy is significantly better, which can be attributed to being a subspace method. In SLP, deflation of the first eigenvalues does not seem to introduce large errors for subsequent ones.

- SLP is significantly slower than NLEIGS, taking more that twice its time in these examples. As pointed out previously, the reason is that it requires computing many factorizations while NLEIGS just computes one. We can also see that the number of linear solves (the number of times $T(\sigma)^{-1}$ or $T(\sigma)^{-*}$ is applied to a vector) is much larger in SLP.

We conclude this section by analyzing the parallel scalability of the two solvers. Figure 3 shows the execution time of both two-sided NLEIGS and two-sided SLP when solving the two problems with increasing number of nodes (with 4 MPI processes per node). The plots for the dimer problem show very good scaling up to 128 processes, while for the gun problem, the trend is reasonably good, but the problem size is too small for maintaining good parallel efficiency after a certain number of processes.

6. Concluding remarks

We have presented the details for implementing a two-sided SLP method for computing a few eigen-triples of large-scale nonlinear eigenvalue problems, and we have compared it with a two-sided variant of the NLEIGS eigensolver in terms of efficiency and accuracy. According to the computational experiments of section 5, SLP cannot compete with NLEIGS in any aspect. NLEIGS is faster and returns more accurate eigen-triples. Furthermore, NLEIGS is more reliable with respect to retrieving eigenvalues that are closest to the target, while SLP might converge erratically to other eigenvalues that are not the closest ones.

The upside of two-sided SLP is its simplicity, which makes it appealing for easy implementation. Also, if the number of wanted eigenvalues is small, it could require a computational effort similar to NLEIGS in some problems. On the other hand, the SLP method solves the original problem while NLEIGS replaces it with an approximation, which is not guaranteed to be good in all cases.

As a future work, we will explore the feasibility of deriving a two-sided variant of Nonlinear Arnoldi [12] via an oblique projection using subspaces containing approximations of left and right eigenvectors.

Acknowledgements

This work was partially funded by the Spanish Agencia Estatal de Investigación under grant PID2019-107379RB-I00. The computational experiments of section 5 were carried out on the supercomputer Tirant III belonging to Universitat de València.

References

[1] J. C. Araujo, C. Campos, C. Engström, and J. E. Roman. Computation of scattering resonances in absorptive and dispersive media with applications to metal-dielectric nano-structures. *J. Comput. Phys.*, 407:109220, 2020.

[2] C. Campos and J. E. Roman. NEP: a module for the parallel solution of nonlinear eigenvalue problems in SLEPc. *ACM Trans. Math. Software*, (to appear), 2021. Preprint available arXiv:1910.11712.

[3] G. Demésy, A. Nicolet, B. Gralak, C. Geuzaine, C. Campos, and J. E. Roman. Non-linear eigenvalue problems with GetDP and SLEPc: Eigenmode computations of frequency-dispersive photonic open structures. *Comput. Phys. Commun.*, 257:107509, 2020.

[4] C. Effenberger. Robust successive computation of eigenpairs for nonlinear eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, 34(3):1231–1256, 2013.

[5] S. Güttel and F. Tisseur. The nonlinear eigenvalue problem. *Acta Numerica*, 26:1–94, 2017.

[6] S. Güttel, R. van Beeumen, K. Meerbergen, and W. Michiels. NLEIGS: A class of fully rational Krylov methods for nonlinear eigenvalue problems. *SIAM J. Sci. Comput.*, 36(6):A2842–A2864, 2014.

[7] V. Hernandez, J. E. Roman, and V. Vidal. SLEPc: A scalable and flexible toolkit for the solution of eigenvalue problems. *ACM Trans. Math. Software*, 31(3):351–362, 2005.

[8] T.-M. Huang, W.-W. Lin, and V. Mehrmann. A Newton-type method with nonequivalence deflation for nonlinear eigenvalue problems arising in photonic crystal modeling. *SIAM J. Sci. Comput.*, 38(2):B191–B218, 2016.

[9] B.-S. Liao, Z. Bai, L.-Q. Lee, and K. Ko. Nonlinear Rayleigh-Ritz iterative method for solving large scale nonlinear eigenvalue problems. *Taiwan. J. Math.*, 14(3A):869–883, 2010.

[10] J. E. Roman, C. Campos, L. Dalcin, E. Romero, and A. Tomas. SLEPc users manual. Technical Report DSIC-II/24/02–Revision 3.15, D. Sistemes Informàtics i Computació, Universitat Politècnica de València, 2021.

[11] A. Ruhe. Algorithms for the nonlinear eigenvalue problem. *SIAM J. Numer. Anal.*, 10(4):674–689, 1973.

[12] H. Voss. An Arnoldi method for nonlinear eigenvalue problems. *BIT*, 44(2):387–401, 2004.

[13] I. N. Zwaan and M. E. Hochstenbach. Krylov–Schur-type restarts for the two-sided Arnoldi method. *SIAM J. Matrix Anal. Appl.*, 38(2):297–321, 2017.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 105–108)

Fractionary iterative methods for solving nonlinear problems

Giro Candelario[1], Alicia Cordero[2], Juan R. Torregrosa[2], María P. Vassileva[1]
1. Instituto Tecnológico de Santo Domingo, Dominican Republic
2. Universitat Politècnica de València, Spain

Abstract

In recent years, some point-to-point fractional Newton-type methods have been proposed to find roots of nonlinear equations using fractional derivatives. We present several Newton-type methods based on Caputo fractional derivative. For each case, the order of convergence of the proposed methods is stated, and some numerical tests are carried out in order to observe their performance, in practice. Convergence to different roots depending on the order of the derivative is observed and also differences among the methods in terms of the percentage of converging starting points.

1. Introduction

Leibnitz and L'Höpital created the concept of the semi-derivative at 1695, giving birth to fractional calculus. Also Riemann, Liouville and Euler were interested in this idea. From then, fractional calculus has evolved from theoretical aspects to the applications in many real world problems (see [2, 5, 8, 10, 12]): medicine, mechanical engineering, economics, ... In numerical analysis, we are focused in the area of research of iterative methods for solving nonlinear equations $f(x) = 0$. A large amount of these procedures are Newton-like, that is, they involve in their iterative expressions the evaluation of the nonlinear function f and its first derivative f' at each iterate. In this context, we question ourselves how would affect to the convergence order p of these schemes the replacement of integer derivatives by fractional ones. In particular, we introduce the Caputo fractional derivative, and study the convergence of these fractional methods. We would like to answer this and other questions for both point-to-point and multipoint schemes.

For the sake of completeness, we introduce in what follows some concepts about fractional derivatives and the series developments necessary to prove the convergence results.

1.1. Preliminary concepts

Now, we introduce some general concepts such as the Caputo fractional derivative [10,11] and the fractional Taylor series [4,9].

The first concept that we define is the Gamma function, as:

$$\Gamma(x) = \int_0^{+\infty} u^{x-1} e^{-u} du,$$

whenever $x > 0$. This function is a generalization of the factorial function to the complex plane, taking into account that $\Gamma(1) = 1$ and $\Gamma(n+1) = n!$, when $n \in \mathbb{N}$. As we will see in the following section, it appears in the iterative expressions of fractional iterative methods, being necessary for reaching the order of convergence of the iterative scheme.

Definition 1.1 (Caputo fractional derivative of order α) Let $f : I \subseteq \mathbb{R} \to \mathbb{R}$ be an element of $C^{+\infty}([a,x])$ $(-\infty < a < x < +\infty)$, with $\alpha \geq 0$ and $n = [\alpha] + 1$, being $[\alpha]$ the integer part of α. Then, the Caputo fractional derivative of order α of $f(x)$ is defined as follows:

$$(cD_a^{\alpha} f)(x) = \begin{cases} \dfrac{1}{\Gamma(n-a)} \displaystyle\int_a^x \frac{d^n f(t)}{dt^n} \frac{dt}{(x-t)^{\alpha-n+1}}, & \alpha \notin \mathbb{N}, \\ \dfrac{d^{n-1} f(x)}{dx^{n-1}}, & \alpha = n-1 \in \mathbb{N} \cup \{0\}. \end{cases} \tag{1.1}$$

Moreover, to prove the order of convergence of the iterative fractional methods, a generalization of the classical Taylor series expansion of $f(x)$ around the zero of the nonlinear function, $\bar{x}$, is needed. Further on, this development also uses the Caputo fractional derivatives, see [9] (with $\rho = 1$).

Theorem 1.2 (*Taylor series expansion by using Caputo fractional derivatives* [9]) *Let us suppose that* $cD_a^{j\alpha} f(x) \in C([a,b])$, *for* $j = 1, 2, \ldots, n+1$, *where* $\alpha \in (0,1]$, *then we have*

$$f(x) = \sum_{i=0}^{n} cD_a^{i\alpha} f(a) \frac{(x-a)^{i\alpha}}{\Gamma(i\alpha+1)} + cD_a^{(n+1)\alpha} f(\xi) \frac{(x-a)^{(n+1)\alpha}}{\Gamma((n+1)\alpha+1)}, \tag{1.2}$$

with $a \leq \xi \leq x$, *for all* $x \in (a,b]$ *where* $cD_a^{n\alpha} = cD_a^{\alpha} \cdot cD_a^{\alpha} \cdots cD_a^{\alpha}$ (*n times composition*).

We develop the work in the following order. In Section 2 we state our iterative methods and expose their convergence results. In the next section, numerical tests are performed, paying special attention to the convergence rates and the roots the methods converge to. Finally, expose the conclusions obtained and some open questions.

2. Iterative methods designed by using fractional derivatives

In this section, we introduce high-order one-point and multi-point fractional iterative methods based on the methods of Newton and Traub methods, stating the conditions that must be assued in order to achieve their order of convergence, which depend on the order of the fractional derivative.

Theorem 2.1 ([1]) *Let* $f : D \subset \mathbb{R} \to \mathbb{R}$ *be a continuous function whose fractional derivatives of order* $k\alpha$ *are defined for any positive integer* k *and any* α, $0 < \alpha < 1$, *on the interval* D *containing the zero* $\bar{x}$ *of* $f(x)$ *and let the fractional derivatives of Caputo type,* $cD_a^{\alpha} f(x)$, *be continuous and non-singular at* $\bar{x}$. *Also, let us suppose that* x_0 *is an initial approximation close enough to* $\bar{x}$. *Then the order of local convergence of Newton's fractional method*

$$x_{k+1} = x_k - \Gamma(\alpha+1) \frac{f(x_k)}{cD_a^{\alpha} f(x_k)}, \ k = 0, 1, \ldots, \tag{2.1}$$

of Caputo type is at least 2α, *where* $0 < \alpha \leq 1$, *with the error equation*

$$e_{k+1} = \frac{\Gamma(2\alpha+1) - (\Gamma(\alpha+1))^2}{(\Gamma(\alpha+1))^3} C_2 e_k^{2\alpha} + O[e_k^{3\alpha}).$$

We denote the iterative method (2.1) as CFN1. However, another kind of fractional iterative method can be designed, fixing the order of convergence to be at least one, as can be seen in the following result.

Theorem 2.2 ([3]) *Let* $f : D \subset \mathbb{R} \to \mathbb{R}$ *be a continuous function with fractional derivatives of order* $k\alpha$ *defined for any positive integer* k *and* $\alpha \in (0,1]$ *defined on the open interval* D *containing the zero* $\bar{x}$ *of* $f(x)$. *Additionally, let us suppose that* $cD_a^{\alpha} f(x)$ *is continuous and not zero at* $\bar{x}$. *Then, the order of convergence of the Caputo type fractional Newton method with iterative scheme*

$$x_{k+1} = x_k - \left(\Gamma(\alpha+1) \frac{f(x_k)}{cD_a^{\alpha} f(x_k)} \right)^{1/\alpha}, \ k = 0, 1, 2, \ldots \tag{2.2}$$

(*denoted by CFN*2) *is at least* $\alpha + 1$, *and its error equation is*

$$e_{k+1} = \frac{\Gamma(2\alpha+1) - (\Gamma(\alpha+1))^2}{\alpha(\Gamma(\alpha+1))^2} C_2 e_k^{\alpha+1} + O[e_k^{2\alpha+1}].$$

On the other hand, higher-order iterative schemes can be designed, following this structure. A Traub-type fractional-order method can be defined also by means of Caputo derivatives. In the following result, the convergence conditions and its fractional order of convergence are stated.

Theorem 2.3 ([3]) *Let* $f : D \subset \mathbb{R} \to \mathbb{R}$ *be a the continuous function with fractional derivatives of order* $k\alpha$, *for any positive integer* k *and* $\alpha \in (0,1]$, *in the open interval* D *holding the zero of* $f(x)$, *denoted by* $\bar{x}$. *Let us suppose* $cD_a^{\alpha} f(x)$ *is continuous and not null at* $\bar{x}$. *Additionally, let us consider an initial estimation* x_0, *close enough to* $\bar{x}$. *Therefore, the local convergence order of CFT method with iterative expression*

$$x_{k+1} = y_k - \left(\Gamma(\alpha+1) \frac{f(y_k)}{cD_a^{\alpha} f(x_k)} \right)^{1/\alpha}, \ k = 0, 1, \ldots, \tag{2.3}$$

where y_k *is obtained using* (2.2), *being* $\alpha^2 + 2\alpha + 1 < 3\alpha + 1$, *is at least* $2\alpha + 1$. *Its error equation is*

$$e_{k+1} = -\frac{\Gamma(2\alpha+1)}{\alpha^2(\Gamma(\alpha+1))^2} \frac{(\Gamma(\alpha+1))^2 - \Gamma(2\alpha+1)}{(\Gamma(\alpha+1))^2} C_2^2 e_k^{2\alpha+1} + O[e_k^{\alpha^2+2\alpha+1}].$$

3. Numerical results

In the following section, the numerical performance of these schemes is tested. We are going to test a nonlinear function in order to make a comparison between the presented methods. It is important to point out that in any case a comparison is being made with the classical methods (when $\alpha = 1$).

To get these results, we have used Matlab R2018b with double precision arithmetics, $|x_{k+1} - x_k| < 10^{-8}$ or $|f(x_{k+1})| < 10^{-8}$ as stopping criteria, and a maximum of 500 iterations. For the calculation of the gamma function, $\Gamma(x)$, we use the program presented in [6], where gamma function is calculated with 15 digits of accuracy along the real axis and 13 elsewhere in $\mathbb{C}$. Moreover, in all the numerical tests, we used $a = 0$.

Our test function is $f(x) = -12.84x^6 - 25.6x^5 + 16.55x^4 - 2.21x^3 + 26.71x^2 - 4.29x - 15.21$ with roots $\bar{x}_1 = 0.82366 + 0.24769i$, $\bar{x}_2 = 0.82366 - 0.24769i$, $\bar{x}_3 = -2.62297$, $\bar{x}_4 = -0.584$, $\bar{x}_5 = -0.21705 + 0.99911i$ and $\bar{x}_6 = -0.21705 - 0.99911i$.

We observe that Newton-type methods (Table 1) with Caputo derivative, for the same value of x_0 and the same values of α, converge to the different roots in more iterations than fractional Traub's methods. It also can be observed that Newton- and Traub-type schemes require approximately the same values of α to converge. Also, it has been observed in practice that, for wide ranges of initial guesses, the same x_0 defines a sequence converging to different roots of the nonlinear function depending on the value of α.

	CFN1 method				CFN2 method			
α	$\bar{x}$	$\lvert x_{k+1} - x_k\rvert$	$\lvert f(x_{k+1})\rvert$	iter	$\bar{x}$	$\lvert x_{k+1} - x_k\rvert$	$\lvert f(x_{k+1})\rvert$	iter
0.6	-	0.29821	28.343	500	-	1.7603e-07	0.0035619	500
0.65	-	0.17488	11.329	500	-	4.1154e-08	6.7515e-04	500
0.7	-	0.058499	2.98929	500	$\bar{x}_4$	9.9926e-09	1.1322e-04	432
0.75	$\bar{x}_4$	9.6537e-09	4.1645e-07	151	$\bar{x}_4$	9.8524e-09	4.6756e-05	230
0.8	$\bar{x}_4$	8.5475e-09	3.0465e-07	50	$\bar{x}_4$	9.6579e-09	1.8943e-05	124
0.85	$\bar{x}_4$	9.468e-09	2.606e-07	28	$\bar{x}_4$	9.9396e-09	7.7541e-06	67
0.9	$\bar{x}_4$	3.9203e-09	7.3851e-08	19	$\bar{x}_4$	9.109e-09	2.6706e-06	37
0.95	$\bar{x}_4$	2.5822e-09	2.4894e-08	13	$\bar{x}_4$	7.3622e-09	6.4461e-07	20
1	$\bar{x}_4$	3.0876e-06	8.8694e-10	6	$\bar{x}_4$	3.0876e-06	8.8694e-10	6

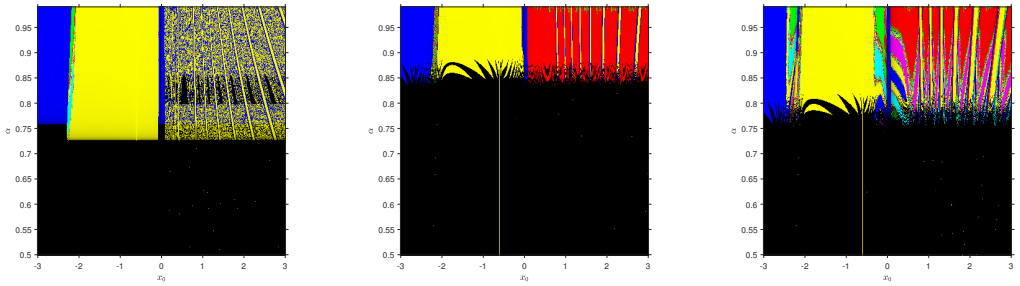
Tab. 1 Fractional Newton results for $f(x)$ with Caputo derivative and initial estimation $x_0 = -1.5$

	CFT method			
α	$\bar{x}$	$\lvert x_{k+1} - x_k\rvert$	$\lvert f(x_{k+1})\rvert$	iter
0.6	-	6.2898e-08	0.0012681	500
0.65	-	1.1562e-08	1.8867e-04	500
0.7	$\bar{x}_4$	9.9588e-09	6.9453e-05	268
0.75	$\bar{x}_4$	9.9889e-09	2.7995e-05	138
0.8	$\bar{x}_4$	9.5606e-09	1.0693e-05	73
0.85	$\bar{x}_4$	9.4657e-09	4.0225e-06	39
0.9	$\bar{x}_4$	6.8084e-09	1.0286e-06	22
0.95	$\bar{x}_4$	5.2078e-09	1.8928e-07	12
1	$\bar{x}_4$	2.2023e-10	5.329e-15	5

Tab. 2 Fractional Traub results for $f(x)$ with Caputo derivative and initial estimation $x_0 = -1.5$

Now, we are going to analyze the dependence on the initial estimation of Newton- and Traub-type methods by using convergence planes defined in [7]. In them (see, for example, Figure 1a) the abscissa axis corresponds to the starting guess and the fractional index α appears in the ordinate axis. A mesh of 400×400 points is used. Points that are not painted in black color correspond to those pairs of initial estimations and values of α that converge to one of the roots with a tolerance of 10^{-3}. Different colors mean convergence to different roots. Therefore, when a point is painted in black, this shows that no root is found in a maximum of 500 iterations. Moreover, for all convergence planes, the percentage of convergent pairs (x_0, α) is calculated, in order to compare the performance of the methods.

In Figure 1, we can see that CFN1 and CFT methods have higher percentage of convergence than CFN2. We can also see, that there are intervals for x_0 such that the same fractional iterative method with different values of the order of the fractional derivative can lead us to converge to different roots. It can be useful in order to find all the roots of a function with few computational effort.



(a) CFN1, 49.54% convergence

(b) CFN2, 27.15% convergenc

(c) CFT, 41.5% convergence

Fig. 1 Convergence planes of proposed methods on $f(x)$ with $-3 \leq x_0 \leq 3$

4. Concluding remarks

Fractional Newton- and Traub-type schemes have been designed by using Caputo derivatives. The convergence properties of these procedures imply always (at least) linear convergence, reaching order 2α, $1+\alpha$ and $1+2\alpha$, respectively. Some numerical tests have been done, and the dependence on the initial estimation has been observed.

It can be concluded that Traub-type procedures can improve Newton-type ones, not only because they require fewer iterations, higher or similar percentages of convergence. Moreover, the test made have shown that, for some problems, the methods using fractional derivatives reach different solutions with the same initial estimations.

Acknowledgements

This research was partially supported by Ministerio de Ciencia, Innovación y Universidades PGC2018-095896-B-C22 (MCIU/AEI/FEDER, UE).

References

[1] Akgül, A., Cordero, A., Torregrosa, J.R., A fractional Newton method with 2αth-order of convergence and its stability. Appl. Math. Lett. 2019, 98, 344–351.

[2] Atanackovic, T.M.; Pilipovic, S.; Stankovic, B.; Zorica, D. Fractional Calculus with Applications in Mechanics: Wave Propagation, Impact and Variational Principles; Wiley: London, UK, 2014.

[3] Candelario, G., Cordero A., Torregrosa JR. Multipoint Fractional Iterative Methods with $(2\alpha+1)$th-Order of Convergence for Solving Nonlinear Problems. Mathematics 2020;8. Url: https://www.mdpi.com/2227-7390/8/3/452/htm.

[4] Jumarie, G., Modified Riemann-Liouville Derivative and Fractional Taylor Series of Nondifferentiable Functions Further Results, Computers and Mathematics with Applications 51 (2006) 1367-1376.

[5] Khan, M.A.; Ullah, S.; Farhan, M. The dynamics of Zika virus with Caputo fractional derivative. AIMS Math. 2019, 4, 134–146.

[6] Lanczos, C. A precision approximation of the gamma function. *SIAM J. Numer. Anal*. **1964**, *1*, 86–96.

[7] Magreñán, Á.A. A new tool to study real dynamics: The convergence plane. *Appl. Math. Comput.* **2014**, *248*, 215–224.

[8] Mathai, A.M.; Haubold, H.J. Fractional and multivariable calculus, model building and optimization problems, In *Springer Optimization and Its Applications*, Springer: Berlin, Germany, 2017, Volume 122.

[9] Benjemaa, M. Taylor's formula involving generalized fractional derivatives. *Appl. Math. Comput.* Volume, **2018**, *335*, 182–195.

[10] Oldham, K. B., Spanier, J., The Fractional Calculus, Academic Press, California, 1974.

[11] Podlubny, I., Fractional differential equations, volume 198 of Mathematics in Science and Engineering. Academic Press Inc., San Diego, CA, 1999.

[12] Ross, B. A brief history and exposition of the fundamental theory of fractional calculus. In *Fractional Calculus and Its Applications*; Ross, B., Ed.; Lecture Notes in Mathematics; Springer: Berlin/Heidelberg, Germany, 1975; Volume 457, pp. 1–36.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 109–113)

Well posedness and numerical solution of kinetic models for angiogenesis

Ana Carpio [1], **Elena Cebrián** [2], **Gema Duro** [3]
1. ana_carpio@mat.ucm.es *Universidad Complutense de Madrid, Spain*
2. elenac@ubu.es *Universidad de Burgos, Spain*
3. gema.duro@uam.es *Universidad Autónoma de Madrid, Spain*

Abstract

Angiogenesis processes including the effect of stochastic branching and spread of blood vessels can be described coupling a (nonlocal in time) integrodifferential kinetic equation of Fokker-Planck type with a diffusion equation for the angiogenic factor. Well posedness studies underline the importance of preserving positivity when constructing approximate solutions. We devise order one positivity preserving schemes for a reduced model and show that soliton-like asymptotic solutions are correctly captured. We also find good agreement with the original stochastic model from which the deterministic kinetic equations are derived working with ensemble averages. Higher order positivity preserving schemes can be devised combining WENO and SSP procedures.

1. Angiogenesis model

Angiogenesis (growth of blood vessels) is fundamental for tissue repair and development. A host of immune, inflammatory and malignant diseases are triggered by angiogenic disorders. We study here a deterministic integrodifferential model for the development of the stochastic vessel network.

Denoting by p and C the density of blood vessel tips and the concentration of angiogenic factor released by hypoxic cells, their time evolution is governed by a system of nondimensional equations [1]:

$$\begin{aligned}\frac{\partial}{\partial t}p(\mathbf{x},\mathbf{v},t) &= \alpha(C(\mathbf{x},t))\delta_{\sigma_v}(\mathbf{v}-\mathbf{v}^0)p(\mathbf{x},\mathbf{v},t) - \Gamma p(\mathbf{x},\mathbf{v},t)\int_0^t ds\int d\mathbf{v}' p(\mathbf{x},\mathbf{v}',s)\\ &\quad -\mathbf{v}\cdot\nabla_{\mathbf{x}}p(\mathbf{x},\mathbf{v},t) + \beta \mathrm{div}_{\mathbf{v}}(\mathbf{v}p(\mathbf{x},\mathbf{v},t)) +\\ &\quad -\mathrm{div}_{\mathbf{v}}\left[\beta\mathbf{F}\left(C(\mathbf{x},t)\right)p(\mathbf{x},\mathbf{v},t)\right] + \frac{\beta}{2}\Delta_{\mathbf{v}}p(\mathbf{x},\mathbf{v},t), \qquad (1.1)\\ \frac{\partial}{\partial t}C(\mathbf{x},t) &= \kappa\Delta_{\mathbf{x}}C(\mathbf{x},t) - \chi C(\mathbf{x},t)j(\mathbf{x},t), \qquad (1.2)\\ p(\mathbf{x},\mathbf{v},0) &= p_0(\mathbf{x},\mathbf{v}), \quad C(\mathbf{x},0) = C_0(\mathbf{x}), \qquad (1.3)\end{aligned}$$

where

$$\alpha(C(\mathbf{x},t)) = A\frac{C(\mathbf{x},t)}{1+C(\mathbf{x},t)}, \quad \mathbf{F}(C(\mathbf{x},t)) = \frac{\delta_1}{(1+\Gamma_1 C(\mathbf{x},t))^{q_1}}\nabla_{\mathbf{x}}C(\mathbf{x},t), \qquad (1.4)$$

$$j(\mathbf{x},t) = \int_{\mathbb{R}^N}\frac{|\mathbf{v}|}{1+e^{(|\mathbf{v}-\mathbf{v}^0|^2-\eta)/\epsilon}}p(\mathbf{x},\mathbf{v},t)\,d\mathbf{v}, \quad \rho(\mathbf{x},t) = \int_{\mathbb{R}^N}p(\mathbf{x},\mathbf{v},t)\,d\mathbf{v}, \qquad (1.5)$$

for $\mathbf{x}\in\Omega\subset\mathbb{R}^N$, $\mathbf{v}\in\mathbb{R}^N$, $N=2,3$, $t\in[0,\infty)$. The parameters $\beta,\Gamma,\kappa,\chi,A,\Gamma_1,\delta_1,\eta,\epsilon$ and q_1 are dimensionless and positive. Typical values are listed in Table 1. Here, δ_{σ_v} is a regularized delta function, such as

$$\delta_{\sigma_v}(\mathbf{v}-\mathbf{v}_0) = \frac{1}{\pi\sigma_v^2}e^{-|\mathbf{v}-\mathbf{v}_0|^2/\sigma_v^2}. \qquad (1.6)$$

In dimension two, these models can be adapted to describe angiogenesis problems causing retinopathies [2].

δ_1	β	A	Γ	Γ_1, q_1	κ	χ	η	ϵ	σ_v
0.255	5.88	22.42	0.135	1	0.0045	0.002	15	0.001	0.08

Tab. 1 Dimensionless parameters.

Existence, uniqueness and stability of positive solutions is established in the whole space resorting to iterative procedures. The key idea for an existence theory is to include the integrodifferential terms in the reference linear operators [3]: we consider iterative schemes in which the velocity integrals ρ and j are fixed from one

step to the next. Then, we construct solutions of the linearized Fokker-Plank and heat problems by means of fundamental solutions. This guarantees nonnegativity of the solutions, a crucial property to obtain preliminary uniform estimates. Existence follows from compactness arguments, using sharp estimates on the force field $\mathbf{F}(C)$ and on the anastomosis terms. Passing to the limit in the equations, we obtain a global in time solution of the original problem for initial data decaying at infinite, as well as stability bounds in terms of the norms of the initial data. Gronwall type inequalities yield uniqueness. More precisely, the following result is proven in [3]

Theorem 1. *Let us assume that:*

$$p_0 \geq 0, C_0 \geq 0, \tag{1.7}$$

$$C_0 \in L^\infty(\mathbb{R}^N), \nabla_{\mathbf{x}} C_0 \in L^\infty(\mathbb{R}^N) \cap L^2(\mathbb{R}^N), \tag{1.8}$$

$$(1+|\mathbf{v}|^2)^{\beta/2} p_0 \in L^\infty(\mathbb{R}^N \times \mathbb{R}^N), \quad \beta > N, \tag{1.9}$$

$$(1+|\mathbf{v}|^2)^{\beta/2} p_0 \in L^1(\mathbb{R}^N \times \mathbb{R}^N), \quad \beta > N. \tag{1.10}$$

Then, there exists a nonnegative solution (p, C) of (1.1)-(1.5) satisfying:

$$C \in L^\infty(0,T; L^\infty(\mathbb{R}^N)), \nabla_{\mathbf{x}} C \in L^\infty(0,T; L^\infty \cap L^2(\mathbb{R}^N)), \tag{1.11}$$

$$p \in L^\infty(0,T; L^\infty \cap L^1(\mathbb{R}^N \times \mathbb{R}^N)), \nabla_{\mathbf{v}} p \in L^2(0,T; L^2(\mathbb{R}^N \times \mathbb{R}^N)), \tag{1.12}$$

$$(1+|\mathbf{v}|^2)^{\beta/2} p \in L^\infty(0,T; L^\infty(\mathbb{R}^N \times \mathbb{R}^N)), \tag{1.13}$$

$$(1+|\mathbf{v}|^2)^{\beta/2} p \in L^\infty(0,T; L^1(\mathbb{R}^N \times \mathbb{R}^N)), \tag{1.14}$$

$$p \in L^\infty(0,T; L^\infty_{\mathbf{x}}(\mathbb{R}^N, L^1_{\mathbf{v}}(\mathbb{R}^N)), \tag{1.15}$$

with norms bounded in terms of the norms of the data.
If $\nabla_{\mathbf{v}} p_0 \in L^\infty_{\mathbf{x}}(\mathbb{R}^N, L^1_{\mathbf{v}}(\mathbb{R}^N))$, then $\nabla_{\mathbf{v}} p \in L^\infty(0,T; L^\infty_{\mathbf{x}}(\mathbb{R}^N, L^1_{\mathbf{v}}(\mathbb{R}^N))$ and the solution is unique.

When the spatial domain Ω is bounded, we need to impose boundary conditions. We consider the slab $(0, L)\times\mathbb{R}$, and set $\mathbf{x} = (x_1, x_2)$, $\mathbf{v} = (v_1, v_2)$. On $x_1 = 0$ (initial blood vessel) and $x_1 = L$ (hypoxic region), we impose Neumann boundary conditions for C:

$$\frac{\partial}{\partial \mathbf{n}} C(0, x_2, t) = 0, \quad \frac{\partial}{\partial \mathbf{n}} C(L, x_2, t) = c_L(t) e^{-a^2 x_2^2}, \quad t > 0, x_2 \in \mathbb{R}, \tag{1.16}$$

where $c_L(t) > 0$ represents the influx of angiogenic factor produced at the hypoxic region and $1/a$ is a characteristic length. This function decreases as blood vessels reach the hypoxic region. We impose nonlocal boundary conditions hold on p:

$$p^+(0, x_2, v_1, v_2, t) = \frac{e^{-|\mathbf{v}-\mathbf{v}_0|^2}}{\int_0^\infty \int_{-\infty}^{+\infty} v_1' \, e^{-|\mathbf{v}'-\mathbf{v}_0|^2} dv_1' dv_2'} \Big[j_0(x_2, t)$$
$$- \int_{-\infty}^0 \int_{-\infty}^{+\infty} v_1' \, p^-(0, x_2, v_1', v_2', t) dv_1' dv_2' \Big] = S_0(p), \quad t, v_1 > 0, v_2, x_2 \in \mathbb{R}, \tag{1.17}$$

$$p^-(L, x_2, v_1, v_2, t) = \frac{e^{-|\mathbf{v}-\mathbf{v}_0|^2}}{\int_{-\infty}^0 \int_{-\infty}^{+\infty} e^{-|\mathbf{v}'-\mathbf{v}_0|^2} dv_1' dv_2'} \Big[\rho_L(x_2, t)$$
$$- \int_0^{+\infty} \int_{-\infty}^{+\infty} p^+(L, x_2, v_1', v_2', t) dv_1' dv_2' \Big] = S_L(p), \quad t > 0, v_1 < 0, x_2, v_2 \in \mathbb{R}, \tag{1.18}$$

where p^+ denotes the values of p for positive v_1 and p^- the values of p for negative v_1. For a fixed $\mathbf{v}_0 = (v_{1,0}, v_{2,0})$

$$j_0(x_2, t) = v_{1,0} \, \alpha(C(0, x_2, t)) \, p(0, x_2, v_{1,0}, v_{2,0}, t), \tag{1.19}$$

$$\rho_L(x_2, t) = \rho(L, x_2, t) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(L, x_2, v_1', v_2', t) dv_1' dv_2'. \tag{1.20}$$

Existence results are based on similar iterative schemes as those employed in the whole space. However, lacking explicit fundamental solutions, proofs exploit balance equations, estimates of velocity decay and compactness results for kinetic operators, combined with gradient estimates of heat kernels for Neumann problems (see [4] for detailed proofs of existence results in an annulus).

We aim to devise robust schemes for this kind of problems. In principle, we could rely on the iterative schemes used for existence, and apply schemes for linear kinetic and heat equations to each iterate. However, the convergence of the iterative scheme may be slow and the order of the resulting procedure would be uncontrolled. Instead, we will discuss how to discretize the original nonlinear problem. To simplify, we will illustrate the ideas on a two dimensional reduction that captures soliton-like solutions. Section 2 describes the discretization procedure. Section 3 presents some numerical results.

2. Positivity preserving high order schemes

In the limit as $\beta \to \infty$, the marginal density $\rho(\mathbf{x}, t)$ and the concentration satisfy the equations [1]:

$$\frac{\partial \rho}{\partial t} + \mathrm{div}_{\mathbf{x}}(\mathbf{F}\rho) - \frac{1}{2\beta}\Delta_{\mathbf{x}}\rho = \mu\,\rho - \Gamma\rho \int_0^t \rho(\mathbf{x}, s)\, ds, \tag{2.1}$$

$$\mu = \frac{\alpha}{\pi}\left[1 + \frac{\alpha}{2\pi\beta(1+\sigma_v^2)}\ln\left(1 + \frac{1}{\sigma_v^2}\right)\right], \tag{2.2}$$

$$\frac{\partial}{\partial t}C(\mathbf{x}, t) = \kappa\Delta_{\mathbf{x}}C(\mathbf{x}, t) - \chi_1\, C(\mathbf{x}, t)\,\rho(\mathbf{x}, t), \tag{2.3}$$

$$\chi_1 = \frac{\chi}{\pi}\int_0^{\infty}\int_{-\pi}^{\pi} \frac{\sqrt{1 + V^2 + 2V\cos\varphi}}{1 + e^{(V^2-\eta)/\epsilon}} e^{-V^2} V\, dV d\varphi. \tag{2.4}$$

To leading order, the density and the marginal density are related by

$$p(\mathbf{x}, \mathbf{v}, t) \sim \frac{1}{\pi} e^{-|\mathbf{v}-\mathbf{v}_0|^2}\rho(\mathbf{x}, t). \tag{2.5}$$

A positivity preserving order one scheme follows by explicit forward time discretization, upwind treatment of transport terms, and standard centered schemes for the Laplacians. Integral terms are discretized by means of composite Simpson rules. The integral $I(\mathbf{x}, t) = \int_0^t \rho(\mathbf{x}, s)\, ds$ is transformed in an additional equation

$$I'(\mathbf{x}, t) = \rho(\mathbf{x}, t), \quad I(\mathbf{x}, 0) = 0. \tag{2.6}$$

To obtain a higher order scheme, we apply a positivity preserving WENO5 scheme to spatial operators, combined with three point Legendre quadrature rules [7]. In spite of their order, these schemes may degenerate to order two in practice. To preserve positivity and stability, we consider strong stability preserving (SSP) time discretizations. Standard choices for third order accuracy are the third order SSP multistep method [7]

$$u(t_{n+1}) = \tfrac{16}{27}\left(u(t_n) + 3\delta t\, r(u(t_n))\right) + \tfrac{11}{27}\left(u(t_{n-3}) + \tfrac{12}{11}\delta t\, r(u(t_{n-3}))\right), \tag{2.7}$$

and the third order Runge Kutta method [6]

$$\begin{array}{l} u^{(1)} = u(t_n) + \delta t\, r(u(t_n)), \\ u^{(2)} = \frac{3}{2}u(t_n) + \frac{1}{4}u^{(1)} + \frac{1}{4}\delta t\, r(u^{(1)}), \\ u(t_{n+1}) = \frac{1}{3}u(t_n) + \frac{2}{3}u^{(2)} + \frac{2}{3}\delta t\, r(u^{(2)}). \end{array} \tag{2.8}$$

The stability of SSP methods is governed by a CFL number c in the following way. If the Euler forward time discretization applied to an equation $u_t = r(u)$ is stable under the condition $\delta t \le \delta t_0$, then the higher order SSP time discretization is stable when $\delta t \le d\,\delta t_0$. For the multistep method we have $d = 1/3$ while $d = 1$ ($d_{eff} = 1/3$) for the RK3 (2.8). For second order accuracy, the RK2 scheme is

$$\begin{array}{l} u^{(1)} = u(t_n) + \delta t\, r(u(t_n)), \\ u(t_{n+1}) = \frac{1}{2}u(t_n) + \frac{1}{2}u^{(1)} + \frac{1}{2}\delta t\, r(u^{(1)}) \end{array} \tag{2.9}$$

with $d = 1$ ($d_{eff} = 1/2$). The spatial operator $r(u(t_n))$ would be the operator obtained discretizing the space variables, time excluded. These schemes can be extended to the whole model [5].

3. Numerical results

In this section, we present numerical solutions for appropriate values of the parameters as listed in Table 1.

Figure 1 shows the evolution of the marginal tip density (1.5) at four different times as the angiogenic network moves towards the hypoxic region on the right, obtained by combining a WENO5 discretization in space and RK2 in time. Figures 2 depicts the angiogenic factor concentration. We observe that the active vessel tips evolve as a patch as they consume the concentration of theangiogenic factor as they advance. The tip density profile forms a soliton-like pattern, with slightly varying profile. The soliton forms at the initial stage and then advances keeping its appearance but changing its size and height. This numerically observed soliton can be described asymptotically as explained in [1].

Acknowledgements

This research has been partially supported by the FEDER/Ministerio de Ciencia, Innovación y Universidades -Agencia Estatal de Investigación grant No. MTM2017-84446-C2-1-R.

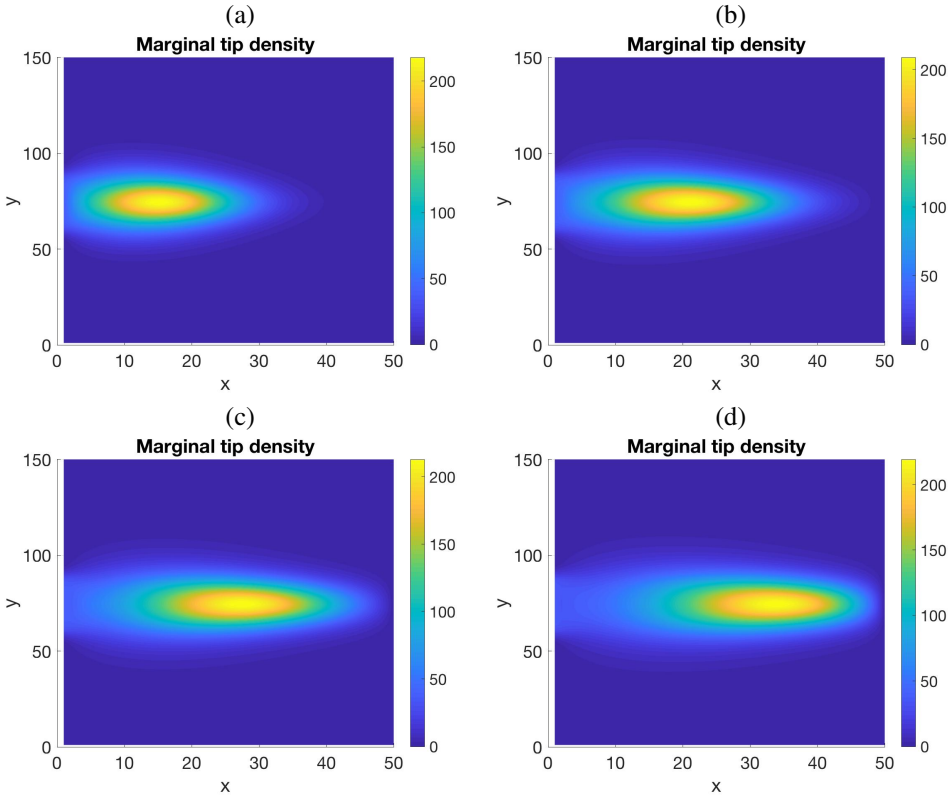


Fig. 1 Snapshots of the time evolution of the marginal tip density.

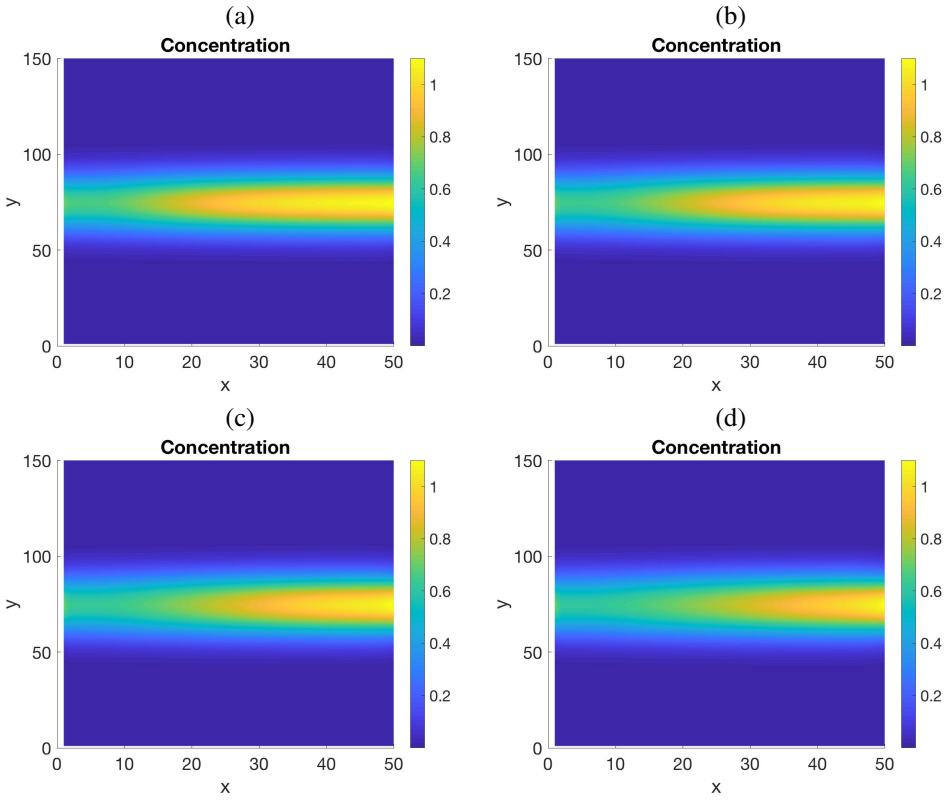


Fig. 2 Snapshots of the time evolution of the concentration.

References

[1] LL Bonilla, A Carpio, M Carretero, G Duro, M Negreanu, F Terragni. A convergent numerical scheme for integrodifferential kinetic models of angiogenesis. *J. Comput. Phys.*, 375:1270–1294, 2018.

[2] V. Capasso, D. Morale, G. Facchetti. Randomness in self-organized phenomena. A case study: Retinal angiogenesis. *BioSystems*, 112:292-297, 2013.

[3] A. Carpio, G. Duro Well posedness of an integrodifferential kinetic model of Fokker-Planck type for angiogenesis. *Nonlinear Analysis-Real World Applications*, 30:184-212, 2016.

[4] A. Carpio, G. Duro, M. Negreanu, Constructing solutions for a kinetic model of angiogenesis in annular domains *Applied Mathematical Modelling*, 45:303-322, 2017.

[5] A. Carpio, E. Cebrián, High order positivity preserving schemes for kinetic models of angiogenesis. Preprint, 2020.

[6] C.W. Shu, Total-variation diminishing time discretizations, SIAM J. Sci. Statist. Comput. 9, 1073-1084, 1988.

[7] X. Zhang, Y. Liu, C.W. Shu. Maximum-principle-satisfying high order finite volume weighted essentially nonoscillatory schemes for convection-diffusion equations. *SIAM J. Sci. Comput.*, 34:A627-A658, 2012.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 114–121)

Variable time-step modal methods to integrate the time-dependent neutron diffusion equation

A. Carreño[1], A. Vidal-Ferràndiz[2], D. Ginestar[2], G. Verdú[1]
1. amcarsan@iqn.upv.es, *ISIRYM, Universitat Politècnica de València, Spain*
2. *IMM, Universitat Politècnica de València, Spain*

Abstract

The time-dependent neutron diffusion equation can describe the power evolution inside a nuclear reactor core. One approach to integrate this time-dependent equation is the modal method. This methodology is based on assuming that the solution can be decomposed as a finite sum of time-dependent amplitudes multiplied by shape functions (obtained by solving a partial eigenvalue problem), which are updated along the transient. In this work, different controls, that adapt the time-step according to the state of the transient, are implemented. Several benchmark problems show the competitiveness of the methodology.

1. Introduction

The evolution of the neutron power inside of reactor core can be described by the time-dependent multigroup neutron diffusion equation [11]. This equation is an approximation of the neutron transport equation that assumes that the neutron current is proportional to the gradient of the scalar neutron flux by means of a diffusion coefficient. For two energy-groups and without considering up-scattering, this model can be expressed as

$$\begin{aligned} \mathcal{V}\frac{d}{dt}\Phi + \mathcal{L}\Phi &= (1-\beta)\mathcal{F}\Phi + \sum_{k=1}^{K}\lambda_k^d \chi C_k, \\ \frac{d}{dt}C_k &= \beta_k \mathcal{F}_1\Phi - \lambda_k^d C_k, \qquad k = 1,\dots,K\,, \end{aligned} \tag{1.1}$$

where the time-dependent operators are

$$\begin{aligned} \mathcal{L} &= \begin{pmatrix} -\vec{\nabla}\cdot(D_1\vec{\nabla}) + \Sigma_{a_1} + \Sigma_{12} & 0 \\ -\Sigma_{12} & -\vec{\nabla}\cdot(D_2\vec{\nabla}) + \Sigma_{a_2} \end{pmatrix}, \\ \mathcal{F} &= \begin{pmatrix} \nu\Sigma_{f_1} & \nu\Sigma_{f_2} \\ 0 & 0 \end{pmatrix}, \qquad \mathcal{F}_1 = \begin{pmatrix} \nu\Sigma_{f_1} & \nu\Sigma_{f_2} \end{pmatrix}, \\ \mathcal{V} &= \begin{pmatrix} 1/v_1 & 0 \\ 0 & 1/v_2 \end{pmatrix}, \qquad \chi = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \Phi = \begin{pmatrix} \Phi_1 \\ \Phi_2 \end{pmatrix}. \end{aligned} \tag{1.2}$$

In the previous expressions, D is the diffusion coefficient, Σ_a is the absorption cross-section, Σ_{12} is the scattering cross-section from the first group to the second group, Σ_f is the fission cross-section, ν is the average number of neutrons produced in each fission and v_g is the velocity of the neutrons. Subindex $g(=1,2)$ denotes the energy group. The first and the second group are known as fast and thermal group, respectively. Thus, Φ_1 and Φ_2 are the fast and thermal neutron fluxes. The concentration of delayed neutron precursors is represented by C_k, where subindex k denotes the delayed group k. β_k is the fraction of delayed neutrons that satisfies $\sum_{k=1}^{K}\beta_k = \beta$. λ_k^d is the neutron decay constant. All magnitudes and variables are, in general, time and space-dependent.

A Galerkin finite element method is applied for the spatial discretization of the neutron diffusion equation to obtain a semi-discrete system of ordinary differential equations [13]. Usually, this system of differential equations is stiff due to, among other things, the fast variation of the neutron flux and the presence of both prompt and delayed neutrons that leads to time scales of different orders of magnitude.

Several methodologies of different types have been studied to integrate this semi-discrete equation. First, one can use implicit differential schemes such as the backward differential method or higher order differential schemes [5, 14]. There are also many works that apply a quasi-static method that decomposes the solution as a product of two functions: an amplitude function that only depends on the time and a shape function that depends on space and time but its variation in time is assumed to be slow. These functions are approximated with two

different schemes that are coupled. In this work, to integrate the time-dependent neutron diffusion equation, a modal method is used. This approach assumes that the solution can be described as the sum of several amplitude functions multiplied by shape functions or modes. This expansion has a strong interest to approximate the solution for some types of transient problems, such as the ones defined from out-of-phase oscillations or local perturbations, where more than one shape function is necessary. The shape functions are obtained computing the eigenfunctions associated with the dominant eigenvalues (larger in magnitude) of the λ-modes problem

$$\mathcal{L}\phi_m = \frac{1}{\lambda_m}\mathcal{F}\phi_m\,, \tag{1.3}$$

where $\mathcal{L}$ and $\mathcal{F}$ are the operators defined in Equation (1.2) for a given configuration of the reactor. This generalized eigenvalue problem is obtained forcing the criticality of the system and describes the steady-state, which is the initial condition for the problem (1.1). Other possibilities for the shape functions have been studied, but they are not as efficient numerically as the λ-modes [4].

The shape functions can be fixed along the transient and one can use the modes associated with the static problem but this implies, in some transients, the necessity of using a high number of them in the expansion to obtain accurate results. To reduce the number of modes, Miró et al. proposed an updated modal method where the shape functions are updated at some time-steps [10]. In this last work, the time-step to update the modes was a fix value that had to be selected before beginning the computation. This strategy leads, in some cases, to use a too small time-step to assure the accuracy of the computations, which implies an unnecessary large computational cost. In this work, we propose some adaptive time-step controls that estimate an updating time-step depending on the transient analysed such that the results obtained are accurate enough with reasonable computational demands.

The rest of the paper is organized as follows. Section 2 briefly describes the finite element method used for the spatial discretization. Section 3 exposes the updated modal method. Section 4 presents the adaptive time-step controls analysed. Section 5 contains the numerical results obtained to test the proposed methodology. Finally, Section 6 collects the main conclusions of this work.

2. Spatial discretization. Finite element method

To approximate the solution, the differential system (1.1) is discretized. For the spatial discretization, a continuous Galerkin finite element method (FEM) is applied to obtain the semi-discrete system of differential equations (see details in [14])

$$\begin{aligned} V\frac{\mathrm{d}\tilde{\Phi}}{\mathrm{d}t} + L\tilde{\Phi} &= (1-\beta)F\tilde{\Phi} + \sum_{k=1}^{K}\lambda_k^d XC_k, \\ \frac{\mathrm{d}XC_k}{\mathrm{d}t} &= \beta_k F\tilde{\Phi} - \lambda_k^d XC_k, \qquad k = 1,\dots,K\,, \end{aligned} \tag{2.1}$$

where L, F, V and X are the matrices obtained from the spatial discretization of operators $\mathcal{L}$, $\mathcal{F}$, $\mathcal{V}$ and $\mathcal{X}$, respectively. Vectors $\tilde{\Phi}$ and C_k are the corresponding coefficients of Φ and C_k in terms of the Lagrange polynomials, which are the polynomials used in the finite element method. The FEM has been implemented by using the open source finite elements library Deal.II ([2]). Henceforth, the notation has been simplified by removing the tildes of the discrete operators from the original notation to the vectors Φ and ϕ.

For the λ-modes problem (1.3), the algebraic problem associated with the spatial discretization has the following structure

$$L\phi_m = \frac{1}{\lambda_m}F\phi_m, \tag{2.2}$$

where $\tilde{\phi}_m$ are the algebraic vectors of coefficients associated with the functions ϕ_m.

Associated with the λ-modes problem one can define the adjoint problem [6]

$$L^T\phi_l^\dagger = \frac{1}{\lambda_l}F^T\phi_l^\dagger, \tag{2.3}$$

where L^T and F^T are the matrices obtained from the spatial discretization of the adjoint operators $\mathcal{L}^\dagger$ and $\mathcal{F}^\dagger$. They also correspond to the transpose matrices of L and F. The solutions of the adjoint modes problem $\phi_l^\dagger$, $l = 1,\dots,q$ satisfy the biorthogonality condition

$$\langle\phi_l^\dagger, F\phi_m\rangle = \langle\phi_m^\dagger, F\phi_m\rangle\delta_{l,m}, \quad l,m = 1,\dots,q, \tag{2.4}$$

where $\langle\,,\rangle$ denotes the inner product for vectors and $\delta_{l,m}$ is the Kronecker's delta.

3. Time discretization. Updated modal method

From the finite element discretization, a semi-discrete system of ordinary differential equations is obtained that must be integrated over the time. In this work, the updated modal expansion is used for this purpose. This is a generalization of the traditional modal method that updates the eigenfunctions used in the expansions to avoid using a high number of modes [10].

To apply this method, the time domain is divided into several intervals $[t_i, t_i + \Delta t_i] = [t_i, t_{i+1}]$ and the neutron flux in this interval is decomposed in terms of q dominant λ-modes as

$$\Phi^i(\vec{r}, t) = \sum_{m=1}^{q} n_m^i(t)\phi_m^i(\vec{r}), \quad t \in [t_i, t_{i+1}], \tag{3.1}$$

where $\phi_m^i(\vec{r})$ is the unitary eigenvector associated with the m-th dominant eigenvalue of the λ-modes problem (1.3) associated with the configuration of the reactor in time $t = t_i$

$$L^i\phi_m^i = \frac{1}{\lambda_m^i}F^i\phi_m^i, \tag{3.2}$$

and $n_m^i(t)$ is the amplitude coefficient associated, that is only time-dependent. The matrices L^i and F^i correspond to the matrices L and F at time $t = t_i$.

Along the transient, it is assumed that only the magnitudes included in the operators L and F are time-dependent. The precursor data and the velocities are considered constant. The matrices L and F are expressed as

$$L(t) = L^i + \delta L^i(t), \qquad F(t) = F^i + \delta F^i(t), \quad t \in [t_i, t_{i+1}]. \tag{3.3}$$

The transient is initialized by forcing the criticality of the reactor dividing the fission cross-sections by λ_1^0 and using the steady-state neutron flux as initial condition.

To apply the modal methodology, the expressions (3.1) and (3.3) are substituted in the discretized neutron diffusion equation (2.1) and the resulting expression is then collapsed on the left by the adjoint λ-modes to obtain a system of $q(K + 1)$ ODEs

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{N}^i = \mathbf{T}^i\mathbf{N}^i, \tag{3.4}$$

where

$$\mathbf{N}^i = \begin{pmatrix} n_1^i \cdots n_q^i & c_{11}^i \cdots c_{q1}^i & \cdots & c_{1K}^i \cdots c_{qK}^i \end{pmatrix}^T, \tag{3.5}$$

$$\mathbf{T}^i = \left(\begin{array}{c|ccc} \Lambda_i^{-1}((1-\beta)I - [\lambda_i]^{-1} - \Delta L^i + (1-\beta)\Delta F^i) & \Lambda_i^{-1}\lambda_1^d & \cdots & \Lambda_i^{-1}\lambda_K^d \\ \hline \beta_1(I + \Delta F^i) & -\lambda_1^d I & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \beta_K(I + \Delta F^i) & 0 & \cdots & -\lambda_K^d I \end{array}\right), \tag{3.6}$$

and

$$\begin{aligned} &\Lambda_{lm}^i = \langle \phi_l^{\dagger,i}, V\phi_m^i\rangle, \qquad \Delta L_{lm}^i = \langle \phi_l^{\dagger,i}, \delta L^i\phi_m^i\rangle, \\ &\Delta F_{lm}^i = \langle \phi_l^{\dagger,i}, \delta F^i\phi_m^i\rangle, \quad c_{lk}^\lambda = \langle \phi_l^{\dagger,i}, XC_k\rangle. \end{aligned} \tag{3.7}$$

The matrix block $[\Lambda]_i$ is a diagonal matrix whose elements are the dominant λ-modes λ_m^i. The initial conditions at $t = 0$ are

$$\begin{aligned} &n_1^0(0) = 1, && n_m^0(0) = 0,\ m = 2, \ldots, q, \\ &c_{1k}^0(0) = \frac{\beta_k}{\lambda_k^d}\langle \phi_1^{\dagger,0}, F_0\phi_1^0\rangle, && c_{mk}^0(0) = 0,\ m = 2, \ldots, q, \quad k = 1, \ldots, K, \end{aligned} \tag{3.8}$$

with ϕ_1^0 and $\phi_1^{\dagger,0}$, the corresponding eigenvector and its adjoint associated with the dominant eigenvalue λ_1^0. That are obtained from the problem in the initial configuration.

The initial conditions at t_i to integrate the system in the interval $[t_i, t_{i+1}]$ must be defined to ensure the continuity of the solution. These initial conditions will be calculated from the solution in the previous interval $[t_{i-1}, t_i]$, the eigenvectors associated with direct modes (ϕ_m^i) and the adjoint modes ($\phi_l^{\dagger,i}$). Therefore, the computation of the solution in the interval $[t_i, t_{i+1}]$ uses the solution of the previous interval $[t_{i-1}, t_i]$.

First, the initial conditions for n_m^i in the interval $[t_i, t_{i+1}]$ are defined. The vector $\Phi(t_i)$ can be computed by using the expansion in the interval $[t_{i-1}, t_i]$ as

$$\Phi^{i-1}(t_i) = \sum_{m=1}^{q} n_m^{i-1}(t_i)\phi_m^{i-1}. \tag{3.9}$$

Assuming the continuity of $\Phi(t)$ on all its domain, that is $\Phi^{i-1}(t_i) = \Phi^i(t_i)$, and collapsing the expression (3.1) at $t = t_i$ by the adjoint modes it is obtained that the amplitude coefficients must be equal to

$$n_m^i(t_i) = \frac{\langle \phi_m^{\dagger,i}, F^i \Phi^{i-1}(t_i) \rangle}{\langle \phi_m^{\dagger,i}, F^i \phi_m^i \rangle}. \tag{3.10}$$

To compute the initial conditions for the concentration of the precursor k at time t_i, $c_{l,k}^i(t_i)$, the coefficients computed in the previous integration for $t \in [t_{i-1}, t_i]$ and the adjoint modes are used. It is supposed that

$$\phi_l^{\dagger,i} = \sum_{m=1}^{q} a_{lm} \phi_m^{\dagger,i-1}. \tag{3.11}$$

Using the biorthogonality relation of the adjoint λ-modes and Equation (3.11) it is obtained that

$$a_{lm} = \frac{\langle \phi_l^{\dagger,i}, F^{i-1} \phi_m^{i-1} \rangle}{\langle \phi_m^{\dagger,i-1}, F^{i-1} \phi_m^{i-1} \rangle}. \tag{3.12}$$

Therefore, the precursors coefficients at time t_i can be computed as

$$c_{l,k}^i(t_i) = \langle \phi_l^{\dagger,i}, XC_k \rangle(t_i) = \sum_{m=1}^{q} a_{lm} \langle \phi_m^{\dagger,i-1}, XC_k \rangle(t_i) = \sum_{m=1}^{q} a_{lm} c_{m,k}^{i-1,\lambda}(t_i). \tag{3.13}$$

Note that the system of ODEs (3.4) is much smaller than the system (2.1) if a moderate number of modes is used in expansion (3.1). In this work, this stiff system is integrated with a backward differentiation formula implemented in the CVODE solver from the SUNDIALS library [1, 7]. This code has implemented an adaptive time step and it is initialized with a time step of 10^{-3}s.

4. Adaptive time-step control

Previous works [10], update the modes with a fix time-step that is selected before beginning the computation. This implies the necessity of selecting a time-step that can lead to results with unpredictable errors. If a small time-step is used to reduce the errors, the computational cost also increases and this small time-step may be not necessary in some stages of the transient. Consequently, it is interesting to have an algorithm to adapt the time-step during the transient. To define it, two fundamental points must be studied: an error estimation and a control to select the time-step based on this error.

4.1. Estimation of local error

The error obtained for the modal expansion essentially comes from the assumption that the neutron flux can be described as a finite linear combination of the spatial modes, because the set of q modes do not form a complete basis of the function space. Therefore, larger spatial variations in the flux will imply larger errors in the modal method. We define several types of errors to estimate these spatial variations.

Modal difference error One way to estimate how the neutron flux will change in the interval $[t_i, t_{i+1}]$ is to compute the modes in the next time t_{i+1} and observe the differences between the modes at t_i and the modes at t_{i+1} as

$$\varepsilon_{md}^i = \max_{m=1,\dots,q} \frac{\|\phi_m^{i-1} - \phi_m^i\|_1}{\|\phi_m^{i-1}\|_1} k_{md}\,. \tag{4.1}$$

Modal residual error Other possibility to estimate this change is computing the residual error of the modes at t_i on the problem corresponding to the time t_{i+1} as

$$\varepsilon_{mr} = \max_{m=1,\dots,q} \frac{\|F^i \phi_m^{i-1} - \lambda_m^{i-1} L^i \phi_m^{i-1}\|_1}{\|\phi_m^{i-1}\|_1} k_{mr}\,.$$

Cross-section perturbation error In nuclear reactor systems the neutron flux shape change will depend on the variation in the cross-sections. Thus, we define an error

$$\varepsilon_{xs} = \sum_c \frac{\|\mathrm{XS}^{i-1}(c) - \mathrm{XS}^i(c)\|_1}{\|\mathrm{XS}^{i-1}(c)\|_1} k_{xs},$$

where c denotes the different cells of the spatial discretization of the reactor and XS is one type of cross-section that depends on the perturbation applied to generate the transient.

In the previous error estimations the constants k_{md}, k_{mr}, k_{xs}, are defined to adjust the accuracy of the approximation and their values will depend on the transient analyzed.

4.2. Time-step control

Two strategies are defined to compute the time-step from the error estimations. Both compute the new time-step Δt_i from the previous one Δt_{i-1} .

Banded control The first control computes this time-step in a fixed way as

$$\Delta t_i = \begin{cases} \Delta 2 t_{i-1}, & \varepsilon < 1.0, \\ \Delta t_{i-1}, & 1.0 < \varepsilon < 2.0, \\ \Delta t_{i-1} \,/\, 2, & 1.0 < \varepsilon, \end{cases} \tag{4.2}$$

where ε is one of the error estimations presented in Section 4.1.

Dynamic control It is based on control algorithms of other differential methods implemented for stiff problems [15]. In particular, the time-step Δt_i is computed as

$$\Delta t_i = \Delta t_{i-1} \min\{2.0, \max\{0.5, \sqrt{1.0/\varepsilon}\}\}, \tag{4.3}$$

where ε is some error defined in Section 4.1.

Finally, a minimum time-step and maximum time-step to avoid using very high or very small time-steps are used. These values are defined as $\Delta t_{\min} = \Delta t_0/2$, $\Delta t_{\max} = 50\Delta t_0$, where Δt_0 is the initial Δt.

5. Numerical results

The performance of the variable time-step updated modal methodology is tested using two type of reactor transients. In the finite element method, Lagrange polynomials of degree 3 are used because usually, this degree gives accurate results for usual reactor calculations [14].

The solution of the partial eigenvalue problems has been computed with a hybrid method by using a residual error of 10^{-7} (see more details in [3]). Moreover, for the implementation, a matrix-free technique is applied where the matrices of the system are not assembled. Matrix-vector products are computed 'on the fly' in a cell-based interface.

To analyzed the results, different relative errors for the neutron power are computed. The neutron power distribution, P, is defined as

$$\mathrm{P}(\vec{r}, t) = \Sigma_{f1}\Phi_1(\vec{r}, t) + \Sigma_{f2}\Phi_2(\vec{r}, t).$$

The Local Error (LE) at time t and Mean Power Error (MPE) in the interval $[t_0, t_N]$ are given by

$$\mathrm{LE}(t) = \frac{\|\mathrm{P}(t) - \mathrm{P}^{\mathrm{ref}}(t)\|_1}{\|\mathrm{P}^{\mathrm{ref}}(t)\|_1} \cdot 100, \qquad \mathrm{MPE} = \frac{1}{(t_N - t_0)} \sum_{n=1}^{N} \mathrm{LE}(t_n)(t_n - t_{n-1}),$$

where $\mathrm{P}^{\mathrm{ref}}(t)$ is the reference power at time t, that is computed from the solution obtained with a backward differential method (BKM) of first order and time-step of 0.001 s. [14].

This methodology has been implemented in C++ based on data structures provided by the library Deal.ii [2], PETSc [1]. It has been incorporated to the open-source neutronic code `FEMFFUSION`. It approximates the neutron diffusion equation and the steady-state $\mathrm{SP_N}$ equations by using a high order finite element method. The full description and the source code of `FEMFFUSION` is available in [12].

The computer used in the computations has been an Intel® Core™ i7-4790 @3.60GHz×8 processor with 32Gb of RAM running on Ubuntu GNU/Linux 18.04 LTS.

5.1. Langenbuch OOP transient

The Langenbuch reactor is a small LWR core with 77 fuel assemblies and two types of fuel [9]. The Langenbuch-OOP transient is defined from two local sinusoidal perturbations that are out-of-phase between them. They are expressed as,

$$\Sigma_{fg}(t) = \Sigma_{fg}(0) + \delta\Sigma_{fg}(t) \qquad g = 1, 2. \tag{5.1}$$

The perturbation 1 ($\mathrm{P_1}$) and the perturbation 2 ($\mathrm{P_1}$), marked in the Figure 1 with dash pattern, are given by

$$\delta\Sigma_{fg}^{P_1}(t) = 5 \cdot 10^{-4} \sin(2\pi t), \qquad \delta\Sigma_{fg}^{P_2}(t) = 5 \cdot 10^{-4} \sin(2\pi t + \pi), \qquad g = 1, 2. \tag{5.2}$$

This transient is followed during 2 s. The number of modes for the modal method has been set to $q = 3$ because the out-of-phase perturbations cannot be described with only one mode.

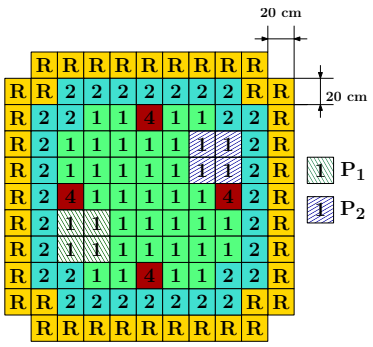
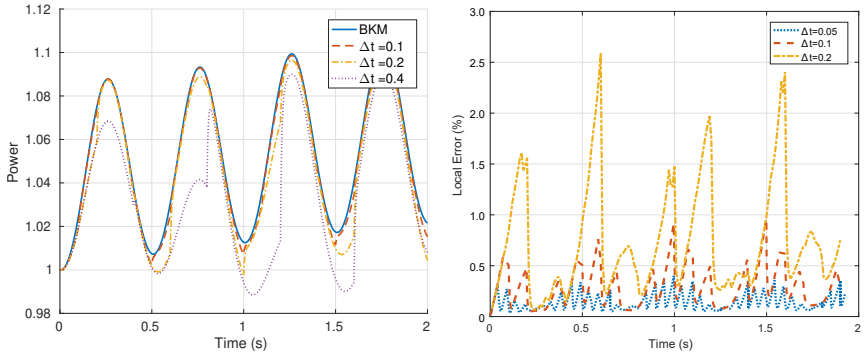


Fig. 1 Radial location of the perturbation areas for the Langenbuch-OPP transient.



(a) Relative power (b) Local Error

Fig. 2 Relative power and local error (%) obtained with the updated modal method with 3 modes for the Langenbuch-OPP transient.

Figure 2(a) displays the evolution of the relative global power computed with the BKM and the updated modal method with several fixed time-steps (Δt). Figure 2(b) displays the local error (LE). Large errors between the BKM and the updated modal method are produced when the perturbations reach their maximum value. However, these differences are reduced for smaller time-steps. Errors are not constant along the transient and the use of a control for the updating time-step is convenient to improve the efficiency of the method.

First, the adaptive control is analysed. The time-step to initialize the time-step control is set to $\Delta t_0 = 0.05$ s. The fission cross-section is used for the *cross-section perturbation error*. Table 1 shows Mean Power Errors and CPU times obtained by setting the different error estimations, control errors and accuracy coefficients k. The *modal difference error* (ε_{md}) is very expensive because it needs to compute the modes at the start of each time interval. The *cross-section perturbation error* (ε_{xs}) gives lower errors than the *modal residual error* (ε_{mr}), but by requiring more time. In the type of controls, the dynamic control gives similar approximations than the banded control, but also using more CPU time.

Tab. 1 Mean Power Errors (MPE) and CPU time obtained with the adaptive time-step modal method for the Langenbuch-OOP transient.

Type of Error	**Banded Control**		**Dynamic Control**	
k	**MPE (%)**	**CPU Time**	**MPE (%)**	**CPU Time**
ε_{md}				
1.0	3.901	3 min	3.748	5 min
2.0	2.902	22 min	1.681	14 min
5.0	1.641	32 min	1.519	20 min
ε_{mr}				
50	6.383	3 min	2.117	3 min
100	1.011	7 min	1.162	8 min
200	1.531	7 min	0.713	18 min
ε_{xs}				
0.5	1.338	4 min	0.809	7 min
1.0	0.647	11 min	0.656	12 min
2.0	0.617	13 min	0.607	13 min

Table 2 compares the mean power errors for the updated modal method with several fixed time-steps Δt and for the adaptive modal method with the *cross-section perturbation residual error*, *dynamic* control time-step and

$k_{xs} = 0.5$. The updated modal method with fix $\Delta t = 0.2$ s and the adaptive modal method uses similar CPU times, but the approximation obtained with the adaptive control is a 20 % more accurate.

Tab. 2 Errors and CPU time obtained to integrate the Langenbuch-OOP transient.

	BKM	**Updated Modal** $\Delta t = 0.2$ s	**Updated Modal** $\Delta t = 0.1$ s	**Updated Modal** $\Delta t = 0.05$ s	**Adaptive Modal**
MPE (%)		0.988	0.558	0.187	0.809
CPU Time (min)	195	7	12	30	7

5.2. AER-DYN-001 transient

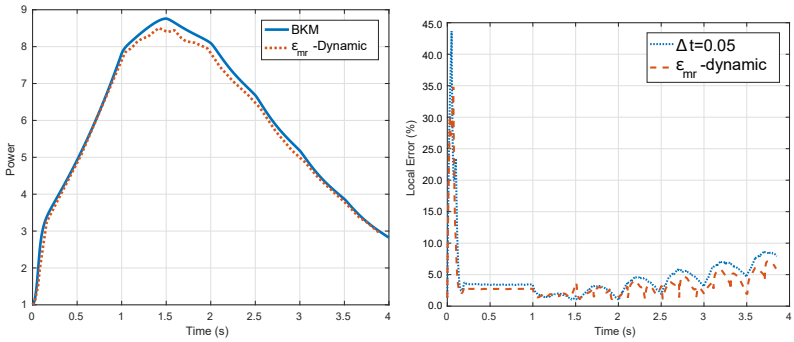
The AER-DYN-001 problem was introduced in [8]. It corresponds to an asymmetric control rod ejection accident without any feedback in a large hexagonal VVER440 reactor. The discretization of the reactor core is composed of 15 156 cells to have a system of 3 361 970 degrees of freedoms for a degree in the FEM equal to 3. Two dominant modes are used for the modal method because one eigenvalue gives non-accurate approximations [8].

Table 3 shows the mean power error (MPE) and the CPU time obtained with the updated modal method with fix time-steps and the adaptive control. Very small time-steps are necessary in the updated modal method to approximate accurately the drop out of the bar at the beginning of the transient, but then these small values are not necessary. Thus, different time-steps are interesting to be used along the transient. The adaptive updated modal method with modal residual error, dynamic control time-step and $k_{mr} = 100$ is applied. This solution has smaller mean power error than the rest of the solutions computed with the fixed updated modal method and this is also computed in less CPU time.

Tab. 3 Comparison of the BKM and the updated modal method for the AER-DYN-001 transient.

Method	**Δ*t*** (s)	**MPE** (%)	**CPU Time** (h)
BKM	0.01	-	89
Updated modal method	0.01	5.90	140
Updated modal method	0.05	4.60	38
Updated modal method	0.10	4.86	23
Adaptive updated modal method	-	3.59	17

Figure 3(a) shows the relative power obtained with the updated modal method with adaptive time step and the BKM. Figure 3(b) displays the local errors of the updated modal method using a fix time step $\Delta t = 0.05$ s and the adaptive update modal method. It observed that the adaptive modal method reduces the local error in the first times, but also reduces the local error beyond $t = 1$ s.



(a) Relative power (b) Local Error

Fig. 3 Relative power, local error (%) and time-step (Δt) obtained with the updated modal method and the adaptive updated modal method with 2 modes for the AER-DYN-001 transient.

6. Conclusions

An updated modal method with a variable time-step is proposed to integrate the neutron diffusion equation, where the updating time-step is selected in function on different types of errors and controls.

Numerical results show that the *modal residual error* and the *cross-section perturbation error* are good estimators to control the time-step update. However, the *modal difference error* has been shown computationally very expensive. In the type of controls, the dynamic control error is more adapted to the local errors, but there are not relevant differences between the dynamic and the banded time step control. Moreover, different coefficients k are defined to adjust the accuracy obtain in the different errors. Values of $k_{md} \approx 2.0$, $k_{mr} \approx 100$ and $k_{xs} \approx 0.5$ are recommended.

Finally, numerical results show that the time-step control for the updated modal method decreases the errors with similar or smaller CPU times than if the updated modal method is applied with fix time-step.

Acknowledgements

This work has been partially supported by Spanish Ministerio de Economía y Competitividad under projects ENE2017-89029-P and MTM2017-85669-P. Furthermore, this work has been financed by the Generalitat Valenciana under the project PROMETEO/2018/035.

References

[1] S. Balay, S. Abhyankar, M. Adams, J. Brown, P. Brune, K. Buschelman, L. Dalcin, A. Dener, et al. Petsc users manual. 2019.

[2] W. Bangerth, T. Heister, and Kanschat G. `deal.II` *Differential Equations Analysis Library*. `http://www.dealii.org`.

[3] A. Carreño, A. Vidal-Ferràndiz, D. Ginestar, and G. Verdú. Block hybrid multilevel method to compute the dominant λ-modes of the neutron diffusion equation. *Annals of Nuclear Energy*, 121:513–524, 2018.

[4] A. Carreño, A. Vidal-Ferràndiz, D. Ginestar, and G. Verdú. Modal methods for the neutron diffusion equation using different spatial modes. *Progress in Nuclear Energy*, 115:181–193, 2019.

[5] D. Ginestar, G. Verdú, V. Vidal, R. Bru, J. Marín, and J. L. Munoz-Cobo. High order backward discretization of the neutron diffusion equation. *Annals of Nuclear Energy*, 25(1-3):47–64, 1998.

[6] A.F. Henry. *Nuclear-reactor analysis*, volume 4. MIT press Cambridge, Massachusetts, 1975.

[7] A.C. Hindmarsh, P.N. Brown, K.E. Grant, S. L. Lee, R. Serban, D.E. Shumaker, and C.S. Woodward. Sundials: Suite of nonlinear and differential/algebraic equation solvers. *ACM Transactions on Mathematical Software (TOMS)*, 31(3):363–396, 2005.

[8] A. Keresztri and M. Telbisz. Dynamic Benchmark 1. `http://aerbench.kfki.hu/aerbench/`, 2009. [Dyn001.doc].

[9] S. Langenbuch, W. Maurer, and W. Werner. Coarse-mesh flux-expansion method for the analysis of space-time effects in large light water reactor cores. *Nuclear Science and Engineering*, 63(4):437–456, 1977.

[10] R. Miró, D. Ginestar, G. Verdú, and D. Hennig. A nodal modal method for the neutron diffusion equation. Application to BWR instabilities analysis. *Annals of Nuclear Energy*, 29(10):1171–1194, 2002.

[11] W.M. Stacey. *Nuclear reactor physics*, volume 2. Wiley Online Library, 2007.

[12] A. Vidal-Ferràndiz, A. Carreño, D. Ginestar, and G. Verdú. `FEMFFUSION`: A finite element method code for the neutron diffusion equation. `https://www.femffusion.imm.upv.es`, 2020.

[13] A. Vidal-Ferrandiz, R. Fayez, D. Ginestar, and G. Verdú. Solution of the lambda modes problem of a nuclear power reactor using an h–p finite element method. *Annals of Nuclear Energy*, 72:338–349, 2014.

[14] A. Vidal-Ferràndiz, R. Fayez, D. Ginestar, and G. Verdú. Moving meshes to solve the time-dependent neutron diffusion equation in hexagonal geometry. *Journal of computational and applied mathematics*, 291:197–208, 2016.

[15] G. Wanner and E. Hairer. *Solving ordinary differential equations II*. Springer Berlin Heidelberg, 1996.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 122–129)

Homoclinic bifurcations in the unfolding of the nilpotent singularity of codimension 4 in $\mathbb{R}^4$

Pablo S. Casas[1], **Fátima Drubi**[2], **Santiago Ibáñez**[2]
1. pabloscasas@uniovi.es *Universidad de Oviedo, Spain*
2. Universidad de Oviedo, Spain

Abstract

The rich variety of homoclinic phenomena exhibited by the limit family of any generic unfolding of a four-dimensional nilpotent singularity of codimension-four is discussed. Specifically, numerical techniques based on the Taylor integrator and the expansion of the invariant manifolds were designed for this family. A partial bifurcation diagram which includes, besides a suggestive catalogue of local bifurcations of equilibria, folds and period doublings of periodic orbits is also given. These results are certainly the first steps towards a much more ambitious goal: to achieve a general understanding of these codimension-four unfoldings.

1. Introduction

Let X be a C^∞ vector field on $\mathbb{R}^n$ with $X(0) = 0$ and 1-jet at 0 linearly conjugate to $\sum_{i=1}^{n-1} x_{i+1}\,\partial/\partial x_i$. Vector fields satisfying this assumption make up a set of codimension n in the space of germs of singularities in $\mathbb{R}^n$ (see [19] for definitions). As argued in [8], working with appropriate coordinates, X can be written as the following differential equation

$$\begin{cases} x_i' &= x_{i+1} \qquad \text{for} \quad i = 1, \dots, n-1, \\ x_n' &= f(x), \end{cases}$$

with $x = (x_1, \dots, x_n)$ and $f(x) = O(\|x\|^2)$. We say that 0 (or X itself) is a n-dimensional nilpotent singularity of codimension n when the condition

$$\frac{\partial^2 f}{\partial x_1^2}(0) \neq 0$$

is satisfied.

Consider now a C^∞-family of vector fields X_ν, with $\nu = (\nu_1, \dots, \nu_n) \in \mathbb{R}^n$, such that X_0 is a n-dimensional nilpotent singularity of codimension n. As proved in [8], under generic assumptions, X_ν can be written as follows

$$\begin{cases} x_i' &= x_{i+1} \qquad \text{for} \quad i = 1, \dots, n-1, \\ x_n' &= \nu_1 + \nu_2 x_2 + \ldots + \nu_n x_n + x_1^2 + h(\nu, x), \end{cases}$$

where $\nu_1, \dots, \nu_n$ and the coefficient in front of x_1^2 represent exact coefficients in a Taylor expansion with respect to x and h is of order $O(\|(\nu, x)\|^2)$ and $O(\|(x_2, \dots, x_n)\|^2)$.

As proved in [1], when rescaling parameters and variables by the equations

$$\begin{aligned} \nu_1 &= \varepsilon^{2n} \bar{\nu}_1, \\ \nu_k &= \varepsilon^{n-k+1} \bar{\nu}_k \qquad \text{for} \quad k = 2, \dots, n, \\ x_k &= \varepsilon^{n+k-1} \bar{x}_k \qquad \text{for} \quad k = 1, \dots, n, \end{aligned} \tag{1.1}$$

with $\varepsilon > 0$ and $\bar{\nu}_1^2 + \ldots + \bar{\nu}_n^2 = 1$, X_ν becomes the system

$$\begin{cases} \bar{x}_i' &= \bar{x}_{i+1} \qquad \text{for} \quad i = 1, \dots, n-1, \\ \bar{x}_n' &= \bar{\nu}_1 + \bar{\nu}_2 \bar{x}_2 + \ldots + \bar{\nu}_n \bar{x}_n + \bar{x}_1^2 + O(\varepsilon), \end{cases}$$

after division by ε. Variable $\bar{x} = (\bar{x}_1, \dots, \bar{x}_n)$ can be assumed to belong to any arbitrarily large compact in $\mathbb{R}^n$.

Understanding the bifurcation diagram of the limit family ($\varepsilon = 0$) is essential to study the dynamics emerging from the singularity, that is, its unfolding. The limit family when $n = 2$ is a key piece in the study of the Bogdanov-Takens bifurcation ([3,20]). The limit family corresponding to the case $n = 3$ was studied in [11–14]. Among other results, it was proved in [14] that any generic unfolding of the 3-dimensional nilpotent singularity of codimension

3 exhibits strange attractors. Finally, the limit family corresponding to the 4-dimensional nilpotent singularity of codimension four

$$\begin{cases} \bar{x}_1' &= \bar{x}_2, \\ \bar{x}_2' &= \bar{x}_3, \\ \bar{x}_3' &= \bar{x}_4, \\ \bar{x}_4' &= \bar{\nu}_1 + \bar{\nu}_2\bar{x}_2 + \bar{\nu}_3\bar{x}_3 + \bar{\nu}_4\bar{x}_4 + \bar{x}_1^2, \end{cases} \tag{1.2}$$

was studied in [1, 2, 8]. Most notably, it was proved in [1] that any generic unfolding of the singularity contains a bifurcation hypersurface corresponding to bifocal homoclinic orbits. Even so, all the mentioned papers only offer very preliminary results. Consequently, the study of the dynamics exhibited by the limit family in the 4-dimensional case continues to be an interesting and enormous challenge.

In this work we delve into the study of (1.2). In Section 2 we propose directional rescalings that facilitate the study. The numerical methods employed along the paper are described in Section 3. The core of this paper is Section 4 where we provide results related to the existence of homoclinic connections. In addition, a first approximation to the complex structure of bifurcations of periodic orbits displayed in the family is presented in Section 5. We conclude with a brief discussion on related topics of interest.

2. Directional rescalings and reversible case

In what follows, we consider the family (1.2). When $\bar{\nu}_1 > 0$, it can be proven that the function

$$L(\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4) = \bar{x}_4 - \bar{\nu}_2\bar{x}_1 - \bar{\nu}_3\bar{x}_2 - \bar{\nu}_4\bar{x}_3$$

is strictly increasing along orbits. Therefore, there are no bounded orbits when $\bar{\nu}_1 > 0$ and the interesting dynamics only emerges for $\bar{\nu}_1 \leq 0$. When $\bar{\nu}_1 = 0$, there is a unique equilibrium point at the origin and for $\bar{\nu}_1 < 0$ there exist two equilibrium points at $p_\pm = (\pm\sqrt{-\bar{\nu}_1}, 0, 0, 0)$.

On the other hand, family (1.2) is invariant with respect to the transformation:

$$(\bar{\nu}_1, \bar{\nu}_2, \bar{\nu}_3, \bar{\nu}_4, x_1, x_2, x_3, x_4, t) \longmapsto (\bar{\nu}_1, -\bar{\nu}_2, \bar{\nu}_3, -\bar{\nu}_4, x_1, -x_2, x_3, -x_4, -t).$$

This allows to restrict the study to the case $\bar{\nu}_4 \leq 0$. In particular, the divergence of each vector field in the family is given by $\bar{\nu}_4$ so the existence of repellors is not feasible for $\bar{\nu}_4 \leqslant 0$ and attractors also do not exist when $\bar{\nu}_4 = 0$. Additionally, the vector fields are time-reversible when $\bar{\nu}_2 = \bar{\nu}_4 = 0$. As argued in [1, 2, 8], understanding the dynamics for the subfamily of time-reversible vector fields becomes essential. The dynamics of the linear part is simple around p_+ and richer around p_- (see [1]). The linear part at p_+ always have a pair of real eigenvalues and a pair of complex eigenvalues with non-zero real part. However, the linear part at p_- has

- a double zero eigenvalue and a pair of pure imaginary eigenvalues when $\bar{\nu}_3 = -1$ and $\bar{\nu}_1 = \bar{\nu}_2 = \bar{\nu}_4 = 0$ (we denote this bifurcation point as HBT),
- two double pure imaginary eigenvalues $\pm i(-\bar{\nu}_3/2)^{1/2}$ when $\bar{\nu}_3^2 - 8\sqrt{-\bar{\nu}_1} = 0$, $\bar{\nu}_3 < 0$ and $\bar{\nu}_2 = \bar{\nu}_4 = 0$ (we denote this bifurcation point as HH),
- two double real eigenvalues $\pm(\nu_3/2)^{1/2}$ when $\bar{\nu}_3^2 - 8\sqrt{-\bar{\nu}_1} = 0$, $\bar{\nu}_3 > 0$ and $\bar{\nu}_2 = \bar{\nu}_4 = 0$ (we denote this bifurcation point as BD),
- a double zero eigenvalue and eigenvalues ± 1 when $\bar{\nu}_3 = 1$ and $\bar{\nu}_1 = \bar{\nu}_2 = \bar{\nu}_4 = 0$ (we denote this bifurcation point as BT).

In between bifurcation points HBT and HH on the circumference $\bar{\nu}_1^2 + \bar{\nu}_3^2 = 1$, the linear part at p_- has four pure imaginary eigenvalues $\pm\omega_k i$, with $k = 1, 2$, and $\omega_1 \neq \omega_2$. For parameter values between bifurcation points HH and BD, it has four complex eigenvalues $\rho \pm \omega i$ and $-\rho \pm \omega i$ with non-zero real part ($\rho \neq 0$). Finally, in between bifurcation points BD and BT, all eigenvalues are real.

Remark 2.1 1. Since $\bar{\nu}_1 = 0$ at the bifurcation points HBT and BT, $p_\pm = (0, 0, 0, 0)$ is the only equilibrium.

2. Although the linearization at the origin has a double zero eigenvalue for the point BT, it is not a generic Bogdanov-Takens point because the vector field is conservative. In the same way, it occurs at the point HBT, where the linearization at the origin matches with a Hopf-Bogdanov-Takens point. Despite this, it should be notice that these bifurcations are generically unfolded in the original family.

3. As the item above suggests, the notation was chosen based on the type of linearization at the equilibrium point. In this sense, the linearization at p_- is related to a Hopf-Hopf bifurcation at the point HH and to a Belyakov-Devaney bifurcation at the point BD.

As usual, when dealing with limit families, it can be more convenient to consider directional rescalings. Namely, we can take $\bar{\nu}_i = +1$ (or $\bar{\nu}_i = -1$) and $(\bar{\nu}_1, \dots, \bar{\nu}_{i-1}, \bar{\nu}_{i+1}, \dots, \bar{\nu}_n) \in \mathbb{R}^{n-1}$ in (1.1). Bearing in mind the study of (1.2) close to the time-reversible subfamily, we consider a directional rescaling with $\bar{\nu}_1 = -1$ to get the family

$$\begin{cases} \bar{x}_1' = \bar{x}_2, \\ \bar{x}_2' = \bar{x}_3, \\ \bar{x}_3' = \bar{x}_4, \\ \bar{x}_4' = -1 + \bar{\nu}_2\bar{x}_2 + \bar{\nu}_3\bar{x}_3 + \bar{\nu}_4\bar{x}_4 + \bar{x}_1^2, \end{cases} \tag{2.1}$$

with $(\bar{\nu}_2, \bar{\nu}_3, \bar{\nu}_4) \in \mathbb{R}^3$.

Remark 2.2 To obtain a complete picture, directional rescalings with $\bar{\nu}_3 = \pm 1$ may be useful. This means to look at the limit family from the bifurcation points BT and HBT.

To compare with results previously obtained in the literature it is better to translate the equilibrium point $p_- = (-1, 0, 0, 0)$ to the origin and rescale variables and parameters as follows:

$$x_1 = \frac{\bar{x}_1 + 1}{2}, \quad x_2 = \frac{\bar{x}_2}{2^{5/4}}, \quad x_3 = \frac{\bar{x}_3}{2^{3/2}}, \quad x_4 = \frac{\bar{x}_4}{2^{7/4}}, \qquad \eta_2 = \frac{\bar{\nu}_2}{2^{3/4}}, \quad \eta_3 = \frac{\bar{\nu}_3}{2^{1/2}}, \quad \eta_4 = \frac{\bar{\nu}_4}{2^{1/4}},$$

to obtain the expression:

$$\begin{cases} x_1' = x_2, \\ x_2' = x_3, \\ x_3' = x_4, \\ x_4' = -x_1 + \eta_2 x_2 + \eta_3 x_3 + \eta_4 x_4 + x_1^2, \end{cases} \tag{2.2}$$

after division by $2^{1/4}$. As already explained, we only have to study the dynamics of system (2.2) around the origin varying $(\eta_2, \eta_3, \eta_4) \in \mathbb{R}^3$ with $\eta_4 \leqslant 0$.

In system (2.2), we first restrict parameters to the reversibility set:

$$\mathcal{T} = \left\{(\eta_2, \eta_3, \eta_4) \in \mathbb{R}^3 \mid \eta_2 = \eta_4 = 0\right\}$$

which, taking $u = x_1$ and $P = -\eta_3$, is equivalent to the fourth order ODE:

$$u^{(4)} + Pu'' + u - u^2 = 0. \tag{2.3}$$

This ODE arises, for instance, in applications to elasticity or fluid problems, and has been widely studied [4–6]. In this case, (2.3) can be expressed by means of Hamilton's equations with Hamiltonian [4]:

$$H = \frac{1}{2}x_1^2 - \frac{1}{3}x_1^3 - \frac{\eta_3}{2}x_2^2 + x_2x_4 - \frac{1}{2}x_3^2. \tag{2.4}$$

At the same time, (2.3) is a time-reversible system, with reversor:

$$R(x_1, x_2, x_3, x_4) = (x_1, -x_2, x_3, -x_4)$$

such that $R \circ \phi_t = \phi_{-t} \circ R$, being ϕ_t the flow associated to (2.3). When $P < 2$, the origin is an hyperbolic stationary solution, meanwhile it is non-hyperbolic for $P \geqslant 2$. The point $P = 2$ (respectively, $P = -2$) corresponds to the bifurcation point HH (respectively, BD).

We have reproduced numerically some findings of previous works concerning homoclinic orbits of (2.3). The set

$$\mathrm{Fix}(R) = \left\{(x_1, x_2, x_3, x_4) \in \mathbb{R}^4 \mid x_2 = x_4 = 0\right\}$$

plays an important role in their computation. Note that in the hyperbolic case ($P < 2$), when one of the invariant manifolds

$$W^s(0) = \left\{x \in \mathbb{R}^4 \mid \lim_{t\to\infty} \phi_t(x) = 0\right\} \quad \text{or} \quad W^u(0) = \left\{x \in \mathbb{R}^4 \mid \lim_{t\to-\infty} \phi_t(x) = 0\right\}$$

intersects transversally $\mathrm{Fix}(R)$ at x_0, the orbit through x_0 is homoclinic to the origin [6]. We use this property to locate homoclinic trajectories. We describe below the numerical procedures.

3. Numerical approach

3.1. Numerical integration

Equation (2.3), or more generally system (2.2), is integrated in time by means of Taylor method [16]. We apply it by restricting errors bellow 10^{-15}, which consequently impose the order of the Taylor polynomial used in every time step. In the reversible case (2.3), H in (2.4) is a conserved quantity. We use this fact as a test for the numerical integration. In addition, homoclinic orbits belong to the set of zero energy, $\{H = 0\}$, since $H(0) = 0$.

Another important fact is the selection of a proper Poincaré section, in order to classify orbits. We choose $\Sigma = \{x_2 = 0\}$ as the main Poincaré section in our computations.

3.2. Invariant manifolds approximation

In the reversible equation (2.3), $R(W^s(0)) = W^u(0)$ and $\dim(W^u(0)) = \dim(W^s(0)) = 2$ hold for $P < 2$ (see [6]). In the general system (2.2), we still have $\dim(W^u(0)) = \dim(W^s(0)) = 2$ for (η_2, η_3, η_4) close to the line $\{(0, \eta_3, 0) \mid \eta_3 > -2\}$. For this reason, we can consider each of the invariant manifolds, say W, expressed as follows

$$W = \left\{(x_1, x_2, x_3, x_4) \in \mathbb{R}^4 \mid x_3 = a(x_1, x_2), x_4 = b(x_1, x_2)\right\}$$

for certain unknown functions, a and b, smooth enough. We use the Taylor's expansion in power series around the origin:

$$x_3 = \sum_{M=1}^{\infty} \sum_{i=0}^{M} a_{M-i,i} x_1^{M-i} x_2^i \qquad \text{and} \qquad x_4 = \sum_{M=1}^{\infty} \sum_{i=0}^{M} b_{M-i,i} x_1^{M-i} x_2^i \tag{3.1}$$

where $a_{M-i,i}, b_{M-i,i}$ are coefficients corresponding to degree M, to be determined. As W is an invariant manifold, we can impose (3.1) to satisfy system (2.2). With standard but lengthy computations, we obtain a $2(M+1)\times 2(M+1)$ system in the unknowns $a_{M-s,s}, b_{M-s,s}$, for each $M = 1, 2, \ldots$ and $s = 0, 1, \ldots, M$:

$$\begin{cases} \sum_{k=1}^{M} a_{k-1,1} a_{M-k+1,0} = b_{M,0} & \\ (M-s+1) a_{M-s+1,s-1} + c_{M-s,s} = b_{M-s,s} & s = 1, \ldots, M \\ \sum_{k=1}^{M} b_{k-1,1} a_{M-k+1,0} = \eta_3 a_{M,0} + \eta_4 b_{M,0} - \delta_{M1} + \delta_{M2} & \\ (M-s+1) b_{M-s+1,s-1} + d_{M-s,s} = \eta_3 a_{M-s,s} + \eta_4 b_{M-s,s} + \eta_2 \delta_{Ms1} & s = 1, \ldots, M \end{cases} \tag{3.2}$$

with

$$c_{M-s,s} = \sum_{k=1}^{M} \sum_{i=l_k}^{L_k} i a_{k-i,i} a_{M-k+i-s,s+1-i} \quad \text{and} \quad d_{M-s,s} = \sum_{k=1}^{M} \sum_{i=l_k}^{L_k} i b_{k-i,i} a_{M-k+i-s,s+1-i}$$

where $l_k = \max\{s + k - M, 1\}$, $L_k = \min\{s + 1, k\}$ and $\delta_{xy} = \delta_{xyz} = 1$ only when $x = y = z$ but 0 otherwise. Those systems are solved in increasing order for $M = 1, 2, \ldots$. Once we compute $a_{M-s,s}$ and $b_{M-s,s}$, we can only evaluate (3.1) on a certain disk of convergence centered at $(0, 0)$, in the plane defined by (x_1, x_2). Accordingly, in order to approximate a point in an invariant manifold W, we fix (x_1, x_2) not very distant to $(0, 0)$. Using the series (3.1) up to a certain order M, we finally find x_3 and x_4 such that $(x_1, x_2, x_3, x_4) \in W$, up to the truncation error.

4. Time-reversible case

In this section we restrict our study to the time-reversible case, i.e. $\eta_2 = \eta_4 = 0, \eta_3 \in \mathbb{R}$ in (2.2). We apply the tools described above, namely: numerical integrator for system (2.2) and approximation of the invariant manifolds W at the origin. Because W is invariant by the flow ϕ_t, an orbit Γ is included in W, provided there exists $x_0 \in \Gamma \cap W$. By means of the Taylor series (3.1), we approximate $x_0 = (x_1^0, x_2^0, a(x_1^0, x_2^0), b(x_1^0, x_2^0)) \in W$ (with a and b defined in the above section) and, using the numerical integrator, we estimate $\Gamma(x_0)$. In Figure 1, we represent different orbits $\Gamma(x_0(\theta))$ for

$$x_0(\theta) = (x_1, x_2, a(x_1, x_2), b(x_1, x_2)), \quad x_1 = r\cos\theta, \quad x_2 = r\sin\theta, \quad \theta \in [0, 2\pi), \quad \text{and} \quad r = 1/10. \tag{4.1}$$

The value of r is chosen so that series (3.1) are convergent. In fact, the system (3.2) for $M = 1$ gives rise to two solutions corresponding, respectively, to the stable and unstable manifolds.

Homoclinic solutions for ODE (2.3) were analyzed in [4–6]. As stated in the final part of §2, homoclinic orbits corresponds to trajectories with a point in $\mathrm{Fix}(R) \cap W$. In order to find an orbit $\Gamma \subset W$ such that $\Gamma \cap \mathrm{Fix}(R) \neq \emptyset$, we consider initial conditions $x_0(\theta)$ as in (4.1). First, we find $t = t(\theta)$ so that $\phi_t(x_0(\theta)) = (x_1, x_2, x_3, x_4)(\theta, t)$ crosses the Poincaré section $\Sigma = \{x_2 = 0\}$ for a fixed number of times, k. If $x_4(\theta, t) = 0$, then $\phi_t(x_0(\theta)) \in \mathrm{Fix}(R)$ and the orbit through $x_0(\theta)$ is homoclinic. Otherwise, we apply a secant method on θ to vanish $x_4(\theta, t(\theta))$. In Figure 2, we

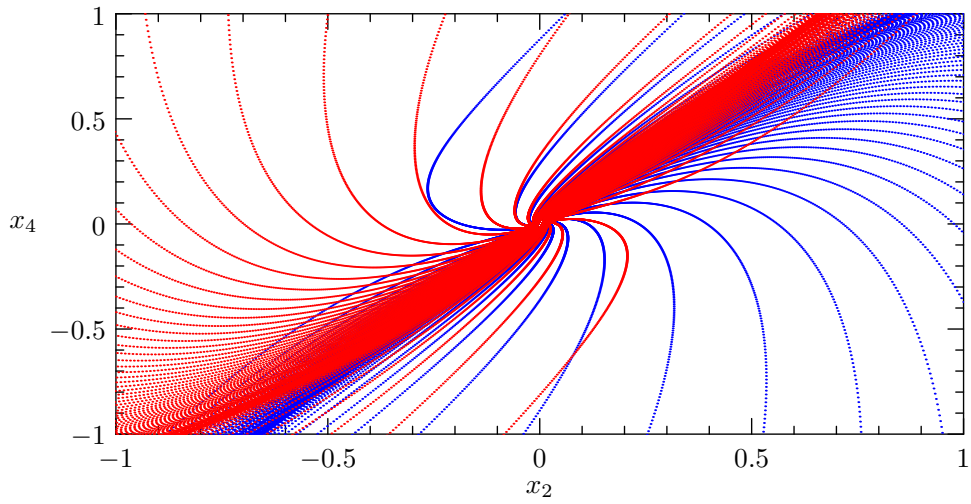


Fig. 1 Different orbits in family (2.2) which make up the stable (blue) and unstable (red) invariant manifolds, close to the origin for $\eta_3 = 1.8$ and projected on the plane (x_2, x_4).

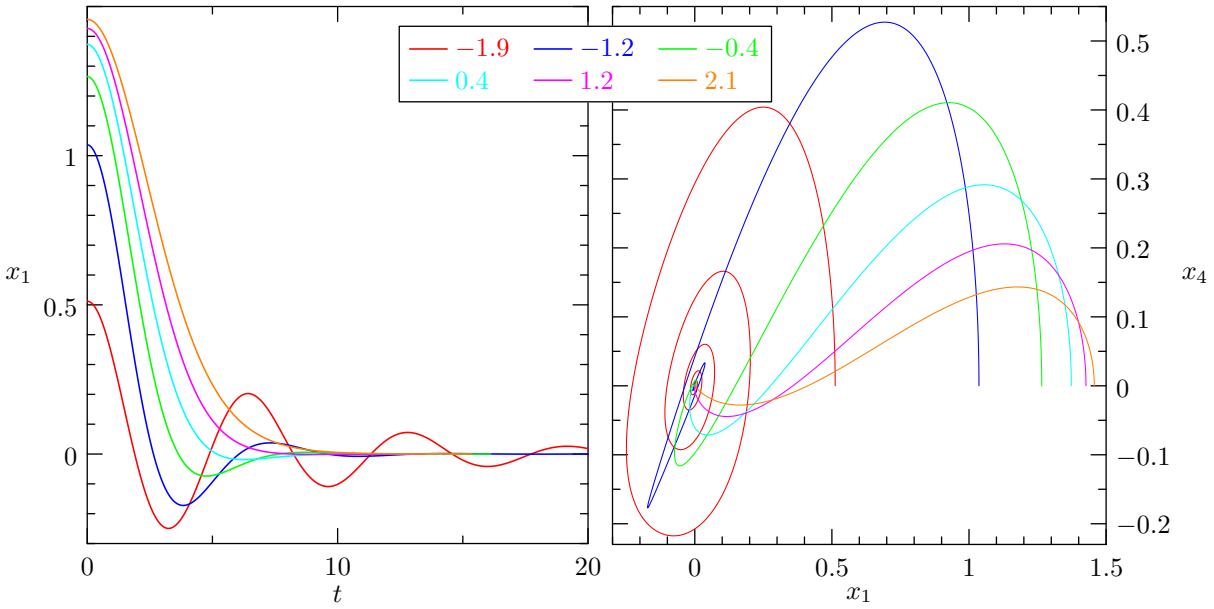


Fig. 2 Homoclinic orbits in family (2.2) for values of η_3 specified by colors. Each orbit starts at $\mathrm{Fix}(R)$ for $t = 0$. Half of the orbit is missing by symmetry. For a given color, different coordinates of the same orbit are shown on the left and right plots. Points in the vertical axis ($t = 0$) on the left panel are in correspondence with points on $x_4 = 0$ on the right panel.

present different homoclinic orbits varying $\eta_3 \in [-1.9, 2.1]$, for $k = 1$. Since we are in the time-reversible case, we have $R \circ \phi_t = \phi_{-t} \circ R$. If the initial condition $\bar{x}_0 \in \mathrm{Fix}(R)$, then:

$$R\phi_t(\bar{x}_0) = \phi_{-t}(R(\bar{x}_0)) = \phi_{-t}(\bar{x}_0)$$

and the orbit is R-symmetric. For this reason, we only plot values for $t \geqslant 0$ in Figure 2.

To improve the plot of the invariant manifolds in Figure 1, we present the curve $W^u(0) \cap \Sigma$ in Figure 3 (left). Each intersection of this curve with $\mathrm{Fix}(R)$ (in red) leads to a homoclinic orbit, which is represented in Figure 3 (right). Particularly, the 11 depicted homoclinic orbits cross Σ a variable number $k = 1, \ldots, 5$ of times. The findings for these homoclinics may not be exhaustive, but they give an idea of the dynamics complexity.

All the homoclinic orbits in the time-reversible case belong to the hypersurface $\{H = 0\}$. Taking initial conditions $x_0 \notin \{H = 0\}$, it is not difficult to meet another kind of invariant orbits. For instance, we obtain the solutions plotted in Figure 4 which correspond to invariant tori.

5. Insights in the non-zero divergence case

In the above section, we have analyzed numerically part of the homoclinic phenomena arising in family (2.2) when $\eta_2 = \eta_4 = 0$. In particular, we used the Taylor integrator and the expansion of the invariant manifolds that we

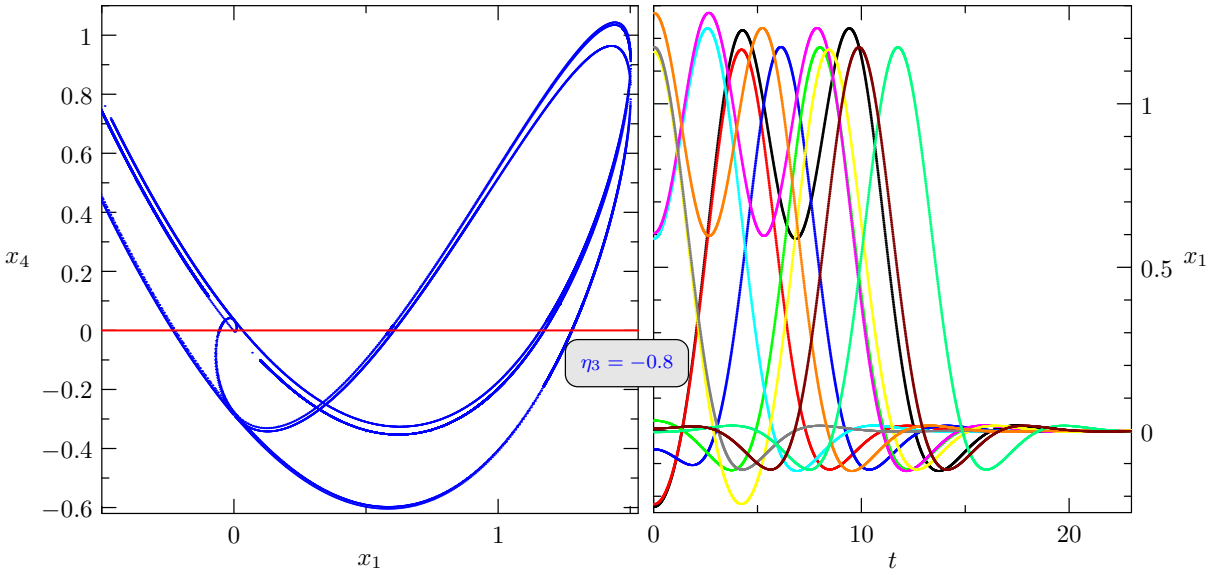


Fig. 3 Invariant manifolds and homoclinic orbits in family (2.2). Left: Curve $W^u(0) \cap \Sigma$ projected on the plane (x_1, x_4). The red line depicts $\mathrm{Fix}(R)$. Right: 11 different homoclinic orbits starting at the respective initial conditions on $W^u(0) \cap \mathrm{Fix}(R)$, located in the left figure as crosses with $\mathrm{Fix}(R)$.

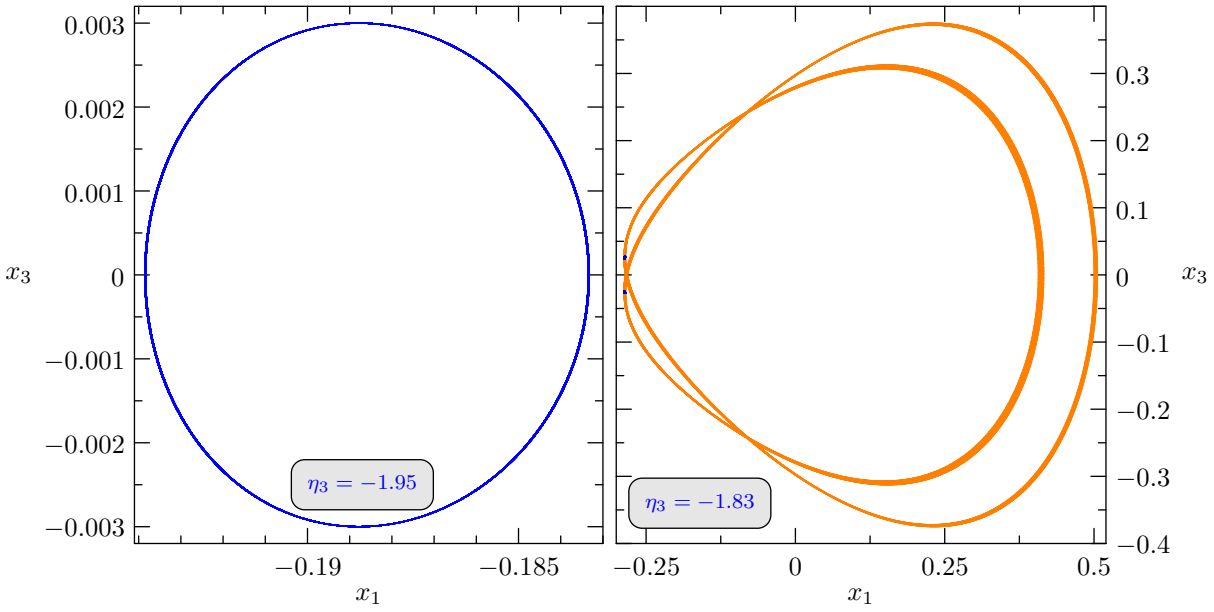


Fig. 4 Quasiperiodic orbits close to periodic, represented on the (x_1, x_3) plane, for family (2.2). Left: Each point of the orbit is only plotted when it crosses Σ. Right: The whole orbit at discrete time values is represented. As a reference, Σ is likewise in this plot, traced on its left side as two little blue curves.

designed specifically for this model. To study the bifurcation diagram around that axis, we can also use numerical continuation methods.

A first overview of the bifurcation complexity is shown in the two bifurcation diagrams in Figure 5, that we obtained using MATCONT [7]. For this analysis, we fix $\bar{\nu}_3 = -3$ and $\bar{\nu}_2 = -0.6$ in family (2.1) and find a Hopf bifurcation at $p_- = (-1, 0, 0, 0)$ when $\bar{\nu}_4 = -0.3$. The continuation of the limit cycle arising at p_- is shown in Figure 5 (left panel, at the bottom). First, an attracting limit cycle emerges from the Hopf bifurcation and loses its stability at a period doubling bifurcation. The periodic orbit recovers stability through another period doubling bifurcation, but loses stability again at a Neimark-Sacker bifurcation where an attracting invariant torus emerges. Finally, the limit cycle disappears at a Hopf bifurcation which occurs at the other equilibrium point $p_+ = (+1, 0, 0, 0)$.

The Hopf bifurcation curve occurring at p_- and the continuation of the Neimark-Sacker bifurcation are represented in Figure 5 (right). Above all, both period doubling bifurcation points belong to the same bifurcation curve, as depicted in Figure 5 (right) where we show a solid red line consisting of two loops.

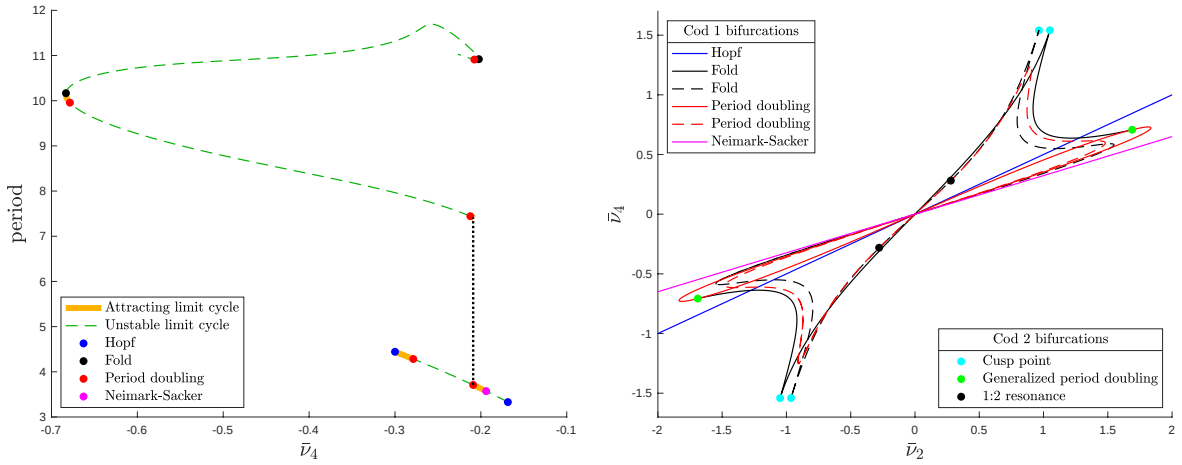


Fig. 5 Left: Continuation of periodic orbits in family (2.1) with $\bar{\nu}_3 = -3$ and $\bar{\nu}_2 = -0.6$. On the one hand, continuation of a periodic orbit emerging from a Hopf bifurcation when $\bar{\nu}_4 = -0.3$ (at the bottom). On the other hand, continuation of a periodic orbit emerging at a period doubling bifurcation (at the top). Right: Partial bifurcation diagram of family (2.1) with $\bar{\nu}_3 = -3$ fixed.

In Figure 5 (left panel, at the top), we also show the continuation of the limit cycle with doubled period that emerges from the period doubling bifurcation point placed on the right side of the continuation curve at the bottom. The attracting limit cycle loses its stability almost immediately due to a period doubling bifurcation. Along the curve we see two fold bifurcation points (black) which belong to the double loop bifurcation curve displayed in Figure 5 (right panel, dashed black line). The limit cycle in between the fold points and the period doubling bifurcation points closer to them is an attractor. These two period doubling points belong to the double loop shown in Figure 5 (right panel, dashed red line).

Additionally, Figure 5 (right) includes a fold bifurcation curve that joins two generalized period doubling bifurcation points (green). Other codimension-two points are the cusp bifurcations of periodic orbits (cyan) and the two point of resonance $1:2$ (black). A description of the bifurcations mentioned can be found in [17].

6. Discussion

In conclusion, there exist thorough studies [4–6] regarding the complex tangle of homoclinic orbits exhibited by system (2.2) when $\eta_2 = \eta_4 = 0$. Nevertheless, an exhaustive picture is not yet available (see conjectures in [5]). Numerical techniques, which we take advantage of to explore the intersections of the invariant manifolds with a transverse section, are tools that, perhaps for technical reasons, have not been fully exploited in the literature. In this case, despite the fact that the scenario is quite different, our numerical study revives an old paper [18] where the heteroclinic connections unfolded in a reversible three-dimensional system with two equilibrium points of saddle-focus type and different stability indices were studied. Ultimately, the analysis of the traces left by invariant manifolds in a cross section is our most immediate interest. Techniques used in [15] to study Poincaré return maps around a homoclinic orbit to bifocus equilibrium will be useful to describe the geometry of such intersections.

Beyond the homoclinic framework, it is fundamental to analyze the conservative dynamics. In particular, the one that emerges in family (2.2) when $\eta_2 = \eta_4 = 0$ and $\eta_3 < -2$, that is, when the equilibrium point $p_- = (-1, 0, 0, 0)$ is a Hopf-Hopf singularity. In this context, it is also crucial to delve into the dynamics of the family around the point HBT as well as in the surroundings of the point HH. However, these are longer-term goals. In fact, the study of generic unfoldings of Hopf-Bogdanov-Takens singularities has started very recently [9, 10]. Furthermover, we must recall that the limit family is not a generic unfolding of the HBT singularity. The process of reaching a complete theoretical support seems too involved and long. Therefore, all the information that we can collect with continuation tools such as those illustrated in this study will be very useful.

Acknowledgements

The authors has been partially supported by the project MINECO-18-MTM2017-87697-P.

References

[1] P. G. Barrientos, S. Ibáñez and J. A. Rodríguez. Heteroclinic cycles arising in generic unfoldings of nilpotent singularities. *Journal of Dynamics and Differential Equations*, 23(4):999–1028, 2011.

[2] P. G. Barrientos, S. Ibáñez and J. A. Rodríguez. Robust cycles unfolding from conservative bifocal homoclinic orbits. *Dynamical Systems*, 31(4):546–579, 2016.

[3] R. I. Bogdanov. The versal deformation of a singular point of a vector field on the plane in the case of zero eigenvalues. *Trudy Sem. Petrovsky*, 2:37–65, 1976.

[4] B. Buffoni, A. R. Champneys and J. F. Toland. Bifurcation and coalescence of a plethora of homoclinic orbits for a hamiltonian system. *Journal of Dynamics and Differential Equations*, 8(2):221–279, 1996.

[5] A. R. Champneys and J. F. Toland. Bifurcation of a plethora of multi-modal homoclinic orbits for autonomous hamiltonian systems. *Nonlinearity*, 6(5):665–721, 1993.

[6] A. R. Champneys and A. Spence. Hunting for homoclinic orbits in reversible systems: a shooting technique. *Advances in Computational Mathematics*, 1(1):81–108, 1993.

[7] A. Dhooge, W. Govaerts, Yu. A. Kuznetsov, H. G. E. Meijer and B. Sautois. New features of the software MatCont for bifurcation analysis of dynamical systems. *Mathematical and Computer Modelling of Dynamical Systems*, 14(2):147–175, 2008.

[8] F. Drubi, S. Ibáñez and J. A. Rodríguez. Coupling leads to chaos. *Journal of Differential Equations*, 239(2):371–385, 2007.

[9] F. Drubi, S. Ibáñez and D. Rivela. A formal classification of Hopf-Bogdanov-Takens singularities of codimension three. *Journal of Mathematical Analysis and Applications*, 480(2):123408, 2019.

[10] F. Drubi, S. Ibáñez and D. Rivela. Chaotic behavior in the unfolding of Hopf-Bogdanov-Takens singularities. *Discrete and Continuous Dynamical Systems - Series B*, 25(2):599–615, 2020.

[11] F. Dumortier, and S. Ibáñez. Nilpotent singularities in generic 4-parameter families of 3- dimensional vector fields. *Journal of Differential Equations*, 127(2):590–647, 1996.

[12] F. Dumortier, S. Ibáñez and H. Kokubu. New aspects in the unfolding of the nilpotent singularity of codimension three. *Dynamical Systems*, 16(1):63–95, 2001.

[13] F. Dumortier, S. Ibáñez and H. Kokubu. Cocoon bifurcation in three-dimensional reversible vector fields *Nonlinearity*, 19(2):305–328, 2006.

[14] S. Ibáñez and J. A. Rodríguez. Shil'nikov configurations in any generic unfolding of the nilpotent singularity of codimension three on $\mathbb{R}^3$. *Journal of Differential Equations*, 208(1):147–175, 2005.

[15] S. Ibáñez and A. Rodrigues. On the dynamics near a homoclinic network to a bifocus: switching and horseshoes. *International Journal of Bifurcation and Chaos*, 25(11):1530030, 2015.

[16] A. Jorba and M. Zou. A software package for the numerical integration of odes by means of high-order taylor methods. *Experimental Mathematics*, 14(1):99–117, 2005.

[17] Y. Kuznetsov. Elements of Applied Bifurcation Theory. *Applied Mathematical Sciences*, 112, Springer-Verlag New York 2004.

[18] Y.-T. Lau. The cocoon bifurcation in three.dimensional systems with two fixed points. *International Journal of Bifurcation and Chaos*, 2(3):543–558, 1992.

[19] F. Takens. Singularities of vector fields. *Publications mathématiques de l'I.H.É.S*, 43:47–100, 1974.

[20] F. Takens. Forced oscillations and bifurcations. In Applications of global analysis I, *Comm. Math. Inst. Rijksuniv. Utrecht*, 3:1–59, 1974.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 130–134)

Different Approximations of the Parameter for Low-Order Iterative Methods with Memory

Francisco I. Chicharro[1], **Neus Garrido**[1], **Íñigo Sarría**[1], **Lara Orcos**[2]
1. *Escuela Superior de Ingeniería y Tecnología, Universidad Internacional de La Rioja, Spain*
2. *Facultad de Educación, Universidad Internacional de La Rioja, Spain*

Abstract

A technique for generating iterative methods for solving nonlinear equations with memory can be constructed from a method without memory that includes a parameter, provided the parameter is present in the error equation.

Generally, the parameter depends on the evaluation of the function and its derivatives in the solution. However, this information is not available. So this parameter is approximated using interpolation techniques, taking the current iterate x_k and the previous iterates $x_{k-1}, x_{k-2}, \ldots$

In this paper we explore different interpolation techniques to obtain both the convergence order of the new methods and their stability characteristics.

1. Introduction

Many phenomena in applied sciences do not respond to a linear pattern. Nonlinearities are present in most fields, such as physics, fluid mechanics, economics or ecology, among others. In this case, these phenomena can be modeled by means of a nonlinear equation $f(x) = 0, f : I \subseteq \mathbb{R} \to \mathbb{R}$, or by means of a system of nonlinear equations $F(x) = 0, F : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$. The desired solution x^* of these problems is a closed-form analytic expression. However, there are problems whose analytic solution is hardly available. Obtaining approximate solutions becomes an alternative, by applying numerical methods based on iterative algorithms.

Numerical methods for solving nonlinear equations can be sorted by different criteria. Single-step methods respond to the scheme $x_{k+1} = \phi(x_k)$, while multi-step methods are those that match with $y_k = \phi_1(x_k), x_{k+1} = \phi_2(x_k, y_k)$. A quantitative comparison between methods can be perfomed by the order of convergence p and the efficiency index [17] $I = p^{1/d}$, where d stands for the number of functional evaluations in each step. Kung-Traub's conjecture [15] states that there exists an upper bound for the order of convergence that is $p \leq 2^{d-1}$; thus, the iterative method is optimal when $p = 2^{d-1}$. There is an interesting overview of these methods in [12].

Kung-Traub's conjecture sets an upper bound for the order of convergence in numerical methods without memory. However, this restriction can be overcome by using iterative methods with memory. These kind of methods are defined as

$$x_{k+1} = \phi(x_k, x_{k-1}, \ldots, x_{k-m}).$$

In other words, the current iterate is calculated taking into account the last $m + 1$ iterates. This idea was introduced by Traub [23], including memory from Steffensen's method. In the last years, many schemes of iterative methods with memory have been presented. A key overview can be found in [18, 19].

One technique for the design of a method with memory consists of the inclusion of an accelerating parameter in the expression of a method without memory. This technique has been widely adopted in the research of this kind of methods for both nonlinear equations [5, 6, 10], and nonlinear systems of equations [7, 16, 20].

Once the parameter has been included in the iterative expression, the next step is the analysis of the error equation. When the parameter is present in the lower term of this equation, the goal is the replacement of the parameter by an expression that cancels this error term. There are different techiques for the approximation of the parameter.

In this paper, we analyze the most common techniques of replacing the parameter, as well as other novel techniques. In [4] the authors introduced the general form of one-step iterative methods using the weight function technique given by

$$x_{k+1} = x_k - H(t_k), \quad k = 0, 1, 2, \ldots, \tag{1.1}$$

where $t_k = f(x_k)/f'(x_k)$. Family (1.1) has quadratic convergence when $H(t)$ satisfies $H(0) = 0$, $H'(0) = 1$ and $|H''(0)| < \infty$. The error equation of members of family (1.1) is

$$e_{k+1} = \left(c_2 - \frac{H''(0)}{2}\right) e_k^2 + O(e_k^3), \tag{1.2}$$

where $e_k = x_k - x^*$ and $c_j = \frac{f^{(j)}(x^*)}{j!f'(x^*)}$, $j \geq 2$. Note that $H(t) = t + \alpha\frac{t^2}{2}$ satisfies the conditions of quadratic convergence of (1.1) for $H(t)$, resulting in

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} - \alpha\frac{f^2(x_k)}{2(f'(x_k))^2}, \tag{1.3}$$

and its error equation is

$$e_{k+1} = \left(c_2 - \frac{\alpha}{2}\right)e_k^2 + O(e_k^3). \tag{1.4}$$

For $\alpha = 2c_2$, the second order error term vanishes. However, the value of $c_2 = \frac{f''(x^*)}{2f'(x^*)}$ is not known. Therefore, some approximations of $f'(x^*)$ and $f''(x^*)$ must be applied.

2. The approximations of f and the convergence analysis

In order to obtain an approximation of f, we compare the approximation of different interpolatory structures. The most of papers apply Newton's interpolation polynomial of different degrees [11, 14, 24]. Let us denote by $N(t)$ the interpolation polynomial of Newton of second degree, whose expression is

$$N(t) = f(x_k) + f[x_{k-1}, x_k](t - x_k) + f[x_{k-2}, x_{k-1}, x_k](t - x_k)(t - x_{k-1}), \tag{2.1}$$

where $f[\cdot,\cdot]$ and $f[\cdot,\cdot,\cdot]$ are the divided differences of orders one and two. The lower degree of the polynomial in order to avoid that $N''(t)$ vanishes is two. Approximating

$$\left\{\begin{array}{rcl} f'(x^*) & = & f'(x_k), \\ f''(x^*) & = & N''(x_k), \end{array}\right.$$

the value of the parameter is

$$\alpha_k = 2\frac{f[x_{k-2}, x_{k-1}, x_k]}{f'(x_k)}. \tag{2.2}$$

Then, parameter α_k is replaced in (1.3), resulting in an iterative method with memory. Note that this method requires the knowledge of three previous iterates and two new functional evaluations.

The Taylor expansion of a function can also give an approximation for the value of α. From the regressive Taylor expansion at node x_{k-1} of order $O((x_{k-1} - x_k)^2)$ the parameter can be approximated by

$$\alpha_k = \frac{2}{(x_{k-1} - x_k)^2}\left(\frac{f(x_{k-1}) - f(x_k)}{f'(x_k)} - (x_{k-1} - x_k)\right). \tag{2.3}$$

In this case, the method requires the value of the two last iterates, and three evaluations of f.

Another option for the approximation of the parameter is the use od Padé's approximant. It has been applied for solving nonlinear equations [9, 21], but –up to our knowledge– it has not been used for methods with memory. Let $P(t)$ be the Padé's approximant

$$P(t) = \frac{a_0 + a_1(t - x_k)}{1 + a_2(t - x_k)}. \tag{2.4}$$

The values of a_0, a_1 and a_2 can be obtained when (2.4) satisfies

$$\left\{\begin{array}{rcl} P(x_k) & = & f(x_k), \\ P(x_{k-1}) & = & f(x_{k-1}), \\ P'(x_k) & = & f'(x_k). \end{array}\right.$$

The approximation of the parameter in this case has the expression

$$\alpha_k = \frac{P''(x_k)}{f'(x_k)} = 2\frac{f'(x_k)\,(f(x_{k-1} - f(x_k) + f'(x_k)(x_k - x_{k-1}))}{(f(x_k) - f(x_{k-1})\,(x_k - x_{k-1})}. \tag{2.5}$$

The resulting method only requires two iterates for the approximation of the parameter and three functional evaluations.

Theorem 2.1 gathers the analysis of the R-order of convergence of the previous methods.

Theorem 2.1 *Let x^* be a simple zero of a sufficiently differentiable function $f : I \subseteq \mathbb{R} \to \mathbb{R}$ in an open interval I. If x_0 is close enough to x^* and α_0 is given, then the R-orders of method* (1.3) *replacing α_k by expressions* (2.2), (2.3) *and* (2.5) *are* $1 + \sqrt{2}$.

Table 1 collects the comparison of the main values of each technique.

Let us remark from Table 1 that every method has the same order of convergence, while the number of functional evaluations is lower for Taylor and Padé's approximant.

Technique	Newton	Taylor	Padé
Iterates	3	2	2
d	4	3	3
p	$1+\sqrt{2}$	$1+\sqrt{2}$	$1+\sqrt{2}$

Tab. 1 Quantitative comparison of the parameter approximation

3. Real multidimensional dynamical analysis

The dynamics of an iterative method analyses their stability in terms of the amount of initial guesses that converge to the expected solution. Some fundamentals about dynamics of iterative methods without memory can be found in [1, 13], while for the iterative methods with memory the basics are in [2, 3].

The fixed points, for real multidimensional dynamics, involves the definition of an auxilary function $G : \mathbb{R}^2 \to \mathbb{R}^2$ such that

$$G(z, x) = (x, g(z, x)),$$

where $g : \mathbb{R}^2 \to \mathbb{R}$ is the iterative expression $x_{k+1} = g(x_{k-1}, x_k)$, $z = x_{k-1}$ and $x = x_k$. Therefore, the fixed points are defined as $G(z_F, x_F) = (z_F, x_F)$. Fixed points that does not match with the roots of f are named strange fixed points. They affect the unstability of the method. A T-periodic point is defined as $G^T(z_T, x_T) = G(z_T, x_T)$, satisfying $G^t(z_T, x_T) \neq (z_T, x_T), t < T$; note that for $T = 1$, the periodic point is a fixed point. The asymptotical behavior of T-periodic points is defined in [22]. Theorem 3.1 collects the asymptotical behavior for $T = 1$.

Theorem 3.1 *Let $G : \mathbb{R}^2 \to \mathbb{R}^2$ be C^2. Let μ_1, μ_2 be the eigenvalues of the Jacobian matrix G' on a fixed point (z_F, x_F). Then*

1. *If $|\mu_1| < 1$ and $|\mu_2| < 1$, then (z_F, x_F) is attracting.*
2. *If $|\mu_1| > 1$ and $|\mu_2| > 1$, then (z_F, x_F) is repelling.*
3. *If $|\mu_1| < 1$ and $|\mu_2| > 1$, or $|\mu_1| > 1$ and $|\mu_2| < 1$, then (z_F, x_F) is unstable.*

The attracting fixed points are denoted by (z^+, x^+). The basin of attraction of an attracting fixed point $\mathcal{A}(z^+, x^+)$ is the set of points that satisfy

$$\mathcal{A}(z^+, x^+) = \{(z, x) \in \mathbb{R}^2 : G^n(z, x) \to (z^+, x^+), n \to \infty\}.$$

The dynamical analysis is performed applying the expressions of α on (1.3) for the solution of $f(x) = x^2 - \lambda$.

In order to make a reasonable comparison, we are analysing the resulting methods of Taylor's and Padé's approximations of α. Note that these methods only require the two last iterates, while Newton's approximation requires three previuos iterates.

The comparison is performed via the representation of the basins of attraction, in a similar manner as described in [8]. In this particular case, the basins of $(z^+, x^+) = \sqrt{\lambda}(1, 1)$ are represented in orange, the basins of $(z^+, x^+) = -\sqrt{\lambda}(1, 1)$ are represented in blue, and the convergence to a different point than $(z^+, x^*) = \pm\sqrt{\lambda}(1, 1)$ is represented in black. The fixed attracting points are represented with white stars.

3.1. Taylor's approximation

Replacing (2.3) in (1.3), the auxiliary function is

$$T(z, x) = \left(x, \frac{3x^4 + 6x^2\lambda - \lambda^2}{8x^3}\right).$$

There are two fixed attracting points $(z^+, x^+) = \pm\sqrt{\lambda}(1, 1)$ and two unstable points $(z, x) = \pm\sqrt{\frac{\lambda}{5}}(1, 1)$.

Figure 1 represents the basins of attraction of $T(z, x)$ for different values of λ. Since $T(z, x)$ does not have dependence on the value of $z = x_{k-1}$, the dynamical planes are vertical bands. Note that every initial guess converge to an attracting fixed point, and bands are wider as the value of λ increases.

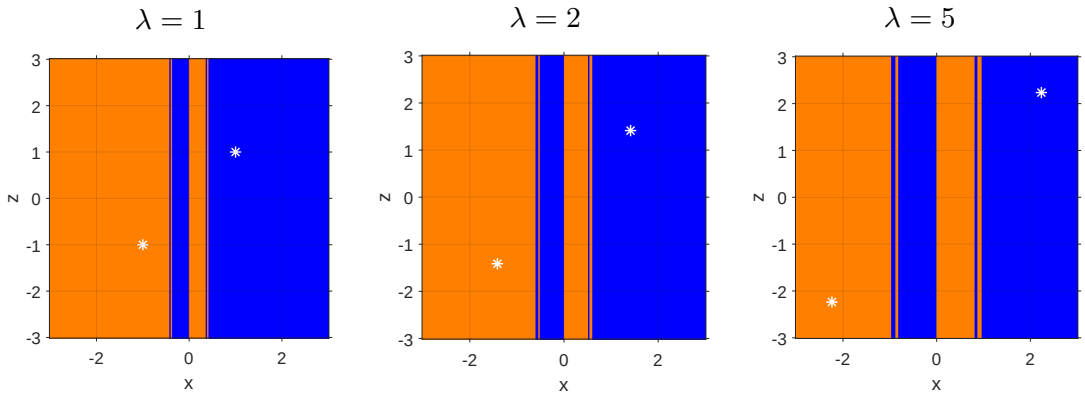


Fig. 1 Dynamical planes using Taylor's approximation of α

3.2. Padé's approximation

Replacing (2.5) in (1.3), the auxiliary function is

$$P(z,x) = \left(x, \frac{x^2 - \frac{(x^2-\lambda)^2}{x+z} + \lambda}{2x} \right).$$

There are two fixed attracting points $(z^+, x^+) = \pm\sqrt{\lambda}(1,1)$ and two unstable points $(z,x) = \left(-1 \pm \sqrt{1+\lambda}\right)(1,1)$.

Figure 2 represents the basins of attraction of $P(z,x)$ for different values of λ. In this case, $P(z,x)$ depends on both $z = x_{k-1}$ and $x = x_k$, so dynamical planes are not vertical bands. There are regions of convergence to the roots of f, but there are other regions that diverge or converge to another point, as black areas represent. Moreover, as λ increases, the width of black central region also does.

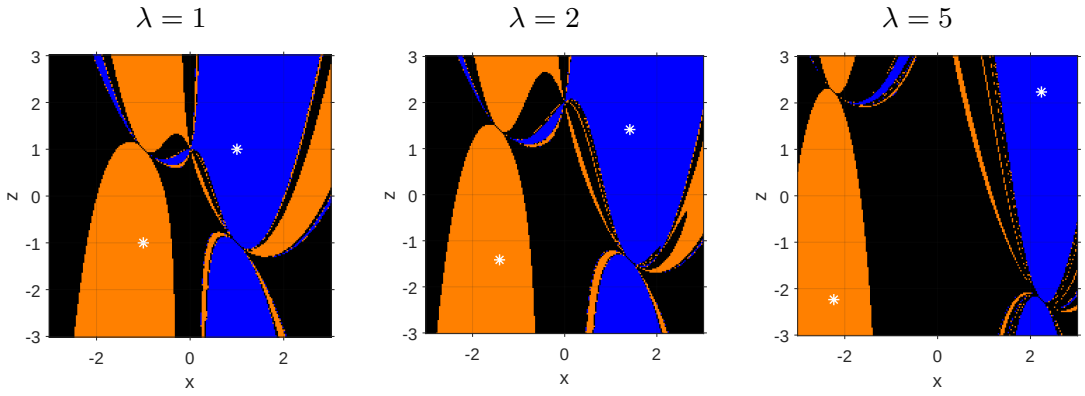


Fig. 2 Dynamical planes using Padé's approximation of α

4. Conclusions

Three new techniques have been introduced for the approximation of the self-accelerating parameter in a low-order iterative method. The order of convergence for the three cases have increased from 2 to $1 + \sqrt{2}$. In order to make a reasonable comparison for the stability counterpart, two approximations that involve the same number of previous iterates have been taken. Taylor's approximation results in vertical dynamical planes, because of the independence of $T(z,x)$ with z. In addition, Taylor's approximation results in more stable dynamical planes than Padé's approximation.

Acknowledgements

The authors were supported by the internal research project ADMIREN of Universidad Internacional de La Rioja (UNIR). The first author was also partially supported by PGC2018-095896-B-C22 (MCIU/AEI/FEDER, UE).

References

[1] P. Blanchard. Complex analytic dynamics on the riemann sphere. *Bulletin of the American Mathematical Society*, 11:85–141, 1984.

[2] B. Campos, A. Cordero, J. R. Torregrosa, and P. Vindel. A multidimensional dynamical approach to iterative methods with memory. *Applied Mathematics and Computation*, 271:701–715, 2015.

[3] B. Campos, A. Cordero, J. R. Torregrosa, and P. Vindel. Stability of king's family of iterative methods with memory. *Journal of Computational and Applied Mathematics*, 318:504–514, 2017.

[4] F. I. Chicharro, A. Cordero, N. Garrido, and J. R. Torregrosa. Generating root-finder iterative methods of second order: Convergence and stability. *Axioms*, 8(2):55, 2019.

[5] F. I. Chicharro, A. Cordero, N. Garrido, and J. R. Torregrosa. Anomalies in the convergence of traub-type methods with memory. *Computational and Mathematical Methods*, 2:e1060, 2020.

[6] F. I. Chicharro, A. Cordero, N. Garrido, and J. R. Torregrosa. On the choice of the best members of the kim family and the improvement of its convergence. *Mathematical Methods in the Applied Sciences*, 43:8051–8066, 2020.

[7] F. I. Chicharro, A. Cordero, N. Garrido, and J. R. Torregrosa. On the improvement of the order of convergence of iterative methods for solving nonlinear systems by means of memory. *Applied Mathematics Letters*, 104:106277, 2020.

[8] F. I. Chicharro, A. Cordero, and J. R. Torregrosa. Drawing dynamical and parameters planes of iterative families and methods. *The Scientific World Journal*, 2013:780153, 2013.

[9] F. I. Chicharro, A. Cordero, and J. R. Torregrosa. Dynamics and fractal dimension of steffensen-type methods. *Algorithms*, 8:271–279, 2015.

[10] N. Choubey, A. Cordero, Jaiswal J. P., and J. R. Torregrosa. Dynamical techniques for analyzing iterative schemes with memory. *Complexity*, 1231341:1–13, 2018.

[11] A. Cordero, H. Ramos, and J. R. Torregrosa. Some variants of halley's method with memory and their applications for solving several chemical problems. *Journal of Mathematical Chemistry*, 58:751–774, 2020.

[12] A. Cordero and J. R. Torregrosa. *On the Design of Optimal Iterative Methods for Solving Nonlinear Equations*, pages 79–111. Springer International Publishing, 2016.

[13] R. L. Devaney. *An Introduction to Chaotic Dynamical Systems*. Addison-Wesley, New York, USA, 1964.

[14] F. W. Khdhr, F. Soleymani, R. K. Saeed, and A. Agkül. An optimized steffensen-type iterative method with memory associated with annuity calculation. *European Physical Journal Plus*, 134:146, 2019.

[15] H. T. Kung and J. F. Traub. Optimal order of one-point and multipoint iteration. *Journal of the Association for Computing Machinery*, 21:643–651, 1974.

[16] M. Narang, S. Bhatiua, A. S. Alshomrani, and V. Kanwar. General efficient class of steffensen type methods with memory for solving systems of nonlinear equations. *Journal of Computational and Applied Mathematics*, 352:23–39, 2019.

[17] A. M. Ostrowski. *Solution of Equations and Systems of Equations*. Academic Press, New York, USA, 1960.

[18] M. S. Petković, B. Neta, L. D. Petković, and J. Dzŭnic. *Multipoint Methods for Solving Nonlinear Equations*. Academic Press, The Netherlands, 2013.

[19] M. S. Petković, B. Neta, L. D. Petković, and J. Dzŭnic. Multipoint methods for solving nonlinear equations: a survey. *Applied Mathematics and Computation*, 226:635–660, 2014.

[20] M. S. Petković and J. R. Sharma. On some efficient derivative-free iterative methods with memory for solving systems of nonlinear equations. *Numerical Algorithms*, 71:457–474, 2016.

[21] P. Praks and D. Brkić. One-log call iterative solution of the colebrook equation for flow friction based on padé polynomials. *Energies*, 11:1825, 2018.

[22] R. C. Robinson. *An Introduction to Dynamical Systems: Continuous and Discrete*. American Mathematical Society, Rhode Island, USA, 2012.

[23] J. F. Traub. *Iterative Methods for the Solution of Equations*. Prentice-Hall, New York, USA, 1964.

[24] X. Wang and Q. Fan. A modified ren's method with memory using a simple self-accelerating parameter. *Mathematics*, 8:540, 2020.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 135–140)

Designing new derivative-free memory methods to solve nonlinear scalar problems

Alicia Cordero[1], **Neus Garrido**[2], **Juan R. Torregrosa**[1], **Paula Triguero**[1]
1. *Instituto de Matemática Multidisciplinar, Universitat Politècnica de València, València, Spain*
2. *Escuela Superior de Ingeniería y Tecnología, Universidad Internacional de La Rioja, Logroño, Spain*

Abstract

In this work, two families of iterative methods without derivatives have been designed using the composition of iterative schemes and the inclusion of weight functions. The convergence analysis of the two-step and the three-step families is presented, showing the necessary conditions that must be satisfied by the weight functions to have order four and six, respectively. From them, two methods with memory have been derived, improving their order of convergence and their efficiency. All the methods are tested and compared with other known methods in the approxitation of the roots of different nonlinear functions. The results show the improvement of the classes after including memory.

1. Introduction

It is becoming a need in many scientific and technological disciplines to solve a nonlinear equation or a system of nonlinear equations. We describe this nonlinear problem for the scalar case as $f(x) = 0$, where $f : D \subset \mathbb{R} \longrightarrow \mathbb{R}$ and D is an open set. Due to the lack of analytical methods for solving these type of nonlinear problems, the implementation of iterative processes to solve them has become more frequent.
The use of iterative methods for solving nonlinear problems has increased in the recent decades. These iterative processes generate a sequence of points closer and closer to the solution, so that an approximation to the root with the required precision is obtained as a solution to the problem.
There is a wide literature related to iterative schemes for approximating simple roots of nonlinear functions (see [8], [1] and [5] and the references therein). Among them, the most classical iterative algorithm is Newton's method, with iterative expression

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \qquad k = 0, 1, 2, \ldots$$

Given an initial estimation x_0 to the root of $f(x) = 0$, Newton's method converges with quadratic order.

There are several ways to quantify the quality of an iterative method, in particular the speed of convergence and its computational cost. One point methods, such as Newton's scheme, are known for their simplicity and low computational cost, but have slow convergence. For this reason, the number of multipoint methods designed to increase the order of convergence has grown exponentially.

The most common iterative schemes are those that use only the previous iteration to obtain the next approximation. They are called methods without memory. However, Kung and Traub conjectured in [4] that the order of these scalar methods can not be grater than 2^{d-1}, where d is the number of different functional evaluations performed on each iteration of the method. However, iterative methods with memory, that is, methods that use more than one previous iterate, do not have any upper bound on their order of convergence.

On the other hand, many nonlinear functions have a derivative that is difficult to calculate or whose expression is not known. For this reason, Steffensen [7] proposed to approximate the derivative of Newton's method by

$$f'(x_k) \approx f[x_k + f(x_k), x_k] = \frac{f(x_k + f(x_k)) - f(x_k)}{f(x_k)}$$

and then replacing it on its iterative structure, obtaining the well-known Steffensen's method:

$$x_{k+1} = x_k - \frac{f(x_k)^2}{f(x_k + f(x_k)) - f(x_k)}, \qquad k = 0, 1, 2, \ldots,$$

with quadratic order.
In addition, it is known that the composition of two iterative methods with orders p_1 and p_2 results in a method with order of convergence $p_1 \cdot p_2$, but whose computational cost increases significantly. However, the use of

weight functions in the composition of iterative schemes allows the design of methods that increase the order of convergence without adding a high number of new functional evaluations.

In this paper, we propose the use of the above techniques to design methods with higher order of convergence and adding the minimal computational cost. For this purpose, we have organized the contents as follows. In Section 2 we present a new derivative-free iterative family with two steps and a real parameter. We analyse its convergence and the possibility of increasing the order by using previous iterations. Section 3 is devoted to extend the two-step family to a three-step family holding the same iterative structure. The convergence of this family and an approximation for the parameter to increase the order of convergence leading to a method with memory are also analysed. In Section 4 we test the performance of the methods studied in this work for solving different nonlinear equations. Finally, the conclusions of the study are summarized in section 5.

2. Derivative-free iterative family with two steps

In this section, we will use several techniques to design new iterative schemes that improve the performance of Newton and Steffensen's methods and also have the possibility of increasing the order of convergence. First, we propose a derivative-free family obtained by composing methods and using weight functions. Then, we will extend this family to higher order memory methods without adding new functional evaluations.

The starting family of iterative methods presented in this work is obtained by the composition of Steffensen's method and the addition of a real parameter β and a weight function H. The proposed scheme is as follows:

$$\begin{aligned} y_k &= x_k - \frac{f(x_k)}{f[w_k, x_k]}, \\ x_{k+1} &= y_k - H(\mu_k)\frac{f(y_k)}{f[y_k, x_k]}, \end{aligned} \qquad k = 0, 1, 2, \ldots, \tag{2.1}$$

where $w_k = x_k + \beta f(x_k)$, $\beta \in \mathbb{R} - \{0\}$, and the weight function variable is defined by $\mu = \frac{f(y)}{f(w)}$. We denote $M4_\beta$ the iterative family (2.1).

Theorem 2.1 shows the conditions that the weight function must satisfy to obtain order four for any value of the parameter.

Theorem 2.1 *Let $f : D \subset \mathbb{R} \longrightarrow \mathbb{R}$ be a real sufficiently differentiable function in a convex set D and let $\alpha \in D$ be a simple root of $f(x) = 0$. If x_0 is close enough to α and $H(\mu)$ satisfies $H(0) = H'(0) = 1$ and $|H''(0)| < \infty$, then sequence $\{x_k\}$ generated by family $M4_\beta$ converges to α with order of convergence 4 for any value of $\beta \in \mathbb{R}$, $\beta \neq 0$, being its error equation*:

$$e_{k+1} = \frac{1}{2}c_2(1 + \beta f'(\alpha))(-2c_3(1 + \beta f'(\alpha)) + c_2^2(6 + 4\beta f'(\alpha) - H_2))e_k^4 + O(e_k^5), \tag{2.2}$$

where $H_2 = H''(0)$, $e_k = x_k - \alpha$ and $c_j = \frac{1}{j!}\frac{f^{(j)}(\alpha)}{f'(\alpha)}$, $j \geq 2$.

From the error equation (2.2), we can observe that family $M4_\beta$ is fourth-order convergent for any value of β. According to the Kung and Traub conjecture [4], $M4_\beta$ is a family of optimal iterative methods as the number of different functional evaluations is three, i.e, $f(x_k)$, $f(y_k)$ and $f(w_k)$, so the maximum value $4 = 2^{3-1}$ is reached. Moreover, the value of $\beta = -\frac{1}{f'(\alpha)}$ leads to a method of the family with order five. As the solution α is unknown, we can not use this value to fix the parameter and increase the order of convergence. Following the guideliness in [3], we can approximate the derivative of f in the solution as

$$f'(\alpha) \approx f[x_k, x_{k-1}] = \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}},$$

so we propose the following approximation of the parameter, which varies on each iteration of the method:

$$\beta_k = -\frac{1}{f[x_k, x_{k-1}]}. \tag{2.3}$$

Let us note that with parameter β_k defined in (2.3) the number of different functional evaluations has not been increased, because $f(x_k)$ and $f(x_{k-1})$ are functional evaluations that were already being performed by the method at iterations k and $k-1$, respectively.

The replacement of (2.3) in the iterative structure (2.1) of $M4_\beta$ gives a method with memory, denoted $MM1$, that belongs to the family and also with higher order of convergence than the original family as Theorem 2.2 states.

Theorem 2.2 *Let $f : D \subset \mathbb{R} \longrightarrow \mathbb{R}$ be a real sufficiently differentiable function in a convex set D and let $\alpha \in D$ be a simple root of $f(x) = 0$. Let us suppose that $H(\mu)$ satisfies $H(0) = H'(0) = 1$, $H''(0) = 2$ and $|H'''(0)| < \infty$. If x_0 is close enough to α, method $MM1$ converges to α with order of convergence*:

$$p = 2 + \sqrt{6} \approx 4.4495.$$

In addition to the approximation with memory considered, we could develop approximations using higher order interpolating polynomials. However, we have studied the resulting iterative family after adding a new step in the iterative scheme of family $M4_\beta$.

3. Derivative-free iterative family with three steps

Following the same iterative structure than family $M4_\beta$, we propose to add a step in order to accelerate the convergence. In this sense, we propose to replicate the last step of the family and add a new weight function G, so we obtain the following three-step family of iterative schemes:

$$\begin{aligned} y_k &= x_k - \frac{f(x_k)}{f[w_k, x_k]}, \\ z_k &= y_k - H(\mu_k)\frac{f(y_k)}{f[y_k, x_k]}, \qquad k = 0, 1, 2, \ldots, \\ x_{k+1} &= z_k - G(\nu_k)\frac{f(z_k)}{f[z_k, y_k]} \end{aligned} \tag{3.1}$$

where $w_k = x_k + \beta f(x_k)$, $\beta \in \mathbb{R} - \{0\}$, and the weight function variables are $\mu = \frac{f(y)}{f(w)}$ and $\nu = \frac{f(z)}{f(y)}$. We denote the resulting three-step family as family $M6_\beta$. The analysis of its order if convergence is shown below.

Theorem 3.1 *Let $f : D \subset \mathbb{R} \longrightarrow \mathbb{R}$ be a real sufficiently differentiable function in a convex set D and let $\alpha \in D$ be a simple root of $f(x) = 0$. Let us suppose that the weight functions $H(\mu)$ and $G(\nu)$ hold*:

- $H(0) = 1$, $H'(0) = 1$, $|H''(0)| < \infty$.
- $G(0) = 1$, $|G'(0)| < \infty$.

Then, if x_0 is close enough to α, the sequence $\{x_k\}$ generated by family $M6_\beta$ converges to α with order of convergence 6 for any value of $\beta \in \mathbb{R}$, $\beta \neq 0$, The error equation of the family is given by:

$$\begin{aligned} e_{k+1} = & -\frac{c_2}{4}(1 + \beta f'(\alpha))\left(-2c_3(1 + \beta f'(\alpha)) + c_2^2(6 + 4\beta f'(\alpha) - H_2)\right) \\ & \cdot \left(-2c_3(1 + \beta f'(\alpha))G_1 + c_2^2(-2 + 6G_1 + 2\beta f'(\alpha)(-1 + 2G_1) - G_1 H_2)\right) e_k^6 + O(e_k^7), \end{aligned} \tag{3.2}$$

where $G_1 = G'(0)$, $H_2 = H''(0)$ and $c_j = \frac{1}{j!}\frac{f^{(j)}(\alpha)}{f'(\alpha)}$, $j \geq 2$. In addition, if we set $H_2 = 2$, the error equation (3.2) *turns into*

$$e_{k+1} = c_2(2c_2^2 - c_3)(c_2^2(1 - 2G_1) + c_3 G_1)(1 + \beta f'(\alpha))^3 e_k^6 + O(e_k^7). \tag{3.3}$$

In the same way as family $M4_\beta$, the term $1 + \beta f'(\alpha)$ appears in the error equation (3.3), so we can use the same approximation for the parameter in order to cancel the lower term in the error equation. Then, we replace the parameter $\beta_k = -\frac{1}{f[x_k, x_{k-1}]}$ in (3.1) obtaining a method of family $M6_\beta$. The resulting iterative scheme has been denoted $MM2$ and is a method with memory without additional functional evaluations. The improvement of the order of convergence is described in Theorem 3.2.

Theorem 3.2 *Let $f : D \subset \mathbb{R} \longrightarrow \mathbb{R}$ be a real sufficiently differentiable function in a convex set D and $\alpha \in D$ a simple root of $f(x) = 0$. Let us suppose that $H(\mu)$ and $G(\eta)$ are real functions satisfying*:

- $H(0) = 1$, $H'(0) = 1$, $H''(0) = 2$ *and* $|H'''(0)| < \infty$,
- $G(0) = 1$, $|G'(0)| < \infty$.

Then, if x_0 is close enough to α, method $MM2$ converges to α with order of convergence:

$$p = 3 + 2\sqrt{3} \approx 6.4641.$$

To compare different iterative methods from the point of view of their computational cost, Ostrowski [6] introduced the efficiency index $I = p^{1/d}$, where p is the order of the method and d is the number of funcional evaluations. Table 1 summarises the order and functional evaluations of the proposed methods and their efficiency index. We can see that all the methods improve the efficiency of Newtona and Steffensen's method. In turn, methods with memory $MM1$ and $MM2$ improve the efficiency with respect to the original families $M4_\beta$ and $M6_\beta$, respectively.

Method	p	d	I
Newton	2	2	1.4142
Steffensen	2	2	1.4142
$M4_\beta$	4	3	1.5847
$MM1$	4.4495	3	1.6448
$M6_\beta$	6	4	1.5650
$MM2$	6.4641	4	1.5945

Tab. 1 Efficiency indices

4. Numerical results

In this section, we perform numerical experiments to test the features of the proposed methods. With this aim, they are used to solve different nonlinear problems. We also compare our methods with the classical iterative schemes of Newton and Steffensen.

All the methods require an initial estimation x_0 to the root of the nonlinear function. In addition, to check the numerical performance of families $M4_\beta$ and $M6_\beta$ we have used the weight functions:

$$H(\mu) = 1 + \mu + \mu^2,$$

$$G(\nu) = 1 + \nu + \nu^2,$$

such that they hold the convergence conditions stated in Theorems 2.2 and 3.2. The real parameter β has been set to $\beta = 1$, having Steffensen's method in the first step, and arbitrarily to $\beta = 5$.

The solution of the following nonlinear functions has been approximated:

- $f_1(x) = e^{-x} + 2\sin(x) - x + 3.5$, $\alpha \approx 3.273938$.
- $f_2(x) = \cos(x) - x$, $\alpha \approx 0.73908513$.
- $f_3(x) = (x-1)^3 - 1$, $\alpha = 2$.

In order to compare the theoretical order of convergence of the methods with their practical implementation, we use the approximated computational order of convergence, ACOC, introduced by the authors in [2] and defined by

$$ACOC = \frac{\ln(|x_{k+1} - x_k|/|x_k - x_{k-1}|)}{\ln(|x_k - x_{k-1}|/|x_{k-1} - x_{k-2}|)}, \qquad k = 2, 3, \ldots$$

The numerical implementation has been done using Matlab R2018b with variable precision arithmetics of 2000 digits of mantissa. Tables 2, 3 and 4 show the results obtained for f_1, f_2 and f_3, respectively. For each method, we have shown the number of iterations, the difference between the two last iterations, the value of the function in the last iterate and the ACOC. Taking an initial estimation x_0, the iterative process stops when $|x_{k+1} - x_k| < 10^{-100}$ or $|f(x_{k+1})| < 10^{-100}$, with a maximum of 50 iterations.

We can observe in Tables 2 and 3 that the best results are given by the methods with memory $MM1$ and $MM2$. Both methods approximate the solution with high precision and the lowest number of iterations. In addition, methods belonging to families $M4_\beta$ and $M6_\beta$ are also competitive. In all cases the ACOC is near the theoretical order of convergence, being the higher value, as expected, in method $MM2$.

Finally, in Table 4 we can see an example where Steffensen's method and families $M4_\beta$ and $M6_\beta$ for $\beta = 1$ do not work properly. However, for $\beta = 5$ the performance is good and again methods with memory remain the most competitive.

x_0	Method	$iter$	$\lvert x_{k+1}-x_k\rvert$	$\lvert f(x_{k+1})\rvert$	ACOC
2	Newton	8	2.0845e–87	6.5561e–175	2.0000
	Steffensen	7	1.5294e–56	7.1309e–113	2.0000
	$M4_1$	5	2.3692e–66	2.1794e–264	4.0000
	$M4_5$	5	3.0444e–42	3.0444e–42	4.0041
	$MM1$	4	3.7619e–74	3.3998e–330	4.5071
	$M6_1$	4	4.8657e–61	2.0361e–364	6.0079
	$M6_5$	4	2.1005e–44	4.4813e–262	5.9678
	$MM2$	3	1.3534e–31	5.4909e–205	6.2994

Tab. 2 Numerical results for $f_1(x)$

x_0	Method	$iter$	$\lvert x_{k+1}-x_k\rvert$	$\lvert f(x_{k+1})\rvert$	ACOC
1	Newton	7	1.7955e–83	1.1913e–166	2.0000
	Steffensen	7	5.4267e–89	7.3307e–178	2.0000
	$M4_1$	4	2.4716e–74	1.0299e–296	4.0000
	$M4_5$	5	4.926e–67	1.9443e–265	4.0000
	$MM1$	3	5.6456e–30	7.4958e–133	4.0126
	$M6_1$	3	2.389e–41	4.0033e–247	6.0180
	$M6_5$	4	2.1274e–71	2.6129e–424	6.0041
	$MM2$	3	1.6466e–64	2.667e–416	6.0214

Tab. 3 Numerical results for $f_2(x)$

5. Conclusions

Two new derivative-free families of iterative methods have been introduced. The starting point has been an optimal two-step family with a real parameter and order four. After analyzing its order of convergence, the parameter has been approximated using two previous iterations, resulting in a method with memory with higher order of convergence than the original family and without additional functional evaluations. Then, we have extended the initial family to a three-step scheme with order six following a similar iterative estructure and the same real parameter. A new method with memory has been designed using the approximation of the parameter with memory as in the initial family and also improving the order of convergence. In both cases, the schemes not only improve the order but also the efficiency index with respect to the starting families. Finally, it has been verified that the theoretical analysis carried out in this work is consistent with the practical implementation of the methods. For this purpose, the proposed methods have been used to approximate roots of nonlinear test functions, obtaining the best results in the methods with memory.

Acknowledgements

This research was supported in part by PGC2018-095896-B-C22 (MCIU/AEI/FEDER, UE). The second author was also partially supported by the internal research project ADMIREN of Universidad Internacional de La Rioja (UNIR).

x_0	Method	$iter$	$\lvert x_{k+1} - x_k \rvert$	$\lvert f(x_{k+1}) \rvert$	ACOC
1.5	Newton	10	1.7506e–90	9.1937e–180	2.0000
	Steffensen	nc			
	$M4_1$	40	6.8579e–32	1.7695e–123	3.9979
	$M4_5$	7	2.4187e–60	4.3803e–236	4.0000
	$MM1$	5	8.8702e–78	5.9533e–343	4.4425
	$M6_1$	nc			
	$M6_5$	9	1.1125e–27	2.5886e–158	5.7084
	$MM2$	4	1.3332e–100	1.162e–645	6.5487

Tab. 4 Numerical results for $f_3(x)$

References

[1] S. Amat and S. Busquier. *Advances in iterative methods for nonlinear equations*. Springer, 2016.

[2] A. Cordero and J.R. Torregrosa. Variant's of Newton's method using fifth order quadrature formulas. *Appl. Math. Comput.*, 190:686–698, 2007.

[3] N. Garrido F.I. Chicharro, A. Cordero and J.R. Torregrosa. Stability and applicability of iterative methods with memory. *J. Math. Chem.*, 57:1282–1300, 2019.

[4] H.T. Kung and J.F. Traub. Optimal order of one-point and multipoint iteration. *J. Assoc. Comput. Math.*, 21:643–651, 1974.

[5] L.D. Petković M.S. Petković, B. Neta and J. Džunić. Multipoint methods for solving nonlinear equations: a survey. *Appl. Math. Comput.*, 226, 2014.

[6] A.M. Ostrowski. *Solution of equations and systems of equations*. Academic Press, 1966.

[7] J.F. Steffensen. Remarks on iteration. *Scand. Actuar. J.*, 1:64–72, 1933.

[8] J.F. Traub. *Iterative Methods for the Solution of Equations*. Prentice-Hall, New York, 1964.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 141–147)

Iterative processes with arbitrary order of convergence for approximating generalized inverses

Alicia Cordero[1], **Pablo Soto-Quiros**[2], **Juan R. Torregrosa**[1]
1. *Instituto de Matemática Multidisciplinar, Universitat Politècnica de València, València, Spain*
2. *Escuela de Matemática, Instituto Tecnológico de Costa Rica, Cartago, Costa Rica*

Abstract

A family of iterative schemes for finding approximate inverses of nonsingular matrices is suggested and established analytically. This class of methods can be used for finding the Moore-Penrose inverse of a rectangular complex matrix. The order of convergence is stated in each case, depending on the first non-zero parameter. For different examples, the accessibility of some schemes, that is, the set of initial estimations leading to convergence, is analyzed in order to select those with wider sets. This wideness is related with the value of the first non-zero value of the parameters defining the method. Finally, some numerical examples are provided to confirm the theoretical results and to show the feasibility and effectiveness of the new methods.

1. Introduction

Computing the matrix inverse of nonsingular matrices of higher size is difficult and is a time consuming task. Generally speaking, in wide variety of topics, one must compute the inverse or particularly the generalized inverses to comprehend and realize significant features of the involved problems.

In the last decade, many iterative schemes of different orders have been designed for approximating the inverse or some generalized inverse (Moore-Penrose inverse, Drazin inverse, etc.) of a complex matrix A. In this paper, we focus our attention on constructing a new class of iterative methods, free of inverse operators and with arbitrary order of convergence, for finding the inverse of a nonsingular complex matrix. We also study the proposed class for computing the Moore-Penrose inverse of complex rectangular matrices. The designed family depends on several real parameters, which by taking particular values provide us numerous known methods constructed by other authors with different procedures.

The most known iterative scheme for computing the inverse A^{-1} of a nonsingular complex matrix A is the Schulz's method whose iterative expression is

$$X_{k+1} = X_k(2I - AX_k), \quad k = 0, 1, \ldots \tag{1.1}$$

where I is the identity matrix with the same size of A. Schulz in [8] demonstrated the convergence of sequence $\{X_k\}_{k\geq 0}$, obtained from (1.1), to the inverse A^{-1} is guaranteed if the eigenvalues of matrix $I - AX_0$ are lower than 1. Taking into account that the residuals $E_k = I - AX_k$, $k = 0, 1, \ldots$ satisfy $\|E_{k+1}\| \leq \|E_k\|^2$, expression (1.1) has quadratic convergence. In general, in the Schulz-type methods it is common to use as initial approach $X_0 = \alpha A^*$ or $X_0 = \alpha A$, where $0 < \alpha < 2/\rho(A^*A)$, where A^* is the conjugate transpose of A and $\rho(\cdot)$ the spectral radius. In this paper, we use in the case of inverses and also in generalized inverses, the initial estimation $X_0 = \beta A^*/\|A\|^2$. We also study the values of the parameter β that guarantee convergence.

Li et al. in [5] proposed the family of iterative methods

$$X_{k+1} = X_k\left(\nu I - \frac{\nu(\nu-1)}{2}AX_k + \frac{\nu(\nu-1)(\nu-2)}{3!}(AX_k)^2 - \ldots + (-1)^{\nu-1}(AX_k)^{\nu-1}\right), \quad \nu = 2, 3, \ldots$$

with $X_0 = \alpha A^*$. They proved the convergence of ν-order of $\{X_k\}_{k\geq 0}$ to the inverse of matrix A. This class was used by Chen et al. in [2] and by Li et al. in [19] for approximating the Moore-Penrose inverse.

Soleymani et al. in [18] also constructed a fourth-order iterative scheme for calculating the inverse and the Moore-Penrose inverse, with iterative expression

$$X_{k+1} = \frac{1}{2}X_k(9I - AX_k(16I - AX_k(14I - AX_k(6I - AX_k)))), \quad k = 0, 1, \ldots$$

On the other hand, Stanimirović et al. in [16] designed the following scheme of order eleven for computing the generalized outer inverse $A^{(2)}_{T,S}$

$$X_{k+1} = X_k(I + (R_k + R_k^2)(I + (R_k^2 + R_k^4)(I + R_k^4))), \quad k = 0, 1, \ldots$$

being $R_k = I - AX_k$, $k = 0, 1, \ldots$.

Kaur et al. in [4], by using also the hyperpower iterative method, designed the following scheme of order five for obtaining the weighted Moore-Penrose inverse

$$X_{k+1} = X_k(5I - 10AX_k + 10(AX_k)^2 - 5(AX_k)^3 + (AX_k)^4), \quad k = 0, 1, \ldots$$

These papers are some of the manuscripts that have been published to approximate the inverse of a nonsingular matrix or some of the generalized inverses of arbitrary matrices. In this paper, we design a parametric family of iterative schemes with arbitrary order of convergence that contains many of the methods constructed up to date. For each fixed value of the order of convergence, we still have a class of iterative methods depending on several parameters.

The rest of this manuscript is organized as follows. Section 2 is devoted to the construction of the proposed class of iterative schemes, proving its convergence to the inverse of a nonsingular complex matrix, with arbitrary order of convergence. In Section 3, it is proven that the same family of iterative methods is able to converge to the Moore-Penrose inverse of a complex matrix of size $m \times n$. Some particular cases of this class are found in Section 4, corresponding to existing methods proposed by different authors. A wide range of numerical test are also found in Section 5, checking the robustness and applicability of the proposed methods on different kinds of matrices. With some conclusions and the references used finishes this manuscript.

2. A class of iterative schemes for matrix inversion

In this section, we present a parametric family of iterative schemes for approximating the inverse of nonsingular matrices and we prove the order of convergence of the different members of the family. First, we define the following polynomial matrix.

Definition 2.1 Let $U \in \mathbb{C}^{m\times m}$ be a complex square matrix and $p > 0$ a positive integer number. We define the polynomial matrix $H_p(U)$ as

$$H_p(U) = \sum_{j=1}^{p}(-1)^{j-1}C_p^j U^{j-1} = C_p^1 I - C_p^2 U + C_p^3 U^2 + \ldots + (-1)^{p-1}C_p^p U^{p-1},$$

where C_p^j is the combinatorial number $C_p^j = \binom{p}{j} = \dfrac{p!}{j!(p-j)!}$.

The following technical result can be proven by using mathematical induction with respect to parameter p.

Lemma 2.2 *Let $p > 0$ be a positive integer and $U \in \mathbb{C}^{m\times m}$. Then $(I - U)^p = I - UH_p(U)$.*

Let $A \in \mathbb{C}^{m\times m}$ be a nonsingular matrix and $p > 1$ a positive integer. Let $\{\alpha_1, \alpha_2, \ldots, \alpha_p\}$ be a set of real parameters such that $\alpha_i \in [0, 1]$, for $i = 1, 2, \ldots, p-1$, $\alpha_p \in]0, 1]$ and $\sum_{i=1}^{p}\alpha_i = 1$.

We assume a sequence of complex matrices $\{X_0, X_1, \ldots, X_n, \ldots\}$, of size $m \times m$, satisfying following conditions:

(a) $\|I - AX_0\| = \gamma_0 < 1$,

(b) $I - AX_{k+1} = \sum_{i=1}^{p}\alpha_i(I - AX_k)^i$.

We consider the family of methods with iterative expression

$$X_{k+1} = X_k \sum_{i=1}^{p}\alpha_i H_i(AX_k), \quad k = 0, 1, \ldots \tag{2.1}$$

For each positive integer p, $p > 1$, we have a different class of iterative methods, whose order of convergence depends on the value of parameters α_i, $i = 1, 2, \ldots, p$.

In the following results, the convergence of these schemes to the inverse of matrix A is proven.

Proposition 2.3 *Let $A \in \mathbb{C}^{m\times m}$ be a nonsingular matrix and $p > 1$ a positive integer. Let us consider the sequence of complex matrices constructed as*

$$X_{k+1} = X_k \sum_{i=1}^{p}\alpha_i H_i(AX_k), \quad k = 0, 1, \ldots,$$

where $\alpha_i \in [0,1]$, *for* $i = 1, 2, ..., p-1$, $\alpha_p \in]0,1]$ *and* $\sum_{i=1}^{p} \alpha_i = 1$. *Then, condition*

$$I - AX_{k+1} = \sum_{i=1}^{p} \alpha_i (I - AX_k)^i,$$

is equivalent to

$$X_{k+1} = X_k \sum_{i=1}^{p} \alpha_i \left(\sum_{j=1}^{i} (-1)^{j-1} C_i^j (AX_k)^{j-1} \right). \tag{2.2}$$

By mathematical induction it is also easy to prove the following result.

Proposition 2.4 *Let us consider sequence* $\{X_k\}_{k\geq 0}$ *obtained from expression* (2.1). *If* $\|I - AX_0\| < 1$, *then*

$$\|I - AX_k\| < 1, \quad k = 1, 2, \ldots$$

From these previous results, we can establish the following convergence theorem.

Theorem 2.5 *Let* $A \in \mathbb{C}^{m\times m}$ *be a nonsingular matrix and an initial guess* $X_0 \in \mathbb{C}^{m\times m}$. *Let* $\alpha_1, \ldots, \alpha_p$ *be nonnegative real numbers such that* $\alpha_i \in [0,1]$, $\alpha_p \neq 0$ *and* $\sum_{i=1}^{p} \alpha_i = 1$. *If* $\|I - AX_0\| < 1$, *then sequence* $\{X_k\}_{k\geq 0}$, *obtained by* (2.1), *converges to* A^{-1} *with convergence order* q *for any* $p > 1$, *where* $q = \min_{i=1,2,\ldots,p} \{i \mid \alpha_i \neq 0\}$.

In the next section, we extend the iterative schemes (2.1) for finding the Moore-Penrose inverse of any complex rectangular matrix.

3. A class of iterative schemes for computing Moore-Penrose inverse

Let A be a $m \times n$ complex matrix. The Moore-Penrose inverse of A (pseudoinverse), denoted by $A^\dagger$, is the unique $n \times m$ matrix X satisfying

$$AXA = A, \quad XAX = X, \quad (AX)^* = AX, \quad (XA)^* = XA.$$

This generalized inverse plays an important role in several fields, such as eigenvalue problems and the linear least square problems [3]. It can be obtained, explicitly, from the singular value decomposition of A but, with a high computational cost. Therefore, it is interesting to have efficient iterative methods to approximate this matrix. In this section, we prove how family (2.1) allows us to compute the pseudoinverse with the same order of convergence that in the previous section, where the inverse of a square matrix was calculated. First, we establish the following technical result, that is proven by mathematical induction, although other authors state similar results in the context of outer inverses (see, for example, [17]).

Lemma 3.1 *Let us consider* $X_0 = \alpha A^*$, *where* $\alpha \in \mathbb{R}$, *and sequence* $\{X_k\}_{k\geq 0}$ *generated by family* (2.1). *For any* $k \geq 0$, *it is satisfied*

$$(X_k A)^* = X_k A, \quad (AX_k)^* = AX_k, \quad X_k A A^\dagger = X_k, \quad A^\dagger A X_k = X_k. \tag{3.1}$$

Now, some technical results are presented.

Lemma 3.2 ([2]) *Let* $A \in \mathbb{C}^{m\times n}$ *and* $X_0 = \alpha A^*$ *be, where* $\alpha < \dfrac{1}{\sigma_1^2}$ *and* σ_1 *is the largest singular value of* A. *Then* $\|(X_0 - A^\dagger)A\| < 1$.

Lemma 3.3 *Let* $A \in \mathbb{C}^{m\times n}$ *and* $\{X_k\}_{k\geq 0}$ *be the sequence generated by* (2.1). *Let us consider* $E_k = X_k - A^\dagger$, $k = 0, 1, \ldots$. *Then,*

1. $\|X_k - A^\dagger\| \leq \|E_k A\| \|A^\dagger\|$,
2. $(I - A^\dagger A) E_k A = \mathbf{0}$.

Lemma 3.4 *Let $A \in \mathbb{C}^{m\times n}$ and $\{X_k\}_{k\geq 0}$ be the sequence generated by* (2.1). *By using $E_k = X_k - A^{\dagger}$, defined in the previous lemma, we have*

$$E_{k+1}A = \sum_{i=1}^{p} \alpha_i (-1)^{i-1} (E_k A)^i, \quad k = 0, 1, \ldots \tag{3.2}$$

Finally, we can state the following convergence result.

Theorem 3.5 *Let $A \in \mathbb{C}^{m\times n}$ and $q = \min_{i=1,2,\ldots,p} \{i \mid \alpha_i \neq 0\}$. Then, sequence $\{X_k\}_{k\geq 0}$ generated by* (2.1) *converges to the Moore-Penrose inverse $A^{\dagger}$ with qth-order provided that $X_0 = \alpha A^*$, where $\alpha < \frac{1}{\sigma_1^2}$ is a constant and σ_1 is the largest singular value of A.*

4. Some known members of the proposed class

The family of iterative schemes (2.1) is a generalization of other known methods constructed with different techniques. Now, we describe some of them.

1. For any $p > 1$, if $\alpha_1 = \cdots = \alpha_{p-1} = 0$ and $\alpha_p = 1$, then we obtain the method proposed by Li and Li. (see Eq. (2.3) in [5] for inverse case and Eq. (2.1) in [2] for pseudoinverse one). Recall that method proposed by Li and Li generalizes the Newton-Schultz ($p = 2$) and Chebyshev method ($p = 3$).

2. On the other hand, if $\alpha_i = 0$ for $i = 1, 2, \ldots, 8$, $\alpha_9 = \alpha_{12} = 1/8$ and $\alpha_{10} = \alpha_{11} = 3/8$, then we get the method proposed by Soleymani and Stanimirovic (see Eq. (12) in [9]).

3. Also, expression (2.1) gives us the method proposed by Toutounian and Soleymani (see Eq. (18) in [18]), if $\alpha_4 = 1/2$ and $\alpha_5 = 1/2$ and $\alpha_1 = \alpha_2 = \alpha_3 = 0$.

4. When the only not null parameter is $\alpha_7 = 1$, then we obtain method proposed by Soleymani (see Eq. (18) in [11]).

5. In a similar way, if the only parameter different from zero is $\alpha_6 = 1$, then the method proposed by Soleymani, Stanimirovic and Zaka (see Eq. (2.4) in [14]) is obtained.

6. When $\alpha_i = 0$ for $i = 1, 2, ...7$, $\alpha_8 = \alpha_{10} = 1/4$ and $\alpha_9 = 2/4$, the resulting scheme is that proposed by Soleymani in Eq. (9) in [12].

7. The method proposed by Soleymani et al in [13], Eq. (10), appears if $\alpha_1 = \cdots = \alpha_8 = 0$, $\alpha_9 = 7/9$ and $\alpha_{10} = 2/9$.

8. The scheme proposed by Razavi, Kerayechian, Gachpazan and Shateyi, (see Eq. (16) in [7]) is obtained if we choose $\alpha_1 = ... = \alpha_9 = 0$, $\alpha_{10} = \alpha_{12} = 1/4$ and $\alpha_{11} = 1/2$ in Equation (2.1).

9. When the first eight paramenters are null, $\alpha_9 = 343/729$, $\alpha_{10} = 294/729$, $\alpha_{11} = 84/729$ and $\alpha_{12} = 8/729$, we get the scheme proposed by Al-Fhaid et al in [1], Eq. (5).

10. If $\alpha_7 = 9/16$, $\alpha_8 = 6/16$, $\alpha_9 = 1/16$ and the rest of parameters are zero, then the scheme proposed by Soleymani is found (see Eq. (3.1) in [10]).

11. When $p = 12$ and the only parameters different from zero are $\alpha_9 = \alpha_{12} = 1/8$ and $\alpha_9 = \alpha_{10} = 3/8$, therefore the method proposed by Liu and Cai. see Eq. (4) in [6]) is obtained.

12. If $\alpha_2 = 0$, $\alpha_1 = 1 - \alpha$ and $\alpha_3 = \alpha$, where $\alpha \in (0, 1]$, then we find the method proposed by Srivastava and Gupta in [15]), Eq. (6).

5. Numerical examples

In this section, we check the performance of the proposed schemes, on small and large–scale matrices. Among them, we work with the Hilbert matrix as an example of ill-conditioned matrix. These numerical tests have been made with Matlab R2018b, by using double precision arithmetics. The convergence is checked by means of the stopping criterium of the residual, $\|AX_k - I\| < 10^{-6}$ and a maximum of 200 iterations. In all cases, the initial estimation taken is $X_0 = \beta \frac{A^T}{\|A\|^2}$, being A the matrix whose inverse we are estimating and choosing values of parameter β close to 0.7.

In Tables 1 - 2, elements $\alpha_1 = 0$, $\alpha_1 = 0.6$ and $\alpha_1 = 0.8$ for the class $p = 2$ (all of them with $\alpha_2 = 1 - \alpha_1$) are compared with the members of class $p = 3$ corresponding to $\alpha_1 = 0$ and $\alpha_2 = 0$, $\alpha_2 = 0.6$ and $\alpha_2 = 0.8$, where $\alpha_3 = 1 - \alpha_2$. The comparison is made through the number of iterations needed to converge (it) and the residual $\|AX_k - I\|$, denoted by (res). If the method does not converge (typically giving "NaN"), it is denoted by "nc" in the column of iterations; if the scheme simplify needs more than 200 iterations to converge, it is denoted by > 200.

In Table 1 the numerical results correspond to a Leslie matrix of size 100×100, and in Table 2 the results generated by a Hilbert matrix of size 5×5 are shown.

β	$p = 2$						$p = 3, \alpha_1 = 0$					
	$\alpha_1 = 0$		$\alpha_1 = 0.6$		$\alpha_1 = 0.8$		$\alpha_2 = 0$		$\alpha_2 = 0.6$		$\alpha_2 = 0.8$	
	it	res	it	res	it	res	it	res	it	res	it	res
1	18	6.9e-12	55	8.5e-7	113	9.2e-7	11	2.9e-8	14	6.4e-7	16	2.4e-10
1.5	17	4.2e-9	54	7.7e-7	111	8.8e-7	11	4.8e-12	14	2.4e-9	15	1.4e-7
2	53	3.7e-11	53	8.3e-7	107	9.6e-7	33	8.0e-11	14	1.3e-11	15	1.4e-9
2.5	nc	-	52	9.8e-7	108	9.2e-7	nc	-	13	3.9e-7	nc	-
3	nc	-	52	7.5e-7	107	9.2e-7	nc	-	13	3.7e-8	nc	-
3.5	nc	-	nc	-	106	9.5e-7	nc	-	26	1.1e-12	nc	-
4	nc	-	nc	-	106	8.1e-7	nc	-	nc	-	nc	-
4.5	nc	-	nc	-	105	8.7e-7	nc	-	nc	-	nc	-
5	nc	-	nc	-	104	9.6e-7	nc	-	nc	-	14	1.2e-12
5.5	nc	-	nc	-	104	8.5e-7	nc	-	nc	-	nc	-
6	nc	-	nc	-	> 200	-	nc	-	nc	-	nc	-

Tab. 1 Numerical results for a Leslie matrix of size 100×100

In Table 1, we notice that for large–scale (100×100) Leslie matrix, the numerical results obtained by $p = 2$, $\alpha_1 = 0$ and $\alpha_2 = 0.6$ show convergence to the inverse matrix even when parameter β of the initial estimation is not close to zero. However, in these cases the number of iterations is very high. Regarding the lowest number of iterations needed to converge, the best method corresponds to $p = 3$, $\alpha_1 = 0$ and $\alpha_2 = 0.6$ as it holds low number of iterations and high value of β.

Table 2 corresponds to a test on a 5×5 Hilbert matrix. It is clear that the number of iterations is high due to the bad conditioning of the matrix. Nevertheless, the performance is, in general similar to previous cases.

β	$p = 2$						$p = 3, \alpha_1 = 0$					
	$\alpha_1 = 0$		$\alpha_1 = 0.2$		$\alpha_1 = 0.4$		$\alpha_2 = 0$		$\alpha_2 = 0.6$		$\alpha_2 = 0.8$	
	it	res	it	res	it	res	it	res	it	res	it	res
1	42	3.9e-9	54	5.7e-7	72	4.3e-7	27	2.3e-11	34	9.5e-11	37	1.07e-7
1.5	41	4.9e-7	53	9.3e-7	71	4.8e-7	26	5.1e-8	33	1.5e-7	37	9.8e-11
2	56	5.7e-8	53	4.2e-7	70	6.9e-7	nc	-	33	2.3e-9	36	3.9e-7
2.5	nc	-	nc	-	70	4.5e-7	nc	-	33	4.8e-11	nc	-
3	nc	-	nc	-	nc	-	nc	-	33	1.3e-11	nc	-
3.5	nc	-	nc	-	nc	-	nc	-	32	2.1e-8	nc	-
4	nc	-	nc	-	nc	-	nc	-	nc	-	nc	-
4.5	nc	-	nc	-	nc	-	nc	-	nc	-	nc	-
5	nc	-	nc	-	nc	-	nc	-	nc	-	nc	-
5.5	nc	-	nc	-	nc	-	nc	-	nc	-	nc	-
6	nc	-	nc	-	nc	-	nc	-	nc	-	nc	-

Tab. 2 Numerical results for a Hilbert matrix of size 5×5

Finally, Table includes the results of pseudoinverse calculation for a random matrix of size 300×200. In this case, Chebyshev's method performs better than the most of schemes under study, as it need a very low number of iterations to converge to the pseudoinverse, although $p = 2$ can converge even with values of $\beta = 6$ or higher.

6. Conclusions

In this paper, we have developed a parametric family of iterative methods for computing inverse and pseudoinverse of a complex matrix, having arbitrary order of convergence. Moreover, we have shown in Theorems 2.5 and 3.5 that the order of the suggested method in (2.1) depends on the first non-zero parameter. The proposed parametric

β	$p=2$						$p=3, \alpha_1=0$					
	$\alpha_1=0$		$\alpha_1=0.6$		$\alpha_1=0.8$		$\alpha_2=0$		$\alpha_2=0.6$		$\alpha_2=0.8$	
	it	res	it	res	it	res	it	res	it	res	it	res
1	19	5.4e-8	56	6.7e-7	109	9.7e-7	13	4.9e-15	16	1.2e-9	18	5.7e-14
1.5	19	1.3e-11	55	6.0e-7	107	9.3e-7	12	4.3e-8	16	4.2e-13	17	3.8e-10
2	18	5.4e-8	54	6.5e-7	106	8.1e-7	12	1.5e-10	15	2.1e-8	17	6.8e-13
2.5	nc	-	53	7.7e-7	104	9.7e-7	nc	-	15	6.1e-10	nc	-
3	nc	-	52	9.7e-7	103	9.7e-7	nc	-	15	2.1e-11	nc	-
3.5	nc	-	52	7.7e-7	103	8.0e-7	nc	-	15	3.5e-8	nc	-
4	nc	-	nc	-	102	8.5e-7	nc	-	nc	-	nc	-
4.5	nc	-	nc	-	101	9.2e-7	nc	-	nc	-	nc	-
5	nc	-	nc	-	101	8.1e-7	nc	-	nc	-	nc	-
5.5	nc	-	nc	-	100	9.0e-7	nc	-	nc	-	nc	-
6	nc	-	nc	-	100	8.1e-7	nc	-	nc	-	nc	-

Tab. 3 Numerical results for the estimation of the pseudoinverse of a random matrix of size 300×200 with $X_0 = \beta \dfrac{A^T}{\|A\|^2}$

family in (2.1) is a generalization of other methods which are obtained for particular values of the parameters. The numerical experiments show the feasibility and effectiveness of the new methods, for both nonsingular and rectangular matrices with or without full rank and arbitrary size.

Acknowledgements

This research was supported in part by PGC2018-095896-B-C22 (MCIU/AEI/FEDER, UE) and in part by VIE from *Instituto Tecnológico de Costa Rica* (Research #1440037).

References

[1] AS Al-Fhaid, S Shateyi, M Ullah, and F Soleymani. A matrix iteration for finding Drazin inverse with ninth-order convergence. *Abstract Appl. Anal.*, 2014 ID 137486, 2014.

[2] Haibin Chen and Yiju Wang. A family of higher-order convergent iterative methods for computing the Moore–Penrose inverse. *Appl. Math. Comput.*, 218(8):4012–4016, 2011.

[3] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge University Press, New York, 2012.

[4] M. Kaur, M. Kansal, and S. Kumar. An efficient hyperpower iterative method for computating weighted Moore-Penrose inverse. *AIMS Mathematics*, 5(3):1680–1692, 2020.

[5] Weiguo Li and Zhi Li. A family of iterative methods for computing the approximate inverse of a square matrix and inner inverse of a non-square matrix. *Appl. Math. Comput.*, 215(9):3433–3442, 2010.

[6] Xiaoji Liu and Naping Cai. High-order iterative methods for the DMP inverse. *J. Math.*, 2018 ID 8175935, 2018.

[7] M Kafaei Razavi, Asghar Kerayechian, Mortaza Gachpazan, and Stanford Shateyi. A new iterative method for finding approximate inverses of complex matrices. *Abstract Appl. Anal.*, 2014 ID 563787, 2014.

[8] G. Schulz. Iterative berechnung der reziproken matrix. *Z. Angew. Math. Mech.*, 13:57–59, 1933.

[9] F Soleymani and P S Stanimirović. A higher order iterative method for computing the Drazin inverse. *Sci. World*, 2013 ID 708647, 2013.

[10] Fazlollah Soleymani. A rapid numerical algorithm to compute matrix inversion. *Int. J. Math. Math. Sci.*, 2012 ID 134653, 2012.

[11] Fazlollah Soleymani. On a fast iterative method for approximate inverse of matrices. *Communications of the Korean Mathematical Society*, 28(2):407–418, 2013.

[12] Fazlollah Soleymani. A fast convergent iterative solver for approximate inverse of matrices. *Numer. Linear Algebra Appl.*, 21(3):439–452, 2014.

[13] Fazlollah Soleymani, Hossein Salmani, and Masoud Rasouli. Finding the Moore–Penrose inverse by a new matrix iteration. *J. Appl. Math. Comput.*, 47(1-2):33–48, 2015.

[14] Fazlollah Soleymani, Predrag S Stanimirović, and Malik Zaka Ullah. An accelerated iterative method for computing weighted Moore–Penrose inverse. *Appl. Math. Comput.*, 222:365–371, 2013.

[15] Shwetabh Srivastava and DK Gupta. A third order iterative method for $a^{\dagger}$. *Int. J. Comput. Sci. Math.*, 4(2):140–151, 2013.

[16] P.S. Stanimirović, A. Kumar, and V.W. Katsikis. Further efficient hyperpower iterative methods for the computation of generalized inverses $A^{(2)}_{T,S}$. *RACSAM*, 113:3323–3339, 2019.

[17] P.S. Stanimirović and F. Soleymani. A class of numerical algorithms for computing outer matrices. *Comput. Appl. Math.*, 263:236–245, 2014.

[18] F Toutounian and F Soleymani. An iterative method for computing the approximate inverse of a square matrix and the Moore–Penrose inverse of a non-square matrix. *Appl. Math. Comput.*, 224:671–680, 2013.

[19] L Weiguo, L Juan, and Q Tiantian. A family of iterative methods for computating Moore-Penrose inverse of a matrix. *Linear Algebra Appl.*, 438:47–56, 2013.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 148–154)

FCF formulation of Einstein equations: local uniqueness and numerical accuracy and stability

Isabel Cordero-Carrión[1], Samuel Santos-Pérez[1], Pablo Cerdá-Durán[2]
1. *Departamento de Matemáticas, Universitat de València, E-46100 Burjassot, València, Spain.*
2. *Departamento de Astronomía y Astrofísica, Universitat de València, E-46100 Burjassot, València, Spain.*

Abstract

We present Einstein equations in the so-called Fully Constrained Formulation (FCF). This formulation has two different sectors: the elliptic sector formed by the Hamiltonian and Momentum constraints together with the equations derived from the gauge choice, and the hyperbolic sector which encodes the evolution of the rest of degrees of freedom of the spacetime metric including the gravitational waves. We present a modification of both sectors that keeps local uniqueness properties but has a better behaviour regarding the relativistic expansion of the equations. We also comment on numerical properties of this reformulation.

1. Introduction

Astrophysical scenarios containing compact objects are modeled by complex spacetimes which require, in general, to solve Einstein equations numerically. This is also true in the case of complex cosmological models. In the 3+1 decomposition of Einstein equations, spacetime is foliated through spacelike hypersurfaces. Doing this, the equations are decomposed in a set of elliptic equations, also called constraint equations, and a set of hyperbolic equations, also called evolution equations.

Constraint equations are only solved initially in the case of the approach by free evolution schemes. It is well known that if we evolve in time analytically some given initial data that satisfies the constraint equations using the evolution equations, then this data will also satisfy the constraint equations in posterior times, see [5]. This is true theoretically, but it may not be the case numerically. Formulations that solve the constraint equations on each time step are called constrained schemes. This work focuses on these schemes, and in particular in the so-called Fully Constrained Formulation (FCF) of the Einstein equations [2, 4].

This document is structured as follows. In Section 2 we introduce Einstein equations and the geometry of foliations. Section 3 describes technical details of a new reformulation of the FCF. In Section 4 we present the solution of the spacetime geometry of a neutron star considering the new reformulation of the FCF. We compare our solution with the one obtained with LORENE library, which employs spectral methods. We also make a comparison between our solution with the modification of the FCF equations and other approximate formulation that neglects the hyperbolic sector; this comparison confirms the accuracy improvement in the proposed reformulation of the FCF equations. Finally, in Section 5 we draw some conclusions and comment on future steps. From now on we use geometrical units in which $c = G = 1$, where c denotes the speed of light and G the universal constant of gravitation.

2. Einstein equations and foliations

Einstein Equations tells us how spacetime is curved according to the energy and matter content. These equations read

$$G_{\mu\nu} = 8\pi T_{\mu\nu}, \tag{2.1}$$

where $G_{\mu\nu}$ is the Einstein tensor, representing the information about the geometry of spacetime, and $T_{\mu\nu}$ is the energy-momentum tensor, concerning the distribution of energy and momentum. Einstein equations are a set of 10 non-linear coupled partial differential equations. They have exact solution only in a very few special cases, mostly in presence of symmetries. In general, they need to be solved numerically and this is the goal of Numerical Relativity.

Globally hyperbolic spacetimes allow to chose coordinates (t, x^i) such that level sets t = constant are spacelike hypersurfaces, that is, every tangent vector of these hypersurfaces is spacelike. Spacetime is thus foliated through spacelike hypersurfaces. The normal vector to these hypersurfaces is associated with the so-called Eulerian observer.

One important variable in Numerical Relativity is the lapse function $N(t, x^i)$, which is the factor connecting the lapse of proper time τ of this observer and the lapse of coordinate time t:

$$d\tau = N(t, x^i)dt. \tag{2.2}$$

Another important variable is the shift vector $\beta(t, x^i)$, which can be seen as the velocity between the Eulerian observer and the curves x^i = constant:

$$x^i_{t+dt} = x^i_{dt} - \beta^i(t, x^i)dt. \tag{2.3}$$

As the curves x^i = constant are not associated with any observer in general, the shift vector can be superluminical. This does not represent any physical propagation velocity, but just a foliation choice. The metric tensor of spacetime can be expressed as

$$g_{\mu\nu}dx^\mu dx^\nu = -N^2dt^2 + \gamma^{ij}(dx^i + \beta^i)(dx^j + \beta^j dt), \tag{2.4}$$

where γ^{ij} is the 3-metric in each hypersurface t = constant, also called the spatial metric. Therefore, Einstein equations can be decomposed in a set of evolution equations that have a hyperbolic character, and another set of constraint equations with elliptic character which have to be satisfied in each hypersurface and only depend on the spatial coordinates x^i.

3. Fully Constrained Formalism

The next manipulations are motivated by previous works and ideas described in [5]. First, we introduce a time independent flat background metric f_{ij}, which coincides with γ_{ij} at spatial infinity, and the following conformal decomposition:

$$\gamma^{ij} = \psi^4\tilde{\gamma}^{ij}. \tag{3.1}$$

We call γ_{ij} the conformal metric and $\psi := (\gamma/f)^{1/12}$ is the conformal factor, where $\gamma = \det\gamma_{ij}$ and $f = \det f_{ij}$. Let us denote by K^{ij} the extrinsic curvature on each hypersurface. We define the tensor A^{ij} as the traceless part of K^{ij}:

$$A^{ij} = K^{ij} - \frac{1}{3}K\gamma^{ij}, \tag{3.2}$$

where K represents the trace of extrinsic curvature. We also define $h^{ij} = \tilde{\gamma}^{ij} - f^{ij}$. Moreover, the gauge freedom of Einstein equations allow us to impose 4 extra conditions. In our case these will be $K = 0$ and

$$\mathcal{D}_k\tilde{\gamma}^{ki} = 0, \tag{3.3}$$

where $\mathcal{D}$ is the Levi-Civita connection associated with f^{ij}. The first condition is called maximal slicing and the second one generalized Dirac gauge. The next step is introducing a conformal decomposition of the extrinsic curvature,

$$K^{ij} = \psi^{10}\hat{A}^{ij}, \tag{3.4}$$

and at the same time its longitudinal/transverse decomposition,

$$\hat{A}^{ij} = (LX)^{ij} + \hat{A}^{ij}_{TT}, \tag{3.5}$$

introducing in this way the vector field X^i and the traceless and transverse tensor $\hat{A}^{ij}_{TT}$, $\mathcal{D}_i\hat{A}^{ij}_{TT} = 0$. L is the Killing operator. These last two definitions are motivated by the resolution of local uniqueness issues as it can be checked in [3].

Finally, we introduce two new fields

$$\dot{X}^i = \partial_t X^i,$$
$$V^i = 2N\psi^{-6}X^i - \beta^i,$$

in order to fix accuracy issues as we will discuss later on. These last two variables are introduced originally in this work. The following projections of the energy-momentum tensor are introduced for completeness,

$$S_{ij} = T_{\mu\nu}\gamma^\mu_i\gamma^\nu_j,$$
$$S^i = -\gamma^{i\mu}T_{\mu\nu}n^\nu,$$
$$S = \gamma^{ij}S_{ij},$$
$$E = T_{\mu\nu}n^\mu n^\nu,$$

and also the rescaled quantities $S_i^* = \psi^6 S_i$, $S^* = \psi^6 S$ and $E^* = \psi^6 E$. After all these definitions, we end up with an evolution equation for h^{ij},

$$\begin{aligned}\partial_t h^{ij} &= \beta^k \mathcal{D}_k h^{ij} - h^{ik}\mathcal{D}_k\beta^j - h^{kj}\mathcal{D}_k\beta^i + \frac{2}{3}h^{ij}\mathcal{D}_k\beta^k \\ &+2N\psi^{-6}\hat{A}^{ij}_{TT} + (LV)^{ij} - X^j\mathcal{D}^i(2N\psi^{-6}) - X^i\mathcal{D}^j(2N\psi^{-6}) + \frac{2}{3}f^{ij}X^k\mathcal{D}_k(2N\psi^{-6}),\end{aligned} \tag{3.6}$$

and another one for $\hat{A}^{ij}_{TT}$,

$$\begin{aligned}\partial_t \hat{A}^{ij}_{TT} &= \mathcal{D}_k\left(\beta^k\hat{A}^{ij}\right) - \hat{A}^{kj}\mathcal{D}_k\beta^i - \hat{A}^{ik}\mathcal{D}_k\beta^j + \frac{2}{3}\hat{A}^{ij}\mathcal{D}_k\beta^k + 2N\psi^{-6}\tilde{\gamma}_{kl}\hat{A}^{ik}\hat{A}^{jl} \\ &+\frac{3}{4}N\psi^{-6}\tilde{\gamma}^{ij}\tilde{\gamma}_{lk}\tilde{\gamma}_{nm}\hat{A}^{km}\hat{A}^{ln} + N\psi^2\tilde{R}^{ij}_* - \frac{1}{4}N\psi^2\tilde{R}\tilde{\gamma}^{ij} - \frac{1}{2}(\tilde{\gamma}^{ik}\mathcal{D}_k h^{lj} + \tilde{\gamma}^{kj}\mathcal{D}_k h^{il})\mathcal{D}_l(N\psi^2) \\ &+\mathcal{D}_k\left(\frac{N\psi^2}{2}\right)\tilde{\gamma}^{kl}\mathcal{D}_l h^{ij} - 8\pi N\psi^{10}S^{ij} + 4\pi NS^*\tilde{\gamma}^{ij} \\ &-(L\dot{X})^{ij} + 4\tilde{\gamma}^{ik}\tilde{\gamma}^{jl}\mathcal{D}_k\psi\mathcal{D}_l(N\psi) + 4\tilde{\gamma}^{ik}\tilde{\gamma}^{jl}\mathcal{D}_l\psi\mathcal{D}_k(N\psi) - 2\tilde{\gamma}^{ij}\tilde{\gamma}^{kl}\mathcal{D}_k\psi\mathcal{D}_l(N\psi) \\ &+\frac{N\psi^2}{2}\tilde{\gamma}^{kl}\mathcal{D}_k\left(\mathcal{D}_l h^{ij}\right) - \tilde{\gamma}^{ik}\tilde{\gamma}^{jl}\mathcal{D}_k\mathcal{D}_l(N\psi^2),\end{aligned} \tag{3.7}$$

where

$$\begin{aligned}\tilde{R}^{ij}_* &= \frac{1}{2}\left(-\mathcal{D}_l h^{ik}\mathcal{D}_k h^{jl} - \tilde{\gamma}_{kl}\tilde{\gamma}^{mn}\mathcal{D}_m h^{ik}\mathcal{D}_n h^{jl} + \tilde{\gamma}_{nl}\mathcal{D}_k h^{mn}(\tilde{\gamma}^{ik}\mathcal{D}_m h^{jl} + \tilde{\gamma}^{jk}\mathcal{D}_m h^{il})\right) \\ &+\frac{1}{4}\tilde{\gamma}^{ik}\tilde{\gamma}^{jl}\mathcal{D}_k h^{mn}\mathcal{D}_l\tilde{\gamma}_{mn},\end{aligned}$$

and

$$\tilde{R} = \frac{1}{4}\tilde{\gamma}^{kl}\mathcal{D}_k h^{mn}\mathcal{D}_l\tilde{\gamma}_{mn} - \frac{1}{2}\tilde{\gamma}^{kl}\mathcal{D}_k h^{mn}D_n\tilde{\gamma}_{ml}.$$

There is an issue concerning the post-newtonian order of the variables appearing in these equations. This means to expand variables in powers of $1/c$ in the approximation of low gravity and low velocity of the sources. These orders can be decuded from [1]. For instance, in equation (3.6) $\partial_t h^{ij}$ has leading post-newtonian order of $1/c^5$ which matches with the right hand side post-newtonian order. This is thanks to the introduction of the vector field V^i. If this were not the case, cancellations on the lower-order side must happen theoretically but it may not be the case numerically. A previous expression of (3.6) in [3], undergoes this problem as well as the equation (3.7): the left hand side has order $1/c^6$, meanwhile the terms of the last two lines of the right hand side of this equation have order $1/c^4$. This order correction is still a work in progress, but we already reduced the numbers of terms that should cancel analytically to get a lower order of $1/c^6$.

The constraint equations are the ones from the original elliptic sector of the FCF equations with some modifications and simplifications by the use of the vector field V^i, and additional elliptic equations for the variables V^i, $\dot{X}^i$. The whole elliptic sector is presented hereinafter (the post-newtonian order is placed beside each equation):

$$\Delta X^i + \frac{1}{3}\mathcal{D}^i\mathcal{D}_j X^j = -\tilde{\gamma}^{im}\left(\mathcal{D}_k\tilde{\gamma}_{ml} - \frac{1}{2}\mathcal{D}_m\tilde{\gamma}_{kl}\right)\hat{A}^{kl} + 8\pi\tilde{\gamma}^{ij}(S^*)_j = \mathcal{O}\left(\frac{1}{c^3}\right); \tag{3.8}$$

$$\tilde{\gamma}^{kl}\mathcal{D}_k\mathcal{D}_l\psi = -2\pi\psi^{-1}E^* - \frac{1}{8}\psi^{-7}\tilde{\gamma}_{il}\tilde{\gamma}_{jm}\hat{A}^{lm}\hat{A}^{ij} + \frac{1}{8}\psi\tilde{R} = \mathcal{O}\left(\frac{1}{c^2}\right); \tag{3.9}$$

$$\begin{aligned}\tilde{\gamma}^{ik}\mathcal{D}_i\mathcal{D}_k(N\psi^2) &= 2\psi^{-1}\tilde{\gamma}^{ik}\mathcal{D}_k\psi\mathcal{D}_i(N\psi^2) - 2\psi^{-2}(N\psi^2)\tilde{\gamma}^{ik}\mathcal{D}_k\psi\mathcal{D}_i\psi + \frac{3}{4}\psi^{-8}(N\psi^2)\tilde{\gamma}_{il}\tilde{\gamma}_{jm}\hat{A}^{lm}\hat{A}^{ij} \\ &+\frac{1}{4}(N\psi^2)\tilde{R} + 4\pi\psi^{-2}(N\psi^2)S^* = \mathcal{O}\left(\frac{1}{c^4}\right);\end{aligned} \tag{3.10}$$

$$\begin{aligned}\Delta V^i + \frac{1}{3}\mathcal{D}^i\mathcal{D}_j V^j &= -h^{kl}\mathcal{D}_k\mathcal{D}_l V^i - \frac{1}{3}h^{ik}\mathcal{D}_k\mathcal{D}_j V^j + 2N\psi^{-6}\left(h^{kl}\mathcal{D}_k\mathcal{D}_l X^i + \frac{1}{3}h^{ik}\mathcal{D}_k\mathcal{D}_l X^l\right) \\ \mathcal{D}_k\mathcal{D}_l(2N\psi^{-6})\left(\tilde{\gamma}^{kl}X^i + \frac{1}{3}\tilde{\gamma}^{ik}X^l\right) &+ \mathcal{D}_k(2N\psi^{-6})\left(2\tilde{\gamma}^{kl}\mathcal{D}_k X^i + \frac{1}{3}\tilde{\gamma}^{ik}\mathcal{D}_l X^l + \frac{1}{3}\tilde{\gamma}^{ij}\mathcal{D}_j X^k - \hat{A}^{ik}\right) = \mathcal{O}\left(\frac{1}{c^5}\right);\end{aligned} \tag{3.11}$$

$$
\begin{aligned}
&\Delta \dot{X}^j + \frac{1}{3}\mathcal{D}^j \mathcal{D}_i \dot{X}^i = \\
&\beta^k \mathcal{D}_i \mathcal{D}_k \hat{A}^{ij} - \mathcal{D}_i \hat{A}^{ik} \mathcal{D}_k \beta^j - \hat{A}^{ik} \mathcal{D}_i \mathcal{D}_k \beta^j + \frac{2}{3}\hat{A}^{ij} \mathcal{D}_i \mathcal{D}_k \beta^k + \frac{5}{3}\mathcal{D}_i \hat{A}^{ij} \mathcal{D}_k \beta^k \\
&-\frac{1}{2}N\psi^{-6}\tilde{\gamma}^{jl}\mathcal{D}_l\left(\tilde{\gamma}_{in}\tilde{\gamma}_{km}\hat{A}^{nm}\hat{A}^{ik}\right) - \psi^{-8}\tilde{\gamma}^{jl}\tilde{\gamma}_{in}\tilde{\gamma}_{km}\hat{A}^{nm}\hat{A}^{ik}\mathcal{D}_l(N\psi^2) + 8\psi^{-7}N\tilde{\gamma}^{jl}\tilde{\gamma}_{in}\tilde{\gamma}_{km}\hat{A}^{nm}\hat{A}^{ik}\mathcal{D}_l\psi \\
&+2N\psi^{-6}\mathcal{D}_i(\tilde{\gamma}_{kl}\hat{A}^{ik}\hat{A}^{jl}) - 16\psi^{-7}N\tilde{\gamma}_{kl}\hat{A}^{ik}\hat{A}^{jl}\mathcal{D}_l\psi + 2\psi^{-8}\tilde{\gamma}_{kl}\hat{A}^{ik}\hat{A}^{jl}\mathcal{D}_l(N\psi^2) \\
&-\frac{1}{2}\mathcal{D}_i(N\psi^2)\mathcal{D}_l h^{ik}\mathcal{D}_k h^{jl} - \frac{1}{6}\tilde{\gamma}^{kj}\mathcal{D}_k h^{il}\mathcal{D}_i\mathcal{D}_l(N\psi^2) - \tilde{\gamma}^{ik}\mathcal{D}_i h^{jl}\mathcal{D}_k\mathcal{D}_l(N\psi^2) \\
&-8N\tilde{\gamma}^{ik}\mathcal{D}_i h^{jl}\mathcal{D}_k\psi\mathcal{D}_l\psi + 4N\tilde{\gamma}^{jl}\mathcal{D}_l h^{ik}\mathcal{D}_i\psi\mathcal{D}_k\psi + 4\psi^{-1}\tilde{\gamma}^{ik}\mathcal{D}_i h^{jl}\left(\mathcal{D}_l(N\psi^2)\mathcal{D}_k\psi + \mathcal{D}_k(N\psi^2)\mathcal{D}_l\psi\right) \\
&-4\psi^{-1}\tilde{\gamma}^{jl}\mathcal{D}_l h^{ik}\mathcal{D}_i(N\psi^2)\mathcal{D}_k\psi + \tilde{R}^{ij}_{**}\mathcal{D}_i(N\psi^2) + N\psi^2\mathcal{D}_i\tilde{R}^{ij}_{**} - \frac{1}{2}N\psi^2\tilde{\gamma}^{ij}\mathcal{D}_i\tilde{R} \\
&-8\pi\psi^{-2}E^*\tilde{\gamma}^{jl}\mathcal{D}_l(N\psi^2) + 16\pi\psi^{-1}NE^*\tilde{\gamma}^{jl}\mathcal{D}_l\psi + 16\pi\psi^{-1}NS^*\tilde{\gamma}^{jl}D_l\psi \\
&-8\pi N\psi^{10}\mathcal{D}_i S^{ij} - 8\pi\psi^8 S^{ij}\mathcal{D}_i(N\psi^2) - 64\pi\psi^9 NS^{ij}\mathcal{D}_i\psi = \mathcal{O}\left(\frac{1}{c^4}\right),
\end{aligned}
\tag{3.12}
$$

where

$$
\tilde{R}^{ij}_{**} = \frac{1}{2}\left(-\tilde{\gamma}_{kl}\tilde{\gamma}^{mn}\mathcal{D}_m h^{ik}\mathcal{D}_n h^{jl} + \tilde{\gamma}_{nl}\mathcal{D}_k h^{mn}(\tilde{\gamma}^{ik}\mathcal{D}_m h^{jl} + \tilde{\gamma}^{jk}\mathcal{D}_m h^{il})\right) + \frac{1}{4}\tilde{\gamma}^{ik}\tilde{\gamma}^{jl}\mathcal{D}_k h^{mn}\mathcal{D}_l\tilde{\gamma}_{mn}.
$$

We can see how the elliptic sector can be solved hierarchically and this sector now is decoupled in equations including terms with progressively lower post-newtonian orders: note that in equation (3.10) we solve $N\psi^2$ as we find that it has post-newtonian order $1/c^4$ instead of $1/c^2$ as N, see [1].

In the next Section we present the first tests of the proposed modified equations in this formulation by solving the spacetime geometry of a neutron star.

4. Results and discussion

In the following we set $h^{ij} = 0$ in such a way we can compare our results with the ones obtained with the xCFC formulation which imposes this condition, see [3], and can be seen as an approximation to the FCF. As a test we use a neutron star model with an equatorial radius of $R_e = 12.859$ km, central density $\rho_c = 7.91 \cdot 10^{14}$ g/cm^3 and angular velocity $\omega = 606$ rad/s. We consider that the star is composed of a perfect fluid with polytropic equation of state $p = C\rho^\Gamma$, where p is the pressure, $\Gamma = 2$ and $C = 145731$ (cgs units). For this compact object, spacetime is stationary and we can adapt the coordinate time t to this stationarity, setting the derivatives with respect to t in (3.6) and (3.7) to zero. Moreover, this spacetime is axisymmetric; we use spherical orthonormal coordinates adapting them to this axisymmetry and fixing the rotation axis at $\theta = 0$, being θ the polar angle in spherical orthonormal coordinates.

In order to compare elliptic sectors in both xCFC and modified FCF formulations, we will focus on this work on equations (3.8)–(3.12) regarding the modified FCF scheme. By using spherical orthonormal coordinates and considering axisymmetry, we just solve equations in the 2-dimensions. The mesh consist in 100 points in the radial coordinate r and 32 points in polar angle θ. We use one ghost cell to properly compute the discretization of derivatives close to the numerical domain boundaries. All discretizations of the differential operators are 2nd order and we use the LAPACK library to invert the Laplacian operators. In some equations the variable under resolution appears outside the main Laplacian operator in the source term; here we apply fix-point iterative methods with a relaxation factor.

Concerning boundary conditions, we apply periodicity in the polar angle θ and for the radial coordinate we set $u(r,\theta) = \pm u(r,\pi-\theta)$ for the inner boundary, where u represents a generic variable and the election of the sign depends on the symmetry of this variable. At the outer boundary we impose a Robin condition, assuming $u = u_0 + M/r^n$, which is equivalent to impose $\partial u/\partial r = -n(u-u_0)/r$, and only n and u_0 need to be specified. We have $n = 1$ for scalars fields and $n = 2$ for vector fields. u_0 is the asymptotic value of the variable at spatial infinity $r \to \infty$.

In Figure 1 we show the results obtained for the numerical solutions of the variables X, ϕ and N. Only the angular component of X^i, X^ϕ, is non-zero. We plot the radial profile of these variables at $\theta = \pi/2$. Figure 2 shows the numerical solution of the new introduced vector field V^i, and the shift vector β^i directly computed from V^i. For

this vector fields again only the angular components V^{ϕ} and β^{ϕ} are non-zero. We plot the radial profile of these variables at $\theta = \pi/2$.

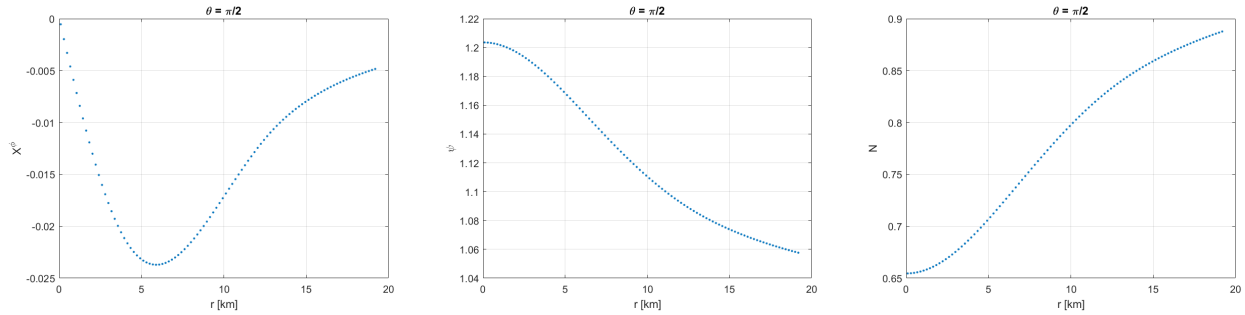


Fig. 1 Radial profiles of X^{ϕ}, the conformal factor ψ and the lapse function N at $\theta = \pi/2$.

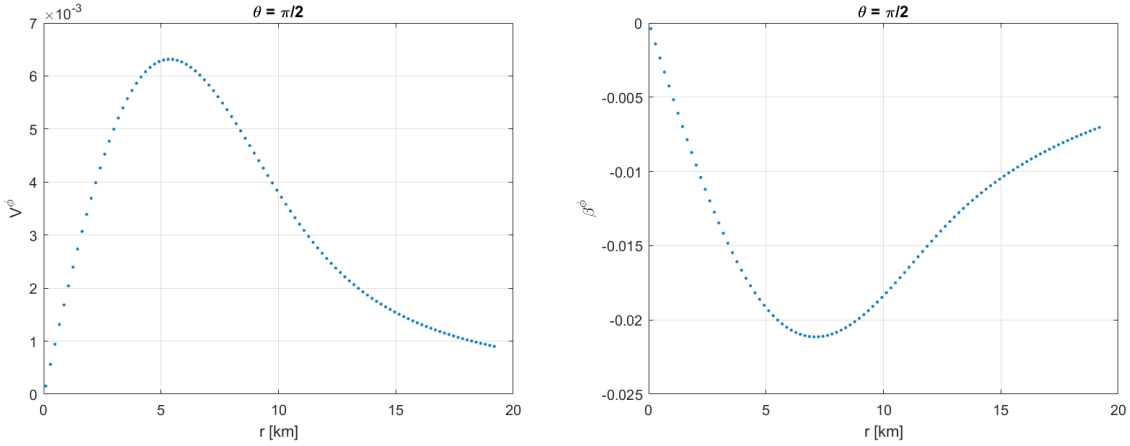


Fig. 2 Radial profiles of V^{ϕ} and third component of the shit vector β^{ϕ} at $\theta = \pi/2$.

Finally, the numerical solution for the other new vector field introduced $\dot{X}^i$ is shown in Figure 3. Here only the angular component, $\dot{X}^{\phi}$, is exactly equal to zero. Therefore, we plot the radial profile of $\dot{X}^r$ and $\dot{X}^{\theta}$ at $\theta = \pi/2$. This is the first time that the vector field $\dot{X}$ is computed numerically. Being a time derivative, it must be zero as our spacetime is stationary. Nevertheless, we get values different from zero due to the fact that the tensor h^{ij} has been neglected. In Figures 1– 3, the radial outer boundary is placed at 1.5 times the equatorial star radius R.

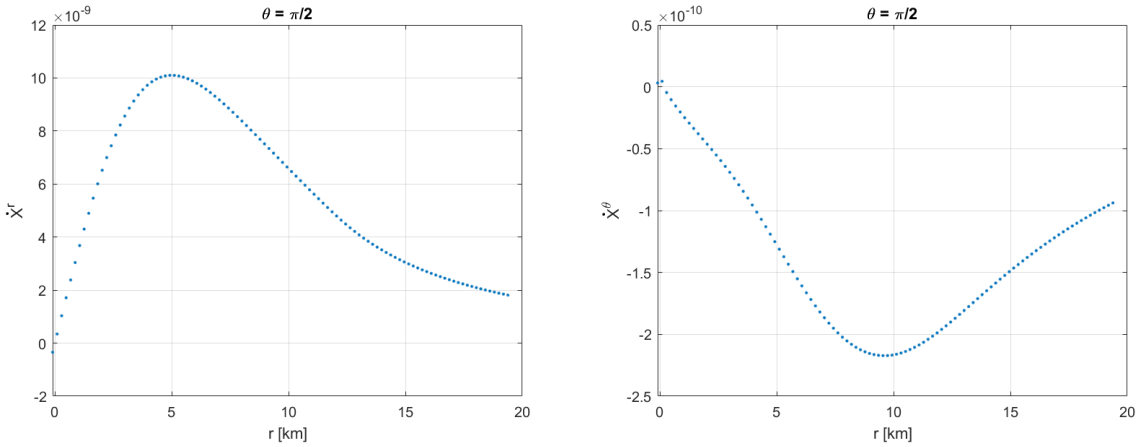


Fig. 3 Radial profiles of $\dot{X}^r$ and $\dot{X}^{\theta}$ at $\theta = \pi/2$.

A deeper analysis of the accuracy has been carried out by considering different resolutions in the radial coordinate as well as different locations of the outer boundary. Placing further the outer boundary, we expect to increase accuracy in the numerical solutions of our variables. This is expected since placing further the outer boundary translate into a closer tend to the main decay of the variables as $r \to \infty$, as we know that spacetime is asymptotically flat.

In other to numerically check the accuracy of our results, we compute the residuals as the comparison between the solution obtained with LORENE [6], and our numerical solutions; specifically,

$$\sigma(f) = \max |f - f_{\text{LORENE}}|, \tag{4.1}$$

where f and f_{LORENE} are the numerical solutions computed by us and by LORENE, respectively.

Figure 4 shows the residuals of the lapse function N, the conformal factor and the shift vector in logarithmic scale in terms of the spatial radial resolution also in logarithmic scale, employing the xCFC and modified FCF equations. Each colour refers to a specific location of the outer boundary; for example, $B/R = 1.5$ means that the outer boundary has been placed at 1.5 times the equatorial coordinate radius of the neutron star, B would be the size of the radial grid.

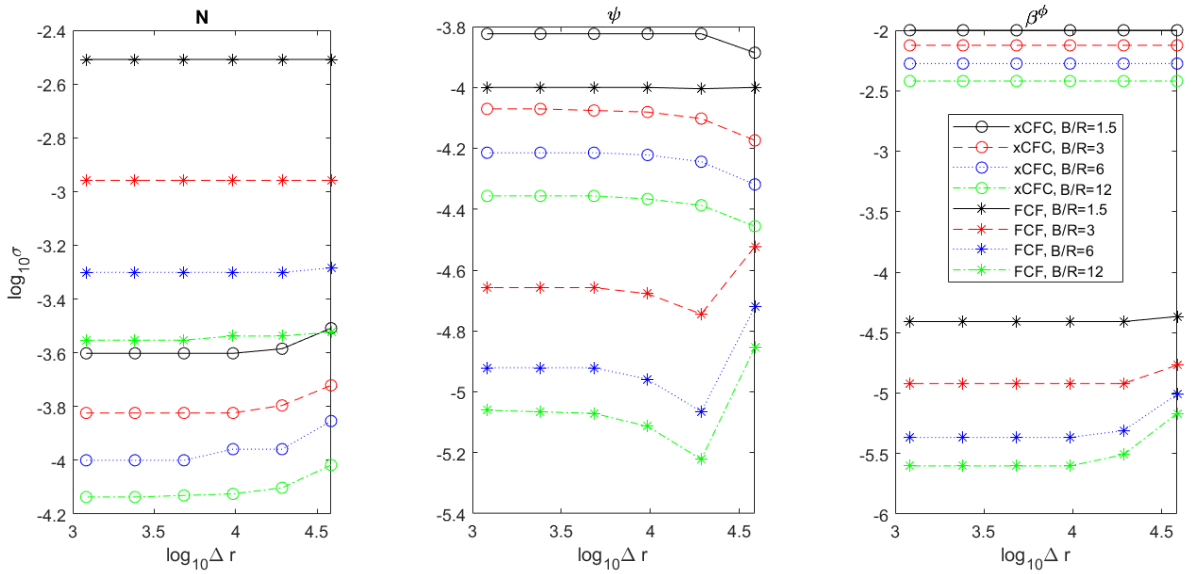


Fig. 4 Residuals of the the lapse function N (left), conformal factor ψ (center) and shift vector β^{ϕ} (right) versus the spatial resolution of the radial coordinate. Logarithmic scale is employed in both axis. Asterisks correspond to xCFC and circles to FCF. Each colour/line is linked with a grid size according to the common legend.

In all cases, the main reason to increase accuracy above a certain radial reasonable resolution is the location of the outer boundary for both the xCFC and the modified FCF equations: the further the outer boundary is located, the better the numerical accuracy for all the variables. This is a key point in the numerical resolution of the elliptic sector and we also expect to find a similar behavior for the numerical resolution of the hyperbolic sector in future works, taking into account that this hyperbolic sector encodes the gravitational radiation of the studied astrophysical scenario.

We lose some accuracy for the lapse function N in the modified FCF case, but this could be because we have to deal with a spatial derivative of the lapse function in the source terms. This fact does not happen in the xCFC case (see [3], to check the exact expressions in the elliptic sector for this formulation). The difference in the residuals decrease when increasing the radius of the location of the outer boundary and are expected to be comparable for good resolutions. We want to explore in more details the possibility of including the spatial derivative on the right hand side of (3.10) in the discretization of the elliptic operator on the left hand side, so checking weather the numerical accuracy of the lapse function can increase. When radial resolution is increased we observe a slightly improvement in the values of the residuals, but they cannot be noticed in Figures 4 due to the logarithmic scales of the vertical axis.

For the vector fields V^i and β^i, the numerical resolution of the modified FCF equations provide a much better accuracy. In the case of the shift vector β^i we gain around two orders of magnitude. This is also quite reasonable since the introduction of the V^i vector is strongly related to the shift vector. We want to remind that for the xCFC case, an elliptic equation for the shift vector is solved, see [3], while in the modified FCF equations the vector V^i is solved and then the shift vector is computed directly from the vector V^i definition.

Going from the xCFC to the modified FCF equations, we also have the possibility of including non zero h^{ij} values in the source of the equations in the elliptic sector, thus getting better accuracy in the computation of our numerical variables. This is indeed the case for a rotating neutron star, where a conformally flat metric cannot reproduce the geometry of this spacetime. This is actually a step for future works: include non-zero values for h^{ij}, include the corresponding hyperbolic sector, and take into account these terms also in the sources of the equations of the elliptic sector.

Despite the lack of reduction of the post-newtonian orders of the source terms in equation (3.7), we can take advantage of the fact that these terms should be neglected up to $1/c^6$ order to get an estimate of the h^{ij} components, by solving the following elliptic equation:

$$(\Delta h)^{ij} = 2(N\psi^2)^{-1}\Big(8Nf^{ik}f^{jl}\mathcal{D}_l\psi\mathcal{D}_k\psi - 2Nf^{ij}f^{kl}\mathcal{D}_k\psi\mathcal{D}_l\psi + f^{ik}f^{jl}\mathcal{D}_k\mathcal{D}_l(N\psi^2) \\ +(L\dot{X})^{ij} + 8\pi N\psi^{10}S^{ij} - 4\pi NS^* f^{ij}\Big).$$

This point should be further analyzed in next works to check if a better proposal can be obtained. Moreover, we also want to numerically solve these equations and check our results.

5. Conclusions

We have been accomplished the first step towards a Fully Relativistic Formulation in the Fully Constrained Formalism. Constrained formulations allow to carry out long term simulations without constraint equations violations. Besides, it posses the properties of local uniqueness, hierarchical resolutions and correct relativistic expansion with the exception of those terms mentioned in equation (3.7).

It remains for the future to check accuracy in complex numerical simulations, as well as include the hyperbolic sector of the evolution equations. This has huge impact in the calibration of gravitational waves templates. Another possible project would be to use leading terms in simplified numerical simulations for cosmological applications, e.g., to compute gravitational waves estimates in cosmological contexts, where requirements of being far away from the source (as in the famous quadrupole formula) do not apply.

References

[1] L. Blanchet, T. Damour & G. Schäefer. Post-Newtonian hydrodynamics and post-Newtonian gravitational wave generation for numerical relativity. Monthly Notices of the Royal Astronomical Society ;242(3):289-305, 1990.

[2] S. Bonazzola, E. Gourgoulhon, P. Grandclément & J. Novak. Constrained scheme for the Einstein equations based on the Dirac gauge and spherical coordinates, Physical Review D 70, 104007, 2004

[3] I. Cordero-Carrión, P. Cerdá-Durán, H. Dimmelmeier, J. L. Jaramillo, J. Novak & E. Gourgoulhon. Improved constrained scheme for the Einstein equations: An approach to the uniqueness issue. Physical Review D, 79(2), 024017, 2009

[4] I. Cordero-Carrión, J.M. Ibáñez, E. Gourgoulhon, J.L. Jaramillo & J. Novak. Mathematical issues in a fully constrained formulation of the Einstein equations. Physical Review D 77, 084007, 2008.

[5] E. Gourgoulhon. 3+1 formalism in general relativity: bases of numerical relativity (Vol. 846). Springer Science & Business Media, 2012.

[6] http://www.lorene.obspm.fr/.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 155–159)

New Galilean spacetimes to model an expanding universe

Daniel de la Fuente Benito
Universidad de Oviedo, Spain

Abstract

We introduce a new family relevant in the context of a generalized Newton-Cartan Theory: the Galilean Generalized Robertson-Walker spacetimes. We study its geometrical structure and analyse the completeness of its inextensible free falling observers. Additionally, we find some sufficient geometric conditions which guarantee a global splitting of a Galilean spacetime as a Galilean Generalized Robertson-Walker spacetime.

1. Introduction

General Relativity is so far the most accurate and successful theory to describe the spacetime structure and the gravitational phenomena. The evolution of the universe on a large scale was aptly described in the first half of the 20th century by means of the Robertson-Walker cosmological models (or fairly, Friedmann-Lemaître-Robertson-Walker models). These models assume that the matter distribution and the "space relative to the family of observers commovil with the matter" are homogeneous and isotropic. These hypotheses may be weakened in order to describe a universe in a more accurate scale. With this objective, much more recently, new cosmological models have been introduced, as the Generalized Robertson-Walker (GRW) spacetimes [5]. This kind of relativistic spacetimes has been intensively studied from a mathematical perspective (see, for instance, [8], [12, 13], [15], [19, 20].)

However, the geometric formulation of the Newtonian's Gravitation, firstly postulated by E. Cartan [10, 11], after the appearance of the Einstein's General Relativity Theory, is still of interest and significant for several reasons.

On one hand, it formulates the classical Newtonian gravitation as a covariant theory and shows that certain results previously considered as characteristic or singular of the theory of Relativity are shared by the (geometric) gravitational Newton-Cartan Theory. In fact, the Newtonian gravity also arises as a consequence of the curvature of a connection in the spacetime, which does not come from any semi-Riemannian metric. Moreover, in the geometric formulation of Newtonian's Gravity Theory, the spacetime structure is dynamical in the sense that it participates in the unfolding of physics rather than being a fixed backdrop against which it unfolds (see [16] and classical references therein).

On the other hand, it allows to establish from an accurate and intrinsic way the limit relation between the Newtonian theory of Gravitation and General Relativity.

The notion of symmetry is clearly basic in Physics. On a geometrical spacetime model, symmetry is usually based on the assumption of the existence of a one-parameter group of transformations generated by a Killing or, more generally, by a conformal vector field (see, [22]). Another important question is that a geometric approach enables possible generalizations of Newtonian Theory, via the assumption of certain symmetries on Galilean spacetimes (see Section 2), which are the geometrical "arena" for the Newton-Cartan gravitation. So, in [17] the author studies the symmetry imposed on a Galilean spacetime by the cosmological principle, obtaining the Galilean model analogous to the relativistic Robertson-Walker spacetimes.

In this work, we introduce a new family of Galilean geometrical models, which generalize the non-relativistic Robertson-Walker spacetimes, in the same way as GRW spacetimes generalize the Friedmann-Lemaître-Robertson-Walker spacetimes: the Galilean Generalized Robertson-Walker (GGRW) spacetimes (Sect. 3). A GGRW spacetime possesses an infinitesimal symmetry given by the existence of a timelike irrotational conformally Leibnizian (ICL) vector field. Several geometrical properties and physical interpretations for this family of spacetimes are given in Section 3, as the possible existence of singularities or the completeness of its free falling observers. Section 4 is devoted to the study of Galilean spacetimes admitting a timelike irrotational conformally Leibnizian vector field . We show that an ICL Galilean spacetime must be locally a GGRW spacetime. Finally, Section 5 is devoted to face the following kind of splitting problems: under what geometrical assumptions an ICL spacetime globally decomposes as a GGRW spacetime.

2. Set up

Recall that a *Leibnizian* structure on a (non-relativistic) spacetime M is a pair (Ω, g) consisting of a differential 1-form $\Omega \in \Lambda^1(M)$, nowhere null ($\Omega_p \neq 0, \forall p \in M$) and a positive definite metric g on its kernel. Specifically, let us

 ISBN 978-84-18482-21-2

denote by $\mathrm{An}(\Omega) = \{v \in TM\,,\ \Omega(v) = 0\}$ the smooth n-distribution induced on M by Ω. If we denote by $\Gamma(TM)$ the set of smooth vector fields on M, we may construct the subset $\Gamma(\mathrm{An}(\Omega)) = \{V \in \Gamma(TM) \,/\, V_q \in \mathrm{An}(\Omega),\ \forall p \in M\}$. So, the map

$$g : \Gamma(\mathrm{An}(\Omega)) \times \Gamma(\mathrm{An}(\Omega)) \longrightarrow C^\infty(M),\ \ (V, W) \mapsto g(V, W),$$

is smooth, bilinear, symmetric and positive definite. Hence, M is endowed with a sub-Riemannian structure defined on the bundle $\mathrm{An}(\Omega)$, i.e., the annihilator of the degenerate metric $\Omega \otimes \Omega$ (see [6] and [7], for details). The triad, $(M, \Omega, g))$ is called Leibnizian spacetime.

Points of M are usually called *events*. The Euclidean vector space $(\mathrm{An}(\Omega_p)\,, g_p)$ is called the *absolute space* at $p \in M$, and the linear form Ω_p is the *absolute clock* at p. A tangent vector $v \in T_pM$ is named *spacelike* if $\Omega_p(v) = 0$ and, otherwise, *timelike*. Additionally, if $\Omega_p(v) > 0$ (resp. $\Omega_p(v) < 0$), v points out the *future* (resp. the *past*).

An *observer* in a Leibnizian spacetime M is a timelike future unit smooth curve $\gamma : J \longrightarrow M$, i.e., its velocity γ' satisfies that $\Omega(\gamma'(s)) = 1$ for all $s \in J$. The parameter s is called the *proper time* of the observer γ. A vector field $Z \in \Gamma(TM)$ with $\Omega(Z) = 1$ is called a *field of observers*, this is, its integral curves are observers.

When the smooth distribution $\mathrm{An}(\Omega)$ is integrable (equivalently, if the absolute clock Ω satisfies $\Omega \wedge d\Omega = 0$), the Leibnizian spacetime (M, Ω, g) is said to be *locally sincronizable*, and making use of the Frobenius Theorem (see [21]), it may be foliated by a family of spacelike hypersurfaces $\{\mathcal{F}_\lambda\}$. In this case, it is well-known that each $p \in M$ has a neighbourhood where $\Omega = f\,dt$, for certain smooth functions $f > 0$, t, and the hypersurfaces $\{t = \text{constant}\}$ locally coincide with a leaf of the foliation $\mathcal{F}$. Thus, any observer may be synchronized through the "compromise time" t, obtained rescaling its proper time. In the more restrictive case $d\Omega = 0$, then the Leibnizian spacetime (M, Ω, g) is called *proper time locally synchronizable*, and one has, locally, $\Omega = dt$. Now, observers are synchronized directly by its proper time (up to a constant). When Ω is exact, $\Omega = dt$ for some function $t \in C^\infty(M)$, which is called the *absolute time function*. In this case, any observer may be assumed to be parametrized by t. Notice that the notion of (local and local proper time) synchronizability is intrinsic to the Leibnizian structure, applicable for every observer, in contrast to the relativistic setting, where the analogous concepts have meanings only for fields of observers.

According to [7], a field of observers is called *Leibnizian* if the stages Φ_s of its local flows are *Leibnizian diffeomorphisms*, that is, they preserve the absolute clock and space, i.e.,

$$\Phi_s^*\Omega = \Omega, \qquad \text{and} \qquad \Phi_s^* g = g.$$

On the other hand, the inertia principle must be codified through a connection on the spacetime. However, a Leibnizian structure has not a canonical affine connection associated. Then, it is required to introduce a compatible connection with the absolute clock Ω and the space metric g, i.e., a connection ∇ such that

(a) $\nabla\Omega = 0$ (equivalently, $\Omega(\nabla_X Y) = X(\Omega(Y))$ for any $X, Y \in \Gamma(TM)$).

(b) $\nabla g = 0$ (i.e., $Z(g(V, W)) = g(\nabla_Z V, W) + g(\nabla_Z W, V)$ for any $Z \in \Gamma(TM)$ and V, W spacelike vector fields).

Such a connection is named *Galilean*. A *Galilean spacetime* (M, Ω, g, ∇) is a Leibnizian spacetime endowed with a Galilean connection ∇. In addition, ∇ is said *symmetric* if its torsion vanishes identically ($\mathrm{Tor}_\nabla(X, Y) = \nabla_X Y - \nabla_Y X - [X, Y] \equiv 0$). From a physical point of view, a symmetric connection is desirable since it is completely determined by its geodesics, i.e., by the free falling observers of M. From now on, we will only consider symmetric Galilean connections on the spacetime.

Given two Galilean spacetimes (M, Ω, g, ∇) and $(M', \Omega', g', \nabla')$, a diffeomorphism $F : M \longrightarrow M'$ is said to be *Galilean* if $F^*\Omega' = \Omega$, $F^*g' = g$ and $F^*\nabla' = \nabla$, i.e., $\nabla'_{dF(X)} dF(Y) = \nabla_X Y$.

For each fixed field of observers Z on a Galilean spacetime (M, Ω, g, ∇), the *gravitational field* induced by ∇ in Z is given by the spacelike vector field $\mathcal{G} = \nabla_Z Z$. The *vorticity* or *Coriolis field* of Z is the 2−form $\omega(Z) = \frac{1}{2}\mathrm{Rot}(Z)$, defined as

$$\omega(Z)(V, W) = \frac{1}{2}\Big(g(\nabla_V Z, W) - g(\nabla_W Z, V)\Big) \qquad \forall V, W \in \Gamma(\mathrm{An}(\Omega)).$$

The main result of [7, Th.5.27] claims that, for a fixed field of observers Z on a Leibnizian spacetime (M, Ω, g) with $d\Omega = 0$, the set of all symmetric Galilean connections is bijectively mapped onto $\Big(\Gamma(TM), \Lambda^2(\mathrm{An}(\Omega))\Big)$. Each symmetric Galilean connection ∇ is mapped to $\Big(\mathcal{G}(Z), \mathrm{Rot}(Z)\Big)$. Thus, the gravitational field and the vorticity of a field of observers determine a unique symmetric Galilean geometry of the spacetime.

Additionally, a Leibnizian field of observers Z in a Galilean spacetime (M, Ω, g, ∇) is named *Galilean* if it is affine for ∇, that is, $L_Z \nabla = 0$, where L denotes the Lie derivative. Finally, a Galilean spacetime is said *Newtonian* if the (symmetric) connection ∇ restricted to the spacelike vectors is flat, and it admits an irrotational Galilean field of observers. This kind of spacetimes has traditionally represented the classical (non-relativistic) geometric model of gravity.

3. Galilean Generalized Robertson-Walker spacetimes

In this section we introduce a new family of Galilean geometric models, which are the classical version of the relativistic Generalized Robertson-Walker spacetimes defined in [5].

Definition 3.1 Let $I \subseteq \mathbb{R}$ be a real interval, (F, h) a n-dimensional connected Riemannian manifold, and $f \in C^\infty(I)$ a smooth positive function on I. A Galilean spacetime (M, Ω, g, ∇) is called *Galilean Generalized Robertson-Walker spacetime (GGRW)* if $M = I \times F$, $\Omega = d\pi_I$, g is the restriction to the bundle $\mathrm{An}(\Omega)$ of the following (degenerate) metric on M,

$$\overline{g} = (f \circ \pi_I)^2 \, \pi_F^* h, \tag{3.1}$$

where π_I, π_F are the canonical projections onto the open interval I and the fiber F respectively, and ∇ is the only symmetric Galilean connection on M such that

$$\nabla_{\partial_t} \partial_t = 0, \qquad \textit{and} \qquad \mathrm{Rot}\, \partial_t = 0, \tag{3.2}$$

where $\partial_t = \partial/\partial t$ is the global coordinate vector field associated to $t := \pi_I$.

The vector field ∂_t defines a field of observers in M ($\Omega(\partial_t) = 1$), which we will call *commovil observers*, by the similarity with the relativistic Robertson-Walker spacetimes. Then, the conditions (3.2) in above definition mean that commovil observers are free falling and they do not rotate. Notice that from [7, Th.5.27], the conditions (3.2) determine the (symmetric) Galilean connection on M.

Example Let us consider a GGRW with $I = \mathbb{R}$ and $F = \mathbb{R}^n$ endowed with the usual Euclidean metric. If $f(t) =$ constant, then the Galilean connection coincides with the standard flat connection of the affine space $\mathbb{R}^{n+1}$. In addition, the commovil observers satisfy the necessary conditions to assure the Newtonian character of this spacetime. More physically relevant examples are given in the next section.

3.1. Completeness of free falling observers in a GGRW spacetime

We now proceed to analyze when the inextensible free falling trajectories in a GGRW spacetime are complete. Physically we are looking for geometric assumptions that guarantee that every free falling observer *lives forever*.

First, we have an analogous result to the geodesic normalization lemma in semi-Riemannian manifolds.

Lemma 3.2 *Let γ be a geodesic in a GGRW spacetime. Then, $\Omega(\gamma')$ is constant along the trajectory of γ.*

The relevant cases correspond with $\Omega(\gamma') = 0$ or 1. The first one ($\Omega(\gamma') = 0$) means that γ is spacelike and contained in a leaf $\mathcal{F}_t$ of the foliation of Ω. As ∇ coincides with the Levi-Civita connection of $(\mathcal{F}_t, f(t)^2 \, h)$, the completeness of this kind of geodesics is equivalent to the geodesic completeness of (F, h). Thus, from now on we will deal with free falling observers (γ geodesic with $\Omega(\gamma') = 1$).

Theorem 3.3 *A GGRW spacetime is geodesically complete if and only if $I = \mathbb{R}$ and the fiber (F, h) is (geodesically) complete.*

4. Irrotational conformally Leibnizian spacetimes

In this section we present a wider family of Galilean spacetimes which locally exhibit the structure of a GGRW spacetime. As a previous step, we introduce the concept of conformally Leibnizian field of observers, generalizing the well-known notion of Leibnizian observer.

Definition 4.1 Let (M, Ω, g) be a Leibnizian spacetime. A vector field X is called *spatially conformally Leibnizian* vector field if

$$L_X \Omega = \mu \, \Omega, \tag{4.1}$$

and the Lie derivative of the absolute space metric satisfies

$$L_X g = 2\lambda \, g, \tag{4.2}$$

for some smooth functions $\lambda, \mu \in C^\infty(M)$. If, additionally, both functions coincide, i.e., $\lambda = \mu$, then X is named *conformally Leibnizian* vector field.

Note that a conformally Leibnizian vector field is Leibnizian if and only if the conformal factor λ is identically zero [7].

Remark 4.2 Condition (4.1) may be also expressed as

$$d\Omega(X,Y) + Y(\Omega(X)) = \mu\,\Omega(Y), \qquad \forall Y \in \Gamma(TM),$$

and means that distribution $\mathrm{An}(\Omega)$ is invariant along the flow of vector field X. So, if this distribution is integrable, the flow of X carries each leaf of the foliation to another one. Analogously, assumption (4.2) is equivalent to

$$X(g(V,W)) = \lambda\, g(V,W) + g([X,V],W) + g([X,W],V), \qquad \forall V, W \in \Gamma(\mathrm{An}(\Omega)).$$

The following result shows that GGRW spacetimes admit a timelike conformally Leibnizian vector field.

Proposition 4.3 *Let $(M = I \times F, \Omega = dt, g, \nabla)$ be a GGRW spacetime with scale factor $f \in C^\infty(I)$. Then, the vector field $K := (f \circ \pi_I)\,\partial_t$ is irrotational and conformally Leibnizian and, consequently, it satisfies the identity*

$$\nabla_X K = (f' \circ \pi_I)\, X, \qquad \forall X \in \Gamma(TM). \tag{4.3}$$

Definition 4.4 Let (M, Ω, g, ∇) be a Galilean spacetime, whose absolute clock is closed ($d\Omega = 0$). If M admits a timelike vector field $K \in \Gamma(TM)$ satisfying

$$\nabla_X K = \rho\, X, \qquad \forall X \in \Gamma(TM),\ \textit{where}\ \rho \in C^\infty(M), \tag{4.4}$$

M is called *Irrotational Conformally Leibnizian spacetime* (ICL).

Remark 4.5 Notice that condition (4.4) directly implies that K is conformally Leibnizian and $\mathrm{Rot}(K)(V,W) = 0$, for all spacelike vector fields V, W.

As a first consequence of Definition 4.4, we obtain that functions $\Omega(K)$ and ρ are constant on each leaf of the foliation induced by Ω.

Lemma 4.6 *Let (M, Ω, g, ∇) be a ICL spacetime with irrotational conformally Leibnizian vector field K and conformal factor ρ. Then*

$$V\big(\Omega(K)\big) = 0 \quad \textit{and} \quad V(\rho) = 0, \qquad \forall V \in \Gamma(\mathrm{An}(\Omega)).$$

We have just seen that each GGRW is an ICL spacetime. Next theorem ensures that any ICL spacetime is locally a GGRW spacetime.

Theorem 4.7 *Let (M, Ω, g, ∇) be an ICL spacetime. For each $p \in M$, there exist an open neighbourhood of p, $\mathcal{U}$, and a Galilean diffeomorphism $\Psi : N \longrightarrow \mathcal{U}$, where N is a GGRW spacetime.*

5. Global GGRW decompositions

We know that an ICL spacetime is locally a GGRW spacetime. Now, our aim here consists in looking for additional assumptions on the geometry of an ICL spacetime which lead to a global splitting as a GGRW spacetime. This type of question has been yet discussed several times in the relativistic setting (see for instance, [8], [14], [15] and [4]), i.e., under what conditions on the geometry of a relativistic spacetime, this admits a global decomposition as a warped product space or, in particular, as a GRW spacetime.

Theorem 5.1 *A Gailean spacetime $\big(M, \Omega, g, \nabla\big)$, whose 1-form Ω is exact, admits a global decomposition as a GGRW spacetime if and only if it is an ICL spacetime with a timelike irrotational conformally vector field K, such that the flow of the associated field of observers, $Z := \frac{1}{\Omega(K)} K$, is well defined and onto in a domain $I \times \mathcal{F}$ for some interval $I \subseteq \mathbb{R}$ and some leaf of the foliation $\mathcal{F}$ induced by Ω.*

Remark 5.2 (i) Note that the hypothesis on the absolute clock Ω automatically holds when the spacetime is simply connected. (ii) Observe that the assumption on the flow of Z trivially holds when Z is complete.

Taking into account the previous Remark, we can assert

Corollary 5.3 *Let (M, Ω, g, ∇) be an ICL spacetime with timelike irrotational conformally Leibnizian vector field K. If the absolute clock Ω is exact and $\frac{1}{\Omega(K)}K$ is complete, then M globally splits as a GGRW spacetime.*

To end this work, we present a global splitting result when the spacetime is spatially compact, that is, when the leaves of the spacelike foliation are compact.

Theorem 5.4 *Let (M, Ω, g, ∇) be an ICL spacetime with Ω exact. If the leaves of the foliation induced by Ω are compact, then M is a GGRW spacetime.*

Acknowledgements

The author was partially supported by Spanish MINECO and ERDF project MTM2016-78807-C2-1-P.

References

[1] David Hilbert. Ueber die nothwendigen und hinreichenden covarianten Bedingungen für die Darstellbarkeit einer binären Form als vollständiger Potenz. *Math. Ann.*, 27(1):158–161, 1886.

[2] Donald E. Knuth. *Tau Epsilon Chi, a system for technical text*. American Mathematical Society, Providence, R.I., 1979. Revised version of Stanford Computer Science report number STAN-CS-78-675.

[3] Laurent Schwartz. Généralisation de la notion de fonction, de dérivation, de transformation de Fourier et applications mathématiques et physiques. *Ann. Univ. Grenoble. Sect. Sci. Math. Phys. (N.S.)*, 21:57–74 (1946), 1945.

[4] J.A. Aledo, A. Romero and R.M. Rubio, The existence and uniqueness of standard static splitting, *Classical Quant. Grav.*, **32** (2015), 105004 (1–9).

[5] L.J. Alías, A. Romero and M. Sánchez, Uniqueness of complete spacelike hypersurfaces of constant mean curvature in Generalized Robertson-Walker spacetimes, *Gen. Relat. Gravit.*, **27** (1995), 71–84.

[6] A.N. Bernal, M. López, M. Sánchez, Fundamental units of length and time, *Found. Phys.*, **32** (2002), 77–108.

[7] A.N. Bernal, M. Sánchez, Leibnizian, Galilean and Newtonian structures of space-time *J. Math. Phys.*, **44** (2003), 1129–1149.

[8] M. Caballero, A. Romero and R.M. Rubio, Constant mean curvature spacelike hypersurfaces in Lorentzian manifolds with a timelike gradient conformal vector field, *Class. Quantum Grav.*, **28** (2011), 145009–145022.

[9] A.M. Candela, A. Romero and M. Sánchez, Completeness of the trajectories of particles coupled to a general force field, *Arch. Rational Mech. Anal.*, **208** (2013), 255–274.

[10] E. Cartan, Les variètés a conexion affine, *Ann. Ec. Norm. Sup.*, **40** (1923), 1–25.

[11] E. Cartan, Les variètés a conexion affine (suite), *Ann. Ec. Norm. Sup.*, **41** (1924), 325–412.

[12] D. De la Fuente, J.A. S. Pelegrín, R.M. Rubio, On the geometry of stationary Galilean spacetimes, *Gen. Relativ. Grav.*, **53**8 (2021), 1–15.

[13] J.L Flores, M. Sánchez, Geodesic connectedness and conjugate points in GRW space-times, *J. Geom. Phys.*, **3**6 (2000), 285–314.

[14] J.L. Flores, The Riemannian and Lorentzian splitting theorems, *Atlantis Trans. Geom.*, Springer **1** (2017), 1–20.

[15] M. Gutierrez and B. Olea, Global decomposition of a Lorentzian manifold as a Generalized Robertson-Walker spacetime, *Diff. Geom. Appl.*, **27** (2009), 146–156.

[16] D. B. Malament, *Topic in the Formulations of General Relativity and Newtonian Gravitation Theory*, Chicago lectures in Physics, University of Chicago Press, 2012.

[17] F. Müler-Hoissen, The cosmological principle and a generalization of Newton's theory of gravitation, *Gen. Relat. Gravitation*, **15** (1983), 1051–1066.

[18] B. O'Neill, *Semi–Riemannian Geometry with Applications to Relativity*, Pure Appl. Math., **103**, Academic Press, New York, 1983.

[19] M. Sánchez, On the geometry of generalized Robertson-Walker spacetimes: geodesics *Gen. Relat. Gravitation* **3**0 (1998), 915–932.

[20] M. Sánchez, On the geometry of generalized Robertson-Walker spacetimes: curvature and Killing fields. *J. Geom. Phys.*, **31** (1999), 1—15.

[21] F.W. Warner, *Foundations of Differentiable Manifolds and Lie Groups*, Grad. Texts Math., Springer-Verlag, 1983.

[22] E. Zafiris, Irreducible decomposition of Einstein's equations in spacetimes with symmetries, *Ann. Phys.* **263** (1998), 155–78.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 160–166)

Numerical approximation of dispersive shallow flows on spherical coordinates

Cipriano Escalante[1], **Manuel J. Castro**[2]
1. cescalante@uco.es *Universidad de Córdoba, Spain*
2. *Universidad de Málaga, Spain*

Abstract

This work aims to develop and implement a numerical model, including dispersion suitable for tsunami simulations. A latitude-longitude coordinate formulation to account for the effects of curvature is presented. A relaxation procedure is then applied to obtain a system of balance laws amenable to be discretized with explicit and efficient numerical methods. Here we follow [5] to develop an efficient finite-volume numerical method. The resulting numerical model has been applied to an experimental test case, which shows the efficiency and accuracy of the method.

1. Introduction

In fluid dynamics, dispersion of waves refers, in general, to frequency dispersion. That means that waves of different wavelengths travel at different celerity. Water waves propagate on the water surface, with gravity and surface tension as the restoring forces. As a result, water with a free surface is considered a dispersive medium. It is well-known that the usual shallow water equations (SW) do not consider the effects of dispersive waves.

Figure 1 illustrates this fact showing snapshots of the evolution of a wave over a plane beach. There, one can see how the (SW) (in black) tend to predict faster velocity for the front of the wave when compared with laboratory data (in red). The Stokes linear theory (or Airy wave theory) explains this situation. It states that the speed of wave propagation, or more precisely the phase velocity C_{Airy}, is a quantity that is given in terms of the typical depth H and the local wave-number k, more explicitly

$$C_{Airy}^2 = gH\frac{\tanh(kH)}{kH}, \tag{1.1}$$

whereas the phase velocity of the (SW) is given by $C_{SW}^2 = gH$. The previous relation also called a linear dispersion relation, reveals the dispersive character of the water wave theory. Therefore (SW) cannot take into account the effects associated with dispersive waves. That also explains the shifting on the computed numerical simulation in Figure 1, since the speed propagation of the (SW), C_{SW}, is faster than the one given by the linear theory, C_{Airy}.

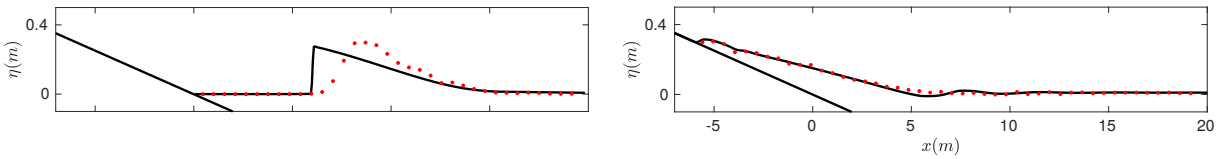


Fig. 1 Comparison of experiments data (red) and simulated ones with (SW) (black) at different times.

Concerning mathematical models that can simulate dispersive water waves, a great effort has been made in recent years to derive systems for shallow water flows that include long, non-linear water waves, such as Tsunami water waves. The development of non-hydrostatic pressure models for coastal water waves has been the topic of many studies over the past 30 years. These models can solve many relevant features of coastal water waves, such as dispersion, non-linearity, shoaling, refraction, diffraction, and run-up. The central hypothesis in the derivation consists of splitting the pressure into a hydrostatic and a non-hydrostatic part (see Casulli [6]). In this work, the non-hydrostatic pressure system derived by Bristeau *et al.* in [3] written in spherical coordinates is considered.

Concerning the nature of non-hydrostatic pressure systems, it is well known that they differ from a hyperbolic system and responds instead to a mixed hyperbolic and elliptic problem. Due to the mixed hyperbolic-elliptic nature of non-hydrostatic systems, the complexity of the corresponding numerical schemes increases. For example, the incompressibility equation appearing in the equations introduced in [3] makes the system a hyperbolic-elliptic problem. This restriction makes that explicit schemes cannot be applied to the system since they may have a very restrictive stability condition, or even worse, it may result in an unconditionally unstable method. Therefore,

implicit schemes must be applied, and several works can be found in the literature (see, for example, [11, 16, 17] and references therein). Numerical methods applied to non-hydrostatic pressure systems typically use a projection-correction type scheme. Usually, it combines finite-volume techniques to solve the underlying hyperbolic part in a first step and finite-differences or finite-elements for solving the elliptic or non-hydrostatic dispersive terms in a second, involving the resolution of a linear system at each time step.

However, there is a recent new alternative to simulate dispersive water waves with hyperbolic PDE systems (see [2, 9, 10, 15] and references therein). In the same vein, we propose a novel first-order system of balance laws in this work that can be seen as a modification of the model. The novel system is obtained using a reformulation of the original governing equations written in spherical coordinates by coupling the divergence constraint of the velocity with the remaining balance laws. That is done with the aid of an evolution equation for the depth-integrated non-hydrostatic pressure, similar to the so-called hyperbolic divergence cleaning considered in [9, 10]. Therefore, the final governing PDE system introduced here is a system of balance laws and is thus amenable for an explicit discretization via high-order numerical schemes.

The organization of this paper is as follows: in the next section, the PDE system in spherical coordinates is introduced. In Section 3, some references for the design of a well-balanced Finite Volume numerical scheme are given. Some numerical comparisons are presented in Section 4 to check the efficiency and the ability of the method to simulate planetary waves or tsunami waves over realistic bathymetry. Finally, some conclusions are drawn.

2. PDE system

2.1. The hyperbolic-elliptic non-hydrostatic pressure system in spherical coordinates

We consider the non-hydrostatic system first derived by Bristeau *et al* in [3] that can model dispersive non-hydrostatic free-surface flows. The governing PDE are obtained by a process of depth-averaging of the incompressible Euler equations with respect to the vertical direction. The total pressure is decomposed into a sum of hydrostatic and non-hydrostatic pressure. The governing equations are given by

$$\begin{cases} \partial_t h + \partial_x q_x + \partial_y q_y = 0, \\ \partial_t q_x + \partial_x \left(\dfrac{q_x^2}{h} + q_p \right) + \partial_y \left(\dfrac{q_x q_y}{h} \right) + (gh + 2p)\, \partial_x \eta - 2p \partial_x h = 0, \\ \partial_t q_y + \partial_x \left(\dfrac{q_x q_y}{h} \right) + \partial_y \left(\dfrac{q_y^2}{h} + q_p \right) + (gh + 2p)\, \partial_y \eta - 2p \partial_y h = 0, \\ \partial_t q_w + \partial_x \left(\dfrac{q_x q_w}{h} \right) + \partial_y \left(\dfrac{q_y q_w}{h} \right) = 2p, \\ \partial_x q_x + \partial_y q_y + \dfrac{q_x}{h} \partial_x (h - 2\eta) + \dfrac{q_y}{h} \partial_y (h - 2\eta) + 2\dfrac{q_w}{h} = 0, \end{cases} \tag{2.1}$$

where g is the gravity; h is the thickness of the water layer; H is the bottom depth and q_x, q_y, q_w are the depth-averaged discharges in the x, y, and z direction respectively. The depth-averaged non-hydrostatic pressure is denoted by $p = \dfrac{q_p}{h}$.

We will describe here a summary of the followed process to write the governing equations in spherical coordinates. First, we will consider the underlying hydrostatic system (SW) that can be obtained from (2.1) by setting $p = 0$ and suppressing the last incompressibility condition. To do that, we follow [5], and the underlying hydrostatic (SW) reads

$$\begin{cases} \partial_t h_\sigma + \dfrac{1}{R} \left(\partial_\theta \left(\dfrac{Q_\theta}{\cos(\varphi)} \right) + \partial_\varphi Q_\varphi \right) = 0, \\ \partial_t Q_\theta + \dfrac{1}{R} \partial_\theta \left(\dfrac{Q_\theta^2}{h_\sigma \cos(\varphi)} \right) + \dfrac{1}{R} \partial_\varphi \left(\dfrac{Q_\theta Q_\varphi}{h_\sigma} \right) - \dfrac{Q_\theta Q_\varphi}{R h_\sigma} \tan(\varphi) + \dfrac{g h_\sigma}{R \cos^2(\varphi)} \partial_\theta \eta_\sigma = 0, \\ \partial_t Q_\varphi + \dfrac{1}{R} \partial_\theta \left(\dfrac{Q_\theta Q_\varphi}{h_\sigma \cos(\varphi)} \right) + \dfrac{1}{R} \partial_\varphi \left(\dfrac{Q_\varphi^2}{h_\sigma} \right) + \left(\dfrac{Q_\varphi^2}{R h_\sigma} + \dfrac{g h_\sigma \eta_\sigma}{R \cos(\varphi)} \right) \tan(\varphi) + \dfrac{g h_\sigma}{R \cos(\varphi)} \partial_\varphi \eta_\sigma = 0, \end{cases} \tag{2.2}$$

where

$$h_\sigma = h \cos(\varphi),\ H_\sigma = H \cos(\varphi),\ \eta_\sigma = h_\sigma - H_\sigma,\ Q_\varphi = q_\varphi \cos(\varphi),\ Q_\theta = q_\theta \cos(\varphi),$$

denote the conserved variables, R is the radius, (θ, φ) the longitude and latitude, and $q_\theta,\ q_\varphi$ are the longitudinal and latitudinal averaged discharges in the normal direction (Fig. 2).

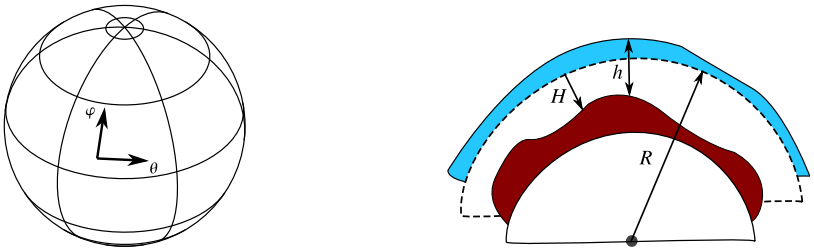


Fig. 2 Sketch of the unknowns for the system in spherical coordinates.

Now, we proceed to include the non-hydrostatic terms as well as the incompressibility condition written in spherical coordinates. To do that, as in [5], we consider the gradient and divergence operators in spherical coordinates:

$$\nabla(f) = \left(\frac{1}{R\cos(\varphi)}\partial_\theta f \quad \frac{1}{R}\partial_\varphi h\right), \quad \nabla\cdot\vec{f} = \frac{1}{R\cos(\varphi)}\left(\partial_\theta f_1 + \partial_\varphi(f_2\cos(\varphi))\right).$$

Taking that into account, the non-hydrostatic pressure system (2.1) can be written in spherical coordinates as follows

$$\begin{cases}
\partial_t h_\sigma + \dfrac{1}{R}\left(\partial_\theta\left(\dfrac{Q_\theta}{\cos(\varphi)}\right) + \partial_\varphi Q_\varphi\right) = 0, & \text{(2.3a)}\\
\partial_t Q_\theta + \dfrac{1}{R}\partial_\theta\left(\dfrac{Q_\theta^2}{h_\sigma\cos(\varphi)} + \dfrac{Q_p}{\cos(\varphi)}\right) + \dfrac{1}{R}\partial_\varphi\left(\dfrac{Q_\theta Q_\varphi}{h_\sigma}\right) - \dfrac{Q_\theta Q_\varphi}{Rh_\sigma}\tan(\varphi) & \text{(2.3b)}\\
\qquad + \dfrac{gh_\sigma + 2p_\sigma}{R\cos^2(\varphi)}\partial_\theta\eta_\sigma - 2\dfrac{p_\sigma}{R\cos^2(\varphi)}\partial_\theta h_\sigma = 0 & \\
\partial_t Q_\varphi + \dfrac{1}{R}\partial_\theta\left(\dfrac{Q_\theta Q_\varphi}{h_\sigma\cos(\varphi)}\right) + \dfrac{1}{R}\partial_\varphi\left(\dfrac{Q_\varphi^2}{h_\sigma} + Q_p\right) & \text{(2.3c)}\\
\qquad + \left(\dfrac{Q_\varphi^2}{Rh_\sigma} + \dfrac{gh_\sigma\eta_\sigma}{R\cos(\varphi)}\right)\tan(\varphi) + \dfrac{gh_\sigma + 2p_\sigma}{R\cos(\varphi)}\partial_\varphi\eta_\sigma - 2\dfrac{p_\sigma}{\cos(\varphi)}\partial_\varphi h_\sigma = 0, & \\
\partial_t Q_w + \dfrac{1}{R}\left(\partial_\theta\left(\dfrac{Q_\theta Q_w}{h}\right) + \partial_\varphi\left(\dfrac{Q_\varphi Q_w}{h}\right)\right) = 2p_\sigma, & \text{(2.3d)}\\
\dfrac{1}{R}\partial_\theta\left(\dfrac{Q_\theta}{\cos(\varphi)}\right) + \partial_\varphi Q_\varphi + \dfrac{1}{R}\left(\dfrac{Q_\theta}{h_\sigma}\partial_\theta\left(\dfrac{h_\sigma - 2\eta_\sigma}{\cos(\varphi)}\right) + \dfrac{Q_\varphi}{h_\sigma}\partial_\varphi\left(\dfrac{h_\sigma - 2\eta_\sigma}{\cos(\varphi)}\right)\right) + 2w_\sigma = 0 & \text{(2.3e)}
\end{cases}$$

where $Q_p = hp_\sigma,\ p_\sigma = p\cos(\varphi)$, and $Q_w = q_w\cos(\varphi)$.

2.2. The relaxed non-hydrostatic pressure system of balance laws in spherical coordinates

Here we follow the standard ideas described in [9, 10], where authors obtain a hyperbolic relaxation system from the hyperbolic-elliptic equations introduced in [3] in Cartesian coordinates. Therefore, we replace the last incompressibility condition in (2.3e) by the relaxed equation

$$\begin{aligned}
&\partial_t Q_p + \frac{1}{R}\left(\partial_\theta\left(\frac{Q_\theta Q_p}{h_\sigma\cos(\varphi)} + c^2\frac{Q_\theta}{\cos(\varphi)}\right) + \partial_\varphi\left(\frac{Q_\varphi Q_p}{h_\sigma\cos(\varphi)} + c^2 Q_\varphi\right)\right)\\
&+\frac{c^2}{R}\frac{Q_\theta}{h_\sigma}\partial_\theta\left(\frac{h_\sigma - 2\eta_\sigma}{\cos(\varphi)}\right) + \frac{c^2}{R}\frac{Q_\varphi}{h_\sigma}\partial_\varphi\left(\frac{h_\sigma - 2\eta_\sigma}{\cos(\varphi)}\right) = -2c^2 w_\sigma,
\end{aligned} \tag{2.4}$$

where $c = \alpha\sqrt{gH_0}$ is a given constant celerity, H_0 being a typical still water depth and $\alpha > 1$. The approximation is based on a modified system in which the divergence constraint on the velocity field is coupled with the other conservation laws following the ideas of the so-called hyperbolic divergence cleaning techniques (see [8–10, 19]). We suggest a formulation in which the divergence errors are transported with a finite speed c.

Remark 1 *Note that when $\alpha \to \infty$, system* (2.3a)-(2.3d) *and* (2.4) *formally converges to system* (2.3a)-(2.3e).

Remark 2 *Note that when $\alpha = 0$, and we consider an initial condition $w = p = 0$, then we recover the classical (SW) in spherical coordinates.*

3. Numerical scheme

In this work, we have adapted the ideas introduced in [5] to obtain an explicit high order well-balanced method for the system of balance laws (2.3a)-(2.3d) and (2.4). A finite volume method is considered based on a first-order path-conservative scheme and high-order reconstruction operator.

Structured grids on the $\theta - \varphi$ plane are considered. We use the HLL scheme written as a Polynomial Viscosity Method following [7] and a third order reconstruction operator described in [12], that has a compact stencil. We use the three step TVD RK method [14] that is also third order accurate in time. Therefore, the resulting scheme is third order accurate in space and time. The CFL condition reads as follows:

$$\Delta t = CFL \min \left\{ \frac{R\Delta_\theta \Delta_\varphi \cos(\varphi_i)}{\left(\frac{Q_{\theta,i}}{h} + \sqrt{gh_i + p_i + c^2}\right)\Delta_\varphi + \left(\frac{Q_{\varphi,i}}{h} + \sqrt{gh_i + p_i + c^2}\right)\Delta_\theta} \right\}, \tag{3.1}$$

$0 \leq CFL \leq 1$, where Δ_θ and Δ_φ are the mesh sizes in the θ and φ directions. To speed up the simulations, a parallel GPU implementation has been performed following the ideas described in [4, 12, 18], and in all the numerical test we set $\alpha = 3$ or $\alpha = 0$ to account for non-hydrostatic or hydrostatic simulations respectively.

4. Numerical results

In this section, we simulate the evolution of a tsunami in the south-western coast of Chile to check the performance of the numerical model and its ability to simulate planetary waves or tsunami waves over a realistic bathymetry.

We remark that although (θ, φ) are expressed in radians in the description of the numerical method, the description of the computational domains will be given in degrees, as it is usual in geophysics. Thus, the notation $(\bar{\theta}, \bar{\varphi})$ will be used to represent the longitude and latitude in degrees.

We consider a uniform Cartesian grid of the rectangular domain $[270, 294] \times [-30, -15]$ in the $\bar{\theta} - \bar{\varphi}$ plane (in degrees) with $\Delta_{\bar{\varphi}} = \Delta_{\bar{\theta}} = 1''$, that is 2880×1800 cells. The mean radius of the Earth is set to $R = 6371009.4\ m$, and the CFL parameter is set to 0.8. Open boundary conditions are prescribed at the four boundaries. The integration time was $[0, T]$, $T = 10000\ s$. The topo-bathymetry (see Fig. 3) of the area has been interpolated from the ETOPO1 Global Relief Model (see [1]). Next, a seafloor deformation generated by an earthquake has been computed using the Okada model. This seafloor deformation is instantaneously transmitted to the water column to generate the initial tsunami profile (see the perturbation on the free-surface in Fig. 3). The initial velocities, as well as the non-hydrostatic pressure, are set to zero. Concerning the numerical treatment of wet/dry fronts, here we follow the ideas described in [13], adapted to the reconstruction operator defined in [12]. Some temporal series provided

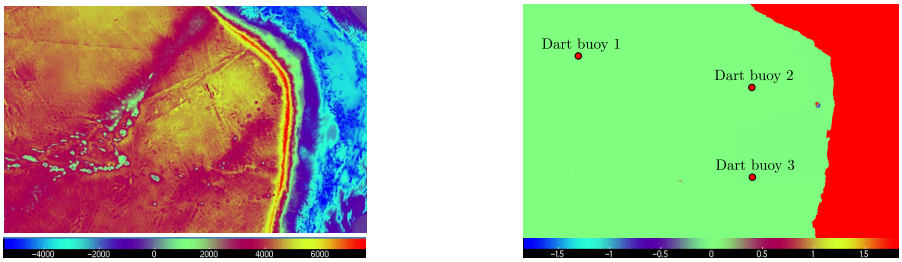


Fig. 3 Left, the topo-bathymetry of the south-western coast of Chile. Right, the free-surface initial condition computed with the Okada model and a sketch of the displacement of the Dart buoys.

by Dart Buoys located at $DB_1 = (273.659, -17.982)$, $DB_2 = (286.571, -20.473)$, $DB_3 = (286.017, -26.743)$ are given (see Fig. 3). We are interested in compare time series provided by the Dart buoys in the three locations against the computed numerical simulations from the new non-hydrostatic model.

Fig. 4 shows the numerical results for the free-surface elevation obtained with the hydrostatic (SW) ($\alpha = 0$) and with the non-hydrostatic model ($\alpha = 3$) at times $T = 1500\ s$ and $T = 4500\ s$. There it can be observed the standard dispersive pattern obtained with a non-hydrostatic dispersive model.

Figs. 5-7 show the comparison between the numerical computed temporal series and the field data provided by the Dart Buoys. The comparison exhibits the ability of the presented non-hydrostatic model to capture high-frequency dispersive waves in contrast with the hydrostatic system (SW). Moreover, the results agree with the Stokes linear theory: the leading wave given by the hydrostatic model tends to propagate faster than the one given by the non-hydrostatic model, and the amplitude of the front wave tends to be more accurate according to the field data. Therefore both amplitude and frequency of the waves are captured on all wave gauges successfully by

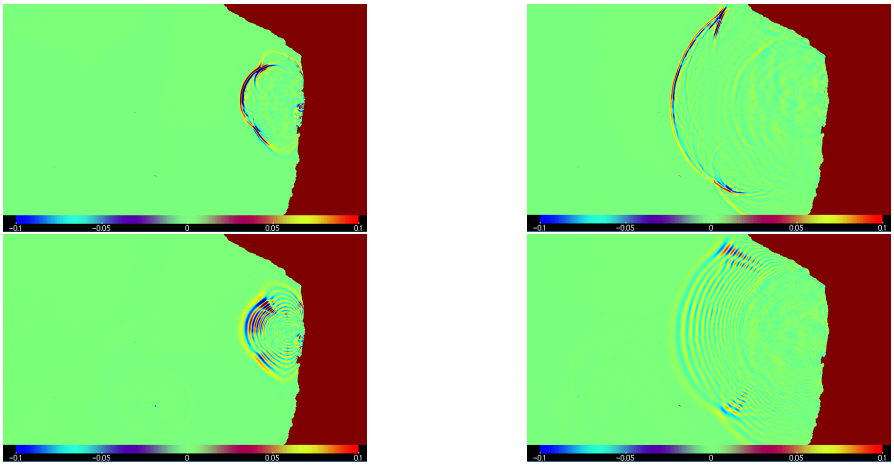


Fig. 4 Free surface elevation at times $T = 1500$ and $T = 4500$ (left and right resp.) The numerical results obtained with the hydrostatic model ($\alpha = 0$) are placed in the upper panels, and the ones obtained with the non-hydrostatic model ($\alpha = 3$) are placed at the lower panels.

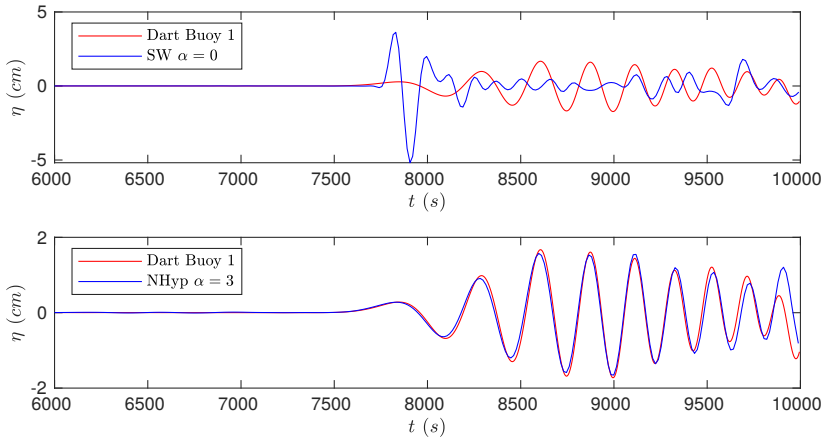


Fig. 5 Time series comparison against field data given by the Dart Buoy DB_1.

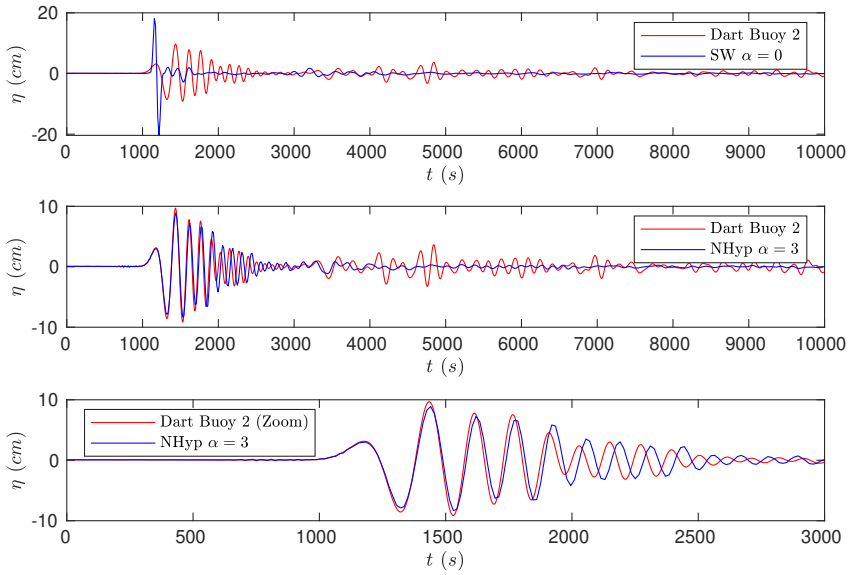


Fig. 6 Time series comparison against field data given by the Dart Buoy DB_2.

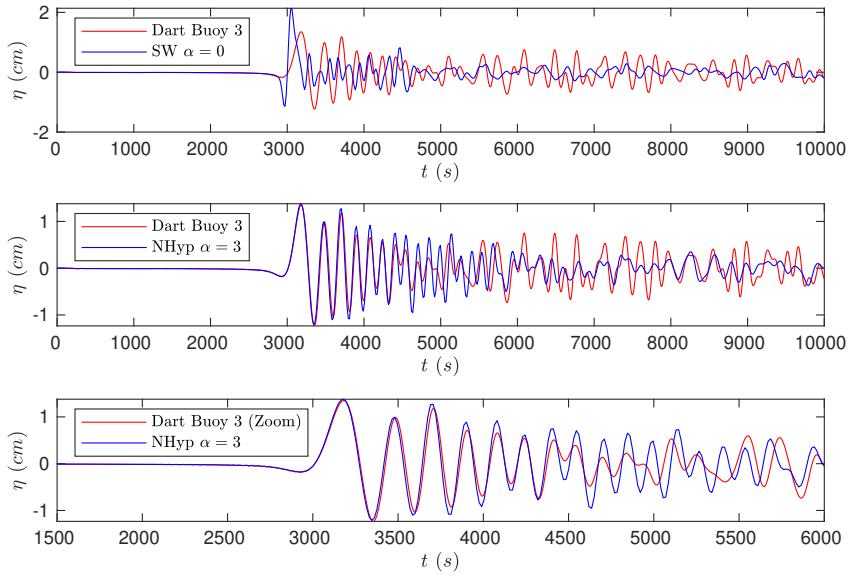


Fig. 7 Time series comparison against field data given by the Dart Buoy DB_3.

Simulated time: 10000 s. Third order scheme

Model	Comput. time	# times FTRT
Hydrostatic SW ($\alpha = 0$)	659.29	15.17
Non-hydrostatic ($\alpha = 3$)	1271.92	7.86

Tab. 1 Computational effort. Wall-clock times on a NVIDIA Tesla V100.

the non-hydrostatic model. The comparison with experimental data emphasizes the need to consider a dispersive model to capture the waves' shape faithfully.

Table 1 shows the execution times on a NVIDIA Tesla V100 GPU for $\alpha = 0$ (SW) and $\alpha = 3$. In view of the obtained results, we can conclude that the non-hydrostatic model can achieve a good computational performance with an additional computational cost that is only about 1.93 times the cost of a simple (SW) simulation.

5. Conclusions

A new first-order system of balance laws for shallow dispersive/non-hydrostatic free surface flows has been proposed to incorporate dispersive effects in the propagation of waves. The model is written in spherical coordinates to take into account the curvature effects of the Earth. The presented model corresponds to a relaxed approximation of the dispersive system derived by Bristeau *et al* in [3] written in spherical coordinates. The relaxation procedure follows ideas presented in [9, 10]. The big advantage of our new reformulation is that it can be easily discretized with explicit and high order accurate numerical schemes for systems of balance laws, without requiring the solution of an elliptic problem at each time step.

The numerical scheme employed here follows the ideas presented in [4,7,12]. As it can be seen in the numerical test, the numerical model can simulate dispersive water waves.

To allow simulations faster than real-time, an efficient GPU implementation of the numerical method has been carried out. The wall-clock times needed for non-hydrostatic simulations with the new model proposed in this paper are at most a factor of 1.93 higher than the wall clock times needed for a simple shallow water model, but which is not able to capture the correct dispersion characteristics of non-hydrostatic water waves.

The proposed numerical model presented in this work provides an efficient and accurate approach to model dispersive effects in the propagation of waves near coastal areas and intermediate waters.

Acknowledgements

This research has been supported by the Spanish Government (SG), the European Regional Development Fund (ERDF), the Regional Government of Andalusia (RGA), and the University of Málaga (UMA) through the projects of reference RTI2018-096064-B-C21 (SG-ERDF), UMA18-Federja-161 (RGA-ERDF-UMA), and P18-RT-3163 (RGA-ERDF).

References

[1] C Amante and B.W. Eakins. ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis. *NOAA Technical Memorandum NESDIS NGDC-24*, 2009.

[2] C. Bassi, L. Bonaventura, S. Busto, and M. Dumbser. A hyperbolic reformulation of the Serre-Green-Naghdi model for general bottom topographies. *Computers and Fluids*, 2020.

[3] M.-O. Bristeau, A Mangeney, J Sainte-Marie, and N Seguin. An energy–consistent depth–averaged Euler system: Derivation and properties. *Discrete and Continuous Dynamical Systems Series B*, 20(4):961–988, 2015.

[4] Manuel Castro, José M. Gallardo, and Carlos Parés. High order finite volume schemes based on reconstruction of states for solving hyperbolic systems with nonconservative products. Applications to shallow-water systems. *Mathematics of Computation*, 75(255):1103–1135, 2006.

[5] Manuel J. Castro, Sergio Ortega, and Carlos Parés. Reprint of: Well-balanced methods for the shallow water equations in spherical coordinates. *Computers and Fluids*, 169:129–140, 2018.

[6] V Casulli. A semi–implicit finite difference method for non–hydrostatic free surface flows. *Numerical Methods in Fluids*, 30(4):425–440, 1999.

[7] Marc de la Asunción, Manuel J. Castro, E. D. Fernández-Nieto, José M. Mantas, Sergio Ortega Acosta, and José Manuel 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 and Fluids*, 80(1):441–452, 2013.

[8] A. Dedner, F. Kemm, D. Kröner, C. D. Munz, T. Schnitzer, and M. Wesenberg. Hyperbolic divergence cleaning for the MHD equations. *Journal of Computational Physics*, 175(2):645–673, 2002.

[9] C. Escalante and T. Morales de Luna. A General Non-hydrostatic Hyperbolic Formulation for Boussinesq Dispersive Shallow Flows and Its Numerical Approximation. *Journal of Scientific Computing*, 2020.

[10] C. Escalante, M. Dumbser, and M. J. Castro. An efficient hyperbolic relaxation system for dispersive non-hydrostatic water waves and its solution with high order discontinuous Galerkin schemes. *Journal of Computational Physics*, 394:385–416, 2019.

[11] C. Escalante, T. Morales de Luna, and M. J. Castro. Non-hydrostatic pressure shallow flows: GPU implementation using finite volume and finite difference scheme. *Applied Mathematics and Computation*, 338:631–659, 2018.

[12] José M. Gallardo, Sergio Ortega, Marc De La Asunción, and José Miguel Mantas. Two-dimensional compact third-order polynomial reconstructions. Solving nonconservative hyperbolic systems using GPUs. *Journal of Scientific Computing*, 48(1-3):141–163, jul 2011.

[13] José M. Gallardo, Carlos Parés, and Manuel Castro. On a well-balanced high-order finite volume scheme for shallow water equations with topography and dry areas. *Journal of Computational Physics*, 227(1):574–601, 2007.

[14] Sigal Gottlieb and Chi-Wang Shu. Total variation diminishing Runge-Kutta schemes. *Mathematics of Computation of the American Mathematical Society*, 67(221):73–85, 1998.

[15] Jean Luc Guermond, Bojan Popov, Eric Tovar, and Chris Kees. Robust explicit relaxation technique for solving the Green-Naghdi equations. *Journal of Computational Physics*, 399:108917, 2019.

[16] M. Kazolea, A. I. Delis, and C. E. Synolakis. Numerical treatment of wave breaking on unstructured finite volume approximations for extended boussinesq-type equations. *Journal of Computational Physics*, 271:281–305, 2014.

[17] Gangfeng Ma, Fengyan Shi, and James T. Kirby. Shock-capturing non-hydrostatic model for fully dispersive surface wave processes. *Ocean Modelling*, 43-44:22–35, 2012.

[18] José Miguel Mantas, Marc De la Asunción, and Manuel J. Castro. An introduction to GPU computing for numerical simulation. In *SEMA SIMAI Springer Series*. 2016.

[19] C. D. Munz, P. Omnes, R. Schneider, E. Sonnendrücker, and U. Voß. Divergence Correction Techniques for Maxwell Solvers Based on a Hyperbolic Model. *Journal of Computational Physics*, 161(2):484–511, 2000.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 167–171)

New contributions to the control of PDEs and their applications

Enrique Fernández-Cara[1]
Universidad de Sevilla, Spain

Abstract

This paper deals with some recent achievements in control theory. Specifically, we will consider the null controllability problem for a quasi-linear parabolic PDE. We present some theoretical and numerical results. We also exhibit the results of a numerical experiment.

1. Introduction

Let $\Omega \subset \mathbb{R}^N$ be an open bounded regular domain ($N \leq 3$) and let $T > 0$ be given. We will mainly consider the system

$$\begin{cases} y_t - \nabla \cdot (a(y)\nabla y) = v\tilde{1}_\omega, \quad (x,t) \in Q := \Omega \times (0,T), \\ y = 0, \quad (x,t) \in \Sigma := \partial\Omega \times (0,T), \\ y(x,0) = y_0(x), \quad x \in \Omega. \end{cases} \tag{1.1}$$

Here, we assume that $\omega \subset\subset \Omega$ is a nonempty open set (the control domain), $\tilde{1}_\omega \in C_0^\infty(\Omega)$ satisfies $0 < \tilde{1}_\omega \leq 1$ in ω and $\tilde{1}_\omega = 0$ outside ω and $a \in C^3(\mathbb{R})$ possesses bounded derivatives of order ≤ 3 and satisfies

$$0 < m \leq a(r) \leq M \quad \forall r \in \mathbb{R}.$$

Obviously, we can interpret the control $v = v(x,t)$ as a heat source term and the state $y = y(x,t)$ as the associated temperature distribution in Q.

We will be concerned with the theoretical and numerical local null controllability of (1.1). Specifically, we will establish the existence of null controls when the initial state y_0 is small, we will present a related iterative algorithm and we will describe a numerical method for their computation.

The ideas and results that follow have been taken from [6] and [3]. They can be adapted to the solution of many other control problems; see for instance [2, 4] and the references therein. In particular, they serve to compute null controls for the Navier-Stokes and other similar equations, see [5].

2. The existence of null controls

The first main result in this contribution is the following:

Theorem 2.1 *Under the previous assumptions on the coefficient a, there exists $\varepsilon > 0$ such that, if $y_0 \in H_0^1(\Omega) \cap L^\infty(\Omega)$ and $\|y_0\|_{H^1} + \|y_0\|_{L^\infty} \leq \varepsilon$, there exists a control $v \in L^2(\omega \times (0,T))$ and an associated solution to the nonlinear system* (1.1) *satisfying*

$$y(x,T) = 0 \ \text{ in } \ \Omega. \tag{2.1}$$

In order to prove Theorem 2.1, we can employ a technique relying on the so called *Liusternik's Inverse Function Theorem,* see [1].

Thus, in view of the regularizing effect, we can assume that $y_0 \in H^3(\Omega) \cap H_0^1(\Omega)$ and is small in this space. Then, we consider the linearized system at zero

$$\begin{cases} y_t - a(0)\Delta y = v\tilde{1}_\omega + h(x,t), \quad (x,t) \in Q, \\ y = 0, \quad (x,t) \in \Sigma, \\ y(x,0) = y_0(x), \quad x \in \Omega. \end{cases} \tag{2.2}$$

It is well known that, under some appropriate assumptions on h, (2.2) is null-controllable. More precisely, the adjoint of (2.2) is given by

$$\begin{cases} -\varphi_t - a(0)\Delta\varphi = F(x,t), \quad (x,t) \in Q, \\ \varphi = 0, \quad (x,t) \in \Sigma, \\ \varphi(x,T) = \varphi_T(x), \quad x \in \Omega, \end{cases} \tag{2.3}$$

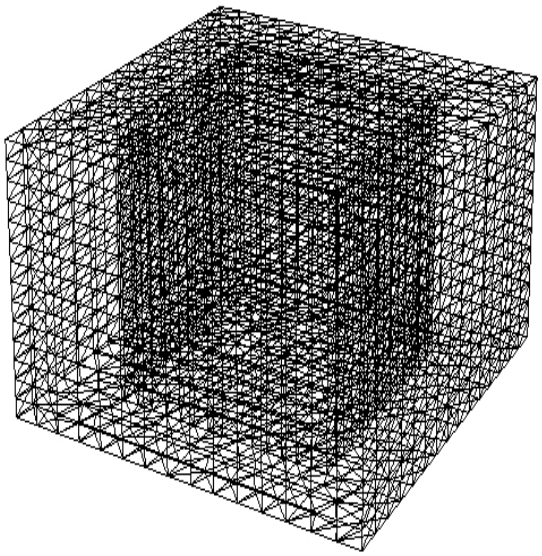


Fig. 1 The mesh. Number of vertices: 7425. Number of tetrahedrons: 38976.

where $\varphi_T \in L^2(\Omega)$; the announced null controllability property is implied by a well known Carleman inequality that can be established for any solution to a system of the form (2.3).

In a second step, we rewrite the null controllability problem for (1.1) as an equation in a well chosen space of "admissible" state-control pairs:

$$\mathcal{H}(y, v) = (0, y_0), \quad (y, v) \in Y. \tag{2.4}$$

Here, Y is a space of couples (y, v) satisfying, among other things, the following properties

$$\iint_Q \rho^2 |y|^2 + \iint_{\omega\times(0,T)} \rho_0^2 |v|^2 < +\infty$$

and

$$\iint_Q \hat{\rho}^2 |y_t - a(0)\Delta y - v\tilde{1}_\omega|^2 < +\infty,$$

where ρ, ρ_0 and $\hat{\rho}$ are appropriate weight functions that blow up to $+\infty$ as $t \to T$. Formally, the definition of $\mathcal{H}$ is the following:

$$\mathcal{H}(y, v) := (y_t - \nabla \cdot (a(y)\nabla y) - v\tilde{1}_\omega, y(\cdot, 0)) \quad \forall (y, v) \in Y.$$

Then, we apply Liusternik's Theorem and we deduce the (local) desired result. To this purpose, we previously have to establish some nontrivial estimates for the null controls and the associated states of (2.2).

3. A convergent algorithm

The computation of a null control of (1.1) is not a simple task; here, we will argue as in [2, 6], taking advantage of the surjectivity of $\mathcal{H}'(0, 0)$.

Thus, let Y be the Hilbert space where we can find a solution (y, v) to (2.4). We introduce the following iterative algorithm:

ALG 1:

1. Choose $(y^0, v^0) \in Y$.
2. Then, for given $n \geq 0$ and $(y^n, v^n) \in Y$, compute

$$(y^{n+1}, v^{n+1}) = (y^n, v^n) - \mathcal{H}'(0, 0)^{-1}(\mathcal{H}(y^n, v^n) - (0, y_0)). \tag{3.1}$$

In these iterates, we use $\mathcal{H}'(0, 0)^{-1}$, which is by definition an inverse to the left of $\mathcal{H}'(0, 0)$.

Note that **ALG 1** is an elementary quasi-Newton method and consequently has the following interesting property: the finite dimensional approximations of the iterates lead to a set of algebraic systems whose coefficient matrices are always the same.

In our second main result, we prove the convergence of **ALG 1** and we furnish some estimates:

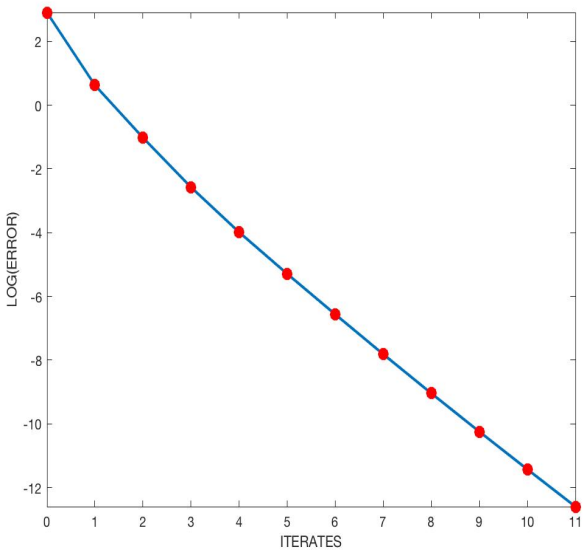


Fig. 2 Evolution of the error at logarithmic scale.

Theorem 3.1 *Let $y_0 \in H_0^1(\Omega) \cap L^\infty(\Omega)$ be given with $\|y_0\|_{H_0^1} + \|y_0\|_{L^\infty} \leq \varepsilon$ (ε is furnished by Theorem 2.1). There exists $\kappa \in (0,1)$ such that, if $(y^0, v^0) \in Y$ and*

$$\|(y^0, v^0) - (y, v)\|_Y \leq \kappa,$$

then the (y^n, v^n) converge to (y, v) and satisfy

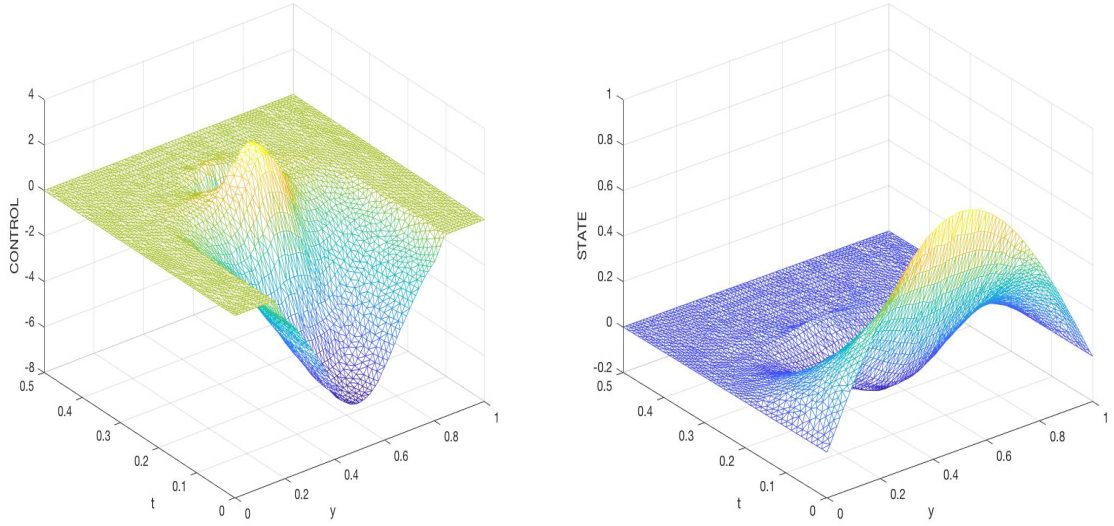
$$\|(y^{n+1}, v^{n+1}) - (y, v)\|_Y \leq \theta \|(y^n, v^n) - (y, v)\|_Y \tag{3.2}$$

for all $n \geq 0$ for some $\theta \in (0,1)$.

Remark 3.2 A natural question is whether Theorems 2.1 and 3.1 also hold for similar systems with PDEs of the form

$$y_t - \nabla \cdot (a(x,t;y)\nabla y) = v\tilde{1}_\omega \quad \text{and/or} \quad y_t - \nabla \cdot (a(x,t;\nabla y)\nabla y) = v\tilde{1}_\omega,$$

that is, with nonlinear diffusion coefficients nonhomogeneous in space and time and eventually depending on the gradient. In both cases, the answer is yes, provided they are regular enough, see [4]. □



(a) Computed control

(b) Associated state

Fig. 3 The computed control and the associated state at $x_1 = 0.68$.

4. A numerical method for the solution of (3.1)

The computation of (y^{n+1}, v^{n+1}) in (3.1) can be achieved following the Fursikov-Imanuvilov method [7].

The strategy is to take

$$y = \rho^{-2} L^* p, \quad v = -\rho_0^2 \, \tilde{1}_\omega p\big|_{\omega\times(0,T)}, \tag{4.1}$$

where $L^* p := -p_t - a(0)\Delta p$ and p is the unique solution to the Lax-Milgram problem

$$\begin{cases} \displaystyle\iint_Q \left(\rho^{-2} L^* p L^* p' + \tilde{1}_\omega p p'\right) = \iint_Q h p' + \int_\Omega y_0(x) p'(x,0)\, dx \\ \forall p' \in P, \quad p \in P \end{cases} \tag{4.2}$$

(P is an appropriate Hilbert space of functions p with $L^* p \in L^2_{loc}(Q)$). Accordingly, we set $\mathcal{H}'(0,0)^{-1}(h, y_0) = (y, v)$, with y and v respectively given by (4.1) and (4.2).

Note that (4.2) is the weak formulation of a boundary-value problem for p that is second-order in time and fourth-order in space.

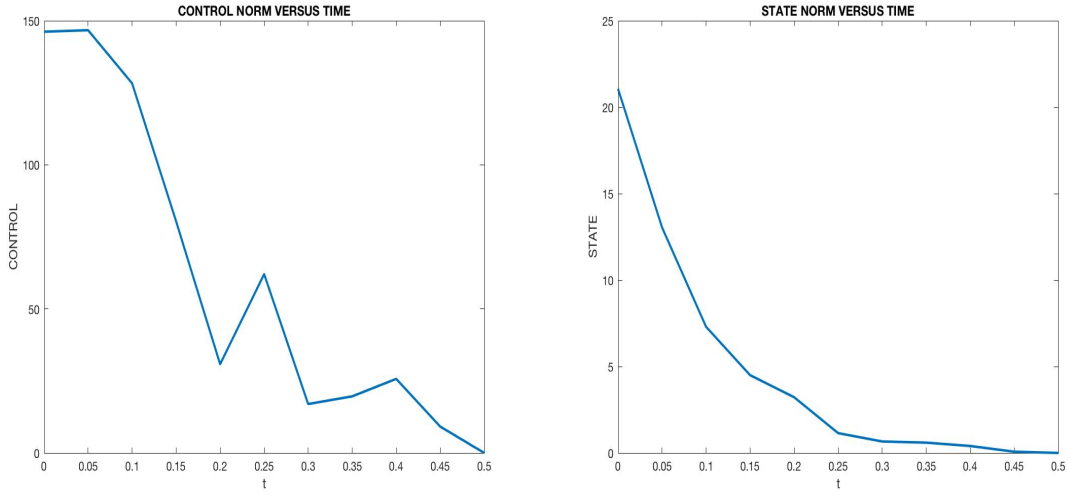
Unfortunately, it is not easy to construct and handle finite dimensional spaces $P_h \subset P$ (except in the particular case $N = 1$). Thus, it is convenient to introduce a mixed formulation (as in [5, 6]) and then reduce to finite dimension. This leads to numerical approximations with P_ℓ-piecewise continuous functions that furnish good results; see the details in [3].

5. A numerical experiment

The quasi-Newton method has been applied to the solution to the null controllability problem for (1.1) with the following data:

- $N = 2$, $\Omega = (0,1)\times(0,1)$, $\omega = (0.2, 0.8)\times(0.2, 0.8)$, $T = 0.5$.
- $y_0(x_1, x_2) = \sin(\pi x_1)\sin(\pi x_2)$.
- $a(s) = \exp(-2\exp(-0.3s))$.

The computations have been performed with the FreeFem++ package; see `http://www.freefem.org//ff++`. The stopping criterion for **ALG 1** has been $\|y^{n+1} - y^n\|_{L^2}/\|y^{n+1}\|_{L^2} \le \varepsilon_0$, where y^n is the computed state and $\varepsilon_0 = 10^{-5}$. The mesh, the error evolution, the computed control and state and their spatial L^2 norms are displayed in Fig. 1–4.



(a) Computed control **(b)** Associated state

Fig. 4 Evolution in time of the L^2 norms of the control and the state.

Acknowledgements

The author was partially supported by Grant MTM2016-76990-P of MICINN (Spain).

References

[1] Alekseev V.M., Tikhomorov V.M., Formin S.V., *Optimal Control,* Consultants Bureau, New York, 1987.

[2] Clark H.R., Fernández-Cara E., Límaco J., Medeiros L.A., *Theoretical and numerical local null controllability for a parabolic system with local and nonlocal nonlinearities,* Applied Mathematics and Computation, 223, 483–505, 2013.

[3] Fernández-Cara E., Límaco J., Marín-Gayte I., *Theoretical and numerical local null controllability of a quasi-linear parabolic equation in dimensions 2 and 3,* to appear in J. Franklin Institute.

[4] Fernández-Cara E., Límaco J., Menezes D., Thamsten Y., *Theoretical and numerical controllability results concerning a nonlinear diffusion model,* submitted.

[5] Fernández-Cara E., Münch A., Souza D.A., *On the numerical Controllability of the two-dimensional heat, Stokes and Navier-Stokes equations,* J. Sci Comput., 70, 78–85, 2017.

[6] Fernández-Cara E., Nina-Huamán D., Núñez-Chávez M.R., Vieira F.B., *On the theoretical and numerical control of a 1D nonlinear parabolic PDE,* J. Optim. Theory Appl., 175(3), 652–682, 2017.

[7] Fursikov A.V., Imanuvilov O.Yu., *Controllability of Evolution Equations,* Lecture Notes Series, Seoul National University, Research Institute of Mathematics, Global Analysis Research Center, Seoul, 34,1996.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 172–176)

Saddle-node bifurcation of canard limit cycles in piecewise linear systems

Soledad Fernández-García[1], Victoriano Carmona[1], Antonio E. Teruel[2]
1. Universidad de Sevilla, Spain
2. Universidad de las Islas Baleares, Spain

Abstract

We study saddle-node bifurcations of canard limit cycles in PWL systems by using singular perturbation theory tools. We distinguish two cases: the subcritical and the supercritical. In the subcritical case, we find saddle-node bifurcations of canard cycles both with head and without head. Moreover, we detect a transition between them. In the supercritical case, we find situations with two saddle-node bifurcations, which take place exponentially close in the parameter space; one of headless canards and another of canards with head. There, three canard cycles can coexist.

1. Introduction

The classical canard explosion is a phenomenon that occurs in limit cycles of planar slow-fast systems. It was discovered and analyzed by Benoit et al. in 1981 [2] in the Van der Pol oscillator and consists on the fast transition, by changing one parameter of the system, from a small amplitude Hopf-like limit cycle to a relaxation oscillation cycle.

The analysis of the slow-fast dynamics is done by using tools from Geometric Singular Perturbation Theory. The main idea consist on reconstructing the global behavior by splitting and then joining, in a suitable way, the fast and slow dynamics. Under hyperbolicity conditions, Fenichel Theorem describes the existence of invariant slow manifolds close to compact parts of the fast nullcline and also describes the stability properties of these slow manifolds [5]. However, when the fast nullcline folds, which is the case in the canard phenomenon, normal hyperbolicity is lost, and Fenichel Theorem cannot be applied. Different sophisticated techniques have been developped in order to analyze this behavior around the fold, such as, for instance, the blow up technique [7].

On the other hand, some authors have analyzed the possibility of reproducing the canard phenomenon in systems more amenable to study, such as, piecewise linear (PWL) systems. Even when some dynamical aspects of the slow-fast behavior had been observed in PWL systems, it has taken some time to understand the way of reproducing the slow-fast dynamics properly, see [4] and references therein.

In [6], the authors reproduced part of the canard explosion phenomenon in the PWL context, in particular the one involving hyperbolic headless canards. Here we present the main results obtained in [3], where we consider an extension of the system analyzed in [6], which allows for the existence of both canards with and without head and both, hyperbolic and non-hyperbolic canard cycles. In particular, the system is able to reproduce saddle-node bifurcations of canard limit cycles.

The obtained results in [3] are comparable with those obtained for smooth vector fields, by Krupa and Szmolyan in [7]. Furthermore, we have found new scenarios that, as far as we are concerned, had not been previously reported in the smooth framework. Surprisingly, we find situations where two saddle-node bifurcations of canard cycles take place, one of headless canards and another one of canards with head. In such a case, we show the coexistence of three canard limit cycles.

The outline of this work is given as follows. First, in Section 2, we review the canard explosion and saddle-node canard cycles in the smooth case. Second, in Section 3, we introduce the PWL systems which we focus on and we present the Main Results. Finally, Section 4 is devoted to present some conclusions.

2. Background on canard cycles: canard explosion.

Canard solutions take place in planar differential slow-fast systems, that is, systems of the form [2, 7],

$$\begin{cases} \varepsilon \dot{x} = f(x, y, a, \varepsilon), \\ \dot{y} = g(x, y, a, \varepsilon), \end{cases} \tag{2.1}$$

where $f, g \in C^r, r \geq 3, a \in \mathbb{R}, 0 < \varepsilon \ll 1$ and the dot denotes the derivative with respect to the temporal variable τ. After the rescaling in time $t = \tau/\varepsilon$, system (2.1) writes as

$$\begin{cases} x' = f(x, y, a, \varepsilon), \\ y' = \varepsilon g(x, y, a, \varepsilon), \end{cases} \tag{2.2}$$

where the prime denotes the derivative with respect to the fast time t. Systems (2.1) and (2.2) are equivalent through the identity when $\varepsilon > 0$, but they have not the same limit for $\varepsilon = 0$. In fact, the limit of system (2.1), called *slow subsystem*, is a semi-explicit Differential Algebraic Equation (DAE), where the relation between the variables is given by

$$S = \{(x, y) : f(x, y, a, 0) = 0\}.$$

Assuming that $f_y(x, y, a, 0) \neq 0$ it follows that S is the graph of a differentiable function $y = \varphi_a(x)$, and the DAE reduces to the differential equation

$$f_x(x, \varphi_a(x), a, 0)\dot{x} = -f_y(x, \varphi_a(x), a, 0)g(x, \varphi_a(x), a, 0), \tag{2.3}$$

which is called the *reduced equation*. On the other hand, the limit for $\varepsilon = 0$ of system (2.2), called *fast subsystem*, is a differential equation having S as the locus of every equilibrium point. From here S is called the *critical manifold*.

Canard cycles develop along a branch born at a Hopf bifurcation, at $a = a_H$, and the canard explosion takes place at a value which is at a distance of $O(\varepsilon)$ from the a_H. This means that very close to the bifurcation point a_H, before the explosion, the cycles have the characteristics of typical Hopf cycles. This Hopf bifurcation arises only for $\varepsilon > 0$ and is usually known as a singular Hopf bifurcation [1].

The existence of saddle-node bifurcation of canard cycles in the smooth framework has been analyzed in [7]. There, the authors consider two different cases, depending whether the Hopf bifurcation where the cycle is born is supercritical or subcritical. Thus, after proving the existence of the maximal canard, they distinguish two different scenarios:

- *Supercritical case*: In Theorem 3.3, authors state the existence of a family of periodic orbits. These periodic orbits can be stable Hopf-type limit cycles, canard limit cycles or relaxation oscillations. To analyze the stability of the canard limit cycles, they use the *way in-way out function* $R(s)$, which is the limit of the integral of the divergence along the slow manifolds when $\varepsilon \to 0$. In Theorem 3.4, assuming that this function is negative, the authors state that the canard limit cycles of the family are stable.

- *Subcritical case*: In Theorem 3.5, authors state the existence of other family of periodic orbits. The orbits of that family can be unstable Hopf-type limit cycles, canard limit cycles or relaxation oscillations. Again, to analyze the stability of canard cycles, they use the *way in-way out function* $R(s)$. In Theorem 3.6, assuming that this function has exactly one simple zero at $s = s_{lp,0}$, the authors state that there exists a function $s_{lp}(\sqrt{\varepsilon})$ having limiting point at $s_{lp,0}$ when $\varepsilon \to 0$, such that canard limit cycles are unstable for $s < s_{lp}(\sqrt{\varepsilon})$ and stable for $s > s_{lp}(\sqrt{\varepsilon})$.

3. Statement of the piecewise linear system and Main Results.

In this section, first we introduce the family of PWL differential systems considered in [3] and after that we stay the main results, whose proofs can be consulted there.

The class of planar differential systems considered in [3] reads,

$$\begin{cases} x' = y - f(x, a, k, m, \varepsilon), \\ y' = \varepsilon(a - x), \end{cases} \tag{3.1}$$

where the prime denotes the derivative with respect to the time t, $(x, y)^T \in \mathbb{R}^2$, $0 < \varepsilon \ll 1$, and the x-nullcline is defined by the graph of the continuous PWL function with four segments given by

$$f(x, a, k, m, \varepsilon) = \begin{cases} x + 1 - k(\sqrt{\varepsilon} - 1) - m(\sqrt{\varepsilon} + a), & \text{if } x < -1 \\ -k(x + \sqrt{\varepsilon}) - m(\sqrt{\varepsilon} + a), & \text{if } -1 < x \leq -\sqrt{\varepsilon}, \\ m(x - a), & \text{if } |x| \leq \sqrt{\varepsilon}, \\ x - \sqrt{\varepsilon} + m(\sqrt{\varepsilon} - a), & \text{if } x > \sqrt{\varepsilon}, \end{cases} \tag{3.2}$$

with $k > 0$, $a \in \mathbb{R}$ and $|m| < 2\sqrt{\varepsilon}$.

Note that, the phase space is splitted into four regions: the lateral half-planes $LL = \{(x, y) : x \leq -1\}$ and $R = \{(x, y) : x \geq \sqrt{\varepsilon}\}$, and the central bands $L = \{(x, y) : -1 \leq x \leq -\sqrt{\varepsilon}\}$ and $C = \{(x, y) : |x| \leq \sqrt{\varepsilon}\}$. Restricted to any of these regions, the vector field is linear.

We proceed now to present the main results in [3]. These results concern to the existence of a one parameter family of canard limit cycles in the PWL system (3.1)-(3.2), and to the description about how this family organizes along a curve in the plane (x, a), where x is the width of the canard limit cycle and a is the parameter value. The results also provide information about the stability of the limit cycles.

In the first result in [3], it is assured that the starting point of the curve organizing the family of limit cycles exhibited by system (3.1)-(3.2) takes place at a Hopf-like bifurcation. At this bifurcation a limit cycle appears after

the change of stability of the singular point, just like in the Hopf bifurcation. The difference between both kind of bifurcations is the relation between the amplitude of the limit cycle and the bifurcation value, this relation is linear in the Hopf-like bifurcation and a square root in the Hopf bifurcation.

Next theorem, which we include subsequently, is devoted to the existence of the maximal canard trajectory, that is, a trajectory connecting the attracting and the repelling branches of the slow manifold.

Theorem 3.1 *Set $m = \pm\sqrt{\varepsilon}$. There exist a value $\varepsilon_0 > 0$ and a function $a = \tilde{a}(k, \varepsilon; m)$, analytic as a function of $(k, \sqrt{\varepsilon})$, defined in the open set $U = (0, +\infty) \times (0, \varepsilon_0)$ and such that, for $(k, \varepsilon) \in U$, a solution of system* (3.1)-(3.2) *starting in the attracting branch of the slow manifold, μ_R, connects to the repelling branch of the slow manifold, μ_L, if and only if $a = \tilde{a}(k, \varepsilon; m)$. In such case, the time of flight of the transition is $\tau_C(k, \varepsilon; m) > 0$. First terms of the expansions of $\tilde{a}(k, \varepsilon; m)$ and $\tau_C(k, \varepsilon; m)$ are given as follows,*

$$\tilde{a}(k,\varepsilon;m) = \begin{cases} \dfrac{e^{\frac{\pi}{\sqrt{3}}}-1}{e^{\frac{\pi}{\sqrt{3}}}+1}\sqrt{\varepsilon} - \dfrac{e^{\frac{\pi}{\sqrt{3}}}}{\left(e^{\frac{\pi}{\sqrt{3}}}+1\right)^2}\left(\dfrac{1-k^2}{k^2}\right)\varepsilon^{3/2} + O(\varepsilon^2), & \text{if } m = -\sqrt{\varepsilon}, \\ -\dfrac{e^{\frac{\pi}{\sqrt{3}}}-1}{e^{\frac{\pi}{\sqrt{3}}}+1}\sqrt{\varepsilon} - \dfrac{e^{\frac{\pi}{\sqrt{3}}}}{\left(e^{\frac{\pi}{\sqrt{3}}}+1\right)^2}\left(\dfrac{1-k^2}{k^2}\right)\varepsilon^{3/2} + O(\varepsilon^2), & \text{if } m = \sqrt{\varepsilon}, \end{cases} \tag{3.3}$$

and

$$\tau_C(k,\varepsilon;m) = \begin{cases} \dfrac{2\pi}{\sqrt{3}}\dfrac{1}{\sqrt{\varepsilon}} - \dfrac{1+k}{k} - \dfrac{1-k^2}{2k^2}\sqrt{\varepsilon} + O(\varepsilon), & \text{if } m = -\sqrt{\varepsilon}, \\ \dfrac{2\pi}{\sqrt{3}}\dfrac{1}{\sqrt{\varepsilon}} - \dfrac{1+k}{k} + \dfrac{1-k^2}{2k^2}\sqrt{\varepsilon} + O(\varepsilon), & \text{if } m = \sqrt{\varepsilon}. \end{cases} \tag{3.4}$$

The existence of the maximal canard trajectory, together with the divergence of the flow in a neighborhood of the slow manifold, provide the arguments used in [3] to prove the following result about the existence of canard cycles of any suitable width. To state the result in a proper way we introduce the following values

$$x_r = -(1+k) + k\sqrt{\varepsilon} - \lambda_L^s(\sqrt{\varepsilon} + a), \quad x_s = -\sqrt{\varepsilon} - \lambda_L^s(\sqrt{\varepsilon} + a). \tag{3.5}$$

These values correspond with the end points of the interval such that limit cycles having width contained in (x_r, x_s) are canard limit cycles. In fact, limit cycles having width $x < x_r$ are relaxation oscillations whereas limit cycles having width $x > x_s$ are still under the effect of the Hopf-like bifurcation.

Theorem 3.2 *Fix ε_0 sufficiently small and set $m = \pm\sqrt{\varepsilon}$. There exists a function $a = \hat{a}(k, \varepsilon, x_0; m)$, C^∞ function of $(k, \sqrt{\varepsilon}, x_0)$, defined in the open set $U = (0, +\infty) \times (0, \varepsilon_0) \times (x_r, x_s)$, fulfilling*

$$\begin{aligned} |\hat{a}(k,\varepsilon,x_0;m) - \tilde{a}(k,\varepsilon;m)| &\approx |x_0| e^{-\frac{x_0}{\varepsilon^{3/2}}} && x_0 \in [-1, x_s), \\ |\hat{a}(k,\varepsilon,x_0;m) - \tilde{a}(k,\varepsilon;m)| &\approx |x_0 - x_r| e^{-\frac{x_0 - x_r}{\varepsilon}} && x_0 \in (x_r, -1), \end{aligned}$$

with $\tilde{a}(k, \varepsilon; m)$ the function defined in Theorem 3.1*, and such that, for $(k, \varepsilon, x_0) \in U$ and $a = \hat{a}(k, \varepsilon, x_0; m)$ system* (3.1)-(3.2) *possesses a canard limit cycle, Γ_{x_0}, passing through $(x_0, f(x_0))$. The canard limit cycle is headless if $x_0 \in (-1, x_s)$ and with head if $x_0 \in (x_r, -1)$.*

Previous result describes the canard explosion taking place in the PWL framework. There, it can be observed that the slope of the explosion is different before and after the maximal canard.

In the following result, the stability of the canard limit cycles obtained in the previous theorem is established. The results are divided into two theorems, depending on whether the Hopf-like bifurcation is supercritical or subcritical.

Theorem 3.3 *Set $\varepsilon > 0$ small enough, $m = -\sqrt{\varepsilon}$, $x_0 \in (x_r, x_u) \cup [-1, x_s)$ and $a = \hat{a}(k, \varepsilon, x_0; m)$. Let Γ_{x_0} be the canard limit cycle of system* (3.1)-(3.2) *whose existence has been proved in Theorem* 3.2*. The following statements hold*:

a) *For $k \leq 1$, the canard limit cycle Γ_{x_0} is hyperbolic and stable.*

b) *For $k > 1$, there exist exactly two values $x_1 \in (-1, x_s)$ and $x_2 \in (x_r, x_u)$ such that the canard limit cycle Γ_{x_0} is hyperbolic and stable if $x_0 \in (x_r, x_2) \cup (x_1, x_s)$, hyperbolic and unstable if $x_0 \in (x_2, x_u) \cup (-1, x_1)$, and a saddle-node canard cycle if $x_0 = x_1$ and $x_0 = x_2$.*

Theorem 3.4 *Set $\varepsilon > 0$ small enough, $m = \sqrt{\varepsilon}$, $x_0 \in (x_r, x_u) \cup [-1, x_s)$ and $a = \hat{a}(k, \varepsilon, x_0; m)$. Let Γ_{x_0} be the canard limit cycle of system* (3.1)-(3.2) *whose existence has been proved in Theorem* 3.2*. The following statements hold*:

a) *For $k < 1$, there exists exactly one value $x_1 \in (-1, x_s)$ such that Γ_{x_0} is an hyperbolic limit cycle, if $x_0 \in (x_r, x_u) \cup (-1, x_s) \setminus \{x_1\}$, and a saddle-node canard cycle, if $x_0 = x_1$. Moreover, Γ_{x_0} is stable if $x_0 < x_1$ and unstable if $x_0 > x_1$.*

b) *For $k = 1$, the canard limit cycle Γ_{x_0} is hyperbolic and stable if $x_0 \in (x_r, x_u)$ and hyperbolic and unstable if $x_0 \in (-1, x_s)$.*

c) *For $k > 1$, there exists exactly one value $x_2 \in (x_r, x_u)$ such that Γ_{x_0} is hyperbolic, if $x_0 \in (x_r, x_u) \cup (-1, x_s) \setminus \{x_2\}$, and a saddle-node canard cycle, if $x_0 = x_2$. Moreover, Γ_{x_0} is stable if $x_0 < x_2$ and unstable if $x_0 > x_2$.*

Subsequently, in the last main result, it is stated that for every width between the smallest canard cycle and the relaxation oscillation cycle, that is for every $x_0 \in (x_r, x_u) \cup [-1, x_s)$, there exist values of the parameters such that system (3.1)-(3.2) exhibits a saddle-node canard limit cycle Γ_{x_0} of width x_0.

Theorem 3.5 *Consider system* (3.1)-(3.2) *with $m = -\sqrt{\varepsilon}$ or $m = \sqrt{\varepsilon}$. For each $x_0 \in (x_r, x_u) \cup (-1, x_s)$, there exists a value ε_0 and a function $k_{x_0}(\varepsilon)$ defined for $\varepsilon \in (0, \varepsilon_0)$, such that system* (3.1)-(3.2) *with parameters $k = k_{x_0}(\varepsilon)$ and $a = \hat{a}(k_{x_0}(\varepsilon), \varepsilon, x_0; m)$ exhibits the saddle-node canard Γ_{x_0} whose existence has been stated in Theorem* 3.3 *for $m = -\sqrt{\varepsilon}$ and in Theorem* 3.4 *for $m = \sqrt{\varepsilon}$, respectively.*

4. Conclusions.

In [3], we have analyzed the existence of saddle-node bifurcation of canard cycles in PWL systems. We have revised in the PWL context the known results in the smooth framework [7]. Let us point out the similarities and differences that we have found:

Canard cycles in [7] develop along a branch born at a Hopf bifurcation, at $a = a_H$, and the canard explosion takes place at a value which is at a distance of $O(\varepsilon)$ from the a_H. In the PWL context, we have checked that the canard explosion takes place at a value which is at a distance of $O(\sqrt{\varepsilon})$ from the a_H.

In the *Supercritical case, $m = -\sqrt{\varepsilon}$*: System (3.1)-(3.2) is able to reproduce the dynamics in the smooth case with $k \leq 1$, that is, the existence of a family of stable canard cycles. By letting k increase, we have found new scenarios that have not been reported in the smooth framework. Specifically, when $k > 1$, we find situations where two saddle-node bifurcations of canard cycles take place, one of headless canards and another one of canards with head. In this case, three canard limit cycles can coexist.

In the *Subcritical case, $m = \sqrt{\varepsilon}$*: In this case, system (3.1)-(3.2) can reproduce the dynamics in the smooth case, with the benefit that in the PWL case we can control the different behaviors that appear in an easier way. Concretely, we have proved the existence of saddle-node bifurcation of headless canards for $k < 1$, and of canards with head for $k > 1$.

It has been stated in Theorem 3.5 that in both subcritical and supercritical cases, for every height between the smallest canard cycle and the relaxation oscillation cycle there exist parameters k and ε such that a saddle-node canard limit cycle with this height exists.

The use of this simpler family of slow-fast systems to reproduce canard dynamics bring us some information which could be interesting when revisiting the smooth context. In particular, conditions $k < 1$ and $k > 1$ organizing the dynamics in the main results, suggest the importance of the ratio between the slopes of the fast nullcline in order to exhibit or not saddle-node canard cycles with head. Bearing this in mind, we believe that only saddle-node canard cycles with head can appear when the slope of the repelling branch of the critical manifold is larger than the slope of the attracting branches of the critical manifold. As this is not the case in the Van der Pol system, we can expect only headless saddle-node canard cycles there.

Last, we would like to point out that some quantitative information obtained in [3] could be relevant for applications. For instance, we highlight the period of the canard cycles and the location of the saddle-node canards in terms of the parameter. Finally, the dependence between the height of a canard cycle and the bifurcation parameter a at which it appears could be approximated from the estimation $|\tilde{a} - \hat{a}|$ appearing in Theorem 3.2.

Acknowledgements

First author is supported by Ministerio de Ciancia, Innovación y Universidades, through the project PGC2018-096265-B-I00. Second author is supported by Ministerio de Ciencia, Innovación y Universidades through the project RTI2018-093521-B-C31. Third author is supported by Ministerio de Economía y Competitividad through the project MTM2017-83568-P (AEI/ERDF, EU).

References

[1] B. Braaksma, Singular Hopf bifurcation in systems with fast and slow variables, em J. Nonlinear Sci. **8**(5), 457–490, 1998.

[2] E. Benoit, J.-L. Callot, F. Diener, et al, Chasse au Canard. *Collect Math.* **32**(1-2), 37–119, 1981.

[3] Victoriano Carmona, Soledad Fernández-García and A. E. Teruel. Saddle-node canard cycles in planar PWL differential systems. *arXiv*:2003.14112v2, 2020.

[4] M. Desroches, S. Fernández-García, M. Krupa, R. Prohens and A. E. Teruel, Piecewise-linear (PWL) canard dynamics: Simplifying singular perturbation theory in the canard regime using piecewise-linear systems, *Nonlinear Systems*, Vol. 1 : Mathematical Theory and Computational Methods, Springer, 2018.

[5] N. Fenichel, Geometric singular perturbation theory for ordinary differential equations, *J. Differ. Equations.* **31**(1), 53–98, 1979.

[6] Soledad Fernández-García, M. Desroches, M. Krupa and A. E. Teruel. Canard solutions in planar piecewise linear systems with three zones. *Dynam. Syst.*, **31**(2), 173–197, 2016.

[7] M. Krupa and P. Szmolyan, Relaxation Oscillation and Canard Explosion, *Journal of Differential Equations*, **174**, 312-.368, 2001.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 177–183)

On the Amplitudes of Spherical Harmonics of Gravitational Potencial and Generalised Products of Inertia

Luis Floría[1]
Universidad de Zaragoza, Spain

Abstract

The vector field of the force of gravitational attraction due to an extended rigid body (of arbitrary irregular geometrical shape, and with an arbitrary internal mass distribution inside it) at any point outside the body can be derived from the gradient of a scalar field, its gravitational potential. In terms of spherical polar coordinates (distance from the origin, colatitude or latitude, and longitude) that potential can be expanded as an absolutely convergent series of spherical harmonics, involving Legendre polynomials and associated Legendre functions of the first kind depending on the colatitude (or the latitude) and circular functions depending on the longitude.

In the present contributed paper we establish, in terms of the so–called "integrals of inertia" (or "generalised products of inertia") of the body, general formulae for the amplitudes (i.e., for the coefficients) of the different zonal, tesseral, and sectorial harmonics of any degree and order in the said series expansion of the gravitational potential outside the body.

Key words and expressions: Celestial Mechanics, Potential Theory, extended rigid body, gravitational potential, Legendre functions, spherical harmonics, inertia integrals (generalised products of inertia).

Mathematics Subject Classification (MSC) 2020: 70 F 15, 33 C 55, 42 C 10, 86 A 20.

1. Introduction: Theoretical Context and Scope

We consider the *usual model of three–dimensional space* $\mathbf{R}^3$, endowed with the well–known algebraic, geometric and topological structures of a linear, affine and Euclidean space over the field $\mathbf{R}$ of the real numbers.

We also consider a *rigid body* of arbitrary geometrical shape containing a mass distribution inside its volume. A simple mathematical model for this situation is provided by a *bounded, connected open subset* $\mathcal{D}$ in ordinary space $\mathbf{R}^3$, delimited by a closed and sufficiently smooth surface $\mathcal{S} = \partial\,\mathcal{D}$ (the *boundary* of $\mathcal{D}$). We further assume that this distribution of matter is characterised by an arbitrary scalar function of position describing the *local density of mass* at each point of the body (say, in a neighbourhood of the point); although in principle this function can be supposed to be bounded and Riemann–integrable over the volume of the body, for certain purposes it should be assumed to be of the class $C^{(1}\left(\mathcal{D} \subseteq \mathbf{R}^3\,,\,\mathbf{R}\right)$ over the said volume of the body.

In particular, this model provides us with a first approach to the study of the force field of gravitational attraction created by many celestial bodies (and, more specifically, the Earth).

It is a well–known fact in some branches of Space Technology and Mathematical and Physical Sciences (e. g., Vector Analysis, Potential Theory, Celestial Mechanics, Astrodynamics, Physical Geodesy, Geophysics) that the vector field corresponding to the gravitational force of attraction created by a mass distribution confined inside an open, bounded and connected set contained in ordinary, three dimensional space can be expressed in terms of the gradient of a single scalar function of position, known as the *(scalar) potential* of that vector field.

Moreover, if spherical polar coordinates $(\,r\,,\,\theta\,,\,\lambda\,)$ are chosen to analyse this issue, that scalar potential can be expressed in the form of an *absolutely convergent series of spherical harmonics* in which products of associated Legendre functions of the first kind (depending on $\cos\theta$, the cosine of the colatitude θ) and elementary circular functions (namely, cosine and sine functions) of integer multiples of the longitude λ are involved.

The so–called *integrals of inertia* (also known as *inertial integrals* or *inertia integrals*) were introduced as a generalisation of the triple integrals (taken over the whole volume of the body) that define the position vector of the centre of mass of the body and its moments and products of inertia. For this reason they are also called *generalised products of inertia*. Accordingly, the volume integrals defining both the centre of mass and the moments and products of inertia of the body are viewed as particular instances of inertia integrals.

The preceding statements and comments, as well as most of the theoretical background concerning this paper, can be documented in detail and justified with the help of some pertinent bibliographical references. For example (just to mention but a few of them), Brouwer and Clemence, [2], Chapter III, pp. 115–133; Cid and Ferrer, [3], Chapter 7, pp. 185–216, and Appendix B, pp. 443-479; Fitzpatrick, [4], Chapter 12, pp. 265–309; Heiskanen and

Moritz, [5], Chapter 1, pp. 1–45, and Chapter 2, §2.5–§2.6, pp. 57–63; MacMillan, [6], Chapter II, pp. 24–95, and Chapter VII, pp. 325–406; Roy, [7], Chapter 7, §7.5, pp. 201–206, and Chapter 11, §11.7, p. 342.

In the present paper we derive *general expressions, in terms of inertia integrals, for the coefficients of the diverse (zonal, tesseral, and sectorial) spherical harmonics* of any degree n and order k occurring in the series expansion of the gravitational potential.

2. Some Basic Concepts and Notations

• We consider the usual affine and Euclidean three–dimensional space $\mathbf{R}^3$, and *a fixed, inertial Cartesian reference frame* $O\,x_1\,x_2\,x_3$, or $O\,x\,y\,z$, with its origin at a point O in $\mathbf{R}^3$. This rectangular coordinate system is determined by the choice of point O and an ordered basis $\{\,\mathbf{i}_1\,,\,\mathbf{i}_2\,,\,\mathbf{i}_3\,\}$ that we also suppose orthonormal and positively oriented (right–handed or dextrorse basis) in Euclidean vector space $\mathbf{R}^3$. This spatial coordinate frame is also denoted $\{\,O\,,\,\{\,\mathbf{i}_1\,,\,\mathbf{i}_2\,,\,\mathbf{i}_3\,\}\}$.

• Given a point P in $\mathbf{R}^3$, it **position vector** *with respect to this Cartesian reference frame* $O\,x\,y\,z$ is

$$\overrightarrow{O\,P} \;\equiv\; \mathbf{r} \;\equiv\; \mathbf{x} \;=\; x\,\mathbf{i}_1 \,+\, y\,\mathbf{i}_2 \,+\, z\,\mathbf{i}_3 \;\equiv\; (\,x\,,\,y\,,\,z\,)_{(\,\mathbf{i}_1\,,\,\mathbf{i}_2\,,\,\mathbf{i}_3\,)} \;\equiv\; (\,x\,,\,y\,,\,z\,)\;. \tag{2.1}$$

• Let $\mathcal{D} \subseteq \mathbf{R}^3$ be a bounded *domain* or bounded *region* (a connected open subset) in $\mathbf{R}^3$, delimited by a closed and smooth surface $\mathcal{S} = \partial\,\mathcal{D}$ (the boundary of $\mathcal{D}$), and $\overline{\mathcal{D}} = \mathcal{D} \cup \partial\,\mathcal{D}$ its topological closure.

• Let $Q \in \mathcal{D}$ be an arbitrary point in this domain, located in space by its position vector (relative to the above Cartesian coordinate system), $\overrightarrow{O\,Q} \;=\; \xi\,\mathbf{i}_1 \,+\, \eta\,\mathbf{i}_2 \,+\, \zeta\,\mathbf{i}_3 \;\equiv\; (\,\xi\,,\,\eta\,,\,\zeta\,)$. The Euclidean distance between Q and P, that is, the Euclidean norm of the position vector of P relative to Q, is

$$\overline{Q\,P} \;=\; ||\,\overrightarrow{Q\,P}\,|| \;=\; \sqrt{(\,x\,-\,\xi\,)^{\,2} \,+\, (\,y\,-\,\eta\,)^{\,2} \,+\, (\,z\,-\,\zeta\,)^{\,2}}\;. \tag{2.2}$$

• In what follows the **notations** for the position variables of the orthogonal curvilinear system of *spherical polar coordinates* will be $(\,r\,,\,\theta\,,\,\lambda\,)$, where $r \,=\, ||\,\overrightarrow{O\,P}\,|| \,=\, ||\,\mathbf{r}\,||$ stands for the radius vector of P (**Euclidean distance** of point P from the origin O of the coordinate system), θ designates the **colatitude** of P (that is, the polar angle of the radius vector, measured from the positive part of the $O\,z \,\equiv\, O\,x_3$ coordinate axis), and λ is the **longitude** of P (azimuthal angle –measured from the positive part of the $O\,x \,\equiv\, O\,x_1$ coordinate axis– that locates the plane that contains point P and is orthogonal to the coordinate plane $O\,x\,y \,\equiv\, O\,x_1\,x_2$.

Accordingly, $r \,\geq\, 0$, i.e., $r \,\in\, [\,0\,,\,+\infty\,) \,=\, \mathbf{R}_+ \,\cup\, \{\,0\,\}$; $0 \,\leq\, \theta \,\leq\, \pi$, that is, $\theta \,\in\, [\,0\,,\,\pi\,]$; and $0 \,\leq\, \lambda \,<\, 2\,\pi$, or $\lambda \,\in\, [\,0\,,\,2\,\pi\,)$.

• For any point $Q \,\in\, \mathcal{D}$. its position in space will be characterized by means of its Cartesian coordinates $(\,\xi\,,\,\eta\,,\,\zeta\,)$, related to its spherical polar coordinates $(\,\rho\,,\,\Theta\,,\,\Lambda\,)$ by means of the equations

$$\xi \;=\; \rho\,\sin\Theta\,\cos\Lambda\;,\quad \eta \;=\; \rho\,\sin\Theta\,\sin\Lambda\;,\quad \zeta \;=\; \rho\,\cos\Theta\;,\quad \text{with}\quad \xi^{\,2} \,+\, \eta^{\,2} \,+\, \zeta^{\,2} \;=\; \rho^{\,2}\;. \tag{2.3}$$

• In a similar way, let $P \,\in\, \mathbf{R}^3 \,\setminus\, \overline{\mathcal{D}}$ be an exterior point, with Cartesian and spherical polar coordinates $(\,x\,,\,y\,,\,z\,)$ and $(\,r\,,\,\theta\,,\,\lambda\,)$, respectively,

$$x \;=\; r\,\sin\theta\,\cos\lambda\;,\quad y \;=\; r\,\sin\theta\,\sin\lambda\;,\quad z \;=\; r\,\cos\theta\;,\quad \text{and}\quad x^{\,2} \,+\, y^{\,2} \,+\, z^{\,2} \;=\; r^{\,2}\;. \tag{2.4}$$

• We consider a distribution of matter confined in the bounded domain $\mathcal{D}$. The Newtonian potential of the gravitational attraction created at point $P\,(\,x\,,\,y\,,\,z\,)$, at which a mass m_P is located outside of the body, by a systems of material points $Q\,(\,\xi\,,\,\eta\,,\,\zeta\,)$ contained in a domain $\mathcal{D}$, is given by ([2], Ch. III, §2, Eqs. (5)–(6), p. 117; [3], Ch. 7, §7.1, Eq. (7.1.2), p. 185, and §7.2, §§7.2.1, Eq. (7.2.6), p. 187; [4], Ch. 12, §12.4, Eq. (12.4.6), p. 290; [5], Ch. 1, §1.2, Eq. (1.11), pp. 3–4; [6], Ch. II, §20, Eq. (1), p. 24; [7], Ch. 7, §7.5, p. 202)

$$V\,(\,P\,) \;=\; \mathcal{G}\,m_P \int\int\int_{\mathcal{D}} \frac{d\,m\,(\,Q\,)}{||\,\overrightarrow{Q\,P}\,||} \;=\; \mathcal{G}\,m_P \int\int\int_{\mathcal{D}} \frac{\rho_{\,vol.}\,(\,Q\,)}{||\,\overrightarrow{Q\,P}\,||}\;d\,v\,(Q)\;, \tag{2.5}$$

where $\mathcal{G}$ is the *universal gravitational constant*, while $d\,m\,(\,Q\,)$ is the *differential element of mass* (or elementary mass) at point Q, and $||\,\overrightarrow{Q\,P}\,||$ is the Euclidean *distance* between Q and P. In practice one takes $m_P \,=\, 1$, the unit mass. As for $\mathcal{G}$, its value in SI units is $\mathcal{G} \;\approx\; 6.67259 \times 10^{\,-\,11}\;\mathrm{N\,m^2\,/\,kg^2} \;=\; 6.67259 \times 10^{\,-\,11}\;\mathrm{m^3\,/\,kg\,s^2}$. If $\rho_{\,vol.}\,(\,Q\,)$ is the *local density of mass* at point Q, and $d\,v\,(\,Q\,)$ the *differential element of volume* in the neighbourhood of Q, then the differential element of mass can be expressed as $d\,m\,(\,Q\,) \,=\, \rho_{\,vol.}\,(\,Q\,)\;d\,v\,(\,Q\,)$.

• This function V turns out to be *harmonic at points outside the domain* $\mathcal{D}$ (and consequently satisfies Laplace's equation $\Delta\,V \,=\, \nabla^{\,2}\,V \,=\, 0$ outside $\mathcal{D}$), while it satisfies Poisson's equation $\Delta\,V \,=\, \nabla^{\,2}\,V \,=\, -\,4\,\pi\,\mathcal{G}\,\rho$ in $\mathcal{D}$.

• Let n be a non–negative integer number, and k a non–negative integer between 0 and n, that is, $n \in \mathbf{N} \cup \{0\}$, and $k \in \{0, 1, 2, \cdots, n\}$. The *Legendre polynomial of degree* n in the independent variable t is denoted $P_n(t)$, while $P_n^k(t)$ will be the *associated Legendre function of the first kind of degree* n *and order* k. Note that for $k = 0$, $P_n^0(t) = P_n(t)$. For the purposes of the present paper, the scalar variable t will be taken as the *cosine function of the colatitude*.

• Within the framework of this theory of (surface) spherical harmonics, terms of the form $P_n(t)$ are called *zonal harmonics* of degree n. Terms $P_n^k(t)\cos(k\lambda)$ and $P_n^k(t)\sin(k\lambda)$ with $0 \neq k \neq n$ are *tesseral harmonics* of degree n and order k, while $P_n^n(t)\cos(n\lambda)$ and $P_n^n(t)\sin(n\lambda)$ are known as *sectorial harmonics* of degree n (and order n).

3. On the Series Expansion of the Gravitational Potential in Terms of Spherical Harmonics

• Taking $m_P = 1$, the gravitational potential given in Eq. (2.5) can be recast in the form ([3], Ch. 7, §7.6, §§7.6.1, p. 206, Eq. (7.6.4); [4], Chapter 12, §12.1, p. 275, Eqs. (12.1.23)–(12.1.24), and §12.2, p. 279, Eq. (12.2.5); [5] Ch. 2, §2.5, Eqs. (2.37)–(2.38) and (2.39)–(2.40), pp. 59–60, with notations as in Ch. 1, §1.13, Eqs. (1.67), p. 29; [7] Ch. 11 §11.7 p. 342)

$$V = \frac{\mathcal{G}M}{r}\left\{1 + \frac{1}{M}\sum_{n=1}^{\infty}\int\int\int_{\mathcal{D}}\left(\frac{\rho}{r}\right)^n \left[P_n(\cos\theta)\,P_n(\cos\Theta) + 2\sum_{k=1}^{n}\frac{(n-k)!}{(n+k)!}P_n^k(\cos\theta)\,P_n^k(\cos\Theta)\cos\{k(\Lambda-\lambda)\}\right] dm\right\}, \tag{3.1}$$

where M stands for the total mass of the distribution of matter contained in the domain $\mathcal{D}$.

• Introducing an auxiliary quantity $R = \sup\left\{\rho = ||\overrightarrow{OQ}|| = \text{distance}(O, Q)\,/\,Q \in \mathcal{D}\right\}$ ([3], Ch. 7, §7.6, §§7.6.1, p. 206; [4], Ch. 12, §12.1, p. 275), and defining the following *coefficientes* ([3], pp. 206–207, Eqs. (7.6.5); [5] Ch. 2, §2.5, Eqs. (2.38), p. 59, and Eqs. (2.40), p. 60),

$$J_n = -\frac{1}{M}\int\int\int_{\mathcal{D}}\left(\frac{\rho}{R}\right)^n P_n(\cos\Theta)\,dm\,, \tag{3.2}$$

$$C_n^k = -\frac{2}{M}\frac{(n-k)!}{(n+k)!}\int\int\int_{\mathcal{D}}\left(\frac{\rho}{R}\right)^n P_n^k(\cos\Theta)\cos k\Lambda\,dm\,, \tag{3.3}$$

$$S_n^k = -\frac{2}{M}\frac{(n-k)!}{(n+k)!}\int\int\int_{\mathcal{D}}\left(\frac{\rho}{R}\right)^n P_n^k(\cos\Theta)\sin k\Lambda\,dm\,, \tag{3.4}$$

the preceding potential (3.1) takes on the form ([3], Eq. (7.6.6)), p. 207; [5], Eqs. (2.39)–(2.40), pp. 59–60)

$$\begin{aligned} V &= \frac{\mathcal{G}M}{r}\left\{1 - \sum_{n=1}^{\infty}\left(\frac{R}{r}\right)^n\left[J_n P_n(\cos\theta) + \sum_{k=1}^{n}P_n^k(\cos\theta)\left(C_n^k\cos k\lambda + S_n^k\sin k\lambda\right)\right]\right\} \\ &= \frac{\mathcal{G}M}{r}\left\{1 - \sum_{n=1}^{\infty}\left(\frac{R}{r}\right)^n\left[J_n P_n(\cos\theta) + \sum_{k=1}^{n}\left(C_n^k P_n^k(\cos\theta)\cos k\lambda + S_n^k P_n^k(\cos\theta)\sin k\lambda\right)\right]\right\} \\ &= \frac{\mathcal{G}M}{r}\left\{1 - \sum_{n=1}^{\infty}\left(\frac{R}{r}\right)^n\left[J_n P_n(\cos\theta) + \sum_{k=1}^{n}\left(C_n^k\left\{P_n^k(\cos\theta)\cos k\lambda\right\} + S_n^k\left\{P_n^k(\cos\theta)\sin k\lambda\right\}\right)\right]\right\} \\ &= \frac{\mathcal{G}M}{r}\left\{1 - \sum_{n=1}^{\infty}\left(\frac{R}{r}\right)^n\left[J_n P_n(\cos\theta) + \sum_{k=1}^{n}\left(C_n^k C_n^k(\theta,\lambda) + S_n^k\mathcal{S}_n^k(\theta,\lambda)\right)\right]\right\}, \end{aligned} \tag{3.5}$$

where the **notations** $C_n^k(\theta,\lambda)$ and $\mathcal{S}_n^k(\theta,\lambda)$ represent the *surface spherical harmonics*, namely

$$C_n^k(\theta,\lambda) = P_n^k(\cos\theta)\cos k\lambda\,, \qquad \mathcal{S}_n^k(\theta,\lambda) = P_n^k(\cos\theta)\sin k\lambda\,. \tag{3.6}$$

• The above constants J_n, C_n^k, and S_n^k, introduced in Eqs. (3.2)–(3.4), are *measures of the amplitudes of the various harmonics* $C_n^0(\theta,\lambda) = P_n(\cos\theta)$, $C_n^k(\theta,\lambda)$, and $\mathcal{S}_n^k(\theta,\lambda)$, respectively ([7], §11.7, p. 342).

♣ In what follows we will propose general expressions for the adimensional coefficients (3.2), (3.3), and (3.4), in terms of inertia integrals (4.5) of the body (see below).

4. Mathematical Formulae and Results Invoked in the Derivation of General Expressions for the Amplitudes

In this Section we collect some formulae and results to which we resort in our considerations and developments leading to the construction of general expression for the coefficients of the spherical harmonics of the gravitational potential.

4.1. Legendre Functions of the First Kind

• To calculate those coefficients, we will start from the following algebraic expression of the *associated Legendre function of the first kind of degree* n *and order* k ([3], App. B, §B.2, §§B.2.1, p. 453; [5], Ch. 1, §1.11, Eq. (1.62), p. 24; [6], Ch. VII, §197, Eq. (8), p. 370),

$$P_n^k(t) \;=\; \frac{\left(1-t^2\right)^{k/2}}{2^n} \sum_{j=0}^{s} (-1)^j \, \frac{(2n-2j)!}{j!\,(n-j)!\,(n-k-2j)!} \; t^{n-k-2j} \;, \tag{4.1}$$

where s is the greatest integer number $\le (n-k)/2$; i.e., $s = (n-k)/2$ or $s = (n-k-1)/2$, whichever is an integer. That is, $s = (n-k)/2$ or $s = (n-k-1)/2$ according as $n-k$ is even or odd. In other words, number s is the *integer part* of $(n-k)/2$.

• Taking $t = \cos\Theta$ allows us to rewrite the preceding algebraic expression (4.1) in the *trigonometric form*

$$P_n^k(\cos\Theta) \;=\; \frac{\sin^k\Theta}{2^n} \sum_{j=0}^{s} (-1)^j \, \frac{(2n-2j)!}{j!\,(n-j)!\,(n-k-2j)!} \; (\cos\Theta)^{n-k-2j}\,\Theta \;, \tag{4.2}$$

• In particular, when $k = 0$, the *Legendre polynomial of degree* n ([3], App. B, §B.2, §§B.2.1 Eq. (B.2.6), p. 452;) reads

$$P_n(t) \;=\; \frac{1}{2^n} \sum_{\ell=0}^{r} (-1)^\ell \, \frac{(2n-2\ell)!}{\ell!\,(n-\ell)!\,(n-2\ell)!} \; t^{n-2\ell} \;, \tag{4.3}$$

$$P_n(\cos\Theta) \;=\; \frac{1}{2^n} \sum_{\ell=0}^{r} (-1)^\ell \, \frac{(2n-2\ell)!}{\ell!\,(n-\ell)!\,(n-2\ell)!} \; (\cos\Theta)^{n-2\ell} \;, \tag{4.4}$$

where r is the integer part of $n/2$.

4.2. Integrals of Inertia of a Body

• The general definition of this concept (also known as *inertial integrals*, *integrals of inertia*, or *generalised products of inertia* of the body) obeys the formula ([4], Ch. 12, §12.4, p. 293; [6], Ch. II, §50, p. 89)

$$I_{i,j,k} \;=\; I_{ijk} \;=\; \int\int\int_{\mathcal{D}} \xi^i \, \eta^j \, \zeta^k \, dm \;, \quad \text{where } i, j, k \; i \text{ are non–negative integers.} \tag{4.5}$$

For a given non–negative integer number n, integrals $I_{i,j,k}$, with i, j, k non–negative integers such that $i + j + k = n$, are also called *moments of orden* n.

• When dealing with spherical harmonics of degree n we will furthermore consider that $i + j + k = n =$ degree of the harmonic .

• Some authors use these integrals only in the case of harmonics of low degree ([2], Ch. III, §6, p. 126 for $n = 3$, and §7, p 126, for $n = 4$; [7], Ch. 7, §7.5, p. 204 for $n = 3$).

• Obviously ([4], p. 293), $I_{0,0,0} = M =$ total mass of the body . Fitzpatrick ([4], Ch. 12, §12.7, Eqs. (12.7.1)–(12.7.4), pp. 306–307) gives explicit expressions for the terms of the gravitational potential up to degree 3, in terms of spherical coordinates, with the amplitudes of the spherical harmonics, namely coefficients

$$J_1, C_1^1, S_1^1, J_2, C_2^1, S_2^1, C_2^2, S_2^2, J_3, C_3^1, S_3^1, C_3^2, S_3^2, C_3^3, S_3^3,$$

represented in terms of inertia integrals as linear combinations of the said integrals up to order three.

♣ Here we shall establish that the coeficientes J_n, C_n^k, and S_n^k of the harmonics of degree n depend on linear combinations of integrals $I_{i,j,k}$ with $i + j + k = n =$ degree of the harmonic .

4.3. Multiple–angle Formulae

According to Weisstein [9], for a positive integer k,

$$\sin(k\alpha) = \sum_{p=0}^{h} (-1)^p \binom{k}{2p+1} \sin^{2p+1}\alpha \cos^{k-2p-1}\alpha \quad , \tag{4.6}$$

$$\cos(k\alpha) = \sum_{p=0}^{H} (-1)^p \binom{k}{2p} \sin^{2p}\alpha \cos^{k-2p}\alpha \quad , \tag{4.7}$$

where h is the integer part of $(k-1)/2$, and H denotes the integer part of $k/2$.

4.4. The Multinomial Theorem

• The *Multinomial Theorem* (attributed to Johann Bernoulli and Leibniz) is a generalisation of Newton´s Binomial Theorem that provides us with a formula for the non–negative entire powers of a polynomial (say, multinomial) expression. Let m be a positive integer number, and n a non–negative integer.

• Consider a multinomial expression $(a_1 + a_2 + \cdots + a_m)$ with m terms (m monomials). Then, from Abramowitz and Stegun ([1], Ch. 24, §24.1, §§24.1.2, §§§24.1.2.I, p. 823), and Weisstein [8],

$$(a_1 + a_2 + \cdots + a_m)^n = \left(\sum_{i=1}^{m} a_i\right)^n$$

$$= \sum_{n_1+n_2+\cdots+n_m=n} \frac{n!}{n_1!\,n_2!\cdots n_m!}\, a_1^{n_1} a_2^{n_2} \cdots a_m^{n_m}$$

$$= \sum_{n_1+n_2+\cdots+n_m=n} \binom{n}{n_1, n_2, \cdots, n_m} \prod_{i=1}^{m} a_i^{n_i} \quad , \tag{4.8}$$

where the sum of the (non–negative) exponents $n_i \in \mathbf{N} \cup \{0\}$ is n: $\sum_{i=1}^{m} n_i = n$. Note that the sum is taken over all combinations of non–negative integers $n_1, n_2, \cdots, n_m$ such that $n_1 + n_2 + \cdots + n_m = n$.

• The *multinomial coefficients* (or *multinomial numbers*) are

$$\binom{n}{n_1, n_2, \cdots, n_m} = \frac{n!}{n_1!\,n_2!\cdots n_m!} \quad . \tag{4.9}$$

• The *number of monomials in the above sums* is

$$\frac{(n+m-1)!}{n!\,(m-1)!} \quad . \tag{4.10}$$

• In particular we are interested in the special case of the multinomial formula for $m = 3$. More specifically, the *trinomial expansion* of $\rho^2 = \xi^2 + \eta^2 + \zeta^2$ (MacMillan [6], Ch. VII, §204, p. 383), namely

$$\rho^{2\ell} = \left(\rho^2\right)^{\ell} = \left(\xi^2 + \eta^2 + \zeta^2\right)^{\ell} = \sum_{\ell_1+\ell_2+\ell_3=\ell} \frac{\ell!}{\ell_1!\,\ell_2!\,\ell_3!}\, \xi^{2\ell_1} \eta^{2\ell_2} \zeta^{2\ell_3} \quad , \tag{4.11}$$

$\ell_j \geq 0$ integer numbers. The number of terms of an expanded trinomial is

$$\frac{(\ell+3-1)!}{\ell!\,(3-1)!} = \frac{(\ell+2)!}{\ell!\,2!} = \frac{(\ell+2)(\ell+1)}{2} \quad , \tag{4.12}$$

where ℓ is the exponent to which the trinomial is raised.

5. Formulae for the Amplitudes of Spherical Harmonics in terms of Inertia Integrals

Theorem 5.1 *Coefficients of zonal harmonics. Let r be the integer part of $n/2$. Then*

$$J_n = -\frac{1}{2^n M R^n} \sum_{\ell=0}^{r} (-1)^{\ell} \frac{(2n-2\ell)!}{(n-\ell)!\,(n-2\ell)!} \left(\sum_{\ell_1+\ell_2+\ell_3=\ell} \frac{1}{\ell_1!\,\ell_2!\,\ell_3!}\, I_{2\ell_1, 2\ell_2, n-2\ell_1-2\ell_2} \right) . \tag{5.1}$$

Proof From the definition (3.2) and the trigonometric form (4.4) of the Legendre polynomial of degree n,

$$\begin{aligned} J_n &= -\frac{1}{M}\int\int\int_{\mathcal{D}} \left(\frac{\rho}{R}\right)^n P_n(\cos\Theta)\, dm \\ &= -\frac{1}{MR^n}\frac{1}{2^n}\int\int\int_{\mathcal{D}} \rho^n \sum_{\ell=0}^{r} \frac{(-1)^\ell (2n-2\ell)!}{\ell!\,(n-\ell)!\,(n-2\ell)!} (\cos\Theta)^{n-2\ell}\, dm \\ &= -\frac{1}{2^n MR^n}\int\int\int_{\mathcal{D}} \sum_{\ell=0}^{r} \frac{(-1)^\ell (2n-2\ell)!}{\ell!\,(n-\ell)!\,(n-2\ell)!} \rho^n \cos^{n-2\ell}\Theta\, dm \\ &= -\frac{1}{2^n MR^n}\int\int\int_{\mathcal{D}} \mathcal{I}_{J_n}\, dm = -\frac{1}{2^n MR^n}\mathbf{I}_{J_n} . \end{aligned}$$

The integrand $\mathcal{I}_{J_n}$ of $\mathbf{I}_{J_n}$ will be treated in the following way,

$$\begin{aligned} \mathcal{I}_{J_n} &= \sum_{\ell=0}^{r} \frac{(-1)^\ell (2n-2\ell)!}{\ell!\,(n-\ell)!\,(n-2\ell)!} (\rho\cos\Theta)^{n-2\ell} \left(\rho^2\right)^\ell \\ &= \sum_{\ell=0}^{r} \frac{(-1)^\ell (2n-2\ell)!}{\ell!\,(n-\ell)!\,(n-2\ell)!} \zeta^{n-2\ell} \left(\xi^2+\eta^2+\zeta^2\right)^\ell \\ &= \sum_{\ell=0}^{r} \frac{(-1)^\ell (2n-2\ell)!}{\ell!\,(n-\ell)!\,(n-2\ell)!} \zeta^{n-2\ell} \left(\sum_{\ell_1+\ell_2+\ell_3=\ell} \frac{\ell!}{\ell_1!\ell_2!\ell_3!} \xi^{2\ell_1}\eta^{2\ell_2}\zeta^{2\ell_3} \right) \\ &= \sum_{\ell=0}^{r} \frac{(-1)^\ell (2n-2\ell)!}{(n-\ell)!\,(n-2\ell)!} \left(\sum_{\ell_1+\ell_2+\ell_3=\ell} \frac{1}{\ell_1!\ell_2!\ell_3!} \xi^{2\ell_1}\eta^{2\ell_2}\zeta^{n-2\ell+2\ell_3} \right) \\ &= \sum_{\ell=0}^{r} \frac{(-1)^\ell (2n-2\ell)!}{(n-\ell)!\,(n-2\ell)!} \left(\sum_{\ell_1+\ell_2+\ell_3=\ell} \frac{1}{\ell_1!\ell_2!\ell_3!} \xi^{2\ell_1}\eta^{2\ell_2}\zeta^{n-2\ell_1-2\ell_2} \right) . \end{aligned}$$

Note that $n-2\ell+2\ell_3 = n-2\ell_1-2\ell_2-2\ell_3+2\ell_3 = n-2\ell_1-2\ell_2$. And integral $\mathbf{I}_{J_n}$ reads

$$\begin{aligned} \mathbf{I}_{J_n} &= \int\int\int_{\mathcal{D}} \sum_{\ell=0}^{r} \frac{(-1)^\ell (2n-2\ell)!}{\ell!\,(n-\ell)!\,(n-2\ell)!} \rho^n \cos^{n-2\ell}\Theta\, dm \\ &= \sum_{\ell=0}^{r} \frac{(-1)^\ell (2n-2\ell)!}{(n-\ell)!\,(n-2\ell)!} \left(\sum_{\ell_1+\ell_2+\ell_3=\ell} \frac{1}{\ell_1!\ell_2!\ell_3!} I_{2\ell_1,2\ell_2,n-2\ell_1-2\ell_2} \right), \end{aligned}$$

from which (5.1) follows.

□

Theorem 5.2 *Coefficients of tesseral and sectorial harmonics of the* $C_n^k(\theta,\lambda)$ *type*:

$$\begin{aligned} C_n^k = -\frac{2}{2^n MR^n}\frac{(n-k)!}{(n+k)!} \Bigg(\sum_{j=0}^{s} (-1)^j \frac{(2n-2j)!}{(n-j)!\,(n-k-2j)!} \\ \times \left[\sum_{\ell=0}^{p} (-1)^\ell \binom{k}{2\ell} \left\{ \sum_{j_1+j_2+j_3=j} \frac{1}{j_1!\,j_2!\,j_3!} I_{k-2\ell+2j_1,\,2\ell+2j_2,\,n-k-2j_1-2j_2} \right\} \right] \Bigg) , \end{aligned} \tag{5.2}$$

with s = *integer part of* $(n-k)/2$, *and* p = *integer part of* $k/2$.

Theorem 5.3 *Coefficients of tesseral and sectorial harmonics of the* $\mathcal{S}_n^k(\theta,\lambda)$ *type*:

$$S_n^k = -\frac{2}{2^n M R^n}\frac{(n-k)!}{(n+k)!}\left(\sum_{j=0}^{s}(-1)^j\frac{(2n-2j)!}{(n-j)!\,(n-k-2j)!}\right.$$
$$\left[\sum_{\ell=0}^{q}(-1)^\ell\binom{k}{2\ell+1}\right.$$
$$\left.\left.\left\{\sum_{j_1+j_2+j_3=j}\frac{1}{j_1!\,j_2!\,j_3!}\;I_{k-2\ell+2j_1-1,\,2\ell+2j_2+1,\,n-k-2j_1-2j_2}\right\}\right]\right)\,, \tag{5.3}$$

where $s =$ *integer part of* $(n-k)/2$ *, and* $q =$ *integer part of* $(k-1)/2$ *.*

Remark 5.4 The proof of these last theorems follows the approach and treatment of the case of the coefficients of zonal harmonics.

Acknowledgements

The author has been partially supported by Project E24–20 (Government of Aragón, European Social Fund).

References

[1] Milton Abramowitz and Irene A. Stegun (Editors). *Handbook of Mathematical Funcions with Formulas, Graphs, and Mathematical Tables*. Dover Publications, Inc., New York, 1965. Ninth Dover printing, 1970.

[2] Dirk Brouwer and Gerald M. Clemence. *Methods of Celestial Mechanics*. Academic Press, Inc., New York and London, 1961.

[3] Rafael Cid Palacios and Sebastián Ferrer Martínez. *Geodesia. Geométrica, Física y por Satélites*. Instituto Geográfico Nacional, Ministerio de FomentoI, Madrid, 1997.

[4] Philip M. Fitzpatrick. *Principles of Celestial Mechanics*. Academic Press, Inc., New York and London, 1970. Chapter 12, §12.4, §12.6, and §12.7, and Exercice 15 (p. 309) of that Chapter 12.

[5] Weikko A. Heiskanen and Helmut Moritz. *Physical Geodesy* W. H. Freeman and Company, San Francisco and London, 1967

[6] William Duncan MacMillan. *Theoretical Mechanics. The Theory of the Potential*. Dover Publications, Inc., New York, New York, 1958. Unabridged and unaltered republication of the first edition, 1930. Chapter VII, §204, pp. 382–384.

[7] Archie E. Roy. *Orbital Motion*. IoP Institute of Physics Publishing Ltd., Bristol and Philadelphia, 2005. Fourth Edition.

[8] Eric W. Weisstein. *Multinomial Series*. From *MathWorld*–A Wolfram Web Resource. `http://mathworld.wolfram.com/MultinomialSeries.html`

[9] Eric W. Weisstein. *Multiple–Angle Formulas*. From *MathWorld*–A Wolfram Web Resource. `http://mathworld.wolfram.com/Multiple-AngleFormulas.html`

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 184–191)

Turing instability analysis of a singular cross-diffusion problem

Gonzalo Galiano[1], **Víctor González-Tabernero**[2]
1. galiano@uniovi.es, *Universidad de Oviedo, Spain*
2. victor.gonzalez.tabernero@rai.usc.es, *Universidad de Santiago de Compostela, Spain*

Abstract

The population model of Busenberg and Travis is a paradigmatic model in ecology and tumour modelling due to its ability to capture interesting phenomena like the segregation of populations. Its singular mathematical structure enforces the consideration of regularized problems to deduce properties as fundamental as the existence of solutions. In this article we perform a weakly nonlinear stability analisys of a general class of regularized problems to study the convergence of the instability modes in the limit of the regularization parameter. We demonstrate with some specific examples that the pattern formation observed in the regularized problems, with unbounded wave numbers, is not present in the limit problem due to the amplitude decay of the oscillations. We also check the results of the stability analysis with direct finite element simulations of the problem. In this short communication, we omit the proofs of our results.

In [4], Busenberg and Travis introduced a class of singular cross-diffusion problems under the assumption that the spatial relocation of each species is due to a diffusion flow which depends on the densities of all the involved species. In the case of two species, if u_1, u_2 denote their densities, the flow, in its simplest form, may be assumed to be determined by the total population $u_1 + u_2$, and thus the conservation laws for both species lead to the system

$$\partial_t u_1 - \operatorname{div}\left(u_1(\nabla u_1 + \nabla u_2)\right) = f_1(u_1, u_2), \tag{0.1}$$
$$\partial_t u_2 - \operatorname{div}\left(u_2(\nabla u_1 + \nabla u_2)\right) = f_2(u_1, u_2). \tag{0.2}$$

The functions f_1 and f_2 capture some ecological features of the populations, such as growth, competition, etc. As usual, the equations (0.1)-(0.2) are complemented with non-negative initial data and non-flow boundary conditions.

The system (0.1)-(0.2) is called a *cross-diffusion* system because the flow of each species depend upon the densities of the other species. We call it *singular* because the resulting diffusion matrix is singular. Indeed, when rewritting (0.1)-(0.2) in matrix form, for $\mathbf{u} = (u_1, u_2)$, we get the equation

$$\partial_t \mathbf{u} - \operatorname{div}(\mathcal{A}(\mathbf{u})\nabla \mathbf{u}) = \mathbf{f}(\mathbf{u}), \quad \text{with} \quad \mathcal{A}(\mathbf{u}) = \begin{pmatrix} u_1 & u_1 \\ u_2 & u_2 \end{pmatrix},$$

where the divergence is applied by rows. The full and singular structure of $\mathcal{A}$ introduces serious difficulties in the mathematical analysis of the problem, as we shall comment later.

In his seminal paper [14], Turing introduced a mechanism explaining how spatially uniform equilibria may evolve, small perturbations mediating, into stable equilibria with non-trivial spatial structure. He considered a system of the type

$$\partial_t u_1 - \Delta u_1 = f_1(u_1, u_2), \tag{0.3}$$
$$\partial_t u_2 - \sigma \Delta u_2 = f_2(u_1, u_2), \tag{0.4}$$

with $\sigma > 0$, and proved that when σ is small or large enough then the stable equilibria of the dynamical system

$$\partial_t v_1 = f_1(v_1, v_2), \tag{0.5}$$
$$\partial_t v_2 = f_2(v_1, v_2), \tag{0.6}$$

are not stable for the diffusion system (0.3)-(0.4) and that, in their place, non-uniform equilibria with spatial structure become the new stable solutions. This mechanism is known as *Turing instability* or *Turing bifurcation*.

In this article we study Turing instability for the cross-diffusion singular system (0.1)-(0.2). We already know that some cross-diffusion systems, such as the paradigmatic SKT model introduced by Shigesada, Kawasaki and Teramoto [13], exhibit Turing instability when cross-diffusion coefficients are large in comparison with self-diffusion coefficients, see e.g. [10, 11]. However, the singularity of the diffusion matrix of the system (0.1)-(0.2), not present in the SKT model, introduces important mathematical difficulties to the analysis of this system.

Regarding the existence of solutions of (0.1)-(0.2), it has been proved only in some special situations: for a bounded spatial domain $\Omega \subset \mathbb{R}$ (Bertsch et al. [2]) and for $\Omega = \mathbb{R}^n$ (Bertsch et al. [3]). In their proofs, the following

observation is crucial: adding the two equations of (0.1)-(0.2) shows that if a solution of this system does exist then the total population, $u = u_1 + u_2$, satisfies the porous medium type equation

$$\partial_t u - \operatorname{div}(u\nabla u) = f(u), \tag{0.7}$$

for which the theory of existence and uniqueness of solutions is well established. In particular, if the initial data of the total population is bounded away from zero and if f is regular enough with $f(0) \geq 0$, it is known that the solution of (0.7) remains positive and smooth for all time. This allows to introduce the change of unknowns $w_i = u_i/u$, for $i = 1, 2$, into the original problem (0.1)-(0.2) to deduce the equivalent formulation

$$\partial_t u - \operatorname{div}(u\nabla u) = F_1(u, w_1), \tag{0.8}$$
$$\partial_t w_1 - \nabla u \cdot \nabla w_1 = F_2(u, w_1), \tag{0.9}$$

for certain well-behaved functions F_1 and F_2. Being the structure of the system (0.8)-(0.9) of parabolic-hyperbolic nature, parabolic regularization of the system by adding the term $-\delta\Delta w_1$ to the left hand side of (0.9), and the consideration of the characteristics defined by the field ∇u are the main ingredients of the proofs made by Bertsch et al. [2,3].

In [9] we followed a different approach to prove the existence of solutions of the original system (0.1)-(0.2) for a bounded domain $\Omega \subset \mathbb{R}$. We directly performed a parabolic regularization of the system by introducing a cross-diffusion perturbation term while keeping the porous medium type equation satisfied by u. More concretely, we considered the system

$$\partial_t u_1 - \operatorname{div}\left(u_1(\nabla u_1 + \nabla u_2)\right) - \frac{\delta}{2}\Delta(u_1(u_1 + u_2)) = f_1(u_1, u_2), \tag{0.10}$$
$$\partial_t u_2 - \operatorname{div}\left(u_2(\nabla u_1 + \nabla u_2)\right) - \frac{\delta}{2}\Delta(u_2(u_1 + u_2)) = f_2(u_1, u_2), \tag{0.11}$$

and then used previous results for cross-diffusion systems [6, 7, 12] to establish the existence of solutions of the approximated problems. Then, BV estimates similar to those obtained in [2] allowed to prove the convergence of the sequence $(u_1^{(\delta)}, u_2^{(\delta)})$ to a solution of the original problem. Let us finally mention that the system (0.1)-(0.2) is a limit case of a general type of problems with diffusion matrix given by

$$\mathcal{A}(\mathbf{u}) = \begin{pmatrix} a_{11}u_1 & a_{12}u_1 \\ a_{21}u_2 & a_{22}u_2 \end{pmatrix},$$

for which, if $a_{ii} > 0$, for $i = 1, 2$, and $a_{11}a_{22} > a_{12}a_{21}$ then the existence of solutions in ensured for any spatial dimension of Ω, see [8]. In addition, it has been shown that this kind of systems, when set in the whole space $\Omega = \mathbb{R}^n$, may be obtained as mean field limits [5].

Concerning Turing instability, since the diffusion matrix, $\mathcal{A}(\mathbf{u})$, corresponding to the system (0.1)-(0.2) is singular, the linearization of this system about an equilibrium of the dynamical system (0.5)-(0.6) does not provide any information on the behaviour of the equilibrium in the spatial dependent case. Thus, our approach to the investigation of Turing instability for the system (0.1)-(0.2) relies on the study of this property for approximating problems like (0.10)-(0.11) and its limit behaviour.

We prove that linear instability is always present in the limit $\delta \to 0$, which is the case when the sequence of solutions of the approximated problems (0.10)-(0.11) converges to the solution of the original problem (0.1)-(0.2). Interestingly, the linear analysis also establishes that the main instability wave number is unbounded as $\delta \to 0$.

For a clearer understanding of this convergence of a increasingly oscillating sequence of functions to a *BV* function (the solution of (0.1)-(0.2) ensured in [2, 9]), we perform a weakly nonlinear analysis (WNA) which allows to gain insight into the behaviour of the amplitude of the main instability mode as $\delta \to 0$. As expected, we find that the amplitude of the instability modes vanishes in the limit $\delta \to 0$ resulting, therefore, coherent with the *BV* convergence. In addition, this result also suggests that the uniform equilibrium is stable for the original problem. We furthermore check these analytical results by numerically comparing the WNA approximation to a FEM approximation of the nonlinear problem.

1. Main results

For simplicity, we study Turing instability for the one-dimensional spatial setting which has also the advantage of a well stablished existence theory for the case of a bounded domain [2, 9]. By redefining the functions f_1, f_2, we

can fix without loss of generality $\Omega = (0, \pi)$ and then rewrite problem (0.1)-(0.2) together with the usual auxiliary conditions as

$$\partial_t u_1 - \partial_x\big(u_1(\partial_x u_1 + \partial_x u_2)\big) = f_1(u_1, u_2) \qquad \text{in } Q_T, \tag{1.1}$$
$$\partial_t u_2 - \partial_x\big(u_2(\partial_x u_1 + \partial_x u_2)\big) = f_2(u_1, u_2) \qquad \text{in } Q_T, \tag{1.2}$$
$$u_1(\partial_x u_1 + \partial_x u_2) = u_2(\partial_x u_1 + \partial_x u_2) = 0 \qquad \text{on } \Gamma_T, \tag{1.3}$$
$$u_1(0,\cdot) = u_{10}, \quad u_2(0,\cdot) = u_{20} \qquad \text{in } \Omega, \tag{1.4}$$

where $Q_T = (0,T)\times\Omega$ and the initial data u_{10}, u_{20} are non-negative functions. We assume a competitive Lotka-Volterra form for the reaction term, this is, $f_i(u_1,u_2) = u_i(\alpha_i - \beta_{i1}u_1 - \beta_{i2}u_2)$, for $i = 1,2$, and for some non-negative parameters α_i, β_{ij}, for $i,j = 1,2$.

In order to deal with several types of regularized problems we introduce, for positive δ and b, the uniformly parabolic cross-diffusion system

$$\partial_t u_1 - \partial_x\big(d_{11}^\delta(\mathbf{u})\partial_x u_1 + d_{12}^\delta(\mathbf{u})\partial_x u_2\big) = f_1^b(\mathbf{u}) \qquad \text{in } Q_T, \tag{1.5}$$
$$\partial_t u_2 - \partial_x\big(d_{21}^\delta(\mathbf{u})\partial_x u_1 + d_{22}^\delta(\mathbf{u})\partial_x u_2\big) = f_2^b(\mathbf{u}) \qquad \text{in } Q_T, \tag{1.6}$$
$$d_{11}^\delta(\mathbf{u})\partial_x u_1 + d_{12}^\delta(\mathbf{u})\partial_x u_2 = d_{21}^\delta(\mathbf{u})\partial_x u_1 + d_{22}^\delta(\mathbf{u})\partial_x u_2 = 0 \qquad \text{on } \Gamma_T, \tag{1.7}$$
$$u_1(0,\cdot) = u_{10}, \quad u_2(0,\cdot) = u_{20} \qquad \text{in } \Omega, \tag{1.8}$$

where the diffusion matrix $D^\delta(\mathbf{u}) = (d_{ij}^\delta(\mathbf{u}))$ and the Lotka-Volterra function $\mathbf{f}^b(\mathbf{u}) = (f_1^b(\mathbf{u}), f_2^b(\mathbf{u}))$ satisfy the assumptions H_D:

1. $D^\delta(\mathbf{u})$ is linear in $\mathbf{u}$ and affine in δ, so that it allows the decompositions

$$D^\delta(\mathbf{u}) = D^0(\mathbf{u}) + \delta D^1(\mathbf{u}) = D^{\delta 1}u_1 + D^{\delta 2}u_2,$$

for some matrices $D^{\delta i}$ for $i = 1,2$, being the coefficients of $D^\delta(\mathbf{u})$ given by

$$d_{ij}^\delta(\mathbf{u}) = d_{ij}^{10}u_1 + d_{ij}^{11}u_1\delta + d_{ij}^{20}u_2 + d_{ij}^{21}u_2\delta,$$

for some non-negative constants d_{ij}^{mn}, for $i,j,m = 1,2$ and $n = 0,1$.

2. We assume that $d_{ii}^\delta(\mathbf{u}) > 0$ for $i = 1,2$, and that $\det(D^\delta(\mathbf{u}))$ is an increasing function with respect to δ satisfying $\det(D^\delta(\mathbf{u})) > 0$ if $\delta > 0$ and $\mathbf{u} \in \mathbb{R}^2_+$.

3. For $i,j = 1,2$, $f_i^b(u_1,u_2) = u_i(\alpha_i^b - \beta_{i1}^b u_1 - \beta_{i2}^b u_2)$ for some non-negative α_i^b, β_{ij}^b such that $\alpha_i^b \to \alpha_i$ and $\beta_{ij}^b \to \beta_{ij}$ as $b \to 0$. Moreover, using the notation $\alpha_i^0 = \alpha_i$ and $\beta_{ij}^0 = \beta_{ij}$, we assume, for $b \geq 0$,

$$\begin{aligned} &\beta_{22}^b\alpha_1^b - \beta_{12}^b\alpha_2^b > 0, \quad \beta_{11}^b\alpha_2^b - \beta_{21}^b\alpha_1^b > 0, \\ &\det(B^b) > 0, \quad \operatorname{tr}(B^b) \geq 0, \quad \text{where } B^b = \left(\beta_{ij}^b\right). \end{aligned} \tag{1.9}$$

Observe that (1.9) ensures the existence of a stable coexistence equilibrium for the dynamical system (0.5)-(0.6), given by

$$\mathbf{u}^* = \left(u_1^*, u_2^*\right) = \left(\frac{\beta_{22}^b\alpha_1^b - \beta_{12}^b\alpha_2^b}{\beta_{11}^b\beta_{22}^b - \beta_{12}^b\beta_{21}^b}, \frac{\beta_{11}^b\alpha_2^b - \beta_{21}^b\alpha_1^b}{\beta_{11}^b\beta_{22}^b - \beta_{12}^b\beta_{21}^b}\right). \tag{1.10}$$

There are two examples of $D^\delta(\mathbf{u})$ in which we are specially interested. The first, due to its simplicity for the calculations. We set

$$D^\delta(\mathbf{u}) = \begin{pmatrix} (1+\delta)u_1 & u_1 \\ u_2 & (1+\delta)u_2 \end{pmatrix}, \tag{1.11}$$

for which $\det(D^\delta(\mathbf{u})) = \delta(2+\delta)u_1u_2$. According to [8], the second hypothesis of H_D guarantees the well-posedness of the problem (1.5)-(1.8) corresponding to this diffusion matrix. The second example corresponds to the approximation used in [9] for proving the existence of BV solutions of the original problem (1.1)-(1.4):

$$D^\delta(\mathbf{u}) = \begin{pmatrix} (1+\delta)u_1 + \frac{\delta}{2}u_2 & (1+\frac{\delta}{2})u_1 \\ (1+\frac{\delta}{2})u_2 & \frac{\delta}{2}u_1 + (1+\delta)u_2 \end{pmatrix}, \tag{1.12}$$

for which $\det(D^\delta(\mathbf{u})) = \frac{1}{2}\delta(1+\delta)(u_1+u_2)^2$.

The approximation of the reaction terms introduced in the system (1.5)-(1.8) is not essential. Its aim is to support the specific example we deal with in Theorem 1.3, but can be ignored ($b = 0$) in the general linear and weakly nonlinear analysis of Theorems 1.1 and 1.2. Nevertheless, we state these results taking it into account. Our first result gives conditions under which linear instability arises. The following notation is used:

$$K = D\mathbf{f}^b(\mathbf{u}^*) = \begin{pmatrix} -\beta^b_{11}u^*_1 & -\beta^b_{12}u^*_1 \\ -\beta^b_{21}u^*_2 & -\beta^b_{22}u^*_2 \end{pmatrix}.$$

Theorem 1.1 (Linear instability) *Assume H_D, with $b \geq 0$. Let $\mathbf{u}^*$ be the coexistence equilibrium defined by* (1.10). *If*

$$\operatorname{tr}(K^{-1}D^\delta(\mathbf{u}^*)) > 0 \quad \textit{for all} \quad \delta \geq 0 \tag{1.13}$$

then there exists $\delta_c > 0$ such that if $\delta < \delta_c$ then $\mathbf{u}^*$ is a linearly unstable equilibrium for problem (1.5)-(1.8). *In such situation, the wave number of the main instability mode tends to infinity as $\delta \to 0$.*

Condition (1.13) is equivalent to

$$d^\delta_{11}(\mathbf{u}^*)\beta^b_{22}u^*_2 + d^\delta_{22}(\mathbf{u}^*)\beta^b_{11}u^*_1 < d^\delta_{12}(\mathbf{u}^*)\beta^b_{21}u^*_2 + d^\delta_{21}(\mathbf{u}^*)\beta^b_{12}u^*_1 \tag{1.14}$$

and introduces a further restriction on the matrix of competence coefficients. Roughly speaking, for B^b to fulfil both (1.9) and (1.14), its elements must be such that intra-population joint competence is larger than inter-population joint competence (condition (1.9)) and one of the inter-population competence coefficients is large in comparison with the others (condition (1.14)). A numeric example we shall work with along the article is

$$B^b = \begin{pmatrix} 1 & \dfrac{b}{2} \\ 2 & 1 \end{pmatrix}, \quad \text{with} \quad b \in (0, \tfrac{1}{2}). \tag{1.15}$$

Assuming the forms of $D^\delta(\mathbf{u}^*)$ given in Examples 1 and 2, see (1.11) and (1.12), we have that the conditions (1.9) and (1.14) on B^b are satisfied if $\delta < b/4$ (Example 1) or $\delta < bu^*_1u^*_2/(u^*_1+u^*_2)^2$ (Example 2). Therefore, the most meaningful case when δ is close to zero is satisfied by both diffusion matrices.

Our second result allows to estimate not only the instability wave numbers provided by the linear analysis but also the amplitude corresponding to these modes. The approximation of the steady state solution is obtained using a weakly nonlinear analysis (WNA) based on the method of multiple scales.

Theorem 1.2 *Assume the hypothesis of Theorem* 1.1 *and let $\varepsilon^2 = (\delta_c - \delta)/\delta_c$ be a small number. Then, there exist sets of data problem such that the stationary WNA approximation to the solution $\mathbf{u}$ of problem* (1.5)-(1.8) *is given by*

$$\mathbf{v}(x) = \mathbf{u}^* + \varepsilon\boldsymbol{\rho}\sqrt{A_\infty}\cos(k_c x) + \varepsilon^2 A_\infty\big(\mathbf{v}_{20} + \mathbf{v}_{22}\cos(2k_c x)\big) + O(\varepsilon^3), \tag{1.16}$$

where $k_c \in \mathbb{Z}$ is the critical wave number corresponding to δ_c, A_∞ is a positive constant and $\boldsymbol{\rho}, \mathbf{v}_{20}$ and $\mathbf{v}_{22}$ are constant vectors.

Our third result focuses on the limit behaviour of the critical parameters and the amplitude when $\delta \to 0$, this is, when the solutions of the approximated problems converge to the solution of the original singular problem. For the sake of simplicity, we limit our study to the following example:

$$\partial_t u_1 - \partial_x\big(u_1(\partial_x u_1 + \partial_x u_2)\big) = u_1(1-u_1), \tag{1.17}$$
$$\partial_t u_2 - \partial_x\big(u_2(\partial_x u_1 + \partial_x u_2)\big) = u_2(4-(2u_1+u_2)), \tag{1.18}$$

whose solutions we approximate by the two-parameter family of solutions of

$$\partial_t u_1 - \partial_x\big(u_1((1+\delta)\partial_x u_1 + \partial_x u_2)\big) = u_1\Big(1-\big(u_1 + \frac{b}{2}u_2\big)\Big), \tag{1.19}$$
$$\partial_t u_2 - \partial_x\big(u_2(\partial_x u_1 + (1+\delta)\partial_x u_2)\big) = u_2(4-(2u_1+u_2)). \tag{1.20}$$

On one hand, Theorem 1.1 ensures the existence of $\delta_c > 0$ such that, for any $b \geq 0$, the equilibrium $\mathbf{u}^* = \frac{1}{1-b}(1-2b, 2)$ of (1.19)-(1.20) becomes unstable for $\delta < \delta_c$, with an associated critical wave number such that $k_c \to \infty$ as $\delta \to 0$.

On the other hand, for $\delta < b/4$ and $b \to 0$, the sequence of solutions of (1.19)-(1.20) converges to a solution of (1.17)-(1.18) in the space $BV(0,T,L^\infty(\Omega)) \cup L^\infty(0,T;BV(\Omega))$. Therefore, for the approximation (1.16) provided by the weakly nonlinear analysis to remain valid for all $\delta > 0$, the corresponding amplitude A_∞ must vanish in the limit $\delta \to 0$, making in this way compatible the increase of oscillations with its BV regularity.

Theorem 1.3 *Set $\alpha = (1,4)$, and let $D^\delta(\mathbf{u})$ and B^b be given by* (1.11) *and* (1.15)*, respectively, for $b < 1/2$ and $0 < \delta < b/4$. Then, there exists $\delta_c(b) > 0$ such that if $\delta < \delta_c(b)$ then $\mathbf{u}^* = \frac{1}{1-b}(1-2b,2)$ is linearly unstable for problem* (1.5)-(1.8)*. In addition,*

$$\lim_{b\to 0} \delta_c(b) = 0, \quad \lim_{b\to 0} k_c(b) = \infty,$$

and the amplitude provided by the weakly nonlinear analysis satisfies

$$\lim_{b\to 0} A_\infty(b) = 0.$$

In particular, the weakly nonlinear approximation $\mathbf{v}$ given by (1.16) *satisfies $\mathbf{v} \to \mathbf{u}^*$ uniformly in Ω as $b \to 0$.*

2. Numerical experiments

In order to analyze the quality of the approximation provided by the WNA, as well as the properties stated in Theorems 1.1 to 1.3, we compare it to a numerical approximation of the evolution problem computed though the finite element method (FEM).

For the FEM approximation, we used the open source software `deal.II` [1] to implement a time semi-implicit scheme with a spatial linear-wise finite element discretization. For the time discretization, we take in the experiments a uniform time partition of time step $\tau = 0.01$. For the spatial discretization, we take a uniform partition of the interval $\Omega = (0,\pi)$ with spatial step depending on the predicted wave number of the pattern, see Table 1.

Let, initially, $t = t_0 = 0$ and set $(u_1^0, u_2^0) = (u_{10}, u_{20})$. For $n \geq 1$, the discrete problem is: Find $u_1^n, u_2^n \in S^h$ such that

$$\frac{1}{\tau}(u_1^n - u_1^{n-1}, \chi)^h + \left(d_{11}^\delta(\mathbf{u}^n)\partial_x u_1^n + d_{12}^\delta(\mathbf{u}^n)\partial_x u_2^n, \partial_x\chi\right)^h = \left(f_1^b(u_1^n, u_2^n), \chi\right)^h, \tag{2.1}$$

$$\frac{1}{\tau}(u_2^n - u_2^{n-1}, \chi)^h + \left(d_{21}^\delta(\mathbf{u}^n)\partial_x u_1^n + d_{22}^\delta(\mathbf{u}^n)\partial_x u_2^n, \partial_x\chi\right)^h = \left(f_2^b(u_1^n, u_2^n), \chi\right)^h, \tag{2.2}$$

for every $\chi \in S^h$, the finite element space of piecewise $\mathbb{Q}_1$-elements. Here, $(\cdot,\cdot)^h$ stands for a discrete semi-inner product on $C(\overline{\Omega})$.

Since (2.1)-(2.2) is a nonlinear algebraic problem, we use a fixed point argument to approximate its solution, (u_1^n, u_2^n), at each time slice $t = t_n$, from the previous approximation (u_1^{n-1}, u_2^{n-1}). Let $u_1^{n,0} = u_1^{n-1}$ and $u_2^{n,0} = u_2^{n-1}$. Then, for $k \geq 1$ the linear problem to solve is: Find $(u_1^{n,k}, u_2^{n,k})$ such that for for all $\chi \in S^h$

$$\begin{aligned}
&\frac{1}{\tau}(u_1^{n,k} - u_1^{n-1}, \chi)^h + \left(d_{11}^\delta(\mathbf{u}^{n,k-1})\partial_x u_1^{n,k} + d_{12}^\delta(\mathbf{u}^{n,k-1})\partial_x u_2^{n,k}, \partial_x\chi\right)^h \\
&\qquad = \left(u_1^{n,k}(\alpha_1^b - \beta_{11}^b u_1^{n,k-1} - \beta_{12}^b u_2^{n,k-1}), \chi\right)^h, \\
&\frac{1}{\tau}(u_2^{n,k} - u_2^{n-1}, \chi)^h + \left(d_{21}^\delta(\mathbf{u}^{n,k-1})\partial_x u_1^{n,k} + d_{22}^\delta(\mathbf{u}^{n,k-1})\partial_x u_2^{n,k}, \partial_x\chi\right)^h \\
&\qquad = \left(u_2^{n,k}(\alpha_2^b - \beta_{21}^b u_1^{n,k-1} - \beta_{22}^b u_2^{n,k-1})\chi\right)^h.
\end{aligned}$$

We use the stopping criteria

$$\max\left(\|u_1^{n,k} - u_1^{n,k-1}\|_2, \|u_2^{n,k} - u_2^{n,k-1}\|_2\right) < \mathrm{tol}_{FP},$$

for values of tol_{FP} chosen empirically, and set $(u_1^n, u_2^n) = (u_1^{n,k}, u_2^{n,k})$. Finally, we integrate in time until a numerical stationary solution, (u_1^S, u_2^S), is achieved. This is determined by

$$\max\left(\|u_1^{n,1} - u_1^{n-1}\|_2, \|u_2^{n,1} - u_2^{n-1}\|_2\right) < \mathrm{tol}_S,$$

where tol_S is chosen empirically too. In the following experiments we always fix $\mathrm{tol}_{FP} = 1.e{-}07$ and $\mathrm{tol}_S = 1.e{-}12$.

	Simulation 1	Simulation 2	Simulation 3
b	3.85e-02	9.91e-03	4.42e-03
$\delta(b)$	4.53e-05	2.94e-06	5.83e-07
$k_c(b)$	10	20	30
$A_\infty(b)$	1.21e-02	3.1e-03	1.4e-03
Number of nodes	128	256	512
Time steps to stationary	3.e+04	1.9e+05	4.4e+05
Execution time (hours)	1.67	19.26	84.77

Tab. 1 Data set for the Experiment 1. Wave numbers and times are rounded. Execution time measured for a standard laptop with i7 processor.

2.1. Experiment 1

We investigate the behaviour of the instabilities arising in the solutions of the approximated problems (1.5)-(1.8) when $\delta \to 0$. Our main aim is to check if the predictions of the weakly nonlinear analysis stated in Theorem 1.3 are captured by the FEM approximation too. Thus, we use the diffusion matrix $D^\delta(\mathbf{u})$ and the competence parameters B^b given by (1.11) and (1.15), respectively.

We run three simulations according to the choice of b, see Table 1, and fix $\delta = 0.95\delta_c(b)$ in all of them, so that $\mathbf{u}^*$ is unstable and pattern formation follows.

In Fig. 1 we show the typical onset and transmission of disturbances found in all the experiments. In this figure and in the following we plot only the first component of the solution, being the behaviour of the second component similar. After a fast decay of the initial data towards the unstable equilibrium, a perturbation with the wave number predicted by the linear analysis grows from one side of the boundary to the rest of the domain until reaching the steady state, see Fig. 2. In the latter figure, we may check the good accordance between the FEM and the WNA approximations which, in numeric figures, have a relative difference of the order 10^{-5}.

In Fig. 3 we show three interesting behaviours of solutions when $\delta \to 0$. In the left panel, the shrinking amplitude of the stationary patterns while the wave number increases. The equilibrium has been subtracted from the solution to center the pattern in $y = 0$. The center panel shows the time evolution of the amplitude (log scale) as given by the exact solution of the Stuart-Landau equation. We readily see that the stabilization time is a decreasing function of δ. This fact together with the increment of the wave number when $\delta \to 0$ results in very high execution times, see Table 1. Finally, the third panel shows how the variation of the numerical stationary solution

$$\int_\Omega |\partial_x u_1(T, x)| dx$$

is an increasing function of δ and tends to zero as $\delta \to 0$, in agreement with the regularity of solutions stated by the theoretical results.

2.2. Experiment 2

We repeated Experiment 1 replacing the diffusion matrix $D^\delta(\mathbf{u})$ by that defined in (1.12) In Table 2 we show the relative differences in L^p, given by

$$\mathrm{RD}_p(\varphi_1, \varphi_2) = \frac{\|\varphi_1 - \varphi_2\|_{L^p}}{\|\varphi_1\|_{L^p}}, \tag{2.3}$$

of the critical bifurcation parameter, δ_c, the *stationary* solution of the FEM approximation, $\mathbf{u}(T, \cdot)$, the WNA approximation, $\mathbf{v}$, and the pattern amplitude, A_∞, corresponding to both approximations of the original diffusion matrix. We see that although the critical bifurcation parameter is clearly affected by the approximation scheme, the FEM and WNA approximations provided by both schemes are in a very good agreement, as well as the amplitudes of the instability patterns, suggesting that in the limit $\delta \to 0$ both sequences of approximations converge to the same limit.

Acknowledgements

First author supported by the Spanish MEC Project MTM2017-87162-P.

	Simulation 1	Simulation 2	Simulation 3
$\mathrm{RD}_\infty(\delta_c^{(E1)}, \delta_c^{(E2)})$	0.136	0.117	0.113
$\mathrm{RD}_2(\mathbf{u}^{(E1)}(T,\cdot), \mathbf{u}^{(E2)}(T,\cdot))$	3.74e-06	8.68e-07	5.41e-07
$\mathrm{RD}_2(\mathbf{v}^{(E1)}, \mathbf{v}^{(E2)})$	3.46e-06	2.13e-07	4.19e-08
$\mathrm{RD}_\infty(A_\infty^{(E1)}, A_\infty^{(E2)})$	2.90e-03	6.82e-04	2.99e-4

Tab. 2 Comparison between the results obtained with the approximation diffusion matrices corresponding to Example 1 (E1) and Example 2 (E2), given by (1.11) and (1.12) respectively. RD_p denotes the relative difference in L^p, see (2.3).

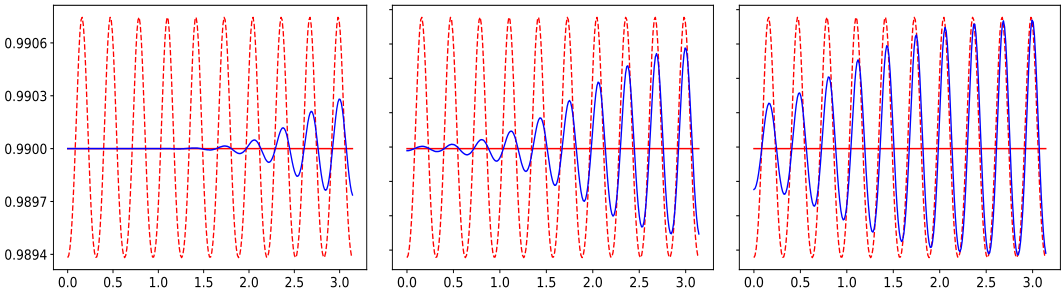


Fig. 1 Typical evolution of disturbances

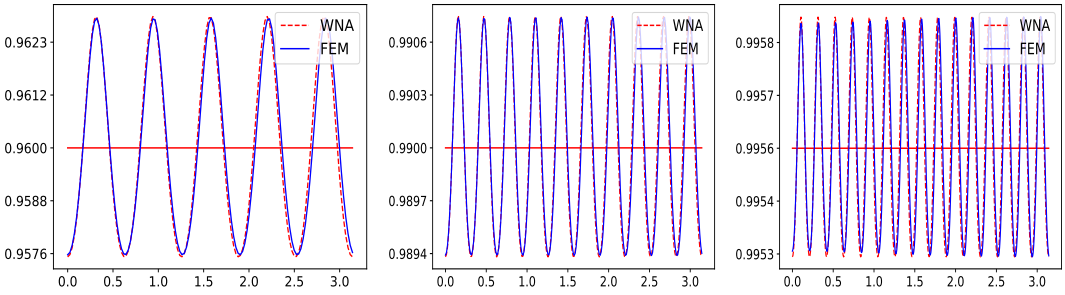


Fig. 2 Experiment 1. WNA and FEM approximations corresponding to Simulations 1 to 3 (left to right). Notice the different scales in the ordinates axis showing the decreasing amplitude of the oscillations.

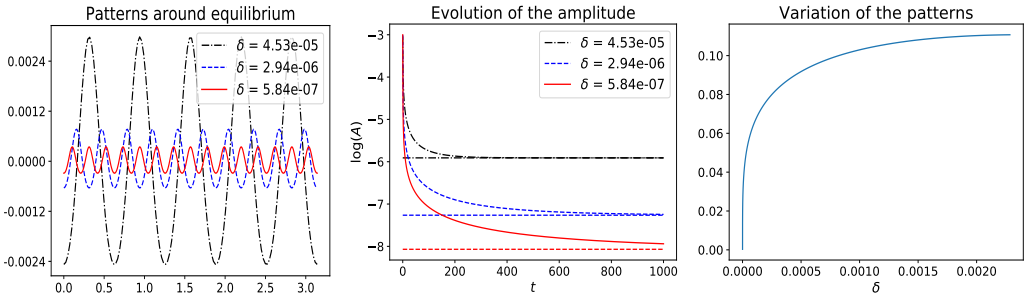


Fig. 3 Experiment 1. Behaviour of the patterns as $\delta \to 0$.

References

[1] W. Bangerth, T. Heister, L. Heltai, G. Kanschat, M. Kronbichler, M. Maier, B. Turcksin, The `deal.II` Library, Version 8.3, Arch. Numer. Software 4(100) (2016) 1–11.

[2] M. Bertsch, R. Dal Passo, M. Mimura, A free boundary problem arising in a simplified tumour growth model of contact inhibition, Interfaces and Free Bound., 12 (2010) 235–250.

[3] M. Bertsch, D. Hilhorst, H. Izuhara, M. Mimura, A nonlinear parabolic-hyperbolic system for contact inhibition of cell-growth, Differ. Equ. Appl. 4(1) (2012) 137-157.

[4] S. N. Busenberg, C. C. Travis, Epidemic models with spatial spread due to population migration, J. Math. Biol. 16 (1983) 181-198.

[5] L. Chen, E. S. Daus, A. Jüngel, Rigorous mean-field limit and cross-diffusion, Z. Angew. Math. Phys. 70 (2019) 122 (2019).

[6] L. Chen, A. Jüngel, Analysis of a multidimensional parabolic population model with strong cross-diffusion, SIAM J. Math. Anal. 36 (2004) 301–322.

[7] G. Galiano, M. L. Garzón, A. Jüngel, Semi-discretization in time and numerical convergence of solutions of a nonlinear cross-diffusion population model, Numer. Math. 93 (2003) 655–673.

[8] G. Galiano, V. Selgas, On a cross-diffusion segregation problem arising from a model of interacting particles, Nonlinear Anal. Real World Appl. 18 (2014) 34–49.

[9] G. Galiano, V. Selgas, Deterministic particle method approximation of a contact inhibition cross-diffusion problem, Appl. Numer. Math. 95 (2015) 229–237.

[10] G. Gambino, M. C. Lombardo, M. Sammartino, Turing instability and traveling fronts for a nonlinear reaction–diffusion system with cross-diffusion Original, Math. Comput. Simul. 82 (2012) 1112–1132.

[11] G. Gambino, M. C. Lombardo, M. Sammartino, Pattern formation driven by cross-diffusion in a 2D domain, Nonlinear Anal. Real World Appl. 14 (2013) 1755–1779.

[12] A. Jüngel, The boundedness-by-entropy method for cross-diffusion systems, Nonlinearity 28 (2015) 1963.

[13] N. Shigesada, K. Kawasaki, E. Teramoto, Spatial segregation of interacting species, J. Theoret. Biol. 79 (1979) 83–99.

[14] A. M. Turing, The chemical basis of morphogenesis, Philosophical Trans. Royal Soc. London. Series B, Biol. Sciences, 237 (1952) 37–72.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 192–196)

Weakly nonlinear analysis of a system with nonlocal diffusion

Gonzalo Galiano[1], **Julián Velasco**[1]
Universidad de Oviedo, Spain

Abstract

We study, through a weakly nonlinear analysis, the pattern formation for a system of partial differential equations of the Shigesada-Kawasaki-Teramoto type with nonlocal diffusion in the one-space dimensional case with periodic boundary conditions. We obtain the pattern of the solutions for values of the bifurcation parameter in the proximity of the onset of instabilities. Finally, we compare the results of the nonlocal model with those of the usual local diffusion model.

1. Introduction

Let $T > 0$. We consider the following nonlocal diffusion problem. Find $u_i : [0,T]\times\mathbb{R} \to \mathbb{R}_+$, for $i = 1, 2$, such that

$$\partial_t u_i(t,x) = \int_{\mathbb{R}} J(x-y)\big(p_i(\mathbf{u}(t,y)) - p_i(\mathbf{u}(t,x))\big)dy + \gamma f_i(\mathbf{u}(t,x)), \tag{1.1}$$

$$u_i(0,x) = u_{0i}(x), \tag{1.2}$$

for $(t,x) \in Q_T = [0,T]\times\mathbb{R}$, and for some $u_{0i} : \mathbb{R} \to \mathbb{R}_+$, periodic functions of period L. The diffusion kernel, $J : \mathbb{R} \to \mathbb{R}_+$, is an even function with compact support included in $(-L/2, L/2)$. We assume $J \in L^\infty(\mathbb{R}) \cap BV(\mathbb{R})$, and $u_{0i} \in L^\infty(0,L) \cap BV(0,L)$. Here, $\mathbb{R}_+ = [0,\infty)$, $\mathbf{u} = (u_1, u_2)$, and, for $i, j = 1, 2$, $i \neq j$, the diffusion and reaction functions are given by

$$p_i(\mathbf{u}) = u_i(c_i + a_i u_i + u_j), \quad f_i(\mathbf{u}) = u_i\big(\alpha_i - (\beta_{i1}u_1 + \beta_{i2}u_2)\big), \tag{1.3}$$

for some non-negative constant coefficients $\gamma,\ c_i,\ a_i,\ \alpha_i,\ \beta_{ij}$.

Problem (1.1)-(2.1) is a nonlocal version of the clasical Shigesada-Kawasaki-Teramoto (SKT) population model, see [4]. For the relationship between local and nonlocal diffusion models, see the monograph by Andreu et al. [1]

2. Existence and uniqueness of solution

Theorem 2.1 *Assume the above conditions and $a_i + \beta_{ii} > 0$ for $i = 1, 2$. Then, there exists a unique strong solution (u_1, u_2) of problem (1.1)-(1.2) with $u_i \geq 0$ a. e. in Q_T and such that, for $i = 1, 2$ and $t \in [0,T]$,*

$$u_i \in W^{1,\infty}(0,T; L^\infty(\mathbb{R})) \cap C([0,T]; L^\infty(\mathbb{R})) \cap BV_{loc}(\mathbb{R}))$$

Proof If u_i are periodic funtions, then the right hand side of equation (1.1) is a periodic funtion with period L and, for fixed x, the integrand is nonzero outside the interval $(x - L/2, x + L/2)$. Defining a periodic extension of J of period L, that we will denote by J_p, the integral in (1.1) may be computed from the corresponding integral in any interval of legth L. Thus, we reformulate the problem (1.1)-(2.1) as defined in the bounded domain $[0, L]$:

$$\partial_t u_i(t,x) = \int_0^L J_p(x-y)\big(p_i(\mathbf{u}(t,y)) - p_i(\mathbf{u}(t,x))\big)dy + \gamma f_i(\mathbf{u}(t,x)), \tag{2.1}$$

$$u_i(0,x) = u_{0i}(x), \tag{2.2}$$

for $(t,x) \in [0,T]\times[0,L]$. Theorem 1 in [2] implies the existence of a unique solution to this problem. The periodic extension of this solution to $\mathbb{R}$ is a solution to our original problem (1.1)−(1.2). The uniqueness follows from the uniqueness result in [2]. □

In the following, we will use the notation J instead of J_p.

3. Linear stability and bifurcation parameter

Under suitable assumptions on the coefficients α_i and β_{ij}, the coexistence equilibrium

$$\tilde{\mathbf{u}} = \Big(\frac{\beta_{22}\alpha_1 - \beta_{12}\alpha_2}{\beta_{11}\beta_{22} - \beta_{12}\beta_{21}}, \frac{\beta_{11}\alpha_2 - \beta_{21}\alpha_1}{\beta_{11}\beta_{22} - \beta_{12}\beta_{21}}\Big) \tag{3.1}$$

is a constant stationary solution of equation (1.1). In what follows, and for simplicity, we will consider $L = 2\pi$. We also introduce the notation $\Delta_{nl}u = J * u - u$.

The linearization of problem (1.1)-(1.2) around $\tilde{\mathbf{u}}$ suggests to look for a solution of the form: $\mathbf{u} = \boldsymbol{\rho} e^{\lambda t}\cos(nx - \Phi_n)$, with $\boldsymbol{\rho} \in \mathbb{R}^2$. The resulting matrix eigenvalue problem is

$$A_n\boldsymbol{\rho} = \lambda\boldsymbol{\rho}, \quad \text{where } A_n = \gamma K - \kappa(n)D,$$

with $\kappa(n) = \int_{-\pi}^{\pi} J(x)\big(1 - \cos(nx)\big)dx \geq 0$, and

$$K := D\mathbf{f}(\tilde{\mathbf{u}}) = \begin{pmatrix} -\beta_{11}\tilde{u}_1 & -\beta_{12}\tilde{u}_1 \\ -\beta_{21}\tilde{u}_2 & -\beta_{22}\tilde{u}_2 \end{pmatrix}, \qquad D = \begin{pmatrix} d_1 + 2a_{11}\tilde{u}_1 + a_{12}\tilde{u}_2 & a_{12}\tilde{u}_1 \\ a_{21}\tilde{u}_2 & d_2 + a_{21}\tilde{u}_1 + 2a_{22}\tilde{u}_2 \end{pmatrix}. \tag{3.2}$$

We are going to assume that the uniform problem in space is stable, so K has two real eigenvalues and his determinant is positive.

Matrix A_n has two different real eigenvalues for any n and one of the eigenvalues is negative.

We want to find a bifurcation parameter such that the problem is stable on one side of the threshold and unestable on the other side, that is, the matrix has zero as eigenvalue for certain value n for the threshold value.

The weakly nonlinear analysis follows the local case in [3].

In the study of the linear stability of the local problem in a bounded domain we are concerned in the arrays $A_k = \gamma K - k^2 D$ with k a natural number.

We want to obtain a condition than implies that the expression $det(A_k)$ attains a minimum value 0 for some value $k > 0$. We oberserve that $Det(A_k)$ is a polynomial of degree 2 in the variable k^2:

$$h(k^2) = det(\gamma K - k^2 D) = \gamma^2 det(K) + \gamma q k^2 + k^4 det(D)$$

where

$$q = \beta_{11}\tilde{u}_1(2a_{22}\tilde{u}_2 + d_2) + \beta_{22}\tilde{u}_2(2a_{11}\tilde{u}_1 + d_1) + a_{12}\tilde{u}_2(\beta_{22}\tilde{u}_2 - \beta_{21}\tilde{u}_1) + a_{21}\tilde{u}_1(\beta_{11}\tilde{u}_1 - \beta_{12}\tilde{u}_2)$$

If we want the parabolyc expression to have a positive root, we need $q < 0$.

The only possible negative terms in q are $\beta_{22}\tilde{u}_2 - \beta_{21}\tilde{u}_1$ and $\beta_{11}\tilde{u}_1 - \beta_{12}\tilde{u}_2$.

It can be proved that these two terms have different sign. In this work we will assume $\beta_{22}\tilde{u}_2 - \beta_{21}\tilde{u}_1 < 0$ and then, we will consider as bifurcation parameter $b := a_{12}$. We will use the superscript b to highlight the dependence on b of the quantities.

In [3] it is proved that there exists a thershold value b^c and a unique value $k_c > 0$ such that $h^{b^c}(k_c^2) = 0$, and for any $b > b^c$ there exist intervals (x_1, x_2) such that $h^b(k) < 0$ for $k^2 \in (k_1^2, k_2^2)$. Furthermore the size of the interval increases in γ

In our nonlocal problem in a bounded domain, instead of k^2 we have to consider $\kappa(n)$ for $n \in \mathbb{N}$. So, for γ big enough, there exist a natural number n and a threshold $b_n^* > b^c$ such that $det(\gamma K - \kappa(n)D^{b_n^*}) = 0$.

In the weakly nonlinear approximation we must consider b_n^* as the threshold value of the bifurcation parameter instead of b^c. However, we will use, for simplicity, b^c in the equations in the following section.

Remark 3.1 We oberve that the temporal evolution of the amplitudes associated to a same wavelength (the cos(nx) term and the sin(nx) term) satisfafy the same equation, so we can consider Φ_n is independent in time.

4. Weakly nonlinear analysis

We follow [3].

Let b^c be the bifurcation threshold. We will consider the expansions

$$b = b^c + \varepsilon b_1 + \varepsilon^2 b_2 + \varepsilon^3 b_3 + O(\varepsilon^4)$$
$$\mathbf{w} = \varepsilon\mathbf{w}_1 + \varepsilon^2\mathbf{w}_2 + \varepsilon^3\mathbf{w}_3 + O(\varepsilon^4)$$
$$\partial_t = \varepsilon\partial_{T_1} + \varepsilon^2\partial_{T_2} + \varepsilon^3\partial_{T_3} + O(\varepsilon^4)$$

then

$$\partial_t \mathbf{w} = \varepsilon^2 \partial_{T_1} \mathbf{w}_1 + \varepsilon^3 (\partial_{T_1} \mathbf{w}_2 + \partial_{T_2} \mathbf{w}_1) + O(\varepsilon^4)$$

We will use the following notation for the nonlocal diffusion terms:

$$\mathcal{L}^b = \mathcal{L}^{b^c} + \sum_{j=1}^{3} \varepsilon^j \begin{pmatrix} b_j \tilde{u}_2 & b_j \tilde{u}_1 \\ 0 & 0 \end{pmatrix} \Delta_{nl}$$

where

$$\mathcal{L}^{b^c} = \gamma K + D^{b^c} \Delta_{nl}$$

Substituting the above expansion into our nonlinear problem and collecting at each order in ε, we obtain the following succesion of linear problems:

Order ε :

The linear system $\mathcal{L}^{b^c} \mathbf{w}_1 = 0$ has solutions:

$$\mathbf{w}_1 = A(T_1, T_2) \rho cos(k_c x - \Phi)$$

with $\rho \in ker\big(\gamma K - \kappa(k_c) D^{b^c}\big)$.

We can select

$$\boldsymbol{\rho} = (1, M)^t, \quad \text{with } M = \frac{-\gamma K_{21} + D_{21}^{b^c} \kappa(k_c)}{\gamma K_{22} - D_{22}^{b^c} \kappa(k_c)} < 0$$

Order ε^2 :

We have the system:

$$\mathcal{L}^{b^c} \mathbf{w}_2 = \partial_{T_1} \mathbf{w}_1 - \frac{1}{2}\big(Q_K(\mathbf{w}_1, \mathbf{w}_1) + \Delta_{nl} Q_D^{b^c}(\mathbf{w}_1, \mathbf{w}_1)\big) - b_1 \begin{pmatrix} \tilde{u}_2 & \tilde{u}_1 \\ 0 & 0 \end{pmatrix} \Delta_{nl} \mathbf{w}_1 =: \mathbf{F}$$

where

$$\partial_{T_1} \mathbf{w}_1 = \partial_{T_1} A(T_1, T_2) \rho cos(k_c x - \Phi)$$

$$Q_K(\mathbf{w}_1, \mathbf{w}_1) = A(T_1, T_2)^2 Q_K(\rho, \rho) cos^2(k_c x - \Phi) = \frac{1}{2} A(T_1, T_2)^2 Q_K(\rho, \rho)\big(1 + cos(2k_c x - 2\Phi)$$

$$Q_D^{b^c}(\mathbf{w}_1, \mathbf{w}_1) = A(T_1, T_2)^2 Q_D^{b^c}(\rho, \rho) cos^2(k_c x - \Phi)$$

since $\Delta_{nl} cos(j k_c x - j\Phi) = -\kappa(j k_c) cos(j k_c x - j\Phi)$, we have

$$\Delta_{nl} Q_D^{b^c}(\mathbf{w}_1, \mathbf{w}_1) = -\frac{1}{2} \kappa(2k_c) A(T_1, T_2)^2 Q_D^{b^c}(\rho, \rho) cos(2k_c x - 2\Phi)$$

$$\Delta_{nl} \mathbf{w}_1 = -\kappa(k_c) A(T_1, T_2) \rho cos(k_c x - \Phi)$$

So, we have the following source term in the system:

$$\mathbf{F} = -\frac{1}{4} A(T_1, T_2)^2 \sum_{j=0,2} \mathcal{M}_j(\rho, \rho) cos(j k_c x) + \big(\partial_{T_1} A \rho + b_1 \kappa(k_c) A(\tilde{u}_2 + \tilde{u}_1 M, 0)^t\big) cos(k_c x - \Phi)$$

with

$$\mathcal{M}_j = Q_K - \kappa(j k_c) Q_D^{b^c}$$

The functions $cos(k_c x - \Phi)$ and $cos(2k_c x - 2\Phi)$ are linearly independent, so the solvability Fredholm condition is satisfied if the source term has no component in $cos(k_c x - \Phi)$.

We deduce, that we need $T_1 = b_1 = 0$.

If we assume this condition we can solve the linear problems and we obtain:

$$\mathcal{L}^{b^c} \mathbf{w}_2 = A(T_2)^2 \sum_{j=0,2} (\gamma K - \kappa(j k_c) D^{b^c}) \mathbf{w}_{2j} cos(j k_c x - j\Phi)$$

and so, $\mathbf{w}_{2j}$ verify the systems:

$$L_j \mathbf{w}_{2j} = -\frac{1}{4} \mathcal{M}_j(\rho, \rho), \text{ para } j = 0, 2$$

with $L_j = \gamma K - \kappa(jk_c)D^{b^c}$, and finally

$$\mathbf{w}_2 = A(T_2)^2(\mathbf{w}_{20} + \mathbf{w}_{22}cos(2k_cx - 2\Phi))$$

Order ε^3:
In this case, the system reads:

$$\mathcal{L}^{b^c}\mathbf{w}_3 = \partial_{T_2}\mathbf{w}_1 - Q_K(\mathbf{w}_1, \mathbf{w}_2) - \Delta_{nl}Q_D^{b^c}(\mathbf{w}_1, \mathbf{w}_2) - \begin{pmatrix}\tilde{u}_2 & \tilde{u}_1\\ 0 & 0\end{pmatrix} b_2\Delta_{nl}\mathbf{w}_1 =: \mathbf{G}, \tag{4.1}$$

where

$$\partial_{T_2}\mathbf{w}_1 = \partial_{T_2}A(T_2)\rho cos(k_cx - \Phi)$$

$$Q_K(\mathbf{w}_1, \mathbf{w}_2) = A(T_2)^3Q_K(\rho, \mathbf{w}_{20})cos(k_cx - \Phi) + A(T_2)^3Q_K(\rho, \mathbf{w}_{22})cos(k_cx - \Phi)cos(2k_cx - 2\Phi)$$

$$= A(T_2)^3\Big(\big(Q_K(\rho, \mathbf{w}_{20}) + \frac{1}{2}Q_K(\rho, \mathbf{w}_{22})\big)cos(k_cx - \Phi) + \frac{1}{2}Q_K(\rho, \mathbf{w}_{22})cos(3k_cx - 3\Phi)\Big)$$

and

$$\Delta_{nl}Q_D^{b^c}(\mathbf{w}_1, \mathbf{w}_2) = -A(T_2)^3\Big(\big(Q_D^{b^c}(\rho, \mathbf{w}_{20}) + \frac{1}{2}Q_D^{b^c}(\rho, \mathbf{w}_{22})\big)\kappa(k_c)cos(k_cx - \Phi)$$

$$+\frac{1}{2}Q_D^{b^c}(\rho, \mathbf{w}_{22})\kappa(3k_c)cos(3k_cx - 3\Phi)\Big)$$

The source term is, in this case

$$\mathbf{G} = \Big(\partial_{T_2}A\rho + A\mathbf{G}_1^{(1)} + A^3\mathbf{G}_1^{(3)}\Big)cos(k_cx - \Phi) + \mathbf{G}_3A^3cos(3k_cx - 3\Phi)$$

with

$$\mathbf{G}_1^{(1)} = (\tilde{u}_2 + \tilde{u}_1M)\kappa(kc)b_2(1, 0)^t$$

$$\mathbf{G}_1^{(3)} = -\big(\mathcal{M}_1(\rho, \mathbf{w}_{20}) + \frac{1}{2}\mathcal{M}_1(\rho, \mathbf{w}_{22})\big)$$

$$\mathbf{G}_3 = -\frac{1}{2}\mathcal{M}_3(\rho, \mathbf{w}_{22})$$

The Fredholm condition is satisfied if $\mathbf{G} \in Ker\big((\mathcal{L}^{b^c})^*\big)^\perp$.

The vectorial space $Ker((\mathcal{L}^{b^c})^*)$ has dimension 1.The elements of this space are multples of $\Psi = \boldsymbol{\eta}cos(k_cx - \alpha)$ with $\alpha \in \mathbb{R}$, $\boldsymbol{\eta} = (1, M^*) \in Ker\big(\gamma K^t - \kappa(k_c)(D^{b^c})^t\big)$.

If we define

$$\hat{\sigma} = -\frac{\mathbf{G}_1^{(1)} \cdot \boldsymbol{\eta}}{\boldsymbol{\rho} \cdot \boldsymbol{\eta}} > 0, \qquad \hat{L} = \frac{\mathbf{G}_1^{(3)} \cdot \boldsymbol{\eta}}{\boldsymbol{\rho} \cdot \boldsymbol{\eta}},$$

We obtain the Stuart-Landau equation in the supercritical case $\hat{L} > 0$:

$$\partial_{T_2}A = \hat{\sigma}A - \hat{L}A^3$$

and the weakly nonlinear approximation of the limit in time of the solution:

$$\mathbf{w} = \varepsilon\boldsymbol{\rho}\sqrt{\frac{\hat{\sigma}}{\hat{L}}}cos(k_cx - \Phi) + \varepsilon^2\frac{\hat{\sigma}}{\hat{L}}(\mathbf{w}_{20} + \mathbf{w}_{22}cos(2k_cx - 2\Phi)) + O(\varepsilon^3).$$

5. Numerical simulations.

We consider a finite difference discretization in space and an explicit discretization in time for problem (1.1)-(1.2). We must use a discrete space length h small enough to have an adequate number or discrete points in the domain of the nonlocal kernel. In Fig. 1, we show the the numerical approximation of the asymptotic solution in time (for a set of parameters) for the nonlinear problem (1.1)-(1.2) and the weakly nonlinear approximation. We took $h = \frac{2\pi}{100}$ and a triangular kernel with support in $[-\frac{\pi}{10}, \frac{\pi}{10}]$ and variance 2.

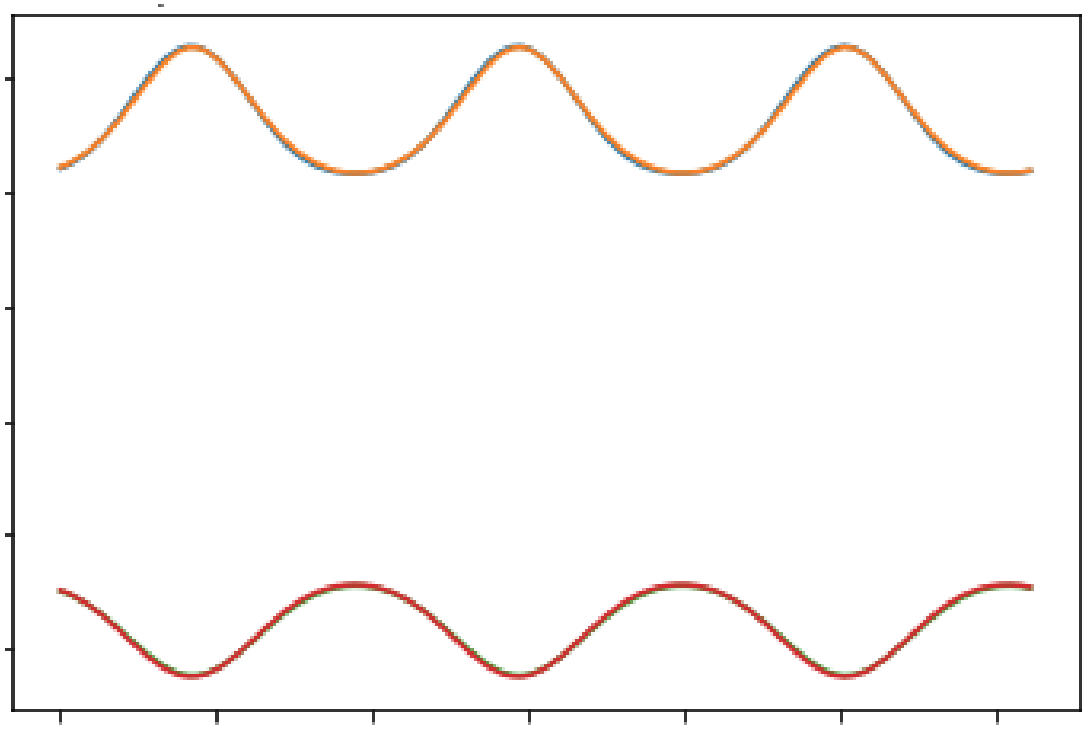


Fig. 1 Nonlocal solution and the weakly nonlinear approximation. The parameters are $d_1 = d_2 = 0.1$, $a_{11} = 1.e-4$, $a_{21} = 0.3$, $a_{22} = 0.1$, $\alpha_1 = 1.2$, $\alpha_2 = 1.0$, $\beta_{11} = 0.5$, $\beta_{12} = 0.4$, $\beta_{21} = 0.38$, $\beta_{22} = 0.41$, $\gamma = 49.75$. With these data, we have $b^c = 5.297$, $n = 3$, $b_3^* = 5.328$. Thus we took $b := a_{12} = 5.5$

Remark 5.1 (Rescaling of the nonlocal kernel) Let u be a $2\pi-$periodic function twice continuously differentiable in $[0, 2\pi]$, and assume that the support of J is contained in [-1,1]. Then, we have:

$$\lim_{\varepsilon\to 0} \frac{c_1}{\varepsilon^3} \int_{x_0-\pi}^{x_0+\pi} J\Big(\frac{y-x}{\varepsilon}\Big)(u(y)-u(x))dy = u_{xx}(x_0)$$

where $c_1^{-1} = \frac{1}{2}\int_{-1}^{1} J(x)x^2dx$. We consider the rescaling $J_\varepsilon(x) = \frac{c_1}{\varepsilon^3}J(\frac{x}{\varepsilon})$, so that, for fixed n,

$$\kappa_\varepsilon(n) = \int_{-1}^{1} J_\varepsilon(x)\big(1-\cos(nx)\big)dx = \frac{c_1}{\varepsilon^3}\int_{-\varepsilon}^{\varepsilon} J\Big(\frac{x}{\varepsilon}\Big)\big(1-\cos(nx)\big)dx = c_1\int_{-1}^{1} J(t)\frac{1-\cos(n\varepsilon t)}{\varepsilon^2}dt \to n^2$$

as $\varepsilon \to 0$. Thus, the weakly nonlinear approximation of the rescaled nonlocal diffusion problem converges, when $\varepsilon \to 0$, to the weakly nonlinear approximation of the corresponding local diffusion problem.

Acknowledgements

The authors were supported by the Spanish MCI Project MTM2017-87162-P.

References

[1] F. Andreu-Vaillo, J.M. Mazón, J.D. Rossi, J.J. Toledo-Melero. Nonlocal Diffusion Problems. *Mathematical Surveys and Monographs*, 165, American Mathematical Society, 2010.

[2] G. Galiano, J. Velasco. Well-Posedness of a Cross-Diffusion Population Model with Nonlocal Diffusion. *SIAM J. Math. Anal.*, 51(4), 2884–2902, 2019

[3] G. Gambino, M.C. Lombardo and M. Sammartino. Turing instability and traveling fronts for a nonlinear reaction–diffusion system with cross-diffusion. *Mathematics and Computer in Simulation*, 82(6), 1112-1132, 2012.

[4] N. Shigesada, K. Kawasaki, E. Teramoto, Spatial segregation of interacting species, J. Theoret. Biol. 79 (1979) 83–99.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 197–200)

What is the humanitarian aid required after tsunami?

José M. González-Vida[1], Sergio Ortega[1], Jorge Macías[2], Manuel J. Castro[2], Alberto Michelini[3], Adriano Azzarone[3]

1. jgv@uma.es, sergio.ortega@uma.es. *Dpto. Matemática Aplicada. Universidad de Málaga, Spain*
2. jmacias@uma.es, mjcastro@uma.es *Dpto. Anál. Matemático, Estadística e I.O y Mat. Aplicada. Universidad de Málaga, Spain*
3. *Istituto Nazionale di Geofisica e Vulcanologia, Roma, Italy*

Abstract

In this work it is illustrated how the urgent computing (UC) capabilities in the tsunami natural hazard framework are strengthening the monitoring and analysis functions of the European Emergency Response Coordination Centre (ERCC) and its Situational Awareness Sector (SAS) by helping to design the multi-hazard advice service at global level and on a 24/7 operational basis. In this context, the ARISTOTLE-eENHSP project (All Risk Integrated System TOwards Trans-boundary hoListic Early-warning - enhanced European Natural Hazards Scientific Partnership) has been designed to offer a flexible and scalable system that can provide new hazard-related services to the ERCC.

1. Introduction

"When a disaster strikes, every minute counts. An immediate, coordinated and pre-planned response saves lives". The Emergency Response Coordination Centre (ERCC) has been established exactly for this reason: to enable the EU and its Member States to respond to overwhelming natural and man-made disasters in a timely and efficient manner. This is one of the key messages presented by the ERCC in the European Civil Protection and Humanitarian Aid Operations (ECHO) factsheet and it is on the basis goal of the ARISTOTLE-eENHSP project (see http://aristotle.ingv.it).

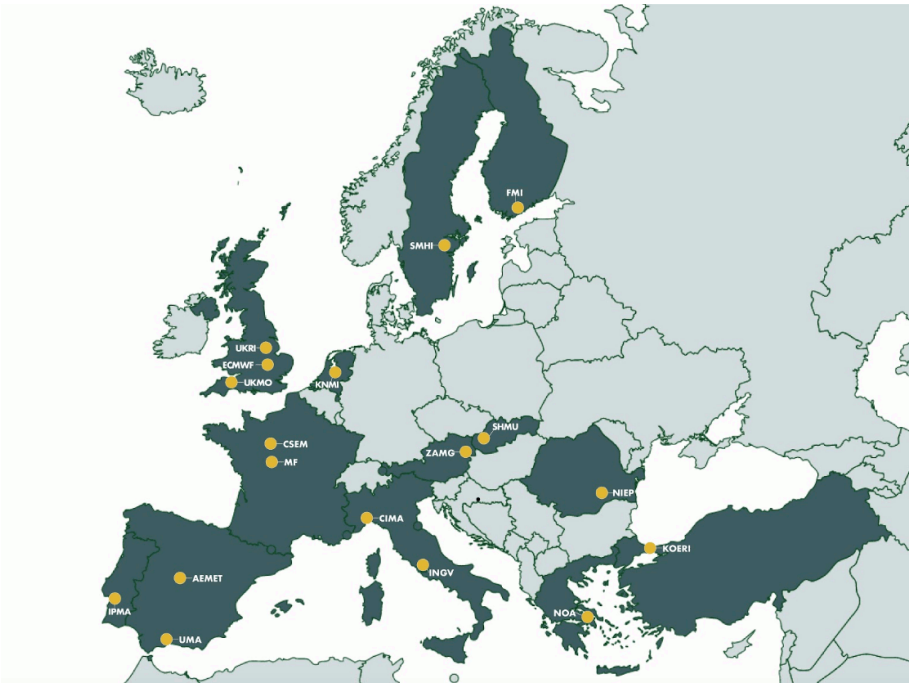


Fig. 1 Geographical distribution of the ARISTOTLE-eENHSP Consortium.

ARISTOTLE-eENHSP has been designed to offer a flexible and scalable system that can provide new hazard-related services to the ERCC. The ARISTOTLE consortium includes 18 partner institutions operating in the Meteorological and Geophysical domains (see Fig. 1). It builds on a consolidated and multi-disciplinary partnership consisting of world-leading scientific centres in the areas of Earth and Climate sciences (see Fig. 2), providing operational and monitoring services, early warning and information systems as well as contributing to innovation

and research actions. The ARISTOTLE-eENHSP Consortium is currently providing advice in a multi-hazard fashion for the following inter-related hazards:

- Earthquakes,
- Tsunamis,
- Volcanic eruptions,
- Severe Weather events,
- Flooding events and
- Wildfires/Forest fires.

The operational capacity of the ARISTOTLE-eENHSP service relies on a blueprint in which experts from each of the different hazard groups provide the collective expertise into a multi-hazard virtual 24/7 operational centre, named "Multi-Hazard Board" (MHB). The MHB, composed of a representative of each hazard group, brings together the best information from the multi-hazard perspective in a single and unified multi-hazard scientific assessment of the ongoing and future events provided to the ERCC and SAS in two different modes, the *Emergency* and the *Routine* modes. Both modes facilitate the accommodation to the different temporal scales of both, the hazards considered, as well as the relative preparatory actions or feedback required for effective emergency response.

The EDANYA Group of the University of Málaga participates in this consortium since 2018 providing services of Faster than Real Time (FTRT) tsunami computations with the numerical model Tsunami-HySEA.

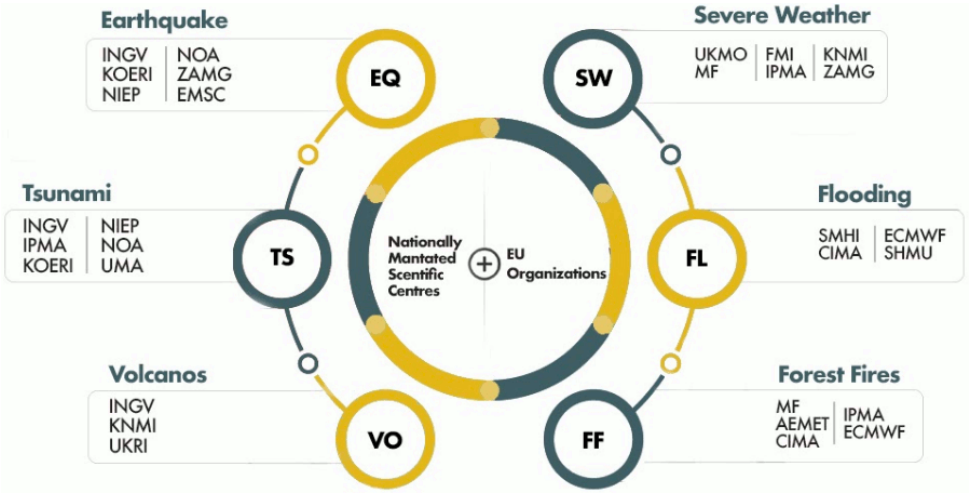


Fig. 2 The experienced multi-hazard scientific ARISTOTLE-eENHSP Partnership and related hazards.

2. The tsunami service

Tsunami-HySEA (see [1, 2]) is a finite-volume numerical model that solves the 2D non-linear shallow water equations in spherical coordinates. It has been developed by the EDANYA group of the University of Málaga specifically for simulations of seismically induced tsunamis. Implemented for graphic processing unit (GPU) architecture, Tsunami-HySEA is a robust, reliable and accurate model capable of simulating the generation, propagation and inundation of a tsunami in a region covered by a grid with several million cells in only a few minutes. This model has been extensively validated under the standard benchmarks proposed by the National Tsunami Hazards Mitigation Program (NTHMP) of the U.S.A. (see [3,4]) and has been extensively tested in several scenarios and compared with other well-established tsunami models (see [5,6]).

Tsunami-HySEA has been implemented using CUDA and MPI in order to take advantage of the massive parallel architecture of multi-GPU clusters, so that the computing time required could be dramatically reduced with respect to the use of a single CPU core or even a multi-core processor and, at the same time, numerical resolution could be increased still computing extremely fast. Many features are included in Tsunami-HySEA, such as the possibility of using nested meshes, direct output of time series, the computation of the initial seafloor deformation using the Okada (1992) model, support for rectangular or triangular faults, etc. By means of a very efficient implementation, the model is able to simulate 8 hours of real time tsunami in the Mediterranean Sea (in a mesh with 10 million volumes and a resolution of 30 arc-sec) in 257 seconds using two NVidia Tesla P100, or even in 284 seconds with one NVidia Tesla V100.

The ARISTOTLE tsunami service (TS) is integrated in the SPADA (Scientific Products Archiving and Document Assembly) IT platform that gathers the scientific, exposure and preliminary impact informations which are used by the multi-hazard operational board to assembly the reports. This platform relies on existing and newly developed web services. The TS workflow (see Fig. 3) consists on four steps: the system is triggered by an end-user who is on duty in the service. Using the earthquake parameters that can be provided by different seismic monitoring sources, the scenario parameterization is defined and it is sent to the supercomputation resources (in this case located at INGV and the University of Málaga). In this step a message passing system (RabbitMQ) is used between the SPADA system and the supercomputing services. Depending on the earthquake epicenter location the system is able to automatically select an optimal computational grid size and refinement level depending on the seismological parameters. For example, as it will be presented in the next section, if the epicenter is located in the Mediterranean Sea, the system automatically performs 8 hours of wallclock simulation in a 2 arc-min resolution grid, then detects the limits where the tsunami waves has arrived and later performs a second simulation in a new domain with more resolution (30 arc-sec). Depending on the event magnitude, the computation time can last from a few seconds to the order of minutes. The current outputs of the tsunami service system are: maximum wave height in the considered domain, wave arrival time and maximum wave height along coast locations.

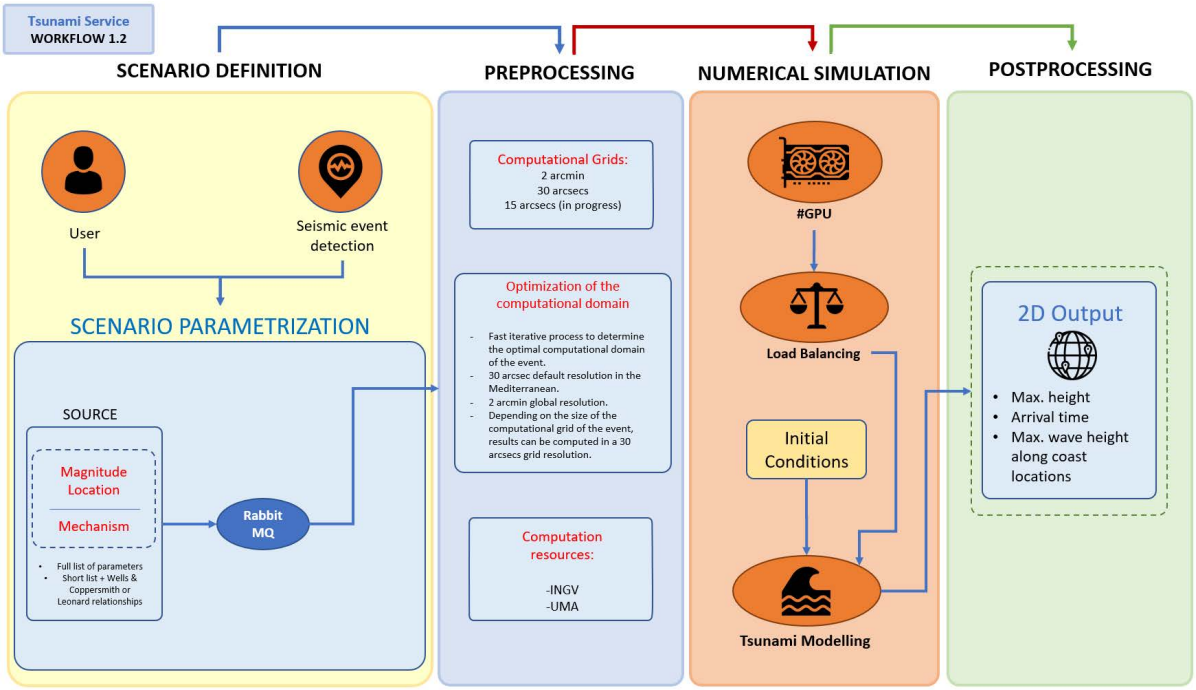


Fig. 3 Scheme of the tsunami service workflow implementation in the ARISTOTLE.

This workflow is scalable depending on different aspects, like the computation resources or the Digital Terrain Models (DTM) available. As consequence, the numerical computation output could be improved in different ways: for instance by improving the grid resolution (even using nested meshes in specific areas of interest), or even providing not only one scenario output but considering an ensemble of cases that could deliver even a Probabilistic Tsunami Forecast (PTF). We are studying these possibilities in H2020 European projects like ChEESE (Center of Excellence (CoE) in Solid Earth) (UE-H2020. Grant agreement: N^o 823844), or more recently, eFlows4HPC (Enabling dynamic and Intelligent workflows in the future EuroHPC ecosystem) (UE -H2020-JTI-EuroHPC-2019-1. Reference: 955558).

3. Outputs and computation time

The system outputs are delivered to the European Emergency Response Coordination Centre (ERCC) in a multi-hazard report providing expert analysis made by an expert panel in the different involved hazards. In our case, the tsunami service outputs are relevant in the sense that they have to be easily readable by the endusers. These aspects have been agreed with all the consortium components related with the TS. For instance, an enhanced semaphore colorbar has been designed where each semaphore color: green, yellow, orange and red has been subdivided into three subcolors. The output is clear even for endusers with a basic information (see Fig. 4. Left). The arrival times figure output has been also improved with the addition of a jet colormap that completes the information given by the isochrons.

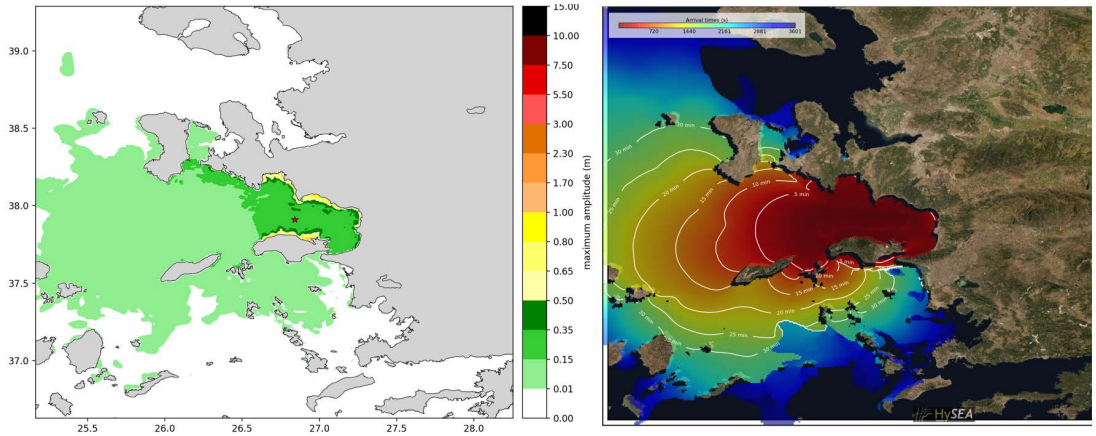


Fig. 4 Izmir (Turkey) 2020 tsunami. Left: max. wave amplitude. Right: arrival times.

To illustrate the computational efficiency of the system, in the next table are shown the computation times required to perform the process described in the previous section. In this case it is simulated one hour of wallclock propagation of the tsunami in the 2 arc-min resolution and then the same time in a 30 arc-min resolution grid adapted to the domain where the tsunami waves have arrived. The total time is 38.48 seconds in a single P100 NVidia GPU.

Scenario	Resolution	Wallclock simul. time	Nr. cells	Comp. Time	Postproc. time	Total time
Izmir	2 arc-min	60 mins (3600 secs)	1305 x 480 = 626400	1.15 secs		
Izmir	30 arc-sec	60 mins (3600 secs)	460 x 336 = 154560	2.22 secs	35.11 secs	38.48 secs

4. Conclusions

The ARISTOTLE-eENHSP project is strengthening the European response capacities to assist Member States in responding to natural disasters when national capacities are overwhelmed. ARISTOTLE has its own operational multi-hazard capacities in order to ensure that the EU can provide better crisis emergency support with maximum efficiency and minimal bureaucracy. The EDANYA group role in this project is related to the Tsunami Service with the develop and tuning of the system. In fact, the 8-9 technology readiness level (TRL) achieved with Tsunami-HySEA in this service makes it operational and at the same time scalable to incorporate the new state of the art techniques when they are available.

Acknowledgements

This work was partially supported by ARISTOTLE-eENHSP. Reference: ECHO/2020/OP/0001. Funding institution: EC-DG ECHO (European Community). Directorate General: European Civil Protection and Humanitarian Aid Operations.

References

[1] Castro, M.J., González-Vida, J.M., Macías, J., Ortega, S. and de la Asunción, M. Tsunami-HySEA: A GPU-based model for Tsunami Early Warning Systems. *Proceedings of the XXIV Congress on Differential Equations and Applications. XIV Congress on Applied Mathematics.* pp. 1-6, 2015.

[2] De la Asunción, M., Castro, M.J., Fernández-Nieto, E. D., Mantas, J.M., S. Ortega Acosta and González-Vida, J. M. Efficient GPU implementation of a two waves TVD WAF method for the two-dimensional one layer shallow water system on structured meshes *Computers & Fluids*, 80:441-452, 2013.

[3] Lynett, P.J., Gately, K., Wilson R., Montoya, L., Arcas, D., Aytore, B., Bai, Y., Bricker, J. D., Castro, M. J., Fai Cheung, K., David, C. G., Dogan, G. G., Escalante, C., González-Vida, J. M., Grilli, S. T., Heitmann, T. W., Horrillo, J., Kânoğlu, U., Kian, R., Kirby, J. T., Li, W., Macías, J., Nicolsky, D. J., Ortega, S., Pampell-Manis, A., Park, Y. S., Roeber, V., Sharghivand, N., Shelby, M., Shi, F., Tehranirad, B., Tolkova, E., Thio, H. K., Velioğlu, D., Yalçıner, A. C., Yamazaki, Y., Zaytsev, A., Zhang, Y. J. Inter-model analysis of tsunami-induced coastal currents. *Ocean Modeling*, 114: 14-32, 2017.

[4] Macías, J., Castro, M.J., Ortega, S., Escalante, C., González-Vida, J.M. Performance benchmarking of Tsunami-HySEA model for NTHMP's inundation mapping activities. *Pure and Applied Geophysics*, 1-37, 2017.

[5] Macías, J., Mercado, A., González-Vida, J.M., Ortega, S., and Castro, M.J. Comparison and numerical performance of Tsunami-HySEA and MOST models for LANTEX 2013 scenario. Impact assessment on Puerto Rico coasts. *Pure and Applied Geophysics*, 173(12), 3973–3997, 2016.

[6] Molinari, I., Tonini, R., Piatanessi, A., Lorito, S., Romano, F., Melini, D., González-Vida, J.M., Macías, J., Castro, M.J., and de la Asunción, M. Fast evaluation of tsunami scenarios: uncertainty assessment for a Mediterranean Sea database. *Natural Hazards and Earth Systems Sciences*, 16:2593-2602, 2016.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 201–207)

On Keller-Segel systems with fractional diffusion

Rafael Granero-Belinchón[1]
Departamento de Matemáticas, Estadística y Computación, Universidad de Cantabria, Santander, Spain

Abstract

In this note we review some recent results on the parabolic-elliptic and parabolic-parabolic Keller-Segel model with fractional diffusion. Furthermore, we consider with particular emphasis the parabolic-hyperbolic case with a fractional Laplacian. In particular we prove certain new results. Namely, first, we prove the existence of solution for analytic initial data and then we establish the finite time singularity formation for initial data in this class. Finally, we obtain a new equation describing unidirectional wave propagation for the parabolic-hyperbolic Keller-Segel system.

1. The parabolic-parabolic and parabolic-elliptic Keller-Segel systems

Even if as of 2021 no “standard model" of the origin of life has yet emerged, most currently accepted models state that life arose on Earth between 3800 and 4100 million years ago. These first forms of life where single-celled organisms. For most of the History of life on Earth there were only unicellular organisms. However, now there are many different fungi, algae, plants and animals that are multicellular organisms (they are formed by aggregations of cells working together). Thus, even if we know that these single-celled organism eventually formed multicellular organisms (around 1500 millions of years ago), the origins of multicellularity are one of the most interesting topics in biology because we still do not know the details of how multicellularity arised.

A particular situation where cells form a cluster, in a process known as *cell aggregation*, arises when the motion of the cells is driven by a chemical gradient, *i.e.* the cells attempt to move towards higher (or lower) concentration of some chemical substance. This process is usually called *chemotaxis*. Then, multicellular aggregates and eventually tissue-like assemblies are formed when individual cells attach to each other as a consequence of the chemotactic movement and when this aggregation leads to subsequent cellular differentiation. That is, for instance, the case of the slime mold *Dictyostelium Discoideum* and bacterial populations, such as of *Escherichia coli* and *Salmonella typhimurium*.

A preliminary step towards a better understanding of chemotaxis and cell aggregation, was given by Keller & Segel with their 1970 classical paper [18] (see also the prior work by Patlak [21]). In this paper, Keller and Segel proposed a nonlinear system of PDEs of cross-diffusive type. After a number of simplifications, the following PDE system appears as a model of cell aggregation as a consequence of chemotactic movement:

$$\partial_t u = -(-\Delta)^{\alpha/2} u + \nabla \cdot (u \nabla v) + f(u, v) \tag{1.1}$$

$$\tau \partial_t v = \nu \Delta v + u - \lambda v. \tag{1.2}$$

Here $(-\Delta)^{\alpha/2}$ denotes the fractional laplacian, u and v denote the cells and chemical concentration, respectively, and $\tau, \nu, \lambda \geq 0$ are fixed constants. In most of the applications, the forcing is typically $f(u, v) = 0$ or $f(u) = ru(1-u)$. System (1.1)-(1.2) is known as the parabolic-parabolic Keller-Segel equation (ppKS).

Diffusions given by $-(-\Delta)^{\alpha/2} u$ for $\alpha < 2$ arise naturally when studying feeding strategies of some organisms in certain situations. For instance, these fractional diffusions have been reported for amoebas [19], microzooplancton [4], flying ants [23], fruit flies [12], jackals [2] and even humans! [22].

Following the works by T. Hillen, K. Painter & M. Winkler [17] and references therein, equations (1.1)-(1.2) also appear in the modelization of cancer invasion of healthy tissues. In particular, the model (1.1)-(1.2) with $f(u, v) = ru(1-u)$ appears as a valid approximation of the three-component urokinase plasminogen invasion.

The case when $\tau = 0$, (1.1)-(1.2) is known as the parabolic-elliptic Keller-Segel equations (peKS) and, besides its interest in regards to aggregation and chemotaxis, this equation also arises as a model of other physical phenomena. In particular, the (peKS) equation (with $\alpha = 2$, $\nu = 1$, $\lambda = 0$ and $f \equiv 0$) is, formally, similar to the two-dimensional, incompressible Navier-Stokes written in vorticity form $\omega = \text{curl} v$:

$$\partial_t \omega = \Delta \omega + \nabla \cdot (\omega \nabla^{\perp} \psi), \ -\Delta \psi = \omega.$$

The (peKS) equation also arises as a model of gravitational collapse and star formation [1]. Thus, every new mathematical result regarding equations (1.1)-(1.2) has potential implications in Applied Sciences.

Using the Keller-Segel model (1.1)-(1.2), aggregation is mathematically equivalent to a finite time singularity of the type

$$\lim_{t\to T}\max_{x} u(x,t) = \infty. \tag{1.3}$$

Then, the question we want to answer is the following: **Can the (ppKS) or the (peKS) systems develop a finite time singularity of the type of** (1.3)**?**

For the simpler case of the (peKS) equations this question is nowadays well understood. Specifically, it is known that the aggregation of cells is very sensitive to changes in the dimension of the spatial domain and the order of the diffusion, α. In particular, in the one dimensional case, (peKS) equation was first studied by Escudero [13]. He proved the global existence of solution in the case $1 < \alpha \leq 2$. This result was later improved by Bournaveas & Calvez [6], where the authors proved finite time singularities for the case $0 < \alpha < 1$ and the existence of $K > 0$ such that, for the case $\alpha = 1$ and initial data satisfying the smallness restriction $\int u_0(x_1)dx_1 \leq K$, there exists a global smooth solution. Furthermore, based on numerical simulations, Bournaveas & Calvez, reported the existence of finite time singularities in the case $\alpha = 1$ for *large* initial data. This conjecture is in agreement with the two dimensional case and $\alpha = 2$, where there are finite time singularities depending on the mass of the initial data

$$M = \int_{\mathbb{R}^2} u_0(x_1, x_2)dx_1dx_2.$$

More precisely, when $M < 8\pi$, the solutions are globally defined and smooth while if $M > 8\pi$ the solutions develop a finite time singularity (see for instance the paper by A. Blanchet, J. Dolbeault & B. Perthame [5] and the references therein). The singularity formation for the two dimensional case with $\alpha < 2$ has been proved by D. Li, J. Rodrigo & X. Zhang [20]. Systems akin to (peKS) with a nonlinear fractional diffusion have also been studied in [16] and [7].

In a fruitful collaboration with Jan Burczak, we were able to

- **Lack of threshold behavior for the critical** ($\alpha = 1$) **(peKS) equation.** One of our results [8] for the one-dimensional (peKS) equations with $\alpha = 1$ is that smooth solutions can be defined for all later times, *i.e.* every initial data leads to a global solution. Furthermore, the solutions are globally bounded (in fact, they decay to the homogeneous steady state) if the initial mass is small enough (with a smallness condition of order $O(1)$). This was a very surprising result that disproved the previous conjecture by Bournaveas & Calvez [6].

- **Global existence for the supercritical** ($\alpha < d$) **(peKS) equation with logistic forcing.** In the case where a logistic forcing of the type

  $$f(u) = ru(1-u),\ 0 < r$$

  is considered into the (peKS) equations, we proved that, if $d = 1, 2$ denotes the spatial dimension, we could find a range

  $$d - c_1(r) < \alpha \leq d, \tag{1.4}$$

  with $c_1(r)$ explicit such that there exist global in time classical solutions to the SP equations with α in the range (1.4) [9, 11]. Furthermore, we could also find a second range,

  $$d - c_2(r) < \alpha \leq d, \tag{1.5}$$

  where $c_2(r)$ is explicit, such that there exist global in time weak solutions to the SP equations with α in the range (1.5). Note that, in d dimensions, the case $\alpha = d$ corresponds to the critical case with respect to the total mass

  $$\int u(x)dx.$$

  Consequently, these results are the first global well-posedness results in supercritical ranges (1.4) and (1.5) for a Keller-Segel type equation. As a consequence of these results, we obtained that aggregation is not possible for α in (1.4) if there is a logistic growth. In the one dimensional case $d = 1$, these results are also interesting when compared to other classical nonlinear, one-dimensional drift-diffusion equations as the Burgers equation, where $\alpha \geq 1$ is a required for global classical solution (with large initial data) to exist.

  In the case of two spatial dimensions $d = 2$, the proof of the global existence of classical solutions is based in a new delicate pointwise estimate for the two-dimensional fractional Laplacian

  $$(-\Delta)^{\alpha/2}u(\bar{x}_1, \bar{x}_2) \geq C_{\alpha,\delta}\frac{u(\bar{x}_1, \bar{x}_2)^{1+\frac{\alpha}{2-\delta}}}{\|\phi\|^{1+\frac{\alpha}{2-\delta}}_{C^\delta}},$$

where $\partial_{x_1}\partial_{x_2}\phi = u$ and $(\bar{x}_1, \bar{x}_2)$ is such that $\max_{y_1,y_2} u(y_1, y_2) = u(\bar{x}_1, \bar{x}_2)$. This inequality is interesting by itself as it may be applied in many other PDE problems.

- **Dynamical properties of (ppKS) equation with logistic forcing.** In [10] we study dynamical properties of the (ppKS) system with logistic forcing. Remarkably, this model exhibits a spatio-temporal chaotic behavior, where a number of peaks emerge. We were able to prove the existence of an attractor and provide an upper bound on the number of peaks that the solution may develop. Finally, we perform a numerical analysis suggesting that there is a finite time blowup if the diffusion is weak enough, even in presence of a damping logistic term.

2. The parabolic-hyperbolic Keller-Segel system

In what follows we study the following system arising in tumor angiogenesis

$$\partial_t u = -(-\Delta)^{\alpha/2} u + \partial_x(uq), \text{ for } x \in \mathbb{T},\ t \geq 0, \tag{2.1}$$

$$\partial_t q = u^{r-1}\partial_x u, \text{ for } x \in \mathbb{T},\ t \geq 0, \tag{2.2}$$

where u is a non-negative scalar function, q is a zero-mean function, $\mathbb{T}$ denotes the domain $[-\pi, \pi]$ with periodic boundary conditions, $0 < \alpha \leq 2$, $1 \leq r \leq 2$ and $(-\Delta)^{\alpha/2} = \Lambda^\alpha$ is the fractional Laplacian.

This system was proposed by Othmers & Stevens [24] based on biological considerations as a model of tumor angiogenesis. In that context, u is the density of vascular endothelial cells and $q = \partial_x \log(v)$ where v is the concentration of the signal protein known as vascular endothelial growth factor (VEGF).

Equation (2.1) appears as a singular limit of the following Keller-Segel model of aggregation of the slime mold *Dictyostelium discoideum* [18] (see also Patlak [21])

$$\begin{cases} \partial_t u = -(-\Delta)^{\alpha/2} u - \chi \nabla \cdot (u \nabla G(v)), \\ \partial_t v = \nu \Delta v + \left(\dfrac{u^r}{r} + \lambda \right) v, \end{cases} \tag{2.3}$$

when $G(v) = \log(v)$ and the diffusion of the chemical v is negligible.

This system was studied in [14, 15]. In particular

- **Local existence and decay.** In [15] the local well-posedness for arbitrary H^3 non-negative initial data, $0 \leq \alpha \leq 2$ and $1 \leq r \leq 2$ was proved. We would like to emphasize that that the sign of the initial data plays the role of a stability condition and helps us to avoid derivative loss. Furthermore, the solution verifies the following global bound: for $r > 1$

$$\frac{\|u(t)\|_{L^r}^r}{r(r-1)} + \frac{\|q(t)\|_{L^2}^2}{2} + \frac{1}{r-1}\int_0^t \int_{\mathbb{T}} (-\Delta)^{\alpha/2} u u^{r-1} dx ds \leq \frac{\|u_0\|_{L^r}^r}{r(r-1)} + \frac{\|q_0\|_{L^2}^2}{2}, \tag{2.4}$$

while, if $r = 1$,

$$\int_{\mathbb{T}} (u\log(u) - u + 1) dx + \frac{\|q(t)\|_{L^2}^2}{2} + \int_0^t \int_{\mathbb{T}} (-\Delta)^{\alpha/2} u \log(u) dx ds \leq \int_{\mathbb{T}} (u_0 \log(u_0) - u_0 + 1) dx + \frac{\|q_0\|_{L^2}^2}{2}. \tag{2.5}$$

- **Global well-posedness for arbitrary initial data in the critical regime for $r = 2$.** Notice that the equations (2.1)-(2.2) verify the following scaling symmetry: for every $\lambda > 0$

$$u_\lambda(x,t) = \lambda^{\frac{2\alpha-2}{r}} u\left(\lambda x, \lambda^\alpha t\right), \quad q_\lambda(x,t) = \lambda^{\alpha-1} q\left(\lambda x, \lambda^\alpha t\right).$$

In [14], the global well-posedness for arbitrary H^2 non-negative initial data and the critical diffusion $\alpha = 3/2$ was proved. Similarly, the two-dimensional case is also studied for the critical value $\alpha = 2$ and global well-posedness is also presented. Due to the hyperbolic character of the equation for q, prior available global existence results of classical solution for equation (2.1) impose several assumptions [25–27] and the references therein.

Our results removed some of the previous conditions. On the one hand, we prove global existence for arbitrary data in the cases $d = 1$ and $\alpha \geq 1.5$ and $d = 2$ and $\alpha = 2$. On the other hand, in the cases where we have to impose size restrictions on the initial data, the Sobolev spaces are bigger than H^2. A question that remains open is the trend to equilibrium. From (2.4) is clear that the solution $(u(t), q(t))$ tends to the homogeneous state, namely $(\langle u_0 \rangle, 0)$. However, the rate of this convergence is not clear.

2.1. Local well-posedness for analytic initial data

Equations (2.1)-(2.2) can be written as

$$\partial_t h = -\Lambda^\alpha h + \partial_x(hq) + \langle u_0\rangle \partial_x q, \text{ for } x \in \mathbb{T},\ t \geq 0, \tag{2.6}$$

$$\partial_t q = (h + \langle u_0\rangle)^{r-1}\partial_x h, \text{ for } x \in \mathbb{T},\ t \geq 0, \tag{2.7}$$

where $h = u - \langle u_0\rangle$. Without lossing generality we consider $\langle u_0\rangle = 1$. Then we have that

Theorem 1 *Define*

$$\nu(t) = 1 - \theta t,$$

for

$$\theta > 1 + \|h_0\|_{\nu(0)} + C_r(1 + \|h_0\|^{r-1}_{\nu(0)}) + \|q_0\|_{\nu(0)}.$$

Let us consider (h_0, q_0) *such that*

$$\|h(t)\|_{\nu(t)} = \sum_{n=-\infty}^{\infty} |\hat{h}(n,t)|e^{\nu(t)|n|} < \infty,$$

$$\|q(t)\|_{\nu(t)} = \sum_{n=-\infty}^{\infty} |\hat{q}(n,t)|e^{\nu(t)|n|} < \infty.$$

Then, there exist a sufficiently short time and a unique local solution which is analytic in a complex strip with sufficiently small width.

Proof We define the scale of spaces

$$\mathbb{A}_{\nu(t)} = \left\{u \in L^2, e^{\nu(t)|n|}\hat{u}(n) \in \ell^1\right\}$$

with norm

$$\|u\|_{\nu(t)} = \|e^{\nu(t)|n|}\hat{u}(n)\|_{\ell^1}.$$

We observe that the previous spaces are a Banach Algebra

$$\|fg\|_{\nu(t)} \leq \|f\|_{\nu(t)}\|g\|_{\nu(t)}.$$

We compute

$$\begin{aligned}\frac{d}{dt}\|F\|_{\nu(t)} &= \sum_{n=-\infty}^{\infty} \nu'(t)|n|e^{\nu(t)|n|}|\widehat{F}(n,t)| + \sum_{n=-\infty}^{\infty} e^{\nu(t)|n|}\mathrm{Re}\left(\frac{\partial}{\partial t}\widehat{F}(n,t)\frac{\overline{\widehat{F}(n,t)}}{|\widehat{F}(n,t)|}\right)\\ &\leq \sum_{n=-\infty}^{\infty} \nu'(t)|n|e^{\nu(t)|n|}|\widehat{F}(n,t)| + \left\|\frac{\partial}{\partial t}F\right\|_{\nu(t)}.\end{aligned}$$

Then, if $0 < \nu(t)$ is a decreasing function we find a regularizing contribution coming from ν'. This regularizing contribution is reflecting the fact that the strip of analyticity is shrinking. At this point it is easy to find the estimate

$$\begin{aligned}\frac{d}{dt}(\|h\|_{\nu(t)} + \|q\|_{\nu(t)}) &\leq \nu'(t)(\|\partial_x h\|_{\nu(t)} + \|\partial_x q\|_{\nu(t)}) - \|(-\Delta)^{\alpha/2}h\|_{\nu(t)}\\ &\quad + (1 + \|h\|_{\nu(t)})\|\partial_x q\|_{\nu(t)} + \|\partial_x h\|_{\nu(t)}(C_r(1 + \|h\|^{r-1}_{\nu(t)}) + \|q\|_{\nu(t)})\\ &\leq 0,\end{aligned}$$

where in the last line we have fix

$$\theta > 1 + \|h_0\|_{\nu(0)} + C_r(1 + \|h_0\|^{r-1}_{\nu(0)}) + \|q_0\|_{\nu(0)}.$$

□

We want to remark that the previous result does not require any sign condition on h nor the parabolic term $(-\Delta)^{\alpha/2}$.

2.2. Finite time blow up for the inviscid case

We consider the inviscid system

$$\partial_t u = \partial_x(uq), \text{ for } x \in \mathbb{T},\ t \geq 0, \tag{2.8}$$

$$\partial_t q = u^{r-1}\partial_x u, \text{ for } x \in \mathbb{T},\ t \geq 0. \tag{2.9}$$

and prove the following result:

Theorem 2 *Let us consider $r = 1$. Then there exist smooth initial data such that the corresponding solution to* (2.8)-(2.9) *blows up in finite time.*

Proof Assume that $0 \leq u_0(x)$, is an even function such that $u_0(0) = \partial_x^2 u_0(0) = 0$. Assume also that q_0 is an odd function such that $\partial_x q_0(0) = 0$. We note that the symmetry is preserved, *i.e.* as long as the solution exist, $u(x,t)$ remains even and $q(x,t)$ remains odd.

The proof is similar to the one in [3]. We argue by contradiction: assume that a a global classical solution exists for this initial data. Then we want to prove that some quantity blows up. We define the following quantities

$$U_i(t) = \partial_x^i u(x,t)\Big|_{x=0},\ Q_i(t) = \partial_x^i q(x,t)\Big|_{x=0}.$$

Then we have

$$\begin{aligned}\frac{d}{dt}U_0 &= U_1 Q_0 + Q_1 U_0\\ &= Q_1 U_0,\end{aligned}$$

so

$$U_0(t) = U_0(0)e^{\int_0^t Q_1(s)ds} = 0.$$

In the same way,

$$\frac{d}{dt}Q_1 = U_2.$$

$$\begin{aligned}\frac{d}{dt}U_2 &= U_3 Q_0 + 3U_2 Q_1 + 3U_1 Q_2 + Q_3 U_0\\ &= 3U_2 Q_1\\ &= 3\frac{d}{dt}Q_1 Q_1,\end{aligned}$$

so

$$\frac{d^2}{dt^2}Q_1 = \frac{d}{dt}U_2(t) = \frac{d}{dt}\frac{3}{2}Q_1(t)^2,$$

and that implies the finite time blow up of Q_1 and U_2. □

2.3. Wave propagation

Finally we turn our attention to the wave-like form of the system (2.6)-(2.7). Indeed, we observe that the system (2.6)-(2.7) with

$$\langle u_0 \rangle = r = 1,$$

can be written as the following bidirectional non-local wave equation

$$\partial_t^2 q = -(-\Delta)^{\alpha/2}\partial_t q + \partial_x^2(\partial_x^{-1}\partial_t q q) + \partial_x^2 q. \tag{2.10}$$

Then, if ε is a small parameter and we consider the unknown

$$q = \varepsilon f,$$

we find the equation

$$\partial_t^2 f = -(-\Delta)^{\alpha/2}\partial_t f + \varepsilon\partial_x^2(\partial_x^{-1}\partial_t f f) + \partial_x^2 f. \tag{2.11}$$

To find the equation describing unidirectional waves we change to far-field variables

$$\xi = x - t, \quad \tau = \varepsilon t.$$

We can apply the chain rule to compute

$$\frac{\partial}{\partial t} f(\chi(x,t),\tau(t)), \quad \frac{\partial^2}{\partial t^2} f(\chi(x,t),\tau(t))$$

and, as a consequence, we find that

$$\partial_\xi^2 f - 2\varepsilon\partial_\tau\partial_\xi f + \varepsilon^2\partial_\tau^2 f = (-\Delta)^{\alpha/2}\partial_\xi f - \varepsilon(-\Delta)^{\alpha/2}\partial_\tau f + \varepsilon\partial_\xi^2(\partial_\xi^{-1}(-\partial_\xi f + \varepsilon\partial_\tau f)f) + \partial_\xi^2 f. \tag{2.12}$$

Then, if we neglect terms of order $O(\varepsilon^2)$, we obtain the asymptotic equation

$$-2\varepsilon\partial_\tau\partial_\xi f = (-\Delta)^{\alpha/2}\partial_\xi f - \varepsilon(-\Delta)^{\alpha/2}\partial_\tau f - \varepsilon\partial_\xi^2(\partial_\xi^{-1}\partial_\xi f)f). \tag{2.13}$$

Integrating in ξ,

$$-2\varepsilon\partial_\tau f + \varepsilon(-\Delta)^{(\alpha-1)/2}H\partial_\tau f = (-\Delta)^{\alpha/2} f - 2\varepsilon f\partial_\xi f. \tag{2.14}$$

Changing back to our previous notation for the independent variables, we conclude

$$\partial_t f - \frac{1}{2}(-\Delta)^{(\alpha-1)/2}H\partial_t f = -\frac{1}{2\varepsilon}(-\Delta)^{\alpha/2} f + f\partial_x f. \tag{2.15}$$

Acknowledgements

R.G-B was supported by the project "Mathematical Analysis of Fluids and Applications" with reference PID2019-109348GA-I00/AEI/ 10.13039/501100011033 and acronym "MAFyA" funded by Agencia Estatal de Investigación and the Ministerio de Ciencia, Innovacion y Universidades (MICIU).

References

[1] Y. Ascasibar, R. Granero-Belinchón, and J. M. Moreno. An approximate treatment of gravitational collapse. *Physica D: Nonlinear Phenomena*, 262:71 – 82, 2013.

[2] R. Atkinson, C. Rhodes, D. Macdonald, and R. Anderson. Scale-free dynamics in the movement patterns of jackals. *Oikos*, 98(1):134–140, 2002.

[3] H. Bae and R. Granero-Belinchón. Singularity formation for the Serre-Green-Naghdi equations and applications to abcd-Boussinesq systems *arXiv preprint arXiv:2001.11937*, 2020.

[4] F. Bartumeus, F. Peters, S. Pueyo, C. Marrasé, and J. Catalan. Helical lévy walks: adjusting searching statistics to resource availability in microzooplankton. *Proceedings of the National Academy of Sciences*, 100(22):12771–12775, 2003.

[5] A. Blanchet, J. Dolbeault, and B. Perthame. Two-dimensional Keller-Segel model: optimal critical mass and qualitative properties of the solutions. *Electron. J. Differential Equations*, pages No. 44, 32 pp. (electronic), 2006.

[6] N. Bournaveas and V. Calvez. The one-dimensional Keller-Segel model with fractional diffusion of cells. *Nonlinearity*, 23(4):923, 2010.

[7] J. Burczak and R. Granero-Belinchón. Boundedness of large-time solutions to a chemotaxis model with nonlocal and semilinear flux. *Topological Methods in Nonlinear Analysis.*, 47(1):369–387, 2016.

[8] J. Burczak and R. Granero-Belinchón. Critical Keller-Segel meets Burgers on $\mathbb{S}^1$: large-time smooth solutions. *Nonlinearity*, 29(12):3810, 2016.

[9] J. Burczak and R. Granero-Belinchón. Global solutions for a supercritical drift-diffusion equation. *Advances in Mathematics*, 295:334–367, 2016.

[10] J. Burczak and R. Granero-Belinchón. On a generalized doubly parabolic Keller-Segel system in one spatial dimension. *Mathematical Models and Methods in the Applied Sciences*, 26(1):111–160, 2016.

[11] J. Burczak and R. Granero-Belinchón. Suppression of blow up by a logistic source in 2 d keller-segel system with fractional dissipation. *Journal of Differential Equations*, 263(9):6115–6142, 2017.

[12] B. J. Cole. Fractal time in animal behaviour: the movement activity of drosophila. *Animal Behaviour*, 50(5):1317–1324, 1995.

[13] C. Escudero. The fractional Keller-Segel model. *Nonlinearity*, 19(12):2909, 2006.

[14] R. Granero-Belinchón. Global solutions for a hyperbolic–parabolic system of chemotaxis. *Journal of Mathematical Analysis and Applications*, 449(1):872–883, 2017.

[15] R. Granero-Belinchón. On the fractional fisher information with applications to a hyperbolic–parabolic system of chemotaxis. *Journal of Differential Equations*, 262(4):3250–3283, 2017.

[16] R. Granero-Belinchón and R. Orive-Illera. An aggregation equation with a nonlocal flux. *Nonlinear Analysis: Theory, Methods & Applications*, 108(0):260

[17] T. Hillen, K. J. Painter, and M. Winkler. Convergence of a cancer invasion model to a logistic chemotaxis model. *Math. Models Methods Appl. Sci.*, 23(1):165–198, 2013.

[18] E. Keller and L. Segel. Initiation of slime mold aggregation viewed as an instability. *Journal of Theoretical Biology*, 26(3):399–415, 1970.

[19] M. Levandowsky, B. White, and F. Schuster. Random movements of soil amebas. *Acta Protozoologica*, 36:237–248, 1997.

[20] D. Li, J. Rodrigo, and X. Zhang. Exploding solutions for a nonlocal quadratic evolution problem. *Revista Matematica Iberoamericana*, 26(1):295–332, 2010.

[21] C. Patlak. Random walk with persistence and external bias. *Bulletin of Mathematical Biology*, 15(3):311–338, 1953.

[22] D. A. Raichlen, B. M. Wood, A. D. Gordon, A. Z. Mabulla, F. W. Marlowe, and H. Pontzer. Evidence of Lévy walk foraging patterns in human hunter–gatherers. *Proceedings of the National Academy of Sciences*, 111(2):728–733, 2014.

[23] M. F. Shlesinger and J. Klafter. Lévy walks versus Lévy flights. In *On growth and form*, pages 279–283. Springer, 1986.

[24] A. Stevens, Angela and H. Othmer, SIAM Journal on Applied Mathematics 57(4):1044–1081, 1997.

[25] Z.-A. Wang, Z. Xiang, and P. Yu. Asymptotic dynamics on a singular chemotaxis system modeling onset of tumor angiogenesis. *Journal of Differential Equations*, 260(3):2225–2258, 2016.

[26] Y. Zhang, Z. Tan, and M.-B. Sun. Global existence and asymptotic behavior of smooth solutions to a coupled hyperbolic-parabolic system. *Nonlinear Analysis: Real World Applications*, 14(1):465–482, 2013.

[27] H.-Y. Jin, J. Li, and Z.-A. Wang. Asymptotic stability of traveling waves of a chemotaxis model with singular sensitivity. *Journal of Differential Equations*, 255(2):193–219, 2013.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 208–215)

An arbitrary high order ADER Discontinous Galerking (DG) numerical scheme for the multilayer shallow water model with variable density

E. Guerrero Fernández [1], M.J. Castro Díaz[1], M. Dumbser[2], T. Morales de Luna[3]

1. *Departamento de Análisis Matemático, Facultad de Ciencias, Universidad de Málaga, Campus de Teatinos S/N, 29081 Málaga, Spain.*
2. *Department of Civil, Environmental and Mechanical Engineering, Via Mesiano, 77 - 38123 Trento, Italy.*
3. *Departamento de Matemáticas, Universidad de Córdoba, Campus de Rabanales, 14071 Córdoba, Spain.*

Abstract

In this work, an arbitrary high order numerical discretization for a density dependent multilayer shallow-water model is presented. The model can be written as a system of hyperbolic PDE equations and it is especially suited for simulations of density driven gravity currents within the shallow-water framework. The proposed discretization is composed by an unlimited high order accurate (ADER) Discontinuous Galerkin (DG) method, which is then limited *a posteriori* with the MOOD paradigm, resulting in great resolution capabilities in smooth regions alongside a robust and accurate respond for strong gradients or discontinuities. A numerical strategy to preserve non-trivial stationary solutions is also discussed. Some numerical results are shown including density driven currents where laboratory data is available.

1. Introduction

A wildly used model for the simulation of geophysical flows is shallow-water (or Saint-Venant) model. In shallow-water flows, the vertical component of the velocity is neglected and the horizontal component is assumed to be constant along the vertical direction. In this way, the dimension of the problem is reduced by one, allowing to improve dramatically the computational times for large scale simulations. This approach has been successfully used in many practical applications (see [10, 18, 19]). However, the horizontally constant velocity hypothesis can seriously limit the amount of information that the model is able to provide and that may be relevant for the problem. To address this issue, multilayer shallow-water models are developed, where the vertical direction is subsequently divided in computational layers and the shallow-water hypotheses are performed in each layer individually (see [4, 5]). This allows, for instance, to recover a detailed vertical profile of the velocity and the cost of a slightly higher computational times. Of course, some mechanism for the interaction between layers must be considered. For instance, [7, 17] assume immiscible layers meanwhile for multilayer shallow-water systems considered in [3, 15], a continuous mass and momentum exchange between the layers is considered. The incorporation of the mass and momentum transfer between layers is performed via non-conservative terms. The multilayer shallow-water model considered in this work includes density effects throughout density dependent pressure terms. A full description of the derivation of the model can be found at [16] and [2].

The Discontinuous Galerkin (DG) method itself dates back to the early work by Reed and Hill in [21]. This methods allows to easily reach high order in space. In more recent work, it is combined with an arbitrary high order derivatives (ADER) procedure, which allows to reach arbitrary high order in time (see [12]). The ADER approach is based on the approximated solution of Riemann problems by means of a fixed point algorithm in each element locally. This combination leads naturally to high order, single step and fully discrete numerical schemes. However, this approach is unlimited, in the sense that there is no mechanism to prevent the apparition of spurious oscillations near strong gradients or discontinuities. As a limiting technique, we use a multi-dimensional optimal order detection (MOOD) (see [13]), which is *a posteriori* approach to the problem of limiting. The unlimited solution of the ADER-DG scheme is tested to study its admissibility in terms of spurious oscillations but also other physical criteria like positivity. If the solution is found inadequate, then the MOOD technique will switch to a robust second order accurate finite volume method in order to compute the limited solution.

Another issue of paramount importance for the long time numerical stability of the numerical scheme are the ability of the scheme to preserve stationary solutions ([6, 9]). Indeed, many practical applications often consist on a perturbation of an *equilibria* state, and thus exactly preserving this state is of great importance. Here, we propose a procedure to exactly preserve non-trivial stationary solutions in the ADER-DG framework.

Due to space restrictions, these techniques will be presented in a non-exhaustive manner. However, the interested reader has references available throughout the text.

2. Model description

Here, we briefly present the density dependent multilayer shallow-water model considered in this work. A full description and derivation of the model can be found at [16]. The full system of equations for the model in one dimension is,

$$\begin{cases} \partial_t h + \partial_x \left(h \sum_{\beta=1}^{M} u_\beta \right) = 0, \\ \partial_t (h\theta_\alpha) + \partial_x (h\theta_\alpha u_\alpha) = \dfrac{1}{l_\alpha} \left(\theta_{\alpha+\frac{1}{2}} G_{\alpha+\frac{1}{2}} - \theta_{\alpha-\frac{1}{2}} G_{\alpha-\frac{1}{2}} \right), \\ \partial_t (h\theta_\alpha u_\alpha) + \partial_x (h\theta_\alpha u_\alpha^2) + g h \theta_\alpha \partial_x \eta + \dfrac{g\, l_\alpha}{2} \left(h\partial_x (h\theta_\alpha) - h\theta_\alpha \partial_x h \right) \\ + g \sum_{\beta=\alpha+1}^{M} l_\beta \left(h\partial_x (h\theta_\beta) - h\theta_\alpha \partial_x h \right) = \dfrac{1}{l_\alpha} \left(u_{\alpha+\frac{1}{2}} \theta_{\alpha+\frac{1}{2}} G_{\alpha+\frac{1}{2}} - u_{\alpha-\frac{1}{2}} \theta_{\alpha-\frac{1}{2}} G_{\alpha-\frac{1}{2}} \right), \end{cases} \tag{2.1}$$

where h is the total height of the water column, $\eta = h + z_b$ is the free surface, and z_b is the bathymetry function. Additionally, u_α refers to the horizontal velocity while θ_α is the relative density of the fluid in the α-layer. Finally, $G_{\alpha\pm\frac{1}{2}}$ $\alpha = 1, \ldots, M$, are the mass transference terms between layers.

System (2.1) is obtained under the closure hypothesis that the layer thickness is proportional to the total height, $h_\alpha = l_\alpha h$, with $l_\alpha \in [0, 1]$, $\alpha = 1, \ldots, M$ such that $\sum_{\alpha=1}^{M} l_\alpha = 1$. Under this assumption, we are able to give an expression for the mass transference terms,

$$G_{\alpha+\frac{1}{2}} = \sum_{\beta=1}^{\alpha} l_\beta \left(\partial_t h + \partial_x (h u_\beta) \right) = \sum_{\beta=1}^{\alpha} l_\beta \left(\partial_x (h u_\beta) - \partial_x \left(\sum_{\gamma=1}^{M} l_\gamma h u_\gamma \right) \right). \tag{2.2}$$

We assume no mass transference at the bottom and free surface, $G_{1/2} = G_{M+1/2} = 0$, and $\theta_{\alpha+1/2}$ and $u_{\alpha+1/2}$ are some approximations of u and θ at the layers interfaces, for example a simple arithmetic mean. Note that the full system (2.1) reduces to the standard shallow water equations for the particular case $M = 1$ and $\theta = 1$.

The full PDE system (2.1) has an infinity number of stationary solutions. Indeed, the standard shallow-water stationary solutions with constant free surface η are also solution of the system (2.1) if a homogeneous density profile is considered,

$$\theta_\alpha = \text{cte}, \quad u_\alpha = 0, \quad \text{for } \alpha = 1, \ldots, M, \quad \eta = \text{cte}.$$

However, system (2.1) also admits lake-at-rest stationary solutions corresponding to non-trivial density profiles. Stationary solutions with $u_\alpha = 0$, $\alpha = 1, \ldots, M$ for the system (2.1) correspond to the solutions of the following ODE system,

$$P_\alpha := g h \theta_\alpha \partial_x \eta + \frac{g l_\alpha}{2} \left(h\partial_x (h\theta_\alpha) - h\theta_\alpha \partial_x h \right) + g \sum_{\beta=\alpha+1}^{M} l_\beta \left(h\partial_x (h\theta_\beta) - h\theta_\alpha \partial_x h \right) = 0. \tag{2.3}$$

Once the free surface is fixed, this equation can be solved iteratively by solving first the upper layer and sequentially going downwards throughout the lower layers. In particular we are interested in those with a constant free surface and a vertically stratified density profile, that is,

$$\eta(x) = h(x) + z_b(x) = \text{cte}, \qquad \theta(z) = \theta_{surface} + \gamma(\eta - z). \tag{2.4}$$

Unfortunately, due to the numerical discretization performed on the full system of PDE equations (2.1), this profile is not a stationary solution of (2.3) and cannot be directly preserved unless the bathymetry is the constant function. However, system (2.3) can be solved recursively, which results into a stratified density profile that could be seen as an approximation of (2.4) associated to the multilayer approach. In particular, those solutions are given by the following expression,

$$\begin{gathered} u_\alpha = 0, \ \eta(x) = z_b(x) + h(x) = \text{cte}, \\ \theta_M(x) = \bar{\theta}_M \geq 1, \\ \theta_\alpha(x) = \bar{\theta}_\alpha \, h^{2(M-\alpha)}(x) + \sum_{\beta=\alpha+1}^{M} S_{2(M-\beta)}(M - \alpha + 1) \bar{\theta}_\beta \, h^{2(M-\beta)}(x), \end{gathered} \tag{2.5}$$

with

$$\begin{aligned} S_\beta(\alpha) &= (\beta+1)\cdot A_{\frac{\beta+2}{2}+1}(\alpha), \\ A_p(k) &= \begin{cases} 1 & \text{if } p \geq k, \\ (p-1)\prod_{\gamma=2}^{k-p}(1+(p-2)C_{\gamma-1}) & \text{if } p < k, \end{cases} \\ C_\gamma &= C_{\gamma-1} - \frac{1}{Q_\gamma}, \\ Q_\gamma &= Q_{\gamma-1} + \gamma + 1, \\ C_0 &= Q_0 = 1, \end{aligned}$$

where $\bar{\theta}_\alpha$ is a free choice constant fixed by the initial conditions, that determines the vertical profile of the density.

For more details relative to this model, we refer the reader to [16].

3. Numerical discretization

In this section we provide brief description of the numerical scheme used on (2.1). If the interested reader wants to know more we refer them to [13, 14].

System (2.1) may be written as follows,

$$\partial_t \boldsymbol{w} + \partial_x \boldsymbol{F}_C(\boldsymbol{w}) + \boldsymbol{P}(\boldsymbol{w}, \partial_x \boldsymbol{w}, \partial_x \eta) - \boldsymbol{T}(\boldsymbol{w}, \partial_x \boldsymbol{w}) = \boldsymbol{0}, \tag{3.1}$$

where $\boldsymbol{w}$ is the vector of the conserved variables,

$$\boldsymbol{w} = (h \,|\, h\theta_\alpha \,|\, h\theta_\alpha u_\alpha)^T \in \mathbb{R}^{2M+1}, \tag{3.2}$$

the physical convective flux $F_C(\boldsymbol{w})$ is given by,

$$F_C(\boldsymbol{w}) = \left(hu_\alpha \,|\, h\theta_\alpha u_\alpha \,|\, h\theta_\alpha u_\alpha^2\right)^T \in \mathbb{R}^{2M+1}, \tag{3.3}$$

and $\boldsymbol{P}(\boldsymbol{w}, \partial_x \boldsymbol{w}, \partial_x \eta)$ corresponds to the pressure term, which depends on the relative density θ_α and the free surface η, and has the following form,

$$\boldsymbol{P}(\boldsymbol{w}, \partial_x \boldsymbol{w}, \partial_x \eta) = (\boldsymbol{0} \,|\, \boldsymbol{0} \,|\, P_\alpha) \in \mathbb{R}^{2M+1}, \tag{3.4}$$

where

$$P_\alpha = gh\theta_\alpha \partial_x \eta + \frac{g\, l_\alpha}{2}(h\partial_x(h\theta_\alpha) - h\theta_\alpha \partial_x h) + g\sum_{\beta=\alpha+1}^{M} l_\beta (h\partial_x(h\theta_\beta) - h\theta_\alpha \partial_x h). \tag{3.5}$$

Finally, the term $\boldsymbol{T}(\boldsymbol{w}, \partial_x \boldsymbol{w})$ corresponds to the mass, density, and momentum exchange between layers:

$$\boldsymbol{T}(\boldsymbol{w}, \partial_x \boldsymbol{w}) = \left(0 \,\middle|\, \frac{1}{l_\alpha}\left(\theta_{\alpha+\frac{1}{2}} G_{\alpha+\frac{1}{2}} - \theta_{\alpha-\frac{1}{2}} G_{\alpha-\frac{1}{2}}\right) \,\middle|\, \frac{1}{l_\alpha}\left(u_{\alpha+\frac{1}{2}}\theta_{\alpha+\frac{1}{2}} G_{\alpha+\frac{1}{2}} - u_{\alpha-\frac{1}{2}}\theta_{\alpha-\frac{1}{2}} G_{\alpha-\frac{1}{2}}\right)\right)^T \in \mathbb{R}^{2M+1}. \tag{3.6}$$

The system of equations (3.1) is solved by applying the family of pure Discontinuous Galerkin methods $\mathbb{P}_N\mathbb{P}_N$. The numerical scheme is formulated as a predictor-corrector method: in the first step, a predictor solution, which consist on a high order approximation of the solution at the following time step, is computed by means of a local Cauchy problem, without interaction with the neighbours states. In the next step, the corrector will make use of these predictor solution to compute a high order in space and time approximation of the solution of system (2.1) at the next time step.

The usual one dimensional considerations relative to the domain discretizations into non-overlaping conforming set of elements are considered. The computation domain Ω is discretized into $T_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$, $i = 1, \ldots, N_s$ elements, where N_s is the total number of cells with a constant length $\Delta x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$.

We will make use of the following notation: for any variable f defined on a volume T_i, we will denote by $f_{i\pm\frac{1}{2}}$ the values at the interface, depending on whether it is the right or left side of the cell. However, when the values correspond to projected states into the interface, it will be generally denoted with the super index $f^\pm$, depending on whether they correspond to the left or to the right side of the intercell.

In the following, the discrete solution of the PDE system (3.1) at time t^n is denoted by $\boldsymbol{w}_h(x, t^n)$ and is defined in terms of piecewise polynomials of degree N on the spatial direction. We shall denote by $\mathcal{U}_h$ the space of piecewise polynomials up to degree N so that $\boldsymbol{w}_h(\cdot, t^n) \in \mathcal{U}_h$. In this work, a nodal basis defined by the Lagrange

interpolation polynomials over the $(N+1)$ Gauss-Legendre quadrature nodes on the element T_i is adopted. As usual in the discontinuous Galerkin (DG) approach, the discrete solution $\boldsymbol{w}_h$ may be discontinuous across the intercells, as in finite volume methods. At each cell T_i, the discrete solution is written in terms of the nodal spatial basis functions $\Phi_l(x)$ and some unknown degrees of freedom $\hat{\boldsymbol{w}}_{i,l}^n$,

$$\boldsymbol{w}_h(x,t^n) = \sum_l \hat{\boldsymbol{w}}_{i,l}^n \Phi_l(x) := \hat{\boldsymbol{w}}_{i,l}^n \Phi_l(x), \qquad \text{for } x \in T_i, \tag{3.7}$$

where the Einstein summation convention over two repeated indices has been considered. The spatial basis functions are defined on the reference interval $[0,1]$.

The ADER-DG method results from multiplying the governing PDE system (3.1) with a test function $\Phi_k \in \mathcal{U}_h$ and integrate over the space-time control volume $T_i \times [t^n, t^{n+1}]$. This results in the expression,

$$\int_{t^n}^{t^{n+1}} \int_{T_i} \Phi_k \partial_t \boldsymbol{w}\, dxdt \;+\; \int_{t^n}^{t^{n+1}} \int_{T_i} \Phi_k \left(\partial_x \boldsymbol{F}_C(\boldsymbol{w})\, dxdt + \boldsymbol{P}(\boldsymbol{w}, \partial_x \boldsymbol{w}, \partial_x \eta) - \boldsymbol{T}(\boldsymbol{w}, \partial_x \boldsymbol{w})\right) dxdt \;=\; \boldsymbol{0}. \tag{3.8}$$

The discrete solution $\boldsymbol{w}_h(x,t^n)$ is allowed to jump across element interfaces, which means that the resulting jump terms have to be properly taken into account. In our scheme this is achieved via numerical flux functions in the form of approximate Riemann solvers that follows the path-conservative approach that was developed by Parés and collaborators in the finite volume framework [8, 20] and which has later been extended to the discontinuous Galerkin finite element framework in [11, 22]. The Riemann solver used in this work is detailed in [16].

In the ADER-DG framework, the higher order in time is achieved with the use of an element-local space-time predictor, denoted by $\mathbf{q}_h(x,t)$ in the following, and which will be discussed in more detail later. Using (3.7), and after some computation on (3.8), we arrive to the following weak formulation,

$$\begin{aligned}
&\left(\int_{T_i} \Phi_k \Phi_l\, dx\right)\left(\hat{\boldsymbol{w}}_{i,l}^{n+1} - \hat{\boldsymbol{w}}_{i,l}^{n}\right) - \int_{t^n}^{t^{n+1}} \int_{T_i^\circ} \left(\partial_x \Phi_k \cdot \boldsymbol{F}_C(\mathbf{q}_h)\right) dxdt \\
&+ \int_{t^n}^{t^{n+1}} \Phi_{k,i+\frac{1}{2}} \mathcal{D}^-_{i+\frac{1}{2}}\left(\mathbf{q}^-_{h,i+\frac{1}{2}}, \mathbf{q}^+_{h,i+\frac{1}{2}}, {zb}^-_{h,i+\frac{1}{2}}, {zb}^+_{h,i+\frac{1}{2}}\right) + \Phi_{k,i-\frac{1}{2}} \mathcal{D}^+_{i-\frac{1}{2}}\left(\mathbf{q}^-_{h,i-\frac{1}{2}}, \mathbf{q}^+_{h,i-\frac{1}{2}}, {zb}^-_{h,i-\frac{1}{2}}, {zb}^+_{h,i-\frac{1}{2}}\right) dt \\
&\qquad + \int_{t^n}^{t^{n+1}} \int_{T_i^\circ} \Phi_k \left(\boldsymbol{P}(\mathbf{q}_h, \partial_x \mathbf{q}_h, \partial_x \eta_h) - \boldsymbol{T}(\mathbf{q}_h, \partial_x \mathbf{q}_h)\right) dxdt = \mathbf{0},
\end{aligned} \tag{3.9}$$

where T_i° corresponds to the interior of T_i and f_h stands for the projection of f onto the space $\mathcal{U}_h$. Moreover, $\mathcal{D}^\pm_{i\pm\frac{1}{2}}$ are the numerical flux at the cell interface given by the Riemann solver.

3.1. ADER-DG space-time predictor

We focus now on the computation of the predictor solution $\mathbf{q}_h(x,t)$, based on a weak formulation of the governing PDE system in space-time. The PDE system (2.1) is approximated with a so-called Cauchy problem *in the small*, i.e. without considering the interaction with the neighbour elements. Again, a similar space-time basis is considered to expand the predictor solution,

$$\mathbf{q}_h(x,t) = \sum_l \theta_l(x,t) \hat{\mathbf{q}}_l^i := \theta_l(x,t) \hat{\mathbf{q}}_l^i, \tag{3.10}$$

with the multi-index $l = (l_0, l_1)$ and where the space-time basis functions $\theta_l(x,t) = \varphi_{l_0}(\tau)\varphi_{l_1}(\xi)$ are again generated from the same one-dimensional nodal basis functions as before. Proceeding now similarly to the system (3.9), we multiply (3.1) by a space-time function and integrate over the space-time control volume $T_i \times [t^n, t^{n+1}]$. However, since we are only interested in an element local predictor solution, without interactions with the neighbor elements, the jump terms across interfaces are not taken into account. This leads to,

$$\begin{aligned}
&\int_{T_i} \theta_k(x,t^{n+1}) \mathbf{q}_h(x,t^{n+1})\, dx - \int_{T_i} \theta_k(x,t^n) \mathbf{q}_h^0(x,t^n)\, dx - \int_{t^n}^{t^{n+1}} \int_{T_i} \partial_t \theta_k(x,t) \mathbf{q}_h(x,t)\, dxdt \\
&\qquad = -\int_{t^n}^{t^{n+1}} \int_{T_i} \theta_k(x,t) \left(\partial_x \boldsymbol{F}_C(\mathbf{q}_h) + \boldsymbol{P}(\mathbf{q}_h, \partial_x \mathbf{q}_h, \partial_x \eta_h) - \boldsymbol{T}(\mathbf{q}_h, \partial_x \mathbf{q}_h)\right) dxdt.
\end{aligned} \tag{3.11}$$

Using the local space-time ansatz (3.10), Eq. (3.11) becomes a local nonlinear system for the unknown degrees of freedom $\hat{\mathbf{q}}_l^i$ of the space-time polynomials $\mathbf{q}_h$. The solution to the system can be found via a fixed point algorithm, that will converge, at most, in $N+1$ iterations for linear homogeneous systems. The initial guess $\mathbf{q}_h^0(x,t)$ for the iterative algorithm is simply set as the solution at time t^n, $\boldsymbol{w}_h(x,t^n)$.

This completes the description of the unlimited high order accurate and fully discrete ADER-DG schemes.

3.2. Preserving stationary solutions in the ADER-DG framework

We describe now the techniques developed to construct arbitrary high order ADER-DG numerical schemes that preserve exactly a set of stationary solutions corresponding to a stationary stratified fluid.

The first step consist on determining a local stationary solution $\boldsymbol{u}_{e,i}(x),\ x \in T_i$ of the family (2.5) at each time step. Although the stationary solution is calculated at each time step, we subsequentially drop the time dependence to simplify the notation. Notice that the family of stationary solutions (2.5) with $u_{\alpha,e,i} = 0,\ 1 \leq \alpha \leq M$ are fully determined by setting $h_{e,i}$ and $\theta_{1,e,i}, \dots \theta_{M,e,i}$. Particularly, $\bar{\eta}_i$,

$$\bar{\eta}_i = \frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \left(h_h(x,t^n) + z_{bh}(x)\right)\, dx,$$

where again we have denoted by f_h the discrete representation of f onto the polynomial space $\mathcal{U}_h$. Similarly, the constants $\theta_{1,e,i}, \dots \theta_{M,e,i}$ are computed solving,

$$\frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} (h\theta)_{\alpha,e,i}(x, \bar{\eta}_i, \bar{\theta}_{\alpha,i}, \cdots, \bar{\theta}_{1,i})\, dx = \frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} (h\theta)_{h,\alpha}(x,t^n)\, dx, \quad 1 \leq \alpha \leq M.$$

Using these constant, we are able to compute the stationary solution $\boldsymbol{u}_{e,i}(x)$. Note that this local stationary solutions satisfy the pressure terms (2.3) at each cell,

$$\boldsymbol{P}(\boldsymbol{u}_{e,i}, \partial_x \boldsymbol{u}_{e,i}, \partial_x \bar{\eta}_i) = 0. \tag{3.12}$$

Now, we could replace the numerical scheme (3.9) by the following well-balanced ADER-DG equivalent numerical scheme,

$$\begin{aligned}
&\left(\int_{T_i} \Phi_k \Phi_l\, dx\right)\left(\hat{\boldsymbol{w}}_{i,l}^{n+1} - \hat{\boldsymbol{w}}_{i,l}^{n}\right) - \int_{t^n}^{t^{n+1}} \int_{T_i^\circ} \left(\partial_x \Phi_k \cdot \boldsymbol{F}_C(\mathbf{q}_h)\right)\, dxdt \\
&+ \int_{t^n}^{t^{n+1}} \Phi_{k,i+\frac{1}{2}} \mathcal{D}_{i+\frac{1}{2}}^{-}\left(\mathbf{q}_{h,i+\frac{1}{2}}^{-}, \mathbf{q}_{h,i+\frac{1}{2}}^{+}, {z_b}_{h,i+\frac{1}{2}}^{-}, {z_b}_{h,i+\frac{1}{2}}^{+}\right) + \Phi_{k,i-\frac{1}{2}} \mathcal{D}_{i-\frac{1}{2}}^{+}\left(\mathbf{q}_{h,i-\frac{1}{2}}^{-}, \mathbf{q}_{h,i-\frac{1}{2}}^{+}, {z_b}_{h,i-\frac{1}{2}}^{-}, {z_b}_{h,i-\frac{1}{2}}^{+}\right)\, dt \\
&+ \int_{t^n}^{t^{n+1}} \int_{T_i^\circ} \Phi_k \left(\partial_x \boldsymbol{F}_C(\mathbf{q}_h) - \boldsymbol{T}(\mathbf{q}_h, \partial_x \mathbf{q}_h)\right)\, dxdt \\
&+ \int_{t^n}^{t^{n+1}} \int_{T_i^\circ} \Phi_k \left(\boldsymbol{P}(\mathbf{q}_h, \partial_x \mathbf{q}_h, \partial_x \eta_h) - \boldsymbol{P}((\boldsymbol{u}_{e,i})_h, \partial_x (\boldsymbol{u}_{e,i})_h, \partial_x (\eta_{e,i})_h)\right)\, dxdt = \mathbf{0},
\end{aligned} \tag{3.13}$$

Moreover, the extrapolated values at the intercells, denoted by $\mathbf{q}_{h,i\pm\frac{1}{2}}^{\pm}$, are computed in the following way,

$$\mathbf{q}_{h,i+\frac{1}{2}}^{-} = \boldsymbol{u}_{e,i}(x_{i+\frac{1}{2}}) + \hat{\boldsymbol{q}}_{h,i+\frac{1}{2}}^{-},$$

where $\hat{\boldsymbol{q}}_{h,i+\frac{1}{2}}^{-}$ is the extrapolation on the cell interface of the fluctuation $(\boldsymbol{q}_{h,i} - (\boldsymbol{u}_{e,i})_h)$, that is,

$$\hat{\boldsymbol{q}}_{h,i+\frac{1}{2}}^{-} = (\boldsymbol{q}_{h,i} - (\boldsymbol{u}_{e,i})_h)(x_{i+\frac{1}{2}}).$$

Similarly,

$$\mathbf{q}_{h,i+\frac{1}{2}}^{+} = \boldsymbol{u}_{e,i+1}(x_{i+\frac{1}{2}}) + \hat{\boldsymbol{q}}_{h,i+\frac{1}{2}}^{-},$$

where

$$\hat{\boldsymbol{q}}_{h,i+\frac{1}{2}}^{+} = (\boldsymbol{q}_{h,i+1} - (\boldsymbol{u}_{e,i+1})_h)(x_{i+\frac{1}{2}}).$$

A similar procedure is applied in the ADER step, where a high order local approximation of the solution $\boldsymbol{w}_h(x,t^{n+1})$ is computed by considering a fluctuation with respect to the local stationary solution $\boldsymbol{u}_{e,i}(x)$.

Finally, to clean possible spurious oscillations due to the absence of numerical viscosity in a stationary solution, we could perform the following procedure: first we compute the average of the fluctuation with respect to the local stationary solution,

$$\widehat{\boldsymbol{w}}_{h,i} = \frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \boldsymbol{w}_h(x,t^n) - \boldsymbol{u}_{e,i}(x)\, dx.$$

if $\left|\widehat{\boldsymbol{w}}_{h,i}\right|$ is less than a small threshold, then $\boldsymbol{w}_h(x,t^n)$ is redefined as follows,

$$\boldsymbol{w}_h(x,t^n) = \boldsymbol{u}_{e,i}(x) + \widehat{\boldsymbol{w}}_{h,i}.$$

4. Numerical test

We briefly demonstrate the capacity of the numerical scheme for preserving stationary solutions and to provide accurate results for complex density-driven flows. We first began considering a small perturbation of a lake-at-rest stationary solution with $M = 3$ and with the following non-constant bathymetry and free surface functions,

$$\eta(x,0) = 2 + \frac{1}{10}e^{-5\,x^2}, \qquad z_b(x) = \frac{1}{2}e^{-x^2},$$

defined in the channel with $x \in [-5,5]$ with just 50 elements and a forth order in space and time numerical scheme. Wall type boundary conditions are set and the initial condition for the relative density is given by equation (2.5) with the constant $\bar{\theta}_1 = 1.01$, $\bar{\theta}_2 = 0.02$ and $\bar{\theta}_3 = 0$. Figures 1 to 2 depicts the solution. As expected, a new stratified density profile is reached once a new free surface is achieved, and this new stationary solution is kept for long simulation times.

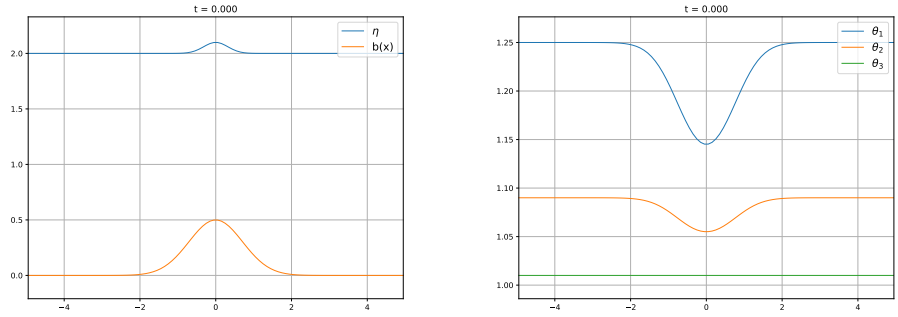


Fig. 1 Perturbation of a lake-at-rest steady state with non-constant density profile at $t = 0$ seconds. Left: free surface and bottom. Right: density profile.

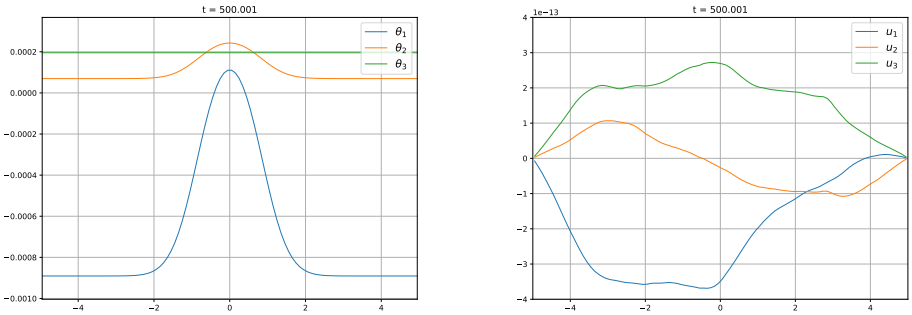


Fig. 2 Perturbation of a lake-at-rest steady state with non-constant density profile. Left: difference of relative densities between times $t = 500$ and $t = 0$ seconds. Right: velocity at time $t = 500$ seconds.

Finally, we show a simulation where a comparison with experimental laboratory data presented in [1]. We consider a flat channel $x \in [0,3]$ and a lock exchange of relative density between two fluids with density $\rho_0 = 1000\,\mathrm{Kg/m^3}$ and $\rho_1 = 1034\,\mathrm{Kg/m^3}$. The fluid with density ρ_1 is within a gatebox of 0.1 meters placed on the left of the channel, which is then released into the fluid ρ_1. The total height of the water is 0.3 meters. Figure 3 (left) depict the initial condition through a heat map of the relative density for a simulation with $M = 30$ layers and just 80 discretization points. To mimic the laboratory experiment in [1], wall-type boundary conditions are considered. Figure 3 (right) shows the simulation at final time $t = 25$ seconds, whereas figure 4 and 5 shows the evolution of the front position as the number of layers M increase. As we can see, we reach outstanding data agreement at approximately $M = 30$ layers.

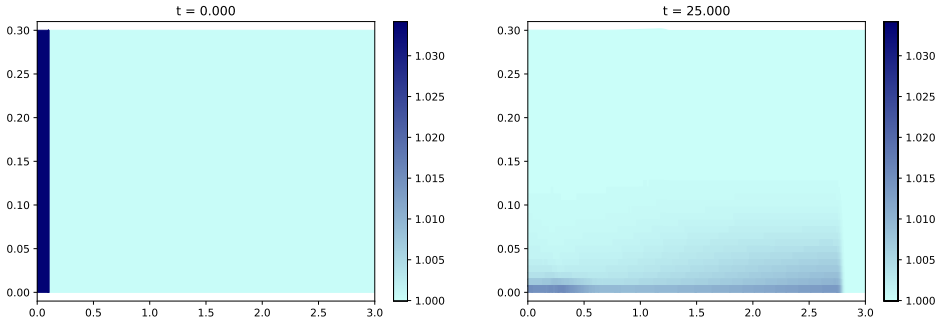


Fig. 3 Lock-exchange experiment in a flat channel: initial condition (left) and relative density at final time $t = 25$ seconds.

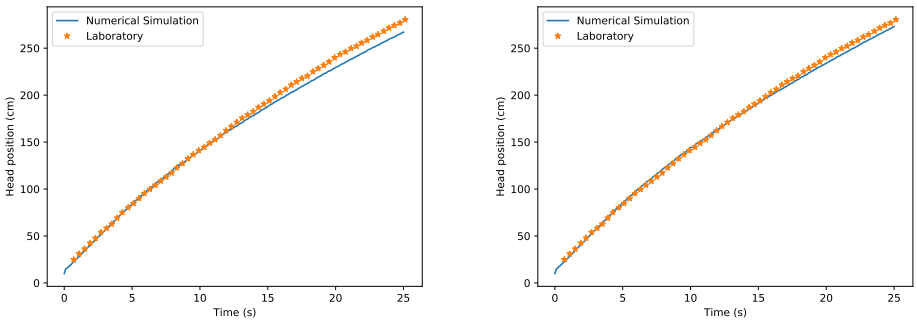


Fig. 4 Lock-exchange experiment in a flat channel: comparison on the evolution of the front position computed with the numerical scheme versus the laboratory data for 20 layers (left) and 25 layers (right).

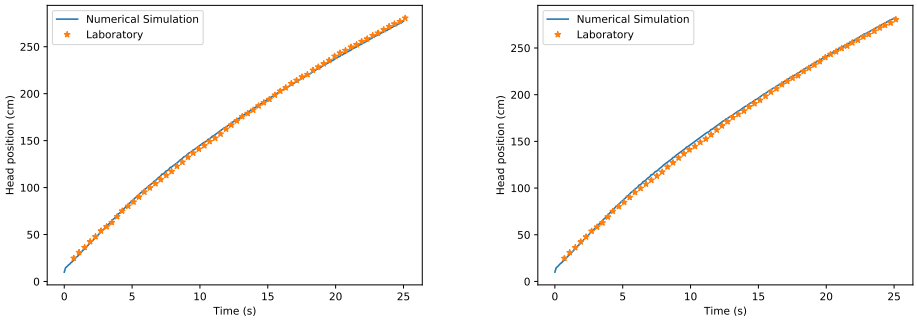


Fig. 5 Lock-exchange experiment in a flat channel: comparison on the evolution of the front position computed with the numerical scheme versus the laboratory data for 30 layers (left) and 40 layers (right).

5. Conclusions

We have briefly presented a novel discretization based on an ADER-DG numerical scheme for a shallow water model with a density dependent pressure term. The numerical scheme is arbitrary high order in space and time and exhibits great accuracy at smooth regions, while providing great results near strong discontinuities thanks to the MOOD strategy combined with a robust path-conservative solver. Finally, a novel strategy for preserving non-trivial stationary solutions in the ADER-DG framework has been presented. The numerical results are promising, showing excellent data agreement, and will help to increase our knowledge of density driven currents.

Acknowledgements

This research has been partially supported by the Spanish Government and FEDER through the coordinated Research project RTI2018-096064-B-C1 and RTI2018-096064-B-C2, The Junta de Andalucía research project P18-RT-3163 and the Junta de Andalucia-FEDER-University of Málaga Research project UMA18-FEDERJA-161.

References

[1] C. Adduce, G. Sciortino, and S. Proietti. Gravity currents produced by lock exchanges: Experiments and simulations with a two-layer shallow-water model with entrainment. *Journal of Hydraulic Engineering*, 138(2):111–121, 2012.

[2] E. Audusse, M.-O. Bristeau, M. Pelanti, and J. Sainte-Marie. Approximation of the hydrostatic navier–stokes system for density stratified flows by a multilayer model: Kinetic interpretation and numerical solution. *Journal of Computational Physics*, 230(9):3453 – 3478, 2011.

[3] Emmanuel Audusse and Marie-Odile Bristeau. A well-balanced positivity preserving "second-order" scheme for shallow water flows on unstructured meshes. *Journal of Computational Physics*, 206(1):311–333, 2005.

[4] Emmanuel Audusse and Marie-Odile Bristeau. Finite-volume solvers for a multilayer saint-venant system. *Applied Mathematics and Computer Science*, 17:311–320, 10 2007.

[5] Emmanuel Audusse, Marie-Odile Bristeau, Benoît Perthame, and Jacques Sainte-Marie. A multilayer saint-venant system with mass exchanges for shallow water flows. derivation and numerical validation. *ESAIM: Mathematical Modelling and Numerical Analysis*, 45(1):169–200, 2011.

[6] Alfredo Bermúdez and Ma Elena Vázquez. Upwind methods for hyperbolic conservation laws with source terms. *Computers & Fluids*, 23(8):1049–1071, 1994.

[7] François Bouchut and Vladimir Zeitlin. A robust well-balanced scheme for multi-layer shallow water equations. *Discrete and Continuous Dynamical Systems-series B*, 13:739–758, 06 2010.

[8] Manuel Castro, José M. Gallardo, and Carlos Parés. High order finite volume schemes based on reconstruction of states for solving hyperbolic systems with nonconservative products. applications to shallow-water systems. *Mathematics of Computation*, 75(255):1103–1134, 2006.

[9] Manuel J Castro and Carlos Parés. Well-balanced high-order finite volume methods for systems of balance laws. *Journal of Scientific Computing*, 82(2):48, 2020.

[10] M. de la Asunción, M.J. Castro, J.M. Mantas, and S. Ortega. Numerical simulation of tsunamis generated by landslides on multiple gpus. *Advances in Engineering Software*, 99:59 – 72, 2016.

[11] M. Dumbser, M. Castro, C. Parés, and E.F. Toro. ADER schemes on unstructured meshes for non-conservative hyperbolic systems: Applications to geophysical flows. *Computers and Fluids*, 38:1731–1748, 2009.

[12] Michael Dumbser and Claus-Dieter Munz. Building blocks for arbitrary high order discontinuous galerkin schemes. *Journal of Scientific Computing*, 27(1-3):215–230, 2006.

[13] Michael Dumbser, Olindo Zanotti, Raphaël Loubère, and Steven Diot. A posteriori subcell limiting of the discontinuous galerkin finite element method for hyperbolic conservation laws. *Journal of Computational Physics*, 278:47–75, Dec 2014.

[14] C. Escalante, M. Dumbser, and M.J. Castro. An efficient hyperbolic relaxation system for dispersive non-hydrostatic water waves and its solution with high order discontinuous galerkin schemes. *Journal of Computational Physics*, 394:385–416, 2019.

[15] Enrique Fernández-Nieto, E H. Koné, and T Chacón Rebollo. A multilayer method for the hydrostatic navier-stokes equations: A particular weak solution. *Journal of Scientific Computing*, 60, 08 2014.

[16] Ernesto Guerrero Fernández, Manuel Jesús Castro-Díaz, and Tomás Morales de Luna. A second-order well-balanced finite volume scheme for the multilayer shallow water model with variable density. *Mathematics*, 8(5):848, 2020.

[17] Nouh Izem, Mohammed Seaid, and Mohamed Wakrim. A discontinuous galerkin method for two-layer shallow water equations. *Mathematics and Computers in Simulation*, 120:12–23, feb 2016.

[18] Miguel Lastra, José M Mantas, Carlos Ureña, Manuel J Castro, and José A García-Rodríguez. Simulation of shallow-water systems using graphics processing units. *Mathematics and Computers in Simulation*, 80(3):598–618, 2009.

[19] Jorge Macias, Manuel J. Castro, José Manuel González-Vida, Marc de la Asunción, and Sergio Ortega. HySEA: An operational GPU-based model for Tsunami Early Warning Systems. In *EGU General Assembly Conference Abstracts*, EGU General Assembly Conference Abstracts, page 14217, May 2014.

[20] C. Parés. Numerical methods for nonconservative hyperbolic systems: a theoretical framework. *SIAM Journal on Numerical Analysis*, 44(1):300–321, 2006.

[21] William H Reed and TR Hill. Triangular mesh methods for the neutron transport equation. Technical report, Los Alamos Scientific Lab., N. Mex.(USA), 1973.

[22] S. Rhebergen, O. Bokhove, and J.J.W. van der Vegt. Discontinuous Galerkin finite element methods for hyperbolic nonconservative partial differential equations. *Journal of Computational Physics*, 227:1887–1922, 2008.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 216–219)

Picard-type iterations for solving Fredholm integral equations

José M. Gutiérrez[1], **Miguel Á. Hernández-Verón**[1]
Universidad de La Rioja, Spain

Abstract

The theoretical solution of Fredholm integral equations involves the calculus of the inverse of an operator. However, for practical purposes, the calculus of this inverse could be not possible or very complicated. For this reason, our aim in this talk is to use iterative methods for approaching such inverse and therefore the solution of the given integral equation. In fact, we use Newton's method to obtain a method with quadratic convergence. In addition, we also use Chebyshev's method to obtain a method with cubic convergence. Next, we extend this idea to iterative methods with a given order of convergence. Finally, we propose the construction of Picard-type iterative methods that do not use derivatives or inverse operators.

1. Introduction

The goal of this work is to obtain an approximate solution of Fredholm integral equations of second kind given by

$$y(x) = f(x) + \lambda \int_a^b K(x,t)y(t)\,dt, \quad s \in [a,b], \quad \lambda \in \mathbb{R}, \tag{1.1}$$

where $f(x) \in C[a,b]$ is a given function and the function $K(x,t)$ is a known function in $[a,b] \times [a,b]$, called kernel of the integral equation. In this equation, $y(x) \in C[a,b]$ is the unknown function to be determined.

We introduce the operator $\mathcal{K} : C[a,b] \to C[a,b]$, given by

$$[\mathcal{K}(y)](s) = \int_a^b K(x,t)y(t)\,dt, \quad s \in [a,b].$$

So, the equation (1.1) can be written in the following form

$$(\mathcal{I} - \lambda\mathcal{K})y(x) = f(x). \tag{1.2}$$

Therefore, its solution is given by obtaining $y(x)$ in (1.2):

$$y(x) = (\mathcal{I} - \lambda\mathcal{K})^{-1} f(x). \tag{1.3}$$

Formula (1.3) provides the exact solution of integral equations (1.1) in a theoretical manner. But in practice, it could be very complicated (or even impossible) the calculus of the inverse $(\mathcal{I} - \lambda\mathcal{K})^{-1}$. To avoid this difficulty, we propose the use of iterative methods for approaching this inverse and therefore the solution of the integral equation.

2. Approximating the inverse $(\mathcal{I} - \lambda\mathcal{K})^{-1}$.

Now, we consider the problem of the approximation of the inverse of the linear operator $A = \mathcal{I} - \lambda\mathcal{K}$ by means of iterative methods for solving nonlinear equations.

To do this, we introduce the following sets:

- $\mathcal{L}(C[a,b], C[a,b])$ is the set of bounded linear operators from the Banach space $C[a,b]$ on itself.
- $GL(C[a,b], C[a,b])$ is the set of invertible operators in $\mathcal{L}(C[a,b], C[a,b])$.

Given a linear operator $A \in GL(C[a,b], C[a,b])$, our target is to solve the equation

$$\mathcal{T}(H) = 0, \text{ where } \mathcal{T}(H) = H^{-1} - A \tag{2.1}$$

by means of iterative methods.

2.1. Newton's method

Our first choice is to use Newton's iteration, that in this case can be written in the following way (see [5] for details):

$$\begin{cases} N_0 \in \mathcal{L}(C[a,b], C[a,b]) \text{ given}, \\ N_{m+1} = N_m - [\mathcal{T}'(N_m)]^{-1}\mathcal{T}(N_m), \quad m \geq 0. \end{cases}$$

The application of this iterative scheme to the integral equation (3.1) allows us to construct a sequence of functions $\{y_m(x)\}_{m\geq 0}$ given by

$$\begin{cases} N_0 \in \mathcal{L}(C[a,b], C[a,b]) \text{ given}, \\ y_0(x) = N_0 f(x), \\ N_m = 2N_{m-1} - N_{m-1}AN_{m-1}, \quad m \geq 0, \\ y_m(x) = N_m f(x), \end{cases} \tag{2.2}$$

that, under adequate conditions, converge to the solution. We would like to highlight that this sequence depends on the choice of a good initial approximation H_0 for the inverse operator $A = \mathcal{I} - \lambda\mathcal{K}$ and on the function $f(x)$ that appears as "independent term" in the integral equation. In [5] we can see local and semilocal convergence results for the iterative scheme (2.2) together with some numerical examples.

2.2. Chebyshev's method

Now we consider Chebyshev's method for solving (2.1),

$$\begin{cases} C_0 \in \mathcal{L}(C[a,b], C[a,b]) \text{ given}, \\ C_{m+1} = C_m - \left[I + \frac{1}{2}L_{\mathcal{T}}(C_m)\right][\mathcal{T}'(C_m)]^{-1}\mathcal{T}(C_m), \quad m \geq 0, \end{cases}$$

where $L_{\mathcal{F}}(C_m) = [\mathcal{F}'(C_m)]^{-1}\mathcal{F}''(C_m)[\mathcal{F}'(C_m)]^{-1}\mathcal{F}(C_m)$. At a first glance, we can think that inverse operators must be used in this algorithm. But we can do the same as in Newton's method to see that Chebyshev's method does not use them (see [3] for more details). Actually, Chebyshev's iteration can be written in the form

$$\begin{cases} \mathcal{F}'(C_m)(P_m - C_m) = -\mathcal{F}(C_m), \quad k \geq 0, \\ \mathcal{F}'(C_m)(C_{m+1} - P_m) = -\frac{1}{2}\mathcal{F}''(C_m)(P_m - C_m)^2. \end{cases} \tag{2.3}$$

Then we can also avoid the use of inverse operators for approximating C_{m+1}.

In consequence, Chebyshev's method for approximating the inverse operator takes the form:

$$\begin{cases} C_0 \in \mathcal{L}(C[a,b], C[a,b]) \text{ given}, \\ C_{m+1} = 3C_m - 3C_mAC_m + C_mAC_mAC_m, \quad m \geq 0. \end{cases} \tag{2.4}$$

The application of the iterative scheme (2.4) to the integral equation (3.1) allows us to construct another sequence of functions $\{y_m(x)\}_{m\geq 0}$ given by

$$\begin{cases} C_0 \in \mathcal{L}(C[a,b], C[a,b]) \text{ given}, \\ y_0(x) = C_0 f(x), \\ C_{m+1} = 3C_m - 3C_mAC_m + C_mAC_mAC_m, \quad m \geq 0, \\ y_m(x) = C_m f(x). \end{cases} \tag{2.5}$$

As in the case of Newton's method (2.2), this sequence depends on the choice of a good initial approximation C_0 for the inverse operator $A = \mathcal{I} - \lambda\mathcal{K}$ and on the function $f(x)$. The increase in computational cost is rewarded with the increase in the order of convergence, passing from quadratic to cubic order of convergence.

2.3. Methods with a prefixed order of convergence

The next step is to generalize the iterative schemes (2.2) and (2.6), obtained from Newton's and Chebyshev's methods respectively. Our idea is to construct iterative schemes, with a prefixed order of convergence, that do not use inverse operators for approximating the inverse of an operator.

For this, we observe that both Newton's and Chebyshev's methods satisfy equalities in the form

$$\mathcal{I} - N_kA = \mathcal{I} - (2N_{k-1} - N_{k-1}AN_{k-1})L = (\mathcal{I} - N_{k-1}A)^2,$$

$$\mathcal{I} - C_k A = \mathcal{I} - (3C_{k-1} - 3C_{k-1}AC_{k-1} + C_{k-1}AC_{k-1}LC_{k-1})A = (\mathcal{I} - C_{k-1}A)^3.$$

Therefore, following the procedure developed in [1] or [6] to generate an iterative scheme with order of convergence $p \geq 2$, we can consider a sequence $T_k = \phi(T_{k-1})$ such that $\mathcal{I} - T_k A = (\mathcal{I} - T_{k-1}A)^p$, that is,

$$T_k A \quad = \quad \mathcal{I} - (\mathcal{I} - T_{k-1}A)^p = \sum_{j=0}^{p-1} \binom{p}{j+1} (-1)^j (T_{k-1}A)^j T_{k-1}A.$$

The application of the previous iterative scheme to the integral equation (3.1) allows us to construct a sequence of functions $\{y_m(x)\}_{m\geq 0}$ that generalizes the previous ones. Actually, for $p \geq 2$, it is given by

$$\begin{cases} T_0 \in \mathcal{L}(C[a,b], C[a,b]) \text{ given}, \\ y_0(x) = T_0 f(x), \\ T_m = \displaystyle\sum_{j=0}^{p-1} \binom{p}{j+1} (-1)^j (T_{m-1}A)^j T_{m-1}, \quad m \geq 1, \\ y_m(x) = T_m f(x). \end{cases} \tag{2.6}$$

Both local and semilocal convergence of these iterative schemes have been studied in [6], together with an analysis of the starting points for the application of these iterative schemes considered.

3. Picard-type iterations

Another procedure to approximate a solution of the integral equation (1.1) is to write it as a functional equation defined between two Banach spaces and to consider different iterative schemes for solving it. So, we introduce the function $F : \Omega \subseteq C[a,b] \longrightarrow C[a,b]$, where Ω is a nonempty convex domain in $C[a,b]$, with

$$F(y)(x) = y(x) - \lambda \int_0^1 K(x,t)y(t)\,dt - f(x). \tag{3.1}$$

We are interested in solving the equation $F(y) = 0$ by means of iterative schemes. Observe that a solution of this equation is a solution of equation (1.1) and vice versa. So, starting from an initial approximation of y^*, a solution of the equation $F(y) = 0$, a sequence $\{y_n\}$ of approximations is constructed such that $\lim_n y_n = y^*$. Note that the sequence $\{\|y_n - y_{n-1}\|\}$ is strictly decreasing and, at every step, a better approximation to the solution y^* is obtained.

We can obtain the sequence of approximations $\{y_n\}$ by different ways, depending on the considered iterative scheme. To approximate such a solution we can apply the well-known method of successive approximations [7], which is also known as Picard's method [4, 8] when it comes to approximate a solution of the equation $F(y) = 0$ and defined by $y_{n+1} = y_n - F(y_n)$, $n \geq 0$, with y_0 given in $\Omega \subseteq C[a,b]$. This iterative scheme has the drawback of its linear convergence, however it is an iterative scheme that does not use derivatives or inverses of operators.

A commonly used iterative scheme is the best-known is Newton's method, whose algorithm is the following:

$$\begin{cases} y_0 \text{ given in } \Omega, \\ y_{n+1} = y_n - [F'(y_n)]^{-1} F(y_n), \quad n = 0, 1, 2 \ldots \end{cases} \tag{3.2}$$

In practice it is not easy to construct the iterative scheme (3.2) for operators defined on infinite dimension spaces. The main difficulties arise, at each step, in the calculus of the inverse of the linear operator $F'(y_n)$ or, equivalently, in solving the associated linear equation. Next, we study the value of the inverse operator of $F'(y)$. So, by applying certain algebraic manipulations [2], we can obtain that:

$$[F'(y)]^{-1}\phi(\xi) = \phi(\xi) + \lambda[\mathcal{I} - \lambda\mathcal{K}]^{-1}\mathcal{K}\phi(\xi), \text{ for } \phi \in C[a,b] \text{ and } \xi \in C[a,b].$$

Now, as a consequence of the last equation, we can rewrite an iteration of the iterative scheme of Newton (2.1) as follows

$$\begin{aligned} y_{n+1}(x) &= y_n(x) - F(y_n)(x) - \lambda[\mathcal{I} - \lambda\mathcal{K}]^{-1}\mathcal{K}F(y_n)(x) \\ &= \lambda\mathcal{K}y_n(x) + f(x) - \lambda[\mathcal{I} - \lambda\mathcal{K}]^{-1}\mathcal{K}F(y_n)(x). \end{aligned}$$

Notice that, if $|\lambda|\|\mathcal{K}\| < 1$, by Banach Lemma, it follows that there exists the operator $(\mathcal{I} - \lambda\mathcal{K})^{-1}$.

Now, as in the previous section 2.1, we approximate $(\mathcal{I} - \lambda\mathcal{K})^{-1}$ by means the Newton sequence, then we define the following Ulm-type algorithm

$$\begin{cases} z_0 \text{ and } N_0 \text{ given in } \Omega \text{ and } \mathcal{L}(C[a,b], C[a,b]) \text{ respectively,} \\ z_{k+1} = \lambda \mathcal{K} z_k + f - \lambda N_k \mathcal{K} F(z_k), \\ N_{k+1} = 2N_k - N_k A N_k \quad n = 0, 1, 2 \dots \end{cases} \tag{3.3}$$

where, as Section 1, we have donoted: $\mathcal{I} - \lambda \mathcal{K} = A$.

Therefore, we present a Ulm-type iterative scheme that shares many properties of Picard iterative scheme, namely it is derivative-free and does not use inverse operators, although preserving the quadratic order of convergence that characterizes Newton's method. These features allow us to design an efficient iterative method. Actually, with a very reduced number of iterations, we can find competitive approximations to the solution of the involved Fredholm integral equation (1.1). This is one of the main targets of our research: to justify that it is enough to consider a few steps in our iterative procedure to reach a good approach to the solution.

Acknowledgements

The authors have been partially supported by Grant PGC2018-095896-B-C21 of the Spanish Ministerio de Ciencia, Innovación y Universidades.

References

[1] S. Amat, J. A. Ezquerro and M. A. Hernández-Verón. Approximation of inverse operators by a new family of high-order iterative methods. *Numer. Linear Algebra Appl.*, 21: 629–644, 2014.

[2] J. A. Ezquerro and M. A. Hernández-Verón. A modification of the convergence conditions for Picard's iteration. *Comp. Appl. Math.*, 23: 55–65, 2004.

[3] J. A. Ezquerro and M. A. Hernández-Verón. An optimization of Chebyshev's method. *J. Complexity*, 25 (4): 343–361, 2009.

[4] A. M. A. El-Sayed, H. H. G. Hashem and E. A. A. Ziada. Picard and Adomian methods for quadratic integral equation. *Comput. Appl. Math.* 29 (3): 447–463, 2010.

[5] J. M. Gutiérrez and M. A. Hernández-Verón. A Picard-type iterative scheme for Fredholm integral equations of second kind. *Mathematics*, 9 (1) 83: 1–15, 2021.

[6] J. M. Gutiérrez, M. A. Hernández-Verón and E. Martínez. Improved iterative solution of linear Fredholm integral equations of second kind via inverse-free iterative schemes. *Mathematics*, 8 (10): 1747: 1–15, 2020.

[7] S. Regmi. Optimized iterative methods with applications in diverse disciplines. *Nova Science Publisher*, New York, 2021.

[8] S. M. Zeynalli, S. V. Abbasova and A. G. Gurbanova. Application of the Picard method of sequential integration in differential inequalities. *Adv. Math. Model. Appl.*, 3 (2): 164–167, 2018.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 220–227)

High-order well-balanced methods for systems of balance laws based on collocation RK ODE solvers

I. Gómez-Bueno[1], M.J. Castro[1], C. Parés[1], G. Russo[2]
1. *University of Málaga, Spain*
2. *University of Catania, Italy*

Abstract

The aim of this work is to develop high-order well-balanced schemes for 1d systems of balance laws. A general methodology for developing such numerical methods was proposed by two of the authors that requires the computation, at every cell and at every time step, of the stationary solution whose cell average is equal to the numerical approximation already obtained. Since solving these problems can be difficult and expensive, our goal is to introduce a general procedure that can be applied to any one-dimensional system of balance laws based on the application of collocation RK methods to approximate the stationary solution with given cell-average.

1. Introduction

Let us consider 1d systems of balance laws of the form

$$U_t + F(U)_x = S(U)H_x, \tag{1.1}$$

where $U(x,t)$ takes value in $\Omega \subset \mathbb{R}^N$, $F : \Omega \to \mathbb{R}^N$ is the flux function; $S : \Omega \to \mathbb{R}^N$; and H is a known function from $\mathbb{R} \to \mathbb{R}$ (possibly the identity function $H(x) = x$), which is supposed to be continuous. We suppose that system (1.1) is strictly hyperbolic, i.e., the Jacobian $J(U)$ of the flux function has N real distinct eigenvalues $\lambda_1(U), \cdots, \lambda_N(U)$ and associated eigenvectors $r_1(U), \cdots, r_N(U)$. Moreover, we suppose that $\lambda_i(U) \neq 0$, $i = 1, \ldots, N$.

The system (1.1) has nontrivial stationary solutions that satisfy the ODE system:

$$F(U)_x = S(U)H_x, \tag{1.2}$$

or

$$J(U)U_x = S(U)H_x. \tag{1.3}$$

A numerical method is said to be well-balanced if it preserves (in some sense) stationary solutions. This property is important when the waves generated by small perturbations of an equilibrium are to be simulated: numerical errors should not break the equilibrium. The research on the idea of constructing numerical schemes that preserve some equilibria has been developed by many authors: see, for instance, [2] , [1], [3], [11], [16], [18], [20], [21], [22], [24], [14], [8], [15], [9], [7]. See [6] and their references for a recent review on this topic.

We consider high-order finite-volume numerical methods for (1.1) of the form:

$$\frac{d\widetilde{U}_i}{dt} = -\frac{1}{\Delta x}\left(F_{i+\frac{1}{2}}(t) - F_{i-\frac{1}{2}}(t)\right) + \frac{1}{\Delta x}S_i, \tag{1.4}$$

where:

- $\widetilde{U}_i(t)$ is the approximation of the average of the exact solution at the ith cell, $I_i = \left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right]$, at time t:

$$\widetilde{U}_i(t) \cong \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} U(x,t)\,dx,$$

  where the length of the cells Δx is supposed to be constant for simplicity;

- $F_{i+\frac{1}{2}} = \mathbb{F}(U^{t,-}_{i+\frac{1}{2}}, U^{t,+}_{i+\frac{1}{2}})$, where $\mathbb{F}$ is a consistent numerical flux and $U^{t,\pm}_{i+\frac{1}{2}}$ are the reconstructed states at the intercells:

$$U^{t,-}_{i+\frac{1}{2}} = P^t_i(x_{i+\frac{1}{2}}), \quad U^{t,+}_{i+\frac{1}{2}} = P^t_{i+1}(x_{i+\frac{1}{2}}).$$

  Here $P^t_i(x)$ is the approximation of the solution at the ith cell given by a reconstruction operator of order p applied to the sequence of cell values $\{\widetilde{U}_i(t)\}$:

$$P^t_i(x) = P_i(x; \{\widetilde{U}_j(t)\}_{j\in\mathcal{S}_i}),$$

  where $\mathcal{S}_i$ denotes the set of indices of the cells belonging to the stencil of the cell I_i;

- finally,

$$S_i \approx \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} S(P_i^t(x))H_x(x)\,dx. \tag{1.5}$$

Given a function U, the following notation

$$\bar{U}_i = \frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} U(x)\,dx, \quad \widetilde{U}_i \approx \bar{U}_i, \quad U^{i+1/2} \approx U(x_{i+1/2}), \quad \forall i,$$

is used to denote its cell averages, the approximations to its cell averages and its point values at the intercells, respectively.

2. Well-balanced numerical methods

The key point in [4], [5] is to transfer the well-balanced property to the reconstruction operator:

Definition 2.1 Given a stationary solution U^*, the reconstruction operator is said to be well-balanced for U^* if

$$P_i(x) = U^*(x), \quad \forall x \in [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}],\ \forall i, \tag{2.1}$$

where P_i is the approximation of U^* given by the reconstruction operator from the vector $\{\bar{U}_i^*\}$ of cell-averages of U^*.

One can easily prove that the numerical method (1.4) with

$$S_i = \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} S(P_i^t(x))H_x(x)\,dx, \tag{2.2}$$

is *exactly well-balanced* if the reconstruction operator is well-balanced for every stationary solution U^*, which means that the vector of its cell-averages $\{\bar{U}_i^*\}$ (or its approximations $\{\widetilde{U}_i^*\}$ if a quadrature formula is used to compute them) is an equilibrium of (1.4).

However, in general a standard reconstruction operator is not expected to be well-balanced. The following algorithm allows us to design a well-balanced reconstruction operator P_i on the basis of a standard operator Q_i, provided that Q_i is exact for the null function (see [4]):

Algorithm 2.2 *Given a family of cell values $\{\bar{U}_i\}$, at every cell I_i:*

1. *Find, if possible, the stationary solution $U_i^*(x)$ in the stencil of cell I_i such that*:

$$\frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} U_i^*(x)\,dx = \bar{U}_i. \tag{2.3}$$

   Otherwise, take $U_i^* \equiv 0$.

2. *Apply the reconstruction operator to the cell values $\{V_j\}_{j\in S_i}$ given by*

$$V_j = \bar{U}_j - \frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} U_i^*(x)\,dx, \quad j \in \mathcal{S}_i,$$

   to obtain:

$$Q_i(x) = Q_i(x; \{V_j\}_{j\in\mathcal{S}_i}).$$

3. *Define*

$$P_i(x) = U_i^*(x) + Q_i(x). \tag{2.4}$$

Another difficulty may come from the use of quadrature formulas to compute the cell-averages and the integral of the source term at the right-hand side of (1.4). In this case, the numerical method is still well-balanced if:

- the quadrature formula is also applied to compute the integrals appearing in the first two steps of Algorithm 2.2;

- S_i is computed as follows:

$$S_i = F\left(U_i^{t,*}(x_{i+\frac{1}{2}})\right) - F\left(U_i^{t,*}(x_{i-\frac{1}{2}})\right) + \Delta x \sum_{m=1}^{s} b_m \left(S(P_i^t(x_i^m)) - S(U_i^{t,*}(x_i^m))\right) H_x(x_i^m), \tag{2.5}$$

  where $U_i^{t,*}$ is the stationary solution found in the first step of the reconstruction procedure at the ith cell and time t, and x_i^m, b_m, $m = 1, \dots, s$ are respectively the nodes and the weights of the selected quadrature formula, whose order of accuracy is bigger or equal than p.

Notice that, at every cell and at every time step, the following nonlinear problem has to be solved:
Find U such that

$$J(U)U_x = S(U)H_x, \quad \frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} U(x)\,dx = \widetilde{U}_i, \tag{2.6}$$

where $\widetilde{U}_i$ is an approximation of the cell-average at the ith cell of the solution of (1.1) that we are looking for. In addition, once the solution U_i^* of (2.6) has been found, one has to solve two Cauchy problems in order to extend it to the cells belonging to the stencil. Specifically, (1.3) with initial condition $U(x_{i+1/2}) = U^*(x_{i+1/2})$ has to be solved forward in space and (1.3) with final condition $U(x_{i-1/2}) = U^*(x_{i-1/2})$ backward in space.

Solving these local nonlinear problems can be difficult if the analytic expression of the solutions of (1.3) are not known either in explicit or implicit form. We propose here to approximate their solutions by solving the following numerical problems:

Problem 2.3 (Local problem (LP)) *Given an index i and a state $\widetilde{U}_i \in \Omega$, find approximations*

$$U_{i,j}^{*,m},\ m = 1, \dots, s,\ j \in \mathcal{S}_i; \quad U_i^{*,i\pm 1/2};$$

of the values

$$U_i^*(x_j^m),\ m = 1, \dots, s,\ j \in \mathcal{S}_i; \quad U_i^*(x_{i\pm 1/2});$$

where U_i^* is the stationary solution that satisfies

$$\sum_{m=1}^{s} b_m U_i^*(x_i^m) = \widetilde{U}_i. \tag{2.7}$$

The numerical methods issues from this strategy are not expected to be exactly well-balanced, but they will be well-balanced according to the following definition:

Definition 2.4 The numerical method (1.4) is said to be well-balanced with order $q \geq p$ if for every stationary solution U^* of (1.1) and for every Δx, there exists an equilibrium $\{\widetilde{U}_{\Delta x,i}^*\}$ of (1.4) such that

$$\bar{U}_i^* = \widetilde{U}_{\Delta x,i}^* + O(\Delta x^q), \quad \forall i. \tag{2.8}$$

The sequence $\{\widetilde{U}_{\Delta x,i}^*\}$ is said to be a discrete stationary solution.

3. RK Collocation methods

We propose here to solve the local problems (LP) using a RK collocation method with Butcher tableau

$$\begin{array}{c|ccc} c_1 & a_{1,1} & \dots & a_{1,s} \\ c_2 & a_{2,1} & \dots & a_{2,s} \\ \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s,1} & \dots & a_{s,s} \\ \hline & b_1 & \dots & b_s. \end{array}$$

Remember that, given a Cauchy problem

$$\begin{cases} U_x = G(x, U), \\ U(x_{i_0-1/2}) = U^{i_0-1/2}, \end{cases} \tag{3.1}$$

and a uniform mesh of nodes $x_{i+1/2} = x_{i-1/2} + \Delta x$, $i = i_0, i_0+1, \dots$, the numerical solutions are computed as follows:

$$U^{i+1/2} = U^{i-1/2} + \Delta x \Phi_{\Delta x}(U^{i-1/2}), \quad i = i_0, i_0+1, \dots \tag{3.2}$$

where

$$\Phi_{\Delta x}(U^{i-1/2}) = \sum_{j=1}^{s} b_j K_i^j.$$

$K_i^1, \ldots, K_i^s$ solve the nonlinear system

$$K_i^j = G\left(x_i^j, U^{i-1/2} + \Delta x \sum_{l=1}^{s} a_{j,l} K_i^l\right), \quad j = 1, \ldots, s, \tag{3.3}$$

where

$$x_i^j = x_{i-1/2} + c_j \Delta x, \quad j = 1, \ldots, s. \tag{3.4}$$

Gauss-Legendre methods will be considered here, in which $x_i^1, \ldots, x_i^s$ and $b_1, \ldots, b_s$ are respectively the quadrature points and the weights of the Gauss quadrature formula in the interval $[x_{i-1/2}, x_{i+1/2}]$. This quadrature formula will be used to compute the averages at the cells.

The key-point of collocation methods is that they can be interpreted as follows:

$$U^{i+1/2} = P_i(x_{i+1/2}),$$

where P_i is the only polynomial of degree s that satisfies:

$$\begin{cases} P_i(x_{i-1/2}) = U^{i-1/2}, \\ P_i'(x_i^j) = G(x_i^j, P_i(x_i^j)), \quad j = 1, \ldots, s. \end{cases} \tag{3.5}$$

Because of this interpretation, it can be shown that these methods are symmetric or reversible in the following sense (see [13]):

$$\Phi_{\Delta x} \circ \Phi_{-\Delta x} = Id, \quad \text{or equivalently} \quad \Phi_{\Delta x} = \Phi_{-\Delta x}^{-1}. \tag{3.6}$$

Let us describe how these methods are used to solve the local problems. Given a cell I_i, let us suppose that its stencil is

$$\mathcal{S}_i = \{i - l, \ldots, i + r\}.$$

The local problem solver based on the collocation RK methods is then as follows:

Algorithm 3.1 *Numerical solver for the local problems (LP) using collocation RK methods.*

- *Find $U^{i-1/2}$, $K_i^1, \ldots, K_i^s$ such that*

$$\begin{cases} J\left(U_i^m\right) K_i^m = S\left(U_i^m\right) H_x(x_i^m), \quad m = 1, \ldots, s, \\ \\ \displaystyle\sum_{m=1}^{s} b_m U_i^m = \widetilde{U}_i, \end{cases}$$

  where

$$U_i^m = U^{i-1/2} + \Delta x \sum_{k=1}^{s} a_{m,k} K_i^k, \quad m = 1, \ldots, s.$$

- *Compute*:

$$U^{i+1/2} = U^{i-1/2} + \Delta x \sum_{m=1}^{s} b_m K_i^m.$$

- *The approximated solution is then obtained at the rest of the stencil from the values at the intercell by applying the RK collocation method backward and forward in space.*

The output of the local solver with the notation of (LP) is then:

$$U_{i,j}^{*,m} = U_j^m, \ m = 1, \ldots, s, \ j = i - l, \ldots i + r; \quad U_i^{*,i-1/2} = U^{i-1/2}, \quad U_i^{*,i+1/2} = U^{i+1/2}.$$

It can be shown that the approximations of Cauchy problems (3.1) with $G(x, U) = J(U)^{-1} S(U) H_x$ using the RK collocation method are discrete stationary solution of the numerical schemes, what proves that they are well-balanced with order $2s$. The reversibility of RK collocation methods plays a crucial role in the proof. Notice that there are not explicit methods which have this property.

Method	Error ($i = 1$)	Error ($i = 2$)	Error ($i = 3$)
SMi	1.34E-3	2.43E-6	1.74E-8
CDWBMi	5.37E-15	5.15E-16	2.51E-14

Tab. 1 Test 1. Errors in L^1 norm for SMi and CDWBMi ($i = 1, 2, 3$) with respect to the stationary solution for the 200-cell mesh at time $t = 5s$.

4. Numerical experiments

The following choices have been made in order to build the well-balanced schemes introduced in this paper:

- For the first and second order well-balanced numerical schemes, the second-order 1-stage Gauss-Legendre collocation method is applied, whereas the 2-stage Gauss-Legendre collocation method is used for the third order schemes.
- The midpoint rule is considered for first and second order schemes, and the 2-point Gauss quadrature rule for third order schemes.
- The Rusanov numerical flux is considered.
- We apply the trivial reconstruction operator for the first order scheme; the MUSCL reconstruction for the second order scheme (see [23]); and the CWENO reconstruction for the third order scheme (see [17], [10]).
- First, second and third order TVD Runge-Kutta methods are used for solving the ODE system (1.4): see [12].

The following notation is introduced to denote the methods considered:

- SMi, $i = 1, 2, 3$: numerical method of order i based on the Rusanov flux and the standard, not well-balanced, reconstruction operators.
- CDWBMi, $i = 1, 2, 3$: numerical method of order i based on the Rusanov flux and the well-balanced reconstruction operator in which the discrete stationary solutions and local problems are obtained by applying the Gauss-Legendre collocation method as described in the previous section.

The numerical experiments have been performed in a computer equipped with Intel(R) Xeon(R) CPU E3-1220 v3 @ 3.10GHz with 8Mb cache using one single core.

4.1. Test 1: Burgers equation with a nonlinear source term

We consider the Burgers equation with the non-linear source term $S(U) = \sin(U)$:

$$\begin{cases} U_t + \left(\dfrac{U^2}{2}\right)_x = \sin(U), \quad x \in \mathbb{R},\ t > 0, \\ U(x, 0) = U_0(x). \end{cases} \tag{4.1}$$

We consider $x \in [-1, 1]$, $t \in [0, 5]$ and CFL= 0.9. As initial condition, we consider the stationary solution which solves the Cauchy problem

$$\begin{cases} \dfrac{dU}{dx} = \dfrac{\sin(U)}{U}, \\ U(-1) = 2. \end{cases}$$

$U(-1, t) = 2$ is imposed at $x = -1$ and free boundary conditions are considered at $x = 1$.

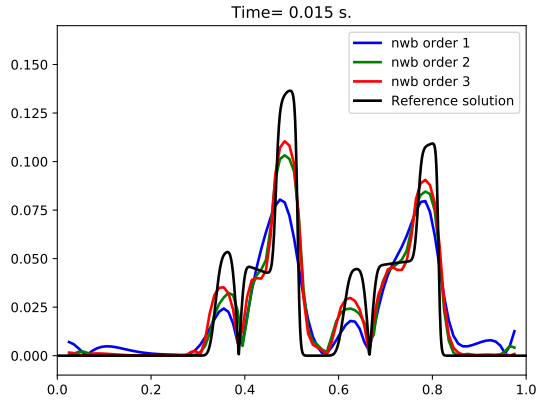
Table 1 shows the errors corresponding to SMi and CDWBMi, $i = 1, 2, 3$ respectively for a 200-cell mesh. As expected, only the well-balanced methods preserve the stationary solutions.

4.2. Test 2: shallow water equations with Manning friction

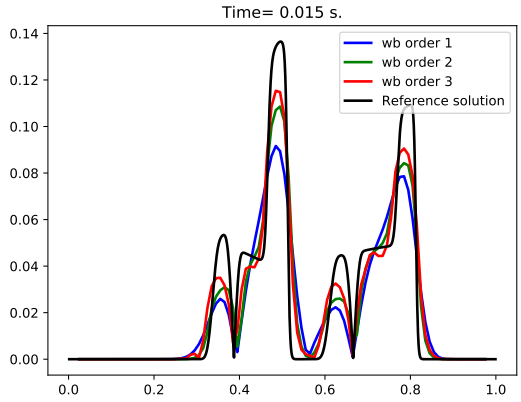
Let us consider the shallow water equations with Manning friction:

$$\begin{cases} h_t + q_x = 0, \\ q_t + \left(\dfrac{q^2}{h} + \dfrac{1}{2} g h^2\right)_x = g h H_x - \dfrac{k q |q|}{h^\eta}. \end{cases} \tag{4.2}$$

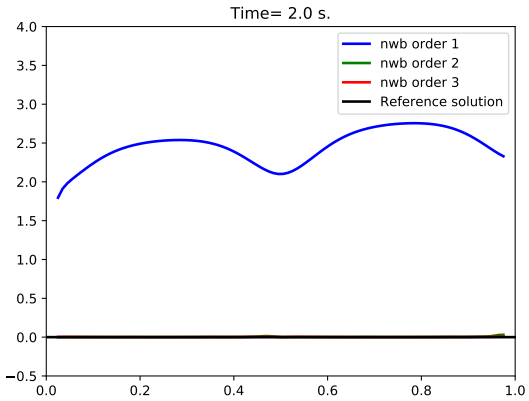
This system is used to model the flow of water in a one-dimensional channel,with a bottom that applies a friction force on the water. Here, the variable x makes reference to the axis of the channel and t is the time; $q(x, t)$ and



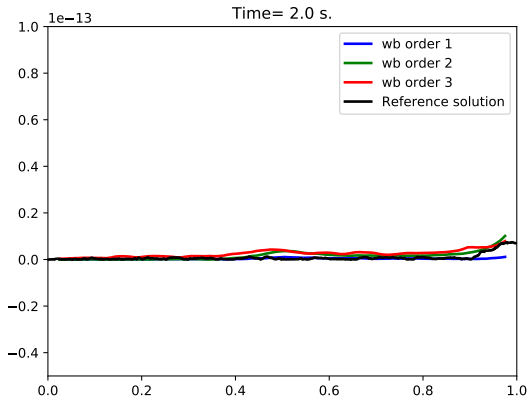
(a) SMi, $i = 1, 2, 3$. $t = 0.015s$.



(b) CDWBMi, $i = 1, 2, 3$. $t = 0.015s$.



(c) SMi, $i = 1, 2, 3..$ $t = 2s$.



(d) CDWBMi, $i = 1, 2, 3..$ $t = 2s$.

Fig. 1 Test 2. Differences between the stationary solution and the reference and numerical solutions at times $t = 0.015, 2s$ for h. Number of cells: 100.

$h(x,t)$ are the discharge and the thickness, respectively; $u = q/h$ is the depth-averaged velocity; g is the gravity; $H(x)$ is the depth function measured from a fixed reference level; k is the Manning friction coefficient; and η is a parameter equal to $\frac{7}{3}$.

Following [19], we consider $x \in [0, 1]$, $k = 0.01$ and the depth function

$$H(x) = 1 - \frac{1}{2}\frac{e^{\cos(4\pi x)} - e^{-1}}{e - e^{-1}}. \tag{4.3}$$

The initial condition $U_0(x) = [h_0(x), q_0(x)]^T$ is

$$h_0(x) = \begin{cases} h^*(x) + 0.05, & \text{if } x \in \left[\frac{2}{7}, \frac{3}{7}\right] \cup \left[\frac{4}{7}, \frac{5}{7}\right], \\ h^*(x), & \text{otherwise,} \end{cases} \quad q_0(x) = \begin{cases} q^*(x) + 0.5, & \text{if } x \in \left[\frac{2}{7}, \frac{3}{7}\right] \cup \left[\frac{4}{7}, \frac{5}{7}\right], \\ q^*(x), & \text{otherwise,} \end{cases} \tag{4.4}$$

where $U^*(x) = [h^*(x), q^*(x)]^T$ is the supercritical stationary solution that satisfies $q(0) = 1$ and $h(0) = 0.3$. The numerical simulation is run until $t = 2s$ using a uniform mesh with 100 cells.

Figure 1 shows the differences between the stationary solution and the numerical solutions at times $t = 0.015$ and $2s$ with SMi, $i = 1, 2, 3$ and CDWBMi, $i = 1, 2, 3$ for h (the graphs are similar for q). A reference solution has been computed with a first order well-balanced scheme on a fine mesh (1600 cells). As expected, only the well-balanced methods are able to recover the stationary solutions. This is clear in Table 2 where the errors in L^1 norm with respect to the stationary solution at time $t = 2s$ are shown for the 100-cell mesh.

Method	Error ($i = 1$)		Error ($i = 2$)		Error ($i = 3$)	
	h	q	h	q	h	q
SMi	2.42	6.12	3.57E-3	4.87E-3	1.39E-3	4.30E-4
CDWBMi	3.73E-16	3.60E-16	1.80E-15	1.99E-15	2.64E-15	8.93E-15

Tab. 2 Test 2. Errors in L^1 norm for SMi and CDWBMi ($i = 1, 2, 3$) with respect to the stationary solution for the 100-cell mesh at time $t = 2s$.

5. Conclusions

Following the methodology introduced in [4], we have described a general strategy in order to build a family of high-order well-balanced numerical methods that can be applied to general 1d systems of balance laws. Due to the difficulty of solving the ODE satisfied by the stationary solutions, sometimes the first step of the reconstruction procedure is required to be numerically solved: the application of the collocation RK methods to deal with these problems have been detailed in this work. The numerical methods have been applied to some systems of balance laws what allows us to check that the well-balanced property is fulfilled.

Acknowledgements

This work has been partially supported by the Spanish Government and FEDER through the coordinated Research project RTI2018-096064-B-C1, The Junta de Andalucía research project P18-RT-3163 and the Junta de Andalucia-FEDER-University of Málaga Research project UMA18-FEDERJA-161. G. Russo acknowledges partial support from ITN-ETN Horizon 2020 Project ModCompShock, "Modeling and Computation of Shocks and Interfaces", Project Reference 642768, and from the Italian Ministry of University and Research (MIUR), PRIN Project 2017 (No. 2017KKJP4X entitled "Innovative numerical methods for evolutionary partial differential equations and applications". I. Gómez-Bueno is also supported by a Grant from "El Ministerio de Ciencia, Innovación y Universidades", Spain (FPU2019/01541).

References

[1] Audusse, E., Bouchut, F., Bristeau, M.-O., Klein, R., and Perthame, B. A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. *SIAM Journal on Scientific Computing 25*, 6 (2004), 2050–2065.

[2] Bermudez, A., and Vazquez, M. E. Upwind methods for hyperbolic conservation laws with source terms. *Computers & Fluids 23*, 8 (1994), 1049–1071.

[3] Bouchut, F. *Nonlinear stability of finite Volume Methods for hyperbolic conservation laws: And Well-Balanced schemes for sources*. Springer Science & Business Media, 2004.

[4] Castro, M., Gallardo, J., López-García, J., and Parés, C. Well-balanced high order extensions of godunov's method for semilinear balance laws. *SIAM J. Numerical Analysis 46* (01 2008), 1012–1039.

[5] Castro, M., and Parés, C. Well-balanced high-order finite volume methods for systems of balance laws. *Journal of Scientific Computing 82* (02 2020).

[6] Castro, M. J., de Luna, T. M., and Parés, C. *Well-balanced schemes and path-conservative numerical methods*, vol. 18. Elsevier, 2017, pp. 131–175.

[7] Chandrashekar, P., and Zenk, M. Well-balanced nodal discontinuous galerkin method for euler equations with gravity. *Journal of Scientific Computing 71*, 3 (2017), 1062–1093.

[8] Cheng, Y., and Kurganov, A. Moving-water equilibria preserving central-upwind schemes for the shallow water equations. *Communications in Mathematical Sciences 14* (01 2016), 1643–1663.

[9] Chertock, A., Cui, S., Kurganov, A., and Wu, T. Well-balanced positivity preserving central-upwind scheme for the shallow water system with friction terms. *International Journal for Numerical Methods in Fluids 78* (04 2015).

[10] Cravero, I., and Semplice, M. On the accuracy of weno and cweno reconstructions of third order on nonuniform meshes. *Journal of Scientific Computing 67* (03 2015).

[11] Gómez-Bueno, I., Castro, M. J., and Parés, C. High-order well-balanced methods for systems of balance laws: a control-based approach. *Applied Mathematics and Computation 394* (2021), 125820.

[12] Gottlieb, S., and Shu, C.-W. Total variation diminishing runge-kutta schemes. *Mathematics of Computation 67* (08 1996).

[13] Haier, E., Lubich, C., and Wanner, G. *Geometric Numerical integration: structure-preserving algorithms for ordinary differential equations*. Springer, 2006.

[14] Klingenberg, C., Puppo, G., and Semplice, M. Arbitrary order finite volume well-balanced schemes for the euler equations with gravity. *SIAM Journal on Scientific Computing 41* (01 2019), A695–A721.

[15] KURGANOV, A. Finite-volume schemes for shallow-water equations. *Acta Numerica 27* (05 2018), 289–351.

[16] KÄPPELI, R., AND MISHRA, S. Well-balanced schemes for the euler equations with gravitation. *Journal of Computational Physics 259* (2014), 199 – 219.

[17] LEVY, D., PUPPO, G., AND RUSSO, G. Compact central weno schemes for multidimensional conservation laws. *Siam Journal on Scientific Computing 22* (09 2000).

[18] LUKACOVA-MEDVIDOVA, M., NOELLE, S., AND KRAFT, M. Well-balanced finite volume evolution galerkin methods for the shallow water equations. *Journal of Computational Physics 221* (01 2007), 122–147.

[19] MICHEL-DANSAC, V., BERTHON, C., CLAIN, S., AND FOUCHER, F. A well-balanced scheme for the shallow-water equations with topography. *Computers & Mathematics with Applications 72* (09 2015).

[20] NOELLE, S., PANKRATZ, N., PUPPO, G., AND NATVIG, J. Well-balanced finite volume schemes of arbitrary order of accuracy for shallow water flows. *Journal of Computational Physics 213* (04 2006), 474–499.

[21] NOELLE, S., XING, Y., AND SHU, C.-W. High-order well-balanced finite volume weno schemes for shallow water equation with moving water. *Journal of Computational Physics 226* (09 2007), 29–58.

[22] RUSSO, G., AND KHE, A. High order well balanced schemes for systems of balance laws. *Proceedings of Symposia in Applied Mathematics 67* (01 2009), 919–928.

[23] VAN LEER, B. Towards the ultimate conservative difference scheme v. a second-order sequel to godunov's method. *Journal of Computational Physics 32* (07 1979), 101–136.

[24] XING, Y., AND SHU, C.-W. High order well-balanced finite volume weno schemes and discontinuous galerkin methods for a class of hyperbolic systems with source terms. *Journal of Computational Physics 214* (05 2006), 567–598.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 228–235)

An algorithm to create conservative Galerkin projection between meshes

P. Gómez-Molina[1], L. Sanz-Lorenzo[2], J. Carpio[1]
1. Departamento de Ingeniería Energética
2. Departamento de Matemática Aplicada a la Ingeniería Industrial
E.T.S.I. Industriales, José Gutiérrez Abascal, 2, 28006 Madrid
Universidad Politécnica de Madrid

Abstract

We present in this paper an algorithm to solve pure-convection problems with a conservative Lagrange-Galerkin formulation in the framework of the finite element method. The integrals obtained from the Lagrange-Galerkin formulation will be computed with an algorithm which leads to conservation of mass up to machine accuracy, when we transfer information from the mesh moved by the characteristic curves of the convection operator to the current mesh. The algorithm to compute the integrals considers the intersection of meshes composed by triangles (2-dimensions) and tetrahedra (3-dimensions) with straight sides. We will illustrate the good features of the method in terms of stability, accuracy and mass conservations in different pure-convection tests with non-divergence-free velocity fields.

1. Introduction

Nowadays, in the resolution of problems related with fluids, such as aerodynamics, combustion and heat transfer, we usually find convection-dominated equations. However, its resolution via finite element methods is not straightforward, since the treatment of the convective terms is a source of numerical problems due to the fact that the standard Galerkin formulation is unstable.

One methodology that brings about the stabilization of the convective term in a natural way is related to the Lagrangian description of the flow. Here, we use the information of the characteristic curves of the convection operator in order to integrate the equation in time. In the so-called Lagrange-Galerkin method (also known as Characteristic-Galerkin or semi-Lagrange-Galerkin method), see [1, 2], we identify each domain point $\mathbf{x}$ as a fluid particle at time t_n and seek backward in time the position of this particle at time t_{n-1}, that we call the foot of the characteristic curve $\mathbf{X}(\mathbf{x}, t_n; t_{n-1})$, where the numerical solution $u_h(\mathbf{x}, t_{n-1}))$ is known. The set of the feet of the characteristic curves defines a backwards convected mesh, and the weak formulation of the problem performs a L^2-projection of the known solution from this convected mesh to the fixed mesh.

In the context of Lagrange-Galerkin schemes, Colera et al. [3] derived a conservative Lagrange-Galerkin formulation to solve pure-convection and convection-diffusion equations in the case of non-divergence-free velocity fields. The method is mainly based on formulating a conservation integral equation for a weighted mass, that can be discretized in time and in space with any order of accuracy, and is posed so that the terms that appear in the formulation can be easily computed by means of standard finite element operations.

Although the weak formulation in [3] leads to mass conservation, the right-hand side of the equation consists on an integral of functions that are defined in different element spaces (associated with the current triangulation and its backwards convected mesh). In [3], the integrals are computed with high-order quadrature rules [4], which is the reason the method proposed there is named "nearly-conservative". Since the basis functions of the fixed mesh are not polynomials, but only piecewise polynomials, over the elements of the convected mesh, the use of quadrature rules over such elements does not produce an exact result.

In this work, we propose an algorithm based on an appropriate mesh intersection procedure to accurately compute the right-hand side integral of the weak formulation. This techniques leads to better accuracy in the mass conservation property and also it improves the stability properties of the Lagrange-Galerkin scheme. Following Farrell et al. [5] we call this technique "supermesh technique".

The layout of the paper is as follows: Section 2 starts with the presentation of the conservative Lagrange-Galerkin formulation of pure convection problems with non-divergence-free velocity fields and concludes with the weak formulation of the problem in the framework of the finite element method. Section 3 constitutes the core of the paper, where we explain the numerical procedure to implement the "supermesh technique" to compute the integral of the right hand side term of the weak formulation of the problem, to transfer information from the moved mesh to the current mesh. In Section 4 we present numerical results to show the good properties of the proposed algorithm, in terms of stability, accuracy and mass conservation. Finally, some conclusions and comments are collected in Section 5.

2. Conservative Lagrange-Galerkin formulation for pure-convection problems

Let us consider the conservative form of a pure-convection equation for the scalar variable $u = u(\mathbf{x}, t)$

$$\begin{cases} \dfrac{\partial u}{\partial t} + \nabla \cdot (\mathbf{a}u) = 0 \quad \text{in } \Omega \times (0, T] \\ u(\mathbf{x}, 0) = u^0(\mathbf{x}) \quad \text{in } \Omega, \end{cases} \tag{2.1}$$

with $\Omega \subset \mathbb{R}^d$ (with $d = 2, 3$) a bounded domain with smooth boundary $\partial\Omega$, and $\mathbf{a}(\mathbf{x}, t)$ a regular velocity field with possible non-null divergence (we do not assume incompressible velocity). To solve numerically the problem (2.1) by means of a Lagrange-Galerkin scheme we have to divide the time interval $\overline{I} = [0, T]$ with a constant step size $\Delta t = t_n - t_{n-1}$. Associated with the velocity field $\mathbf{a}(\mathbf{x}, t)$ we can define the characteristic curves $\mathbf{X}(\mathbf{x}, t_n; t)$ of the convective or material derivative operator $D/Dt = \partial/\partial t + \mathbf{a} \cdot \nabla$ that correspond to the position backward in time of a fluid particle at time $t \leq t_n$ that will reach the domain point $\mathbf{x}$ at instant of time t_n. $\mathbf{X}(\mathbf{x}, t_n; t)$ is the solution to the system of equations

$$\begin{cases} \dfrac{d\mathbf{X}(\mathbf{x}, t_n; t)}{\partial t} = \mathbf{a}\left(\mathbf{X}(\mathbf{x}, t_n; t), t\right) \quad t < t_n, \\ \mathbf{X}(\mathbf{x}, t_n; t_n) = \mathbf{x}. \end{cases} \tag{2.2}$$

To obtain the weak conservative Lagrangian-Galerkin formulation of problem (2.1) we multiply the equation by a test function $v = v(\mathbf{x}, t)$ that satisfies the equation $Dv/Dt = 0$. Then, we obtain the expression

$$\frac{\partial (uv)}{\partial t} + \nabla \cdot (\mathbf{a}uv) = 0. \tag{2.3}$$

Now, we integrate (2.3) in the domain $\widetilde{\Omega}(t)$, that evolves backward in time from t_n to $t < t_n$ according to the velocity field $\mathbf{a}(\mathbf{x}, t)$ and it is defined by the family of characteristic curves $\mathbf{X}(\mathbf{x}, t_n; t)$

$$\widetilde{\Omega}(t) := \left\{ \mathbf{X} \in \mathbb{R}^d : \ \mathbf{X} = \mathbf{X}(\mathbf{x}, t_n; t), \ \mathbf{x} \in \Omega \right\}$$

with $\widetilde{\Omega}(t_n) = \Omega$. Therefore, applying the Gauss theorem to the second term followed by the Reynolds theorem, we obtain the weak formulation of equation (2.3) as a temporal derivative over a integral extended to the fluid volume $\widetilde{\Omega}(t)$. For all test functions we have

$$\frac{d}{dt} \int_{\widetilde{\Omega}(t)} uv dX = 0 \quad \rightarrow \quad \int_{\Omega} u^n(\mathbf{x}) v^n(\mathbf{x}) dx = \int_{\widetilde{\Omega}(t_{n-1})} u^{n-1}(\mathbf{X}) v^{n-1}(\mathbf{X}) dX, \tag{2.4}$$

2.1. Finite element discretizacion and convected finite element space

The equation (2.4) forces us to consider an integration domain $\widetilde{\Omega}(t)$ that moves with the fluid particles, as well as test functions $v(\mathbf{x}, t)$ that remain constant along the fluid trajectories. Then, we chose as integration domain at instant of time t_n the set $\Omega_h := \widetilde{\Omega}(t_n)$, a polygonal domain that approximates Ω and over which we define a regular triangulation $\mathbb{T}_h$ composed of triangles (in 2D) or tetrahedra (in 3D).

Associated with the triangulation $\mathbb{T}_h$ we define a conforming finite element space V_h where the numerical solution $u_h^n(\mathbf{x})$ (shorthand for $u_h(\mathbf{x}, t_n)$) is computed. To do so, we consider a reference element $\hat{K} \in \mathbb{R}^d$ and define $\hat{V}_h$ as the space P_m of polynomial functions of degree less or equal to m and denote it dimension by n_v. For each element $K \in \mathbb{T}_h$ we define the one-to-one affine mapping $F_K : \hat{K} \rightarrow K$

$$\mathbf{F}_K : \hat{K} \longrightarrow K, \ \mathbf{x} = J_K \hat{\mathbf{x}} + \mathbf{b}_K, \quad J_K \in \mathbb{R}^{d \times d} \text{ and } \mathbf{b}_K \in \mathbb{R}^d. \tag{2.5}$$

and we denote V_h the resulting conforming finite element space and N_v the number of mesh nodes.

Now, each element $K \in \mathbb{T}_h$ which composes domain Ω_h is convected backwards in time from t_n to t_{n-1} obtaining $\widetilde{K}^{n-1}$ as the geometric place of the so-called feet of the characteristic curves $\mathbf{X}(\mathbf{x}, t_n; t_{n-1})$ with $\mathbf{x} \in K$, for which we use the shorthand $\mathbf{X}^{n-1}(\mathbf{x})$. We are going to consider an approximation $\widetilde{K}_h^{n-1}$ of $\widetilde{K}^{n-1}$ given by the following isoparametric transformation

$$\widetilde{\mathbf{F}}_K^{n-1} : \widehat{\mathbf{x}} \in \hat{K} \longrightarrow \mathbf{X}_h \in \widetilde{K}_h^{n-1}, \quad \mathbf{X}_h^{n-1}(\hat{\mathbf{x}}) = \sum_{i=1}^{n_v} \mathbf{X}^{n-1}(\mathbf{v}_i) \hat{\varphi}_i(\hat{\mathbf{x}}), \tag{2.6}$$

with $\mathbf{v}_i$ the coordinates of the i-th local node of K and $\hat{\varphi}_i$ the i-th elemental nodal basis function of $\hat{V}_h$. The transformation $\widetilde{\mathbf{F}}_K^{n-1}$ incurs an error in the approximation of $\mathbf{X}^{n-1}(\mathbf{x})$ consistent with the space finite element discretization, i.e.,

$$|\mathbf{X}^{n-1}(\mathbf{x}) - \mathbf{X}_h^{n-1}(\mathbf{x})| = O(h^{m+1}),$$

Note that $\mathbf{X}^{n-1}(\mathbf{v}_i)$ must be computed for the N_v mesh nodes $\{\mathbf{v}_i\}_{i=1}^{N_v}$ solving numerically the differential equation (2.2). Moreover, the transformation $\widetilde{\mathbf{F}}_K^{n-1}$ has the advantage that $\mathbf{X}_h^{n-1}(\mathbf{x})$ and $\mathbf{x}$ share the same natural coordinates in the reference element [1, 6], that is,

$$\mathbf{X}^{n-1}(\mathbf{x}) \simeq \mathbf{X}_h^{n-1}(\mathbf{x}) := \widetilde{\mathbf{F}}_K^{n-1}\left(\mathbf{F}_K^{-1}(\mathbf{x})\right). \tag{2.7}$$

For each $K \in \mathbb{T}_h$, the definition of the isoparametric transformation $\widetilde{\mathbf{F}}_K^{n-1}$ of (2.6) leads to an element $\widetilde{K}_h^{n-1}$ with polynomial edges of degree m, the order of the finite element space approximation $\hat{V}_h$, and which is an aproximation of the real convected element $\widetilde{K}^{n-1}$. Let $\widetilde{\mathbb{T}}_h^{n-1}$ be the mesh composed by these approximate convected elements $\widetilde{\mathbb{T}}_h^{n-1} := \left\{\widetilde{K}_h^{n-1} : K \in \mathbb{T}_h\right\}$. In the present paper we are going to consider linear finite elements, $m = 1$, in order to have approximated convected elements $\widetilde{K}_h^{n-1}$ with straight edges (sides in 3D). We can see Fig. 1 for an explanatory scheme of the above explanation.

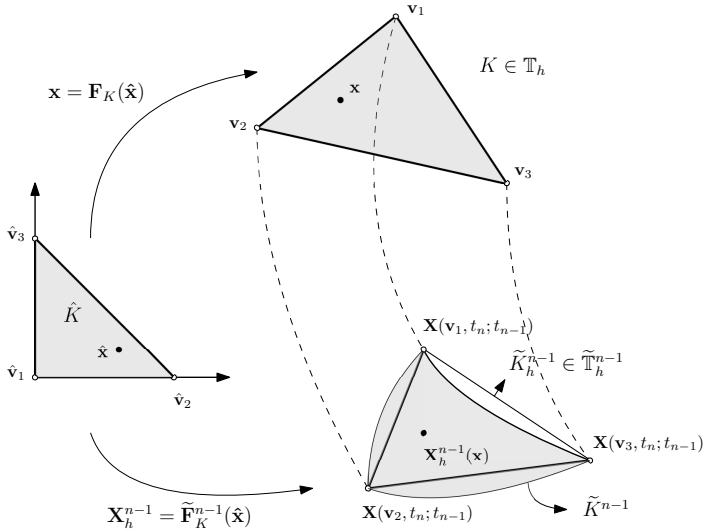


Fig. 1 Convected nodes $\mathbf{X}^{n-1}(\mathbf{v}_i)$ backward in time from triangle $K \in \mathbb{T}_h$. From the convected nodes, we approximate the real convected element by an isoparametric $\widetilde{K}_h^{n-1} \in \widetilde{\mathbb{T}}_h^{n-1}$ element via application $\mathbf{X}_h^{n-1}(\mathbf{x}) = \widetilde{\mathbf{F}}_K^{n-1}\left(\mathbf{F}_K^{-1}(\mathbf{x})\right)$.

Now, let us consider a function $v_h(\mathbf{x}, t)$ which verifies $Dv_h/Dt = 0$ and $v_h^n(\mathbf{x}) \in V_h$. Since v_h is constant along the fluid trajectories, we can make the approximation $v_h^{n-1}\left(\mathbf{X}_h^{n-1}(\mathbf{x})\right) = v_h^n(\mathbf{x})$. Moreover, v_h^{n-1} belongs to the convected finite element space

$$\widetilde{V}_h^{n-1} := \left\{v_h^{n-1} : v_h^{n-1}\left(\mathbf{X}_h^{n-1}(\mathbf{x})\right) = v_h^n(\mathbf{x}) \in V_h\right\},$$

and hence $\widetilde{V}_h^{n-1}$ is also a P_m space, but associated to the mesh $\widetilde{\mathbb{T}}_h^{n-1}$ instead of $\mathbb{T}_h$.

Finally, with the definition of Ω_h, $\widetilde{\Omega}_h^{n-1}$ and their associated triangulations $\mathbb{T}_h$, $\widetilde{\mathbb{T}}_h^{n-1}$ and the finite element spaces V_h, $\widetilde{V}_h^{n-1}$, we can define the numerical approximation of the weak conservative formulation (2.4) via the finite element method

$$\int_{\Omega_h} u_h^n v_h^n dx = \int_{\widetilde{\Omega}_h^{n-1}} u_h^{n-1}(\mathbf{X}) v_h^{n-1}(\mathbf{X}) dX, \quad \forall v_h^n \in V_h \ \text{ (and its associated function } v_h^{n-1} \in \widetilde{V}_h^{n-1}), \tag{2.8}$$

where the initial value u_h^0 is taken as the L^2 projection of the initial condition $u^0(\mathbf{x})$, i.e.,

$$\int_{\Omega_h} u_h^0 v_h^0 dx = \int_{\Omega_h} u^0 v_h^0 dx, \quad \forall v_h^0 \in V_h. \tag{2.9}$$

The computation of the left-hand side of (2.8) lead us to the standard mass matrix associated to the triangulation $\mathbb{T}_h$ and it is straightforward to compute it, however the right-hand side in (2.8) involves the product of u_h^{n-1}, which is defined piecewise in $\mathbb{T}_h$, and v_h^{n-1}, which is defined piecewise in $\widetilde{\mathbb{T}}_h^{n-1}$ (see Fig. 2). Since these two meshes are usually different, to compute this right-hand side we can follow the following strategies:

1) Integrate over the elements in $\widetilde{\mathbb{T}}_h$ with high-order quadrature rules [4] as is done in Colera et al. [3].

2) Develop a mesh intersection technique to accurately compute this term [5, 7, 8], which leads to better accuracy in mass conservation. In the present work we adopt this approach and derive a technique to develop mesh intersection of straight triangles in 2D and straight tetrahedra in 3D of high efficiency using very conventional operations within the finite element methodology.

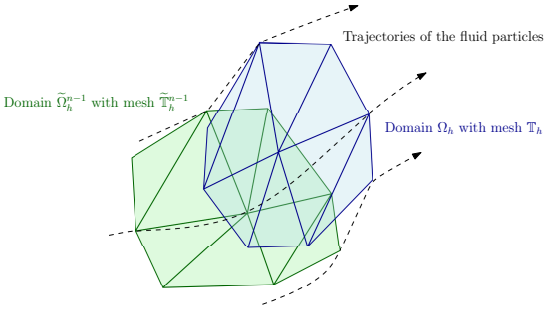


Fig. 2 Scheme that illustrates the domains and meshes that appear in the formulation. The nodes of the elements in $\mathbb{T}_h$ are convected backwards in time with the flow velocity field to form $\widetilde{\mathbb{T}}_h^{n-1}$. The variables u_h^n, v_h^n and u_h^{n-1} are defined piecewise in $\mathbb{T}_h$, whereas v_h^{n-1} is defined piecewise in $\widetilde{\mathbb{T}}_h^{n-1}$.

Note that, if in problem (2.1) the velocity field satisfies $\mathbf{a} \cdot \mathbf{n} = \mathbf{0}$ on the boundary $\partial\Omega_h$, with $\mathbf{n}$ the outward normal vector on the boundary, then the backward convected domain will be the same as the current volume (fluid particles on the boundary $\partial\Omega_h$ do not cross the boundary), i.e., $\Omega_h = \tilde{\Omega}_h^{n-1}$ (but $\mathbb{T}_h \neq \widetilde{\mathbb{T}}_h^{n-1}$ if $\mathbf{a} \neq \mathbf{0}$). That means that the conservation principle is satisfied in the domain Ω_h for all instants of time

$$\int_{\Omega_h} u_h^n(\mathbf{x})dx = \int_{\Omega_h} u_h^0(\mathbf{x})dx, \quad \forall t_n \in [0, T]. \tag{2.10}$$

3. Algorithm for the intersection of meshes with straight elements

To compute (2.8) numerically, we replace the test function v_h^n by each of the N_v basis functions $\varphi_I(\mathbf{x}) \in V_h$ and $u_h^n = \sum_{J=1}^{N_v} U_J^n \varphi_J(\mathbf{x})$. Then, the left hand side reads

$$\int_{\Omega_h} u_h^n(\mathbf{x})\varphi_I(\mathbf{x})dx = \sum_{K \in \mathbb{T}_h} \left(\sum_{J=1}^{N_v} U_J^n \int_K \varphi_I(\mathbf{x})\varphi_J(\mathbf{x})dx \right),$$

which can be computed exactly with a quadrature rule since both $\varphi_I(\mathbf{x})$ and $\varphi_J(\mathbf{x})$ are basic function in V_h. However, for the right hand side we have

$$\int_{\widetilde{\Omega}_h^{n-1}} u_h^{n-1}(\mathbf{X}^{n-1}(\mathbf{x}))\varphi_I^{n-1}(\mathbf{X}^{n-1}(\mathbf{x}))dX = \sum_{\widetilde{K}_h^{n-1} \in \widetilde{\mathbb{T}}_h^{n-1}} \left(\sum_{J=1}^{N_v} U_J^{n-1} \int_{\widetilde{K}_h^{n-1}} \varphi_I^{n-1}(\mathbf{X}^{n-1}(\mathbf{x}))\varphi_J(\mathbf{X}^{n-1}(\mathbf{x}))dX \right), \tag{3.1}$$

and inside the integral there are two kinds of basis functions: $\varphi_J \in V_h$ associated with the elements $K \in \mathbb{T}_h$ and $\varphi_I^{n-1} \in \widetilde{V}_h^{n-1}$ associated with the elements $\widetilde{K}_h^{n-1} \in \widetilde{\mathbb{T}}_h^{n-1}$. Therefore, as the basis functions are piecewise polynomials over each element of their respective meshes, the use of intersection techniques is required to compute this right hand side exactly with quadrature rules, as can be seen in Fig. 2.

In this work we propose a mesh intersection algorithm based in [5]. We are going to simplify the notation in accordance with those introduced in that paper. We define $\mathbb{T}_h$ as the donor mesh, and denote donor elements $K_D \equiv K \in \mathbb{T}_h$, whereas we define as $\widetilde{\mathbb{T}}_h^{n-1}$ the target mesh, so that the target elements are $K_T \equiv \widetilde{K}_h^{n-1} \in \widetilde{\mathbb{T}}_h^{n-1}$. Moreover, as we have considered in Fig. 1 and Fig. 2, we are going to illustrate the main stages of the algorithm with figures of meshes composed of triangles in two-dimensions. For tetrahedra in three-dimensions the situation is analogous, but spatial figures are more difficult to understand. Then, the algorithm to compute integral (3.1) has the following stages:

1. For each $K \in \mathbb{T}_h$, find the set of elements $K_D^T := \{K_{D_1}, K_{D_2}, ...\} \subset \mathbb{T}_h$ such that $K_{Di} \bigcap K_T \neq \emptyset$ with $K_T \in \widetilde{\mathbb{T}}_h^{n-1}$. Those elements are shown in Fig. 3.

2. Compute the supermesh associated with the intersection $K_{Di} \bigcap K_T$ for each $K_{D_i} \in K_D^T$. The supermesh is the set $K_{TD_i} = \{K_{TD_i,1}, K_{TD_i,2}, ...\}$ where K_{TD_i} is included in both K_T and K_{D_i}. Fig. 4 shows a scheme of this stage of the algorithm. To create a triangulation of the intersection zone there are several procedures, and one of the simplest (useful for convex polytopes in two and three dimension) is the Sutherland-Hodgman clipping algorithm [9].

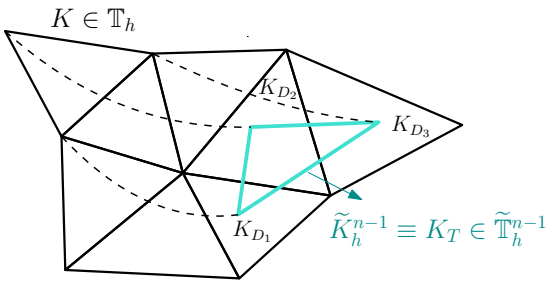


Fig. 3 The element $K_T \in \widetilde{\mathbb{T}}_h^{n-1}$ (convected from $K \in \mathbb{T}_h$) intersects with the triangles of the set $K_D^T = \{K_{D_1}, K_{D_2}, K_{D_3}\} \in \mathbb{T}_h$.

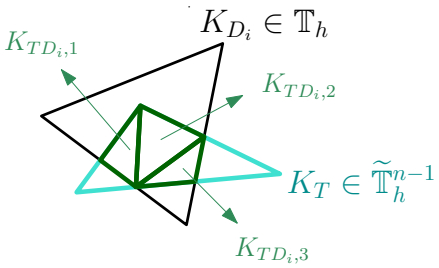


Fig. 4 The intersection of K_T and K_{D_i} is remeshed to obtain the supermesh. In this example that supermesh has three elements: $K_{TD_i,1}$, $K_{TD_i,2}$ and $K_{TD_i,3}$

3. Define the linear transformations $\hat{\mathbf{x}}_{D_i} = \mathbf{g}_{D_iT,j}(\hat{\mathbf{x}})$ and $\hat{\mathbf{x}}_T = \mathbf{g}_{TD_i,j}(\hat{\mathbf{x}})$ between a common reference element and a reference element associated to K_{D_i} and K_T, respectively. This is shown in Fig. 5

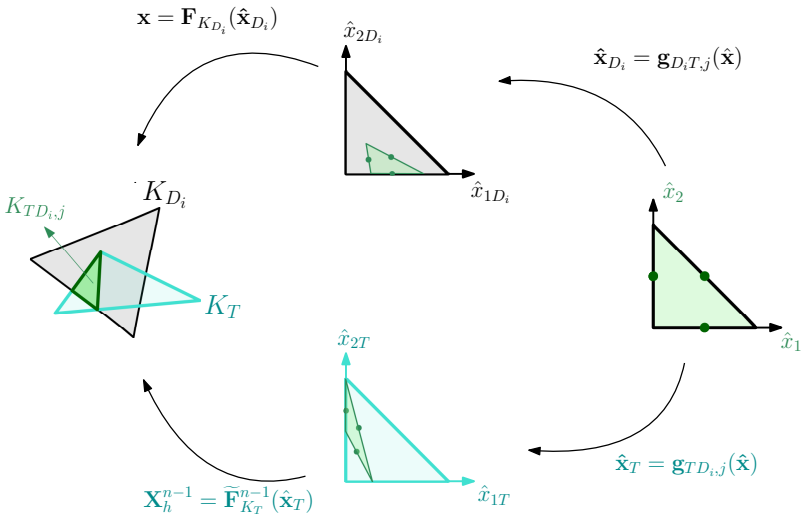


Fig. 5 Linear transformations from the standard reference element $\hat{K}$ to those reference elements $\hat{K}_{D_i}$ and $\hat{K}_T$ associated with the proper elements K_{D_i} and K_T, respectively.

4. Finally, the integral $m_{IJ}|_{K_T} := \int_{\widetilde{K}_h^{n-1}} \varphi_I^{n-1}\left(\mathbf{X}^{n-1}(\mathbf{x})\right) \varphi_J\left(\mathbf{X}^{n-1}(\mathbf{x})\right) dX$ can be computed as:

$$m_{IJ}|_{K_T} = \sum_{K_{D_i} \in K_D^T} \left[\sum_{K_{TD_i,j} \in K_{TD_i}} \int_{K_{TD_i,j}} \varphi_I^{n-1}\left(\mathbf{X}^{n-1}(\mathbf{x})\right) \varphi_J\left(\mathbf{X}^{n-1}(\mathbf{x})\right) dX \right]$$

with

$$\int_{K_{TD_i,j}} \varphi_I^{n-1}\left(\mathbf{X}^{n-1}(\mathbf{x})\right) \varphi_J\left(\mathbf{X}^{n-1}(\mathbf{x})\right) dX = \sum_{s=1}^{n_q} \omega_s \hat{\varphi}_I\left(\mathbf{g}_{TD_i,j}(\hat{\boldsymbol{\xi}}_s)\right) \hat{\varphi}_J\left(\mathbf{g}_{D_iT,j}(\hat{\boldsymbol{\xi}}_s)\right) \det\left(\frac{\partial \widetilde{\mathbf{F}}_{K_T}(\hat{\mathbf{x}}_T)}{\partial \hat{\mathbf{x}}_T}\right) \det\left(\frac{\partial \mathbf{g}_{TD_i,j}(\hat{\mathbf{x}})}{\partial \hat{\mathbf{x}}}\right)$$

where $\hat{\boldsymbol{\xi}}_s$, $s = 1, ..., n_q$ are the quadrature points in the reference element and the product $\det\left(\frac{\partial \widetilde{\mathbf{F}}_{K_T}(\hat{\mathbf{x}}_T)}{\partial \hat{\mathbf{x}}_T}\right) \det\left(\frac{\partial \mathbf{g}_{T D_i,j}(\hat{\mathbf{x}})}{\partial \hat{\mathbf{x}}}\right)$ equals the size of the supermesh element, $\left|K_{T D_i,j}\right|$

4. Numerical test

Next, we present a numerical test to illustrate the performance of the numerical algorithm. Note that the trajectories of the fluid particles can be computed with high accuracy or analytically for these problem, which allows us to obtain the exact solution of pure-convection problems (2.1) through the formula $u(\mathbf{x}, T) = u^0\left(\mathbf{X}(\mathbf{x}, T; 0)\right) \det\left(\frac{\partial \mathbf{X}(\mathbf{x}, T; 0)}{\partial \mathbf{x}}\right)$ [10], which means

$$u(\mathbf{x}, T) = u^0(\mathbf{X}(\mathbf{x}, T; 0)) \exp\left(-\int_0^T \left[(\nabla \cdot \mathbf{a})\right]_{\mathbf{X}(\mathbf{x},T;t)} dt\right).$$

We are going to show numerical results for the projection technique presented in this paper via intersection of meshes. This technique is called "supermesh projection" as opposed to the "standard projection" where we use high-order quadrature rules to compute the integrals with high accuracy. For the supermesh projection we need quadrature rules of order only two (both in 2-dimension ($n_q = 3$) and 3-dimension ($n_q = 4$)), the minimum order needed to integrate exactly the corresponding product of basis functions. For the standard projection we show in figures the number of quadrature points used in each numerical simulation.

The test consists of a pure convection problem (2.1) with the following velocity field and initial condition:

$$\mathbf{a}(\mathbf{x}, t) = \left[0.45 + \sin(t - x_1),\ 0.45 + \sin(t - x_2)\right]^{\mathrm{T}}, \quad \text{and} \quad u^0(\mathbf{x}) = \exp\left(-200\left(2 - \cos(x_1) - \cos(x_2)\right)\right).$$

in a domain $\Omega_h = [-1, 1] \times [-1, 1]$ and final instant of time T=0.5. In this problem $\mathbf{a} \cdot \mathbf{n} \neq \mathbf{0}$, but the solution on the boundary for all instant of time $t \leq T$ is negligible $u|_{\partial\Omega} \simeq 0$ and then the mass of the solution in the domain Ω_h is almost maintained (2.10).

The evolution of the solution with time can be seen in Fig.6. At time $t = 0$ we have a gaussian hill in the middle of the domain. When the time goes on, the solution moves along the line $y = x$ and its width is reduced and the value of its vertex is increased, to satisfy the conservation of mass at all instant of times following (2.10).

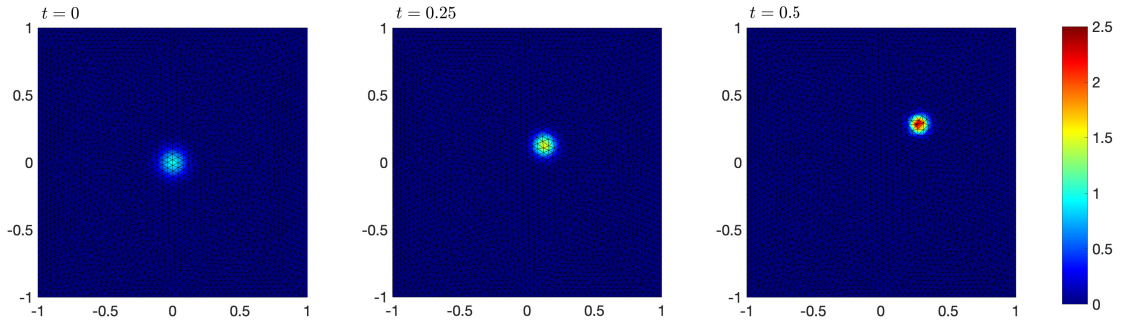


Fig. 6 Numerical solution of test at different instants of time in a uniform triangulation $\mathbb{T}_h$ with $h = 0.085$.

Now, we are going to measure the error in the L^2-norm between the numerical solution and the exact solution and also the mass error, both at the last instant of time. We consider a time step size $\Delta t = 0.005$ and different meshes composed of regular elements of size h. The results can be seen in Fig.7. The L^2 error for supermesh projection and standard projection is similar in both cases, and also its convergence with h is close to the theoretical one shown in the black dashed line ($O(h^2)$ for linear finite elements). However, the difference between supermesh projection and standard projection is remarkable in terms of mass conservation. We can observe in Fig. 7 that supermesh projection is up to twelve orders of magnitude more accurate when computing the mass error.

Finally, in Fig.8 we show an interesting phenomenon when we analyze the L^2-error at the last instant of time, as a function of time step size Δt. Now, we are going to consider a uniform mesh with with $h = 0.085$ and make several numerical simulations with different time step sizes Δt. It is known that standard projection may get unstable when using small time steps if the number of quadrature points used for integration is not high enough. However, for the supermesh projection, this instability does not appear, since the integrals are computed exactly with low number of quadrature points (the minimum needed to integrate exactly the corresponding product between basis functions).

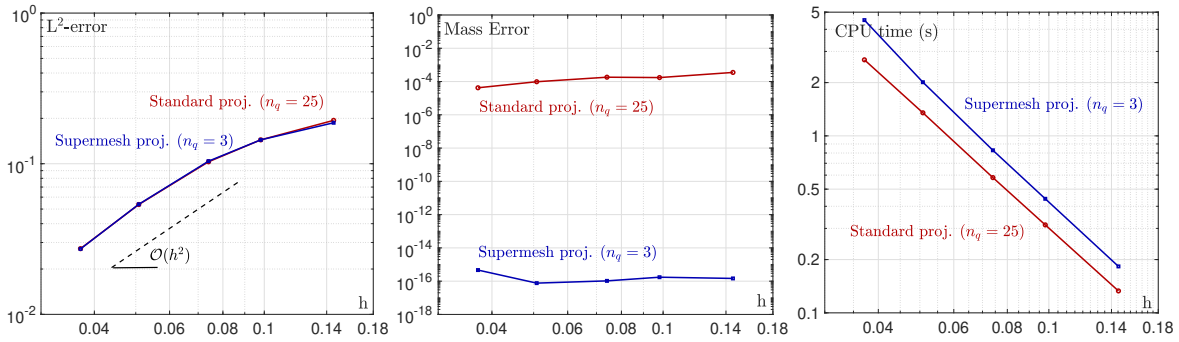


Fig. 7 L^2-error (on the left), mass error (in the middle), and computational requirement in terms of CPU time (on the right) for different uniform meshes with size h and $\Delta t = 0.005$ maintained constant in all the simulations.

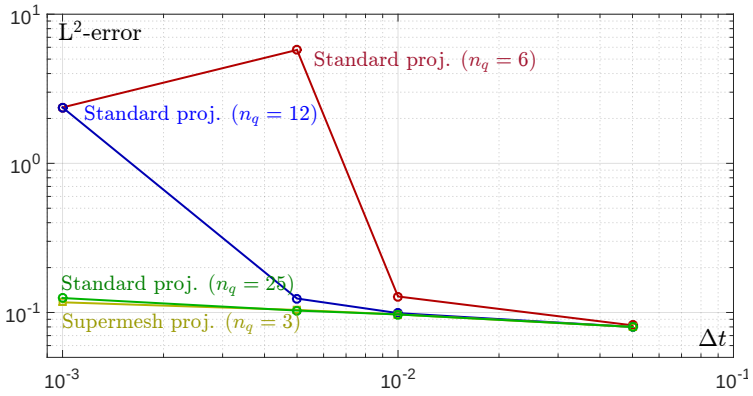


Fig. 8 L^2-error for different time steps and the same mesh with $h = 0.085$.

5. Conclusion

In this work we described in detail an algorithm to compute exactly the integrals that appear in the conservative Lagrange-Galerkin formulation of a pure-convection problem in a linear finite element framework. This technique is called "supermesh projection" as opposed to the "standard projection" where we use high-order quadrature rules to compute the integrals with high accuracy. Both techniques have almost identical L^2-error with moderately large time step sizes Δt and high quadrature rules for the standard projection (for supermesh projection we need quadrature rules of order two, both in 2- and 3-dimensions). However, the main advantage of supermesh projection is that it avoids instabilities when the integrals are computed with enough accuracy in standard projection and when numerical error are accumulated in time (small time step sizes Δt). The main disadvantage of the supermesh projection is that the computational requirement of the right hand side of the weak formulation is larger than standard projection (nearly a factor two when the number of element which conformed the mesh is large). However, in practice this issue is not a significant problem, since the Lagrange-Galerkin formulation usually is used in other more complicated problems as convection-diffusion-reaction equations or Navier-Stokes equations. In these cases, computing the convection terms is usually the less resource consuming step, so this time increase would not be that relevant for the overall process. Moreover, the intersection mesh procedure can be carried out with straightforward parallel programming. As future work, we want to extent the present algorithm of mesh intersection to finite elements of high order ($m > 1$ order). In this case, convected elements $\widetilde{K}_h^{n-1}$ have sides given by a polynomial function of degree m, and the intersection of elements is much more complicated.

Acknowledgments

This research has been partially funded from the "Ministerio de Ciencia, Innovación y Universidades" of Spain and European Regional Development Funds by project PGC-2018-097565-BI00.

References

[1] R. Bermejo, J. Carpio, A semi-Lagrangian-Galerkin projection scheme for convection equations, IMA Journal of Numerical Analysis, Vol. 30(3) (2010) 799-831.

[2] R. Bermejo, P. Galán del Sastre, L. Saavedra, *A second order in time modified Lagrange–Galerkin finite element method for the incompressible navier-stokes equations*, SIAM J. Numer. Anal. 50(6) (2012) 3084–3109.

[3] M. Colera, J. Carpio, R. Bermejo, A nearly conservative high order Lagrange Galerkin method for the resolution of scalar convection-dominated equations in non-divergence-free velocity fields, Computer Methods in Applied Mechanics and Engineering, 372(1) (2020) 113366.

[4] K.W. Morton, A. Priestley, E. Süli, Stability of the Lagrange-Galerkin method with non-exact integration, ESAIM: Math. Model. Numer. Anal. 22(4) (1988) 625-653.

[5] P. Farrell, J. Maddison, Conservative interpolation between volume meshes by local Galerkin projection, Comput. Methods Appl. Mech. Engrg., 200 (2011) 89-100.

[6] J. Carpio, J.L. Prieto, M. Vera, Local anisotropic adaptive algorithm for the solution of low-Mach transient combustion problems, J. Comput. Phys. 306 (2016) 19-42.

[7] F. Alauzet, A parallel matrix-free conservative solution interpolation on unstructured tetrahedral meshes, Comput. Methods Appl. Mech. Engrg., 299 (2016) 116-142.

[8] M. Tabata, S. Uchiumi, A genuinely stable Lagrange-Galerkin scheme for convection-diffusion problems, Japan J. Indust. Appl. Math, 33 (2016) 121-143.

[9] I.E. Sutherland, G.W. Hodgman, Reentrant polygon clipping, Commun. ACM 17(1) (1974) 32-42.

[10] R. Bermejo, J. Conde, A conservative Quasi-Monotone Semi-Lagrangian Scheme, Mon. Wea. Rev., 130 (2002) 423-430.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 236–241)

On iterative schemes for matrix equations

M.A. Hernández-Verón and N. Romero[1]
E-mail: mahernan@unirioja.es, natalia.romero@unirioja.es. Universidad de La Rioja, Spain

Abstract

In this work we focus on solving quadratic matrix equations. We start by transforming the quadratic matrix equation into a fixed point equation. From this transformation, we propose an iterative scheme of stable successive approximations. We study the global convergence of this iterative scheme. In addition, we obtain a result of restricted global convergence to the well-known Picard method using a technique of auxiliary points. From the results obtained, we analyze the location and separation of the solutions of the quadratic matrix equation considered. Finally, we build a hybrid iterative scheme, predictor-corrector, which allow us to approximate a solution of the quadratic matrix equation more efficiently.

1. Introduction

The study of quadratic matrix equation is motivated by the great variety of problems where appears. Quadratic matrix equation arises in many areas of scientific computing and engineering applications. For instance, algebraic Riccati equations arising in control theory [8]. Another important class of quadratic matrix equations is motivated by noisy Wiener-Hopf problems for Markov chains [9].

Although some algebraic Riccati equations are quadratic matrix equation, and vice versa, the two classes of equations require different techniques for analysis and solution in general.

In this study we are interested in the simplest quadratic matrix equation:

$$\mathbb{Q}(X) = X^2 - BX - C = 0, \qquad B, C \in \mathbb{R}^{m\times m}, \tag{1.1}$$

which occurs in a variety of applications, for example, it may arise in the well known quadratic eigenvalue problem:

$$Q(\lambda)x = \lambda^2 Ax + \lambda Bx + Cx = 0, \text{ with } A, B, C \in \mathbb{C}^{m\times m},$$

that arises in the analysis of structural systems and vibration problems [10].

The application of iterative schemes is commonly used to approximate a solution of equation (1.1). We obtain qualitative results about the equation at issue from the study of the convergence. For instance, a solution existence result is obtained for equation (1.1), with the so-called existence ball of an iterative scheme given in [1], which allows us to locate a solution. On the other hand, a result of uniqueness of the solution allows us to separate solutions [2]. Finally, the iterative scheme considered, under the convergence conditions obtained, allows us to approximate a solution of equation (1.1). This is how the three main aims of our work arise: locate, separate and approximate a solution of equation (1.1).

The paper is organized as follows. In Section 2, we present different conditions to locate and separate solutions of equation (1.1) from the study of the convergence of the Successive Approximations and Picard methods. In Section 3, we define a hybrid iterative scheme to approximate a solution of equation (1.1).

2. The Successive Approximations and Picard Methods

In what follows, we suppose that there exists X^* a fixed matrix of T with $T(X) = X$, $T : \mathbb{R}^{m\times m} \to \mathbb{R}^{m\times m}$ in $\overline{B(X^*, R)}$. In this case, we use the following modification of the Banach Fixed Point Theorem.

Theorem 2.1 *If $\Omega \subset \mathbb{R}^{m\times m}$ is convex and compact and $T : \Omega \to \Omega$ is a contraction, then T admits a unique fixed matrix in Ω and it can be approximated from $X_{n+1} = T(X_n)$, $n \geq 0$, for any X_0 given in Ω.*

So, we look for conditions on R so that the Successive Approximations Method is convergent for any starting matrix X_0 in $\overline{B(X^*, R)}$. Thus, we obtain a local convergence result.

Now, we provide a basic technical result whose proof is easily followed taking into account

$$T'(X)Y = -(X - B)^{-1}Y(X - B)^{-1}C.$$

Lemma 2.2 *Let X^* be a fixed matrix of T in $\overline{B(X^*,R)}$ and we suppose that there exists $(X^*-B)^{-1}$ with $\|(X^*-B)^{-1}\| \leq \beta$. For each $X \in \overline{B(X^*,R)}$, with $R < 1/\beta$, are satisfied*:

(*i*) *there exists $(X^*+t(X-X^*)-B)^{-1}$ for $t \in [0,1]$, and $\|(X^*+t(X-X^*)-B)^{-1}\| \leq f_R(t)$, where $f_R(t) = \dfrac{\beta}{1-t\beta R}$,*

(*ii*) $\|T'(X^*+t(X-X^*))\| \leq f_R(t)^2\|C\|$,

(*iii*) $\|T'(X^*+t(X-X^*)-T'(X^*))(X-X^*)\| \leq (f_R(t)^2+f_R(0)^2)\|X-X^*\|\|C\|$.

Now, to apply the modification of the Banach Fixed Point Theorem to operator T, restricted to $\Omega = \overline{B(X^*,R)}$ with $R > 0$, T must be a contraction map of Ω into itself. That happens if $R < \dfrac{1}{\beta} - \sqrt{c}$, where we denote $\|C\| = c$. Notice that, in this case, this implies that $R < \dfrac{1}{\beta}$ and we can prove the following local result.

Theorem 2.3 *Let X^* be a fixed point of T and we suppose that there exists $(X^*-B)^{-1}$ with $\|(X^*-B)^{-1}\| \leq \beta$. If $\beta^2 c < 1$, with $c = \|C\|$, then, the Successive Approximations Method*

$$X_0 \text{ given in } \mathbb{R}^{m\times m}, \quad X_{n+1} = T(X_n), \quad n \geq 0, \tag{2.1}$$

is convergent to X^*, from any starting matrix $X_0 \in \overline{B(X^*,R)}$, with $R \in \left(0, \dfrac{1}{\beta} - \sqrt{c}\right)$. Moreover, X^* is the unique fixed matrix of the operator T in $\overline{B(X^*,R)}$.

We observe that if X^* is a fixed matrix of the operator T, such that there exists $(X^*-B)^{-1}$ with $\|(X^*-B)^{-1}\| \leq \beta$,it follows

$$\|X^*\| = \|T(X^*)\| \leq \|(X^*-B)^{-1}\|\|C\| \leq \beta c.$$

Therefore, we have $X^* \in \overline{B(0,\beta c)}$, where we denote by 0 the null matrix in $\mathbb{R}^{m\times m}$. So, the domain $\overline{B(0,R)}$, with $R \geq \beta c$, can be a convenient domain where to ensure the convergence of the Successive Approximations Method.

Theorem 2.4 *Let $X^* \in \overline{B(0,R)}$ be a fixed matrix of T and we suppose that there exists $(X^*-B)^{-1}$ with $\|(X^*-B)^{-1}\| \leq \beta$. If $\beta^2 c < \dfrac{1}{8}$, and*

$$R \in \begin{cases} \left[\dfrac{1-\sqrt{1-8\beta^2 c}}{4\beta}, \dfrac{1}{2}\left(\dfrac{1}{\beta} - \sqrt{c}\right)\right) & \text{if } \beta^2 c \in \left(0, \dfrac{1}{9}\right), \\ \left[\dfrac{1-\sqrt{1-8\beta^2 c}}{4\beta}, \dfrac{1+\sqrt{1-8\beta^2 c}}{4\beta}\right] & \text{if } \beta^2 c \in \left[\dfrac{1}{9}, \dfrac{1}{8}\right]. \end{cases},$$

Then, from any starting matrix $X_0 \in \overline{B(0,R)}$, the Successive Approximations Method is convergent to X^*. Moreover, X^* is the unique fixed matrix of T in $\overline{B(0,R)}$.

Notice that if there exists B^{-1}, then for $X \in \overline{B(0,R)}$, it follows

$$\|I - (-B^{-1})(X-B)\| \leq \|B^{-1}\|\|X\| \leq \alpha R,$$

with $\|B^{-1}\| \leq \alpha$ and $X \in \overline{B(0,R)}$. Therefore, if $R < 1/\alpha$, then there exists $(X-B)^{-1}$ and $\|(X-B)^{-1}\| \leq \dfrac{\alpha}{1-\alpha R}$ by the perturbation lemma in matrix analysis. From this, we obtain the following restricted global convergence result.

Theorem 2.5 *Suppose that there exists B^{-1}, with $\|B^{-1}\| \leq \alpha$, and $\alpha^2 c \leq 1/4$. Then, from any starting matrix $X_0 \in \overline{B(0,R)}$, with $R \in \left[\dfrac{1-\sqrt{1-4\alpha^2 c}}{2\alpha}, \dfrac{1}{\alpha} - \sqrt{c}\right)$, the Successive Approximations Method is convergent to the fixed matrix X^* of the operator T. Moreover, X^* is the unique fixed matrix of T in $\overline{B(0,R)}$.*

Next, we illustrate the theoretical results obtained above with some examples. Firstly, we examine a simple academic case, where the technique developed can be applied. We consider the particular (QME) with:

$$B = \begin{pmatrix} 2 & 0 \\ 0 & -1 \end{pmatrix}, \quad C = \begin{pmatrix} 2\epsilon(\epsilon-1) & \epsilon(2-3\epsilon) \\ -\epsilon(1+3\epsilon) & \epsilon(2+5\epsilon) \end{pmatrix}, \tag{2.2}$$

where the parameter ϵ is not zero. We find that it has the solution

$$X^* = \begin{pmatrix} \epsilon & -\epsilon \\ -\epsilon & 2\epsilon \end{pmatrix}.$$

For the value $\epsilon = 0.04$, we have $\beta^2 c = 0.16299 < 1$ and, then result of Theorem 2.3 follows immediately. The Successive Approximations Method is convergent to X^*, from any starting matrix $X_0 \in \overline{B(X^*, R)}$, with $R \in (0, 0.564271)$. Moreover, X^* is the unique fixed matrix of the operator T in $\overline{B(X^*, R)}$.

On the other hand, the results of Theorem 2.4 and 2.5 follows for smaller values of the parameter ϵ. For instance, if $\epsilon = 0.025$, then it follows that $\beta^2 c = 0.105551 < 1/8$. Thus, the local result given in Theorem 2.4 states that from any starting matrix $X_0 \in \overline{B(0, R)}$, with $R \in [0.140379, 0.313011]$, the Successive Approximations Method is convergent to X^* and is the unique fixed matrix of T in $\overline{B(0, R)}$. While, in this case, if $\alpha^2 c = 0.113448 < 1/4$, the semilocal result given in Theorem 2.4 states from any starting matrix $X_0 \in \overline{B(0, R)}$, with $R \in [0.116697, 0.296583)$, the Successive Approximations Method is convergent to the fixed matrix X^* of the operator T. Moreover, X^* is the unique fixed matrix of T in $\overline{B(0, R)}$.

It is clear that, from Theorem 2.3 we separate the solution X^* successfully from other possible solutions, despite its poor location. However, the local result obtained in Theorem 2.4 shows a better separation of the solution. On the other hand, the semilocal convergence result obtained in Theorem 2.5 is more applicable than the local result, since it does not need to know X^*. And moreover, the semilocal result is the one that best locates the aforesaid solution.

It is clear that, from Theorem 2.3 we separate the solution X^* successfully from other possible solutions, despite its poor location. However, the local result obtained in Theorem 2.4 shows a better separation of the solution. On the other hand, the semilocal convergence result obtained in Theorem 2.5 is more applicable than the local result, since it does not need to know X^*. And moreover, the semilocal result is the one that best locates the aforesaid solution.

Next, we try to smooth the results obtained. For this, we consider the Picard method:

$$X_0 \text{ given in } \mathbb{R}^{m\times m}, \quad X_{n+1} = P(X_n) = X_n - F(X_n), \quad n \geq 0, \tag{2.3}$$

where $F(X) = (I - T)(X)$, $F : \mathbb{R}^{m\times m} \to \mathbb{R}^{m\times m}$, with $F(X) = X - (X - B)^{-1}C$. Notice that, the iterations obtained by the Picard method are the same as those obtained by the Successive Approximations Method. Both methods are equivalent.

To obtain a global convergence result to the Picard method, we use auxiliary matrices. Moreover, we can establish both semilocal and local convergence results for the Picard method.

Theorem 2.6 *Let $\tilde{X} \in \mathbb{R}^{m\times m}$ such that there exists $(\tilde{X} - B)^{-1}$ with $\|(\tilde{X} - B)^{-1}\| \leq \tilde{\beta}$. We suppose that $\|F(\tilde{X})\| \leq \dfrac{1 + \tilde{\beta}^2 c - 2\tilde{\beta}\sqrt{c}}{\tilde{\beta}}$, with $c = \|C\|$, and $\tilde{\beta}^2 c < 1$. Then, from any starting matrix $X_0 \in B(\tilde{X}, R)$, the Picard method* (2.3) *converges to a solution X^* of equation* (1.1)*. The solution X^* and the iterates X_n belong to $\overline{B(\tilde{X}, R)}$, for $n \geqslant 0$, where*

$$R \in \left[\frac{1 - \tilde{\beta}^2 c + \tilde{\beta}\|F(\tilde{X})\| - \sqrt{\Delta}}{2\tilde{\beta}}, \min\left\{ \frac{1}{\tilde{\beta}} - \sqrt{c}, \frac{1 - \tilde{\beta}^2 c + \tilde{\beta}\|F(\tilde{X})\| + \sqrt{\Delta}}{2\tilde{\beta}} \right\} \right), \tag{2.4}$$

with $\Delta = (1 - \tilde{\beta}^2 c + \tilde{\beta}\|F(\tilde{X})\|)^2 - 4\tilde{\beta}\|F(\tilde{X})\|$. Moreover, X^* is the unique solution of equation (1.1) *in $B(\tilde{X}, \frac{1}{\tilde{\beta}} - \sqrt{c})$.*

Corollary 2.7 *Let X^* be a solution of equation* (1.1) *such that exists $(X^* - B)^{-1}$ with $\|(X^* - B)^{-1}\| \leq \beta$ and $\beta^2 c < 1$. Then, the Picard method* (2.3)*, from any starting at $X_0 \in B(X^*, R)$ converges to X^*, where $R \in \left(0, \dfrac{1}{\beta} - \sqrt{c}\right)$.*

Moreover, X^* is unique in $B\left(X^*, \dfrac{1}{\beta} - \sqrt{c}\right)$.

Next, to obtain a semilocal convergence result for the Picard method, we consider $\tilde{X} = X_0$ from Theorem 2.6.

Corollary 2.8 *Let $X_0 \in \mathbb{R}^{m\times m}$ be such that exists $(X_0 - B)^{-1}$ with $\|(X_0 - B)^{-1}\| \leq \beta_0$. Suppose that $\|F(X_0)\| \leq \dfrac{1 + \beta_0^2 c - 2\beta_0\sqrt{c}}{\beta_0}$, with $c = \|C\|$, and $\beta_0^2 c < 1$. Then, the Picard method* (2.3) *converges to a solution X^* of equation equation* (1.1) *. The solution X^* and the iterates X_n belong to $\overline{B(\tilde{X}, R)}$, for $n \geqslant 0$, where*

$$R \in \left[\frac{1 - \tilde{\beta}^2 c + \tilde{\beta}\|F(X_0)\| - \sqrt{\Delta}}{2\beta_0}, \min\left\{\frac{1}{\beta_0} - \sqrt{c}, \frac{1 - \beta_0^2 c + \beta_0\|F(X_0)\| + \sqrt{\Delta}}{2\beta_0}\right\}\right), \tag{2.5}$$

with $\Delta = (1 - \beta_0^2 c + \beta_0\|F(X_0)\|)^2 - 4\beta_0\|F(X_0)\|$. Moreover, the solution X^* is the unique solution of the equation $F(X) = 0$ in $B\left(X_0, \dfrac{1}{\beta_0} - \sqrt{c}\right)$.

Next, we provide another semilocal convergence result for the Picard method.

Theorem 2.9 *Let $X_0 \in \mathbb{R}^{m\times m}$ such that there exists $(X_0 - B)^{-1}$ with $\|(X_0 - B)^{-1}\| \leq \beta_0$ and $\|F(X_0)\| \leq \eta_0$. We suppose that the scalar equation*

$$\left(1 + \frac{\beta_0^2 c(1 - \beta_0 t)}{1 - \beta_0^2 c - 2\beta_0 t + \beta_0^2 t^2}\right)\eta_0 = t \tag{2.6}$$

has at least one positive solution and we denote by R the smallest positive root. If $R < \dfrac{1}{\beta_0} - \sqrt{c}$ and $\beta_0^2 c < 1$, then, starting at X_0, the Picard method (2.3) *converges to X^* a solution of equation* (1.1)*. Moreover, $X_n, X^* \in \overline{B(X_0, R)}$, for all $n \geq 0$, and X^* is unique in $B\left(X_0, \dfrac{1}{\beta_0} - \sqrt{c}\right)$.*

Next, we illustrate the theoretical results obtained for the Picard method, considering the simple numerical example given in (2.2).

Taking the value $\epsilon = 0.025$, and

$$\tilde{X} = \begin{pmatrix} \epsilon & 0 \\ 0 & \epsilon \end{pmatrix},$$

then there exists $(\tilde{X} - B)^{-1}$ with $\|(\tilde{X} - B)^{-1}\| \leq 1.09917$ and conditions of Theorem 2.6, $\|F(\tilde{X})\| \leq \dfrac{1 + \tilde{\beta}^2 c - 2\tilde{\beta}\sqrt{c}}{\tilde{\beta}}$ and $\tilde{\beta}^2 c < 1$, are satisfied. Thus, from any starting matrix $X_0 \in \overline{B(\tilde{X}, R)}$ with $R \in [0.0699064, 0.608512]$, the Picard method (2.3) converges to a solution X^* of equation (2.2). The solution X^* and X_n belong to $\overline{B(\tilde{X}, R)}$, for $n \geqslant 0$. Moreover, X^* is the unique solution of (QME) in $B(\tilde{X}, 0.608512)$.

In general, both Theorem 2.6 and Theorem 2.9 provide a more precise location of the solution X^* than that obtained by the Successive Approximations Method for which we always obtain balls centered in the null matrix. However, in these results, locating a starting matrix X_0 satisfying the indicated conditions, we locate the solution in a ball centered in the aforesaid X_0.

Now, we choose the starting matrix

$$X_0 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

to compare the results obtained by means of the Successive Approximations and the Picard methods, in Corollary 2.8 and in Theorem 2.9,

So, the hypotheses of Corollary 2.8 with $\|(X_0 - B)^{-1}\| \leq 1.11803$, $\|F(X_0)\| = 0.0895976 \leq \dfrac{1 + \beta_0^2 c - 2\beta_0\sqrt{c}}{\beta_0} = 0.393375$, and $\beta_0^2 c = 0.113448 < 1$ are satisfied. Thus, the Picard method converges to a solution X^* of equation (2.2). The solution X^* and the iterates X_n belong to $\overline{B(X_0, R)}$, for $n \geqslant 0$, with $R \in [0.104128, 0.593166)$ and X^* is unique in $B(\tilde{X}, 0.593166)$.

On the other hand, equation (2.6) has at least one positive solution and the smallest positive root is $R = 0.103032$, such that, $R < \dfrac{1}{\beta_0} - \sqrt{c} = 0.593166$. Thus, starting at X_0, the Picard method converges to X^* a solution of equation (2.2). Moreover, $X_n, X^* \in \overline{B(X_0, 0.10303)}$, for all $n \geq 0$.

As we can observe, we have considered the null matrix as the starting point, the same matrix when applying the Successive Approximations Method. Notice that, the location and the separation of solutions given by the existence and uniqueness ball, respectively, is improved when we apply the Picard method. Namely, both Corollary 2.8 and Theorem 2.9 improve the results obtained to the Successive Approximations Method in Theorems 2.4 and 2.5.

3. Predictor-corrector scheme

Our next goal is the approximation of a solution of equation (1.1). Although both the Successive Approximations and the Picard methods have a linear convergence speed, their applications are useful. This is due to the fact that they have a low operational cost and good accessibility domain associated with them. Now, we propose to build a hybrid iterative scheme through a predictor-corrector method. That is, a hybrid method consisting of two stages. The idea is, firstly to apply a method which has a good accessibility and low operational cost and later, in a second stage, to apply a method that accelerates the convergence as follows:

$$\begin{cases} \begin{cases} X_0 \in \mathbb{R}^{m\times m}, \\ X_{n+1} = \Phi(X_n), \quad n = 1, 2, \ldots, N_0, \end{cases} \\ \begin{cases} Y_0 = X_{N_0+1}, \\ Y_{n+1} = \Psi(Y_n), \quad n \geqslant 0, \end{cases} \end{cases} \tag{3.1}$$

from any two one-point iterative schemes:

$$\begin{cases} X_0 \in \mathbb{R}^{m\times m}, \\ X_{n+1} = \Phi(X_n), \quad n \geqslant 0, \end{cases} \quad \text{and} \quad \begin{cases} Y_0 \in \mathbb{R}^{m\times m}, \\ Y_{n+1} = \Psi(Y_n), \quad n \geqslant 0. \end{cases}$$

The first iterative scheme to be applied Φ, is called the predictor iterative scheme and the second Ψ, the corrector iterative scheme. It is known that high-order iterative schemes have a reduced accessibility domain and, therefore, locating starting points for them is a difficult problem to solve. Therefore, we propose that the hybrid scheme (3.1) be convergent under the conditions that the iterative predictor scheme is. In our case we consider the hybrid iterative scheme formed by the Picard method, as a predictor, and the Newton method as a corrector iterative scheme that accelerates the convergence speed of the Picard method. Note that Newton's method is also an iterative scheme with low operational cost and quadratic convergence. Thus, we propose the following iterative scheme:

$$\begin{cases} \begin{cases} X_0 \in \mathbb{R}^{m\times m}, \\ X_{n+1} = X_n - F(X_n), \quad n = 0, 1, 2, \ldots, N_0 - 1, \end{cases} \\ \begin{cases} Y_0 = X_{N_0}, \\ Y_{n+1} = Y_n - [F'(Y_n)]^{-1}F(Y_n), \quad n \geqslant 0, \end{cases} \end{cases} \tag{3.2}$$

to approximate a solution of equation (1.1), where $F : \mathbb{R}^{m\times m} \to \mathbb{R}^{m\times m}$, with $F(X) = X - (X - B)^{-1}C$. From now on, we use the notation $\{Z_n\}$ to refer to the hybrid method (3.2), such that

$$Z_n = \begin{cases} X_n & \text{for } n = 0, 1, ..., N_0 - 1, \\ Y_n & \text{for } n \geqslant N_0, \end{cases}$$

Secondly, our main is to ensure the convergence of the iterative scheme (3.2) under the same convergence conditions given for the Picard method in Theorem 2.9, locating the value of N_0. This maintains the accessibility of the Picard method for the hybrid iterative scheme (3.2).

Theorem 3.1 *Under conditions of Theorem* 2.9. *We suppose that the scalar equation Therefore, if we suppose that the scalar equation*

$$\frac{2((1-K)^2 - M(t)\eta_0)}{2(1-K)^2 - 3M(t)\eta_0} = t, \tag{3.3}$$

has at least one positive solution and we denote by $\widetilde{R}$ *the smallest positive root, starting at* $X_0 \in \mathbb{R}^{m\times m}$ *and for*

$$N_0 \geqslant 1 + \left[\max \left\{ \frac{\ln\left(\frac{(1-K)^2}{2\eta_0 M(\widetilde{R})}\right)}{\ln K}, \frac{\ln\left(\frac{(1-K)(1/\beta_0 - R)}{\widetilde{R}\eta_0}\right)}{\ln K} \right\} \right], \tag{3.4}$$

where $[x]$ *represents the integer part of the real number x, the hybrid iterative scheme* (3.2) *converges to* Z^*, *a solution of equation* (1.1). *Moreover,* $Z_n, Z^* \in \overline{B(X_0, R + \delta\widetilde{R})}$, *for all* $n \geq 0$.

Following the numerical example given in (2.2) we illustrate the result obtained in Theorem 3.1 for the hybrid iterative scheme (3.2) . Taking $\epsilon = 0.04$ and

$$X_0 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

we are able to apply Theorem2.9. It's easy to see that $\widetilde{R} = 0.412888$ is the smallest positive root of scalar equation (3.3). Moreover, $N_0 \geqslant 1$ and then it is enough to iterate once with the Picard method to ensure a fast convergence with the Newton method to a solution of (QME) given in (2.2). Moreover, $Z_n, Z^* \in \overline{B(X_0, 0.160193)}$, for all $n \geq 0$.

4. Conclusions

From a fixed point type transformation of (QME), we obtain a stable iterative scheme of successive approximations. Using this scheme and the Picard method we carried out a qualitative study of (QME). We obtain domains of existence and uniqueness of solutions that allow us to locate and separate them. Moreover, we construct a hybrid method taking into account, the low operational cost and the good accessibility domain that these linear methods have associated. The numerical examples confirm that the hybrid iterative scheme improves the operational cost that involves the application of Newton's method as a corrector method, when approximating a solution of (QME).

Acknowledgements

This research was funded by the Spanish Ministerio de Ciencia, Innovaci?n y Universidades. PGC2018-095896-B-C21

References

[1] I.K. Argyros, S. George and S.M. Erappa. Ball convergence for an eighth order efficient method under weak conditions in Banach spaces, *SeMA J.*, 74: 513–521, 2017.

[2] I.K. Argyros and S. Hilout. *Numerical Methods in Nonlinear Analysi*. World Scientific Publ. Comp., New Jersey, 2013.

[3] V. Berinde. *Iterative approximation of fixed points*. Springer, New York, 2007.

[4] G. J. Davis. Numerical solution of a quadratic matrix equation. *SIAM J. Sci. Statist. Comput.* , 2:164–175, 1981.

[5] C. H. Guo. On a quadratic matrix equation associated with an M-matrix. *IMA J. Numer. Anal.*, 23:11-?27, 2003.

[6] M. A. Hernández and N. Romero. Numerical analysis for the quadratic matrix equations from a modification of fixed-point type. *Math. Meth. Appl. Sci.*, 42:5856–5866, 2019.

[7] W.A. Kirk and B. Sims. *Handbook of Metric Fixed Point Theory*. Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.

[8] P. Lancaster and L. Rodman. *Algebraic Riccati Equations*. The Clarendon Press, Oxford University Press, New York, 1995.

[9] L. C. G. Rogers. Fluid models in queueing theory and Wiener-Hopf factorization of Markov chains. *Ann. Appl. Probab.*, 4:390–413, 1994.

[10] Z. C. Zheng, G. X. Ren, and W. J. Wang. A reduction method for large scale unsymmetric eigenvalue problems in structural dynamics. *J. Sound Vibration*, 199:253–268, 1997.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 242–246)

A predictor-corrector iterative scheme for improving the accessibility of the Steffensen-type methods

M.A. Hernández-Verón[1], A. A. Magreñán[1], Eulalia Martínez[2], Sukhjit Singh[3]
1. *Dept. of Mathematics and Computation, University of La Rioja. Logroño, Spain.*
2. *Instituto Universitario de Matemática Multidisciplinar, Universitat Politècnica de València. València, Spain.*
3. *Department of Mathematics, Dr BR Ambedkar National Institute of Technology, Jalandhar, India.*

Abstract

Solving equations of the form $H(x) = 0$ is usually done by applying iterative methods. The Steffensen-type methods, defined by means divided differences and derivative free, are usually considered to solve these problems when H is a non-differentiable operator due its accuracy and efficiency. But, in general, the accessibility of iterative methods that use divided differences in their algorithms is reduced. The main interest of this paper is to improve the accessibility, domain of starting points, for Setffensen-type methods. So, by using a predictor-corrector iterative process we can improve this accessibility. For this, we use a predictor iterative process with a good accessibility and after we consider a Steffensen-type iterative method for a good accuracy, since this type of iterative process has quadratic convergence. Thus we will obtain a predictor-corrector iterative process with good accessibility, given by the predictor iterative process, and an accuracy like the Steffensen-type methods. Moreover, we analyze the semilocal convergence of the predictor-corrector iterative process proposed in two cases: when H is differentiable and H is non-differentiable. So, we present a good alternative for the non-applicability of Newton's method to non-differentiable operators. The theoretical results are illustrated with numerical experiments. **CEDYA/CMA 2020**.

1. Introduction

One of the most studied problems in numerical mathematics is the solution of nonlinear systems of equations

$$H(x) = 0, \tag{1.1}$$

where $H : \Omega \subset \mathbb{R}^m \longrightarrow \mathbb{R}^m$ is a nonlinear operator, $H \equiv (H_1, H_2, \ldots, H_m)$ with $H_i : \Omega \subseteq \mathbb{R}^m \to \mathbb{R}$, $1 \leq i \leq m$, and Ω is a non-empty open convex domain. Iterative methods are a powerful tool for solving these equations.

In this paper, we consider iterative processes free of derivatives. But these methods have a serious shortcoming: they have a region of reduced accessibility. In [4], the accessibility of an iterative process is increased by means of an analytical procedure, that is, by modifying the convergence conditions. However, in our work, we will increase accessibility by building an iterative predictor-corrector process. This iterative process has a first prediction phase and a second accurate approximation phase.

Kung and Traub presented in [10] a class of iterative processes without derivatives. These iterative processes considered by Kung and Traub contain Steffensen-type methods as a special case. From the biparametric family of iterative processes given in [2],

$$\begin{cases} x_0 \in \Omega, \quad \alpha, \ \beta \geq 0 \\ y_n = x_n - \alpha H(x_n), \\ z_n = x_n + \beta H(x_n), , \\ x_{n+1} = x_n - [y_n, z_n; H]^{-1} H(x_n), n \geqslant 0. \end{cases} \tag{1.2}$$

The three best-known Steffensen-type methods appear. So, for $\alpha = 0$ and $\beta = 1$ we have the original Steffensen method, the Backward-Steffensen method for $\alpha = 1$ and $\beta = 0$ and the Central-Steffensen method for $\alpha = 1$ and $\beta = 1$.

Notice that, if we consider the Newton's method,

$$x_{n+1} = x_n - [H'(x_n)]^{-1} H(x_n), \quad n \geq 0; \quad x_0 \in \Omega \text{ is given}, \tag{1.3}$$

which is one of the most used iterative methods to approximate a solution x^* of $H(x) = 0$, the Steffensen-type methods are obtained as a special case of this method, where the evaluation of $H'(x)$ in each step of Newton's method is approximated by the divided difference of first order $[x - \alpha H(x), x + \beta H(x); H]$. The Stetffensen-type methods have been widely studied ([1, 3, 6]).

Symmetric divided differences generally perform better. Moreover, maintain the quadratic convergence of Newton's method, by approximating the derivative through symmetric divided differences with respect to the x_n, and the Center-Steffensen method also has the same computational efficiency as Newton's method. But, in order to achieve the second order in practice, we need an iteration close enough to the solution to have a good approximation of the first derivative of H used in Newton's method. In other case, some extra iterations in comparison with Newton's method are required. Basically, when the norm of $H(x)$ is big, the approximation of the divided difference to the first derivative of H is bad. So, in general, the set of starting points of the Steffensen-type methods is poor. This justify that Steffensen-type methods are less used than Newton's method to approximate solutions of equations for differentiable operators.

Thus, two are our main objectives in this work. On the one hand, in the case of differentiable operators, to which Newton's method can also be applied, to construct a predictor-corrector iterative process that has accessibility and efficiency like Newton's method, but using symmetric divided differences. And, secondly, that this predictor-corrector iterative process considered can have a behavior like Newton's method has in the differentiable case, but considering the case of non-differentiable operators where Newton's method cannot be applied.

Following this idea, in this paper we consider the derivative-free point-to-point iterative process given by

$$\begin{cases} x_0 \text{ given in } \Omega, \\ x_{n+1} = x_n - [x_n - \mathbf{Tol}, x_n + \mathbf{Tol}; H]^{-1} H(x_n), \quad n \geq 0, \end{cases} \tag{1.4}$$

where $\mathbf{Tol} = (tol, tol, \ldots, tol) \in \mathbb{R}^m$ for a real number $tol > 0$. Thus, we take a symmetric divided difference to approximate the derivative in Newton's method. Furthermore, by varying the parameter tol, we can approach the value $F'(x_n)$. Notice that, in the differentiable case, for $tol = 0$ we obtain the Newton's method.

However, although reducing the value of tol we can reach a speed of convergence similar to Newton's method, its order of convergence is linear. That is why we will consider this method as a predictor, due to its good accessibility, and we will consider the Center-Steffensen method as a corrector, whose order of convergence is quadratic.

So, we consider the predictor-corrector method given by

$$\begin{cases} \begin{cases} \text{Given an initial guess } \mathrm{u}_0 \in \Omega, \\ u_{j+1} = u_j - [u_j - \mathbf{Tol}, u_j + \mathbf{Tol}; H]^{-1} H(u_j), \quad j = 0, ..., N_0 - 1, \end{cases} \\ \begin{cases} x_0 = u_{N_0}, \\ y_n = x_n - H(x_n),\ n \geqslant 0, \\ z_n = x_n + H(x_n),\ n \geqslant 0, \\ x_{n+1} = x_n - [y_n, z_n; H]^{-1} H(x_n),\ n \geqslant 0, \end{cases} \end{cases} \tag{1.5}$$

where $\mathbf{Tol} = (tol, tol, \ldots, tol) \in \mathbb{R}^m$ for a real number $tol > 0$. Thus, this predictor-corrector method will be a Steffensen-type method with good accessibility and quadratic convergence from an iteration to be determined.

The paper is organized as follows. First, we introduce the motivation of the paper. Next, we establish a semilocal convergence analysis of the new method when operator H is both differentiable and non-differentiable cases.

2. Semilocal convergence

The semilocal study of the convergence is based on demanding conditions to the initial approximation u_0, from certain conditions on the operator H , and provide conditions required to the initial approximation that guarantee the convergence of sequence (1.5) to the solution x^*. To analyze the semilocal convergence of iterative processes that do not use derivatives in their algorithms, the conditions usually are required for the operator divided difference. Although, in the case that the operator H is Fréchet differentiable, the divided difference operator can be defined from the Fréchet derivative of the operator H.

2.1. Differentiable operators

Now, we establish the semilocal convergence of iterative process given in (1.5) for differentiable operators. So, we consider $H : \Omega \subset \mathbb{R}^m \longrightarrow \mathbb{R}^m$ a Fréchet differentiable operator and there exists

$$[v, w; H] = \int_0^1 H'(tv + (1-t)w)\, dt, \tag{2.1}$$

for each pair of distinct points $v, w \in \Omega$. Notice that, as H is Fréchet differentiable $[x, x; H] = H'(x)$.

Now, we suppose the following initial conditions:

(D1) Let $u_0 \in \Omega$ such that exists $\Gamma_0 = [H'(u_0)]^{-1}$ with $\|\Gamma_0\| \leq \beta$ and $\|H(u_0)\| \leq \delta_0$.

(D2) $\|H'(x) - H'(y)\| \leq K\|x - y\|, x, y \in \Omega, K \in \mathbb{R}^+$.

In first place, we obtain some technical results.

Lemma 2.1 *The following items are verified.*

(*i*) *Let $R > 0$ with $B(u_0, R + \|\mathbf{Tol}\|) \subseteq \Omega$. If $\beta K(R + \|\mathbf{Tol}\|) < 1$ then, for each pair of distinct points $y, z \in B(u_0, R + \|\mathbf{Tol}\|)$, there exists $[y, z; H]^{-1}$ such that*

$$\|[y, z; H]^{-1}\| \leq \frac{\beta}{1 - \beta K(R + \|\mathbf{Tol}\|)}. \tag{2.2}$$

(*ii*) *If $u_j, u_{j-1} \in \Omega$, for $j = 0, 1, \ldots, N_0$, then*

$$\|H(u_j)\| \leq \frac{K}{2}(\|\mathbf{Tol}\| + \|u_j - u_{j-1}\|)\|u_j - u_{j-1}\|. \tag{2.3}$$

(*iii*) *If $x_j, x_{j-1} \in \Omega$, , for $j \geqslant 1$, then*

$$\|H(x_j)\| \leq \frac{K}{2}(\|H(x_{j-1}\| + \|x_j - x_{j-1}\|)\|x_j - x_{j-1}\|. \tag{2.4}$$

To simplify the notation, from now on, we denote

$$A_j = [u_j - \mathbf{Tol}, u_j + \mathbf{Tol}; H], \quad B_j = [x_j - H(x_j), x_j + H(x_j); H],$$

and the parameters $a_0 = \beta^2 K\delta_0$ and $b_0 = \beta K\mathbf{Tol}$. Other parameters that we are going to use are the following:

$$M = \frac{L}{2}(b_0 + La_0), \text{ where } L = \frac{1}{1 - b_0 - \beta KR}.$$

Moreover, notice that the polynomial equation $p(t) = 0$, with

$$p(t) = 2a_0(1 - b_0) - (2 + a_0 - 5b_0 + 3b_0^2)\beta Kt + (4 - 5b_0)\beta^2 K^2 t^2 - 2\beta^3 K^3 t^3,$$

has at least a positive real root since that $p(0) > 0$ and $p(t) \to -\infty$ as $t \to \infty$. Then, we denote by R the smallest positive real root of the polynomial equation $p(t) = 0$.
Finally, we denote by $[x]$ the integer part of the real number x.

Theorem 2.2 *Let $H : \Omega \subset \mathbb{R}^m \longrightarrow \mathbb{R}^m$ a Fréchet differentiable operator defined on a nonempty open convex domain Ω. Suppose that conditions $(D1)$ and $(D2)$ are satisfied and there exists $tol > 0$ such that $M < 1$, $R < \dfrac{1 - b_0}{\beta K}$ and $B(u_0, R + \|\mathbf{Tol}\|) \subseteq \Omega$. If we consider*

$$N_0 \geqslant \begin{cases} 1 + \left[\dfrac{log(\|\mathbf{Tol}\|/\delta_0)}{log(M)}\right] & \text{if } \|\mathbf{Tol}\| < \delta_0, \\ 1 & \text{if } \|\mathbf{Tol}\| > \delta_0, \end{cases} \tag{2.5}$$

then the iterative process predictor-corrector (1.5)*, starting at u_0, converges to x^* a solution of $H(x) = 0$. Moreover, $u_j, x_n, x^* \in \overline{B(u_0, R)}$ for $j = 1, \ldots, N_0$ and $n \geqslant 0$.*

Next, we get an uniqueness result for the iterative process predictor-corrector (1.5).

Theorem 2.3 *Under conditions of the previous Theorem, the solution x^* of the equation $H(x) = 0$ is unique in $B(u_0, R)$.*

2.2. Non-differentiable operators

To obtain a result of semilocal convergence for iterative process (1.5) when H is a non-differentiable operator, we must suppose that for each pair of distinct points $x, y \in \Omega$, there exists a first-order divided difference of H at these points. As we consider Ω an open convex domain of $\mathbb{R}^m$, this condition is satisfied ([5], [7]). Moreover, it is also necessary to impose a condition for the first-order divided difference of the operator H. As it appears in [11] and [9], a Lipschitz-continuous condition or a Hölder-continuous can be considered, but in the above cases, it is known [8], that the Fréchet derivative of H exists in Ω. Therefore, these conditions cannot be verified if the operator H is non-differentiable. Then, to establish the semilocal convergence of iterative process given in (1.5) for non-differentiable operator H, we suppose the following conditions:

(ND1) Let $u_0 \in \Omega$ such that A_0^{-1} exists with $||A_0^{-1}|| \leq \beta_0$ and $||H(u_0)|| \leq \delta_0$.

(ND2) $||[x, y; H] - [u, v; H]|| \leq P + K\left(||x - u|| + ||y - v||\right)$, $P, K \geq 0$, with $x, y, u, v \in \Omega, x \neq y, u \neq v$.

To simplify the notation, from now on, we denote

$$\widetilde{M} = \beta_0(P + K(\beta_0\delta_0 + 2\|\mathbf{Tol}\|)) \qquad \text{and} \qquad S = \frac{\widetilde{M}}{1 - \beta_0(P + 2K(R + \|\mathbf{Tol}\|))}$$

In this conditions, we start our study obtaining a technical result, the proof of which is evident from algorithm given in (1.5).

Lemma 2.4 *The following items can be easily verified.*

(*i*) *If $u_j, u_{j-1} \in \Omega$, for $j = 0, 1, \ldots, N_0$, then*

$$H(u_j) = \left([u_j, u_{j-1}; H] - A_{j-1}\right)(u_j - u_{j-1}). \tag{2.6}$$

(*ii*) *If $x_j, x_{j-1} \in \Omega$, for $j \geqslant 1$, then*

$$H(x_j) = \left([x_j, x_{j-1}; H] - B_{j-1}\right)(x_j - x_{j-1}). \tag{2.7}$$

Theorem 2.5 *Under the conditions (ND1)-(ND2), if the real equation*

$$t = \frac{\beta_0\delta_0(1 - \beta_0(P + 2K(t - \|\mathbf{Tol}\|)))}{1 - \beta_0(P + 2K(t + \|\mathbf{Tol}\|)) - \widetilde{M}}, \tag{2.8}$$

has at least one positive real root, the smallest positive root is denoted by R, and there exists $tol > 0$ such that satisfies

$$\widetilde{M} + \beta_0(P + 2K(R + \|\mathbf{Tol}\|)) < 1, \tag{2.9}$$

and $B(u_0, R + \|\mathbf{Tol}\|) \subset \Omega$. If we consider

$$N_0 \geqslant \begin{cases} 2 + \left\lfloor \dfrac{log(\|\mathbf{Tol}\|/\widetilde{M}\delta_0)}{log(M)} \right\rfloor & \text{if } \|\mathbf{Tol}\| < \dfrac{\beta_0\delta_0(P + \beta_0\delta_0 K)}{1 - 2\beta_0\delta_0}, \\ 1 & \text{if } \|\mathbf{Tol}\| > \dfrac{\beta_0\delta_0(P + \beta_0\delta_0 K)}{1 - 2\beta_0\delta_0}, \end{cases} \tag{2.10}$$

then the iterative process predictor-corrector (1.5), starting at u_0, converges to x^* a solution of $H(x) = 0$. Moreover, $u_j, x_n, x^* \in \overline{B(u_0, R)}$ for $j = 1, \ldots, N_0$ and $n \geqslant 0$.

Moreover, x^* is unique in $B(u_0, R) \subset \Omega$.

Theorem 2.6 *Under conditions of the previous Theorem, the solution x^* of the equation $H(x) = 0$ is unique in $B(u_0, R)$.*

Acknowledgements

This research was partially supported by the project PGC2018-095896-B-C21-C22 of Spanish Ministry of Economy and Competitiveness and by the project of Generalitat Valenciana Prometeo/2016/089.

References

[1] Alarcón, V. , Amat, S., Busquier, S. and López, D. J., A Steffensen's type method in Banach spaces with applications on boundary-value problems, J. Comput. Appl. Math. **216** (2008), 243-250.

[2] Amat, S., Ezquerro, J.A. and Hernández-Verón, M.A., On a Steffensen-like method for solving nonlinear equations, Calcolo, 53 (2016) 171–188.

[3] Argyros, I. K., A new convergence theorem for Steffensen's method on Banach spaces and applications, Southwest J. Pure Appl. Math., **1** (1997), 23–29.

[4] Argyros, I.K. and George,S., On the complexity of extending the convergence region for Traub's method, J. Complexity 56 (2020), 101423.

[5] Balazs, M. and Goldner, G., On existence of divided differences in linear spaces. Rev. Anal. Numer. Theor. Approx., 2 (1973), 3–6.

[6] Ezquerro, J.A., Hernández, M.A., Romero, N. and Velasco, A.I., On Steffensen's method on Banach spaces, J. Comput. Appl. Math., **249** (2013), 9–23.

[7] Grau-Sánchez, M., Noguera, M. and Amat, S., On the approximation of derivatives using divided difference operators preserving the local convergence order of iterative methods. J. Comput. Appl. Math., 237 (2013), 363–372.

[8] Hernández, M. A. and Rubio, M. J., A uniparametric family of iterative processes for solving non-differentiable equations, J. Math. Anal. Appl., **275** (2002), 821–834.

[9] Hilout, S., Convergence analysis of a family of Steffensen-type methods for generalized equations, J. Math. Anal. Appl., 329 (2008) 753–761.

[10] Kung, H. T. and Traub, J. F., Optimal order of one-point and multipoint iteration, Carnegie Mellon University. Research Showcase @ CMU. Computer Science Department. Paper 1747. 1973.

[11] Moccari, M. and Lotfi, T., On a two-step optimal Steffensen-type method: Relaxed local and semi-local convergence analysis and dynamical stability, J. Math. Anal. Appl., 468 (2018) 240–269.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 247–252)

Recent developments in modeling free-surface flows with vertically-resolved velocity profiles using moments

Julian Koellermeier[1]
julian.koellermeier@kuleuven.be KU Leuven, Belgium

Abstract

Shallow water moment models are non-linear PDEs in balance law form for free-surface flows that allow for vertical variations in the horizontal velocity. The models are extensions of the standard shallow water equations. However, the models in their original form lack global hyperbolicity. The loss of hyperbolicity already occurs for small vertical variations of the velocity and this leads to instabilities in numerical test cases. We review two recently developed hyperbolic shallow water moment models, which are based on two different linearizations during the derivation. Recently, the models have been extended to consider sediment transport and bottom topographies, for which new well-balanced numerical schemes based on analytical derivation of steady states can be constructed. We summarize the recent developments focusing on analytical properties of the models and their derivation.

1. Introduction

The well-known Shallow Water Equations (SWE), sometimes also called Saint-Venant equations, are a simplified model for free-surface flows and are commonly used to model different physical phenomena. However, the main deficiency of these models is that they assume a constant velocity profile of the horizontal velocity. In fact, the model only takes into account the mean velocity averaged along the vertical axis. This limits the applicability of the SWE model for complex flows and situations in which bottom friction plays an important role such as sediment transport.

One option to include vertical variations of the velocity is the use of multiple layers with piecewise constant velocities [2] leading to a system of equations that is coupled via the interfaces. However, the analysis of the model is difficult and no analytical eigenvalues can be obtained. Additionally, many layers are necessary to accurately describe varying profiles.

A polynomial velocity ansatz was used in [9] and the system of equations for the coefficients can be obtained by projection onto orthogonal test functions. This can be seen as an extension of the standard SWE model using an extended set of variables, so-called moments. These new Shallow Water Moment Equations (SWME) have been applied to several test cases which showed the accuracy and flexibility of the approach.

The main drawback of the SWME model in its original version is that the model looses hyperbolicity even for small variations of the velocity profile, as shown in [7]. This can lead to oscillations and a breakdown of the solution during simulations, which was exemplified using a dam-break test case.

Hyperbolicity was restored using two different linearization of the model in [7] and [6]. We will summarize the derivations of both models in this paper and outline the different analytical properties.

While hyperbolicity is a main ingredient for a stable numerical simulation, different physical phenomena need to be modeled by means of special friction terms or additional equations. We show a recently developed example of sediment transport [3].

2. Shallow Water Moment Models

The standard shallow water equations (SWE) for a Newtonian fluid in one horizontal direction x for water height h and mean velocity u_m using a flat bottom topography are given by

$$\partial_t \begin{pmatrix} h \\ hu_m \end{pmatrix} + \partial_x \begin{pmatrix} hu_m \\ hu_m^2 + \frac{1}{2} g h^2 \end{pmatrix} = -\frac{\nu}{\lambda} \begin{pmatrix} 0 \\ u_m \end{pmatrix}, \tag{2.1}$$

where λ and ν denote the slip length and the kinematic viscosity, respectively.

While the SWE model is efficient to compute approximate solutions of simple flows in very shallow conditions, the model is inaccurate in case of horizontal variations of the vertical velocity. This is due to the fact that only the average velocity u_m is a variable of the model. In [9], the Shallow Water Moment Equations (SWME) were developed to overcome this problem. The derivation is based on two main ideas:

- The first idea is to scale vertical position variable $\zeta(t,x)$ as

$$\zeta(t,x) := \frac{z - h_b(t,x)}{h_s(t,x) - h_b(t,x)} = \frac{z - h_b(t,x)}{h(t,x)},$$

with $h(t,x) = h_s(t,x) - h_b(t,x)$ the water height from the bottom h_b to the surface h_s. This transforms the vertical z-direction from a physical space to a projected space $\zeta : [0,T] \times \mathbb{R} \to [0,1]$, see [9].

- The second idea assumes a polynomial expansion of the velocity variable, in the transformed vertical direction. We thus expand $u : [0,T] \times \mathbb{R} \times [0,1] \to \mathbb{R}$ as

$$u(t,x,\zeta) = u_m(t,x) + \sum_{j=1}^{N} \alpha_j(t,x)\phi_j(\zeta), \tag{2.2}$$

where $u_m : [0,T] \times \mathbb{R} \to \mathbb{R}$ is the mean velocity and $\phi_j : [0,1] \to \mathbb{R}$ are the *scaled Legendre polynomials* of degree j defined by

$$\phi_j(\zeta) = \frac{1}{j!}\frac{d^j}{d\zeta^j}(\zeta - \zeta^2)^j. \tag{2.3}$$

Note that the basis polynomials fulfill $\phi_j(0) = 1$ and they are orthogonal basis functions as

$$\int_0^1 \phi_m \phi_n d\zeta = \frac{1}{2n+1}\delta_{mn}, \tag{2.4}$$

with Kronecker delta δ_{mn} [9].

With $\alpha_j : [0,T] \times \mathbb{R} \to \mathbb{R}$ for $j \in [1,2,\dots,N]$ we denote the corresponding *basis coefficients* at time t and position x. These coefficients are also called *moments*. Different values of the coefficients describe different horizontal velocity profiles, which allows for more complex flows and extends the standard SWE (2.1), where the horizontal velocity is constant. In the expansion, $N \in \mathbb{N}$ is the order of the velocity expansion and at the same time the maximum degree of the Legendre polynomials. A larger N typically enables the representation of more complex flows, whereas $N = 0$ corresponds to the constant velocity profile of the standard SWE (2.1).

To derive evolution equations for the basis coefficients, the expansion is inserted into the Navier-Stokes equations, which have been properly transformed to the new $\zeta(t,x)$ variable, see [9] for details. Then, the equations are projected onto the Legendre polynomials of degree $i = 1,\dots,N$, by multiplication with ϕ_j and integration over ζ, which gives one additional equation for each coefficient in the expansion. The arising integrals of the basis polynomials A_{ijk}, B_{ijk}, C_{ij} are denoted as follows

$$A_{ijk} = (2i+1)\int_0^1 \phi_i \phi_j \phi_k \, d\zeta, \tag{2.5}$$

$$B_{ijk} = (2i+1)\int_0^1 \partial_\zeta \phi_i \left(\int_0^\zeta \phi_j \, d\hat{\zeta}\right) \phi_k \, d\zeta, \tag{2.6}$$

$$C_{ij} = \int_0^1 \partial_\zeta \phi_i \, \partial_\zeta \phi_j \, d\zeta. \tag{2.7}$$

More details can be found in [6, 9].

The model with variables $U = (h, hu, h\alpha_1, \dots, h\alpha_N)^T \in \mathbb{R}^{N+2}$ can be written in compact form as

$$\partial_t U + \frac{\partial F}{\partial U}\partial_x U = Q\partial_x U + S, \tag{2.8}$$

where the conservative flux Jacobian $\frac{\partial F}{\partial U}$ is given by

$$\frac{\partial F}{\partial U} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ gh - u^2 - \sum_{i=1}^{N} \frac{\alpha_i}{2i+1} & 2u & \frac{2\alpha_1}{2\cdot 1+1} & \dots & \frac{2\alpha_N}{2N+1} \\ -2u\alpha_1 - \sum_{j,k=1}^{N} A_{1jk}\alpha_j\alpha_k & 2\alpha_1 & 2u\delta_{11} + 2\sum_{k=1}^{N} A_{11k}\alpha_k & \dots & 2u\delta_{1N} + 2\sum_{k=1}^{N} A_{1Nk}\alpha_k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -2u\alpha_N - \sum_{j,k=1}^{N} A_{Njk}\alpha_j\alpha_k & 2\alpha_N & 2u\delta_{N1} + 2\sum_{k=1}^{N} A_{N1k}\alpha_k & \dots & 2u\delta_{NN} + 2\sum_{k=1}^{N} A_{NNk}\alpha_k \end{pmatrix},$$

and the non-conservative matrix Q reads

$$Q = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & u\delta_{11} + \sum_{k=1}^{N} B_{11k}\alpha_k & \dots & u\delta_{1N} + \sum_{k=1}^{N} B_{1Nk}\alpha_k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & u\delta_{N1} + \sum_{k=1}^{N} B_{N1k}\alpha_k & \dots & u\delta_{NN} + \sum_{k=1}^{N} B_{NNk}\alpha_k \end{pmatrix}.$$

The friction term on the right-hand side $S = (0, S_0, S_1, \dots, S_N)^T \in \mathbb{R}^{N+2}$ is defined in [9] as $S_0 = 0$ and

$$S_i = -(2i+1)\frac{\nu}{\lambda}\left(u + \sum_{j=1}^{N}\left(1 + \frac{\lambda}{h}C_{ij}\right)\alpha_j\right), i = 0, \dots, N. \tag{2.9}$$

The model (2.8) can also be written in the form of

$$\partial_t U + A(U)\partial_x U = S(U), \tag{2.10}$$

where the combined transport matrix $A = \frac{\partial F}{\partial U} - Q$ can easily be obtained from conservative flux Jacobian and the non-conservative terms.

The new SWME model was used for simulation of smooth waves and dam-break scenarios in [9]. The model was more accurate than the standard SWE model and converged towards a reference solution with increasing number of moments/coefficients N.

3. Hyperbolic Regularization

As already noted in [9], the SWME model is not hyperbolic for values $N > 1$. Loosing hyperbolicity may or may not lead to instabilities and non-physical values during numerical simulations. In [7], the hyperbolicity was studied in more detail and a breakdown of hyperbolicity inducing instable oscillations in time could be found for standard simulations.

Two hyperbolic models have recently been developed. The first one called the Hyperbolic Shallow Water Moment Equations (HSWME) from [7] is based on insights from moment models for rarefied gases [1, 5, 8]. The second one called the Shallow Water Linearized Moment Equations (SWLME) from [6] is based on the assumption of small deviations from equilibrium and neglecting small terms in the derivation. Both models are hyperbolic. We will outline the main ideas and state the model equations following both approaches in the next two subsections.

3.1. Hyperbolic Shallow Water Moment Equations

The Hyperbolic Shallow Water Moment Equations (HSWME) [7] overcome the loss of hyperbolicity using a linearization of the SWME model around linear velocity deviations, denoted by the case $N = 1$. This leads to setting all other coefficients $\alpha_i = 0$ for $i = 2, \dots, N$ in the SWME model matrices. Note that the model is still non-linear and includes the dependencies on all other variables h, u_m, α_1 on the dynamics of the other coefficients.

The HSWME system from [7] is written in the same non-conservative form as (2.10) as

$$\partial_t U + A_H(U)\partial_x U = S(U), \tag{3.1}$$

with regularized hyperbolic system matrix $A_H(U) \in \mathbb{R}^{(N+2)\times(N+2)}$ given by

$$A_H(U) = \begin{pmatrix} 0 & 1 & & & & \\ gh - u_m^2 - \frac{1}{3}\alpha_1^2 & 2u_m & \frac{2}{3}\alpha_1 & & & \\ -2u_m\alpha_1 & 2\alpha_1 & u_m & \frac{3}{5}\alpha_1 & & \\ -\frac{2}{3}\alpha_1^2 & 0 & \frac{1}{3}\alpha_1 & u_m & \ddots & \\ & & & \ddots & \ddots & \frac{N+1}{2N+1}\alpha_1 \\ & & & & \frac{N-1}{2N-1}\alpha_1 & u_m \end{pmatrix}. \tag{3.2}$$

In [7], it was shown up to a certain N, that the model has real eigenvalues and is therefore hyperbolic, for all variable states. The proof was recently extended to arbitrary order in [4] which yields the following theorem

Theorem 3.1 *The HSWME model* (3.1) *of arbitrary order N is globally hyperbolic and the eigenvalues are*

$$\lambda_{1,2} = u_m \pm \sqrt{gh + \alpha_1^2},$$
$$\lambda_{i+2} = u_m + r_{i,N}\alpha_1, \quad i = 1, 2, \ldots, N,$$

where $r_{i,N} \in \mathbb{R}$ *is the i-th root of the real polynomial* $p_N(z)$ *of degree N, defined by the recursion* $p_k(z) = zp_{k-1}(z) - b_k p_{k-2}(z)$, *for* $2 \le k \le N$, $p_1(z) = 1$, $b_k = \frac{(k-1)(k+1)}{(2k-1)(2k+1)}$.

3.2. Shallow Water Linearized Moment Equations

The second hyperbolic model called Shallow Water Linearized Moment Equations (SWLME) derived in [6] is based on a careful investigation of non-linear terms in the underlying model equations. One example is the term

$$\int_0^1 \phi_i u^2 \, d\zeta.$$

Using the polynomial velocity expansion (2.2), this terms can be computed according to [6] as

$$\int_0^1 \phi_i u^2 \, d\zeta = \int_0^1 \phi_i \left(u_m + \sum_{j=1}^N \alpha_j \phi_j \right)^2 d\zeta \tag{3.3}$$
$$= u_m^2 \int_0^1 \phi_i \, d\zeta + \sum_{j=1}^N 2u_m\alpha_j \int_0^1 \phi_i\phi_j \, d\zeta + \sum_{j,k=1}^N 2\alpha_j\alpha_k \int_0^1 \phi_i\phi_j\phi_k \, d\zeta \tag{3.4}$$
$$= 0 + \frac{2}{2i+1}u_m\alpha_i + \frac{1}{2i+1}\sum_{j,k}^N A_{ijk}\alpha_j\alpha_k. \tag{3.5}$$

Now the model assumes small deviations from a constant profile, i.e., $\alpha_i = O(\epsilon)$, such that the last term containing the coefficient coupling $\alpha_j\alpha_k = O(\epsilon^2)$ can be neglected in comparison to the first term. The result is the simpler expression

$$\int_0^1 \phi_i u^2 \, d\zeta \approx \frac{2}{2i+1}u_m\alpha_i.$$

Based on this strategy, the SWLME model includes fewer terms than the original (2.10) and reads

$$\partial_t \begin{pmatrix} h \\ hu_m \\ h\alpha_1 \\ \vdots \\ h\alpha_N \end{pmatrix} + \partial_x \begin{pmatrix} hu_m \\ hu_m^2 + g\frac{h^2}{2} + \frac{1}{3}h\alpha_1^2 + \ldots + \frac{1}{2N+1}h\alpha_N^2 \\ 2hu_m\alpha_1 \\ \vdots \\ 2hu_m\alpha_N \end{pmatrix} = Q\partial_x \begin{pmatrix} h \\ hu_m \\ h\alpha_1 \\ \vdots \\ h\alpha_N \end{pmatrix} + P, \tag{3.6}$$

where the non-conservative term simplifies to

$$Q = (0, 0, \mathrm{u_m}, \ldots, \mathrm{u_m}),$$

and the combined transport system matrix of the new SWLME can be written as

$$A_N = \begin{pmatrix} 0 & 1 & 0 & \vdots & 0 \\ gh - u_m^2 - \frac{\alpha_1^2}{3} - \ldots - \frac{\alpha_N^2}{2N+1} & 2u_m & \frac{2\alpha_1}{3} & \ldots & \frac{2\alpha_N}{2N+1} \\ -2u_m\alpha_1 & 2\alpha_1 & u_m & & \\ \vdots & \vdots & & \ddots & \\ -2u_m\alpha_N & 2\alpha_N & & & u_m \end{pmatrix} \in \mathbb{R}^{(N+2)\times(N+2)}. \tag{3.7}$$

It was shown in the following theorem from [6] that the eigenvalues of the SWLME model are indeed real such that the model is hyperbolic

Theorem 3.2 *The SWLME system matrix* $A_N \in \mathbb{R}^{(N+2)\times(N+2)}$ (3.7) *has the following characteristic polynomial*

$$\chi_{A_N}(\lambda) = (u_m - \lambda)\left[(\lambda - u_m)^2 - gh - \sum_{i=1}^N \frac{3\alpha_i^2}{2i+1} \right]$$

and the eigenvalues are given by

$$\lambda_{1,2} = u_m \pm \sqrt{gh + \sum_{i=1}^{N} \frac{3\alpha_i^2}{2i+1}} \quad \textit{and} \quad \lambda_{i+2} = u, \ \textit{for} \ i = 1, \ldots, N. \tag{3.8}$$

The system is thus hyperbolic.

3.3. Steady states of SWLME

Another main benefit of the SWLME model is the possibility of obtaining analytical steady states that generalize the standard SWE Rankine-Hugoniot conditions. According to [6] the steady states can be derived as follows for flat bottom $\partial_x b = 0$ and zero friction:

$$\begin{aligned}
\partial_x \left(hu_m\right) &= 0 && (3.9)\\
\partial_x \left(hu_m^2 + \frac{1}{2}gh^2 + \frac{1}{3}h\alpha_1^2 + \ldots + \frac{1}{2N+1}h\alpha_N^2\right) &= 0 && (3.10)\\
\partial_x \left(2hu_m\alpha_1\right) &= u_m \partial_x \left(h\alpha_1\right) && (3.11)\\
&\vdots && (3.12)\\
\partial_x \left(2hu_m\alpha_N\right) &= u_m \partial_x \left(h\alpha_N\right), && (3.13)
\end{aligned}$$

which first leads to

$$\begin{aligned}
hu_m &= const, && (3.14)\\
u_m = 0 \text{ or } \frac{\alpha_i}{h} &= const, \ \text{for } i = 1, \ldots, N. && (3.15)
\end{aligned}$$

The Rankine-Hugoniot conditions for a shock from a given state $\left(h_0, h_0u_{m,0}, h_0\alpha_{1,0}, \ldots, h_0\alpha_{N,0}\right)$ to a state $(h, hu_m, h\alpha_1, \ldots, h\alpha_N)$ then read

$$(h-h_0)\left[-\frac{u_{m,0}^2}{gh_0} + \frac{1}{2}\left(\left(\frac{h}{h_0}\right)^2 + \left(\frac{h}{h_0}\right)\right) + \sum_{i=1}^{N} \frac{1}{2i+1}\frac{\alpha_{i,0}^2}{gh_0}\left(\left(\frac{h}{h_0}\right)^3 + \left(\frac{h}{h_0}\right)^2 + \left(\frac{h}{h_0}\right)\right)\right] = 0. \tag{3.16}$$

Introducing the dimensionless flow numbers

$$\begin{aligned}
Fr &= \frac{u_{m,0}}{\sqrt{gh_0}}, && (3.17)\\
(M\alpha)_i &= \frac{\alpha_{i,0}}{u_{m,0}}, \quad \text{for } i = 1, \ldots, N && (3.18)
\end{aligned}$$

and writing $y = \frac{h}{h_0}$, leads to the non-dimensional solutions

$$h = h_0 \ \vee \ -Fr^2 + \frac{1}{2}\left(y^2 + y\right) + \sum_{i=1}^{N} \frac{1}{2i+1}\left(M\alpha\right)_i^2 Fr^2 \left(y^3 + y^2 + y\right) = 0. \tag{3.19}$$

This gives rise to a new dimensionless number $M\alpha^2 := \sum_{i=1}^{N} \frac{1}{2i+1}\left(M\alpha\right)_i^2$. According to [6], $M\alpha$ measures the total deviation from equilibrium. Note, that there is at least one non-trivial solution for non-zero Fr and $M\alpha$.

It is also possible to derive steady states for smooth and frictionless flows including bottom topographies that can later be used to derive well-balanced schemes. In the momentum equation, this requires

$$\partial_x \left(\frac{1}{2}u_m^2 + g(h+b) + \frac{3}{2}\sum_{i=1}^{N} \frac{1}{2i+1}\alpha_i^2\right) = 0, \tag{3.20}$$

where $b(x)$ is the bottom topography term.

The full non-trivial steady state solution is then computed by solving

$$\begin{aligned}
hu_m &= const, && (3.21)\\
\frac{1}{2}u_m^2 + g(h+b) + \frac{3}{2}\sum_{i=1}^{N} \frac{1}{2i+1}\alpha_i^2 &= const, && (3.22)\\
\frac{\alpha_i}{h} &= const, \ \text{for } i = 1, \ldots, N. && (3.23)
\end{aligned}$$

The analytically computed equations to determine steady-states are used within a well-balanced numerical scheme to conserve certain steady-states numerically. We refer to [6] for detailed examples.

4. Sediment transport and friction models

In [3], the HSWME model was coupled to an Exner equation [11], modeling sediment transport at the bottom. This means that the bottom topography $b(t, x)$ is also a function of time and evolves according to

$$\partial_t b + \partial_x Q_b = 0, \tag{4.1}$$

where Q_b is the solid transport discharge that can be modeled by the Meyer-Peter & Müller formula [10].

It was shown in [3] that the eigenvalues of the coupled model are a generalization of the eigenvalues of the standard SWE model coupled to the Exner equation. The additional eigenvalues are real such that the model is again hyperbolic. The model leads to a much more realistic sediment transport. Unlike as for the SWE model, the velocity at the bottom is not the same as the average velocity u_m, which means that the coupled sediment equation (4.1) is correctly transported with the bottom velocity according to the polynomial expansion (2.2).

5. Summary and future work

In this paper, recent developments in modeling free-surface flows with vertically resolved velocity profiles were summarized and compared. Based on a polynomial expansion of the velocity profile, the derivation of the Shallow Water Moment Equations was outlined. Two hyperbolic regularizations based on different linearizations of the model are described and the results for the eigenvalues and steady states are given. As one application, a sediment transport model that builds up on the previously discussed models is described.

The recently developed models are a major step forward for the simulation of complex free-surface flows. The models open up a lot of possibilities for future work. Firstly, the inclusion of a coriolis force term and the analytical investigation of wave properties is necessary for applications and to understand the structure of the models. Additional efforts should focus on the numerical simulation of the model equation, e.g., regarding the implementation of wet-dry fronts or asymptotic-preserving schemes for the limits of large friction terms. Lastly, the inclusion of more realistic friction terms of Savage-Hutter type [11] to model granular flows, e.g., for avalanches, land slides, or mud flows would be beneficial for real-world applications.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement no. 888596. The first author is a postdoctoral fellow in fundamental research of the Research Foundation - Flanders (FWO), funded by FWO grant no. 0880.212.840.

References

[1] Y. Fan, J. Koellermeier, J. Li, R. Li, and M. Torrilhon. Model reduction of kinetic equations by operator projection. *Journal of Statistical Physics*, 162(2):457–486, 2016.

[2] E. D. Fernández-Nieto, J. Garres-Díaz, A. Mangeney, and G. Narbona-Reina. A multilayer shallow model for dry granular flows with the μ (i)-rheology: Application to granular collapse on erodible beds. *Journal of Fluid Mechanics*, 798:643–681, 2016.

[3] J. Garres-Díaz, M. J. C. Díaz, J. Koellermeier, and T. M. de Luna. Shallow water moment models for bedload transport problems. *accepted by Adv. Appl. Math. Mech.*, 2021.

[4] Q. Huang, J. Koellermeier, and W.-A. Yong. Equilibrium stability analysis of hyperbolic shallow water moment equations. *submitted*.

[5] J. Koellermeier. *Derivation and numerical solution of hyperbolic moment equations for rarefied gas flows*. PhD thesis, 2017.

[6] J. Koellermeier and E. Pimentel-Garcia. Steady states and well-balanced schemes for shallow water moment equations with topography. *submitted*.

[7] J. Koellermeier and M. Rominger. Analysis and numerical simulation of hyperbolic shallow water moment equations. *Communications in Computational Physics*, 28(3):1038–1084, 2020.

[8] J. Koellermeier, R. P. Schaerer, and M. Torrilhon. A framework for hyperbolic approximation of kinetic equations using quadrature-based projection methods. *Kinetic and Related Models*, 7(3):531–549, 2014.

[9] J. Kowalski and M. Torrilhon. Moment approximations and model cascades for shallow flow. *Communications in Computational Physics*, 25(3):669–702, 2019.

[10] E. Meyer-Peter and R. Müller. Formulas for bed-load transport. Technical report, 1948.

[11] S. B. Savage and K. Hutter. The motion of a finite mass of granular material down a rough incline. *Journal of Fluid Mechanics*, 199(2697):177–215, 1989.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 253–257)

Stability of a one degree of freedom Hamiltonian system in a case of zero quadratic and cubic terms

Víctor Lanchares[1], **Boris Bardin**[2]
1. *Universidad de La Rioja, Spain*
2. *Moscow Aviation Institute, Russia*

Abstract

We consider the stability of the equilibrium position of a periodic Hamiltonian system with one degree of freedom. It is supposed that the series expansion of the Hamiltonian function, in a small neighborhood of the equilibrium position, does not include terms of second and third degree. Moreover, we focus on a degenerate case, when fourth-degree terms in the Hamiltonian function are not enough to obtain rigorous conclusions on stability or instability. A complete study of the equilibrium stability in the above degenerate case is performed, giving sufficient conditions for instability and stability in the sense of Lyapunov. The above conditions are expressed in the form of inequalities with respect to the coefficients of the Hamiltonian function, normalized up to sixth-degree terms inclusive.

1. Introduction

Let us consider a one degree of freedom Hamiltonian system, periodically dependent on time, defined by the canonical differential equations

$$\frac{dx}{dt} = \frac{\partial H}{\partial y}, \quad \frac{dy}{dt} = -\frac{\partial H}{\partial x}. \tag{1.1}$$

We assume that the origin, $x = y = 0$, is an equilibrium position and that the Hamiltonian function $H = H(x, y, t)$ can be expanded in a convergent power series in a sufficiently small neighborhood of the origin. That is,

$$H(x, y, t) = \sum_{k=2}^{\infty} H_k(x, y, t), \qquad H_k(x, y, t) = \sum_{\nu+\mu=k} h_{\nu\mu} x^\nu y^\mu, \tag{1.2}$$

where ν and μ are nonnegative integers and the coefficients $h_{\nu\mu}$ are continuous 2π periodic functions of time, t. We also assume that a resonance of first or second order takes place in system (1.1). That is, the corresponding linear system has multiple characteristic multipliers. In particular, $\rho_{1,2} = 1$, for a first order resonance, and $\rho_{1,2} = -1$, for a second order resonance. In addition, the monodromy matrix is supposed to be diagonal. In the case it is nondiagonal, the problem of stability, in the sense of Lyapunov, has been completely solved [3, 8].

Under these assumptions, the origin is linearly stable, but nonlinear analysis is necessary to obtain a rigorous result about stability in the Lyapunov sense. Thus, terms of order three or higher in the Hamiltonian function $H(x, y, t)$ must be taken into account. It can be seen that, after a series of canonical change variables, the Hamiltonian function $H(x, y, t)$ can be brought to the following form [10, 12]

$$H(q, p, t) = \sum_{k=3}^{N} H_k(q, p) + \sum_{k=N+1}^{\infty} H_k(q, p, t), \qquad H_k = \sum_{\nu+\mu=k} h_{\nu\mu} q^\nu p^\mu, \tag{1.3}$$

where, for $3 \le k \le N$ (N can be set arbitrarily large), the coefficients $h_{\nu\mu}$ in H_k are real numbers, whereas, for $k > N$, they are T-periodic functions of time t.

The stability of the origin for the system (1.1) with the Hamiltonian (1.3), in the case $H_3 \neq 0$, has been studied in [10, 11] and we consider here the case $H_3 \equiv 0$, which appears in the presence of second order resonance. Now, the terms of fourth order in (1.3) play the most important role in the stability analysis of the equilibrium.

After a linear canonical change of variables [10], H_4 can be brought to one of the following nine simple forms:

$$\begin{array}{lll} 1)\ q^4 + aq^2p^2 + p^4,\ a > -2, & 5)\ q^2(q^2 + p^2), & 9)\ q^4. \\ 2)\ q^4 + aq^2p^2 + p^4,\ a < -2 & 6)\ q^2p^2, & \\ 3)\ q^4 + aq^2p^2 - p^4, a \in \mathbb{R}, & 7)\ q^3p, & \\ 4)\ q^2(q^2 - p^2), & 8)\ -q^3p, & \end{array} \tag{1.4}$$

In [10], it is also proved that in the case 1) the equilibrium is stable in the sense of Lyapunov, whereas it is unstable in the cases 2), 3), 4), 7). Cases 5) and 6) are considered in [12] and [9], respectively. In particular, considering terms up to six order, sufficient conditions for stability and instability in the Lyapunov sense are derived.

We concentrate our attention on the case 9), already considered in [7], where partial stability results are given. Our goal is to apply the results developed in [2] to derive complete and rigorous solution of the stability problem in this particular case.

2. Method of study and main result

To study the stability of the origin, it is convenient to introduce polar canonical variables by means of the canonical transformation

$$q = \sqrt{2r}\sin\varphi, \quad p = \sqrt{2r}\cos\varphi. \tag{2.1}$$

Now, the Hamiltonian function (1.3) is written as

$$H = r^2\Psi(\varphi, r) + O(r^{(N+1)/2}), \tag{2.2}$$

where

$$\Psi(\varphi, r) = \sum_{k=4}^{N} r^{\frac{k-4}{2}}\,\Psi_k(\varphi), \tag{2.3}$$

and $\Psi_k(\varphi)$ is a homogeneus function of order k with respect to $\sin\varphi$ and $\cos\varphi$.

It is shown in [5,10,14] that, if the function $\Psi_4(\varphi)$ does not have real roots, then the origin is a stable equilibrium point. This is what happens in the case 1), listed in (1.4). On the other hand, if $\Psi_4(\varphi)$ has a simple real root φ_0, such that $\frac{d\Psi_4}{d\varphi}(\varphi_0) < 0$, there is instability. This situation takes place in cases 2), 3), 4) and 7).

In the cases 5), 6) and 9) the function $\Psi_4(\varphi)$ has only multiple real roots and we say that a degeneracy takes place. To solve now the stability problem, it is necessary to consider the terms of order higher than r^2. To this end, we will use a technique for degenerate cases developed in [2]. The key idea is that simple roots of the function (2.3), coming from a multiple root of $\Psi_4(\varphi)$, play an important role for the stability problem. Thus, it is necessary to determine whether multiple roots of $\Psi_4(\varphi)$ give rise to simple distinct roots, when terms of order higher than r^2 in the Hamiltonian function (2.2) are considered. Even more, we have to ensure that additional terms of higher order cannot destroy the simple real roots of function (2.3). In this way, we introduce the concepts of main part and simple main part of a root (see [2]).

Let φ_0 be a root of multiplicity $M > 1$ of the function $\Psi_4(\varphi)$. Thus, according to the implicit function theorem [6], $\Psi(\varphi, r) = 0$ has exactly M roots approaching φ_0 with $r \to 0$. Lut us denote by $\varphi_*(r)$ one of these roots, which can be represented as a series expansion in fractional powers of r

$$\varphi_*(r) = \varphi_0 + \sum_{j=1}^{\infty} a_j r^{\frac{j}{m}}, \tag{2.4}$$

where m is an even integer ($2 \le m \le 2M$) and a_j are obtained by equating to zero the coefficients of powers of r, after substituting (2.4) into (2.3).

Definition 2.1 Let us consider the finite series

$$\varphi_q(r) = \varphi_0 + \sum_{j=1}^{q} a_j r^{\frac{j}{m}}, \tag{2.5}$$

which is obtained by omitting terms of order higher than q/m in (2.4); q is the maximal integer number such that the equation for a_q is obtained by substituting (2.5) in (2.3) and equating to zero the coefficient of $r^{\frac{\nu}{m}}$, where $\frac{\nu}{m} < \frac{N-3}{2}$. We call finite series (2.3) *main part* of root (2.4).

Definition 2.2 We say that root (2.4) has a *simple main part* if among roots of the equation $\Psi(\varphi, r) = 0$ there is not another root with the same main part.

Taking these two definitions in mind, general conditions for instability in the case of a degeneracy are given by the following theorem [2].

Theorem 2.3 *Let us consider the canonical system defined by Hamiltonian* (2.2)*. Suppose that all real roots of the function* $\Psi_4(\varphi)$ *are multiple and the function* $\Psi(\varphi, r)$ *has a real root* φ_* *of form* (2.4)*. If the root* φ_* *has a simple main part* φ_q *and, for sufficiently small* r*, the inequality* $\frac{\partial\Psi}{\partial\varphi}(\varphi_*, r) < 0$ *is satisfied, then the equilibrium* $r = 0$ *is unstable.*

As it was said previously, in the case 9), all real roots of the function $\Psi_4(\varphi) = 4\sin^4\varphi$ have multiplicity four and the use of Theorem 2.3 will be our main tool to obtain sufficient conditions for instability. To begin our analysis, we perform a series of near identity canonical transformations, in order to simplify the Hamiltonian function. This procedure has been previously introduced by Markeev to study other degenarete cases [9, 11, 12] and applied by Gutiérrez and Vidal [7] in the case we are dealing with.

Let us take $N = 6$ in (1.3). Thus, the Hamiltonian function reads as

$$H = q^4 + H_5(q,p) + H_6(q,p) + H^{(7)}(q,p,t), \tag{2.6}$$

where $H^{(7)}(q,p,t)$ is a convergent series in powers of q and p, starting from terms of degree seven or higher, whose coefficients are T-periodic functions of t.

Let us introduce new canonical variables Q, P by using a generating function $S(q,P)$ of the form

$$S(q,P) = qP + S_3(q,P) + S_4(q,P), \quad S_k(q,P) = \sum_{\nu+\mu=k} s_{\nu\mu} q^\nu P^\mu, \tag{2.7}$$

being $s_{\nu\mu}$ constant coefficients properly chosen in order to simplify the expression of the new Hamiltonian function. Taking into account the relations

$$p = \frac{\partial S}{\partial q}, \qquad Q = \frac{\partial S}{\partial P}, \tag{2.8}$$

we can express the old variables in a power series expansion of the new ones in such a way that the new Hamiltonian function, K, becomes [7]

$$\begin{aligned} K &= Q^4 + K_5(Q,P) + K_6(Q,P) + K^{(7)}(Q,P,t), \\ K_5(Q,P) &= \gamma_{23}Q^2P^3 + \gamma_{14}QP^4 + \gamma_{05}P^5, \\ K_6(Q,P) &= \gamma_{24}Q^2P^4 + \gamma_{15}QP^5 + \gamma_{06}P^6. \end{aligned} \tag{2.9}$$

The coefficients γ_{ij} in (2.9) are related to the coefficients of Hamiltonian (1.3) through the following identities [7]

$$\begin{aligned} &\gamma_{23} = h_{23}, \quad \gamma_{14} = h_{14}, \quad \gamma_{05} = h_{05}, \\ &\gamma_{24} = h_{24} - \tfrac{3}{7}h_{32}^2 + \tfrac{7}{4}h_{50}h_{14} - \tfrac{1}{8}h_{23}h_{41}, \\ &\gamma_{15} = h_{15} - \tfrac{1}{2}h_{32}h_{23} + \tfrac{1}{4}h_{41}h_{14} + \tfrac{5}{2}h_{50}h_{05}, \\ &\gamma_{06} = h_{06} - \tfrac{1}{4}h_{32}h_{14} + \tfrac{5}{8}h_{41}h_{05}. \end{aligned} \tag{2.10}$$

The main result of our stability study can be formulated in terms of the coefficients of the Hamiltonian (2.9) and it is collected in the following Theorem [4].

Theorem 2.4 *Let us consider the Hamiltonian system defined by* (2.9)*, then*

1. *If at least one of the inequalities* $\gamma_{05} \neq 0$, $\gamma_{14} \neq 0$ *or* $\gamma_{23}^2 - 4\gamma_{06} > 0$ *is fulfilled, then the origin is an unstable equilibrium.*
2. *If* $\gamma_{05} = \gamma_{14} = 0$ *and* $\gamma_{23}^2 - 4\gamma_{06} < 0$ *, then the origin is stable in the sense of Lyapunov.*
3. *In the case* $\gamma_{05} = \gamma_{14} = 0$ *and* $\gamma_{23}^2 - 4\gamma_{06} = 0$ *and* $\gamma_{15} \neq 0$ *the origin is unstable.*

3. Sketch of the proof

A complete proof of Theorem 2.4 is given in [4]. Here we outline the main ideas. To begin with, we introduce a scaling of the variables that will help us to see which are the relevant terms contributing to the proper splitting of the multiple roots. In this way, we introduce the following canonical transformation

$$Q = \varepsilon\bar{Q}, \qquad P = \varepsilon^\kappa \bar{P}, \tag{3.1}$$

and the Hamiltonian (2.9) reads as

$$\begin{aligned} K &= Q^4 + K_5(Q,P) + K_6(Q,P) + K^{(7)}(Q,P,t), \\ K_5(Q,P) &= \varepsilon^{3\kappa-2}\gamma_{23}Q^2P^3 + \varepsilon^{4\kappa-3}\gamma_{14}QP^4 + \varepsilon^{5\kappa-4}\gamma_{05}P^5, \\ K_6(Q,P) &= \varepsilon^{4\kappa-2}\gamma_{24}Q^2P^4 + \varepsilon^{5\kappa-3}\gamma_{15}QP^5 + \varepsilon^{6\kappa-4}\gamma_{06}P^6, \end{aligned} \tag{3.2}$$

where bars have been suppressed. The scaling introduces an ordering which depends on the exponent κ. Indeed, a different exponent κ gives rise to a different ordering and, to solve the degeneracy, we look for a proper choice of κ. To this end, we introduce the concept of leading exponent of a monomial.

Definition 3.1 We say that the leading exponent of a monomial $P^\alpha Q^\beta$ is $\lambda(\alpha,\beta)$ if the scaling (3.1), with $\kappa = \lambda(\alpha,\beta)$, places this monomial at the same order than Q^4. That is $\lambda(\alpha,\beta) = \frac{4-\alpha}{\beta}$.

monomial	P^5	QP^4	Q^2P^3	P^6	QP^5	Q^2P^4	P^7	QP^6	Q^2P^5	P^8
$\kappa(\alpha,\beta)$	4/5	3/4	2/3	2/3	3/5	1/2	4/7	1/2	2/5	1/2

Tab. 1 Leading exponent $\lambda(\alpha,\beta)$ for different monomials.

It can be seen that the first term that can solve the degeneracy is the one with the largest leading exponent [4]. Table 1 shows the leading exponent for those monomials appearing in the Hamiltonian function up to six order and also monomials of order seven and eight. We can see that the monomial P^5 has the maximum leading exponent and it is the first term to be taken into account to proper split the multiple root. If this term fails, then the next term to be considered is QP^4 and so on.

Now, we move to polar coordinates (2.1) in order to apply Theorem 2.3. The Hamiltonian function in the form (2.2) is given by

$$\begin{aligned} K &= 4r^2(\sin^4\varphi + r^{1/2}\Psi_5(\varphi) + r\Psi_6(\varphi)) + \tilde{K}(\varphi,r,t),\\ \Psi_5(\varphi) &= \sqrt{2}(\gamma_{23}\sin^2\varphi\cos^3\varphi + \gamma_{14}\sin\varphi\cos^4\varphi + \gamma_{05}\cos^5\varphi),\\ \Psi_6(\varphi) &= 2(\gamma_{24}\sin^2\varphi\cos^4\varphi + \gamma_{15}\sin\varphi\cos^5\varphi + \gamma_{06}\cos^6\varphi). \end{aligned} \tag{3.3}$$

Our goal is to analyze the real roots of the equation

$$\Psi(\varphi,r) \equiv \sin^4\varphi + r^{1/2}\Psi_5(\varphi) + r\Psi_6(\varphi) = 0, \tag{3.4}$$

emanating from multiple roots $\varphi = 0$ and $\varphi = \pi$ of the function $\sin^4\varphi$.

To determine the main part of the roots, we introduce a fractional power series of the form (2.4), where the fractional exponents are chosen according to the leading exponent. In this way, if $\lambda(\alpha,\beta)$ is the maximum leading exponent, the first fractional exponent with nonzero coefficient in (2.4) is given by

$$\frac{j}{m} = \frac{1-\lambda(\alpha,\beta)}{2\lambda(\alpha,\beta)}. \tag{3.5}$$

For instance, if $\gamma_{05} \neq 0$ we consider the series

$$\varphi_1 = a_1 r^{1/8} + a_2 r^{2/8} + \cdots, \qquad \varphi_2 = \pi + b_1 r^{1/8} + b_2 r^{2/8} + \cdots.$$

It is almost straightforward to check that, in all the cases of instability of Theorem 2.4, the conditions of Theorem 2.3 are satisfied and we are done.

To prove item 2. of Theorem 2.4, we introduce proper action-angle variables. In this sense, we rewrite the Hamiltonian function as

$$H = H_0(Q,P) + \hat{H}(Q,P,t)\,, \tag{3.6}$$

where

$$H_0(Q,P) = (Q^2 + \alpha P^3)^2 + \beta P^6, \quad \hat{H}(Q,P,t) = \gamma_{15}QP^5 + \gamma_{24}Q^2P^4 + H^{(7)}(Q,P,t)\,. \tag{3.7}$$

The coefficients α and β read

$$\alpha = \frac{1}{2}\gamma_{23}, \quad \beta = \gamma_{06} - \frac{1}{4}\gamma_{23}^2\,.$$

We note that H_0 is a positive definite function, provided that, under the conditions of item 2., $\beta > 0$. Thus, the origin of the truncted system with Hamiltonian H_0 is stable and it is encircled by a family of closed curves, describing periodic motion in a sufficient small neighborhood. Let be the action variable

$$I = \frac{1}{2\pi}\oint P(Q,h)\,dQ, \tag{3.8}$$

where the integral is calculated along a closed phase trajectory. Every trajectory is completely defined by h, where h is a constant such that $H_0 = h$. A direct calculation shows that

$$I = h^{\frac{5}{12}}J_0\,, \tag{3.9}$$

being J_0 a constant. Then, it follows that H_0 reduces to

$$h(I) = \left(\frac{I}{J_0}\right)^{\frac{12}{5}}. \tag{3.10}$$

Moreover, it can be proved that, in action-angle variables, the Hamiltonian takes the form

$$\Gamma = h(I) + h_1(I, w, t), \tag{3.11}$$

where $h_1(I, w, t) = o(h(I))$. However, the nondegeneracy condition

$$\frac{d^2h}{dI^2} = \frac{84I^{\frac{2}{5}}}{25J_0^{\frac{12}{5}}} \neq 0 \tag{3.12}$$

is fulfilled for $0 < I \ll 1$. Thus, the Arnold-Moser theorem [1, 13] guarantees the stability of the equilibrium position of the original canonical system.

Acknowledgements

The first author acknowledges support from the Spanish Ministry of Science and Innovation through project MTM2017-88137-C2-2-P, and from the University of La Rioja through project REGI 2018751. The second author performed his part of the work at the Moscow Aviation Institute (National Research University) within the framework of the state assignment (project No 3.3858.2017/4.6).

References

[1] V. I. Arnold. Small denominators and problems of stability of motion in classical and celestial mechanics. *Russ. Math. Surv.*, 18(6):85–192, 1963.

[2] B. Bardin and V. Lanchares. On the Stability of Periodic Hamiltonian Systems with One Degree of Freedom in the Case of Degeneracy. *Regul. Chaotic Dyn.*, 20(6):627–648, 2015.

[3] B. S. Bardin. On the Stability of a Periodic Hamiltonian System with One Degree of Freedom in a Transcendental Case. *Doklady Mathematics*, 97(2):161–163, 2018.

[4] B. S. Bardin and V. Lanchares. Stability of a one-degree-of-freedom canonical system in the case of zero quadratic and cubic part of a Hamiltonian. *Regul. Chaotic Dyn.*, 25(3):237–249, 2020.

[5] H. E. Cabral and K. R. Meyer. Stability of equilibria and fixed points of conservative systems. *Nonlinearity*, 12(5):1351–1362, 1999.

[6] E. Goursat. *A course in mathematical analysis Vol. 2, Part 1: Functions of a complex variable*. Dover Publications Inc., New York, 1959.

[7] R. Gutiérrez and C. Vidal. Stability of Equilibrium Points for a Hamiltonian System with One Degree of Freedom in One Degenerate Case. *Regul. Chaotic Dyn.*, 22:880–892, 2017.

[8] A. P. Ivanov and A. G. Sokolsky. On the stability of a nonautonomous Hamiltonian system under a parametric resonance of essential type. *J. Appl. Math. Mech.*, 44(6):963–970, 1980.

[9] A. P. Markeev. On the fixed points stability for the area-preserving maps. *Rus. J. Nonlin. Dyn.*, 11:503–545, 2014. (Russian).

[10] A. P. Markeev. Simplifying the structure of the third and fourth degree forms in the expansion of the Hamiltonian with a linear transformation. *Rus. J. Nonlin. Dyn.*, 10(4):447–464, 2014. (Russian).

[11] A. P. Markeev. On the Birkhoff Transformation in the Case of Complete Degeneracy of the Quadratic Part of the Hamiltonian. *Regul. Chaotic Dyn.*, 20:309–316, 2015.

[12] A. P. Markeev. On the problem of the stability of a Hamiltonian system with one degree of freedom on the bundaries of parametric resonance. *J. Appl. Math. Mech.*, 80:1–6, 2016.

[13] C.L. Siegel and J. K. Moser. *Lectures on celestial mechanics*. Springer-Verlag, New York, 1971.

[14] A. G. Sokolsky. On stability of an autonomous Hamiltonian system with two degrees of freedom under first-order resonance. *J. Appl. Math. Mech.*, 41(1):20–28, 1977.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 258–264)

Minimal complexity of subharmonics in a class of planar periodic predator-prey models

Julián López-Gómez[1], **Eduardo Muñoz-Hernández**[1], **Fabio Zanolin** [2]
1. julian@mat.ucm.es *and* eduardmu@ucm.es *Universidad Complutense de Madrid, Spain*
2. fabio.zanolin@uniud.it *Università degli Studi di Udine, Italy*

Abstract

This contribution analyzes the existence of nT-periodic coexistence states, for $n \geq 1$, in two classes of non-autonomous predator-prey Volterra systems with periodic coefficients. In the first place, when the model is non-degenerate it is shown that the Poincaré–Birkhoff twist theorem can be applied to get the existence of subharmonics of arbitrary order. In the second place, it will be analyzed a *degenerate* predator-prey model introduced in [9] and [5] and, then, deeply studied in [7]. By analyzing the iterates of the Poincaré map of the system, it is shown that it admits nontrivial nT-periodic coexistence states for every $n \geq 2$.

1. Introduction

In this contribution, we study the existence of positive subharmonics of arbitrary order (nT-periodic coexistence states) of the periodic Volterra predator-prey model

$$\begin{cases} u' = \lambda\alpha(t)u(1-v), \\ v' = \lambda\beta(t)v(-1+u), \end{cases} \tag{1.1}$$

where $\lambda > 0$ is regarded as a parameter, and, for some $T > 0$, $\alpha(t)$ and $\beta(t)$ are T-periodic real continuous functions. Throughout this note, we set

$$A := \int_0^T \alpha(s)ds \quad \text{and} \quad B := \int_0^T \beta(s)ds.$$

Two different cases can arise according to whether, or not, the following condition holds

$$\text{supp } \alpha \cap \text{supp } \beta \neq \emptyset. \tag{1.2}$$

In this *non-degenerate* situation, which has been sketched in Figure 1, the existence of subharmonics of arbitrary order, for sufficiently large λ, can be obtained through an updated version of the celebrated Poincaré–Birkhoff twist theorem.

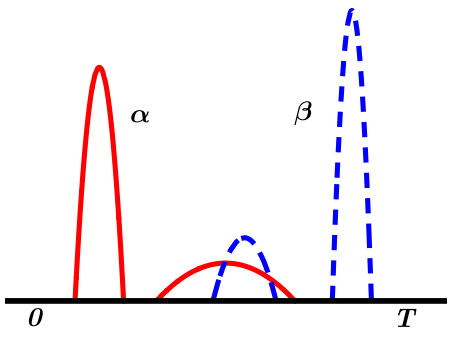


Fig. 1 α (continuous line) and β (dashed line) satisfying (1.2).

However, in the *degenerate* case when, instead of (1.2), the next condition holds

$$\text{supp } \alpha \cap \text{supp } \beta = \emptyset, \tag{1.3}$$

then the Poincaré–Birkhoff theorem is unable to provide, in general, with subharmonics of arbitrary order, unless $\alpha(t)$ and $\beta(t)$ have some special nodal structure. An admissible distribution of α and β is sketched in Figure 2.

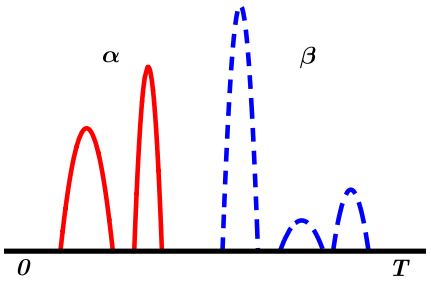


Fig. 2 α (continuous line) and β (dashed line) satisfying (1.3).

2. The non-degenerate case

The non-degenerate case when (1.2) holds has been recently analyzed in [8] by adapting, in a sophisticated way, some original ideas going back to [3] (later revised and applied in [2] and [10]), where a Poincaré–Birkhoff version for Hamiltonian systems was delivered. Note that the change of variables

$$x = \log u, \qquad y = \log v,$$

transforms (1.1) into the planar Hamiltonian system

$$\begin{cases} x' = -\lambda\alpha(t)(e^y - 1), \\ y' = \lambda\beta(t)(e^x - 1). \end{cases} \tag{2.1}$$

The updated version of the Poincaré–Birkhoff twist theorem that will be used reads as follows:

Theorem 2.1 (Poincaré–Birkhoff) *Assume that there exist* $0 < \varrho_0 < \varrho_1$ *and a positive integer* ω *such that*

$$\mathrm{rot}_{\varrho_0}[(x_0, y_0); [0, nT]] > \omega \quad \textit{and} \quad \mathrm{rot}_{\varrho_1}[(x_0, y_0); [0, nT]] < \omega,$$

where

$$\mathrm{rot}_{\rho}[(x_0, y_0); [0, nT]] = \frac{\theta(nT) - \theta(0)}{2\pi}$$

with $||(x_0, y_0)|| = \rho$; $\theta(t)$ *being the angular polar coordinate of the solution starting at* (x_0, y_0)*, say* $(x(t), y(t))$*. Then,* (2.1) *admits, at least, two* nT*-periodic solutions lying in different periodicity classes with rotation number* ω*.*

As a consequence of Theorem 2.1, the following result holds.

Theorem 2.2 *Assume* (1.2)*. Then, for every positive integers* ω *and* n*, there exists* $\lambda_n^\omega > 0$ *such that* (2.1) *possesses, at least, two* nT*-periodic solutions with rotation number* ω *for every* $\lambda > \lambda_n^\omega$*.*

Proof Firstly, attention will be focused into the small solutions of (2.1). Obviously, there exists $\varepsilon > 0$ such that

$$(e^\xi - 1)\xi \geq \frac{\xi^2}{2} \quad \text{if} \quad |\xi| < \varepsilon. \tag{2.2}$$

Choose (x_0, y_0) sufficiently close to $(0, 0)$, say $|(x_0, y_0)| \leq \varrho_0$, so that the solution of (2.1) with $(x(0), y(0)) = (x_0, y_0)$, say $(x(t), y(t))$, satisfy $|(x(t), y(t))| < \varepsilon$ for all $t \in [0, nT]$. This is possible by continuous dependence on the initial conditions.

According to (1.2), there are $\tau \in (0, T)$ and $\delta > 0$ such that $\alpha(t)\beta(t) > 0$ for every $t \in [\tau - \delta, \tau + \delta] \subsetneq [0, T]$. Thus,

$$\zeta := \min_{t \in [\tau - \delta, \tau + \delta]} \{\alpha(t), \beta(t)\} > 0.$$

Consequently, due to (2.1) and (2.2), we obtain that, for every $t \in [0, nT]$,

$$\theta'(t) = \frac{y'(t)x(t) - x'(t)y(t)}{x^2(t) + y^2(t)} = \frac{\lambda\beta(t)(e^{x(t)} - 1)x(t) + \lambda\alpha(t)(e^{y(t)} - 1)y(t)}{x^2(t) + y^2(t)} \geq \frac{\lambda}{2}\frac{\beta(t)x^2(t) + \alpha(t)y^2(t)}{x^2(t) + y^2(t)} \geq \frac{\lambda\zeta}{2}. \tag{2.3}$$

Hence, owing to (2.3),

$$\mathrm{rot}_{\varrho_0}[(x_0,y_0);[0,nT]] = \frac{\theta(nT)-\theta(0)}{2\pi} = \frac{1}{2\pi}\int_0^{nT}\theta'(s)ds \geq \frac{n}{2\pi}\int_{\tau-\delta}^{\tau+\delta}\theta'(s)ds \geq \frac{n\lambda\zeta 2\delta}{2\pi}.$$

Therefore,

$$\mathrm{rot}_{\varrho_0}[(x_0,y_0);[0,nT]] > \omega \quad \text{if} \quad \lambda > \frac{\pi\omega}{n\zeta\delta} =: \lambda_n^{\omega}.$$

On the other hand, sufficiently large solutions do not rotate. Indeed, arguing by contradiction, assume that, for some solution $(x(t),y(t))$, we have that $\mathrm{rot}_{\varrho}[(x_0,y_0);[0,nT]] \geq 1$. Then, e.g., it must cross entirely the third quadrant. So, there exists $[\tau_1,\tau_2]\subset[0,nT]$ such that $y(\tau_1)=0$, $x(\tau_1)<0$, $y(\tau_2)<0$, $x(\tau_2)=0$, and $x(t)<0$ and $y(t)<0$ for all $t\in(\tau_1,\tau_2)$. Hence, for every $t\in[\tau_1,\tau_2]$, we find that

$$|x(t)| = |\lambda\int_t^{\tau_2}\alpha(s)(e^{y(s)}-1)ds| \leq \lambda\int_0^{nT}\alpha(s)ds = \lambda nA,$$

$$|y(t)| = |\lambda\int_{\tau_1}^{t}\beta(s)(e^{x(s)}-1)ds| \leq \lambda\int_0^{nT}\beta(s)ds = \lambda nB.$$

Therefore, if there exists $\tau_0\in[0,nT]$ such that $(x(\tau_0),y(\tau_0))$ lies in the third quadrant and $x^2(\tau_0)+y^2(\tau_0) > R_1^2 := \lambda^2n^2(A^2+B^2)$, then $(x(t),y(t))$ cannot cross the entire third quadrant. Similarly, since $|e^{x(t)}-1|$ (resp. $|e^{y(t)}-1|$) are bounded in the second (resp. fourth) quadrant, there exists $R_2>0$ (resp. $R_3>0$) such that $x^2(t)+y^2(t)<R_2^2$ (resp. $x^2(t)+y^2(t)<R_3^2$) if the solution crosses the second (resp. fourth) quadrant. Therefore, taking $R:=\max\{R_1,R_2,R_3\}$, if $(x(\hat{t}),y(\hat{t}))$ lies in the second, third or fourth quadrants and $x(\hat{t})^2+y(\hat{t})^2>R$ for some $\hat{t}\in[0,nT]$, then, the solution $(x(t),y(t))$ cannot cross the corresponding quadrant.

Finally, let $s_0\in[0,nT]$ be such that $x(s_0)=0$ and $0<y(s_0)\leq R$, and consider the maximal interval $[s_1,s_0]\subset[0,s_0]$ such that $x(t),y(t)\geq 0$ for all $t\in[s_1,s_0]$. By (2.1), $y(t)$ is non-decreasing in $[s_1,s_0]$ and, hence, $0\leq y(t)\leq R$ for all $t\in[s_1,s_0]$. Since $y(t)$ is bounded, $x(t)$ must be bounded too. Thus, there exists a constant $R_*\geq R>0$ such that if $x^2(\tilde{t})+y^2(\tilde{t})>R_*^2$ for some $\tilde{t}\in[0,nT]$ with $(x(\tilde{t}),y(\tilde{t}))$ lying in the first quadrant, then the solution $(x(t),y(t))$ cannot cross neither the second, nor the third and fourth quadrants. Therefore, $x(0)^2+y(0)^2=\varrho_1^2>R_*^2$ implies that $\mathrm{rot}_{\varrho_1}[(x(0),y(0));[0,nT]]<1$ and hence, the hypothesis of Theorem 2.1 holds for every $\lambda>\lambda_n^{\omega}$, which ends the proof. □

Remark 2.3 Although Theorem 2.2 has a counterpart for a more general class of Hamiltonian systems of the type

$$\begin{cases} x' = -\lambda\alpha(t)f(y), \\ y' = \lambda\beta(t)g(x), \end{cases}$$

where f and g satisfy certain boundedness and sign conditions (see [8, Sec. 2]), in this note we are focusing our attention into the predator-prey model (1.1). Thus, we restrict ourselves to consider f and g as they appear in (2.1).

3. The degenerate case

To analyze the problem (1.1) under the condition (1.3), we suppose that either

$$\mathrm{supp}\,\alpha\subset[t_0^1,t_1^1] \quad \text{and} \quad \mathrm{supp}\,\beta\subset[t_0^2,t_1^2], \tag{3.1}$$

or else

$$\mathrm{supp}\,\beta\subset[t_0^1,t_1^1] \quad \text{and} \quad \mathrm{supp}\,\alpha\subset[t_0^2,t_1^2], \tag{3.2}$$

for some partition

$$0\leq t_0^1<t_1^1\leq t_0^2<t_1^2\leq T.$$

By (1.3), the system (1.1) can be integrated. Thus, in case (3.1) we have that, for every $t\in[0,T]$,

$$u(t) = u_0e^{(1-v_0)\lambda\int_0^t\alpha(s)ds}, \quad v(t) = v_0e^{(u(T)-1)\lambda\int_0^t\beta(s)ds},$$

whereas, in case (3.2),

$$u(t) = u_0e^{(1-v(T))\lambda\int_0^t\alpha(s)ds}, \quad v(t) = v_0e^{(u_0-1)\lambda\int_0^t\beta(s)ds},$$

for all $t\in[0,T]$. Hence, in case (3.1), the T-time Poincaré map is

$$(u_1,v_1) := \mathcal{P}_1(u_0,v_0) := (u(T),v(T)) = (u_0e^{(1-v_0)\lambda A}, v_0e^{(u_1-1)\lambda B}).$$

while, in case (3.2), is given through

$$(u_1, v_1) := \mathcal{P}_1(u_0, v_0) := (u(T), v(T)) = (u_0 e^{(1-v_1)\lambda A}, v_0 e^{(u_0-1)\lambda B}).$$

Consequently, iterating n times these maps, it becomes apparent that either

$$\begin{aligned}(u_n, v_n) &:= \mathcal{P}_n(u_0, v_0) = \mathcal{P}_1^n(u_0, v_0) := (u(nT), v(nT)) = (u_{n-1} e^{(1-v_{n-1})\lambda A}, v_{n-1} e^{(u_n-1)\lambda B}) \\ &= (u_0 e^{(n-v_0-v_1-\cdots-v_{n-1})\lambda A}, v_0 e^{(u_1+u_2+\cdots+u_n-n)\lambda B})\end{aligned} \tag{3.3}$$

under condition (3.1), or

$$\begin{aligned}(u_n, v_n) &:= \mathcal{P}_n(u_0, v_0) = \mathcal{P}_1^n(u_0, v_0) := (u(nT), v(nT)) = (u_{n-1} e^{(1-v_n)\lambda A}, v_{n-1} e^{(u_{n-1}-1)\lambda B}) \\ &= (u_0 e^{(n-v_1-v_2-\cdots-v_n)\lambda A}, v_0 e^{(u_0+u_1+\cdots+u_{n-1}-n)\lambda B})\end{aligned} \tag{3.4}$$

under condition (3.2). By the uniqueness for the underlying Cauchy problem, the nT-periodic coexistence states of (1.1) are given by the positive fixed points of $\mathcal{P}_n$. Thus, by (3.3) and (3.4), we are driven to solve the system

$$\begin{cases} n = u_0 + u_1 + \cdots + u_{n-1}, \\ n = v_0 + v_1 + \cdots + v_{n-1}. \end{cases} \tag{3.5}$$

Naturally, the $u_i's$ and the $v_i's$ are different depending on (3.1) or (3.2). Our next result deals with the T-periodic and $2T$-periodic cases.

Theorem 3.1 *Assume* (3.1) *or* (3.2). *Then,* (1.1) *does not admit any non-trivial T-periodic coexistence state. Moreover,* (1.1) *possesses exactly two non-trivial $2T$-periodic coexistence states if, and only if,*

$$\lambda > \frac{2}{\sqrt{AB}}. \tag{3.6}$$

Proof First, suppose (3.1). Then, by (3.3), $(u_1, v_1) = \mathcal{P}_1(u_0, v_0) = (u_0, v_0)$ if, and only if, $v_0 = 1$ and $u_0 = u_1 = 1$. Thus, $(u(t), v(t))$ is a T-periodic coexistence state if, and only if, $(u(t), v(t)) = (1, 1)$, which is the equilibrium of the system (1.1). Similarly,

$$(u_2, v_2) = \mathcal{P}_2(u_0, v_0) = (u_0 e^{(2-v_0-v_1)\lambda A}, v_0 e^{(u_1+u_2-2)\lambda B}) = (u_0, v_0)$$

if, and only if,

$$2 = v_0 + v_1 \quad \text{and} \quad 2 = u_1 + u_2 = u_0 + u_1,$$

or, equivalently,

$$2 = v_0 + v_0 e^{(u_1-1)\lambda B} = v_0 + v_0 e^{(1-u_0)\lambda B} \quad \text{and} \quad 2 = u_0 + u_0 e^{(1-v_0)\lambda A}. \tag{3.7}$$

Hence,

$$u_0 = \frac{2}{1 + e^{(1-v_0)\lambda A}}.$$

Setting $x := v_0$ and substituting u_0 in the first equation of (3.7) it is apparent that the $2T$-periodic coexistence states are given by the zeros of the map

$$\varphi(x) = x\left[e^{\frac{e^{(1-x)\lambda A}-1}{e^{(1-x)\lambda A}+1}\lambda B} + 1\right] - 2.$$

It is easily seen that

$$\varphi(x) < 0 \ \text{ if } \ x \le 0, \quad \varphi(1) = 0, \quad \varphi(x) > 0 \ \text{ if } \ x \ge 2, \quad \text{and} \ \ \varphi'(1) = 2 - \lambda^2 \frac{AB}{2}.$$

By (3.6), we find that $\varphi'(1) < 0$. Thus, there are $0 < x_1 < 1 < x_2 < 2$ such that $\varphi(x_1) = \varphi(x_2) = 0$, i.e., (1.1) has two non-trivial $2T$-periodic coexistence states. The uniqueness follows by analyzing φ'', much like in the proof of [7, Th. 2.1]. Similarly, one can derive the necessity of (3.6). This ends the proof when (3.1) is satisfied.

Now, assume (3.2). Then, by (3.4) and arguing as above, we find that

$$(u_1, v_1) = \mathcal{P}_1(u_0, v_0) = (u_0, v_0)$$

if, and only if, $(u(t), v(t)) = (1, 1)$. Moreover,

$$(u_2, v_2) = \mathcal{P}_2(u_0, v_0) = (u_0 e^{(2-v_1-v_2)\lambda A}, v_0 e^{(u_0+u_1-2)\lambda B}).$$

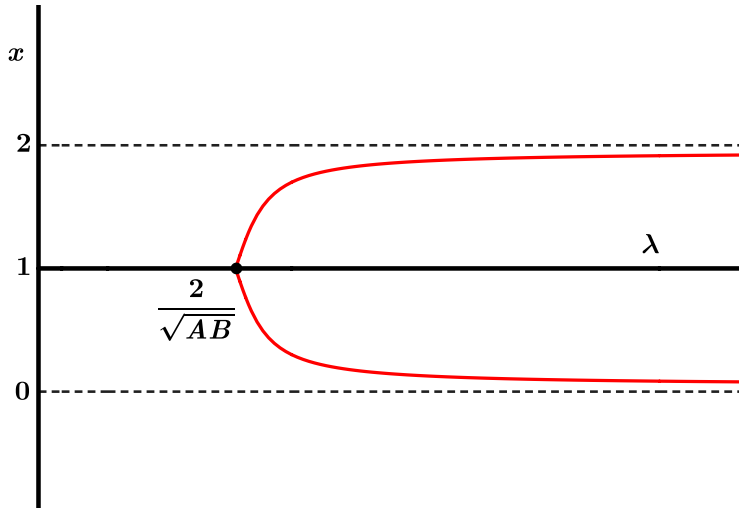


Fig. 3 Global bifurcation diagram to $2T$-periodic coexistence states.

Thus, in this occasion, the $2T$-periodic coexistence states of (1.1) are given by the zeros of the map

$$\psi(x) = x[e^{\frac{1-e^{(x-1)\lambda A}}{1+e^{(x-1)\lambda A}}\lambda B} + 1] - 2.$$

Those with $x \neq 1$ provide us with the non-trivial $2T$-periodic coexistence states of (1.1). Adapting the previous argument, it readily follows the same result as before. This concludes the proof. □

Figure 3 shows the global bifurcation diagram of $2T$-periodic coexistence states of (1.1) in each of the cases (3.1), or (3.2). In both cases, they bifurcate supercritically from the equilibrium $(1, 1)$ at $\lambda = \frac{2}{\sqrt{AB}}$.

Subsequently, we will make explicit the dependence of the functions φ and ψ defined in the proof of Theorem 3.1 on the variables x and λ. Since

$$\varphi(x, \lambda) = \psi(x, -\lambda),$$

dealing with the case when $\lambda > 0$ under condition (3.1) is the same as dealing with the case when $\lambda < 0$ under (3.2), in the sense that the $2T$-periodic coexistence states of (1.1) in each of these cases must coincide. From a biological point of view, this is rather natural. Actually, it is equivalent to inter-exchanging the role of the prey and the predator in the model.

Our last result provides us with the nT-periodic coexistence states of (1.1) when $n \geq 2$ in case (3.1). To get it, we must impose the following condition

$$A = B \quad \text{and} \quad u_0 = v_0 = x. \tag{3.8}$$

Theorem 3.2 *Assume* (3.8). *Then, for every* $\lambda > \frac{2}{A}$, (1.1) *admits, at least,* n *coexistence states with period* nT *if* n *is even, and* $n - 1$ *coexistence states with period* nT *if* n *is odd.*

Proof First, we set $E_0(x) = 1$, and

$$E_n(x) := \begin{cases} e^{[\frac{n+1}{2} - x(E_0(x)+E_2(x)+\cdots+E_{n-1}(x))]\lambda A} & \text{if} \quad n \in 2\mathbb{N}+1, \\ e^{[x(E_1(x)+E_3(x)+\cdots+E_{n-1}(x)) - \frac{n}{2}]\lambda A} & \text{if} \quad n \in 2\mathbb{N}. \end{cases} \tag{3.9}$$

By (3.8), it turns out that

$$\varphi_n(x) = \varphi_{n-1}(x) - 1 + E_{n-1}(x),$$

where $\varphi_1(x) = x - 1$, is the map whose zeros provide us with the nT-periodic coexistence states of (1.1). As the analysis of these maps is fraught with a number of serious technical difficulties, in order to obtain some information concerning the nT-periodic coexistence states of (1.1), we are driven to analyze the variational equations of these maps at the trivial curve $(\lambda, 1)$,

$$p_n(\lambda) := \frac{\partial \varphi_n}{\partial x}(\lambda, 1).$$

It is easy to prove that $p_n(\lambda)$ is a sequence of polynomials in the indeterminate λ that satisfy the recursive formula

$$p_n(\lambda) = [2 - (-1)^n A\lambda]p_{n-1}(\lambda) - p_{n-2}(\lambda),$$

where $p_1(\lambda) = 1$ and $p_2(\lambda) = 2 - A\lambda$. From this recursive formula, it can be shown that any root of p_n is real and algebraically simple. Thanks to these features, for any given $r \in p_n^{-1}(0)$, the transversality condition

$$\frac{dp_n(r)}{d\lambda}(N[p_n(r)]) \oplus R[p_n(r)] = \mathbb{R}$$

holds, where N and R stand for the null space and the rank, respectively, of the underlying one-dimensional operators. Thus, for any given $r \in p_n^{-1}(0)$, the algebraic multiplicity of Esquinas and López-Gómez [4] equals one at every point $(r, 1)$. So, according to Crandall and Rabinowitz [1, Th. 1.7], a local bifurcation occurs at every point $(r, 1)$. Moreover, by the unilateral theorem of López-Gómez [6, Th. 6.4.3], the underlying subcomponents of nT-periodic coexistence states are unbounded in λ, in agrement with Rabinowitz [11]. As the number of positive roots of $p_n(\lambda)$ equals $\frac{n}{2}$ if n is even and $\frac{n-1}{2}$ if n is odd, the result holds. This ends the proof. □

Figure 4 shows the global bifurcation diagram provided by Theorem 3.2. It is an ideal global bifurcation diagram as the local bifurcation directions and the eventual secondary bifurcations have not been analyzed yet. According to [7, Th. 6.1], the local bifurcations of the $3T$-periodic component is transcritical, while the $4T$-periodic component bifurcates subcritically from the trivial curve $(\lambda, 1)$.

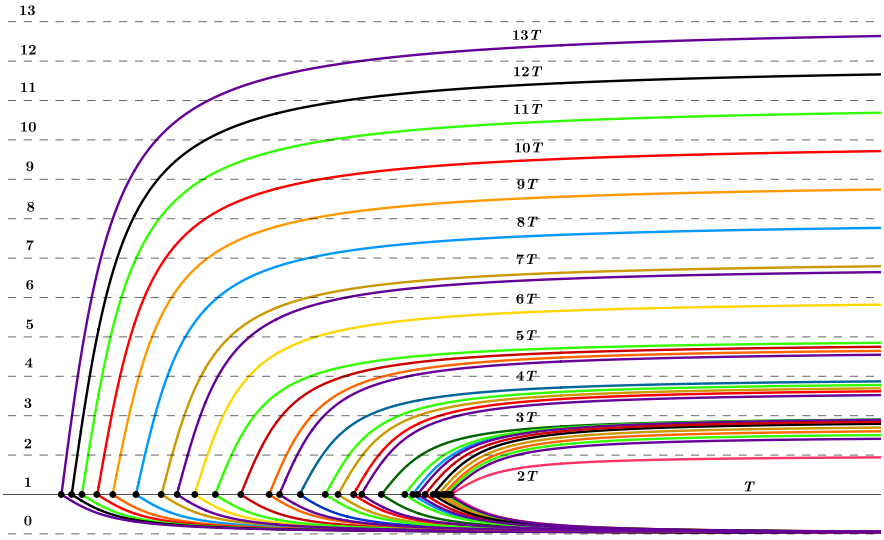


Fig. 4 Bifurcation diagram of (1.1) under condition (3.8).

Acknowledgements

This paper has been written under the joint auspices of the IMI of Complutense University and the Ministry of Science, Technology and Universities of Spain, under Research Grant PGC2018-097104-B-100 (first author), and the INDAM-GNAMPA under the research project SiDiA (Dynamical Systems and Applications) of the University of Udine (third author). The second author has been as well supported by the Contract CT42/18-CT43/18 of Complutense University of Madrid.

References

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

[2] Tongren Ding, Fabio Zanolin. Periodic solutions and subharmonic solutions for a class of planar systems of Lotka–Volterra type *Proc. First World Congress Nonlin. Anal.*: 395–406, 1996.

[3] Wei Yue Ding. Fixed points of twist mappings and periodic solutions of ordinary differential equations *Acta Math. Sinica*, 25: 227–235, 1982

[4] Jesús Esquinas and Julián López-Gómez, Optimal multiplicity in local bifurcation theory I: Generalized Generic Eigenvalues, *J. Diff. Eqns.*, 71, 72–92, 1988.

[5] Julián López-Gómez. A bridge between Operator Theory and Mathematical Biology *Fields Inst.*, 25: 383–397, 2000.

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

[7] Julián López-Gómez and Eduardo Muñoz-Hernández. Global structure of subharmonics in a class of periodic predator-prey models. *Nonlinearity*, 33: 34–71, 2020.

[8] Julián López-Gómez, Eduardo Muñoz-Hernández and Fabio Zanolin. On the applicability of the Poincaré–Birkhoff twist theorem to a class of planar periodic predator-prey models. *Discrete Contin. Dyn. Syst. Series A*, 40 (4): 2393–2419, 2020.

[9] Julián López-Gómez, Rafael Ortega and Antonio Tineo. The periodic predator-prey Lotka–Volterra model *Adv. Diff. Eqns.*, 1: 403–423, 1996.

[10] Alessandro Margheri, Carlota Rebelo and Fabio Zanolin. Maslov index, Poincaré–Birkhoff theorem and periodic solutions of asymptotically linear planar Hamiltonian systems *J. Differential Equations*, 183: 342–367, 2002.

[11] Paul H. Rabinowitz, Some global results for nonlinear eigenvalue problems, *J. Funct. Anal.*, 7, 487–513, 1971.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 265–270)

On a non-linear system of PDEs with application to tumor identification

Faustino Maestre [1], **Pablo Pedregal**[2]
1. *Dpto. Ecuaciones Diferenciales y Análisis Numérico and IMUS, Universidad de Sevilla, Spain*
2. *INEI, Universidad de Castilla La Mancha, Campus de Ciudad Real, Ciudad Real, Spain*

Abstract

In this work we present the analysis of a non-linear system of PDEs in relation with a solution of an inverse problem in conductivity with application in tumor detection. This non-linear system of PDEs corresponds with the Euler-Lagrange optimality systems associated to a non-convex vector variational problem. We study the (quasi-)convexification of the vector variational problem and present an strategy to solve the inverse problem numerically.

1. Introduction

We consider a bounded, regular domain $\Omega \subset \mathbb{R}^2$. We are interested in the following non-linear system of PDEs

$$\operatorname{div}\left(\frac{|\nabla u_2(\mathbf{x})|}{|\nabla u_1(\mathbf{x})|}\nabla u_1(\mathbf{x})\right) = 0 \text{ in } \Omega, \quad u_1 = u_{1,0} \text{ on } \partial\Omega, \tag{1.1}$$

$$\operatorname{div}\left(\frac{|\nabla u_1(\mathbf{x})|}{|\nabla u_2(\mathbf{x})|}\nabla u_2(\mathbf{x})\right) = 0 \text{ in } \Omega, \quad u_2 = u_{2,0} \text{ on } \partial\Omega, \tag{1.2}$$

for a given boundary data $(u_{1,0}, u_{2,0}) \in H^1(\Omega;\mathbb{R}^2)$. The above system can be interpreted, at least formally, as a Euler-Lagrange system associated with the functional

$$I(\mathbf{u}) = \int_\Omega |\nabla u_1(\mathbf{x})|\,|\nabla u_2(\mathbf{x})|\,d\mathbf{x}, \quad \mathbf{u} = (u_1, u_2). \tag{1.3}$$

Then, the optimal solutions of the optimization problem

$$\text{Minimize in } \mathbf{u} \in \mathcal{A}: \quad I(\mathbf{u}) \tag{1.4}$$

for a appropriated class of functions $\mathcal{A}$ with satisfying Dirichlet boundary condition,

$$\mathbf{u} = \mathbf{u}_0 \text{ on } \partial\Omega, \quad \mathbf{u}_0 = (u_{1,0}, u_{2,0}),$$

must to be solutions of the non-linear system (1.1)-(1.2).

In both cases, the system of PDEs and the optimization problems, correspond with vector problems. There are few references in bibliography of problems like these. Our aim is to prove the the existence of solutions of the non-linear system (1.1)-(1.2) proving the existence of global minimizers for the problem (1.4).

The minimization problem (1.4) has not good properties in order to apply the direct method of the Calculus of variations. The cost functional (1.3) is neither coercive, nor quasiconvex (see [7]).

We would like to remark that the solutions of previous minimization problem (1.4) (and in particular the solutions of the non-linear system (1.1)-(1.2)) are linked with an inverse problem in conductivity at the plane. In this sense, for a given boundary data

$$f \in H^{1/2}(\partial\Omega), \quad g \in H^{-1/2}(\partial\Omega)$$

the inverse problem consists in determining the function $\gamma : \mathbb{R} \to \mathbb{R}$ such that the unique solution of

$$\begin{cases} -\operatorname{div}(\gamma\nabla u) = 0 & \text{in } \Omega, \\ u = f & \text{on } \partial\Omega, \end{cases} \tag{1.5}$$

holds the additional Neumann boundary condition

$$\gamma\frac{\partial u}{\partial \nu} = g \quad \text{on } \partial\Omega, \tag{1.6}$$

 ISBN 978-84-18482-21-2

where g, represents the normal component of the outgoing electric current density on the surface, which is prescribed.

If we consider the state equation in (1.5), assuming Ω simply connected, then there exists a function $v \in H^1(\Omega)$ such that the equation in (1.5) is equivalent to the following pointwise equality

$$\gamma(x)\nabla u(x) = R\nabla v(x), \quad \text{a. e. } x \in \Omega, \tag{1.7}$$

where R is the counterclockwise $\pi/2$-rotation in the plane. From the vector equation (1.7) we can deduce three important aspects:

- a conductivity equation for v
$$\operatorname{div}\left[\frac{1}{\gamma(\mathbf{x})}\nabla v(\mathbf{x})\right] = 0 \text{ in } \Omega; \tag{1.8}$$
- a formula for γ in terms of u and v, namely
$$\gamma = \frac{|\nabla v|}{|\nabla u|}; \tag{1.9}$$
- Dirichlet boundary values around $\partial\Omega$ based on the Neumann condition for u
$$\nabla v \cdot \mathbf{t} = \gamma \nabla u \cdot \mathbf{n} \text{ on } \partial\Omega, \tag{1.10}$$
where $\mathbf{n}$ is the outer normal to Ω, and $\mathbf{t} = \mathbf{Rn}$ is the counterclockwise tangential vector to $\partial\Omega$.

Then, having in mind that γ is defined in (1.9), the state equation in (1.5) and (1.8) correspond with equations to the non-linear system (1.1)-(1.2), and in particular γ is solution to the inverse problem associated to the additional Neumann boundary condition given by (1.10).

This kind of inverse problem can be considered in the framework to the Calderon's problem, consisting of determining the conductivity inside of a medium by electrical measurements on its surface. In order to consider a more realist problem we assume that we know the result for M measurements, i.e., we seek a function $\gamma : \mathbb{R} \to \mathbb{R}$ such that the unique solution of

$$\begin{cases} -\operatorname{div}(\gamma\nabla u_m) = 0 & \text{in } \Omega, \\ u_m = u_{0,m} & \text{on } \partial\Omega, \end{cases} \qquad m = 1, \dots, M \tag{1.11}$$

holds the additional Neumann boundary condition

$$\gamma\frac{\partial u_m}{\partial \nu} = v_{0,m} \quad \text{on } \partial\Omega, \qquad m = 1, \dots, M \tag{1.12}$$

with $v_{0,m}$, $m = 1, \dots, M$.

In this case of multi-measurement, we consider the non-linear system of PDEs

$$\mathbf{u}(\mathbf{x}) : \Omega \subset \mathbb{R}^2 \to \mathbb{R}^{2M}$$

becomes

$$\operatorname{div}\left(\frac{|\nabla \mathbf{u}_2|}{|\nabla \mathbf{u}_1|}\nabla u_1^{(j)}\right) = 0 \text{ in } \Omega, \quad u_1^{(j)} = u_{1,0}^{(j)} \text{ on } \partial\Omega, \tag{1.13}$$

$$\operatorname{div}\left(\frac{|\nabla \mathbf{u}_1|}{|\nabla \mathbf{u}_2|}\nabla u_2^{(j)}\right) = 0 \text{ in } \Omega, \quad u_2^{(j)} = u_{2,0}^{(j)} \text{ on } \partial\Omega, \tag{1.14}$$

for $j = 1, 2, \dots, M$, where we are using the notation

$$\mathbf{u} = (\mathbf{u}^{(j)})_{j=1,2,\dots,N} = (u_1^{(j)}, u_2^{(j)})_{j=1,2,\dots,N},$$
$$\mathbf{u}^{(j)} = (u_1^{(j)}, u_2^{(j)}), \quad \mathbf{u}_i = (u_i^{(j)})_{j=1,2,\dots,N}, \quad i = 1, 2.$$

Note how this system is fully coupled because this time the quotient for the conductivity coefficient

$$\gamma = \frac{|\nabla \mathbf{u}_2|}{|\nabla \mathbf{u}_1|}$$

involves all of the components of $\mathbf{u}$.

Where now, the new cost functional for the variational problem is the following

$$I(\mathbf{u}) = \int_\Omega |\nabla \mathbf{u}_1(\mathbf{x})|\,|\nabla \mathbf{u}_2(\mathbf{x})|\,d\mathbf{x}. \tag{1.15}$$

In medical imaging, Calderón's problem is known as Electrical Impedance Tomography. This is a non-invasive technic in order to identify internal anomalies. It is known that, health or ill tissues have different electrical properties, in this way the determination of the internal electrical conductivity of the medium from boundary measurements have important applications in the detection of blood clots or tumor identifications (see [8] for a review of medical applications).

The literature concerning inverse problems of these kind of problems is large (see for instance [1,6]). The main questions concerning these inverse problems are *uniqueness* ([4, 11]) , *stability* ([2,5]) and *reconstruction* ([3]). In this work we show a technic of reconstruction of the internal electrical conductivity of the medium from the solutions of a vector variational problem.

2. On the vector variational problem

We present some important results for the vector variational problem (1.4), the case of one-single measurement. In this case the density of the cost functional is

$$\phi(\mathbf{F}) = |\mathbf{F}^{(1)}|\,|\mathbf{F}^{(2)}|, \quad \mathbf{F} = \begin{pmatrix}\mathbf{F}^{(1)}\\ \mathbf{F}^{(2)}\end{pmatrix} \in \mathbf{M}^{2\times 2}. \tag{2.1}$$

Having in mind (1.7) and (1.9), the state equation can written in matrix notation as

$$\frac{|\mathbf{F}_2|}{|\mathbf{F}_1|}\mathbf{F}_1 + \mathbf{R}\mathbf{F}_2 = \mathbf{0}, \quad \mathbf{F} = \begin{pmatrix}\mathbf{F}_1\\ \mathbf{F}_2\end{pmatrix}. \tag{2.2}$$

We put

$$\psi(\mathbf{F}) = \phi(\mathbf{F}) - \det \mathbf{F}, \tag{2.3}$$

then $\psi(\mathbf{F}) \geq 0$ always, but $\psi(\mathbf{F}) = 0$ precisely when (2.2) holds. In this way, we consider the modified functional

$$I(\mathbf{u}) = \int_\Omega \psi(\nabla \mathbf{u}(\mathbf{x}))\,d\mathbf{x}. \tag{2.4}$$

There are two main advantages of this functional over the old one.

1. Since we have added a null-lagrangian, $-\det \mathbf{F}$, to the old ϕ, the new underlying Euler-Lagrange system remains the same, i.e. our original system of PDEs (1.1)-(1.2).

2. If m is the infimum of I in (2.4), over a class of mappings respecting boundary data, then $m \geq 0$, and $m = 0$ is attained, i.e. $m = 0$ is a minimum, precisely when (2.2) holds for a minimizer (u_1, u_2).

We present the following result where the explicit quasiconvexification for (1.4) is presented.

Theorem 2.1 *The quasiconvexification $Q\phi$ of ϕ in* (2.1) *is given by the jacobian*

$$Q\phi(\mathbf{F}) = |\det \mathbf{F}|.$$

From the above Theorem and by means that $-\det \mathbf{F}$ is a null-lagrangian, it is immediate to argue that

$$Q\psi(\mathbf{F}) = |\det \mathbf{F}| - \det \mathbf{F} = 2\det{}^{-}\mathbf{F}. \tag{2.5}$$

In particular, we can deduce the following result to identify a sufficient condition for the Dirichlet boundary condition, in order to get that the infimum value vanishes.

Theorem 2.2 *Let $u_{i,0} \in H^{1/2}(\partial\Omega)$, $i = 1, 2$. If there is an extension of $(u_{1,0}, u_{2,0})$ to some $\mathbf{u}_0 \in H^1(\Omega;\mathbb{R}^2)$ with $\det \nabla \mathbf{u}_0 > 0$ a.e. in Ω then $m = 0$.*

In order to see the prove of the above results, see [9].

3. An inverse problem: synthetic data

We are interested in finding solutions for the non-linear system of PDEs (1.13)-(1.14), or global minimizer for minimization problem with cost functional (1.15), in order to get solutions to the inverse problem (1.11) - (1.12). We present a way to generate good boundary conditions on $\partial\Omega$, in the sense that there extremal problem has solution, and the minimum vanishes. The strategy is as follows

- Take any density $\gamma \in L^\infty(\Omega)$, and a function $u_{0,1} \in H^{1/2}(\partial\Omega)$.
- Let $\mathbf{R}_\delta : \mathbb{R}^2 \to \mathbb{R}^2$ be the counterclockwise rotation in the plane of angle δ. Take:

$$u_{0,1}^{(j)}(x,y) = u_{0,1}(\mathbf{R}_{\delta_j}(x,y)) \quad \delta_j = 2\pi\frac{j-1}{M} \in [0,2\pi) \quad j = 1,\dots,M.$$

  where δ_j represents different angles of rotation.

- Solve the problem

$$\operatorname{div}(\gamma\nabla u_1^{(j)}) = 0 \text{ in } \Omega, \quad u_1^{(j)} = u_{0,1}^{(j)} \text{ on } \partial\Omega, \quad j = 1,\cdots,M.$$

We consider

$$u_{0,1}^{(j)} = u_1^{(j)}, j = 1,\cdots,M,$$

with $\mathbf{u}_{0,1} \in H^{1/2}(\partial\Omega)^M$. To determine $\mathbf{u}_{0,2} \in H^{1/2}(\partial\Omega)^M$, we solve problems

$$\begin{cases} -\operatorname{div}\left(\frac{1}{\gamma}\nabla u_2^{(j)}\right) = 0 & \text{in } \Omega, \\ \frac{1}{\gamma}\frac{\partial u_2^{(j)}}{\partial\nu} = \nabla u_1^{(j)} \cdot \mathbf{t} & \text{on } \partial\Omega, \end{cases} \qquad j = 1,\dots,M, \tag{3.1}$$

under the normalization condition $\displaystyle\int_\Omega u(x)\,dx = 0$, and take

$$\mathbf{u}_{0,2} = \mathbf{u}_2|_{\partial\Omega} \in H^{1/2}(\partial\Omega)^M$$

where $\mathbf{t}$ is the counterclockwise tangential vector to $\partial\Omega$.

In this way, having in mind Theorem 2.2, we ensure that γ solves inverse problem (1.11) - (1.12) associated with the boundary data

$$\mathbf{u}_0 = (\mathbf{u}_{0,1}, \mathbf{u}_{0,2}) \in H^{1/2}(\partial\Omega;\mathbb{R}^2)^M, \tag{3.2}$$

4. Numerical experiments

We would like to present some numerical evidences to the resolution of inverse problems. In this way we have considered different strategies, the Newton-Raphson scheme for the non-linear system, we have also examined a fixed point algorithm, and a gradient descent algorithms (conjugated gradient or optimal step) in order to approximate optimal solutions for the variational problem. We decided to use the Newton-Raphson scheme becasue it is quickier. The algorithm is the following

1. We choose an admissible initialization $\mathbf{u}^0 \in H^1(\Omega;\mathbb{R}^2)^M$.
2. Iterate until convergence $\left(I(\mathbf{u}^k) < tol \text{ or } \frac{||\mathbf{w}^k||_\infty}{||\mathbf{u}^k||_\infty} < tol\right)$:
   - take $\mathbf{w}^k \in H_0^1(\Omega;\mathbb{R}^2)^M$ such that

$$D\mathcal{L}(\mathbf{u}^k,\mathbf{v})\,\mathbf{w}^k = \mathcal{L}(\mathbf{u}^k,\mathbf{v}),$$

     for every $\mathbf{v} \in H_0^1(\Omega;\mathbb{R}^2)^M$, where $\mathcal{L}(\mathbf{u}^k,\mathbf{v})$ and $D\mathcal{L}(\mathbf{u},\mathbf{v})\,\mathbf{w}$ corresponds with variational formulation of the system of PDEs (1.13)-(1.14) and its derivative, respectively.
   - update

$$\mathbf{u}^{k+1} = \mathbf{u}^k - \mathbf{w}^k.$$

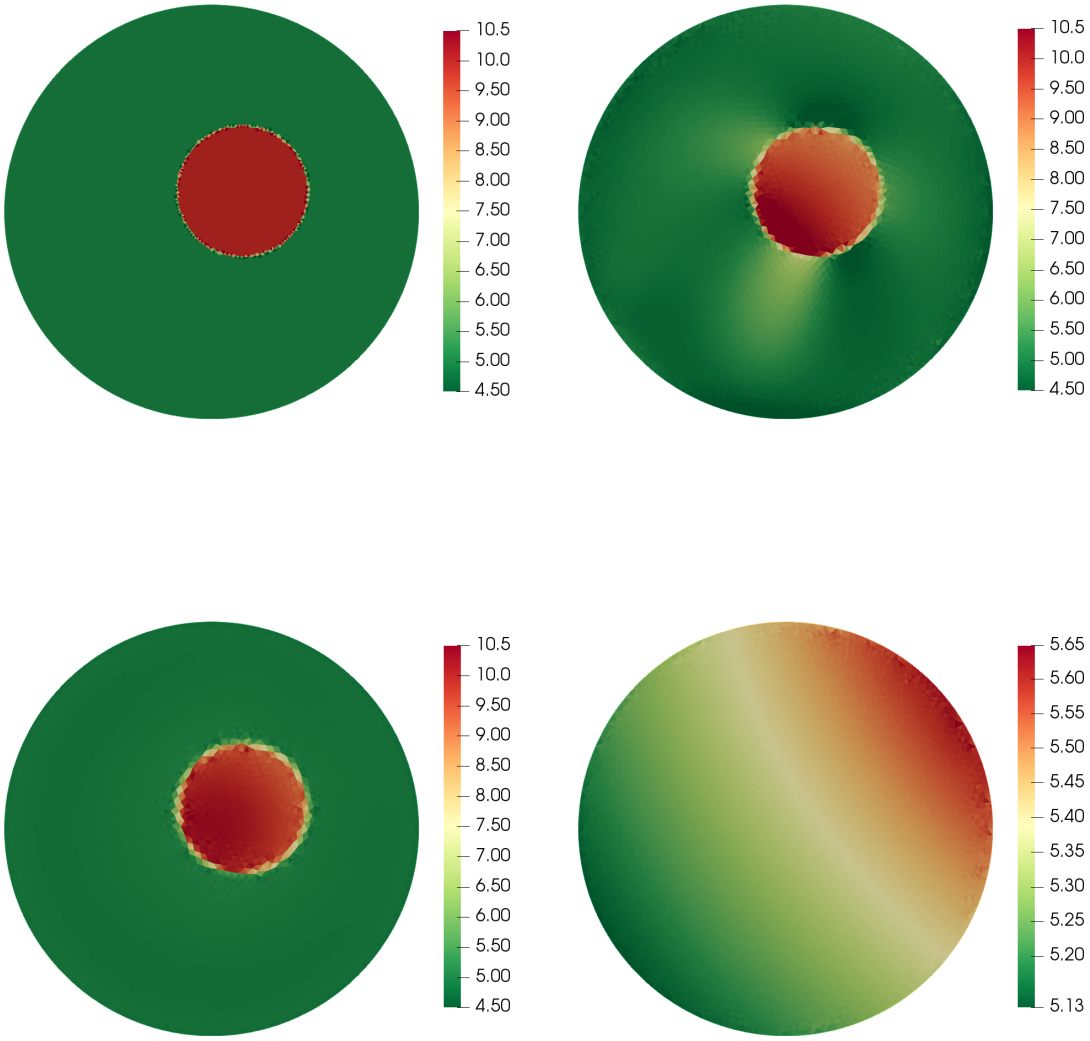


Fig. 1 Example 1 – The target γ (top left) and the computed γ for different numbers of measurements: N=1 (top right), N=3 (bottom left) using initialization 1 and , M=3 using initialization 2 (bottom right).

In our simulations the domain of reference is

$$\Omega = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1\},$$

and, we consider

$$\gamma = \beta\chi_D + \alpha(1 - \chi_D), \quad \text{with} \quad D = \{(x, y) \in \mathbb{R}^2 : (x - 0.15)^2 + (y - 0.1)^2 \leq 0.1\}. \tag{4.1}$$

with $\alpha = 5$ and $\beta = 10$. In order to generate synthetic data we choose

$$u_{0,1}(x, y) = 10x + 5 \sin y, \tag{4.2}$$

We have implemented the numerical simulation using the free software FreeFemm++ v 3.56.

The non-linear character of the problem, and the local character for convergence of Newton-Raphson algorithm suggest an importan aspect to choose the initialization. We consider two different ways to built it. In all our experiments we have considered the value of tolerance for convergence tol=10^{-6}.

In Figure 1 we can observer at top right the value of γ used to build data, this is the numerical target to reconstruction. The pictures at to right and bottom left represent the computed γ in the case of one single

experiment or 3 experiments for an kind of initialization. The picture at bottom right corresponds with the computed γ for the case of three experiments and using another initialization. We have a numerical evidence of the lack of uniqueness of solution of the inverse problem for a finite number of measurements.

In order to see more numerical experiments, and a more extensive analysis for the numerical problem you can see [9].

Acknowledgements

The first author was partially supported by Grant MTM2017-83583-P of the Spanish Government.

References

[1] Alessandrini, G. Stable determination of conductivity by boundary measurements. *Appl. Anal.* 27 (1988), 153-172.

[2] Alessandrini, G., de Hoop, M. V., Gaburro, R., Sincich, E., Lipschitz stability for the electrostatic inverse boundary value problem with piecewise linear conductivities. *J. Math. Pures Appl.* (9) 107 (2017), no. 5, 638-664.

[3] Ammari, H., Kang, H. Reconstruction of small inhomogeneities from boundary measurements. Lecture Notes in Mathematics. Springer-Verlag, Berlin, 2004.

[4] Astala, K., Päivärinta, L. Calderón's inverse conductivity problem in the plane. *Ann. of Math.* (2) 163 (2006), no. 1, 265-299.

[5] Barceló, T., Faraco, D., Ruiz, A., Stability of Caldeórn inverse conductivity problem in the plane. *J. Math. Pures Appl.* (9) 88 (2007), no. 6, 522-556.

[6] Borcea, L. Electrical impedance tomography. *Inverse Problems* 18 (2002), no. 6, R99-R136.

[7] Dacorogna, B., Direct method in the calculus of variations, Springer, 1989.

[8] Dijkstra A., Brown B., Harris N., Barber D., and Endbrooke D., Review: clinical applications of electrical impedance tomography, *J. Med. Eng. Technol.*17 (1993), 89-98.

[9] Maestre, F., Pedregal, P. Some non-linear systems of PDEs related to inverse problems in conductivity *Calc. Var. Partial Differential Equations* 60 (2021), 71

[10] Nachman A., Reconstructions from boundary measurements, *Ann. of Math.* 128 (1988), 531-576.

[11] Nachman A., Global uniqueness for a two-dimensional inverse boundary value problem. *Ann. of Math.* 143 (1996), 71-96.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 271–276)

Fractional evolution equations in dicrete sequences spaces

Pedro J. Miana
E-mail: pjmiana@Unizar.es. Universidad de La Rioja & Universidad de Zaragoza , Spain

Abstract

In this talk, we consider fractional differential equations (in the sense of Caputo) on the sequence Lebesgue spaces $\ell^p(\mathbb{Z})$ with $p \geq 1$. The associated operator to the Cauchy problem is defined by convolution with a sequence of compact support. We use techniques from Functional Analysis to calculate the solution of the problem. In the case of fractional powers of operators, we give explicitly the solution of the problem. As a consequence, we obtain new integral formulae for certain special functions.

1. Introduction

Of concern in this paper is the following semi discrete Cauchy problem

$$\begin{cases} \partial_t u(n,t) = Bu(n,t) + g(n,t), & n \in \mathbb{Z},\ t > 0, \\ u(n,0) = \varphi(n), & n \in \mathbb{Z}, \end{cases} \tag{1.1}$$

where B has the form of a convolution operator in the discrete variable, i.e.

$$Bu(n,t) = \sum_{j \in \mathbb{Z}} b(n-j)u(j,t), \tag{1.2}$$

and b belong to the Banach algebra $\ell^1(\mathbb{Z})$. A typical example is one dimensional discrete Laplacian, Δ_d, which can be obtained taking $b = \delta_{-1} - 2\delta_0 + \delta_1$, where $\delta_i(j)$ denotes the Kronecker delta (or discrete Dirac measure). In such case, equation (1.1) corresponds to the non-homogeneous semi discrete diffusion equation (also known as the semi discrete heat equation or the lattice diffusion equation). The analytical study of such kind of equations have received an increasing interest in the last decade, mainly due to their many applications in diverse areas of knowledge. For instance, in probability theory, the value $u(n,t)$ of (1.1) with $B = \Delta_d$, describes the probability that a continuous-time symmetric random walk on $\mathbb{Z}$ visits a point n at time t; cf. [6, Section 4]. In chemistry, (1.1) describes the flow of a chemical in an infinite system of tanks arranged in a row, where each two neighbors are connected by pipes [10, Section 3] and in transport theory, (1.1) describes the dynamics of an infinite chain of cars, each of them being coupled to its two neighbours. The value $u(n;t)$ is the displacement of car n at time t from its equilibrium position; cf. [5, Example 1]. From an analytical point of view, quite recently Slavik [11] studied the asymptotic behavior of solutions of (1.1) when $B = \Delta_d$, showing that a bounded solution approaches the average of the initial values if the average exists. Note that choosing $b = \delta_{-1} - \delta_0$ in (1.2) we obtain the forward difference operator $B = \Delta$ and hence (1.2) corresponds to the semi discrete transport equation, studied recently by Abadias et.al. [1].

Interestingly, the references [4] and [9] studied fundamental solutions of (1.1) and the second order semi discrete equation

$$\begin{cases} \partial_{tt} u(n,t) = Bu(n,t) + g(n,t), & n \in \mathbb{Z},\ t > 0, \\ u(n,0) = \varphi(n),\ \ u_t(n,0) = \phi(n), & n \in \mathbb{Z}, \end{cases} \tag{1.3}$$

when $B = -(-\Delta_d)^\alpha$ is the discrete fractional Laplacian. Particularly, in [9], the authors combine operator theory techniques with the properties of the Bessel functions to develop a theory of analytic semigroups and cosine operators generated by Δ_d and $-(-\Delta_d)^\alpha$. Also note that the fractional forward difference operator $B = -(-\Delta)^\alpha$ has been studied in [1] where maximum and comparison principles in the context of harmonic analysis are proved.

However, to the best of our knowledge, to date there is no attempt to investigate in an unified way fundamental solutions of the general equation (1.1).

Our key observation in this paper concerning this issue is that the discrete fractional Laplacian can be obtained from (1.2) by allowing fractional powers of b as element of the Banach algebra $\ell^1(\mathbb{Z})$. This original approach, that we provide in this paper, allow us to obtain new insights by introducing a completely new method to analyze both qualitative behavior and fundamental solutions of (1.1) in an unified way.

More generally, and in order to provide simultaneously in our analysis the sub diffusive and super difussive cases associated to the equations (1.1) and (1.3), in this paper we include the representation of the fundamental solutions for the following semi discrete equations:

$$\begin{cases} \mathbb{D}_t^{\beta} u(n,t) = Bu(n,t) + g(n,t), & n \in \mathbb{Z},\ t > 0, \\ u(n,0) = \varphi(n), & n \in \mathbb{Z}, \end{cases} \tag{1.4}$$

in case $0 < \beta \leq 1$ and

$$\begin{cases} \mathbb{D}_t^{\beta} u(n,t) = Bu(n,t) + g(n,t), & n \in \mathbb{Z},\ t > 0, \\ u(n,0) = \varphi(n),\ \ u_t(n,0) = \phi(n), & n \in \mathbb{Z}, \end{cases} \tag{1.5}$$

in case $1 < \beta \leq 2$. In both cases, B is the convolution operator $Bf(n) := (b * f)(n)$ defined on $\ell^p(\mathbb{Z})$, $p \in [1, \infty]$, $b \in \ell^1(\mathbb{Z})$ and $\beta \in (0, 2]$ is a real number. The symbol $\mathbb{D}_t^{\beta}$ denotes the Caputo fractional derivative of order $\beta > 0$. These results have been jointly obtained with Jorge González-Camus and Carlos Lizama from the Universidad de Santiago de Chile to appear in Advances in Difference Equations (2021).

2. A Banach algebra framework

Given $1 \leq p \leq \infty$, we recall that the Banach spaces $(\ell^p(\mathbb{Z}), \|\ \ \|_p)$ are formed by bi-infinite sequences $f = (f(n))_{n\in\mathbb{Z}} \subset \mathbb{C}$ such that

$$\begin{aligned} \|f\|_p : &= \left(\sum_{n=-\infty}^{\infty} |f(n)|^p \right)^{\frac{1}{p}} < \infty, \qquad 1 \leq p < \infty; \\ \|f\|_\infty : &= \sup_{n\in\mathbb{Z}} |f(n)| < \infty. \end{aligned}$$

We remind the natural embeddings $\ell^1(\mathbb{Z}) \hookrightarrow \ell^p(\mathbb{Z}) \hookrightarrow \ell^\infty(\mathbb{Z})$, for $1 \leq p \leq \infty$ and that the dual of $\ell^p(\mathbb{Z})$ is identified with $\ell^{p'}(\mathbb{Z})$ where $\frac{1}{p} + \frac{1}{p'} = 1$ for $1 < p < \infty$ and $p = 1$ if $p' = \infty$.

In the case that $f \in \ell^1(\mathbb{Z})$ and $g \in \ell^p(\mathbb{Z})$, we define

$$(f * g)(n) := \sum_{j=-\infty}^{\infty} f(n-j)g(j), \qquad n \in \mathbb{Z}.$$

From Young's Inequality, it follows that $f * g \in \ell^p(\mathbb{Z})$. Note that $(\ell^1(\mathbb{Z}), *)$ is a commutative Banach algebra with identity, that we denote $\delta_0 := \chi_{\{0\}}$. We observe that $\delta_1 * \delta_1 = \delta_2$ and, in general, $\delta_n * \delta_m = \delta_{n+m}$ for $n, m \in \mathbb{Z}$.

The Gelfand transform associated to $(\ell^1(\mathbb{Z}), *)$, is the discrete Fourier transform $\mathcal{F} : \ell^1(\mathbb{Z}) \to C(\mathbb{T})$ (or Fourier series) where

$$\hat{f}(\theta) := \mathcal{F}(f)(e^{i\theta}) := \sum_{n\in\mathbb{Z}} f(n)e^{in\theta}, \quad \theta \in \mathbb{T}.$$

We recall that the spectrum of f, denoted as $\sigma_{\ell^1(\mathbb{Z})}(f)$, is defined by

$$\sigma_{\ell^1(\mathbb{Z})}(f) := \{\lambda \in \mathbb{C} \ :\ (\lambda\delta_0 - f)^{-1} \in \ell^1(\mathbb{Z})\}.$$

In what follows, we consider the general theory of commutative Banach algebra as framework. We collect the results that will be of our interest in the following theorem.

Theorem 2.1 *The following properties hold*:

(*i*) *The spectrum Spec*$(\ell^1(\mathbb{Z}))$ *is compact and, consequently, homeomorphic to the unit complex circle,* $\mathbb{T} := \{z \in \mathbb{C} : |z| = 1\}$.

(*ii*) $\sigma_{\ell^1(\mathbb{Z})}(f) \subset \{z \in \mathbb{C}\ ;\ |z| < \|f\|_1\}$ *and*

$$(\lambda\delta_0 - f)^{-1} = \sum_{n\geq 0} \lambda^{-n-1} f^n, \qquad \|f\|_1 < |\lambda|. \tag{2.1}$$

(*iii*) *The algebra* $\ell^1(\mathbb{Z})$ *is a semi simple regular Banach algebra and the discrete Fourier transform* $\mathcal{F}$ *is injective.*

(*iv*) $\mathcal{F}(f*g)=\mathcal{F}(f)\mathcal{F}(g)$ *and*

$$\sigma_{\ell^1(\mathbb{Z})}(f)=\mathcal{F}(f)(\mathbb{T}),\qquad f\in\ell^1(\mathbb{Z}). \tag{2.2}$$

We observe that the range of the Gelfand transform is the Wiener algebra $\mathcal{A}(\mathbb{T})$, the pointwise algebra of absolutely convergent Fourier series, i.e., $F(e^{i\theta})=\sum_{n\in\mathbb{Z}} f(n)e^{i\theta n}$, $(\theta\in\mathbb{T})$ with $f\in\ell^1(\mathbb{Z})$. For $F\in\mathcal{A}(\mathbb{T})$, we also write $F(z)=\sum_{n\in\mathbb{Z}} f(n)z^n$, for $|z|\le 1$.

The inverse discrete Fourier transform is given by the following expressions

$$\mathcal{F}^{-1}(F)(n)=\frac{1}{2\pi}\int_{-\pi}^{\pi} F(e^{i\theta})e^{-in\theta}\,d\theta=\frac{1}{2\pi i}\int_{|z|=1} F(z)\frac{dz}{z^{n+1}},\quad n\in\mathbb{Z},$$

for $F\in\mathcal{A}(\mathbb{T})$ (and for other functions in larger sets).

The classical formulation of Wiener's lemma characterizes functions $F\in\mathcal{A}(\mathbb{T})$ which are invertible in $\mathcal{A}(\mathbb{T})$ as follows:

Given $F\in\mathcal{A}(\mathbb{T})$ where $F(e^{i\theta})=\sum_{n\in\mathbb{Z}} f(n)e^{i\theta n}$ for $\theta\in\mathbb{T}$. Then $F(e^{i\theta})\neq 0$ for all $\theta\in\mathbb{T}$ if and only if $1/F\in\mathcal{A}(\mathbb{T})$, i.e., $(1/F)(e^{i\theta})=\sum_{n\in\mathbb{Z}} g(n)e^{i\theta n}$ with $(g(n))_{n\in\mathbb{Z}}\in\ell^1(\mathbb{Z})$; in this case $f*g=\delta_0$. ([7, Theorem 5.5]).

Definition 2.2 Given $\alpha,\beta>0$, we define the vector-valued Mittag-Leffler function, $E_{\alpha,\beta}:\ell^1(\mathbb{Z})\to\ell^1(\mathbb{Z})$, by

$$E_{\alpha,\beta}(a):=\sum_{j=0}^{\infty}\frac{a^j}{\Gamma(\alpha j+\beta)},\qquad a\in\ell^1(\mathbb{Z}).$$

Note that

$$E_{1,1}(a)=\sum_{j=0}^{\infty}\frac{a^j}{j!}=e^a;\qquad E_{2,1}(a)=\sum_{j=0}^{\infty}\frac{a^j}{(2j)!}.$$

The set $\exp(\ell^1(\mathbb{Z})):=\{e^a\,;a\in\ell^1(\mathbb{Z})\}$ is the connected component of δ_0 in the set of regular elements in $\ell^1(\mathbb{Z})$ ([8, Theorem 6.4.1]).

We follow the usual terminology in semigroup theory: the element a is called the generator of the entire group $(e^{za})_{z\in\mathbb{C}}$; a cosine function, $\mathrm{Cos}(z,a):=E_{2,1}(z^2a)$, and a sine function, $\mathrm{Sin}(z,a):=zE_{2,2}(z^2a)$. We have

$$\mathrm{Sin}(z,a)=\int_{[0,z]}\mathrm{Cos}(s,a)ds,\qquad z\in\mathbb{C},$$

for $a\in\ell^1(\mathbb{Z})$, see [2, Sections 3.1 and 3.14]. Moreover, the Laplace transform of a entire group or a cosine function is connected with the resolvent of its generator as follows:

$$\begin{aligned}(\lambda-a)^{-1}&=\int_0^{\infty}e^{-\lambda s}e^{as}ds,\qquad \lambda>\|a\|_1,\\ \lambda(\lambda^2-a)^{-1}&=\int_0^{\infty}e^{-\lambda s}\mathrm{Cos}(s,a)ds,\qquad \lambda>\sqrt{\|a\|_1},\end{aligned} \tag{2.3}$$

see, for example, [2, p. 213].

Example 2.3 For $\alpha,\beta>0$, we have that

$$E_{\alpha,\beta}(z\delta_0)=E_{\alpha,\beta}(z)\delta_0;\qquad E_{\alpha,\beta}(z\delta_1)=\sum_{j=0}^{\infty}\frac{z^j\delta_j}{\Gamma(\alpha j+\beta)}.$$

In particular, $e^{z\delta_1}=\sum_{j=0}^{\infty}\frac{z^j\delta_j}{j!}$ and $\mathrm{Cos}(z,\delta_1)=\sum_{j=0}^{\infty}\frac{z^{2j}\delta_j}{(2j)!}$ are generated by δ_1.

In the next proposition, we collect some basic properties of these vector-valued Mittag-Leffler functions. As usual, we consider Bochner vector-valued integration in the Banach space $\ell^1(\mathbb{Z})$, see for example [12, Section 1.2].

Proposition 2.4 *For $\alpha, \beta > 0$ and $a \in \ell^1(\mathbb{Z})$, we have that*

(*i*) $\|E_{\alpha,\beta}(a)\|_1 \leq E_{\alpha,\beta}(\|a\|_1)$.

(*ii*) $\mathcal{F}(E_{\alpha,\beta}(a)) = E_{\alpha,\beta}(\mathcal{F}(a))$; *in particular* $\mathcal{F}(e^{az}) = e^{z\mathcal{F}(a)}$ *and* $\mathcal{F}(Cos(z,a)) = Cos(\mathcal{F}(z),a)$ *for* $z \in \mathbb{C}$.

(*iii*) $\sigma_{\ell^1(\mathbb{Z})}(E_{\alpha,\beta}(a)) = E_{\alpha,\beta}(\sigma_{\ell^1(\mathbb{Z})}(a))$.

(*iv*) *The following Laplace transform formula holds*

$$\int_0^\infty e^{-\lambda t} t^{\alpha k+\beta-1} E^{(k)}_{\alpha,\beta}(t^\alpha a)dt = k!\lambda^{\alpha-\beta}\left((\lambda^\alpha - a)^{-1}\right)^{(k+1)}, \quad \Re(\lambda) > \|a\|_1^{1/\alpha}, \tag{2.4}$$

for $k \in \mathbb{N} \cup \{0\}$.

(*v*) *For* $0 < \gamma < 1$, $E_{\gamma,1}(a) = \displaystyle\int_0^\infty \Phi_\gamma(t)e^{ta}dt$.

A nice application of the classical Wiener's lemma is the invariance of spectrum for convolution operators defined on $\ell^p(\mathbb{Z})$ for $1 \leq p \leq \infty$. This issue is contained in the following theorem that is the key abstract result in this paper.

Theorem 2.5 *Given $a \in \ell^1(\mathbb{Z})$, we define*

$$A(b)(n) := (a * b)(n), \quad n \in \mathbb{Z}, \quad b \in \ell^p(\mathbb{Z}), \tag{2.5}$$

then $A \in \mathcal{B}(\ell^p(\mathbb{Z}))$ for all $1 \leq p \leq \infty$. Moreover, $\|A\| = \|a\|_1$ and, for all $1 \leq p \leq \infty$, the following identities hold:

$$\sigma_{\mathcal{B}(\ell^p(\mathbb{Z}))}(A) = \sigma_{\ell^1(\mathbb{Z})}(a) = \mathcal{F}(a)(\mathbb{T}). \tag{2.6}$$

For all $a \in \ell^1(\mathbb{Z})$, we have that e^{za} is an entire group in $\ell^p(\mathbb{Z})$ with generator a and for all $1 \leq p \leq \infty$, the following identities hold:

$$\sigma_{\mathcal{B}(\ell^p(\mathbb{Z}))}(e^{za}) = \sigma_{\ell^1(\mathbb{Z})}(e^{za}) = e^{z\mathcal{F}(a)(\mathbb{T})}, \qquad z \in \mathbb{C}. \tag{2.7}$$

The element a in the above theorem is also called the *symbol* of the operator A.

3. Some finite difference operators in $\ell^1(\mathbb{Z})$

An important case of finite difference operators are given by sequences in the set

$$c_c(\mathbb{Z}) := \{a \in \ell^1(\mathbb{Z}) : \exists\, m \in \mathbb{Z}_+ : a(n) = 0, \forall\, |n| > m)\}.$$

In such case, the discrete Fourier Transform of $a \in c_c(\mathbb{Z})$ is a trigonometric polynomial

$$\mathcal{F}(a)(e^{i\theta}) = \sum_{j=-m}^{m} a(j)e^{ij\theta}. \tag{3.1}$$

It is interesting to observe that if $\sum_{j=-m}^m a(j) = 0$ then $0 \in \sigma_{\ell^1(\mathbb{Z})}(a)$. This follows immediately from (2.6).

Definition 3.1 For $f \in \ell^p(\mathbb{Z})$, with $1 \leq p \leq \infty$, we define the following operators

1. $-\Delta f(n) := f(n) - f(n+1) = ((\delta_0 - \delta_{-1}) * f)(n)$;
2. $\nabla f(n) := f(n) - f(n-1) = ((\delta_0 - \delta_1) * f)(n)$;
3. $\Delta_d f(n) := f(n+1) - 2f(n) + f(n-1) = ((\delta_{-1} - 2\delta_0 + \delta_1) * f)(n)$;
4. $\Delta_{dd} f(n) := f(n+2) - 2f(n) + f(n-2) = ((\delta_{-2} - 2\delta_0 + \delta_2) * f)(n)$;

for $n \in \mathbb{Z}$.

We remark that when considering the above defined operators in the context of numerical analysis, the operators $-\Delta$ and ∇ are related to Euler scheme of approximation, and the operator Δ_d corresponds to the second-order central difference approximation for the second order derivative. The operator Δ_{dd} appears in Bateman's paper [3, Page 506] in connection with the equations of Born and Karman on crystal lattices in vibration.

3.1. The operator $-\Delta$

The forward difference operator $\Delta f(n) := f(n+1) - f(n)$ is a classical operator used in approximation theory and in the theory of difference equations. Considered as an operator from $\ell^p(\mathbb{Z})$ to $\ell^p(\mathbb{Z})$, our main result read as follows.

Theorem 3.2 *The operator $-\Delta f = a * f$ where $a := \delta_0 - \delta_{-1}$ enjoys the following properties*

1. *The norm is given by $\|\Delta\| = 2$;*
2. *The Fourier transform is $\mathcal{F}(a)(z) = 1 - z,\ |z| = 1$;*
3. *For all $1 \le p \le \infty$ the spectrum is given by $\sigma_{\mathcal{B}(\ell^p(\mathbb{Z}))}(-\Delta) = \{z \in \mathbb{T} \ : |z-1| = 1\}$;*
4. *For $|\lambda + 1| > 1$,*
$$(\lambda\delta_0 + a)^{-1} = \sum_{j\ge 0} \frac{\delta_{-j}}{(1+\lambda)^{j+1}}.$$
5. *The associated group is $e^{-za}(n) = e^{-z}\dfrac{z^{-n}}{(-n)!}\chi_{-\mathbb{N}_0}(n)$, $z \in \mathbb{C}$, $n \in \mathbb{Z}$ and its generator is $-a$.*
6. *The norm of the group is given by $\|e^{-ta}\|_1 = 1, \quad t > 0$;*
7. *The associated cosine function is $Cos(z,-a)(n) = \frac{\sqrt{\pi}}{(-n)!}\left(\frac{z}{2}\right)^{-n+\frac{1}{2}} J_{-n-\frac{1}{2}}(z)\chi_{-\mathbb{N}_0}(n)$ where $z \in \mathbb{C},\ n \in \mathbb{Z}$.*

Similar results are proved for operator ∇ , Δ_d and Δ_{dd}.

4. Fundamental solution for semidiscrete evolution equations

Given $0 < \beta \le 1$, we first consider the equation

$$\begin{cases} \mathbb{D}_t^\beta u(n,t) = Bu(n,t) + g(n,t), & n \in \mathbb{Z},\ t > 0, \\ u(n,0) = \varphi(n), & n \in \mathbb{Z}. \end{cases} \tag{4.1}$$

We recall that function $E_{\alpha,\beta}(b)$ (with $b \in \ell^1(\mathbb{Z})$) is the vector-valued Mittag-Leffler function given in Definition 2.2. The main result is the following Theorem.

Theorem 4.1 *Let $\varphi, \phi \in \ell^p(\mathbb{Z})$, and $g : \mathbb{Z} \times \mathbb{R}_+ \to \mathbb{C}$ be such that, for each $t \in \mathbb{R}_+$, $g(\cdot,t) \in \ell^p(\mathbb{Z})$ and $\sup_{s\in[0,t]} ||g(\cdot,s)||_p < \infty$ with $1 \le p \le \infty$.*

(i) For $0 < \beta < 1$, the function

$$\begin{aligned} u(n,t) =& (E_{\beta,1}(t^\beta b) * \varphi)(n) \\ &+ \int_0^t (t-s)^{\beta-1}\left(E_{\beta,\beta}((t-s)^\beta b) * g(\cdot,s)\right)(n)ds,\ n \in \mathbb{Z}, \end{aligned}$$

is the unique solution of the initial value problem (4.1)*. Moreover, $u(\cdot,t)$ belong to $\ell^p(\mathbb{Z})$ for $t > 0$.*

(ii) For $1 < \beta < 2$, the function

$$\begin{aligned} u(n,t) =& (E_{\beta,1}(t^\beta b) * \varphi)(n) + t(E_{\beta,2}(t^\beta b) * \phi)(n) \\ &+ \int_0^t (t-s)^{\beta-1}\left(E_{\beta,\beta}((t-s)^\beta b) * g(\cdot,s)\right)(n)ds,\ n \in \mathbb{Z}, \end{aligned}$$

is the unique solution of the initial value problem (1.5)*. Moreover, $u(\cdot,t)$ belong to $\ell^p(\mathbb{Z})$ for $t > 0$.*

5. Applications to special functions

Take $a = \delta_{-1} - \delta_0$ or $a = \delta_1 - \delta_0$.

(i) For $0 < \beta < 1$, $t \in \mathbb{C}$ and $n \in \mathbb{N}_0$, we have

$$E^{(n)}_{\beta,1}(t) = \sum_{j=0}^{\infty} \frac{(j+n)!}{j!} \frac{t^j}{\Gamma(\beta(j+n)+1)} = \int_0^{\infty} \Phi_\beta(\tau) e^{\tau t} \tau^n d\tau.$$

(ii) For $1 < \beta < 2$, $t \in \mathbb{C}$ and $n \in \mathbb{N}_0$, we have

$$(2t)^{n-\frac{1}{2}} \sum_{j=0}^{\infty} \frac{(-1)^j (j+n)!}{j!} \frac{t^{2j}}{\Gamma(\beta(j+n)+1)} = \frac{\sqrt{\pi}}{2} \int_0^{\infty} \Phi_{\frac{\beta}{2}}(\tau) \tau^{n+\frac{1}{2}} J_{n-\frac{1}{2}}(\tau t)\, d\tau. \tag{5.1}$$

Now take $a = \delta_{-1} - 2\delta_0 + \delta_1$ or $a = \delta_{-2} - 2\delta_0 + \delta_2$.

(i) For $0 < \beta < 1$, $t \in \mathbb{C}$ and $n \in \mathbb{N}_0$, we have

$$\sum_{j=0}^{\infty} (-1)^j \binom{2(j+n)}{j} \frac{t^{j+n}}{\Gamma(\beta(j+n)+1)} = \int_0^{\infty} \Phi_\beta(\tau) e^{-2\tau t} I_n(2\tau t) d\tau. \tag{5.2}$$

In particular, when $\beta = \frac{1}{3}$, we get the integral formula for Airy function,

$$\sum_{j=0}^{\infty} (-1)^j \binom{2(j+n)}{j} \frac{t^{j+n}}{\Gamma(\frac{j+n}{3}+1)} = \int_0^{\infty} 3^{\frac{2}{3}} Ai\left(\frac{\tau}{3^{\frac{1}{3}}}\right) e^{-2\tau t} I_n(2\tau t) d\tau,$$

for $t \in \mathbb{C}$ and $n \in \mathbb{N}_0$.

(ii) For $1 < \beta < 2$, $t \in \mathbb{C}$ and $n \in \mathbb{N}_0$, we have

$$\sum_{j=0}^{\infty} (-1)^j \binom{2(j+n)}{j} \frac{t^{2(j+n)}}{\Gamma(\beta(j+n)+1)} = \int_0^{\infty} \Phi_{\frac{\beta}{2}}(\tau) J_{2n}(2\tau t) d\tau. \tag{5.3}$$

Acknowledgements

P.J. Miana has been partially supported by ProjectID2019-105979GBI00, DGI-FEDER, of the MCEI and Project E48-20R, Gobierno de Aragón, Spain

References

[1] L. Abadias, M. de León-Contreras, J.L. Torrea. *Non-local fractional derivatives. Discrete and continuous.* J. Math. Anal. Appl., **449**(1) (2017) 734–755.

[2] W. Arendt, C. Batty, M. Hieber and F. Neubrander. Vector-valued Laplace Transforms and Cauchy Problems. Monographs in Mathematics. **96**. Birkhäuser, Basel, 2001.

[3] H. Bateman. *Some simple differential difference equations and the related functions.* Bull. Amer. Math. Soc. 49 (1943), 494–512.

[4] O. Ciaurri, C. Lizama, L. Roncal and J.L. Varona. *On a connection between the discrete fractional Laplacian and superdiffusion.* Applied Math. Letters, 49 (2015), 119-125.

[5] A. Feintuch, B. Francis. *Infinite chains of kinematic points.* Automatica J. IFAC 48 (2012), 901–908.

[6] M. Friesl, A. Slavik, P. Stehlik. *Discrete-space partial dynamic equations on time scales and applications to stochastic processes.* Appl. Math. Lett. 37 (2014), 86–90.

[7] K. Gröchenig. *Wiener's Lemma: Theme and Variations. An Introduction to Spectral Invariance and Its Applications* in "Four Short Courses on Harmonic Analysis" (edit. B. Forster and P. Massopust), Birkhüusser, Boston, (2010), 175–234.

[8] R. Larsen. Banach Algebras: An Introduction. Marcel Dekker, New York, 1973.

[9] C. Lizama, L. Roncal. *Hölder-Lebesgue regularity and almost periodicity for semidiscrete equations with a fractional Laplacian.* Discr. Cont. Dyn. Syst., Series A, **38** (3) (2018), 1365–1403.

[10] A. Slavik. *Mixing problems with many tanks.* Amer. Math. Monthly **120** (2013), 806–821.

[11] A. Slavik. *Asymptotic behavior of solutions to the semidiscrete diffusion equation.* Appl. Math. Letters **106** (2020), 106392.

[12] A.M. Sinclair. Continuous Semigroups in Banach Algebras.London Mathematical Society, Lecture Note Series **63**, Cambridge University Press, Cambridge, 1982.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 277–283)

KPZ equation approximated by a nonlocal equation

Alexis Molino
Universidad de Almería, Spain

Abstract

Our main concern is the study of several aspects related with solutions of nonlocal problems whose prototype is

$$\begin{cases} u_t = \displaystyle\int_{\mathbb{R}^N} J(x-y)\big(u(y,t)-u(x,t)\big)\mathcal{G}\big(u(y,t)-u(x,t)\big)dy & \text{in } \Omega\times(0,T)\,, \\ u(x,0)=u_0(x) & \text{in } \Omega\,, \\ u(x,t)=h(x,t) & \text{in } \quad (\mathbb{R}^N\setminus\Omega)\times(0,T)\,. \end{cases}$$

where we take, as the most important instance, $\mathcal{G}(s) \sim 1+\frac{\mu}{2}\frac{s}{1+\mu^2 s^2}$ with $\mu\in\mathbb{R}$ as well as $u_0\in L^1(\Omega)$, J is a smooth symmetric function with compact support and Ω is a bounded smooth subset of $\mathbb{R}^N$, with nonlocal Dirichlet boundary condition $h(x,t)$.

The results deal with existence, uniqueness and comparison principle. The main motivation for dealing with these types of equations is that, under a kernel $\mathcal{G}$ rescaled in a suitable way, the unique solution of the above problem converges to a solution of the deterministic Kardar-Parisi-Zhang equation.

1. Introduction

We present some partial results from [15] concerning to the Dirichlet problem. Concretely, existence, uniqueness, comparison principle and rescaling kernel for the following nonlinear parabolic equation with nonlocal diffusion,

$$\begin{cases} u_t(x,t) = \displaystyle\int_{\mathbb{R}^N} J(x-y)\big(u(y,t)-u(x,t)\big)\mathcal{G}\big(u(y,t)-u(x,t)\big)dy & \text{in } \Omega\times(0,T)\,, \\ u(x,0)=u_0(x) & \text{in } \Omega\,, \\ u(x,t)=h(x,t) & \text{in } \quad (\mathbb{R}^N\setminus\Omega)\times(0,T)\,, \end{cases} \tag{1.1}$$

for an appropriate functions J and $\mathcal{G}$ (see below (J) and $(\mathcal{G})$), and its relationship with the deterministic KPZ equation

$$\begin{cases} u_t-\Delta u=\mu|\nabla u|^2 & \text{in } \Omega\times(0,T)\,, \\ u(x,0)=u_0(x) & \text{in } \Omega\,, \\ u(x,t)=h(x,t) & \text{on } \partial\Omega\times(0,T)\,, \end{cases} \tag{1.2}$$

where

1. Ω is a bounded smooth subset of $\mathbb{R}^N$ adding the boundary condition $u(x,t)=h(x,t)$ in $(\mathbb{R}^N\setminus\Omega)\times(0,T)$ for h sufficiently smooth;
2. $T>0$ (possibly infinite) and $\mu\in\mathbb{R}$;
3. u_0 is a smooth enough datum.

1.1. Local problem

The equation $u_t-\Delta u=\mu|\nabla u|^2$, at least for $\mu>0$, is known in the literature as the deterministic Kardar-Parisi-Zhang (KPZ) equation. The KPZ equation was proposed in [13] in the physical theory of growth and roughening of surfaces. Further developments on physical applications of the KPZ equation can be found in [5] (for a survey on more recent aspects we refer to [19]). The deterministic case corresponds to the smoothing from an initially rough surface to a flat one.

The Kardar–Parisi–Zhang equation has given rise to a rich mathematical theory which has had a spectacular recent progress (see [10,11]). From the point of view of Partial Differential Equations, equations having a gradient term with the so-called natural growth have been largely studied in the last decades by many mathematicians: in addition to the classical reference [14] let us just mention the pioneer paper by Aronson and Serrin [3] and also the result due to Boccardo, Murat and Puel [6].

1.2. Nonlocal problem

Nonlocal evolution equations have been extensively studied to model diffusion processes. The prototype example in this framework is the following one

$$u_t(x,t) = \int_{\mathbb{R}^N} K(x,y)(u(y,t) - u(x,t))dy, \tag{1.3}$$

where the kernel $K : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ is a nonnegative smooth function (not necessarily symmetric) satisfying $\int_{\mathbb{R}^N} K(x,y)dx = 1$ for any $y \in \mathbb{R}^N$ (or variations of it, see for instance [2]). If $u(y,t)$ is thought of as a density at location y at time t and $K(x,y)$ as the probability distribution of jumping from place y to place x, then the rate at which individuals from any other location go to the place x is given by $\int_{\mathbb{R}^N} K(x,y)u(y,t)dy$. On the other hand, the rate at which individuals leave the location x to travel to all other places is $-\int_{\mathbb{R}^N} K(y,x)u(x,t)dy = -u(x,t)$. In the absence of external sources this implies that the density must satisfy equation (1.3).

We are especially interested in symmetric kernels that have compact support; it means that the individuals can jump from a place to other, but they cannot go "too far away". On the contrary, for instance, nonlocal operators that allow "long jumps"correspond to a different choice of kernels. It is the case of the fractional Laplacian that involves a kernel that is singular and that does not have compact support (see, for instance [18] for a survey on this latter class of processes and [1] for the KPZ equation in fractional framework).

In particular, we consider $J : \mathbb{R}^N \to \mathbb{R}$ as a nonnegative radial symmetric function such that

$$J \in C_c(\mathbb{R}^n), \quad \text{with} \quad \int_{\mathbb{R}^N} J(z)\,dz = 1.$$

Choosing the kernel as $K(x,y) = J(x-y)$, equation (1.3) changes into a diffusion equation of convolution type, namely

$$u_t(x,t) = (J * u - u)(x,t) = \int_{\mathbb{R}^N} J(x-y)u(y,t)dy - u(x,t), \qquad \text{in } \Omega \times (0,T) \tag{1.4}$$

(see for instance [4, 7, 9]).

1.3. Background

One of the most important features of nonlocal equations is that they can be rescaled to approximate local ones.

In [8] (see also [16] and [17] for the same type of result in a more general case) it has been proved that, under an appropriate rescaling kernel, solutions of (1.4) converge uniformly to solutions of heat equation. To be more specific, solutions of

$$u_t^\varepsilon(x,t) = \frac{C}{\varepsilon^2}\left[\int_{\mathbb{R}^N} J_\varepsilon(x-y)u^\varepsilon(y,t)dy - u^\varepsilon(x,t)\right] \qquad \text{in } \Omega \times (0,T) \tag{1.5}$$

converge uniformly (when $\epsilon \to 0$) to solutions of

$$v_t = \Delta v \qquad \text{in } \Omega \times (0,T)\,,$$

where $C^{-1} = \frac{1}{2}\int_{\mathbb{R}^N} J(z)z_N^2\,dz$ and $J_\varepsilon(s) = \frac{1}{\varepsilon^N}J\left(\frac{s}{\varepsilon}\right)$.

Let us mention that results in this direction, with the presence of a gradient term of convection type can be found, for instance, in [12]: in such a case the equation is the sum of two terms, one corresponding to the diffusion one, the other to the convection term.

In general, we consider nonlocal problems of the type

$$u_t(x,t) = \int_{\mathbb{R}^N} J(x-y)\big(u(y,t) - u(x,t)\big)\,\mathcal{G}\big(u(y,t) - u(x,t)\big)\,dy, \tag{1.6}$$

where $\mathcal{G} : \mathbb{R} \to \mathbb{R}$ is a suitable continuous function. For instance, if $\mathcal{G} \equiv 1$, then we recover problem (1.4). Let us mention that the case $\mathcal{G}(s) = |s|^{p-2}$, with $p \geq 2$ has been treated in [2] where it is proved that solutions to the rescaled nonlocal problem converge to solutions of the Dirichlet problem for the p–Laplacian evolution equation.

On the contrary, the kind of kernels $\mathcal{G}$ we consider does not have the same structure of the previous ones, since they are bounded and do not satisfy any symmetry assumptions (neither odd nor even).

With this background, it is not surprising that problem (1.2) can be approximated by nonlocal equations. The question is to identify what kind of nonlocal equation approximates, under rescaling, problem (1.2).

1.4. Main results

To conclude this introduction we want to state the most relevant results of this work. In order to not enter in technicalities, let us fix a family of kernels $\mathcal{G}_\mu$ that are the easiest (not trivial) example we can consider: for $\mu \in \mathbb{R}$ let

$$\mathcal{G}_\mu(s) = 1 + \frac{\mu s}{2(1+\mu^2 s^2)}, \qquad s \in \mathbb{R}, \quad \mu \in \mathbb{R},$$

and the corresponding family of nonlocal Dirichlet problems

$$\begin{cases} u_t(x,t) = \displaystyle\int_{\mathbb{R}^N} J(x-y)\big(u(y,t)-u(x,t)\big)\,\mathcal{G}_\mu\big(u(y,t)-u(x,t)\big)\,dy & \text{in } \Omega\times(0,T), \\ u(x,0) = u_0(x) & \text{in } \Omega, \\ u(x,t) = h(x,t) & \text{in } (\mathbb{R}^N\setminus\Omega)\times(0,T). \end{cases} \tag{1.7}$$

with Ω a bounded domain and u_0 and h smooth enough (see Definition 2.1 and Definition 2.4 for more precise hypotheses).

After have proved the existence, uniqueness (see Theorem 2.3) and a Comparison Principle (see Theorem 2.5) for solutions of (1.7), we face the problem of *rescaled kernels*.

The result we prove, in this model case, reads like this.

Let u be the unique smooth solution to (1.2)*, with suitable initial data u_0 and smooth enough boundary condition $u(x,t) = h(x,t)$ on $\partial\Omega\times(0,T)$. Then there exists a family of functions $\{u^\varepsilon\}$, $\varepsilon > 0$, such that u^ε solves the approximating nonlocal problem*

$$\begin{cases} u_t^\varepsilon(x,t) = \dfrac{C}{\varepsilon^2}\displaystyle\int_{\Omega_{J_\varepsilon}} J_\varepsilon(x-y)\left[\big(u^\varepsilon(y,t)-u^\varepsilon(x,t)\big) + \frac{\mu}{2}\frac{\big(u^\varepsilon(y,t)-u^\varepsilon(x,t)\big)^2}{1+\mu^2\big(u^\varepsilon(y,t)-u^\varepsilon(x,t)\big)^2}\right]dy & \text{in } \Omega\times(0,T), \\ u^\varepsilon(x,0) = u_0(x) & \text{in } \Omega, \\ u^\varepsilon(x,t) = h(x,t) & \text{in } (\Omega_{J_\varepsilon}\setminus\Omega)\times(0,T), \end{cases}$$

with C a suitable constant, $\Omega_{J_\varepsilon} = \Omega + \operatorname{supp} J_\varepsilon$ and the family $\{u^\varepsilon\}$ satisfies

$$\lim_{\varepsilon\to 0}\ \sup_{t\in[0,T]} \left\| u^\varepsilon(x,t) - u(x,t)\right\|_{L^\infty(\Omega)} = 0.$$

2. Statement of the results

Let us consider the following equation:

$$u_t(x,t) = \int_{\mathbb{R}^N} J(x-y)u(y;x,t)\,\mathcal{G}(x,u(y;x,t))\,dy, \tag{2.1}$$

where $J : \mathbb{R}^N \to \mathbb{R}$ is a nonnegative radial symmetric function such that

$$J \in C_c(\mathbb{R}^n), \quad \text{with} \qquad \int_{\mathbb{R}^N} J(z)\,dz = 1, \tag{J}$$

and where, here and throughout, we denote by $u(y;x,t) := u(y,t) - u(x,t)$ and by

$$C(J) := \int_{\mathbb{R}^N} J(z) z_N^2\,dz < \infty, \quad z = (z_1, z_2, \dots, z_N).$$

As far as the function $\mathcal{G}$ is concerned, we assume that $\mathcal{G} : \mathbb{R}^N\times\mathbb{R} \to \mathbb{R}$ is a nonnegative Carathéodory function (namely, $\mathcal{G}(\cdot,s)$ is measurable for every $s\in\mathbb{R}$ and $\mathcal{G}(x,\cdot)$ is continuous for almost every $x\in\mathbb{R}^N$) satisfying

$$\exists\ \alpha_2 \ge \alpha_1 > 0 : \qquad \alpha_1 \le \frac{\mathcal{G}(x,s)s - \mathcal{G}(x,\sigma)\sigma}{s-\sigma} \le \alpha_2, \qquad \forall s,\sigma\in\mathbb{R}\ s\neq\sigma, \text{ and for a.e. } x\in\mathbb{R}^N. \tag{$\mathcal{G}$}$$

Let us first point out that the above condition implies that $\mathcal{G}$ is a positive bounded function, since taking $\sigma = 0$ in ($\mathcal{G}$), we get

$$0 < \alpha_1 \le \mathcal{G}(x,s) \le \alpha_2, \qquad \text{for any } s\in\mathbb{R} \text{ and for a.e. } x\in\mathbb{R}^N.$$

Moreover observe that the above condition relies to be a sort of uniform ellipticity for the operator, while $(\mathcal{G})$ corresponds to a strong monotonicity.

Anyway, let us stress again that, in contrast with all the known results about nonlocal equations of the above type, in our case we do not require any symmetry (neither odd nor even) assumption on $\mathcal{G}$.

The prototype of $\mathcal{G}$ we have in mind (we will come back on this example later) is the following one:

$$\mathcal{G}_\mu(x,s) = 1 + \frac{\mu(x)\, s}{2(1+\mu(x)^2 s^2)}, \quad x \in \Omega, \quad s \in \mathbb{R},$$

where $\mu : \Omega \to \mathbb{R}$ stands for a measurable function. Notice that this function satisfies $\mathcal{G}_\mu(x,0) = 0$ and $\frac{d}{ds}\mathcal{G}_\mu(x,0) = \mu(x)$.

The first kind of results we want to stablish deals with the existence and uniqueness of solutions. More precisely, consider the following problem in a bounded domain $\Omega \subset \mathbb{R}^N$, $N \geq 1$.

$$\begin{cases} u_t(x,t) = \displaystyle\int_{\mathbb{R}^N} J(x-y)u(y;x,t)\,\mathcal{G}(x,u(y;x,t))\,dy, & \text{in } \Omega \times (0,T) \\ u(x,t) = h(x,t), & \text{in } (\mathbb{R}^N \setminus \Omega) \times (0,T), \\ u(x,0) = u_0(x), & \text{in } \Omega, \end{cases}$$

with $h \in L^1\left((\mathbb{R}^N \setminus \Omega) \times (0,\infty)\right)$ and $u_0 \in L^1(\Omega)$.

Let us first observe that the integral expression vanishes outside of $\Omega_J = \Omega + \mathrm{supp}(J)$. In this way, h has only to be prescribed, in fact, in $\Omega_J \setminus \Omega$ and we can rewrite the above problem as

$$\begin{cases} u_t(x,t) = \displaystyle\int_{\Omega_J} J(x-y)u(y;x,t)\,\mathcal{G}(x,u(y;x,t))\,dy, & \text{in } \Omega \times (0,T), \\ u(x,t) = h(x,t), & \text{in } (\Omega_J \setminus \Omega) \times (0,T), \\ u(x,0) = u_0(x), & \text{in } \Omega, \end{cases} \tag{P}$$

where $T > 0$ may be finite or $+\infty$.

We give now two definitions of solution.

Definition 2.1 Assume that J and $\mathcal{G}$ satisfy (J) and $(\mathcal{G})$, respectively.
For $h(x,t) \in L^1((\Omega_J \setminus \Omega) \times (0,T))$ and $u_0(x) \in L^1(\Omega)$, we define a weak solution of problem (P) as a function $u \in C([0,T); L^1(\Omega))$ such that:

$$u(x,t) = \int_0^t \int_{\Omega_J} J(x-y)u(y;x,\tau)\mathcal{G}(x,u(y;x,\tau))\,dy\,d\tau + u_0(x), \qquad \text{for a.e. x } \in \Omega,\ t \in (0,T), \tag{2.2}$$
$$u(y,t) = h(y,t) \qquad \text{for a.e. } y \in \Omega_J \setminus \Omega \text{ and } t \in (0,T)$$
$$\lim_{t \to 0^+} \|u(x,t) - u_0(x)\|_{L^1(\Omega)} = 0\,.$$

Moreover, if $h(x,t) \in C((\Omega_J \setminus \overline{\Omega}) \times (0,T))$ and $u_0(x) \in C(\overline{\Omega})$, we define a regular solution of problem (P) as a function $u \in C([0,\infty); C(\overline{\Omega}))$ such that:

$$u(x,t) = \int_0^t \int_{\Omega_J} J(x-y)u(y;x,\tau)\mathcal{G}(x,u(y;x,\tau))\,dy\,d\tau + u_0(x), \qquad \text{for any x } \in \overline{\Omega},\ t \in (0,T),$$
$$u(y,t) = h(y,t) \qquad \text{for any } y \in \Omega_J \setminus \overline{\Omega} \text{ and } t \in (0,T)$$
$$\lim_{t \to 0^+} \|u(x,t) - u_0(x)\|_{C(\overline{\Omega})} = 0\,.$$

Some more remarks about the meaning of weak and regular solutions are now in order.

Remark 2.2

i) Observe that, in addition to the different smoothness of the boundary condition and/or the initial datum, the main difference lies on the prescription of data on $\partial\Omega$. Indeed, for weak solutions, h is prescribed in $(\Omega_J \setminus \Omega) \times (0,T)$ and u_0 in Ω, while for regular solutions, h is prescribed in $(\Omega_J \setminus \overline{\Omega}) \times (0,T)$ and u_0 in $\overline{\Omega}$.

ii) As already noticed in [7] (in a different context) the boundary conditions are not understood in a classical way, i.e. it is not true that the solutions of problem (P) pointwise coincide with the prescribed boundary data $h(x,t)$. This is due to the fact that the value at any point $(x,t) \in \partial\Omega \times (0,T)$ depends both on the values of u inside $\overline{\Omega} \times [0,T]$ and on the boundary datum $h(x,t)$, since

$$u(x,t) = \int_0^t \int_{\Omega \cap \mathrm{supp} J} J(x-y)u(y;x,\tau)\, \mathcal{G}\big(x, u(y,\tau) - u(x,\tau)\big)\, dy\, d\tau$$
$$+ \int_0^t \int_{\Omega^c \cap \mathrm{supp} J} J(x-y)\big(h(y,\tau) - u(x,\tau)\big)\, \mathcal{G}\big(x, h(y,\tau) - u(x,\tau)\big)\, dy\, d\tau + u_0(x)\,.$$

Consequently, in contrast with the local case, the equation is solved up to the boundary, depending, near $\partial\Omega$, also of the prescribed boundary condition.

iii) Let us stress that the regularity required in the definition of weak solutions is the less restrictive in order to give sense to the formulation, and to the boundary and initial conditions. Anyway from (2.2) we deduce that the time derivative $u_t(x,t)$ of u also belongs to $C((0,\infty); L^1(\Omega))$.

Let us also point out that the weak solution framework is the more natural one in order to prove the existence of a solution. Indeed we only require an L^1 regularity to prove the existence of a solution.

Finally we want to underline that the nonlocal operator involved in such equation does not have the regularizing effect that is typical of the Laplacian, but leaves unchanged the regularity of the initial and boundary data.

In this framework, the existence result is the following:

Theorem 2.3 [*Existence*] *Consider problem* (P) *and suppose that* (J) *and* $(\mathcal{G})$ *are in force. Then*:

i) For any $u_0 \in L^1(\Omega)$ *and* $h \in L^1((\Omega_J \setminus \Omega) \times (0,T))$ *there exists a unique weak solution*;

ii) For any $u_0 \in C(\overline{\Omega})$ *and* $h \in C((\Omega_J \setminus \overline{\Omega}) \times [0,T))$ *there exists a unique regular solution and moreover its time derivative belongs to* $C(\overline{\Omega} \times (0,T))$.

Once we have deduced the existence of a solution, one important tool is to compare two solutions, or, more generally a sub and a supersolution. Here we recall what we mean by those concepts in our setting.

Definition 2.4 A function $u \in C(\overline{\Omega} \times [0,T])$ is a regular subsolution to problem (P) if it satisfies $u_t \in C(\overline{\Omega} \times (0,T))$ and

$$\begin{cases} u_t(x,t) \le \displaystyle\int_{\Omega_J} J(x-y)u(y;x,t)\, \mathcal{G}(x, u(y;x,t))\, dy, & \text{in } \overline{\Omega} \times (0,T), \\ u(x,t) \le h(x,t), & \text{in } (\Omega_J \setminus \overline{\Omega}) \times (0,T), \\ u(x,0) \le u_0(x), & \text{in } \overline{\Omega}, \end{cases} \tag{2.3}$$

with $u_0(x) \in C(\overline{\Omega})$ and $h(x,t) \in C((\Omega_J \setminus \overline{\Omega}) \times (0,T))$.
As usual, a regular supersolution is defined analogously by replacing "$\le$" with "$\ge$". Clearly, a regular solution is both a regular subsolution and a regular supersolution.

Next, we state the comparison principle:

Theorem 2.5 [*Comparison Principle*] *Let* u *and* v *be a regular subsolution and a regular supersolution of problem* (P)*, respectively, with boundary data* $h_1(x,t)$ *and* $h_2(x,t)$ *and initial data* $u_0(x)$ *and* $v_0(x)$*, respectively. If* $h_1(x,t) \le h_2(x,t)$ *in* $\Omega_J \setminus \overline{\Omega}$ *and* $u_0(x) \le v_0(x)$ *in* $\overline{\Omega}$*, then* $u \le v$ *in* $\overline{\Omega} \times [0,T]$.

Remark 2.6 The existence, uniqueness and comparison principle are also true relaxing the hypotheses on the kernel $J(x-y)$ by considering a more general one of the form $K : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^+$ with compact support in $\Omega \times B(0,\rho)$, with $\rho > 0$ such that

$$0 < \sup_{y \in B(0,\rho)} K(x,y) = R(x) \in L^\infty(\Omega).$$

The next result we want to state relates solutions of local and nonlocal equations. In order to do it, let us fix a Hölder continuous function $\mu : \overline{\Omega} \to \mathbb{R}$ with exponent $\alpha \in (0,1)$, and consider

$$\mathcal{G}_\mu(x,s) = 1 + \frac{\mu(x)\, s}{2(1+\mu(x)^2 s^2)}, \qquad (x,s) \in \overline{\Omega} \times \mathbb{R}. \tag{2.4}$$

The local problem we are interested in is the following

$$\begin{cases} v_t(x,t) = \Delta v(x,t) + \mu(x)|\nabla v(x,t)|^2 & \text{in } \Omega \times (0,T), \\ v(x,t) = h_0(x,t) & \text{on } \partial\Omega \times (0,T), \\ v(x,0) = v_0(x) & \text{in } \Omega\,. \end{cases} \tag{2.5}$$

Observe that if, for the same $0 < \alpha < 1$, we have $\partial\Omega \in C^{2+\alpha}$, $v_0 \in C^{1+\alpha}(\overline{\Omega})$, $h \in C^{1+\alpha,1+\alpha/2}(\partial\Omega \times [0,T])$ with v_0 and h compatible (namely, they are globally $C^{1+\alpha,1+\alpha/2}$ functions of the parabolic boundary of the cylinder) and the equation holds up to the boundary, then Theorem 6.1 of Chapter V in [14] provides a solution $v \in C^{2+\alpha,1+\alpha/2}(\overline{\Omega} \times (0,T])$.

Such a result becomes trivial if we assume $\mu(x) = \mu \in \mathbb{R}$, after the Hopf–Cole transformation, since solutions of the heat equation satisfy the required regularity.

We set here the definition of *classical solution* and then we state our convergence result.

Definition 2.7 We say that $v \in C(\overline{\Omega} \times [0,T]) \cap C^{2+\alpha,1+\alpha/2}(\Omega \times (0,T))$ is a *classical solution* to the Dirichlet problem (2.5) if it satisfies both the equations and the boundary and initial conditions in a pointwise sense.

Finally, the main result of this work establishes that solutions of the deterministic equation KPZ can be uniformly approximated by solutions of nonlocal problems by means of a suitable kernel rescaled.

Theorem 2.8 *Let Ω be a $C^{2+\alpha}$, with $\alpha \in (0,1)$, bounded domain of $\mathbb{R}^N$, $N \geq 1$, and let v be a classical solution of the quasilinear problem* (2.5) *with $h \in C^{1+\alpha}\left(\Omega_{J_\varepsilon} \backslash \Omega \times (0,T]\right)$ such that $h\big|_{\partial\Omega\times(0,T)} = h_0(x,t)$ and $v_0 \in C^{1+\alpha}(\overline{\Omega})$. Assume that J satisfies (J) and that for a.e. x in Ω, $\mathcal{G}(x,s)$ is a $C^{1+\alpha}$ function with respect to the s variable such that that $(\mathcal{G})$ holds true. For any $\varepsilon > 0$, let u^ε denote the unique solution to*

$$\begin{cases} u_t^\varepsilon(x,t) = \dfrac{C(x)}{\varepsilon^2} \displaystyle\int_{\Omega_{J_\varepsilon}} J_\varepsilon(x-y) u^\varepsilon(y;x,t) \mathcal{G}(x, u^\varepsilon(y;x,t)) dy & \text{in } \overline{\Omega} \times (0,T), \\ u^\varepsilon(x,t) = h(x,t) & \text{in } (\Omega_{J_\varepsilon} \setminus \overline{\Omega}) \times (0,T), \\ u^\varepsilon(x,0) = v_0(x) & \text{in } \overline{\Omega}, \end{cases} \tag{2.6}$$

with $C(x)^{-1} = \frac{1}{2} C(J) \mathcal{G}(x,0)$ and $\mu(x) = \dfrac{2\mathcal{G}_s'(x,0)}{\mathcal{G}(x,0)}$ for any a.e. $x \in \Omega$.

Then we have:

$$\lim_{\varepsilon \to 0} \sup_{t \in [0,T]} \left\| u^\varepsilon(x,t) - v(x,t) \right\|_{L^\infty(\Omega)} = 0\,.$$

Acknowledgements

The author is supported under project PGC2018–096422–B–I00 (MCIU/AEI/FEDER, UE) and by Junta de Andalucía FQM-116 (Spain).

References

[1] B. Abdellaoui, I. Peral and A. Primo. On the KPZ equation with fractional diffusion. *preprint. arXiv:1904.04593.*

[2] F. Andreu, J.M. Mazón, J.D. Rossi and J. Toledo. A nonlocal p-Laplacian evolution equation with nonhomogeneous Dirichlet boundary conditions. *SIAM J. Math. Anal.*, 40:1815–1851 (2009).

[3] D.G. Aronson, J. Serrin. Local behavior of solutions of quasilinear parabolic equations. *Arch. Rat. Mech. Anal.*, 25:81–122 (1967).

[4] P. Bates, P. Fife, X. Ren, X. Wang. Travelling waves in a convolution model for phase transitions. Arch. Ration. Mech. Anal., 138:105–136 (1997).

[5] A.L. Barabási and H.E. Stanley. Fractal Concepts in Surface Growth. *Cambridge University Press* (1995).

[6] L. Boccardo, F. Murat, J.P. Puel. Existence results for some quasilinear parabolic equations. *Nonlinear Analysis T.M.A.*, 13:373–392 (1989).

[7] E. Chasseigne, M. Chaves, J.D. Rossi. Asymptotic behavior for nonlocal diffusion equations. *J. Math. Pures Appl.*, 86:271–291 (2006).

[8] C. Cortázar, M. Elgueta and J.D. Rossi. Nonlocal diffusion problems that approximate the heat equation with Dirichlet boundary conditions. *Israel Journal of Mathematics*, 170:53–60 (2009).

[9] C. Cortázar, M. Elgueta, J.D. Rossi, N. Wolanski. Boundary fluxes for non-local diffusion. *J. Differential Equations*, 234:360–390 (2007).

[10] I. Corwin. The Kardar-Parisi-Zhang equation and universality class. *Random Matrices Theory Appl.*, 1:1130001 (2012).

[11] M. Hairer. Solving the KPZ equation. *Annals of Mathematics*, 178:559–664 (2013).

[12] L. Ignat and J.D. Rossi. A nonlocal convection-diffusion equation. *J. Funct. Anal.*, 251:399–437 (2007).

[13] M. Kardar, G. Parisi, Y.C. Zhang. Dynamic scaling of growing interfaces. *Phys. Rev. Lett.*, 56:889–892 (1986).

[14] O.A. Ladyzenskaja, V.A. Solonnikov, N.N. Ural'ceva. Linear and quasi-linear equations of parabolic type. *Translations of Math. Monographs*, vol 23, Providence, 1968.

[15] T. Leonori, A. Molino and S. Segura de León. Parabolic equations with natural growth approximated by nonlocal equations. *Communications in Contemporary Mathematics*, 23(1): 32 pp. (2021).

[16] A. Molino and J.D. Rossi. Nonlocal diffusion problems that approximate a parabolic equation with spatial dependence. *Z. Angew. Math. Phys.*, 67:41 (2016).

[17] A. Molino and J.D. Rossi. Nonlocal Approximations to Fokker-Planck Equations. *Funkcialaj Ekvacioj*, 62:35–60 (2019).

[18] E. Valdinoci. From the long jump random walk to the fractional Laplacian. *Bol. Soc. Esp. Mat. Apl. SeMA*, 49:33–44 (2009).

[19] H. S. Wio, C. Escudero, J. A. Revelli, R. R. Deza, M. S. de La Lama. Recent developments on the Kardar-Parisi-Zhang surface-growth equation. *Phil. Trans. R. Soc. A* 369:396 (2011).

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 284–288)

Symmetry analysis and conservation laws of a family of non-linear viscoelastic wave equations

Almudena del Pilar Márquez[1], **María de los Santos Bruzón**[1]
Department of Mathematics, University of Cadiz, Spain

Abstract

This work considers a non-linear viscoelastic wave equation with non-linear damping and source terms. We analyze the partial differential equation from the point of view of Lie symmetries. Firstly, we apply Lie's method to obtain new symmetries. Hence, we transform the partial differential equation into an ordinary differential equation, by using the symmetries. Moreover, new solutions are derived from the ordinary differential equation. Finally, by using the direct method of multipliers, we construct low-order conservation laws depending on the form of the damping and source terms.

1. Introduction

Recently, several viscoelastic wave equations have been studied. The single viscoelastic wave equation of the form

$$u_{tt} - \Delta u + \int_0^1 h(t-s)\Delta u(x,s)ds + f(u_t) = g(u)$$

in $\Omega \times (0,\infty)$, where Ω is a bounded domain of $\mathbb{R}^N$ $(N \geq 1)$, with initial and boundary conditions, has been extensively studied. Many results concerning non-existence and blow-up in finite time have been proved [3–7, 10].

Furthermore, the non-linear viscoelastic wave equation with damping and source terms

$$u_{tt} - u_{xx} + f(u_t) = g(u), \qquad x \in \Omega,\ t > 0, \tag{1.1}$$

has also been very studied obtaining similar results. As in the single viscoelastic wave equation, in the absence of the source term ($g = 0$), it is well-known that the damping term $f(u_t)$ assures global existence and decay of the solution energy for arbitrary initial data. In the same way, in the absence of the damping term, the source term causes finite time blow-up of solutions with a large initial data (negative initial energy). Here, the interaction between the damping term and the source term makes the problem more interesting.

The aim of this work is to obtain the Lie point symmetries of equation (1.1). Afterwards, we present the reductions obtained from the symmetries, transforming the PDE into an ODE. Moreover, we obtain traveling wave solutions by the comparison between equation (1.1) and similar equations studied previously [1, 2, 8]. Finally, we give a complete classification of the conservation laws admitted by equation (1.1).

2. Lie point symmetries and reductions

It is considered a one-parameter Lie group of infinitesimal transformations in (x,t,u) given by

$$\begin{aligned} x^* &= x + \epsilon\xi(x,t,u) + O(\epsilon^2), \\ t^* &= t + \epsilon\tau(x,t,u) + O(\epsilon^2), \\ u^* &= u + \epsilon\eta(x,t,u) + O(\epsilon^2), \end{aligned} \tag{2.1}$$

where ϵ is the group parameter. These transformations leave invariant the set of solutions of equation (1.1). The associated Lie algebra of infinitesimal symmetries is given by the infinitesimal generator

$$X = \xi(x,t,u)\partial_x + \tau(x,t,u)\partial_t + \eta(x,t,u)\partial_u. \tag{2.2}$$

Each infinitesimal generator (2.2) generates a transformation obtained by solving the system of ODEs

$$\frac{\partial\hat{x}}{\partial\epsilon} = \xi(\hat{x},\hat{t},\hat{u}), \qquad \frac{\partial\hat{t}}{\partial\epsilon} = \tau(\hat{x},\hat{t},\hat{u}), \qquad \frac{\partial\hat{u}}{\partial\epsilon} = \eta(\hat{x},\hat{t},\hat{u}),$$

satisfying the initial conditions

$$\hat{x}|_{\epsilon=0} = x, \qquad \hat{t}|_{\epsilon=0} = t, \qquad \hat{u}|_{\epsilon=0} = u,$$

with ϵ the group parameter.

The symmetry variables are found by solving the invariant surface condition

$$\Phi \equiv \xi(x,t,u)u_x + \tau(x,t,u)u_t - \eta(x,t,u) = 0.$$

For equation (1.1), a PDE with two independent variables, a single group reduction transforms the PDE into ODEs, easier to solve than the original equation.

We require that the transformation (2.1) leaves invariant the set of solutions of equation (1.1). This leads to an overdetermined linear system of equations for the infinitesimals $\xi(x,t,u)$, $\tau(x,t,u)$ and $\eta(x,t,u)$, generated by requiring that

$$\mathrm{pr}^{(2)} X \left(u_{tt} - u_{xx} + f(u_t) - g(u)\right) = 0,$$

where $\mathrm{pr}^{(2)} X$ is the 2-th order prolongation of the vector field X defined by

$$\mathrm{pr}^{(2)} X = X + \eta_x \frac{\partial}{\partial u_x} + \eta_t \frac{\partial}{\partial u_t} + \eta_{xx} \frac{\partial}{\partial u_{xx}} + \eta_{xt} \frac{\partial}{\partial u_{xt}} + \eta_{tt} \frac{\partial}{\partial u_{tt}},$$

with the coefficients

$$\begin{aligned}
\eta_x &= D_x\eta - u_t D_x\tau - u_x D_x\xi, \\
\eta_t &= D_t\eta - u_t D_t\tau - u_x D_t\xi, \\
\eta_{xx} &= D_x(\eta_x) - u_{xt} D_x\tau - u_{xx} D_x\xi, \\
\eta_{xt} &= D_t(\eta_x) - u_{xt} D_x\tau - u_{xx} D_t\xi, \\
\eta_{tt} &= D_t(\eta_t) - u_{tt} D_t\tau - u_{xt} D_t\xi,
\end{aligned}$$

where D_x and D_t are the total derivatives of x and t, respectively.

Applying the previous condition to equation (1.1), we get a system of equations for the infinitesimals. Then, by solving the system, we can make a Lie symmetries classification.

Theorem 2.1 *The Lie point symmetries of the non-linear viscoelastic wave equation* (1.1)*, with $f(u_t)$ and $g(u)$ arbitrary functions, are generated by the operators*

$$X_1 = \partial_x, \quad X_2 = \partial_t.$$

The symmetries of Theorem 2.1 yield to the one-parameter symmetry transformation groups

$$\begin{aligned}
(\hat{x},\hat{t},\hat{u})_1 &= (x+\epsilon,t,u), \quad \text{space translation,} \\
(\hat{x},\hat{t},\hat{u})_2 &= (x,t+\epsilon,u), \quad \text{time translation.}
\end{aligned}$$

From the generator $\lambda X_1 + X_2$, we obtain the traveling wave reductions

$$z = x - \lambda t, \qquad u(x,t) = h(z), \tag{2.3}$$

where $h(z)$ satisfies

$$(\lambda^2 - 1)h'' + f(-\lambda\, h') - g(h) = 0. \tag{2.4}$$

3. Traveling wave solutions

Let us consider the second-order equation (2.4)

$$h'' = \frac{1}{1-\lambda^2} f(-\lambda h') + \frac{1}{1-\lambda^2} g(h). \tag{3.1}$$

We can find equation (2.4) studying other mathematical models. For instance, the general solution of a second-order ODE of the form

$$h'' = \frac{1}{\lambda}\left(\mu h' + \frac{1}{2}h^2 - \omega h - c_0\right), \tag{3.2}$$

with c_0 an arbitrary constant and λ, μ, ω satisfying $\omega = \frac{6\mu^2}{25\lambda}$, was obtained by Kudryashov [8]. The general solution is given in terms of the Weierstrass elliptic function, with invariants $g_2 = 0$ and $g_3 = c_1$,

$$h(z) = \omega_k + \frac{6\alpha^2}{25\beta} - \exp\left\{\frac{2z\alpha}{5\beta}\right\} \mathcal{P}\left(c_2 - \frac{5\beta}{\alpha\sqrt{12\beta}} \exp\left\{\frac{z\alpha}{5\beta}\right\}, 0, c_1\right),$$

where c_1 and c_2 are arbitrary constants.

The comparison between equation (3.1) and equation (3.2) shows that these equations are the same if

$$\begin{aligned} f(-\lambda h') &= \tfrac{1-\lambda^2}{\lambda}\,\mu h', \\ g(h) &= \tfrac{1}{\lambda}\left(\tfrac{1}{2}h^2 - \omega h - c_0\right). \end{aligned}$$

Hence, the solutions of equation (3.1) and equation (3.2) are equal with the previous condition. Finally, by undoing the change of variables (2.3), a exact solution of the non-linear viscoelastic wave equation (1.1) is

$$u(x,t) = \omega + \frac{6\alpha^2}{25\beta} - \exp\left\{\frac{2(x-\lambda t)\alpha}{5\beta}\right\}\mathcal{P}\left(c_2 - \frac{5\beta}{\alpha\sqrt{12\beta}}\exp\left\{\frac{(x-\lambda t)\alpha}{5\beta}\right\},\, 0,\, c_1\right). \tag{3.3}$$

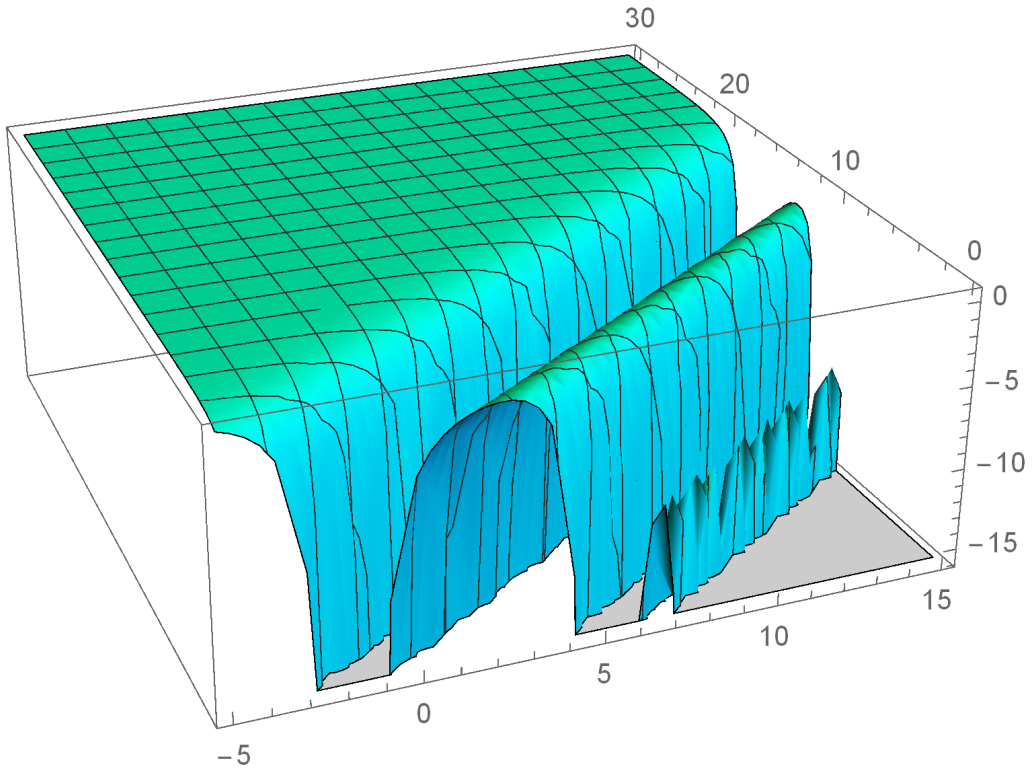


Fig. 1 Solution (3.3) for $\lambda = \alpha = \beta = c_1 = c_2 = 1$.

Solution (3.3) is a soliton (see Fig. 1).

4. Conservation laws

A conservation law admitted by equation (1.1) satisfies the divergence identity

$$D_t T + D_x X = (u_{tt} - u_{xx} + f(u_t) - g(u))Q,$$

called the characteristic equation for the conserved density T and the conserved flux X.

However, the general form for low-order multipliers Q in terms of u and derivatives of u is given by those variables that can be derived to obtain a leading derivative of the equation. Clearly, u_{tt} can be obtained by the derivative of u_t with respect to t, and u_{xx} can be obtained by the derivative of u_x with respect to x.

This determines

$$Q(t, x, u, u_t, u_x)$$

as the general form for a low-order multiplier for equation (1.1).

All low-order multipliers can be found by solving the determining equation

$$E_u((u_{tt} - u_{xx} + f(u_t) - g(u))Q) = 0, \tag{4.1}$$

where E_u represents the Euler operator with respect to u [9], that is

$$E_u = \partial_u - D_x\partial_{u_x} - D_t\partial_{u_t} + D_xD_t\partial_{u_{xt}} + D_x^2\partial_{u_{xx}} + \cdots .$$

Hence, we write and split the determining equation (4.1) with respect to u_{xx}, u_{tt}, u_{tx}, yielding an overdetermined system in Q, $f(u_t)$, $g(u)$. The multipliers are found by solving the system with the same algorithmic method used for the determining equation for infinitesimal symmetries. Thus, we obtain a complete classification of multipliers and conservation laws.

Tab. 1 Multipliers admitted by equation (1.1), with $f(u_t) \neq 0$.

$f(u_t)$	$g(u)$	Q
$f_0u_t + f_1$	arbitrary	$u_x\, e^{f_0 t}$
f_0	arbitrary	$u_x,\ u_t$
f_0	$g_1\, e^{g_0 u} - f_0$	$u_t,\ u_x,\ tu_t + xu_x + \frac{2}{g_0}$
$-g_0 - \frac{1}{f_0u_t+f_1}$	g_0	$f_0u_tu_x + f_1u_x$
$-\frac{4f_0}{u_t+f_1} + f_2$	$\frac{4f_0}{f_1} - f_2$	$u_t + f_1$

The multipliers with $f(u_t) \neq 0$ are shown in Tab. 1.

Theorem 4.1 *All non-trivial low-order conservation laws admitted by the non-linear viscoelastic wave equation* (1.1)*, with* $f(u_t) \neq 0$*, are given below.*

1. *For* $f(u_t) = f_0$*,* $g(u)$ *arbitrary function and* $Q = u_t$*, we obtain the conservation law*

$$\begin{aligned} T &= \frac{1}{2}\,{u_x}^2 + \frac{1}{2}\,{u_t}^2 + \int g\,(u) + f_0\,\mathrm{d}u, \\ X &= -u_t u_x. \end{aligned}$$

2. *For* $f(u_t) = f_0$*,* $g(u)$ *arbitrary function and* $Q = u_x$*, we obtain the conservation law*

$$\begin{aligned} T &= u_t u_x, \\ X &= -\frac{1}{2}\,{u_x}^2 - \frac{1}{2}\,{u_t}^2 + \int g\,(u) + f_0\,\mathrm{d}u. \end{aligned}$$

3. *For* $f(u_t) = f_0$*,* $g(u) = g_1\, e^{g_0 u} - f_0$ *and* $Q = tu_t + xu_x + \frac{2}{g_0}$*, we obtain the conservation law*

$$\begin{aligned} T &= \frac{1}{2g_0}\,2\,t\mathrm{e}^{ug_0}g_1 + \left(t{u_t}^2 + t{u_x}^2 + 2\,u_x x u_t\right) g_0 + 4\,u_t, \\ X &= \frac{1}{2g_0}\,2\,x\mathrm{e}^{ug_0}g_1 + \left(-2\,tu_tu_x - x{u_t}^2 - {u_x}^2 x\right) g_0 - 4\,u_x. \end{aligned}$$

4. *For* $f(u_t) = f_0u_t + f_1$*,* $g(u)$ *an arbitrary function and* $Q = u_x\, e^{f_0 t}$*, we obtain the conservation law*

$$\begin{aligned} T &= u_x\mathrm{e}^{f_0\,t}u_t, \\ X &= \int \mathrm{e}^{f_0\,t}\,(g\,(u) + f_1)\,\mathrm{d}u + \frac{1}{2}\left(-{u_t}^2 - {u_x}^2\right)\mathrm{e}^{f_0\,t}. \end{aligned}$$

5. *For* $f(u_t) = -g_0 - \frac{1}{u_t f_0 + f_1}$*,* $g(u) = g_0$ *and* $Q = f_0u_tu_x + f_1u_x$*, we obtain the conservation law*

$$\begin{aligned} T &= \frac{1}{6}\,f_0\,{u_x}^3 + \frac{1}{2}\,f_0\,{u_t}^2u_x + f_1\,u_xu_t, \\ X &= -\frac{1}{2}\,f_0\,u_t{u_x}^2 - \frac{1}{2}\,{u_x}^2f_1 - \frac{1}{6}\,f_0\,{u_t}^3 - u - \frac{1}{2}\,{u_t}^2f_1. \end{aligned}$$

6. *For* $f(u_t) = -\frac{4f_0}{u_t+f_1} + f_2$*,* $g(u) = \frac{4f_0}{f_1} - f_2$ *and* $Q = u_t + f_1$*, we obtain the conservation law*

$$\begin{aligned} T &= \frac{1}{2}\,{u_x}^2 + \frac{1}{2}\,{u_t}^2 + f_1\,u_t + 4\,\frac{f_0\,u}{f_1}, \\ X &= (-u_t - f_1)\,u_x. \end{aligned}$$

5. Conclusions

In this work, we have obtained some Lie point symmetries for the viscoelastic wave equation (1.1). However, for a future paper we will study the Lie point symmetries complete classification of equation (1.1), in the presence of damping and source terms, for different expressions of the functions f and g. Then, we have constructed the corresponding reduced equation. This reduction makes easier the resolution of the viscoelastic wave equation (1.1), in order to obtain solutions of physical interest such as solitons. Moreover, we have obtained a traveling wave solution from the reduced equation by the comparison between equation (1.1) and comparable equations studied before by other authors. Furthermore, classical Lie symmetries are not the only ones that can be studied. Another symmetries, such as non-classical or potential symmetries, can also be studied in the future. Finally, we have derived the non-trivial low-order conservation laws by using the direct multiplier method developed by Anco and Bluman.

Acknowledgments

The financial support of the *Plan Propio de Investigación y Transferencia de la Universidad de Cádiz* is gratefully acknowledged.

References

[1] Bruzón MS, Garrido TM, Recio E, Rosa R. Lie symmetries and travelling wave solutions of the nonlinear waves in the inhomogeneous Fisher-Kolmogorov equation. *Math. Meth. Appl. Sci.* , 2:1–9, 2019.

[2] Bruzón MS, Gandarias ML. Travelling wave solutions for a generalized double dispersion equation. *Nonlinear Analysis* , 71:2109–2117, 2009.

[3] Ball J. Remarks on blow up and nonexistence theorems for nonlinear evolutions equations. *Quart. J. Math. Oxford* , 28(2):473–486, 1977.

[4] Haraux A, Zuazua E. Decay estimates for some semilinear damped hyperbolic problems. *Arch. Ration. Mech. Anal.* , 150:191–206, 1988.

[5] Kafini M, Messaoudi SA. A blow up result for a viscoelastic system in $\mathbb{R}^N$. *Electron. J. Differential Equations* , 113:1–7, 2006.

[6] Kafini M, Messaoudi SA. A blow up result in a Cauchy viscoelastic problem. *Appl. Math. Lett.* , 21(6):549–553, 2008.

[7] Kalantarov VK, Ladyzhenskaya OA. The occurrence of collapse for quasilinear equations of parabolic and hyperbolic type. *J. Soviet Math.* , 10:53–70, 1978.

[8] Kudryashov N. On "new travelling wave solutions" of the KdV and the KdV-Burgers equations. *Commun. Nonlinear Sci. Numer. Simulat.* , 14(5):1891–1900, 2009.

[9] Olver PJ. *Applications of Lie groups to differential equations*. Verlag: Springer, 1986.

[10] Wang Y. A global nonexistence theorem for viscoelastic equations with arbitrary positive initial energy. *Appl. Math. Lett.* , 22:1394–1400, 2009.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 289–294)

Flux-corrected methods for chemotaxis equations

Alba M. Navarro Izquierdo[1], **M. Victoria Redondo Neble**[1], **J. Rafael Rodríguez Galván**[1]
Universidad de Cádiz, Spain

Abstract

The aim of this work is to review flux correction methods for chemotaxis equations with special emphasis in two directions. Firstly, to study a possible extension to the Keller–Segel equations of some recent research available on literature about well-posedness and error order of flux correction schemes. And secondly, to test the validity of the low order scheme in some practical numerical examples.

1. Introduction

The importance of mathematics to understand biological processes and the number of mathematicians studying biological and medical phenomena has been continuously increasing in recent years. In particular, for chemotaxis phenomena, which model the property of living organisms to migrate in response to chemical gradients. The celebrated classical chemotaxis model was introduced in [8, 9] and, together with all its numerous variants, has attracted significant interest from the theoretical point of view (see e.g. the compilations [4, 7]).

On the other hand, whereas there are very few numerical results in the literature, mathematical modeling of chemotaxis is a challenging task and it has developed into a relatively large and diverse discipline. In fact, the solutions exhibit interesting mathematical properties which are not easily adapted to a classical discrete methods for solving partial differential equations (EDP) like finite elements or other Galerkin methods. For instance, solutions to the Keller–Segel equations satisfy lower bounds (positivity) and enjoy an energy law, which is obtained by testing the equations against non linear functions. Generally speaking, cross-diffusion mechanisms governing the chemotactic phenomena makes them difficult to analyze not only theoretically but also numerically.

2. Setting of the Problem

In this work we focus on the numerical analysis and simulation of some discrete schemes for the classical Keller-Segel system on chemotaxis, which is given by the following equations:

$$\begin{cases} u_t = \alpha_0 \Delta u - \alpha_1 \nabla \cdot (u \nabla v), & x \in \Omega,\ t > 0, \\ v_t = \alpha_2 \Delta v - \alpha_3 v + \alpha_4 u, & x \in \Omega,\ t > 0, \\ \nabla u \cdot \mathbf{n} = \nabla v \cdot \mathbf{n} = 0, & x \in \partial\Omega,\ t > 0, \\ u(x,0) = u_0(x),\ v(x,0) = v_0(x), & x \in \Omega. \end{cases} \tag{2.1}$$

Here u and v are non-negative functions in $\Omega \times [0,T]$ representing the density of cells and chemical-signal, respectively, $T > 0$ is a fixed time, and Ω is a bounded domain in $\mathbb{R}^d$, $d = 2$ or $d = 3$, where the boundary $\partial\Omega$ is Lipschitz and $\mathbf{n}$ is the unit outward normal vector.

A lot of research on this topic has been recently made from an analytical point of view (see e.g. [4, 7] and references therein). Global in time existence and boundedness of the solution has been show if the initial data is small enough, while blow-up in some solutions of (2.1) occurs in many other interesting cases.

The following well-known properties can be higlighted: positivity,

$$u(x,t) > 0, \quad (x,t) \in \Omega \times [0,T], \tag{2.2}$$

and conservtion of the total mass,

$$\int_\Omega u(x,t)\,dx = \int_\Omega u_0(x)\,dx. \tag{2.3}$$

Developing numerical schemes which satisfy the discrete versions of these properties has been the object of many authors, most of whom have focused on finite volume schemes, that in principle fit well with the chemotactic cross difusion term present in (2.1). Among them we can stand out the works of Saad and coworkers [1, 12] and also of Kurganov [5] and coworkers. Some works also use Galerkin discrete schemes, for instance discontinuous Galerkin methods (Kurganov and Epshtyn [6]) and standard finite elements (Saito [13]). In all cases, the strategy is to introduce some linearization of the chemotaxis term in (2.1) and then to use some upwind technique to preserve property (2.2).

 ISBN 978-84-18482-21-2

In this work we focus in the flux correction technique for the following time-stepping numerical scheme that uncouple cells equation from chemical-signal equation: guive a partition of the time interval $[0, T]$ into subintervals of size $k > 0$, at each time step t^{m+1}, we approximate $u(t^{m+1})$ and $v(t^{m+1})$ as follows:

$$\begin{cases} (1/k)v^{m+1} - \Delta v^{m+1} + v^{m+1} = (1/k)v^m + u^m \\ (1/k)u^{m+1} - \Delta u^{m+1} + \nabla \cdot (u^{m+1}\nabla v^{m+1}) = (1/k)u^m. \end{cases} \tag{2.4}$$

Note that, for the sake of simplicty, we have taken $\alpha_0 = \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1$ and also a semi-implicit Euler scheme is introduced, altough some results will generalized to Crank-Nicolson.

Flux correction (see e.g. [11] and references therein) has been investigated for decades for transport equations with the name of FCT (flux-corrected transport). These schemes have applied in many other context like the discretization of time-dependent convection–diffusion or turbulent flows. Also they have been applied specifically to chemotaxis equations [14–16]. But, the difficulty for practical implementation on standard finitite elements libraries and and specifically the lack of solid analytical results for existence of solution and a priori error estimates have made this methods little used excepting a small circle of computer scientists. This situation may have changed because a theory has started to be developed in recent years in a serie of papers of Barrenechea et al [2,3]. Although this theory covers only the steady time-independent case for divergence-free convection-diffusion equations, the purpose of this work has been its exploration and testing the low order solutions in the framework of chemotaxis equations.

3. Algebraic Flux Correction in Chemotaxis

At each time iteration, the system of algebraic equations (2.4) consists of two decoupled subproblems for the unknowns v^{m+1} and u^{m+1}:

$$\left[\boldsymbol{M} + k\boldsymbol{L} + k\boldsymbol{M}\right]v^{m+1} = \boldsymbol{M}v^m + k\boldsymbol{M}u^m, \tag{3.1}$$

$$\left[\boldsymbol{M} + k\boldsymbol{L} + k\boldsymbol{K}(\boldsymbol{\beta}^{m+1})\right]u^{m+1} = \boldsymbol{M}u^m, \tag{3.2}$$

where $\boldsymbol{M}$, $\boldsymbol{L}$ and $\boldsymbol{K}(\boldsymbol{\beta}^{m+1})$ are respectively the mass, diffusion and convection matrices with elements defined as

$$m_{ij} = \int_\Omega \varphi_i\varphi_j\,dx, \quad l_{ij} = \int_\Omega \nabla\varphi_i\nabla\varphi_j\,dx, \quad k_{ij}(\boldsymbol{\beta}^{m+1}) = -\int_\Omega \varphi_j\boldsymbol{\beta}^{m+1}\nabla\varphi_i\,dx.$$

We denote $\boldsymbol{\beta}^{m+1} = \nabla v^{m+1}$ and φ_i is a P_1 piecewise-polynomial basis. The algebraic flux correction technique consists of a conservative manipulation of the matrices $\boldsymbol{M}$ and $\boldsymbol{K}(\boldsymbol{\beta}^{m+1})$ in order to enforce at the discrete level the positivity of the system (3.1)–(3.2). Specifically, the *consistent* mass matrix $\boldsymbol{M}$ is approximated by the diagonal matrix $\boldsymbol{M}_L$ using the well-know mass lumping technique:

$$\boldsymbol{M}_L = \mathrm{diag}(m_i), \quad m_i := \sum_j m_{ij}.$$

On the other hand, negative off-diagonal elements of $\boldsymbol{K}(\boldsymbol{\beta}^{m+1})$ are eliminated by adding an artificial diffusion operator $\boldsymbol{D}$, defined as the symmetric matrix with elements

$$d_{ij} = \max\{-k_{ij}, 0, -k_{ji}\}, \;\; j \neq i, \qquad d_{ii} = -\sum_{j\neq i} k_{ij}.$$

The result is a low order positivity-preserving discretization which, in the 1D case, transforms the linear finite element convection system (3.2) into a first-order upwind difference [10]. The 2D case is much more complicated although error estimates have been recently derived for the steady case [2]. If we denote $\boldsymbol{A}_v = \boldsymbol{M} + k\boldsymbol{L} + k\boldsymbol{M}$ and $\boldsymbol{A}_u = \boldsymbol{M} + k\boldsymbol{L} + k\boldsymbol{K}(\boldsymbol{\beta}^{m+1})$, this low-order system corresponds to

$$\left[\boldsymbol{A}_v + \boldsymbol{D}_v\right] v^{m+1} = \boldsymbol{M}v^m + k\boldsymbol{M}u^m \tag{3.3}$$

$$\left[\boldsymbol{A}_u + \boldsymbol{D}_u\right] u^{m+1} = \boldsymbol{M}u^m \tag{3.4}$$

with $\boldsymbol{D}_v = \boldsymbol{M}_L - \boldsymbol{M}$, $\boldsymbol{D}_u = \boldsymbol{M}_L - \boldsymbol{M} + k\boldsymbol{D}$. And, since the row sums of the matrix $\boldsymbol{D}$ vanish, the error with respect to the original consistent system (3.1)–(3.2) can be written in terms of two vectors

$$f_i^v = \sum_{j\neq i} f_{ij}^v \qquad f_i^u = \sum_{j\neq i} f_{ij}^u$$

where the anti-diffusive fluxes f_{ij}^v and f_{ij}^u are computed from the mass lumping and the artificial diffusion received by each node i:

$$\left((\boldsymbol{M} - \boldsymbol{M}_L)u^{m+1}\right)_i = \sum_{j\neq i} m_{ij}(u_j - u_i), \qquad \left(\boldsymbol{D}u^{m+1}\right)_i = \sum_{j\neq i} d_{ij}(u_j - u_i). \tag{3.5}$$

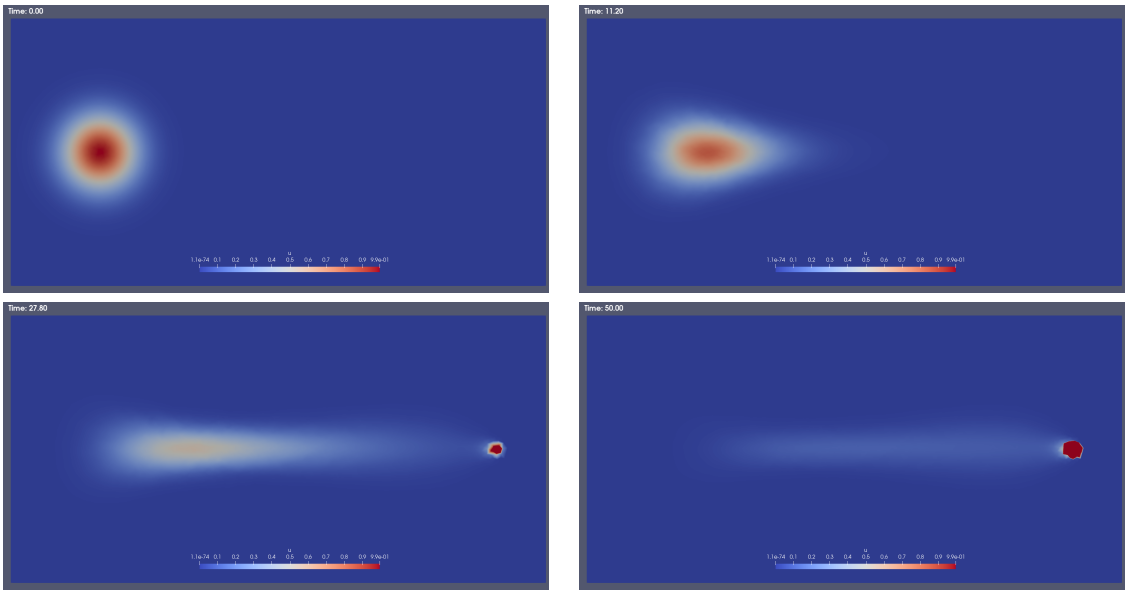


Fig. 1 Chemotatic migration towards fixed chemical concentration. Low order scheme.

Now, instead of adding this fluxes to the right hand side of (3.3), (3.4) and thus obtain the original consistent system, the idea of algebraic flux correction schemes is to limit those anti-diffusive fluxes f_{ij} that would otherwise get back to the original high order solution and cause spurious oscillations. Each flux is multiplied by a solution-dependent correction factor $\alpha_{ij} \in [0, 1]$. The original Galerkin discretization correspond to selecting $\alpha_{ij} = 1$ and may be applied where the solution is smooth, while $\alpha_{ij} = 0$ may be set in the neighborhood of steep fronts, where adding diffusion is appropriate. We impose $\alpha_{ij} = \alpha_{ji}$ to guarantee that the scheme is conservative.

Thus the final form of the algebraic flux correction scheme corresponds to the following system of nonlinear equations:

$$\boldsymbol{A}_v v_i^{m+1} + \sum_{j=1}^{N} (1 - \alpha_{ij}^v) d_{ij}^v (v_j^{m+1} - v_i^{m+1}) = \boldsymbol{M} v_i^m + k \boldsymbol{M} u_i^m, \quad i = 1, \ldots, N, \tag{3.6}$$

$$\boldsymbol{A}_u u_i^{m+1} + \sum_{j=1}^{N} (1 - \alpha_{ij}^u) d_{ij}^u (u_j^{m+1} - u_i^{m+1}) = \boldsymbol{M} u_i^m, \quad i = 1, \ldots, N, \tag{3.7}$$

where d_{ij}^v and d_{ij}^u are respectively the entries of $\boldsymbol{D}^u$ and $\boldsymbol{D}^v$ while $\alpha_{ij}^v = \alpha_{ij}^v(v^{m+1})$ and $\alpha_{ij}^u = \alpha_{ij}^u(u^{m+1})$ are in $[0, 1]$, being $\alpha_{ij}^v = \alpha_{ji}^v$ and $\alpha_{ij}^u = \alpha_{ji}^u$.

For the choice of the flux limiters α_{ij}^u and α_{ij}^v, we are going to us consider the widely used Zalesak limiters (see e.g. [2, 10, 11, 17]). Other appropriate limiters can be set, see e.g. [3]. At the present time, the authors of this work claim that, under some restrictions in the time and space discretization, the following result can be shown:

Theorem 3.1 *Assuming $u_0 \geq 0$ and $v_0 \geq 0$:*

1. *The low order discrete solution (u^{m+1}, v^{m+1}) (obtained from $\alpha_{ij}^u = \alpha_{ij}^v = 0$) is positive for all $m \geq 0$.*
2. *If $\alpha_{ij}^u, \alpha_{ij}^v$ are the Zalesak limiters*
   (a) *There exists a solution of the nonlinear problem* (3.6)–(3.7).
   (b) *The high order solution (u^{m+1}, v^{m+1}) is positive for all $m \geq 0$.*
3. *The approximate solutions (u^{m+1}, v^{m+1}) converge to the exact solution (u, v) with at least suboptimal error order.*

4. Numerical Tests

4.1. Chemotactic transport

In our first numerical test we take a static (non time-dependent) chemical source v and study the migration of the biological organisms u toward high gradients of v. Specifically, the domain Ω is taken as the rectangle

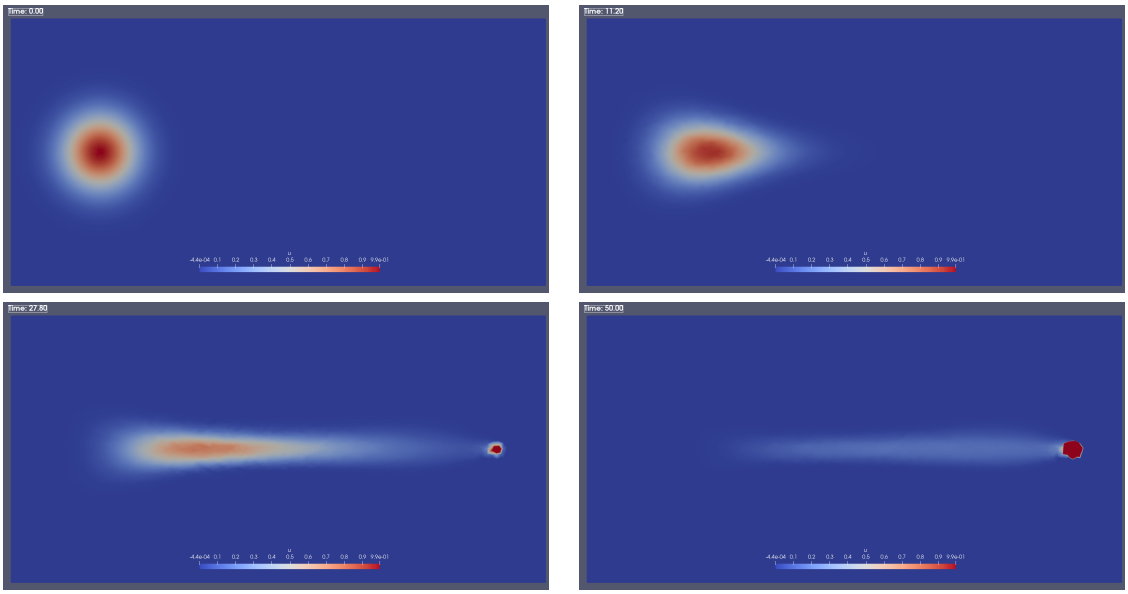


Fig. 2 Chemotatic migration towards fixed chemical concentration. High order scheme.

$[0,5]\times[-1,1]$. we introduce an unstructured mesh with size $h \simeq 0.01$. The time interval $[0,50]$ is discretized so that the CFL condition $k/h < 1$ is verified. On the other hand, v is set as the gaussian function

$$v(x,t) = v(x) = e^{-C((x-4)^2+y^2)},$$

with $C = 50$, which approximately verifies $\nabla v \cdot \mathbf{n} = 0$ on $\partial\Omega$, and u is the solution of $(2.1)_a$ with the less favorable no diffusion condition $\alpha_0 = 0$, $\alpha_1 = 1$. On the other hand, $\nabla u \cdot \mathbf{n} = 0$ on $\partial\Omega$ and the following initial state is chosen:

$$u_0(x) = e^{-C((x-1)^2+y^2)}.$$

Figure 1 shows the result obtained at times $t = 0$, $t = 11.20$, $t = 27.80$ and $t = 50$ with the low order scheme ($\alpha^u_{ij} = \alpha^v_{ij} = 0$). This initial state is transported toward maximum concentration of v. Positivity of solution is maintained strictly and no spurious ripples appear. Figure 2 shows the same test but the high-order scheme (Zalesak limiters) is introduced. In this case, fixed point iterations are introduced to avoid nonlinearity of the scheme. The results are similar although lesser diffusion can be appreciated at intermediate time steps.

4.2. Neuroblast Migration in the Brain

Secondly, we show a numerical test dealing with the migration of neuroblasts (precursor cells of neurons) in the adult brain. This test is part of a project we are working in, together with researchers of Universidad de Sevilla and Universidad de Cádiz. In recent decades, it has been known that neuroblasts are born in a specific part of the adult brain (the Subventricular Zone, SVZ) migrate to other zones: to the Olfactory Bulb (OB) and eventually to lesions in the brain). Some specific parts of the brain (the Corpus Callosum) influences the migration, obstructing it. This process can be modeled by a chemotaxis-like process.

A low order flux correction scheme has been applied in this context (together with other numerical schemes, all of which will be published in a forthcoming work). The results for different time steps are presented in Figure 3, where a source of neuroblasts starts from the SVZ and, bording the Corpus Callosum (represented as a light spot) goes to the OB, located at the left side of the brain.

4.3. Blow up for Keller–Segel System

One of the more challenging characteristics of Keller–Segel equations (2.1) is the fact that finite-time blow-up occurs in many interesting cases, for 2D and 3D, if initial data is not small enough [4]. It has been deeply investigated and many authors have worked in obtaining numerical schemes that maintain positivity and are free of spurious oscillations in blow-up regime. In particular, flux correction and chemotaxis have been studied in [14–16].

Here we just show a validation focused on the low order scheme, which can be programmed in standard finite element libraries without too much additional difficulty. In particular, the low order scheme avoids the necessity of solving nonlinear scheme and its much less computing demanding, Making feasible the use of finer meshes.

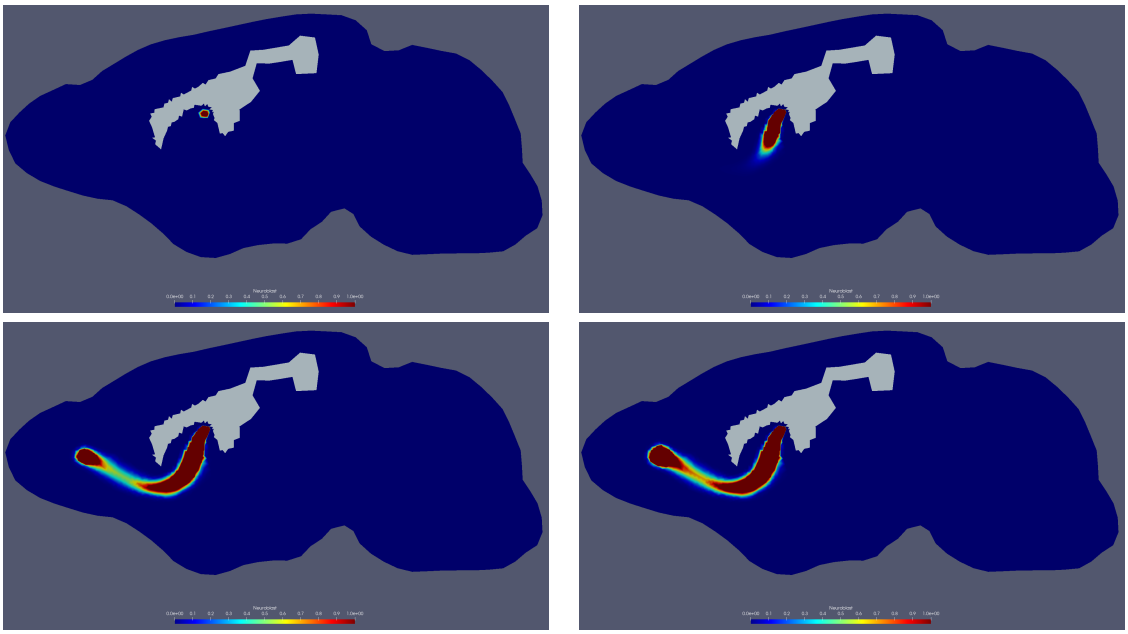


Fig. 3 Chemotatic migration towards fixed chemical concentration. Different time steps (from top left to bottom right). Low order scheme.

Specifically, we consider the numerical test studied in [5], where the domain $\Omega = (-1/2, 1/2)^2$ is meshed with $h \approx 1/100$, $h \approx 1/200$ and $h \approx 1/400$. System (2.1) is solved with $\alpha_0 = \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1$ and

$$v_0 = 0, \quad u_0 = 1000e^{-100(x^2+y^2)}.$$

The blow up which, according to the theoretical results, is expected for this initial data, is conjectured to occur at some $t^* \in (4.4 \times 10^{-5}, 10^{-4})$, where maximum values of u are around 10^4 or 10^5.

Our numerical test with the low-order scheme, maintain the positivity of u and v and reaches the following values:

- At $t = 4.4 \times 10^{-5}$: $\max(u) = 2.58586e+04$, $\min(u) = 1.41495e-18$, $\max(v) = 4.95975e+02$, $\min(v) = 1.41495e-18$.
- At $t = 10 \times 10^{-4}$: $\max(u) = 1.27965e+05$, $\min(u) = 6.35863e-18$, $\max(v) = 4.92729e+02$, $\min(v) = 6.35863e-18$.

It is interesting no observe that mass is conserved, with a constant value of $3.14159e+01$ for u in Ω.

Acknowledgements

Authors are grateful to F. Guillén González, Pedro Núñez Abades, (Universidad de Sevilla, Spain), Carmen Castro González and Daniel Acosta Soba (Universidad de Cádiz) for their support as part of the project of migration of neuroblasts in the brain.

The second author has been supported by *Proyecto PGC2018-098308-B-I00, funded by FEDER/Ministerio de Ciencia e Innovación - Agencia Estatal de Investigación, Spain*.

References

[1] Boris Andreianov, Mostafa Bendahmane, and Mazen Saad. Finite volume methods for degenerate chemotaxis model. *Journal of Computational and Applied Mathematics*, 235(14):4015–4031, May 2011.

[2] Gabriel R. Barrenechea, Volker John, and Petr Knobloch. Analysis of Algebraic Flux Correction Schemes. *SIAM Journal on Numerical Analysis*, 54(4):2427–2451, January 2016.

[3] Gabriel R. Barrenechea, Volker John, Petr Knobloch, and Richard Rankin. A unified analysis of algebraic flux correction schemes for convection–diffusion equations. *SeMA Journal*, 75(4):655–685, December 2018.

[4] N. Bellomo, A. Bellouquid, Y. Tao, and M. Winkler. Toward a mathematical theory of Keller–Segel models of pattern formation in biological tissues. *Mathematical Models and Methods in Applied Sciences*, 25(09):1663–1763, August 2015.

[5] Alina Chertock and Alexander Kurganov. A second-order positivity preserving central-upwind scheme for chemotaxis and haptotaxis models. *Numerische Mathematik*, 111(2):169–205, December 2008.

[6] Yekaterina Epshteyn and Alexander Kurganov. New Interior Penalty Discontinuous Galerkin Methods for the Keller–Segel Chemotaxis Model. *SIAM Journal on Numerical Analysis*, 47(1):386–408, January 2009.

[7] Dirk Horstmann and others. From 1970 until present: the Keller-Segel model in chemotaxis and its consequences. 2003.

[8] Evelyn F. Keller and Lee A. Segel. Initiation of slime mold aggregation viewed as an instability. *Journal of Theoretical Biology*, 26(3):399–415, March 1970.

[9] Evelyn F. Keller and Lee A. Segel. Model for chemotaxis. *Journal of Theoretical Biology*, 30(2):225–234, February 1971.

[10] D. Kuzmin and S. Turek. Flux Correction Tools for Finite Elements. *Journal of Computational Physics*, 175(2):525–558, January 2002.

[11] Dmitri Kuzmin, Rainald Löhner, and Stefan Turek, editors. *Flux-Corrected Transport: Principles, Algorithms, and Applications*. Scientific Computation. Springer Netherlands, Dordrecht, 2012.

[12] Georges Chamoun Saad Mazen and Talhouk Raafat. Finite volume scheme for isotropic Keller-Segel model with general scalar diffusive functions. *ESAIM: Proceedings and Surveys*, 45:128–137, September 2014.

[13] Norikazu Saito. Conservative upwind finite-element method for a simplified Keller–Segel system modelling chemotaxis. *IMA Journal of Numerical Analysis*, 27(2):332–365, April 2007.

[14] R. Strehl, A. Sokolov, D. Kuzmin, and S. Turek. A Flux-Corrected Finite Element Method for Chemotaxis Problems. *Computational Methods in Applied Mathematics*, 10(2), 2010.

[15] Robert Strehl, Andriy Sokolov, Dmitri Kuzmin, Dirk Horstmann, and Stefan Turek. A positivity-preserving finite element method for chemotaxis problems in 3D. *Journal of Computational and Applied Mathematics*, 239:290–303, February 2013.

[16] Robert Strehl, Andriy Sokolov, and Stefan Turek. Efficient, accurate and flexible finite element solvers for chemotaxis problems. *Computers & Mathematics with Applications*, 64(3):175–189, August 2012.

[17] Steven T Zalesak. Fully multidimensional flux-corrected transport algorithms for fluids. *Journal of Computational Physics*, 31(3):335–362, June 1979.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 295–299)

Ejection-collision orbits in two degrees of freedom problems

Mercè Ollé[1], Martha Alvarez-Ramírez[2], Esther Barrabés[3], Mario Medina[2]
1. merce.olle@upc.edu, *Universitat Politècnica de Catalunya and IMTech, Spain*
2. mar@xanum.uam.mx,mvmg@xanum.uam.mx, *Metropolitan Autonomous University, Mexico*
3. barrabes@imae.udg.edu, *Universitat de Girona, Spain*

Abstract

We study the mechanism that explains the birth of ejection-collision orbits (ECO) in a general Hamiltonian problem of two degrees of freedom with some specific properties. The model is a generalization of some N body problems, in which ejection-collision orbits (ECO) are solutions that start at (eject from) the total collision of the particles involved and go back to a total collision after some time. We describe the main tools to prove the existence of such ECO and discuss some results.

1. Introduction

The generalized model considered in this talk has a singularity that represents the total collision in the models coming from the Celestial Mechanics, and two more singularities that can be regarded as two kind of partial collisions when not all the bodies are involved. In suitable variables that regularize the singularities, the total collision is a manifold C topologically equivalent to a sphere minus four points, and the knowledge of the dynamics on C is crucial to prove the existence of ejection-collision orbits.

Some models of the Celestial Mechanics fit in this setting, for example the Symmetric Collinear Four Body Problem (SC4BP, see [1, 5, 9]), the Collinear Three Body Problem (C3BP, see [7]), the Rectangular Four Body Problem (R4BP, see [9]) or the Rhomboidal Four Body Problem (Rh4BP, see [6]).

In order to tackle the problem of the behavior near the total collision, the work of McGehee in 1974, [7], on the triple-collision behavior in the collinear three-body problem was essential. Later, Devaney in 1980 ([3]), used McGehee's work to describe the dynamics on the collision manifold in the planar isosceles three-body problem. We review the ideas introduced in the cited woirks that allow to prove the existence of the ejection-collision orbits in the generalized model, and we show that the same mechanisms apply provided that the potential satisfies certain conditions.

More concretely, we will describe the main characteristics of the dynamics of the general model, and we will prove the existence and give a classification of the ECO that can be obtained.

2. Description of the model

We consider the general system of ODE with two degrees of freedom:

$$\left\{\begin{array}{rcl} \dot{\mathbf{q}} &=& A^{-1}\mathbf{p}, \\ \dot{\mathbf{p}} &=& \nabla U(\mathbf{q}), \end{array}\right. \tag{2.1}$$

where $\mathbf{q} \in \mathbb{R}^2 \setminus \Delta$, $\mathbf{p} \in \mathbb{R}^2$, A is a diagonal constant matrix, $A = \mathrm{diag}(a_1, a_2)$, $a_1, a_2 > 0$, and U is an homogeneous function of degree -1 on $\mathbb{R}^2 \setminus \Delta$, singular in Δ that corresponds to all the possible collisions between the bodies in the context of an N-body problem. In particular $\mathbf{q} = 0 \in \Delta$ corresponds to the total collision of all the bodies.

We introduce McGehee's coordinates (see [7]) (r, v, θ, u) together with a scaling in time (being t the old time and τ the new one) to obtain the system

$$\left\{\begin{array}{rcl} \dfrac{dr}{d\tau} &=& rv, \\ \dfrac{dv}{d\tau} &=& \dfrac{v^2}{2} + u^2 - V(\theta), \\ \dfrac{d\theta}{d\tau} &=& u, \\ \dfrac{du}{d\tau} &=& \dfrac{-vu}{2} + V'(\theta), \end{array}\right. \tag{2.2}$$

where $V(\theta) = rU(\mathbf{q})$ and $V' = dV/d\theta$. Clearly, the system is well defined for $r = 0$.

As previously mentioned, we aim at proving the existence of ECO under certain conditions for the potential $V(\theta)$. Next result states the key hypothesis on the potential.

Proposition 2.1 *Assume that $V(\theta)$ is such that*

$$V(\theta) = \frac{c_b}{\sin(\theta_b - \theta)} + \frac{c_a}{\sin(\theta - \theta_a)} + \widetilde{V}(\theta),$$

where $\theta \in (\theta_a, \theta_b)$ for fixed values θ_a, θ_b such that $0 < \theta_b - \theta_a \leq \pi$, and

- *$c_a > 0$, $c_b \geq 0$ are constants, and $c_b = 0$ if and only if $\theta_b - \theta_a = \pi$;*
- *$\widetilde{V}(\theta) > 0$ is a smooth bounded function in $[\theta_a, \theta_b]$;*
- *$V(\theta)$ has only one non-degenerate critical value at $\theta = \theta_c \in (\theta_a, \theta_b)$, which is a minimum.*

Then, the system of equations (2.2) has two equilibrium points, denoted by $E^{\pm}$, given by $r = 0$, $v = \pm v_c$, $\theta = \theta_c$, $u = 0$, where $v_c^2 = 2V(\theta_c)$. Both equilibrium points $E^{\pm}$ are saddle points, and there exist invariant unstable and stable manifolds $W^{u/s}(E^{\pm})$. Restricted to a fixed energy level $H = h$, $\dim(W^u(E^-)) = 1$, $\dim(W^s(E^-)) = 2$ and $\dim(W^u(E^+)) = 2$, $\dim(W^s(E^+)) = 1$.

As we have mentioned, the removed singularity $r = 0$ (when all particles of the system collide simultaneously) corresponds to the manifold defined as

$$C := \{(r, \theta, v, u) |\, r = 0,\ \theta_a < \theta < \theta_b,\ u^2 + v^2 = 2V(\theta)\}, \tag{2.3}$$

called the *total collision manifold* which is invariant under the flow given by equations (2.2), and contains the equilibrium points $E^{\pm} \in C$.

From Proposition 2.1 we can define ECO as follows:

Definition 2.2 We say that an orbit is a collision orbit if it is contained in $W^s(E^-)$ and an ejection orbit if it is contained in $W^u(E^+)$. An orbit is an ejection–collision orbit (ECO) if it is contained in $W^s(E^-) \cap W^u(E^+)$.

Notice that we have removed the singularity $r = 0$ from system (2.2) but still it has two additional singularities at $\theta = \theta_a$ and $\theta = \theta_b$ that can be regarded as two binary collisions between two pairs of bodies. They can be removed simultaneously through a Sundman type regularization: consider the functions

$$W(\theta) = f(\theta)V(\theta) \quad \text{and} \quad F(\theta) = \frac{f(\theta)}{\sqrt{W(\theta)}},$$

where $f(\theta) = \sin(\theta - \theta_a)\sin(\theta_b - \theta)$ if $\theta_b - \theta_a \neq \pi$, and $f(\theta) = \sin(\theta_b - \theta)$ otherwise. Then, introducing a new variable and the change of time

$$w = F(\theta)u, \qquad d\tau = F(\theta)ds,$$

the system of equations (2.2) transforms into

$$\begin{cases} \dfrac{dr}{ds} = rvF(\theta), \\ \dfrac{dv}{ds} = F(\theta)\left(2hr - \dfrac{v^2}{2}\right) + \sqrt{W(\theta)}, \\ \dfrac{d\theta}{ds} = w, \\ \dfrac{dw}{ds} = -F(\theta)\dfrac{vw}{2} + \dfrac{W'(\theta)}{W(\theta)}\left(f(\theta) - \dfrac{w^2}{2}\right) + f'(\theta)\left(1 + \dfrac{f(\theta)}{W(\theta)}(2hr - v^2)\right), \end{cases} \tag{2.4}$$

where $W' = dW/d\theta$.

3. Key ingredients

Now we shortly mention the main ingredients necessary to prove the existence of different types of ECO.

Symmetry

On the one hand, notice that system (2.4) exhibits the symmetry

$$(r, v, \theta, w, s) \to (r, -v, \theta, -w, -s). \tag{3.1}$$

This can be phrased in terms of solutions as follows: if $\gamma(s) = (r(s), v(s), \theta(s), w(s))$ is a solution then $\overline{\Gamma}$ defined as:

$$\overline{\gamma}(s) = (r(-s), -v(-s), \theta(-s), -w(-s)) \tag{3.2}$$

is also a solution.

Homothetic ECO

On the other hand, a first important result is the existence of an orbit (with $r(s) > 0$) that connects both equilibrium points $E^{\pm}$ (see [8]). It is called the homothetic solution because the configuration maintains the same shape along its evolution for all the time, only changing its size.

Proposition 3.1 *For every fixed level of energy $H = h < 0$, there exists a solution of the system of equations* (2.4) *of the form*

$$\gamma_h(s) = (r(s), \theta = \theta_c, v(s), w = 0),$$

such that $r(s) \underset{s\to\pm\infty}{\longrightarrow} 0$.

Notice that this is an ejection-collision orbit since it starts and ends at $r = 0$.

Poincaré section and map

A convenient Poincaré section Σ, which will be a keystone to show the existence of ECO, is the set where partial collisions occur, that is, where $\theta = \theta_{a,b}$. We denote by Σ the union of two half planes $\Sigma_{a,b}$:

$$\Sigma = \Sigma_a \cup \Sigma_b = \{(r, v, \theta, w) \mid r \geq 0,\ w = 0,\ \theta = \theta_a\} \cup \{(r, v, \theta, w) \mid r \geq 0,\ w = 0,\ \theta = \theta_b\}\,.$$

We consider the Poincaré map (in forward time) defined on Σ

$$\mathcal{P} : \Sigma \longrightarrow \Sigma, \tag{3.3}$$

as $\mathcal{P}(\mathbf{Z}) = \Phi_s(\mathbf{Z})$, where Φ_s is the flow associated to the system (2.4), and s is the first positive time needed to reach the section Σ starting at $\mathbf{Z}$. In a similar way we define $\mathcal{P}^{-1}$, the Poincaré map in backward time.

Collision manifold

The already defined total collision ($r = 0$) manifold C in (2.3), is a 2-dimensional manifold, topologically equivalent to a sphere minus four points, independent of the total energy h, see Figure 1. It is invariant under the flow (2.4), and we remark that is gradient-like with respect the variable v, that is, $dv/ds \geq 0$.

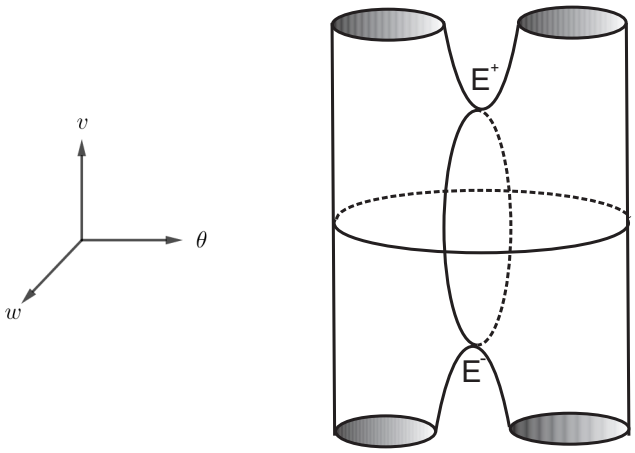


Fig. 1 Qualitative scheme of the total collision manifold C. The two equilibrium points $E^{\pm}$ are also shown.

Dynamics on the invariant manifolds

As stated in Proposition 2.1, the equilibrium points are hyperbolic and, for a fixed value of the energy h, each one has associated two invariant manifolds: one of dimension one, the other of dimension two. The behavior of $W^{u/s}(E^{\pm})$ determines the existence and type of ECO.

In Figure 2 we show the behavior of the one dimensional manifolds $W^u(E^-)$ (by symmetry, we obtain $W^s(E^+)$). Any given branch is an orbit that revolves around C and then it can only exhibit two different behaviors: either it tends to E^+ (becoming an heteroclinic connection, degenerate case) or the trajectory "escapes" towards $v \to +\infty$ along one of the upper legs of the total collision manifold. In the latter, there are four different types of behavior: types I, II, III and IV (shown in Figure 2).

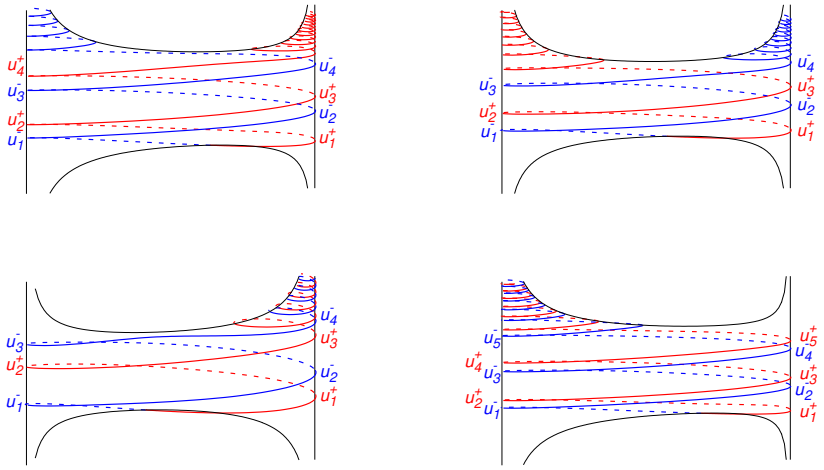


Fig. 2 Branches of the invariant manifolds $W^u_+(E^-)$ (in red) and $W^u_-(E^-)$ (in blue) on the collision manifold. The plots show the four different scenarios in the non-degenerate cases: types I, II, III, IV.

We can provide a classification for the 1-dimensional manifolds $W^u_\pm(E^-)$ and $W^s_\pm(E^+)$ as follows. Let $\mathcal{S}$ be the set of all possible sequences, just taking into account the elements a and b. We define

$$\begin{array}{rcl} \mathcal{I}^+ : W^u(E^-) & \longrightarrow & \mathcal{S} \\ \Gamma & \longrightarrow & \sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n, \ldots) \end{array}$$

where

$$\sigma_j = \begin{cases} a & \text{if the } j\text{-th intersection of } \Gamma \text{ with } \Sigma \text{ is at } \Sigma_a, \\ b & \text{if the } j\text{-th intersection of } \Gamma \text{ with } \Sigma \text{ is at } \Sigma_b, \end{cases} \qquad \text{for } j \geq 1.$$

The sequence $\mathcal{I}^+(\Gamma)$ codes the partial collisions (intersections with Σ) forwards in time for the unstable manifold. Similarly, we can define $\mathcal{I}^-$ on $W^s(E^+)$, obtaining a sequence of partial collisions backwards in time. Using the symmetry of the problem we have that

$$\mathcal{I}^+(W^u_+(E^-)) = \mathcal{I}^-(W^s_-(E^+)) \qquad \text{and} \qquad \mathcal{I}^+(W^u_-(E^-)) = \mathcal{I}^-(W^s_+(E^+)).$$

We classify the behavior of the 1-dimensional manifolds $W^u_\pm(E^-)$ and $W^s_\pm(E^+)$ using the number of full turns of each branch, their intersections with the section Σ and the map $\mathcal{I}^+$. In what follows, the sequence $\underline{\star, \bullet}, \overset{n)}{\ldots}, \star, \bullet$ denotes that the sequence $\star, \bullet$ is repeated n times. For example, the sequence $(\underline{b, a}, \overset{n)}{\ldots}, b, a, b, b, b, \ldots)$ represents an orbit with a sequence of n pairs of collisions b, a (a collision of type b followed by a collision of type a) and then the orbit only has collisions of type b forwards in time. Analogous interpretations are given for other sequences.

With respect to the two dimensional invariant manifolds, they are glued to the total collision manifold C and their intersections are

$$W^u(E^+) \cap C = \Gamma^u_\pm \quad \text{and} \quad W^s(E^-) \cap C = \Gamma^s_\pm, \tag{3.4}$$

where each $\Gamma^u_\pm$ (resp. $\Gamma^s_\pm$) is an orbit that escapes forwards (respectively backwards) in the v-direction, $v \to +\infty$ (resp. $v \to -\infty$) through one of the legs of C. See Figure 3 where the homothetic orbit is also shown.

4. Existence of ECO

We can prove the existence of the different types of ECO depending on the behavior of the 1-dimensional invariant manifolds. Actually, we characterize the ECO by its finite number of successive binary collisions with Σ. Let σ be a finite sequence of collisions of type a and b. Following the notation given above, we will say that an ECO is of type σ if its orbit describes forwards in time the finite sequence of binary collisions encoded by σ.

Using the behavior of the invariant manifolds, we are able to prove the following results.

Theorem 4.1 *For any natural number $m \geq 1$, the system* (2.4) *has an ejection-collision orbit of type*

$$(\underline{a}, \overset{m)}{\ldots}, a), \quad \text{and} \quad (\underline{b}, \overset{m)}{\ldots}, b).$$

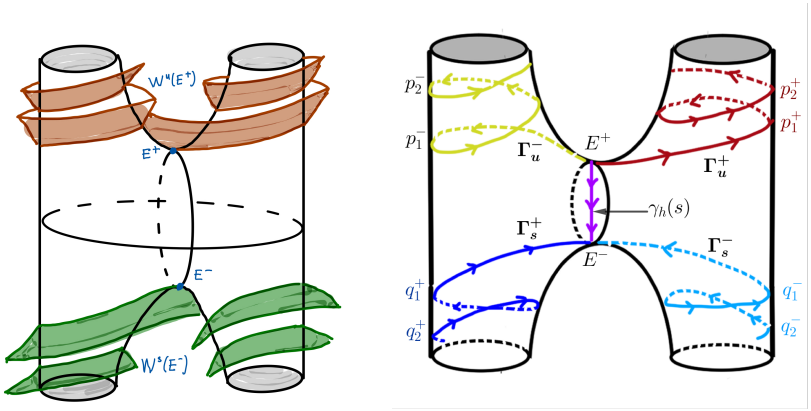


Fig. 3 Left: Qualitative behavior of the 2D-invariant manifolds $W^u(E^+)$ and $W^s(E^-)$. Right: Behavior of $\Gamma_\pm^{u/s}$ and the homothetic solution $\gamma_h(t)$.

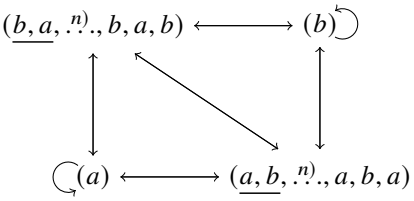
In the case that the 1-d invariant manifolds associated with the equilibrium points are non-degenerate, we can prove the following result for type I. Similar results for types II, III and IV can be found in [2].

Theorem 4.2 *Suppose that $W_\pm^u(E^-)$ are of type I, and let $n \geq 1$ be the number of full turns performed by the branches of the 1D-invariant manifolds before escaping through different arms of C. Then*:

(*a*) *There exist ejection-collision orbits exhibiting $2n+1$ collisions of types*

$$(\underline{b,a},\overset{n)}{\ldots},b,a,b) \qquad \text{and} \qquad (\underline{a,b},\overset{n)}{\ldots},a,b,a).$$

(*b*) *There exist ejection-collision orbits exhibiting any sequence that can be obtained by the following graph*:



Acknowledgements

M. Alvarez-Ramírez is partially supported by Programa Especial de Apoyo a la Investigación de UAM (Mexico) grant number I5-2019. E. Barrabés has been supported by grants MTM2016-80117-P, PGC2018-100928-B-100 (MINECO/FEDER, UE) and Catalan (AGAUR) grant 2017 SGR 1374. M. Ollé has been supported by grant PGC2018-100928-B-100 (MINECO/FEDER, UE) and the Catalan (AGAUR) grant 2017 SGR1 049.

References

[1] M. Alvarez-Ramírez, E. Barrabés, M. Medina, and M. Ollé. Ejection-collision orbits in the symmetric collinear four-body problem. *Commun. Nonlinear Sci. Numer. Simul.*, 71:82–100, 2019.

[2] M. Alvarez-Ramírez, E. Barrabés, M. Medina, and M. Ollé. Ejection-collision orbits in two degrees of freedom problems in celestial mechanics. To appear in *J. Nonlinear Science*, 2021.

[3] R. L. Devaney. Triple collision in the planar isosceles three-body problem. *Invent. Math.*, 60(3):249–267, 1980.

[4] E. Lacomba, Mouvements voisins de collision quadruple dans le problème trapezoidal des 4 corps. *Celest. Mech. Dyn. Astron.*, 31 (1983) 23–41.

[5] E. Lacomba, M. Medina, Symbolic dynamics in the symmetric collinear four-body problem. *Qual. Theory Dyn. Syst.*, 5 (1) (2004) 75–100.

[6] E.A. Lacomba, E. Pérez-Chavela. Motions close to escapes in the rhomboidal four body problem *Celest. Mech. Dyn. Astron.*, 57 (1993) 411–437.

[7] R. McGehee, Triple collision in the collinear three-body problem. *Invent. Math.*, 27 (1974) 191–227.

[8] R. Martínez. On the existence of doubly symmetric "Schubart-like" periodic orbits. *Discrete Contin. Dyn. Syst. Ser. B*, 17(3) (2012) 943–975.

[9] C. Simó, E. Lacomba. Analysis of some degenerate quadruple collisions. *Celest. Mech.* 28 (1-2) (1982) 49–62.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 300–306)

Teaching experience in the Differential Equations Semi-Virtual Method course of the Tecnológico de Costa Rica

Norberto Gerardo Oviedo Ugalde[1]
noviedo@itcr.ac.cr *Instituto Tecnológico, Costa Rica*

Abstract

In 2018, the Differential Equations course for Engineering students is incorporated into the CEDA-TEC Digital Teaching Vice-Rector's Project. The main objective of this work is to disclose my teaching experience in the hard labor of planning, implementation, creation of materials and evaluation of the aspects that are carried out and implemented in this course under the semi-virtual modality.

Finally, it shows some relevant results obtained in an assessment questionnaire about the course applied to the students.

1. Introduction

In 2018 the idea was born from CEDA (Academic Development Center)-Tecdigital to virtualize the course of differential equations in the engineering area therefore, this article shares teaching experience on its planning, creation-development of teaching materials and support, implementation in the classroom, evaluative aspects and also some results obtained in evaluation questionnaire applied to students of the course under which this modality was applied.

2. Semi-virtual mode course planning (bimodal)

It is important to know that in our case a semi-virtual course is understood as the one in which we work every week under two types of classes:

In Person, It consists of a two-hour classroom class, where the teacher is simply a guide or mediator of the process and the student is the main participant and builder of knowledge. During these face-to-face sessions, the teacher makes a theoretical presentation with illustrative examples of the topics covered, clarifies doubts in semi-virtual class and plans the learning activities in the classroom in such a way that they allow the student to achieve collaborative work and through a type of flipped learning.

Virtual, in this type of class, the student works at home (or another place of study that he/she deems appropriate) on the course contents that were previously assigned by the teacher through guidance from said class, this can be based on carrying out a didactic guide, assigned work exercises, view videos and some support material such as books or applications made in software.

The Tecdigital platform will be used as a means of communication on the planification, visualization and interaction to realize the distinct activities inside or outside of the classroom each week.

As a first stage in the first semester of 2018, a course on planning and instructional design is carried out to teachers from various areas of the CEDA (Center for Teaching Development) department of training of the Technological of Costa Rica. The initial objective of the course is to know about basic principles of instructional design for virtual learning environments, learning experiences and educational resources for virtuality, evaluation of learning in virtuality and all the above with the great purpose of planning and developing the design or planning of the own specific course in our case of differential equations under this modality. In it, the following aspects are considered:

- **Population**

  Students of the course are typical of engineering careers, said student population requires as requirements to have passed courses in differential calculus and linear algebra, students attend between the IV and V semester of their study plans, maximum per group is 32 students.

- **Duration**

  The course lasts 16 weeks with a weight of 4 credits for 12 hours per week, distributed as follows: 2 face-to-face hours, 2 face-to-face consultation and 8 work independently at home or another appropriate place.

- **Curriculum design and organization**

  In this segment, aspects of instructional planning or design are organized and stipulated in a matrix form in the most explicit and detailed way.

Week #	Objectives	Contents	Learning Experiences	Means, materials and resources	Assessment

3. Creation of teaching materials and support

Once the instructional design has been prepared and reviewed by experts in the area, the preparation of the didactic and support materials embodied in such planning proceeds in a second stage. Each of them is briefly described below.

3.1. Teaching guides

A few weeks back, didactic guides were developed with the purpose of educating under guidance given by the teacher to put into practice knowledge acquired on topics under study. These guides are resolved individually prior to the face-to-face class in their homes or another place that each student considers appropriate.

3.2. Theoretical summaries

As an introductory activity to each face-to-face class, the teacher provides students with a brief theoretical summary in order to quickly return to the concepts covered during the week. It is understood that the students prior to the face-to-face class must have studied the theoretical concepts of the week at home or in another appropriate place.

3.3. Written support material

• At the Tecnológico de Costa Rica (TEC), various mathematics teachers have developed theoretical support materials for the different courses, for ours on differential equations there are five brochures by Sharay Meneses Rodríguez, MSc namely:

- First-order ordinary differential equations
- Application problems using differential equations
- Higher-order differential equations
- Vibratory motion problems
- Laplace transformed

The previous brochures have served as the basis for the student of the Differential Equations Department for years; however, in addition to them, they have been supplemented with other materials from other of our school's teachers.

- First Order Ordinary Differential Equations with interactive support and illustrative videos. (MSc. Norberto Oviedo Ugalde MSc, 2019)
- Differential Equations (Luis Alejandro Acuña Prado, Ana Marcela Rojas Loaiza, María Nazarelle Rojas Machado, 2019)

In the weekly orientations and in accordance with the provisions of the instructional design, the teacher assigns the students the sessions that are to be studies based on the materials previously mentioned.

3.4. Classroom exercise Videos

As a way of providing support on the various topics studied in the course, 46 videos on solved exercises are produced, in which detailed explanations of them are provided. In order to understand the videos, the learner is instructed to have previously studied the concepts discussed there, since some of the videos include exercises that are not so traditional or with a medium degree of difficulty.

During the process of the creation of the videos, as we had no experience in it, at the beginning, options such as recorders were tested through tablets with a pencil, which apart from having a high cost, was not appropriate in our case because I did not feel comfortable and handwriting was difficult, that is why in a first stage it is then chosen to carry out the exercises in Power Point presentations in which the resolution of the exercises can be presented in a more attractive and dynamic way. For the elaboration of the slide template, we had the support of a graphic designer from TEC, Ing. Luis Carlos Guzmán Arias, likewise the fingering and animation of the resolution processes of the exercises embodied in the presentations was overseen by the assistant from the School of Mathematics, Dayana Calderón Prado.

In the second stage, one proceeds to the review and pertinent adjustments to leave the suitable presentation of every exercise and without errors and then to proceed to the recording of the explanation of the presentation by means of a screen and audio capturer.

In a third stage, we proceed to look for appropriate screen recorders that feel comfortable and are appropriate to our needs. In our case, it began with recorders, such as Screen Record, Apowersoft, Screencasr-o-Matic, Debut, of which, the paid version of Debut is chosen because of its ease in adjustment and comfort.

In the fourth stage, once the first versions of the videos have been recorded, they were reviewed in terms of audio, editing errors and text, which is why, the video was edited and an assistant producer was used, Jonnathan Ramirez of TecDigital. In some cases, it goes to a second or third revision and editing to obtain the final product. In the final stage, REA forms are filled out in order to obtain a final review of the form and thus be published in the TEC repository through videos on YouTube to be shared by students in general. Some examples of video cover are shown in figure 1 and visualized through the links: `https://youtu.be/M-Kz91akjko`, `https://youtu.be/kGrPaa2esqM`.

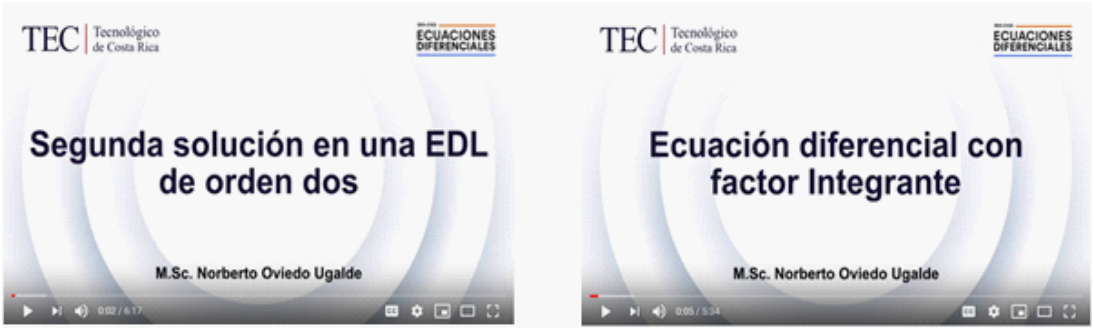


Fig. 1 Covers of videos developed in the semi-virtual course

3.5. Interactive notebooks using Mathematics software in CDF player form

For six years, I have worked with the non-free Mathematica software, through a license given to teachers from the University of Costa Rica, where I have worked in parallel with Tecnológico de Costa Rica during that time. As product of my graduate work in my Educative Mathematics Masters from the University of Costa Rica, I have created some interactive pages on my field of study in differential equations of the first order, where the student can interact and visualize, step by step, the involved processes of different ordinary equations of the first order (ED01).

Given, that the Tecnológico de Costa Rica does not possess the license for said Mathematic software, I begin by investigating how to use or readjust the pages that have been already worked on, in the free software. This is how, the free format of the Mathematica CDF software (Documents in computable format, freely downloadable from `https//www.wolfram.com/cdf-player/index.es.html?footerlang`), player, which offers the interactive and dynamic option of computable documents worked in Mathematica software in a free format. The respective adjustments and adjustments are made to the programming codes elaborated by means of the Mathematica software

and saved in CDF format until the pages or interactive notebooks of free access are made. For the differential equations semivirtual course, three interactive notebooks were divided in the following manner:

I. Interactive notebook on first-order ordinary differential equations (published in the TEC School of Mathematic magazine and can be viewed on `https://tecdigital.tec.ac.cr/revistamatematica/material_didactico/revisado/`

II. Interactive notebook on superior order linear, differential equations. (`https://tecdigital.tec.ac.cr/revistamatematica/material_didactico/revisado/`).

III. Interactive notebook on differential equations by Laplace transform. (`https://tecdigital.tec.ac.cr/revistamatematica/material_didactico/revisado/`).

Each interactive notebook comprises interactive pages on the different topics of interest, in which a brief theoretical summary (definitions, theorems and resolution processes) of the different topics to be studied is presented, along with predetermined and resolved examples in which the reader will be able to visualize his resolution process step by step in a dynamic and interactive way and serves as a support tool for the study of the topics presented there.

3.6. Evaluations online (quizzes) using GAAP

One form of online evaluation developed consists of the creation of virtual quizzes every two weeks on the different topics studied in them. To do this, a series of single-selection and false-true questions are drawn up and developed in Latex code (mathematical text editor), which later serve as the basis for the creation, programming and revision of virtual quizzes by means of the TEC's own platform through the learning activities manager of the same.

In general, the virtual quizzes consist of 5 single-choice questions that are worth 1 point each, where theoretical concepts are directly evaluated and two or three true or false questions where more concepts must be developed in order to give the correct answer. These quizzes are enabled online to students on weekends for a maximum of 12 hours, where once entered, there is no option to return and leave incomplete, meaning that each student must have previously studied the theoretical concepts. Once the virtual quiz has been completed, the student automatically receives the grade.

4. Course implementation

Once the materials and support resources described in the previous section have been developed and reviewed, we proceed to the implementation of what is embodied in the instructional design, that is, to put it into practice as planned. Since 2017 to the present, I have been teaching the differential equations course, in which as the various materials were developed, they were being implemented. I clarify that it was until the second half of 2019 that has been fully applied to a group under the semi virtual modality.

The Tecnológico de Costa Rica uses its own platform, TecDigital, where the course was developed and organized and presented in an orderly way, with week-by-week sections of course presentation, weekly orientations, information from the department chair, teaching materials, evaluations of GAAP tests and other documents that are used for the course. This serves as a means of information and teacher/student interaction. In figure 2, you can see a screenshot of the components of the TecDigital platform.

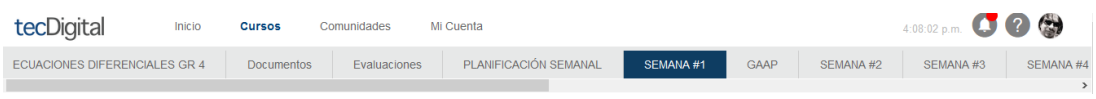


Fig. 2 Semi-virtual course portal group 04 of the I semester 2020

5. Aspectos evaluativos

The semi-virtual modality course carries out the same three partial exams of the other groups of the chair, which are theoretical and developmental in person in a classroom on non-teaching days. The three partials with a total value of 81% and a percentage weight of each of 27%. Furthermore, there will be a percentage of 19%, destined for

perhaps short, virtual evaluations every two weeks through the learning activities manager (GAAP) of TecDigital and tasks during the weeks that there is no virtual quiz.

Of a formative nature, in some weeks the student, prior to the face-to-face classes, will make didactic guides, and sometimes in the class, a contributory work in pairs on the weekly subject.

6. Results of the course evaluation tools

In order to learn about student information and also the appreciation and assessment of experience on the implementation of this semivirtual course, two questionnaires were applied:

I. Teacher questionnaire

It consists of a teacher's own questionnaire based on 15 closed questions in which it is first intended to know information about the group's students, such as whether they have repeated the course, weighted average enrollment, how many hours they dedicate to the course, etc., and others such as those shown in Figure 3 that allows direct appreciation of aspects of planning, organization and implementation of the course under this semi-virtual modality, which are important forms of feedback in order to improve.

13.	¿Con respecto a planteamiento y organización del curso en el portal del TEC digital, qué calificación le daría a este de curso de ecuaciones diferenciales bajo la modalidad semivirtual? ( ) muy mala ( ) mala ( ) buena ( ) muy buena
14.	¿ Los materiales de apoyo brindados como material escrito, videos, aplicaciones de software en CDF player, le beneficiaron y sirvieron de apoyo en el aprendizaje de los distintos contenidos del curso? ( ) nada ( ) aveces ( ) casi siempre ( ) siempre
15.	¿En términos generales qué calificación le daría a este de curso de ecuaciones diferenciales bajo la modalidad semivirtual? ( ) muy mala ( ) mala ( ) buena ( ) muy buena

Fig. 3 Semi virtual student appreciation questionnaire questions

With regard to results obtained by applying group instrument 09 of the second semester of 2019 where 30 students participated, in general they indicated, that average group age ranges from 20 to 23 years, only 1 work, two repeat courses, on average dedicate 4 to 6 hours to the course weekly, weighted average of average enrollment of 80 to 90. Questions 13, 14 and 15 in Figure 3 also have the following:

Question #13			**Question #14**			**Question #15**	
	Frecuency			Frecuency			Frecuency
Very bad	0		Very bad	0		Very bad	0
Bad	0		Bad	0		Bad	0
Good	7		Good	9		Good	10
Very good	23		Very good	21		Very good	20

The results obtained in these questions show a very good acceptance and assessment on the part of the students of the semi virtual modality course, that is, they consider there was a very good planning, organization of the course in the TecDigital portal, the materials provided (theoretical material, CDF player software applications and videos supported them and helped them learn the different topics studied in the course.

II. Course planning and portal questionnaire – generic

This questionnaire consists of an instrument with 9 closed questions on general student information, 32 closed questions on a Likert scale from 1 to 10, 1 being the lowest score and 10 the highest and 1 open question. The same is applied during week 14 by the person in charge of the project through the TecDigital in charge of Julia Espinoza, which has the objective of assessing these aspects: general student information, objectives, contents,

learning activities, educational and evaluation materials, everything related to the planification, organization and evaluation of the course portal. Below, there is a sample of the main results obtained from the 29 students who filled out the questionnaire are shown by means of tables uch as those shown in Figure 4.

Instructional Design Component	Average	Standard Deviation
Objectives	9.38	0.98
Contents	9.41	1.05
Contents	8.61	1.90
Support Materials	9.61	0.78
Evaluation of Learning	9.19	9.18
General Average	**9.18**	

QUESTION	Answer	
	Yes	No
When enrolling this course, did you know that it would be taught bimodally or semivirtually?	6	23
Based on your current experience in this course, would you take a course in this modality again?	26	3

QUESTION	Answer	
	Yes	No
Is the information structure of educational resources clear?	29	0
Was the amount of information (textual and graphic) provided in the educational resources sufficient to understand the topic?	28	1
Do the multimedia elements used give correct support to the didactic content?	29	0
Does the organization of the course portal allow easy navigation between its contents?	29	0
Does the course portal organization make it easy to locate available learning materials and resources?	28	1

Fig. 4 Main results obtained from the 29 students who filled out the questionnaire

Student comments (verbatim copy, no spell check):

- All very well, the only thing I don't like is the idea of only coming one day a week.
- As a suggestion more practice could be added.
- They should use a different evolution to the face-to-face courses, as in some of the CDI that the exams are worth 60.
- Complementing the semi-virtual modality with the teacher who taught the course makes the course simple thanks to the large amount of content available digitally and face-to-face classes.
- The CDF-player material is very good. Excellent service.

7. Conclusions-recommendations

1. Some resistance persists to this type of course modality.
2. Instructions in weekly guidance and platform should be timely.
3. Requires discipline on the part of the student to carry out step by step what is indicated in each week.
4. Strengthen evaluations in line with GAAP or other related evaluations.
5. To promote diverse, classroom activities for major interaction on behalf of the student.
6. Have a pdf in which you can share more quickly and directly what is captured on the TECdigital platform.

References

[1] Abell, Martha L. y Braselton James P. *Differential Equations with Mathematica*. Elsevier Science & Technology Books, Academic Press, 2016.

[2] Ayres, Frank Jr. *Ecuaciones diferenciales*. McGraw Hill-Serie Schaum, México, 1991.

[3] Boyce, W. E. e Diprima, R. C. *Ecuaciones diferenciales y problemas con valores en la frontera*. Editora Limusa Wiley, 4a edición, México.

[4] Coddington E. *Introducción a las ecuaciones diferenciales ordinarias*. Compañía Editorial Continental S.A, 1968.

[5] Lomen D. y Lovelock D. *Ecuaciones Diferenciales a través de gráficas, modelos y datos*. Primera edición. Compañía editorial Continental, México, 2000.

[6] Zill, Dennis G. *Ecuaciones Diferenciales con Aplicaciones de Modelado*. Editorial Thompson, sétima edición, México, 2002.

[7] CDF: Formato de documento computable. *Los documentos cobran vida con la potencia de la computación*. `https://www.wolfram.com/cdf/` .Consultado 10/01/2017 .

[8] Wolfram Mathematica.demostrations Projects. *Interactives demostrations.*. `http://demonstrations.wolfram.com/` .Consultado 24/09/2015.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 307–312)

Nonlinear Analysis in Lorentzian Geometry: The maximal hypersurface equation in a Generalized Robertson-Walker spacetime

José A. S. Pelegrín[1]
jpelegrin@ugr.es *Universidad de Granada, Spain*

Abstract

In this article we obtain a uniqueness result for the maximal hypersurface equation in a spatially open Generalized Robertson-Walker spacetime by means of Bochner's technique and a generalized maximum principle.

1. Introduction

Maximal hypersurfaces have played a key role in the study of General Relativity since they describe the physical space that can be measured in the transition from an expanding to a contracting phase of the universe. Maximal hypersurfaces constitute a useful initial set for the Cauchy problem in General Relativity [20]. Namely, Lichnerowicz proved that a Cauchy problem with initial conditions on a maximal hypersurface is reduced to a second order nonlinear elliptic differential equation and a first order linear differential system [10]. Moreover, the existence of constant mean curvature spacelike hypersurfaces (and in particular maximal) is necessary for the study of the structure of singularities in the space of solutions of Einstein's equations [2].

From a mathematical standpoint, maximal hypersurfaces enable us to understand the structure of the spacetime. Indeed, for some asymptotically flat spacetimes the existence of a foliation by maximal hypersurfaces was proved in [4]. As a matter of fact, maximal hypersurfaces appear as critical points of the area functional (see for instance [3]).

The study of maximal hypersurfaces from a mathematical perspective was boosted by the discovery of new nonlinear elliptic problems associated to these geometric objects. Indeed, the function defining a maximal graph in the $(n+1)$-dimensional Lorentz-Minkowski spacetime satisfies a second order PDE known as the maximal hypersurface equation in $\mathbb{L}^{n+1}$. Furthermore, the well-known Calabi-Bernstein theorem states that the only entire solutions to the maximal hypersurface equation in $\mathbb{L}^{n+1}$ are the affine functions. This result was proved by Calabi [5] for $n \leq 4$ and later extended to arbitrary dimension by Cheng and Yau [6].

These Calabi-Bernstein type results for the maximal hypersurface equation have been a subject of study in recent years, being extended to several ambient spacetimes such as standard static spacetimes [18, 19], pp-waves [16], doubly warped product spacetimes [8], among others. In this article we will focus on the models known as Generalized Robertson-Walker (GRW) spacetimes. These spacetimes were introduced in [1] to extend the classical notion of Robertson-Walker spacetime to the case where the fiber does not necessarily have constant sectional curvature. In particular, we will deal with the spatially open case, i.e., the case where the fiber is a complete non-compact Riemannian manifold. This is due to the fact that some experimental observations and theoretical arguments suggest that spatially open models provide a more accurate description of our current universe [7]. Furthermore, spatially closed universes lead to a violation of the holographic principle, making spatially open spacetimes compatible with a possible theory that unifies gravity and quantum mechanics [13].

Consequently, our aim in this article will be to particularize some of the results obtained in [14] to the maximal case, which will enable us to obtain uniqueness results for the maximal hypersurface equation in spatially open GRW spacetimes. The technique that will be used is based on combining Bochner formula with a generalized maximum principle (see [9, 15] for different ways of using these ideas to obtain parametric uniqueness results).

2. Preliminaries

Let (F, g_F) be an $n(\geq 2)$-dimensional (connected) Riemannian manifold, I an open interval in $\mathbb{R}$ and f a positive smooth function defined on I. Consider now the product manifold $\overline{M} = I \times F$ endowed with the Lorentzian metric

$$\overline{g} = -\pi_I^*(dt^2) + f(\pi_I)^2\, \pi_F^*(g_F), \tag{2.1}$$

where π_I and π_F denote the projections onto I and F, respectively. The Lorentzian manifold $(\overline{M}, \overline{g})$ is a warped product (in the sense of [12, Chap. 7]) with base $(I, -dt^2)$, fiber (F, g_F) and warping function f. Endowing

$(\overline{M}, \overline{g})$ with the time orientation induced by $\partial_t := \partial/\partial t$ we can call it, following the terminology introduced in [1], an $(n+1)$-dimensional Generalized Robertson-Walker (GRW) spacetime.

In any GRW spacetime there is a distinguished timelike and future pointing vector field, $K := f(\pi_I)\partial_t$ that satisfies

$$\overline{\nabla}_X K = f'(\pi_I)\, X \tag{2.2}$$

for any $X \in \mathfrak{X}(\overline{M})$, where $\overline{\nabla}$ is the Levi-Civita connection of the Lorentzian metric (2.1). Thus, K is conformal and its metrically equivalent 1-form is closed.

Given an n-dimensional manifold M, an immersion $\psi : M \to \overline{M}$ is called spacelike if the Lorentzian metric (2.1) induces a Riemannian metric g on M through ψ. In this codimension one case, M is called a spacelike hypersurface. Along this article, we will denote the restriction of π_I along ψ by τ. It can be easily seen that its gradient is given by $\nabla\tau = -\partial_t^\top$, where $\partial_t^\top$ is the tnagential component of ∂_t along ψ. In addition, we also have $\sinh^2\varphi = |\nabla\tau|^2$.

Furthermore, the time-orientation of $\overline{M}$ allows to globally define on each spacelike hypersurface M in $\overline{M}$ a unique unitary timelike vector field $N \in \mathfrak{X}^\perp(M)$ with the same time-orientation as ∂_t .

Denoting by A the shape operator associated to N, the mean curvature function associated to N is $H := -(1/n)\mathrm{trace}(A)$. A spacelike hypersurface with identically zero constant mean curvature is called maximal hypersurface.

Among the family of spacelike hypersurfaces in a GRW spacetime we should highlight the subfamily of spacelike graphs. Given an $n(\geq 2)$-dimensional Riemannian manifold (F, g_F) and a smooth function $f : I \longrightarrow \mathbb{R}^+$ we can consider in the GRW spacetime $\overline{M} = I \times_f F$ the graph

$$\Sigma_u = \{(u(p), p) : p \in \Omega\},$$

where $\Omega \subseteq F$, $u \in C^\infty(\Omega)$ and $u(\Omega) \subseteq I$. The induced metric on Ω from the Lorentzian metric on $\overline{M}$, via the graph Σ_u is given by

$$g_u = -du^2 + f(u)^2 g_F.$$

Note that g_u is positive definite (i.e., Σ_u is spacelike) if and only if u satisfies

$$|Du| < f(u).$$

In this case,

$$N = \frac{1}{f(u)\sqrt{f(u)^2 - |Du|^2}}\left(f(u)^2\partial_t + Du\right)$$

is a future pointing unit normal vector field on Σ_u and when $\Omega = F$ the spacelike graph is said to be entire. From [12, Prop. 7.35] we obtain that the mean curvature function of a spacelike graph associated to N is

$$H = \mathrm{div}\left(\frac{Du}{nf(u)\sqrt{f(u)^2 - |Du|^2}}\right) + \frac{f'(u)}{n\sqrt{f(u)^2 - |Du|^2}}\left(n + \frac{|Du|^2}{f(u)^2}\right), \tag{2.3}$$

where div represents the divergence operator in (F, g_F).

Our aim in this article will be to obtain a uniqueness result for the solutions of the maximal hypersurface equation in a spatially open GRW spacetime. Namely, we are interested in the solutions on (F, g_F) of the following second order nonlinear elliptic PDE:

$$\mathrm{div}\left(\frac{Du}{nf(u)\sqrt{f(u)^2 - |Du|^2}}\right) + \frac{f'(u)}{n\sqrt{f(u)^2 - |Du|^2}}\left(n + \frac{|Du|^2}{f(u)^2}\right) = 0, \tag{E.1}$$

$$|Du| < \lambda f(u), \;\; 0 < \lambda < 1, \tag{E.2}$$

3. Main results

In order to obtain our uniqueness result for equation (E) we will first deal with the parametric version of the problem, considering maximal hypersurfaces in a spatially open GRW spacetime which are not necessarily graphs. To prove our main uniqueness results we will need the following lemma, which bounds the Laplacian of the hyperbolic angle of these hypersurfaces.

Lemma 3.1 *Let $\psi : M \to \overline{M}$ be a complete maximal hypersurface in a GRW spacetime $\overline{M} = I \times_f F$. Then, the hyperbolic angle of M satisfies*

$$\begin{aligned}\frac{1}{2}\Delta \sinh^2\varphi &= \cosh^2\varphi\left(\mathrm{Ric}^F(N^F,N^F) - n(\log f)''(\tau)\sinh^2\varphi\right) + |\mathrm{Hess}(\tau)|^2 \\ &\quad + \frac{f'(\tau)^2}{f(\tau)^2}\sinh^2\varphi(n+\sinh^2\varphi) + |\nabla\cosh\varphi|^2.\end{aligned} \tag{3.1}$$

where Ric^F *denotes the Ricci tensor of the fiber F and N^F is the projection of N on F.*

Proof The crucial idea of this proof is to compute the Laplacian of the function $\cosh\varphi$, which is defined by

$$\cosh\varphi = -\overline{g}(N,\partial_t).$$

Using (2.2) we can compute this function's gradient, obtaining

$$\nabla\cosh\varphi = A\partial_t^\top + \frac{f'(\tau)}{f(\tau)}\cosh\varphi\,\partial_t^\top. \tag{3.2}$$

Choosing a local orthonormal reference frame $\{E_1,\dots,E_n\}$ on TM we can obtain the Laplacian of $\cosh\varphi$ using (3.2) as follows

$$\Delta\cosh\varphi = \sum_{i=1}^n g(\nabla_{E_i}(A\partial_t^\top),E_i) + \sum_{i=1}^n g\left(\nabla_{E_i}\left(\frac{f'(\tau)}{f(\tau)}\cosh\varphi\,\partial_t^\top\right),E_i\right). \tag{3.3}$$

In fact, we can rewrite (3.3) as

$$\begin{aligned}\Delta\cosh\varphi &= \sum_{i=1}^n g((\nabla_{E_i}A)\partial_t^\top,E_i) + \sum_{i=1}^n g(\nabla_{E_i}\partial_t^\top,AE_i) - \frac{f''(\tau)}{f(\tau)}\cosh\varphi\sinh^2\varphi \\ &\quad + 2\frac{f'(\tau)^2}{f(\tau)^2}\cosh\varphi\sinh^2\varphi + \frac{f'(\tau)}{f(\tau)}g(A\partial_t^\top,\partial_t^\top) \\ &\quad + \frac{f'(\tau)}{f(\tau)}\cosh\varphi\sum_{i=1}^n g(\nabla_{E_i}\partial_t^\top,E_i).\end{aligned} \tag{3.4}$$

where we have used that $(\nabla_X A)Y = \nabla_X(AY) - A(\nabla_X Y)$ for all $X,Y\in\mathfrak{X}(M)$. On the other hand, using Codazzi equation $\overline{g}(\overline{\mathrm{R}}(X,Y)N,Z) = \overline{g}((\nabla_Y A)X,Z) - \overline{g}((\nabla_X A)Y,Z)$ (where $\overline{\mathrm{R}}$ denotes the curvature tensor of $\overline{M}$) and choosing our local frame in T_pM satisfying $\left(\nabla_{E_j}E_i\right)_p = 0$ we deduce from (3.4)

$$\begin{aligned}\Delta\cosh\varphi &= -\overline{\mathrm{Ric}}(\partial_t^\top,N) + 2\frac{f'(\tau)}{f(\tau)}g(A\partial_t^\top,\partial_t^\top) + \cosh\varphi\,\mathrm{trace}(A^2) \\ &\quad - \frac{f''(\tau)}{f(\tau)}\cosh\varphi\sinh^2\varphi + 3\frac{f'(\tau)^2}{f(\tau)^2}\cosh\varphi\sinh^2\varphi + n\frac{f'(\tau)^2}{f(\tau)^2}\cosh\varphi,\end{aligned} \tag{3.5}$$

where $\overline{\mathrm{Ric}}$ is the Ricci tensor of $\overline{M}$. Decomposing N as $N = N^F - \overline{g}(N,\partial_t)\partial_t$, being N^F the projection of N on the fiber F, we can use [12, Cor. 7.43] to write

$$\overline{\mathrm{Ric}}(\partial_t^\top,N) = -\cosh\varphi\left(\mathrm{Ric}^F(N^F,N^F) + (n-1)(\log f)''(\tau)\sinh^2\varphi\right). \tag{3.6}$$

Now, (3.6) can be used in (3.5) to obtain

$$\begin{aligned}\Delta\cosh\varphi &= \cosh\varphi\left(\mathrm{Ric}^F(N^F,N^F)-(n-1)(\log f)''(\tau)\sinh^2\varphi\right) \\ &\quad +2\,\frac{f'(\tau)}{f(\tau)}g(A\partial_t^\top,\partial_t^\top)+\cosh\varphi\,\mathrm{trace}(A^2)-\frac{f''(\tau)}{f(\tau)}\cosh\varphi\sinh^2\varphi \\ &\quad +3\,\frac{f'(\tau)^2}{f(\tau)^2}\cosh\varphi\sinh^2\varphi+n\,\frac{f'(\tau)^2}{f(\tau)^2}\cosh\varphi.\end{aligned} \tag{3.7}$$

If we now compute $|\mathrm{Hess}(\tau)|^2$ we have

$$\begin{aligned}|\mathrm{Hess}(\tau)|^2 &= \sum_{i=1}^n g(\nabla_{E_i}\partial_t^\top,\nabla_{E_i}\partial_t^\top)=\frac{f'(\tau)^2}{f(\tau)^2}\left(n-1+\cosh^4\varphi\right) \\ &\quad +\cosh^2\varphi\,\mathrm{trace}(A^2)+2\,\frac{f'(\tau)}{f(\tau)}\cosh\varphi\,g(A\partial_t^\top,\partial_t^\top).\end{aligned} \tag{3.8}$$

Combining (3.7) and (3.8) leads to

$$\begin{aligned}\cosh\varphi\,\Delta\cosh\varphi &= \cosh^2\varphi\left(\mathrm{Ric}^F(N^F,N^F)-n(\log f)''(\tau)\sinh^2\varphi\right) \\ &\quad +|\mathrm{Hess}(\tau)|^2+\frac{f'(\tau)^2}{f(\tau)^2}\sinh^2\varphi(n+\sinh^2\varphi).\end{aligned} \tag{3.9}$$

We conclude the proof noticing that

$$\frac{1}{2}\Delta\sinh^2\varphi=\cosh\varphi\,\Delta\cosh\varphi+|\nabla\cosh\varphi|^2,$$

and using (3.9) to obtain (3.1). □

To prove our main results we also need the following lemma that extends [17, Lemma 3] and gives a bound for the Ricci curvature of constant mean curvature spacelike hypersurfaces in GRW spacetimes.

Lemma 3.2 *Let $\psi: M\to\overline{M}$ be a maximal hypersurface in a GRW spacetime $\overline{M}=I\times_f F$ whose warping function satisfies $(\log f)''(\tau)\le 0$. If either the fiber F has non-negative sectional curvature, M has bounded hyperbolic angle and the sectional curvature of the fiber F is bounded from below, then the Ricci curvature of M is bounded from below.*

Proof Given a point $p\in M$, let us choose a local orthonormal reference frame $\{E_1,\dots,E_n\}$ around p. From the Gauss equation we have that the Ricci curvature of M, Ric, satisfies

$$\mathrm{Ric}(X,X)\ge\sum_{i=1}^n\overline{g}(\overline{\mathrm{R}}(X,E_i)E_i,X)$$

for all $X\in\mathfrak{X}(M)$. Using [12, Prop. 7.42] we obtain

$$\begin{aligned}\sum_{i=1}^n\overline{g}(\overline{\mathrm{R}}(X,E_i)E_i,X) &= \sum_{i=1}^n g_F(\mathrm{R}^{\mathrm{F}}(X^F,E_i^F)E_i^F,X^F)+(n-1)\frac{f'(\tau)^2}{f(\tau)^2}|X|^2 \\ &\quad -(n-2)(\log f)''(\tau)\,g(X,\nabla\tau)^2-(\log f)''(\tau)|\nabla\tau|^2|X|^2,\end{aligned}$$

being R^{F} the curvature tensor of F and X^F and E_i^F the projections of X and E_i on the fiber. If $(\log f)''(\tau)\le 0$ and F has non-negative sectional curvature, we see that the Ricci curvature of M is bounded from below. On the other hand, if φ is bounded and the sectional curvature of F is bounded from below we can consider $X\in\mathfrak{X}(M)$ such that $|X|^2=1$ and decompose it as

$$X=-\overline{g}(X,\partial_t)\partial_t+X^F.$$

Moreover, we can also see that

$$|X^F|^2|E_i^F|^2 = \left(1 + g(X, \nabla\tau)^2\right)\left(1 + g(E_i, \nabla\tau)^2\right),$$

as well as

$$\overline{g}(X^F, E_i^F)^2 = g(X, E_i)^2 + g(X, \nabla\tau)^2 g(E_i, \nabla\tau)^2 + 2g(X, E_i)g(X, \nabla\tau)g(E_i, \nabla\tau).$$

Thus, if the sectional curvature of F is bounded from below by a constant C the above expressions yield

$$\sum_{i=1}^{n} g_F(\mathrm{R}^{\mathrm{F}}(X^F, E_i^F)E_i^F, X^F) \geq C\left(n - 1 + \sinh^2\varphi + (n-2)g(X, \nabla\tau)^2\right). \tag{3.10}$$

Thus, if the hyperbolic angle of M is bounded the classical Schwarz inequality guarantees that the left hand side of (3.10) is bounded from below by a constant. Therefore, we conclude again that if $(\log f)''(\tau) \leq 0$, then the Ricci curvature of M is bounded from below. □

In addition, we will make use of the following consequence of the Omori-Yau maximum principle obtained by Cheng and Yau in [6].

Lemma 3.3 [6, 11] *Let M be a complete Riemannian manifold whose Ricci curvature is bounded from below. If $u \in C^2(M)$ is a non-negative function that satisfies $\Delta u \geq Cu^2$ for a positive constant C, then u vanishes identically on M.*

Taking these three lemmas into account, we are now in a position to prove our main parametric uniqueness result.

Theorem 3.4 *Let $\psi : M \to \overline{M}$ be a complete maximal hypersurface in a GRW spacetime $\overline{M} = I \times_f F$ whose fiber F has non-negative sectional curvature. If the warping function satisfies*

$$\sup_M (\log f)''(\tau) < 0, \tag{A}$$

then M is a totally geodesic spacelike slice.

Proof Under these assumptions, we deduce from Lemma 3.1 that φ satisfies

$$\frac{1}{2}\Delta \sinh^2\varphi \geq -n\,(\log f)''(\tau)(1 + \sinh^2\varphi)\sinh^2\varphi.$$

Moreover, (A) allows us to use Lemma 3.2 to guarantee that the Ricci curvature of M is bounded from below as well as ensures the existence of a positive constant C such that

$$\Delta \sinh^2\varphi \geq C \sinh^4\varphi.$$

Finally, we can use Lemma 3.3 to conclude that M is a totally geodesic spacelike slice. □

Remark 3.5 Note that assumption (A) cannot be omitted in order to obtain these results in spatially open GRW spacetimes. For instance, in the Lorentz-Minkowski spacetime of arbitrary dimension $\mathbb{L}^{n+1}$ this assumption does not hold and there is no analogous uniqueness result for complete maximal spacelike hypersurfaces.

As a consequence of Theorem 3.4 we can obtain our main non parametric result for the maximal hypersurface equation in a GRW spacetime whose fiber has non-negative sectional curvature.

Corollary 3.6 *Let $f : I \longrightarrow \mathbb{R}^+$ be a smooth function such that* $\inf f > 0$ *and* $\sup (\log f)'' < 0$. *Then, the only entire solutions to the equation*

$$\operatorname{div}\left(\frac{Du}{nf(u)\sqrt{f(u)^2 - |Du|^2}}\right) + \frac{f'(u)}{n\sqrt{f(u)^2 - |Du|^2}}\left(n + \frac{|Du|^2}{f(u)^2}\right) = 0, \tag{E.1}$$

$$|Du| < \lambda f(u), \;\; 0 < \lambda < 1, \tag{E.2}$$

on a complete Riemannian manifold F with non-negative sectional curvature are the constant functions $u = t_0$, with $t_0 \in I$ such that $f'(t_0) = 0$.

Proof Note that constraint (E.2) implies that the hyperbolic angle of the graph Σ_u satisfies

$$\cosh\varphi < \frac{1}{\sqrt{1-\lambda^2}}. \tag{3.11}$$

Furthermore, using the classical Schwarz inequality we deduce

$$g_u(v,v) \geq |Du|^2 g_u(v,v) + f(u)^2 g_F(d\pi_F(v), d\pi_F(v)), \text{ for all } v \in T\Sigma_u. \tag{3.12}$$

Hence, from (3.12) we obtain

$$g_u(v,v) \geq \frac{f(u)^2}{\cosh^2\varphi} g_F(d\pi_F(v), d\pi_F(v)). \tag{3.13}$$

Denoting by $L_F(\gamma)$ and $L_u(\gamma)$ the length of a smooth curve γ on F with respect to the metrics g_F and g_u, respectively, from (3.13) and (3.11) we have

$$L_u(\gamma) \geq (1-\lambda^2)(\inf f(u)^2) L_F(\gamma). \tag{3.14}$$

Thus, since (F, g_F) is complete and $\inf f > 0$ we obtain that the metric g_u is also complete. This fact and the rest of our assumptions enable us to apply Theorem 3.4 to end the proof. □

Acknowledgements

The author is partially supported by Spanish MINECO and ERDF project MTM2016-78807-C2-1-P.

References

[1] L.J. Alías, A. Romero and M. Sánchez. Uniqueness of complete spacelike hypersurfaces of constant mean curvature in Generalized Robertson-Walker spacetimes, *Gen. Relat. Gravit.*. **27** (1995), 71–84.

[2] J.M. Arms, J.E. Marsden and V. Moncrief. The structure of the space of solutions of Einstein's equations. II. Several Killing fields and the Einstein-Yang-Mills equations, *Ann. Phys.*, **144** (1982), 81–106.

[3] A. Brasil and A.G. Colares. On constant mean curvature spacelike hypersurfaces in Lorentz manifolds, *Mat. Contemp.*, **17** (1999), 99–136.

[4] D. Brill and F. Flaherty. Isolated maximal surfaces in spacetime, *Commun. Math. Phys.* **50** (1984), 157–165.

[5] E. Calabi. Examples of Bernstein problems for some nonlinear equations, *P. Symp. Pure Math.*, **15** (1970), 223–230.

[6] S.Y. Cheng and S.T. Yau. Maximal spacelike hypersurfaces in the Lorentz-Minkowski spaces, *Ann. of Math.*, **104** (1976), 407–419.

[7] H.Y. Chiu. A cosmological model of the universe, *Ann. Phys.*, **43** (1967), 1–41.

[8] G. Colombo, J.A.S. Pelegrín and M. Rigoli. A note on spacelike hypersurfaces and timelike conformal vectors, *Recent Advances in Pure and Applied Mathematics*, RSME Springer Series, 2020, (135–147).

[9] J.M. Latorre and A. Romero. Uniqueness of noncompact spacelike hypersurfaces of constant mean curvature in Generalized Robertson-Walker spacetimes, *Geometriae Dedicata*, **93** (2002), 1–10.

[10] A. Lichnerowicz. L'integration des équations de la gravitation relativiste et le problème des n corps, *J. Math. Pure Appl.*, **23** (1944), 37–63.

[11] S. Nishikawa. On maximal spacelike hypersurfaces in a Lorentzian manifold, *Nagoya Math. J.*, **95** (1984), 117–124.

[12] B. O'Neill. *Semi-Riemannian Geometry with applications to Relativity*, Academic Press, New York, 1983.

[13] J.A.S. Pelegrín. From holography to the geometry of the spacetime, *Int. J. Geom. Methods M.*, **16** (2019), 1950010 (1–11).

[14] J.A.S. Pelegrín. Calabi-Bernstein type results for complete constant mean curvature spacelike hypersurfaces in spatially open Generalized Robertson-Walker spacetimes, *RACSAM Rev. R. Acad. A*, **114** (2020), 2 (1–13).

[15] J.A.S. Pelegrín and M. Rigoli. Constant mean curvature spacelike hypersurfaces in spatially open GRW spacetimes, *J. Geom. Anal.*, **29** (2019), 3293-3307.

[16] J.A.S. Pelegrín, A. Romero and R.M. Rubio. On maximal hypersurfaces in Lorentz manifolds admitting a parallel lightlike vector field, *Classical Quant. Grav.*, **33** (2016), 055003 (1–8).

[17] J.A.S. Pelegrín, A. Romero and R.M. Rubio. Uniqueness of complete maximal hypersurfaces in spatially open $(n+1)$-dimensional Robertson-Walker spacetimes with flat fiber, *Gen. Relat. Gravit.*, **48** (2016), 1–14.

[18] J.A.S. Pelegrín, A. Romero and R.M. Rubio. An extension of Calabi's correspondence between the solutions of two Bernstein problems to more general elliptic nonlinear equations, *Math. Notes*, **105** (2019), 122–126.

[19] J.A.S. Pelegrín, A. Romero and R.M. Rubio. Spacelike hypersurfaces in spatially parabolic standard static spacetimes and Calabi-Bernstein-type problems, *Mediterr. J. Math.*, **16** (2019), 34 (1–15).

[20] H. Ringström. *The Cauchy problem in General Relativity*, ESI Lectures in Mathematics and Physics, 2009.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 313–320)

Well-balanced algorithms for relativistic fluids on a Schwarzschild background

Ernesto Pimentel-García[1], **Carlos Parés**[1], **Philippe G. LeFloch**[2]
1. *University of Málaga, Spain*
2. *Laboratoire Jacques-Louis Lions, Sorbonne Université, Paris, France*

Abstract

A class of well-balanced finite volume methods with first and higher order of accuracy is designed for two spherical symmetric fluid models on a Schwarzschild curved background: the Burgers-Schwarzschild model and the Euler-Schwarzschild model. We take advantage of the explicit or implicit forms available for the stationary solutions of these models to design numerical methods that preserve them. These methods are then used to investigate the late time behaviour of the flows.

1. Introduction

We are interested in the numerical approximation and the long time behaviour of relativistic compressible fluid flows on a Schwarzschild black hole background. The flow is assumed to enjoy spherical symmetry and therefore we deal with nonlinear hyperbolic systems of partial differential equations (PDEs) in one space variable. The objective is two-fold: on the one hand, designing and testing numerically finite volume algorithms that are well-balanced; on the other hand, to perform a thorough investigation of the behavior of the solutions and numerically infer definite conclusions about the long-time behavior of such flows. Our study should provide first and useful insights for, on the one hand, further development concerning the mathematical analysis of the models and, on the other hand, further investigations to the same problem in higher dimensions without symmetry restriction.

We consider first the *relativistic Burgers-Schwarzschild model* (see [12, 13]):

$$v_t + F(v,r)_r = S(v,r), \qquad t \geq 0, \quad r > 2M, \tag{1.1a}$$

where $v = v(t,r) \in [-1,1]$ is the unknown function and the flux and source terms read

$$F(v,r) = \Big(1 - \frac{2M}{r}\Big)\frac{v^2-1}{2}, \qquad S(v,r) = \frac{2M}{r^2}\,(v^2-1), \tag{1.1b}$$

while the constant $M > 0$ represents the mass of the black hole. The speed of propagation for this scalar balance law reads

$$\partial_v F(v,r) = \Big(1 - \frac{2M}{r}\Big)v, \tag{1.2}$$

which vanishes at the boundary $r = 2M$, so that no boundary condition is required in order to pose the Cauchy problem.

Next, we consider the *relativistic Euler-Schwarzschild model* (as it is called in [12, 13]):

$$V_t + F(V,r)_r = S(V,r), \qquad t \geq 0, \quad r > 2M, \tag{1.3a}$$

whose unknowns are the fluid density $\rho = \rho(t,r) \geq 0$ and the normalized velocity $v = v(t,r) \in (-1,1)$. These functions are defined for all $r > 2M$ and the limiting values $v = \pm 1$ can be reached at the boundary $r = 2M$ only, and

$$V = \begin{pmatrix} V^0 \\ V^1 \end{pmatrix} = \begin{pmatrix} \dfrac{1+k^2v^2}{1-v^2}\rho \\ \dfrac{1+k^2}{1-v^2}\rho v \end{pmatrix}, \qquad F(V,r) = \begin{pmatrix} \Big(1-\dfrac{2M}{r}\Big)\dfrac{1+k^2}{1-v^2}\rho v \\ \Big(1-\dfrac{2M}{r}\Big)\dfrac{v^2+k^2}{1-v^2}\rho \end{pmatrix}, \tag{1.3b}$$

$$S(V,r) = \begin{pmatrix} -\dfrac{2}{r}\Big(1-\dfrac{2M}{r}\Big)\dfrac{1+k^2}{1-v^2}\rho v \\ \dfrac{-2r+5M}{r^2}\dfrac{v^2+k^2}{1-v^2}\rho - \dfrac{M}{r^2}\dfrac{1+k^2v^2}{1-v^2}\rho + 2\dfrac{r-2M}{r^2}k^2\rho \end{pmatrix}, \tag{1.3c}$$

with

$$v = \frac{1+k^2-\sqrt{(1+k^2)^2 - 4k^2\left(\frac{V^1}{V^0}\right)^2}}{2k^2\frac{V^1}{V^0}}, \qquad \rho = \frac{V^1(1-v^2)}{v(1+k^2)}. \tag{1.3d}$$

Here, $k \in (-1, 1)$ denotes the (constant) speed of sound. The eigenvalues of the Jacobian of the flux function are

$$\mu_\pm = \Big(1 - \frac{2M}{r}\Big) \frac{v \pm k}{1 \pm k^2 v}, \tag{1.4}$$

so that the system is strictly hyperbolic. As usual, a state (ρ, v), by definition, is said to be *sonic* if one of the eigenvalues vanishes, i.e. if $|v| = |k|$, *supersonic* if both eigenvalues have the same sign, i.e. if $|v| > |k|$, or *subsonic* if the eigenvalues have different signs, i.e. if $|v| < |k|$. Both eigenvalues $\mu_\pm$ vanish at the boundary $r = 2M$, so that no boundary condition is required in order to pose the Cauchy problem.

In order to be able of running reliable and accurate numerical simulations for these two models, we design shock-capturing, high-order, and well-balanced finite volume methods of first- and second-order of accuracy (and even third-order accurate for (1.1)). Specifically, we extend to the present problem the well-balanced methodology proposed recently by Castro and Parés [7] for nonlinear hyperbolic systems of balance laws. For earlier work on well-balanced schemes we also refer to [5, 16, 17] and, concerning the design of geometry-preserving schemes, we refer for instance to [1–3, 6, 8–10, 15, 19] and the references therein.

The properties of the stationary solutions play a fundamental role in the design of well-balanced schemes, as well as in the study of the long time behavior of solutions. We thus also built here upon earlier investigations by LeFloch and collaborators [11–13] on the theory and approximation of the relativistic Burgers- and Euler-Schwarzschild model (1.1) and (1.3). Remarkably, the stationary solutions to both models are available in explicit or implicit form.

2. Well-balanced methodology

Both problems of interest are of the form

$$V_t + F(V, r)_r = S(V, r), \qquad r > 2M, \tag{2.1}$$

with unknown $V = V(t, r) \in \mathbb{R}^N$ and $N = 1$ or 2. Systems of this form have non-trivial stationary solutions, which satisfy the ODE

$$F(V, r)_r = S(V, r). \tag{2.2}$$

Our goal is to introduce a family of numerical methods that are well-balanced, i.e. that preserve the stationary solutions in a sense to be specified. We follow the strategy in [7] to which we refer for further details and arguments of proof.

We consider semi-discrete finite volume numerical methods of the form

$$\frac{dV_i}{dt} = -\frac{1}{\Delta r}\Big(F_{i+\frac12} - F_{i-\frac12} - \int_{r_{i-\frac12}}^{r_{i+\frac12}} S(\mathbb{P}_i^t(r), r)\, dr\Big), \tag{2.3}$$

where the following notation is used.

- $I_i = [r_{i-\frac12}, r_{i+\frac12}]$ denote the computational cells, whose length Δr is assumed to be constant for simplicity.
- $V_i(t)$ denotes the approximate average of the exact solution in the ith cell at the time t, that is,

$$V_i(t) \cong \frac{1}{\Delta r} \int_{r_{i-\frac12}}^{r_{i+\frac12}} V(r, t)\, dr. \tag{2.4}$$

- $\mathbb{P}_i^t(r)$ denotes the approximation of the solution in the ith cell given by a high-order reconstruction operator based on the cell averages $\{V_i(t)\}$, that is, $\mathbb{P}_i^t(r) = \mathbb{P}_i^t(r; \{V_j(t)\}_{j\in\mathcal{S}_i})$. Here, $\mathcal{S}_i$ denotes the set of cell indices associated with the stencil of the ith cell.
- The flux terms are denoted by $F_{i+\frac12} = \mathbb{F}\Big(V_{i+\frac12}^{t,-}, V_{i+\frac12}^{t,+}, r_{i+\frac12}\Big)$, where $V_{i+\frac12}^{t,\pm}$ are the reconstructed states at the interfaces, i.e.

$$V_{i+\frac12}^{t,-} = \mathbb{P}_i^t(r_{i+\frac12}), \quad V_{i+\frac12}^{t,+} = \mathbb{P}_{i+1}^t(r_{i+\frac12}). \tag{2.5}$$

  Here, $\mathbb{F}$ is a consistent numerical flux, i.e. a continuous function $\mathbb{F} : \mathbb{R}^N \times \mathbb{R}^N \times (2M, +\infty) \to \mathbb{R}^N$ satisfying $\mathbb{F}(V, V, r) = F(V, r)$ for all V, r.

Furthermore, given a stationary solution V^* of (2.2), we use the following terminology.

- The numerical method (2.3) is said to be well-balanced for V^* if the vector of cell averages of V^* is an equilibrium of the ODE system (2.3).

- The reconstruction operator is said to be well-balanced for V^* if we have $\mathbb{P}_i(r) = V^*(r)$ for all $r \in [r_{i-\frac{1}{2}}, r_{i+\frac{1}{2}}]$, where $\mathbb{P}_i$ is the approximation of V^* obtained by applying the reconstruction operator to the vector of cell averages of V^*.

It is easily checked that, if the reconstruction operator is well-balanced for a continuous stationary solution V^* of (2.2) then the numerical method is also well-balanced for V^*. The following strategy to design a well-balanced reconstruction operator $\mathbb{P}_i$ on the basis of a standard operator $\mathbb{Q}_i$ was introduced in [5]:

Given a family of cell values $\{V_i\}$, in every cell $I_i = [r_{i-\frac{1}{2}}, r_{i+\frac{1}{2}}]$ we proceed as follows.

1. Seek, (whenever possible), a stationary solution $V_i^*(x)$ defined in the stencil of cell I_i $(\cup_{j\in\mathcal{S}_i} I_j)$ such that

$$\frac{1}{\Delta r}\int_{r_{i-\frac{1}{2}}}^{r_{i+\frac{1}{2}}} V_i^*(r)\,dr = V_i. \tag{2.6}$$

   If such a solution does not exist, take $V_i^* \equiv 0$.

2. Apply the reconstruction operator to the cell values $\{W_j\}_{j\in\mathcal{S}_i}$ given by

$$W_j = V_j - \frac{1}{\Delta r}\int_{r_{j-\frac{1}{2}}}^{r_{j+\frac{1}{2}}} V_i^*(r)\,dr, \qquad j \in \mathcal{S}_i, \tag{2.7}$$

   in order to obtain $\mathbb{Q}_i(r) = \mathbb{Q}_i(r; \{W_j\}_{j\in\mathcal{S}_i})$. We consider the MUSCL reconstruction operator (see [18]) in the second-order case and the CWENO3 (see [14]) in the third-order case.

3. Define finally

$$\mathbb{P}_i(r) = V_i^*(r) + \mathbb{Q}_i(r). \tag{2.8}$$

It can be then easily shown that the reconstruction operator $\mathbb{P}_i$ in (2.8) is well-balanced for every stationary solution provided that the reconstruction operator $\mathbb{Q}_i$ is exact for the zero function. Moreover, if $\mathbb{Q}_i$ is conservative then $\mathbb{P}_i$ is conservative, in the sense that

$$\frac{1}{\Delta r}\int_{r_{i-\frac{1}{2}}}^{r_{i+\frac{1}{2}}} \mathbb{P}_i(r)\,dr = V_i, \tag{2.9}$$

and $\mathbb{P}_i$ has the same accuracy as $\mathbb{Q}_i$ if the stationary solutions are sufficiently regular.

If a quadrature formula (whose order of accuracy must be greater or equal to the one of the reconstruction operator)

$$\int_{r_{i-\frac{1}{2}}}^{r_{i+\frac{1}{2}}} f(x)\,dx \approx \Delta r \sum_{l=0}^{q} \alpha_l f(r_{i,l})$$

where $\alpha_0, \ldots, \alpha_q, r_{i,0}, \ldots, r_{i,q}$ represent the weights and the nodes of the formula, is used to compute the averages of the initial condition, namely $V_{i,0} = \sum_{l=0}^{q} \alpha_l V_0(r_{i,l})$, the reconstruction procedure has to be modified to preserve the well-balanced property: Steps 1 and 2 have to be replaced by the following ones

1. Seek, if possible, the stationary solution $V_i^*(x)$ defined in the stencil of cell I_i $(\cup_{j\in\mathcal{S}_i} I_j)$ such that

$$\sum_{l=0}^{q} \alpha_l V_i^*(r_{i,l}) = V_i. \tag{2.10}$$

   If this solution does not exist, take $V_i^* \equiv 0$.

2. Apply the reconstruction operator to the cell values $\{W_j\}_{j\in\mathcal{S}_i}$ given by

$$W_j = V_j - \sum_{l=0}^{q} \alpha_l V_i^*(r_{j,l}), \qquad j \in \mathcal{S}_i.$$

For first- or second-order methods, if the midpoint rule is selected to compute the initial averages, i.e. $V_{i,0} = V_0(r_i)$, then at the first step of the reconstruction procedure, the problem (2.10) reduces to finding the stationary solution satisfying

$$V_i^*(r_i) = V_i. \tag{2.11}$$

The well-balanced property of the method can be lost if the quadrature formula is used to compute the integral appearing at the right-hand side of (2.3). In order to circumvent this difficulty, in [7] it is proposed to rewrite the methods as follows:

$$\begin{aligned}\frac{dV_i}{dt} = &-\frac{1}{\Delta r}\left(F_{i+\frac{1}{2}} - F\left(V_i^{t,*}(r_{i+\frac{1}{2}}), r_{i+\frac{1}{2}}\right) - F_{i-\frac{1}{2}} + F\left(V_i^{t,*}(r_{i-\frac{1}{2}}), r_{i-\frac{1}{2}}\right)\right)\\ &+ \frac{1}{\Delta r}\int_{r_{i-\frac{1}{2}}}^{r_{i+\frac{1}{2}}} \left(S(\mathbb{P}_i^t(r), r) - S(V_i^{t,*}(r), r)\right)\, dr,\end{aligned} \tag{2.12}$$

where $V_i^{t,*}$ is the function selected in Step 1 for the ith cell at time t. In this equivalent form, a quadrature formula can be applied to the integral without losing the well-balanced property, and this leads to a numerical method of the form:

$$\begin{aligned}\frac{dV_i}{dt} = &-\frac{1}{\Delta r}\left(F_{i+\frac{1}{2}} - F\left(V_i^{t,*}(r_{i+\frac{1}{2}}), r_{i+\frac{1}{2}}\right) - F_{i-\frac{1}{2}} + F\left(V_i^{t,*}(r_{i-\frac{1}{2}}), r_{i-\frac{1}{2}}\right)\right)\\ &+ \sum_{l=0}^{q} \alpha_l \left(S(\mathbb{P}_i^t(r_{i,l}), r_{i,l}) - S(V_i^{t,*}(r_{i,l}), r_{i,l})\right).\end{aligned} \tag{2.13}$$

First-order well-balanced methods are obtained by selecting the trivial constant piecewise reconstruction operator as the standard one, i.e.

$$\mathbb{Q}_i(r, V_i) = V_i, \qquad r \in [r_{i-\frac{1}{2}}, r_{i+\frac{1}{2}}]. \tag{2.14}$$

It can be easily checked that the numerical method then reduces to

$$\frac{dV_i}{dt} = -\frac{1}{\Delta r}\left(F_{i+\frac{1}{2}} - F\left(V_i^{t,*}(r_{i+\frac{1}{2}}), r_{i+\frac{1}{2}}\right) - F_{i-\frac{1}{2}} + F\left(V_i^{t,*}(r_{i-\frac{1}{2}}), r_{i-\frac{1}{2}}\right)\right), \tag{2.15}$$

where $F_{i+\frac{1}{2}} = \mathbb{F}\left(V_i^*(r_{i+\frac{1}{2}}), V_{i+1}^*(r_{i+\frac{1}{2}}), r_{i+\frac{1}{2}}\right)$.

Notice that the implementation of these methods requires to find a stationary solution with prescribed average at Step 1 of the reconstruction procedure. In the case of the Burgers-Schwarzschild model, the explicit expression of the stationary solutions is available

$$v^*(r) = \pm\sqrt{1 - K^2\left(1 - \frac{2M}{r}\right)}, \quad K > 0. \tag{2.16}$$

and it can be easily checked that (2.10) and (2.11) have always a unique solution. In the case of the Euler-Schwarzschild model, the following implicit form of the stationary solutions is available

$$\frac{\mathrm{sgn}(v)(1-v^2)|v|^{\frac{2k^2}{1-k^2}} r^{\frac{4k^2}{1-k^2}}}{\left(1 - \frac{2M}{r}\right)} = C_1, \qquad r(r-2M)\rho\frac{v}{1-v^2} = C_2, \tag{2.17}$$

where C_1, C_2 are constants. Once the constants are fixed by imposing (2.11), a nonlinear system has to be solved to evaluate the stationary solution at a point of the stencil. This system can have 0, 1, or 2 solutions. If there is no solution the standard reconstruction is used. When there are two solutions, one of them is supersonic and the other is subsonic: the one whose regime is equal to that of V_i is selected.

3. Numerical tests

First-, second- and third-order methods for the Burgers-Schwarzschild and first- and second-order methods for Euler-Schwarzschild have been implemented. Several numerical test are presented here to show the relevance of the well-balanced property for the investigation of the asymptotic behaviour of the flows.

3.1. Burgers-Schwarzschild

We consider the spatial interval $[2M, L]$ with $M = 1$ and $L = 4$, a 256-point uniform mesh, and the CFL number equal to 0.5. At $r = 2M$, $F_{-\frac{1}{2}} = 0$ is imposed. At $r = L$, a transmissive boundary condition based on the use ghost-cells is used. The following numerical flux is considered:

$$F_{i+\frac{1}{2}} = \mathbb{F}(v_i, v_{i+1}, r_{i+\frac{1}{2}}) = \left(1 - \frac{2M}{r_{i+\frac{1}{2}}}\right)\frac{q^2(0; v_i, v_{i+1}) - 1}{2},$$

where $q(\cdot; v_L, v_R)$ is the self-similar solution of the Riemann problem for the standard Burgers equation with the initial condition

$$v_0(r) = \begin{cases} v_L, & r < 0,\\ v_R, & r > 0.\end{cases}$$

In order to check the relevance of the well-balanced property, the well-balanced methods will be compared with standard ones based on the same numerical fluxes and the standard first-, second-, or third-order reconstructions.

Positive stationary solution We consider the initial condition

$$v_0(r) = \sqrt{\frac{3}{4} + \frac{1}{2r}} \tag{3.1}$$

corresponding to a positive stationary solution. Table 1 shows the error in L^1 norm between the initial condition and the numerical solution at time $t = 50$. Figure 1 compares the numerical solutions obtained with the well-balanced and the non-well-balanced methods: it can be seen how the latter are unable to capture the stationary solution. After a time that decreases with the order, the numerical solutions depart from the steady state.

Scheme (256 cells)	Error (1st)	Error (2nd)	Error (3rd)
Well-balanced	1.13E-14	8.72Ee-17	7.22E-14
Non well-balanced	1.89	1.61	8.78E-02

Tab. 1 Well-balanced versus non-well-balanced schemes: L^1 errors at $t = 50$ for the Burgers model with the initial condition (3.1).

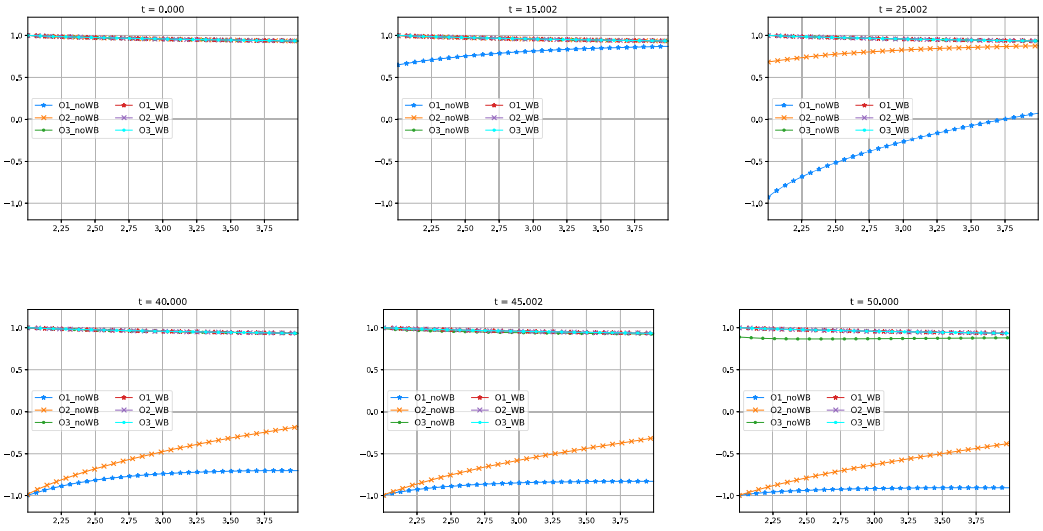


Fig. 1 Burgers-Schwarzschild model with the initial condition (3.1): first-, second-, and third-order well-balanced and not-well-balanced methods at various times.

Perturbation of a steady shock solution In this test case we consider the initial condition:

$$\tilde{v}_0(r) = v_0(r) + p_L(r), \tag{3.2}$$

where v_0 is the steady shock solution given by

$$v_0(r) = \begin{cases} \sqrt{\dfrac{3}{4} + \dfrac{1}{2r}}, & 2 < r < 3, \\ -\sqrt{\dfrac{3}{4} + \dfrac{1}{2r}}, & \text{otherwise}, \end{cases} \tag{3.3}$$

and

$$p_L(r) = \begin{cases} -\dfrac{1}{5} e^{-200(r-2.5)^2}, & 2.2 < r < 2.8, \\ 0, & \text{otherwise}. \end{cases} \tag{3.4}$$

The first-, second-, and third-order well-balanced methods have been applied to this problem. In Figure 2 it can be observed that, after the wave generated by the initial perturbation leaves the computational domain, the stationary solution (3.3) is not recovered: a different stationary solution is obtained whose shock is placed at a different location. Observe that all the three methods capture the same stationary solution.

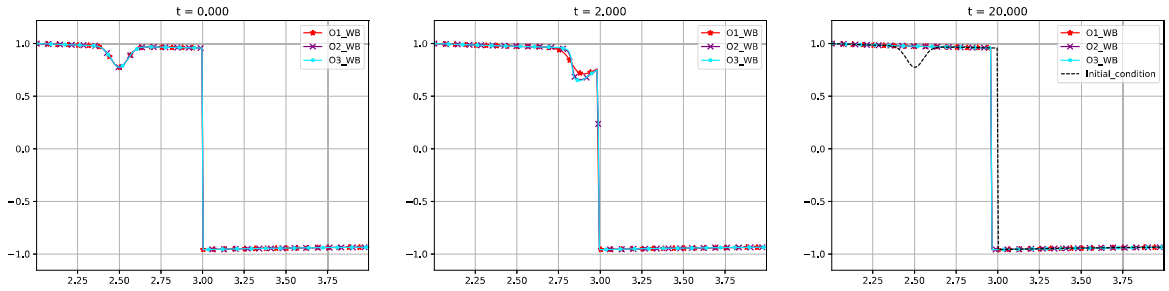


Fig. 2 Burgers-Schwarzschild model with the initial condition (3.2)-(3.3)-(3.4): first-, second-, and third-order well-balanced methods at selected times.

3.2. Euler-Schwarzschild

We consider the spatial interval $[2M, L]$ with $M = 1$ and $L = 10$, a 500-point uniform mesh, $k = 0.3$, and the CFL number equal to 0.5. At $r = 2M$ we impose $F_{-\frac{1}{2}} = 0$ as boundary condition since $\left(1 - \frac{2M}{r}\right) = 0$. The boundary conditions are the same as in the previous test case. A HLL-like numerical flux in PVM form (see [4]) will be used:

$$F_{i+\frac{1}{2}} = \frac{1}{2}(F(V_i) + F(V_{i+1})) - \frac{1}{2}\left(\alpha_0(V_{i+1} - V_i) + \alpha_1(F(V_{i+1}) - F(V_i))\right), \tag{3.5}$$

with

$$\alpha_0 = \frac{\overline{\lambda_2}|\overline{\lambda_1}| - \overline{\lambda_1}|\overline{\lambda_2}|}{\overline{\lambda_2} - \overline{\lambda_1}}, \qquad \alpha_1 = \frac{|\overline{\lambda_2}| - |\overline{\lambda_1}|}{\overline{\lambda_2} - \overline{\lambda_1}}, \tag{3.6}$$

where $\overline{\lambda_1}$ and $\overline{\lambda_2}$ are the eigenvalues of some intermediate matrix $J_{i+\frac{1}{2}}$ of the form

$$J_{i+\frac{1}{2}} = \left(1 - \frac{2M}{r_{i+\frac{1}{2}}}\right)\begin{bmatrix} 0 & 1 \\ \dfrac{k^2 - v_m^2}{1 - k^2 v_m^2} & \dfrac{2(1-k^2)v_m}{1 - k^2 v_m^2} \end{bmatrix} \tag{3.7}$$

where v_m is some intermediate value between v_i^n and v_{i+1}^n.

Discontinuous stationary entropy weak solution We consider the initial condition

$$V_0(r) = \begin{cases} V_-^*(r), & r \le 6, \\ V_+^*(r), & \text{otherwise}, \end{cases} \tag{3.8}$$

where $V_-^*(r)$ is the supersonic stationary solution such that

$$\rho_-^*(6) = 4, \quad v_-^*(6) = 0.6 \tag{3.9}$$

and $V_+^*(r)$ is the subsonic one such that

$$\rho_+^*(6) = \frac{\rho_-^*(6)(v_-^*(6)^2 - k^4)}{k^2(1 - v_-^*(6)^2)}, \quad v_+^*(6) = \frac{k^2}{v_-^*(6)}. \tag{3.10}$$

V_0 is an entropy weak stationary solution of the system: see [12,13]. Table 2 shows the error in L^1 norm between the numerical solution at time $t = 50$ and Figure 3 shows the comparison of the numerical results obtained with well-balanced and non-well-balanced methods at selected times. The numerical results of this section put on evidence, as for the Burgers-Schwarzschild system, the relevance of using well-balanced methods for the Euler-Schwarzschild model.

Relation between the perturbation and the displacement of the shock In order to study the relationship between the amplitude of the perturbation and the distance between the initial and the final shock locations, we consider the family of initial conditions:

$$\widetilde{V}_0(r) = \widetilde{V}^*(r) + \delta(\alpha, r), \tag{3.11}$$

Scheme (500 cells)	Error v (1st)	Error ρ (1st)	Error v (2nd)	Error ρ (2nd)
Well-balanced	2.20E-13	1.25E-11	1.92E-13	1.03E-11
Non well-balance	0.89	3.94	0.89	3.92

Tab. 2 Well-balanced versus non-well-balanced schemes: L^1 errors at time $t = 50$ for the Burgers-Schwarzschild model with the initial condition (3.8)

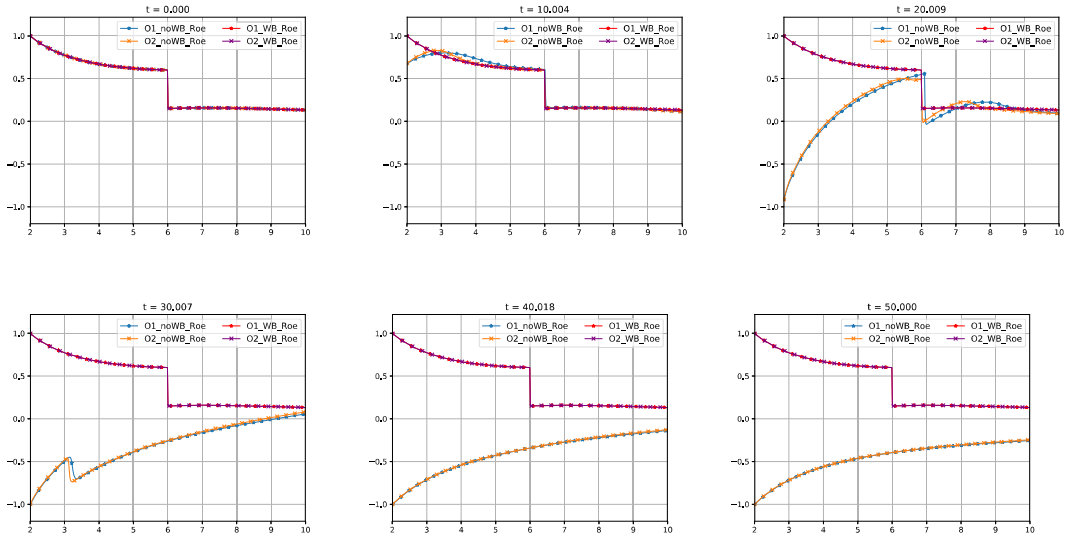
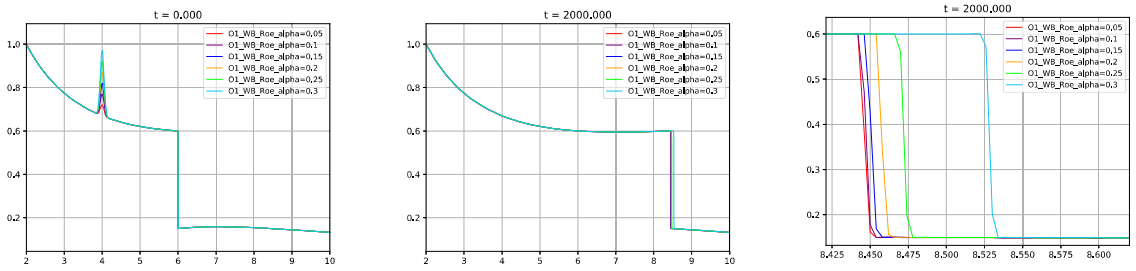


Fig. 3 Euler-Schwarzschild model with the initial condition (3.8): first- and second-order well-balanced and non-well-balanced methods at selected times for the variable v.

where $\widetilde{V}^*$ is the steady shock solution given by (3.8)-(3.10) and

$$\delta(\alpha, r) = [\delta_v(\alpha, r), \delta_\rho(\alpha, r)]^T = \begin{cases} [\alpha e^{-200(r-4)^2}, 0]^T, & 3 < r < 5, \\ [0, 0]^T, & \text{otherwise}, \end{cases} \tag{3.12}$$

with $\alpha > 0$. In this case we will also use a 2000-point uniform mesh. Figure 4 shows the numerical solution for different values of α and we observe that depending on the amplitude of the perturbation the numerical solutions converge in time to different steady shock solutions.



(a) Zoom.

Fig. 4 Euler-Schwarzschild model with the initial condition (3.11): first-order well-balanced method taking different values of α for variable v.

4. Conclusions

The procedure introduced in [5] and recalled in [7] is extended to the relativistic fluid flows in the Schwarzschild background. More precisely, we develop first and higher order well-balanced schemes for the relativistic Burgers and Euler systems. Several numerical tests are used to validate the schemes and to highlight the relevance of the well-balanced property when dealing with these relativistic flows. We also use these schemes to perform a

systematic numerical study of these two PDE systems in order to be able to extract general conclusions about the long time behavior of the flow. Such a study is expected to be a useful tool to direct the mathematical analysis of the models and the study of more complex relativistic models.

Acknowledgements

The authors were supported by an Innovative Training Network (ITN) under the grant 642768 "ModCompShock". The research of CP and EPG was also partially supported by the Spanish Government(SG), the European Regional Development Fund (ERDF), the Regional Government of Andalusia (RGA), and the University of Málaga (UMA) through the projects RTI2018-096064-B-C21 (SG-ERDF), UMA18-Federja-161 (RGA-ERDF-UMA), and P18-RT-3163 (RGA-ERDF).

References

[1] A. Beljadid, P.G. LeFloch, and M. Mohamadian, Late-time asymptotic behavior of solutions to hyperbolic conservation laws on the sphere, *Computer Methods in Applied Mechanics and Engineering*, 349:285–311 2019.

[2] F. Bouchut, Nonlinear stability of finite Volume Methods for hyperbolic conservation laws: And Well-Balanced schemes for sources, *Springer Science & Business Media*, 2004.

[3] A. Beljadid and P.G. LeFloch, A central-upwind geometry-preserving method for hyperbolic conservation laws on the sphere, Commun. Appl. Math. Comput. Sci. 12 (2017), 81–107.

[4] M.J. Castro and E. Fernández-Nieto, A class of computationally fast first order finite volume solvers: PVM methods. *SIAM Journal on Scientific Computing*, 34(4):A2173–A2196, 2012.

[5] M.J. Castro, J.M Gallardo, J.A López-García, and C. Parés, Well-balanced high order extensions of Godunov method for semilinear balance laws. *SIAM Journal on Numerical Analysis*, 46(2):1012–1039, 2008.

[6] M.J. Castro, T. Morales de Luna and C. Parés, Well-balanced schemes and path-conservative numerical methods. In *Handbook of Numerical Analysis*, 18:131–175, 2017.

[7] M.J. Castro and C. Parés, Well-balanced high-order finite volume methods for systems of balance laws. *Journal of Scientific Computing*, 82(2):1–48, 2020.

[8] S. Dong and P.G. LeFloch, Convergence of the finite volume method on a Schwarzschild background. *ESAIM: Mathematical Modelling and Numerical Analysis*, 53(5):1459–1476, 2019.

[9] D. Dziuk, D. Kröner, and T. Müller, Scalar conservation laws on moving hypersurfaces. *Interfaces and Free Boundaries*, 15(2):203–236, 2013.

[10] J. Giesselman and P.G. LeFloch, Formulation and convergence of the finite volume method for conservation laws on spacetimes with boundary. *Numerische Mathematik*, 144:751–785, 2020.

[11] P.G. LeFloch and H. Makhlof, A geometry-preserving finite volume method for compressible fluids on Schwarzschild spacetime. *Communications in Computational Physics*, 15(3):827–852, 2014.

[12] P.G. LeFloch and S. Xiang, A numerical study of the relativistic Burgers and Euler equations on a Schwarzschild black hole exterior. *Applied Mathematics and Computational Science*, 13(2):271–301, 2018.

[13] P.G. LeFloch and S. Xiang, Weakly regular fluid flows with bounded variation on the domain of outer communication of a Schwarzschild black hole spacetime. II, J. Math. Pure Appl. 122 (2019), 272–317.

[14] D. Levy, G. Puppo, and G. Russo. Compact central WENO schemes for multidimensional conservation laws. *SIAM Journal on Scientific Computing*, 22(2):656–672, 2000.

[15] A. Rossmanith, D.S. Bale, and R.J. LeVeque, A wave propagation algorithm for hyperbolic systems on curved manifolds. *Journal of Computational Physics*, 199(2):631–662, 2004.

[16] G. Russo, Central schemes for conservation laws with application to shallow water equations. *Trends and Applications of Mathematics to Mechanics*, Springer, p. 225–246, 2005.

[17] G. Russo, High-order shock-capturing schemes for balance laws.*Numerical solutions of partial differential equations, Advanced Courses in Mathematics — CRM Barcelona Centre de Recerca Matemàtica, Birkhäuser, Basel*, p. 59–147, 2009.

[18] B. Van Leer. Towards the ultimate conservative difference scheme. ii. monotonicity and conservation combined in a second-order scheme. *Journal of computational physics*, 14(4):361–370, 1974.

[19] M.-E. Vázquez-Cendón, Improved treatment of source terms in upwind schemes for the shallow water equations in channels with irregular geometry. *Journal of computational physics*, 148(2): 497–526, 1999.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 321–324)

Asymptotic analysis of the behavior of a viscous fluid between two very close mobile surfaces

José M. Rodríguez [1], **Raquel Taboada-Vázquez**[1]
Universidade da Coruña, Departamento de Matemáticas, Spain

Abstract

The aim of this work is the study of the behavior of an incompressible viscous fluid moving between two closely spaced surfaces, also in motion. To carry out this work we use the asymptotic expansion method that allows us to formally justify two different models starting from the same initial problem: a lubrication model and a shallow water model. The type of model that yields depends on whether the fluid is"pressure dominated" and on the boundary conditions imposed. We discuss in detail under what conditions each of the models would be applicable.

1. Introduction

In this work, we are interested, in a first step, in justifying using the asymptotic development technique, a lubrication model in a thin domain with curved mean surface.

The asymptotic analysis method is a mathematical tool that has been widely used to obtain and justify reduced models, both in solid and fluid mechanics, when one or two of the dimensions of the domain are much smaller than the others. In particular, in fluid mechanics, these technique has been applied to derive lubrication models, shallow water models, tube flow models, etc. (see, for example, [1]- [9], [11]- [12], and many others). Here, we follow the steps of [2], but changing the starting point.

We consider a three-dimensional thin domain, Ω_t^ε, filled by a fluid, that varies with time $t \in [0,T]$, given by

$$\begin{aligned}\Omega_t^\varepsilon \quad = \quad & \left\{(x_1^\varepsilon, x_2^\varepsilon, x_3^\varepsilon) \in \mathbb{R}^3 : x_i(\xi_1,\xi_2,t) \le x_i^\varepsilon \le x_i(\xi_1,\xi_2,t) + h^\varepsilon(\xi_1,\xi_2,t)N_i(\xi_1,\xi_2,t),\right. \\ & \left.(i=1,2,3),\ (\xi_1,\xi_2) \in D \subset \mathbb{R}^2\right\} \end{aligned} \tag{1.1}$$

where $\vec{X}_t(\xi_1,\xi_2) = \vec{X}(\xi_1,\xi_2,t) = (x_1(\xi_1,\xi_2,t), x_2(\xi_1,\xi_2,t), x_3(\xi_1,\xi_2,t))$ is the lower bound surface parametrization, $h^\varepsilon(\xi_1,\xi_2,t)$ is the gap between the two surfaces in motion, and $\vec{N}(\xi_1,\xi_2,t)$ is the unit normal vector:

$$\vec{N}(\xi_1,\xi_2,t) = \frac{\dfrac{\partial \vec{X}}{\partial \xi_1} \times \dfrac{\partial \vec{X}}{\partial \xi_2}}{\left\| \dfrac{\partial \vec{X}}{\partial \xi_1} \times \dfrac{\partial \vec{X}}{\partial \xi_2} \right\|} \tag{1.2}$$

The lower bound surface is assumed to be regular and the gap is assumed to be small with regard to the dimension of the bound surfaces. We take into account that the fluid film between the surfaces is thin by introducing a small non-dimensional parameter ε, and setting that

$$h^\varepsilon(\xi_1,\xi_2,t) = \varepsilon h(\xi_1,\xi_2,t) \tag{1.3}$$

where

$$h(\xi_1,\xi_2,t) \ge h_0 > 0, \quad \forall\, (\xi_1,\xi_2) \in D \subset \mathbb{R}^2,\ \forall\, t \in [0,T]. \tag{1.4}$$

Let us suppose that the fluid motion is governed by Navier-Stokes equations since we consider that it is an incompressible newtonian fluid,

$$\rho_0 \left(\frac{\partial u_i^\varepsilon}{\partial t^\varepsilon} + \frac{\partial u_i^\varepsilon}{\partial x_j^\varepsilon} u_j^\varepsilon \right) = -\frac{\partial p^\varepsilon}{\partial x_i^\varepsilon} + \mu \left(\frac{\partial^2 u_i^\varepsilon}{\partial (x_1^\varepsilon)^2} + \frac{\partial^2 u_i^\varepsilon}{\partial (x_2^\varepsilon)^2} + \frac{\partial^2 u_i^\varepsilon}{\partial (x_3^\varepsilon)^2} \right) + \rho_0 f_i^\varepsilon, \quad (i=1,2,3) \tag{1.5}$$

$$\frac{\partial u_j^\varepsilon}{\partial x_j^\varepsilon} = 0 \tag{1.6}$$

where repeated indices indicate summation (j takes values from 1 to 3), ρ_0 is the fluid density, $\vec{u}^\varepsilon$ is the velocity, p^ε is the pressure and $\vec{f}^\varepsilon$ is the density of applied volume forces.

Now, we consider a reference domain independent of ε and t

$$\Omega = D \times [0,1] \tag{1.7}$$

related to Ω_t^ε by the following change of variable:

$$t^\varepsilon = t \tag{1.8}$$
$$x_i^\varepsilon = x_i(\xi_1,\xi_2,t) + \varepsilon\xi_3 h(\xi_1,\xi_2,t)N_i(\xi_1,\xi_2,t) \tag{1.9}$$

where $(\xi_1,\xi_2) \in D$ and $\xi_3 \in [0,1]$, and we make a change of basis to a new basis $\{\vec{a}_1,\vec{a}_2,\vec{a}_3\}$, where

$$\vec{a}_1(\xi_1,\xi_2,t) = \frac{\partial \vec{X}(\xi_1,\xi_2,t)}{\partial \xi_1} \tag{1.10}$$
$$\vec{a}_2(\xi_1,\xi_2,t) = \frac{\partial \vec{X}(\xi_1,\xi_2,t)}{\partial \xi_2} \tag{1.11}$$
$$\vec{a}_3(\xi_1,\xi_2,t) = \vec{N}(\xi_1,\xi_2,t) \tag{1.12}$$

The details about the change of variable and basis can be found in [13].

The velocity and the applied forces $(\vec{u}^\varepsilon, \vec{f}^\varepsilon)$ are written in the new basis (1.10)-(1.12) as follows:

$$\vec{u}^\varepsilon = u_i^\varepsilon \vec{e}_i = u_k(\varepsilon)\vec{a}_k \tag{1.13}$$
$$\vec{f}^\varepsilon = f_i^\varepsilon \vec{e}_i = f_k(\varepsilon)\vec{a}_k \tag{1.14}$$

so

$$u_i^\varepsilon = (u_k(\varepsilon)\vec{a}_k)\cdot\vec{e}_i = u_k(\varepsilon)a_{ki} \tag{1.15}$$
$$f_i^\varepsilon = (f_k(\varepsilon)\vec{a}_k)\cdot\vec{e}_i = f_k(\varepsilon)a_{ki} \tag{1.16}$$

where $a_{ki} = \vec{a}_k\cdot\vec{e}_i$, and we assume that the velocity, the pressure and the applied forces can be developed in powers of ε as in [2], [1], [6], [11] and [12]:

$$u_i(\varepsilon) = u_i^0 + \varepsilon u_i^1 + \varepsilon^2 u_i^2 + \cdots \quad (i=1,2,3) \tag{1.17}$$
$$p(\varepsilon) = \varepsilon^{-2}p^{-2} + \varepsilon^{-1}p^{-1} + p^0 + \varepsilon p^1 + \varepsilon^2 p^2 + \cdots \tag{1.18}$$
$$f_i(\varepsilon) = f_i^0 + \varepsilon f_i^1 + \varepsilon^2 f_i^2 + \cdots \quad (i=1,2,3) \tag{1.19}$$

Taking into account (1.15)-(1.16), equations (1.5)-(1.6) yield ($i=1,2,3$):

$$\rho_0\left(\frac{\partial(u_k(\varepsilon)a_{ki})}{\partial t^\varepsilon} + \frac{\partial(u_k(\varepsilon)a_{ki})}{\partial x_j^\varepsilon}(u_k(\varepsilon)a_{kj})\right) = -\frac{\partial p(\varepsilon)}{\partial x_i^\varepsilon}$$
$$+\mu\left(\frac{\partial^2(u_k(\varepsilon)a_{ki})}{\partial(x_1^\varepsilon)^2} + \frac{\partial^2(u_k(\varepsilon)a_{ki})}{\partial(x_2^\varepsilon)^2} + \frac{\partial^2(u_k(\varepsilon)a_{ki})}{\partial(x_3^\varepsilon)^2}\right) + \rho_0 f_k(\varepsilon)a_{ki} \tag{1.20}$$
$$\frac{\partial(u_k(\varepsilon)a_{kj})}{\partial x_j^\varepsilon} = 0 \tag{1.21}$$

Next, we substitute developments (1.17)-(1.19) in Navier-Stokes equations written in the reference domain ((1.20)-(1.21)) and we identify the terms multiplied by the same power of ε. In this way we obtain a series of equations that will allow us to determine the terms of the previous developments.

In the next two sections we summarize the results obtained in [13].

2. A new generalized lubrication model

If we assume that the fluid slips at the lower surface ($\xi_3 = 0$), and at the upper surface ($\xi_3 = 1$), but there is continuity in the normal direction, so the tangential velocities at the lower and upper surfaces are known, and the normal velocity of each of them must match the fluid velocity, we derive the following generalized lubrication equation:

$$\frac{1}{\sqrt{A^0}}\mathrm{div}\left(\frac{h^3}{\sqrt{A^0}}M\nabla p^{-2}\right) = 12\mu\frac{\partial h}{\partial t} + 12\mu\frac{hA^1}{A^0}\left(\frac{\partial\vec{X}}{\partial t}\cdot\vec{N}\right)$$
$$-6\mu\nabla h\cdot(\vec{W}^0 - \vec{V}^0) + \frac{6\mu h}{\sqrt{A^0}}\mathrm{div}(\sqrt{A^0}(\vec{W}^0 + \vec{V}^0)) \tag{2.1}$$

where the pressure is approximated by $\varepsilon^{-2}p^{-2}$

$$A^0 = EG - F^2 \tag{2.2}$$
$$A^1 = -eG - gE + 2fF \tag{2.3}$$
$$M = \begin{pmatrix} G & -F \\ -F & E \end{pmatrix} \tag{2.4}$$

and $E, F, G,\ e, f, g$ are the coefficients of the first and second fundamental forms, respectively, of the surface parametrized by $\vec{X}$. $\vec{V}^0 = (V_1^0, V_2^0)$ and $\vec{W}^0 = (W_1^0, W_2^0)$ are the approximations of order 0 on ε of the tangential velocity at the lower and upper surfaces respectively.

Once p^{-2} is calculated we have the following approximation of the three components of the velocity:

$$u_1^0 = \frac{h^2(\xi_3^2 - \xi_3)}{2\mu A^0}\left(G\frac{\partial p^{-2}}{\partial \xi_1} - F\frac{\partial p^{-2}}{\partial \xi_2}\right) + \xi_3(W_1^0 - V_1^0) + V_1^0 \tag{2.5}$$
$$u_2^0 = \frac{h^2(\xi_3^2 - \xi_3)}{2\mu A^0}\left(E\frac{\partial p^{-2}}{\partial \xi_2} - F\frac{\partial p^{-2}}{\partial \xi_1}\right) + \xi_3(W_2^0 - V_2^0) + V_2^0 \tag{2.6}$$
$$u_3^0 = \frac{\partial \vec{X}}{\partial t} \cdot \vec{N} \tag{2.7}$$

If we consider the classic assumptions to derive Reynolds equations (domain independent of time, $x_3 = 0$ in (1.1), upper surface fixed, lower surface moving in the x_1-direction with constant velocity), we re-obtain the classic Reynolds equation (see [10]) from (2.1).

3. A new thin fluid layer model

During this process we have observed that, depending on the boundary conditions, other models can be obtained. In this section, we change the boundary conditions that we imposed in the first case: instead of assuming that we know the velocities on the upper and lower boundaries of the domain, we assume that we know the tractions on these upper and lower boundaries. In particular, we assume that the normal component of the traction on $\xi_3 = 0$ and on $\xi_3 = 1$ are known pressures (π_0^ε and π_1^ε), and that the tangential component of the traction on these surfaces are friction forces depending on the value of the velocities on ∂D.

Under these assumptions we derive a shallow water model that allow us to determine h, V_1^0 and V_2^0:

$$\frac{\partial h}{\partial t} + \frac{h}{\sqrt{A^0}}\mathrm{div}\left(\sqrt{A^0}\vec{V}^0\right) + \frac{hA^1}{A^0}\left(\frac{\partial \vec{X}}{\partial t} \cdot \vec{N}\right) = 0 \tag{3.1}$$

$$\begin{aligned}\frac{\partial V_i^0}{\partial t} &+ \sum_{l=1}^{2}\left(V_l^0 - C_l^0\right)\frac{\partial V_i^0}{\partial \xi_l} + \sum_{k=1}^{2}\left(R_{ik}^0 + \sum_{l=1}^{2} H_{ilk}^0 V_l^0\right)V_k^0 = -\frac{1}{\rho_0}\left(\alpha_i^0\frac{\partial \pi_0^0}{\partial \xi_1} + \beta_i^0\frac{\partial \pi_0^0}{\partial \xi_2}\right) \\ &+ \nu\left\{\sum_{m=1}^{2}\sum_{l=1}^{2}\frac{\partial^2 V_i^0}{\partial \xi_m \partial \xi_l}J_{lm}^0 + \sum_{k=1}^{2}\sum_{l=1}^{2}\frac{\partial V_k^0}{\partial \xi_l}(L_{kli}^0 + \psi(h)_{ikl}^0)\right. \\ &\left. + \sum_{k=1}^{2} V_k^0(S_{ik}^0 + \chi(h)_{ik}^0) + \hat{\kappa}(h)_i^0\right\} + F_i^0(h) - Q_{i3}^0\left(\frac{\partial \vec{X}}{\partial t} \cdot \vec{N}\right) \quad (i = 1, 2)\end{aligned} \tag{3.2}$$

where the coefficients $\alpha_i^0, \beta_i^0, C_l^0, H_{ilk}^0, J_{lm}^0, L_{kli}^0, Q_{i3}^0, R_{ik}^0, S_{ik}^0$ depend only on the lower bound surface parametrization, $\vec{X}$ while the coefficients $F_i^0(h), \psi(h)_{ikl}^0, \chi(h)_{ik}^0), \hat{\kappa}(h)_i^0$ depend both on the parametrization and on the gap h. The detailed definition of these coefficients is given in [13].

Let π_0^0 be the approximation of order 0 on ε of the pressure π_0^ε. Then, we obtain the following approximations of the velocity and the pressure:

$$u_i^0 = W_i^0 = V_i^0 \quad i = 1, 2 \tag{3.3}$$
$$u_3^0 = \frac{\partial \vec{X}}{\partial t} \cdot \vec{N} \tag{3.4}$$
$$p^0 = \frac{2\mu}{h}\frac{\partial h}{\partial t} + \pi_0^0 \tag{3.5}$$

4. Conclusions

Thus, two new models that can not be found in the literature, as far as we know, are presented here. Both models have been derived starting from the same initial problem, an incompressible viscous fluid moving between two closely spaced surfaces.

The method used to justify them allows us to answer the question of when each of them is applicable. We reach the conclusion that the magnitude of the pressure differences at the lateral boundary of the domain is key when deciding which of the two models best describes the fluid behavior.

Boundary conditions tell us which of the two models should be used when simulating the flow of a thin fluid layer between two surfaces: if the fluid pressure is dominant (that is, it is of order $O(\varepsilon^{-2})$), and the fluid velocity is known on the upper and lower surfaces, we must use the lubrication model; if the fluid pressure is not dominant (that is, it is of order $O(1)$), and the tractions are known on the upper and lower surfaces, we must use the shallow water model. In the first case we will say that the fluid is "driven by the pressure" and in the second that it is "driven by the velocity".

Acknowledgements

This work has been partially supported by Ministerio de Economía y Competitividad of Spain, under grant MTM2016-78718-P with the participation of FEDER, and the European Union's Horizon 2020 Research and Innovation Programme, under the Marie Sklodowska-Curie Grant Agreement No 823731 CONMECH.

References

[1] A. Assemien, G. Bayada, M. Chambat. Inertial effects in the asymptotic behavior of a thin film flow. *Asymptotic Analysis*, 9(3): 177–208, 1994. https://doi.org/10.3233/ASY-1994-9301.

[2] G. Bayada, M. Chambat. The Transition Between the Stokes Equations and the Reynolds Equation: A Mathematical Proo. *Appl. Math. Optim.*, 14: 73–93, 1986. https://doi.org/10.1007/BF01442229.

[3] D. Bresch, P. Noble. Mathematical justification of a shallow water model. *Methods and Applications of Analysis*, 14(2): 87-118, 2007. https://dx.doi.org/10.4310/MAA.2007.v14.n2.a1.

[4] G. Castiñeira, J. M. Rodríguez. Asymptotic Analysis of a Viscous Fluid in a Curved Pipe with Elastic Walls. F. Ortegón Gallego, M. Redondo Neble, J. Rodríguez Galván (eds), *Trends in Differential Equations and Applications*, SEMA SIMAI Springer Series 8, Springer, Cham.: 73-87, 2016. https://doi.org/10.1007/978-3-319-32013-7_5.

[5] G. Castiñeira, E. Marušić-Paloka, I. Pažanin, J. M. Rodríguez. Rigorous justification of the asymptotic model describing a curved-pipe flow in a time-dependent domain. *Z Angew Math Mech.*, 99(1): 99:e201800154, 2019. https://doi.org/10.1002/zamm.201800154.

[6] G. Cimatti. A rigorous justification of the Reynolds equation. *Quart. Appl. Math.*, 45: 627–644, 1987. https://doi.org/10.1090/qam/917014.

[7] H. Dridi. Comportement asymptotique des équations de Navier-Stokes dans des domaines "aplatis". *Bull. Sc. Math.*, 106: 369-385, 1982. https://zbmath.org/?q=an:0512.35015.

[8] I. Moise, R. Temam, M. Ziane. Asymptotic analysis of the Navier-Stokes equations in thin domains. *Topol. Methods Nonlinear Anal.*, 10(2): 249-282, 1997. https://projecteuclid.org/euclid.tmna/1476842206.

[9] S. A. Nazarov. Asymptotic solution of the Navier-Stokes problem on the flow of a thin layer of fluid. *Sib Math J*, 31(2): 296-307, 1990. https://doi.org/10.1007/BF00970660.

[10] O. Reynolds. On the theory of lubrication and its application to Mr Beauchamp tower's experiments. *Phil. Trans. Roy Soc. London*, 117: 157–234, 1886 https://www.jstor.org/stable/109480.

[11] J. M. Rodríguez, R. Taboada-Vázquez. Bidimensional shallow water model with polynomial dependence on depth through vorticity. *Journal of Mathematical Analysis and Applications*, 359(2): 556-569, 2009. https://doi.org/10.1016/j.jmaa.2009.06.003.

[12] J. M. Rodríguez, R. Taboada-Vázquez. Derivation of a new asymptotic viscous shallow water model with dependence on depth. *Applied Mathematics and Computation*, 219(7): 3292-3307, 2012. https://doi.org/10.1016/j.amc.2011.08.053.

[13] J. M. Rodríguez, R. Taboada-Vázquez. Asymptotic analysis of a thin fluid layer flow between two moving surfaces. *arXiv:2101.07862* (https://arxiv.org/abs/2101.07862), 2021, sent to Journal of Mathematical Analysis and Applications.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 325–331)

Convergence rates for Galerkin approximation for magnetohydrodynamic type equations

María Ángeles Rodríguez-Bellido[1], **Marko Antonio Rojas-Medar**[2], **Alex Sepúlveda-Cerda**[3]
1. angeles@us.es *Universidad de Sevilla, Spain*
2. mmedar@academicos.uta.cl *Universidad de Tarapacá, Chile*
3. alex.sepulveda@ufrontera.cl *Universidad de La Frontera, Chile*

Abstract

The motion of incompressible electrical conducting fluids can be modeled by magnetohydrodynamics equations, which consider the Navier-Stokes equations coupled with Maxwell's equations. For the classical Navier-Stokes system, there exists an extensively study of the convergence rate for the Galerkin approximations. Here, we extend the estimates rates of spectral Galerkin approximations for the magnetohydrodynamic equations. We prove optimal error estimates in the $L^2(\Omega)$ and $H^1(\Omega)$-norms, we obtain a result similar to the Rautmann for the $H^2(\Omega)$-norm, and we reach basically the same level of knowledge as in the case of the classical Navier-Stokes.

1. Introduction

The motion of incompressible electrical conducting fluids can be modeled by the so-called equations of magnetohydrodynamics, which can be described as the coupling of the Navier-Stokes equations and the Maxwell's equations. To describe these equations, we consider a bounded domain $\Omega \subset \mathbb{R}^3$, $T > 0$, denoted $Q_T \equiv \Omega \times (0,T)$ and $S_T \equiv \partial\Omega \times (0,T)$. In the case where there is free motion of heavy ions, not directly due to the electric field (see [11], [19], [20]), these equations can be reduced to the form:

$$\left\{\begin{array}{rcll} \dfrac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u} - \dfrac{\eta}{\rho_m}\Delta\mathbf{u} + \dfrac{1}{\rho_m}\nabla\left(p^* + \dfrac{\mu}{2}\mathbf{h}^2\right) - \dfrac{\mu}{\rho_m}\mathbf{h}\cdot\nabla\mathbf{h} &=& \mathbf{f}, & \text{in } Q_T, \\ \dfrac{\partial \mathbf{h}}{\partial t} - \dfrac{1}{\mu\sigma}\Delta\mathbf{h} + (\mathbf{u}\cdot\nabla)\mathbf{h} - (\mathbf{h}\cdot\nabla)\mathbf{u} + \nabla q &=& 0, & \text{in } Q_T, \\ \operatorname{div}\mathbf{u} = \operatorname{div}\mathbf{h} &=& 0, & \text{in } Q_T, \end{array}\right. \tag{1.1}$$

together with the following boundary and initial conditions:

$$\left\{\begin{array}{rclrcll} \mathbf{u} &=& \mathbf{0}, & \mathbf{h} &=& \mathbf{0} & \text{on } S_T, \\ \mathbf{u}(x,0) &=& \mathbf{u}_0(x), & \mathbf{h}(x,0) &=& \mathbf{h}_0(x) & \text{in } \Omega, \end{array}\right. \tag{1.2}$$

Here, $\mathbf{u}$ and $\mathbf{h}$ are unknown velocity and magnetic field, respectively, p^* is an unknown hydrostatic pressure, q is an unknown function related to the heavy ions (in such way that the density of electric current, $\mathbf{j}_0$, generated by this motion satisfies the relation $\operatorname{rot}\mathbf{j}_0 = -\sigma\nabla q$), ρ_m is the density of mass of the fluid (assumed to be a positive constant), $\mu > 0$ is a constant magnetic permeability of the medium, $\sigma > 0$ is a constant electric conductivity, $\eta > 0$ is a constant viscosity of the fluid and $\mathbf{f}$ is a given external force field.

There are an extensive literature on the magnetohydrodynamic system (1.1)–(1.2): Lassner [9], by using the semigroup results of Kato and Fujita [7], proved the existence and uniqueness of strong solutions, local in time for any data and global in time for small data. Boldrini and Rojas-Medar [3] studied the existence of weak solutions and the reproductive property using the Galerkin method. The same authors improved this result to local and global strong solutions by using the spectral Galerkin method in [4, 5]. Damázio and Rojas-Medar [6] studied the regularity of weak solutions, and Notte-Cuello and Rojas-Medar [10] used an iterative approach to show the existence and uniqueness of the strong solutions. The initial value problem in time dependent domains was studied by Rojas-Medar and Beltrán-Barrios in [17], and by Berselli and Ferreira in [1]. The problem in unbounded domains with boundary uniformly of C^3-class was studied by Zhao in [22].

On the other hand, for the classical Navier–Stokes system there exists an extensively study of the convergence rate for the Galerkin approximations. The first work in this way was given by Rautmann in [12], where he proved the optimal convergence in the $H^1(\Omega)$-norm, but the optimal convergence in the $L^2(\Omega)$-norm was left as an open problem in [12] and was answered by Salvi in [18] (see also [2]). Applying the same method and assuming the uniform boundedness in time of the $L^2(\Omega)$-norm of the gradient of the velocity and the exponential stability in

the $H^1(\Omega)$-norm of the solution, Heywood [8] was able to derive optimal uniform in time error estimates for the velocity in the $H^1(\Omega)$-norm. Also, without explicitly assuming $H^1(\Omega)$-exponential stability, Boldrini and Rojas–Medar [2] proved optimal uniform in time error estimates for the spectral Galerkin approximations in the $H^1(\Omega)$ and $L^2(\Omega)$-norms, assuming that the external force field has a mild form of decay.

The study of the convergence rate in the $\mathbf{H}^2(\Omega)$-norm is difficult, because estimates of higher order derivates spatial of solution are required, which needs a compatibility condition to be satisfied by the initial value of the solution. The work of Rautmann [15] give an answer to the question *"how smooth a Navier–Stokes solution can be at time $t = 0$ without any compatibility condition"*. Making use of this result, Rautmann [13], [14] proved the convergence rate in the $\mathbf{H}^2(\Omega)$-norm of the spectral Galerkin approximation to the solution without any compatibility condition.

The aim of this work is to extend the estimates rates of spectral Galerkin approximations for the the Navier-Stokes system to the magnetohydrodynamic equations (1.1)-(1.2). We prove optimal error estimates in the $L^2(\Omega)$ and $H^1(\Omega)$-norms and obtain a result similar to the Rautmann in [13], [14] for $H^2(\Omega)$-norms. In this way, we reach the same level of knowledge as in the case of the classical Navier-Stokes equations. The complete proofs of all the results contained in this manuscript can be consulted in [16].

2. Function Spaces and framework

Throughout this paper we will use the following notation: Vector functions will be written in bold letters. The H^m norm is denoted by $\|\cdot\|_m$. Here $H^m = W^{m,2}(\Omega)$ $(m > 0)$ are the usual Sobolev spaces. H_0^1 denotes the closure of $C_0^\infty(\Omega)$ in the H^1–norm. Let

$$\mathbf{C}^\infty_{0,\sigma}(\Omega) := \{\mathbf{v} \in (C_0^\infty(\Omega))^3 : \operatorname{div} \mathbf{v} = 0 \quad \text{in} \quad \Omega\}, \quad \mathbf{V} = \left\{\text{closure of } \mathbf{C}^\infty_{0,\sigma}(\Omega) \text{ in } \mathbf{H}^1_0(\Omega)\right\},$$

$$\mathbf{H} = \left\{\text{closure of } \mathbf{C}^\infty_{0,\sigma}(\Omega) \text{ in } \mathbf{L}^2(\Omega)\right\} \quad \text{and} \quad \mathbf{V}^* = \{\text{topological dual of } \mathbf{V}\}.$$

In order to give an operator interpretation of problem (1.1)-(1.2), we shall introduce the well known Helmholtz and Weyl decomposition. The Hilbert space $\mathbf{L}^2(\Omega)$ admits the Helmholtz and Weyl decomposition (cf. [21]):

$$\mathbf{L}^2 = \mathbf{H} \oplus \mathbf{H}^\perp,$$

where $\oplus$ denotes direct sum and $\mathbf{H}^\perp = \{\nabla\pi \,:\, \pi \in H^{1,}(\Omega)\}$. Let P be the orthogonal projection from $\mathbf{L}^2(\Omega)$ onto $\mathbf{H}$. Then the operator $A : \mathbf{H} \to \mathbf{H}$ given by $A = -P\Delta$ with domain $D(A) = \mathbf{V} \cap \mathbf{H}^2(\Omega)$ is called the Stokes operator. It is well known that A is a positive self-adjoint operator and is characterized by the following relation:

$$(A\mathbf{w}, \mathbf{v}) = (\nabla\mathbf{w}, \nabla\mathbf{v}) \ \text{ for all } \mathbf{w} \in D(A),\ \mathbf{v} \in \mathbf{V}.$$

From now on, we also denote the inner product in $\mathbf{H}$ by the $\mathbf{L}^2(\Omega)$–inner product $(\cdot,\cdot)$. The general $L^p(\Omega)$-norm will be denoted by $\|\cdot\|_{L^p(\Omega)}$; to make easier the notation, in the case $p = 2$ we simply denote the L^2–norm by $\|\cdot\|$. We shall denote by $\mathbf{w}^k(x)$ and λ_k the eigenfunctions and the eigenvalues of the Stokes operator. It is well known (see [21]) that $\mathbf{w}^k(x)$ are orthogonal in the inner products $(\cdot,\cdot)$, $(\nabla\cdot,\nabla\cdot)$ and $(A\cdot, A\cdot)$ and complete in the spaces $\mathbf{H}$, $\mathbf{V}$ and $\mathbf{V} \cap \mathbf{H}^2(\Omega)$, respectively. For each $k \in \mathbb{N}$, we denote by P_k the orthogonal projection from $\mathbf{L}^2(\Omega)$ onto $\mathbf{V}_k = \operatorname{span}[\mathbf{w}^1(x), \ldots, \mathbf{w}^k(x)]$.

Throughout this work, we will deal with the following notion of strong solution for (1.1)-(1.2).

Definition 2.1 Let $\mathbf{u}_0, \mathbf{h}_0 \in \mathbf{V}$ and $\mathbf{f} \in L^2(0,T;\mathbf{L}^2(\Omega))$. By a strong solution of the problem (1.1)–(1.2), we mean a pair of vector-valued functions $(\mathbf{u}, \mathbf{h})$ such that $\mathbf{u}, \mathbf{h} \in L^\infty(0,T;\mathbf{V}) \cap \mathbf{L}^2(0,T;D(A))$ and that satisfies (1.1)–(1.2).

As a first step to set up and prove the main results of this work, and using the properties of the operator P, we can reformulate the problem (1.1)–(1.2), as follows: find $\mathbf{u}, \mathbf{h}$ in suitable spaces, satisfying:

$$\begin{cases} (\mathbf{u}_t, \mathbf{v}) + (\nabla\mathbf{u}, \nabla\mathbf{v}) + ((\mathbf{u}\cdot\nabla)\mathbf{u}, \mathbf{v}) - ((\mathbf{h}\cdot\nabla)\mathbf{h}, \mathbf{v}) = (\mathbf{f}, \mathbf{v}), & \forall\, \mathbf{v} \in \mathbf{V}, \\ (\mathbf{h}_t, \mathbf{z}) + (\nabla\mathbf{h}, \nabla\mathbf{z}) + ((\mathbf{u}\cdot\nabla)\mathbf{h}, \mathbf{z}) - ((\mathbf{h}\cdot\nabla)\mathbf{u}, \mathbf{z}) = 0, & \forall \mathbf{z} \in \mathbf{V}, \\ \mathbf{u}(x,0) = \mathbf{u}_0(x), & x \in \Omega, \\ \mathbf{h}(x,0) = \mathbf{h}_0(x), & x \in \Omega. \end{cases} \tag{2.1}$$

Observe that, because we do not focus on the dependence of the error on the η, μ, σ or ρ_m, then we consider them all equal to 1.

In order to establish the results concerning estimates for spectral Galerkin approximation, we need to fix some problems. The *spectral Galerkin approximations* for $(\mathbf{u},\mathbf{h})$ are defined for each $k \in \mathbb{N}$ as the solution $(\mathbf{u}^k,\mathbf{h}^k) \in C^2([0,T];\mathbf{V}_k)\times C^2([0,T];\mathbf{V}_k)$ of:

$$\left\{\begin{array}{rcll} (\mathbf{u}_t^k,\mathbf{v})+(\nabla\mathbf{u}^k,\nabla\mathbf{v})+((\mathbf{u}^k\cdot\nabla)\mathbf{u}^k,\mathbf{v})-((\mathbf{h}^k\cdot\nabla)\mathbf{h}^k,\mathbf{v}) &=& (\mathbf{f},\mathbf{v}), & \forall\,\mathbf{v}\in\mathbf{V}_k,\\ (\mathbf{h}_t^k,\mathbf{z})+(\nabla\mathbf{h}^k,\nabla\mathbf{z})+((\mathbf{u}^k\cdot\nabla)\mathbf{h}^k,\mathbf{z})-((\mathbf{h}^k\cdot\nabla)\mathbf{u}^k,\mathbf{z}) &=& 0, & \forall\mathbf{z}\in\mathbf{V}_k,\\ \mathbf{u}(x,0) &=& P_k\mathbf{u}_0(x), & x\in\Omega,\\ \mathbf{h}(x,0) &=& P_k\mathbf{h}_0(x), & x\in\Omega. \end{array}\right. \tag{2.2}$$

Recall that the eigenfunctions expansion of $\mathbf{u}$ and $\mathbf{h}$ can be written, respectively, as:

$$\mathbf{u}(x,t)=\sum_{i=1}^{\infty}a_i(t)\mathbf{w}^i(x) \quad\text{and}\quad \mathbf{h}(x,t)=\sum_{i=1}^{\infty}c_i(t)\mathbf{w}^i(x), \tag{2.3}$$

where $\mathbf{w}^i$ are the eigenfunctions of the Stokes operator. The partial sums of the series for $\mathbf{u}$ and $\mathbf{h}$ will also appear in our study, whose expression are given, respectively, by:

$$\mathbf{v}^k(t)=P_k\mathbf{u}(t)=\sum_{i=1}^{k}a_i(t)\mathbf{w}^i(x) \quad\text{and}\quad \mathbf{b}^k(t)=P_k\mathbf{h}(t)=\sum_{i=1}^{k}c_i(t)\mathbf{w}^i(x). \tag{2.4}$$

3. Known results

By using the spectral Galerkin approximations (2.2), Rojas-Medar and Boldrini ([4], [5]) proved the following results:

Theorem 3.1 *Assume the following condition for the initial data* $\mathbf{u}_0$, $\mathbf{h}_0$, *and the external force* $\mathbf{f}$ *of* (1.1)-(1.2):

$$\mathbf{u}_0,\,\mathbf{h}_0\in\mathbf{V},\quad \mathbf{f}\in L^2(0,T;\mathbf{L}^2(\Omega)) \tag{3.1}$$

Then, on a (*possibly small*) *time interval* $[0,T_1]$, $0<T_1\le T$, *problem* (1.1)-(1.2) *has a unique strong solution* $(\mathbf{u},\mathbf{h})$. *This solution belongs* $C([0,T_1];\mathbf{V})\times C([0,T_1];\mathbf{V})$. *Moreover, there exist* C^1*-functions* $F(t)$ *and* $G(t)$ *such that for any* $t\in[0,T_1]$, *there hold*:

$$\begin{aligned} \|\nabla\mathbf{u}(t)\|^2+\|\nabla\mathbf{h}(t)\|^2+\int_0^t(\|A\mathbf{u}(s)\|^2+\|A\mathbf{h}(s)\|^2)ds &\le F(t),\\ \int_0^t(\|\mathbf{u}_t(s)\|^2+\|\mathbf{h}_t(s)\|^2)ds &\le G(t). \end{aligned}$$

Moreover, the same kind of estimates holds uniformly in $n\in\mathbb{N}$ *for the Galerkin approximations* $(\mathbf{u}^n,\mathbf{h}^n)$.

Theorem 3.2 *Assume* (3.1) *and*

$$\mathbf{u}_0,\,\mathbf{h}_0\in D(A),\quad \mathbf{f}_t\in L^2(0,T;\mathbf{L}^2(\Omega)). \tag{3.2}$$

Then:

$$\begin{aligned} \|\mathbf{u}_t(t)\|^2+\|\mathbf{h}_t(t)\|^2+\int_0^t(\|\nabla\mathbf{u}_t(s)\|^2+\|\nabla\mathbf{h}_t(s)\|^2)ds &\le H_0(t),\\ \|A\mathbf{u}(t)\|^2+\|A\mathbf{h}(t)\|^2 &\le H_1(t),\\ \int_0^t(\|\mathbf{u}_{tt}(s)\|_{V^\star}^2+\|\mathbf{h}_{tt}(s)\|_{V^\star}^2)ds &\le H_2(t), \end{aligned}$$

for any $t\in[0,T_1]$, *where* $H_i(t)$, $i=0,1,2$ *are continuous functions* $t\in[0,T_1]$. *Therefore*:

$$\mathbf{u}(t),\mathbf{h}(t)\in C^1([0,T_1];\mathbf{V})\cap C([0,T_1];D(A)).$$

Moreover, the same kind of estimates holds uniformly in n for the Galerkin approximations $(\mathbf{u}^n,\mathbf{h}^n)$.

Referring to the Navier–Stokes equations, the following lemma can be found in the Rautmann's paper [12].

Lemma 3.3
If $\mathbf{u} \in \mathbf{V}$*, then there holds*:

$$\|\mathbf{u} - P_k\mathbf{u}\|^2 \leq \frac{1}{\lambda_{k+1}}\|\nabla\mathbf{u}\|^2.$$

Also, if $\mathbf{u} \in \mathbf{V} \cap \mathbf{H}^2(\Omega)$*, we have*:

$$\|\mathbf{u} - P_k\mathbf{u}\|^2 \leq \frac{1}{\lambda_{k+1}^2}\|A\mathbf{u}\|^2, \qquad \|\nabla\mathbf{u} - \nabla P_k\mathbf{u}\|^2 \leq \frac{1}{\lambda_{k+1}}\|A\mathbf{u}\|^2.$$

Some of the classical Sobolev interpolation inequalities, considered in this manuscript, can be found in the following result:

Lemma 3.4 *The following estimates are true*:

- $\|\mathbf{v}\|_{\mathbf{L}^\infty(\Omega)} \leq C\|A\mathbf{v}\|, \quad \forall \mathbf{v} \in \mathbf{V} \cap \mathbf{H}^2(\Omega)$,
- $\|\mathbf{v}\|_{\mathbf{L}^6(\Omega)} \leq C\|\nabla\mathbf{v}\|, \quad \forall \mathbf{v} \in \mathbf{V}$,
- $\|\mathbf{v}\|_{\mathbf{L}^3(\Omega)} \leq C\|\mathbf{v}\|^{1/2}\|\nabla\mathbf{v}\|^{1/2}, \quad \forall \mathbf{v} \in \mathbf{V}$
- $\|\mathbf{v}\|_{\mathbf{L}^4(\Omega)} \leq C\|\mathbf{v}\|^{1/4}\|\nabla\mathbf{v}\|^{3/4}, \quad \forall \mathbf{v} \in \mathbf{V}$.

4. Estimates for the solution in $\mathbf{H}^1(\Omega)$

Our first result on error estimates read as follows:

Theorem 4.1 *Assume hypothesis* (3.1) *for the data. Then, the approximations* $(\mathbf{u}^k, \mathbf{h}^k)$ *satisfy*:

$$\|\mathbf{u}(t) - \mathbf{u}^k(t)\|^2 + \|\mathbf{h}(t) - \mathbf{h}^k(t)\|^2 + \int_0^t (\|\nabla\mathbf{u}(s) - \nabla\mathbf{u}^k(s)\|^2 + \|\nabla\mathbf{h}(s) - \nabla\mathbf{h}^k(s)\|^2)\, ds \leq \frac{C}{\lambda_{k+1}}.$$

In addition, if we assume that (3.2)*, then the approximations* $(\mathbf{u}^k, \mathbf{h}^k)$ *satisfy*:

$$\|\mathbf{u}(t) - \mathbf{u}^k(t)\|^2 + \|\mathbf{h}(t) - \mathbf{h}^k(t)\|^2 \leq \frac{C}{\lambda_{k+1}^2}.$$

Theorem 4.2 *If in addition to* (3.1) *we assume* (3.2)*, then we have that there exists a constant* $C > 0$ *such that*:

$$\|\nabla\mathbf{u}(t) - \nabla\mathbf{u}^k(t)\|^2 + \|\nabla\mathbf{h}(t) - \nabla\mathbf{h}^k(t)\|^2 \leq \frac{C}{\lambda_{k+1}}.$$

Corollary 4.3 *Under the hypothesis of Theorem* 4.2*, there exists a positive constant* $C > 0$ *such that*:

$$\int_0^t (\|\mathbf{u}_t(s) - \mathbf{u}_t^k(s)\|^2 + \|\mathbf{h}_t(s) - \mathbf{h}_t^k(s)\|^2)\, ds \leq \frac{C}{\lambda_{k+1}}$$

and, if $\mathbf{f} \in L^2(0,T;\mathbf{H}^1(\Omega)$ *then*:

$$\int_0^t (\|A\mathbf{u}(s) - A\mathbf{u}^k(s)\|^2 + \|A\mathbf{h}(s) - A\mathbf{h}^k(s)\|^2)\, ds \leq \frac{C}{\lambda_{k+1}}.$$

Note that these estimates are made in the Sobolev spaces related to the strong regularity of the solution (see Definition 2.1).

In the search of a proof for these theorems (and corollary), we have to use some preliminary results whose proof needs to define the following auxiliary variables and problems:
Using (2.3) and (2.4), we define:

$$\begin{aligned} \mathbf{e}^k(t) &= \mathbf{u}(t) - \mathbf{v}^k(t), \qquad \tilde{\mathbf{e}}^k(t) = \mathbf{h}(t) - \mathbf{b}^k(t), \\ \mathbf{E}^k(t) &= \mathbf{v}^k(t) - \mathbf{u}^k(t), \quad \tilde{\mathbf{E}}^k(t) = \mathbf{b}^k(t) - \mathbf{h}^k(t), \end{aligned} \tag{4.1}$$

where $\mathbf{u}^k$ and $\mathbf{h}^k$ are the k^{th} Galerkin approximations of $\mathbf{u}$ and $\mathbf{h}$ solutions of (2.2), respectively. One of our aim is to "mesure" the distance between the solutions of (2.1) and (2.2), that we split as:

$$\mathbf{u}(t) - \mathbf{u}^k(t) = \mathbf{e}^k(t) + \mathbf{E}^k(t), \quad \text{and} \quad \mathbf{h}(t) - \mathbf{h}^k(t) = \tilde{\mathbf{e}}^k(t) + \tilde{\mathbf{E}}^k(t). \tag{4.2}$$

These variables satisfy the following problem:

$$\begin{cases} (\mathbf{E}_t^k, \mathbf{v}) + (\nabla \mathbf{E}^k, \nabla \mathbf{v}) + ((\mathbf{e}^k \cdot \nabla)\mathbf{u}, \mathbf{v}) + ((\mathbf{E}^k \cdot \nabla)\mathbf{u}, \mathbf{v}) + ((\mathbf{u}^k \cdot \nabla)\mathbf{e}^k, \mathbf{v}) + ((\mathbf{u}^k \cdot \nabla)\mathbf{E}^k, \mathbf{v}) \\ -((\tilde{\mathbf{e}}^k \cdot \nabla)\mathbf{h}, \mathbf{v}) - ((\tilde{\mathbf{E}}^k \cdot \nabla)\mathbf{h}, \mathbf{v}) - ((\mathbf{h}^k \cdot \nabla)\tilde{\mathbf{e}}^k, \mathbf{v}) - ((\mathbf{h}^k \cdot \nabla)\tilde{\mathbf{E}}^k, \mathbf{v}) = 0, \qquad \forall\, \mathbf{v} \in \mathbf{V}_k, \\ (\tilde{\mathbf{E}}_t^k, \mathbf{z}) + (\nabla \tilde{\mathbf{E}}^k, \nabla \mathbf{z}) + ((\mathbf{e}^k \cdot \nabla)\mathbf{h}, \mathbf{z}) + ((\mathbf{E}^k \cdot \nabla)\mathbf{h}, \mathbf{z}) + ((\mathbf{u}^k \cdot \nabla)\tilde{\mathbf{e}}^k, \mathbf{z}) + ((\mathbf{u}^k \cdot \nabla)\tilde{\mathbf{E}}^k, \mathbf{z}) \\ -((\tilde{\mathbf{e}}^k \cdot \nabla)\mathbf{u}, \mathbf{z}) - ((\tilde{\mathbf{E}}^k \cdot \nabla)\mathbf{u}, \mathbf{z}) - ((\mathbf{h}^k \cdot \nabla)\mathbf{e}^k, \mathbf{z}) - ((\mathbf{h}^k \cdot \nabla)\mathbf{E}^k, \mathbf{z}) = 0, \qquad \forall\, \mathbf{z} \in \mathbf{V}_k, \\ \mathbf{E}^k(x, 0) = \tilde{\mathbf{E}}^k(x, 0) = 0, \qquad x \in \Omega. \end{cases}$$

Using adequate estimates (see [16] for more details), the following results can be proved:

Lemma 4.4 *Assume hypothesis* (3.1) *for the data. Then*:

$$\|\mathbf{E}^k(t)\|^2 + \|\tilde{\mathbf{E}}^k(t)\|^2 \leq \frac{C}{\lambda_{k+1}}.$$

In addition, if we assume (3.2)*, then*:

$$\|\mathbf{E}^k(t)\|^2 + \|\tilde{\mathbf{E}}^k(t)\|^2 \leq \frac{C}{\lambda_{k+1}^2}.$$

Corollary 4.5 *Assume hypothesis* (3.1) *for the data. Then*:

$$\int_0^t (\|\nabla \mathbf{E}^k(s)\|^2 + \|\nabla \tilde{\mathbf{E}}^k(s)\|^2 ds \leq \frac{C}{\lambda_{k+1}}.$$

In addition, if we assume (3.2)*, then*:

$$\int_0^t (\|\nabla \mathbf{E}^k(s)\|^2 + \|\nabla \tilde{\mathbf{E}}^k(s)\|^2) ds \leq \frac{C}{\lambda_{k+1}^2}.$$

Lemma 4.6 *Assuming* (3.1) *and* (3.2) *for the data, we have that there exists a constant* $C > 0$ *such that*:

$$\|\nabla \mathbf{E}^k(t)\|^2 + \|\nabla \tilde{\mathbf{E}}^k(t)\|^2 \qquad \leq \frac{C}{\lambda_{k+1}}.$$

Corollary 4.7 *Under the hypotheses of Lemma* 4.6*, here exists a positive constant* $C > 0$ *such that*:

$$\int_0^t (\|\mathbf{E}_t^k(s)\|^2 + \|\tilde{\mathbf{E}}_t^k(s)\|^2) ds \leq \frac{C}{\lambda_{k+1}}.$$

5. $H^2(\Omega)$-error estimates for the velocity and the magnetic field

The objective of this section is to state and sketch the estimates in the $\mathbf{H}^2(\Omega)$-norm for the solutions of (1.1)-(1.2) that we have obtained. Concretely, our result reads as follows:

Theorem 5.1 *Assume* (3.1)-(3.2)*. If moreover* $\mathbf{f} \in C([0,T], \mathbf{H}^1(\Omega))$ *and* $\mathbf{u}_0, \mathbf{h}_0 \in D(A^{1+\epsilon})$*, with* $\epsilon \in (0, \frac{1}{4})$*, then*

$$\begin{aligned} \|A\mathbf{u}(t) - A\mathbf{u}^k(t)\| + \|\mathbf{u}_t(t) - \mathbf{u}_t^k(t)\| &\leq C\left[\frac{C(\alpha+\epsilon)}{\lambda_{k+1}^{\epsilon}} + \frac{1}{\lambda_{k+1}}\right], \\ \|A\mathbf{h}(t) - A\mathbf{h}^k(t)\| + \|\mathbf{h}_t(t) - \mathbf{h}_t^k(t)\| &\leq C\left[\frac{C(\alpha+\epsilon)}{\lambda_{k+1}^{\epsilon}} + \frac{1}{\lambda_{k+1}}\right]. \end{aligned}$$

For the proof of this theorem we will use the writing of $\mathbf{u}(t) - \mathbf{u}^k(t)$ and $\mathbf{h}(t) - \mathbf{h}^k(t)$ in terms of $\mathbf{e}^k(t)$ and $\mathbf{E}^k(t)$ and $\tilde{\mathbf{e}}^k(t)$ and $\tilde{\mathbf{E}}^k(t)$, respectively, given in (4.2). Therefore, if we want to estimate $A\mathbf{u} - A\mathbf{u}^k$ and $A\mathbf{h} - A\mathbf{h}^k$, then we need to estimate $A\mathbf{u} - A\mathbf{v}^k$ and $A\mathbf{E}^k$ and $A\mathbf{h} - A\mathbf{b}^k$ and $A\tilde{\mathbf{E}}^k$. With this objective, we precise, in first time, to estimate $A^\alpha\mathbf{u} - A^\alpha\mathbf{v}^k$ and $A^\alpha\mathbf{E}^k$ and $A^\alpha\mathbf{h} - A^\alpha\mathbf{b}^k$ and $A^\alpha\tilde{\mathbf{E}}^k$ for $\alpha \in [0, 1)$ and then obtain the desired result.

The regularity results for the solution obtained in the Theorems 3.1 and 3.2 will be also necessary in order to obtain our results. Firstly, observe that we can write the following representation of the solution obtained in Theorem 3.1:

$$\begin{aligned} \mathbf{u}(t) &= e^{-At}\mathbf{u}_0 + \int_0^t e^{-(t-s)A}P(\mathbf{f} - (\mathbf{u}(s)\cdot\nabla)\mathbf{u}(s) + (\mathbf{h}(s)\cdot\nabla)\mathbf{h}(s))ds, \\ \mathbf{h}(t) &= e^{-At}\mathbf{h}_0 + \int_0^t e^{-(t-s)A}(-(\mathbf{u}(s)\cdot\nabla)\mathbf{h}(s) + (\mathbf{h}(s)\cdot\nabla)\mathbf{u}(s))ds. \end{aligned} \tag{5.1}$$

Theorem 5.2 *Suppose that* $\mathbf{f} \in C([0,T], \mathbf{H}^1(\Omega))$ *and* $\mathbf{u}_0, \mathbf{h}_0 \in D(A^{1+\epsilon})$*, then the solution* $(\mathbf{u}, \mathbf{h})$ *of* (1.1)-(1.2) *satisfies for* $0 \le \epsilon < 1/4$*,*

$$\mathbf{u}, \mathbf{h} \in C([0,T]; D(A^{1+\epsilon})) \cap C^1([0,T]; D(A^\epsilon)).$$

The proof of Theorem 5.2 is based the properties of $D(A^\alpha)$, the Stokes operator properties and the use of (4.1), (4.2) and (5.1). In particular, the fractional powers A^α with domain of definition $D(A^\alpha) \subset \mathbf{H}$ are defined for any real α by means of the spectral representation of A. For $\alpha < \beta$ the imbedding $D(A^\beta) \subset D(A^\alpha)$ is compact and $D(A^\beta)$ is dense in $D(A^\alpha)$, therefore A is a sectorial operator and A is the infinitesimal generator of an analytic semigroup $\{e^{-tA}\}$. On $D(A^\alpha)$, the operator A^α commute, with e^{-tA}, and satisfies several properties (see [7]).

Acknowledgements

The first author author was partially supported by Project PGC2018-098308-B-I00, financed by FEDER/Ministerio de Ciencia e Innovación - Agencia Estatal de Investigación, Spain. Second author was partially supported by MATH-AMSUD project 21-MATH-03 (CTMicrAAPDEs), CAPES-PRINT 88887.311962/2018-00 (Brazil), Project UTA-Mayor, 4753-20, Universidad de Tarapacá (Chile). Finally, the third author was partially supported by DIUFRO DI15-0021.

References

[1] L.C. Berselli and J. Ferreira. On the magnetohydrodynamic type equations in a new class of non-cylindrical domains. *Boll. Unione Mat. Ital.*, serie **8** vol. 2-B (1999), 365–382.

[2] M.A. Rojas-Medar and J.L. Boldrini. Spectral Galerkin approximations for the Navier-Stokes equations: uniform in time error estimates. *Rev. Mat. Apl.* **14** (1993), 63–74.

[3] M.A. Rojas-Medar and J.L. Boldrini. The weak solutions and reproductive property for a system of evolution equations of magnetohydrodynamic type. *Proyecciones* **13** no. 2 (1994), 85-97.

[4] J.L. Boldrini and M. Rojas-Medar. On a system of evolution equations of magnetohydrodynamic type. *Mat. Contemp.* **8** (1995), 1–19.

[5] M.A. Rojas-Medar and J.L. Boldrini. Global strong solutions of equations of magnetohydrodynamic type. *J. Austral. Math. Soc. Ser. B* **38** no. 3 (1997), 291-306.

[6] P. Damázio and M.A. Rojas-Medar. On some questions of the weak solutions of evolution equations for magnetohydrodynamic type. *Proyecciones* **16** (1997), 83–97.

[7] H. Fujita and T. Kato. On the Navier-Stokes initial value problem. I. *Arch. Rational Mech. Anal.* **16** (1964), 269–315.

[8] J.G. Heywood. An error estimates uniform in time for spectral Galerkin approximations of the Navier-Stokes problem. *Pacific J. Math.* **98** (1982), 333–345.

[9] G. Lassner. Über ein Rand-Anfangswertproblem der Magnetohydrodynamik. (German) *Arch. Rational Mech. Anal.* **25** (1967), 388–405.

[10] E.A. Notte-Cuello and M.A. Rojas-Medar. On a system of evolution equations of magnetohydrodynamic type: an iterational approach. *Proyecciones* **17** (1998), 133–165.

[11] S.B. Pikelner. Fundamentals of Cosmic Electrodynamics [in Russian]. Fizmatgiz, Moscow, (1961). S.B. Pikelner. Fundamentals of Cosmic Electrodynamics. NASA technical translation, Washington DC, NASA, (1964).

[12] R. Rautmann. On the convergence rate of nonstationary Navier-Stokes approximations. In Proc. IUTAM Symp. 1979, Approximations Methods for Navier-Stokes Problem (R. Rautmann ed.), Springer-Verlag, *Lect. Notes in Math.*, **771** (1980), 235–248.

[13] R. Rautmann. On error bounds for nonstationary spectral Navier-Stokes approximations. In Ordinary and partial differential equation (W.N. Everit and B.D. Sleeman eds.). Springer-Verlag. *Lect. Notes in Math.* **964** (1982), 576–583.

[14] R. Rautmann. A semigroup approach to error estimates for nonstationary Navier-Stokes approximations. *Methoden Verfahren Math. Phys.*, v. **27** (1983), 63–77.

[15] R. Rautmann. On optimum regularity of Navier-Stokes at time $t = 0$. *Math. Z.* **184** (1983), 141–149.

[16] M. A. Rodríguez-Bellido, M. A. Rojas-Medar and A. Sepúlveda-Cerda. On the convergence rate for Galerkin approximation for the magnetohydrodynamic type equations Work in progress.

[17] M.A. Rojas-Medar and R. Beltrán-Barrios. The initial value problem for the equations of magnetohydrodynamic type in noncylindrical domains. *Rev. Mat. Univ. Compl. Madrid*, **8** (1995), 229–251.

[18] R. Salvi. Error estimates for the spectral Galerkin approximations of the solutions of Navier-Stokes type equations. *Glasgow Math. J.* **31** (2) (1989), 199–211.

[19] A. Schlüter. Dynamik des plasmas-I - grundgleichungen, plasma in gekreuzten feldern. *Zeitschrift für Naturforschung/A* **5** (1950), 72–78.

[20] A. Schlüter. Dynamik des plasmas-II - plasma mit neutralgas. *Zeitschrift für Naturforschung/A* **6** (1951), 73-79.

[21] R. Temam. Navier-Stokes Equations, Theory and Numerical Analysis, Third edition. North-Holland Publishing Co., Amsterdam, (1984).

[22] C. Zhao. Initial boundary value problem for the evolution system of MHD type describing geophysical flow in three-dimensional domains. *Math. Methods Appl. Sci.* **26** no. 9 (2003), 759–781.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 332–338)

Asymptotic aspects of the logistic equation under diffusion

José C. Sabina de Lis[1], **Sergio Segura de León**[2]
1. *Departamento de Análisis Matemático & IUEA, Universidad de La Laguna, Spain.*
2. *Departament d'Anàlisi Matemàtica, Universitat de València, Spain.*

Abstract

This talk is devoted to describe the nontrivial solutions to

$$\begin{cases} -\Delta_p u = \lambda |u|^{p-2}u - |u|^{q-2}u & x \in \Omega \\ u = 0 & x \in \partial\Omega. \end{cases}$$

Exponents satisfy $1 < p < q$ while $\lambda > 0$ is a bifurcation parameter. We are confining ourselves to the case where Ω is a ball and solutions are radial. More importantly, we are discussing the asymptotic behavior of these solutions as $p \to 1+$. We are further stating not only the existence of such limits but even introducing the limit problem which such limits solve.

1. Introduction

This talk is firstly devoted to describe the nontrivial solutions to the nonlinear eigenvalue problem:

$$\begin{cases} -\Delta_p u = \lambda |u|^{p-2}u - |u|^{q-2}u, & x \in \Omega, \\ u = 0, & x \in \partial\Omega, \end{cases} \tag{1.1}$$

where $\Omega \subset \mathbb{R}^N$ is a bounded smooth domain, ν is outer unit normal, λ is a positive (bifurcation) parameter and $\Delta_p u = \text{div } (|\nabla u|^{p-2}\nabla u)$ is the p–Laplacian operator. The exponents p, q are assumed to satisfy,

$$1 < p < q. \tag{1.2}$$

The case $p = 2$ is the logistic problem, a well–known model in population dynamics (see [17], [6], also [8] for related applications). As for the nonlinear diffusion regime $p \neq 2$, a detailed discussion of its positive solutions has been performed in [10–12], [15] and [9], the latter specially concerned with the one–dimensional case. Regarding the problem (1.1) observed in a N–dimensional domain Ω, see [13] for existence results on a closely related problem.

A further feature we are going to address is the analysis of the *limit perturbation* of problem (1.1) as $p \to 1$. Namely,

$$\begin{cases} -\Delta_1 u = \lambda \dfrac{u}{|u|} - |u|^{q-2}u, & x \in \Omega, \\ u = 0, & x \in \partial\Omega, \end{cases} \tag{1.3}$$

where $\Delta_1 u = \text{div } \left(\dfrac{\nabla u}{|\nabla u|} \right)$ is the one–Laplacian operator. Such operator finds its natural applications in a broader class of fields ranging from image processing ([4], [18]) to torsion theory ([16]).

Due to the fact that the N–dimensional versions of problems (1.1) and (1.3) are plagued of technical obstacles, main emphasis here will be put on their radial versions. In such case, $\Omega = B(0, R) \subset \mathbb{R}^N$ is a N–dimensional ball with $N \geq 2$. It should be remarked that the one–dimensional versions of (1.1) and (1.3),

$$\begin{cases} -(|u_x|^{p-2}u_x)_x = \lambda |u|^{p-2}u - |u|^{q-2}u, & 0 < x < R, \\ u(0) = u(R) = 0, \end{cases} \tag{1.4}$$

and

$$\begin{cases} -\left(\dfrac{u_x}{|u_x|} \right)_x = \lambda \dfrac{u}{|u|} - |u|^{q-2}u, & 0 < x < R, \\ u(0) = u(R) = 0, \end{cases} \tag{1.5}$$

have been recently studied in [21] (problem (1.4) goes back to [14]).

This note is organized as follows. Basic results, specially those concerning the limit problem (1.3) are reviewed in Section 2. A global description of the set of nontrivial solutions to (1.1) in a ball is presented in Section 3. The features on the limit behavior of solutions to (1.1) as $p \to 1+$ are described in Section 4.

2. Background results

By a (weak) solution $u \in W_0^{1,p}(\Omega) \cap L^q(\Omega)$ to (1.1) it is understood that equality

$$\int_\Omega |\nabla u|^{p-2}\nabla u \nabla v = \lambda \int_\Omega |u|^{p-2}uv - \int_\Omega |u|^{q-2}uv,$$

holds for every test function $v \in C_0^1(\Omega)$. In fact it can be checked that such test functions can be allowed to belong to $W_0^{1,p}(\Omega)$ ([22]).

Analysis of (1.1) in a ball $B(0,R)$ is closely linked to the radial eigenvalues to

$$\begin{cases} -\Delta_p u = \lambda |u|^{p-2}u, & x \in B(0,R), \\ u = 0, & x \in \partial B(0,R), \end{cases} \tag{2.1}$$

which will be designated as,

$$0 < \lambda_{1,p} < \lambda_{2,p} < \dots .$$

We refer to [7], [23] and [20] for a detalied account (also [2] for an early source). Eigenvalues in the unit ball $B(0,1)$ are more conveniently expressed as $\lambda_{n,p} = \omega_n^p$ for certain positive numbers ω_n. Thus, eigenvalues in the ball $B(0,R)$ turn out to be $\lambda_{n,p} = R^{-p}\omega_n^p$.

Following the nowadays well settled down approach in [3] and [4] we introduce the concept of a solution to (1.3). Framework space is

$$BV(\Omega) = \{u \in L^1(\Omega) : Du \in C_0(\Omega, \mathbb{R}^N)'\},$$

that is, the space of functions in $L^1(\Omega)$ whose gradient Du is a vectorial zero order distribution, whose components define finite Radon measures $D_i u$, $1 \le i \le N$ (see [1] for a comprehensive source on this space).

To introduce the concept of weak solution to (1.3), the problematic term $\frac{Du}{|Du|}$ must be conveniently replaced with a suitable field $\mathbf{z} \in L^\infty(\Omega, \mathbb{R}^N)$. On the other hand, the formulation of a Green identity is required in order to test with functions $v \in BV(\Omega)$. Anzellotti's theory is instrumental for these purposes. A featured result in [5] is the identity,

$$\int_\Omega (\mathbf{z}, Dv) + \int_\Omega v \operatorname{div} \mathbf{z} = \int_{\partial\Omega} v[\mathbf{z}, \nu]\, ds, \tag{2.2}$$

which holds for every $\mathbf{z} \in L^\infty_{q'}(\Omega, \mathbb{R}^N) := \{\mathbf{z} \in L^\infty(\Omega, \mathbb{R}^N) : \operatorname{div} \mathbf{z} \in L^{q'}(\Omega)\}$ and $v \in BV_q(\Omega) := BV(\Omega) \cap L^q(\Omega)$. To account for every term in (2.2) it is shown in [5] that the normal component $[\mathbf{z}, \nu]$ has a well–defined trace on $\partial\Omega$ which belongs to $L^\infty(\partial\Omega)$. In addition, the scalar product $\mathbf{z} \cdot Du$ is extended as a bilinear mapping $(\mathbf{z}, Du)$, from $C^1(\overline{\Omega}, \mathbb{R}^N) \times W^{1,1}(\Omega)$ to $L^\infty_{q'}(\Omega, \mathbb{R}^N) \times BV_q(\Omega)$ in the following distributional way:

$$\langle (\mathbf{z}, Du), \varphi \rangle = -\int_\Omega u \operatorname{div}(\varphi \mathbf{z}), \qquad \varphi \in C_0^\infty(\Omega).$$

It is shown in [5] that $(\mathbf{z}, Du)$ defines a finite Radon measure in Ω such that

$$|(\mathbf{z}, Du)(B)| \le \|\mathbf{z}\|_\infty |Du|(B),$$

$B \subset \Omega$ being a Borelian and $|Du|$ standing for the total variation of Du.

We are now ready for the next definition.

Definition 2.1 A function $u \in BV_q(\Omega)$ defines a (weak) solution to (1.3) provided that there exist $\mathbf{z} \in L^\infty_{q'}(\Omega, \mathbb{R}^N)$, $\|\mathbf{z}\|_\infty \le 1$, $\beta \in L^\infty(\Omega)$, $\|\beta\|_\infty \le 1$ such that,

i) $-\operatorname{div} \mathbf{z} = \lambda\beta - |u|^{q-2}u$, in $\mathcal{D}'(\Omega)$,

ii) $\beta u = |u|$ and $(\mathbf{z}, Du) = |Du|$, in $\mathcal{D}'(\Omega)$,

iii) $[z, \nu]u = -|u|$ on $L^1(\partial\Omega)$, (boundary condition).

Remark 2.2 Boundary condition in iii) is suggested by two features. First one, the fact that the weak–$*$ limit $u \in BV(\Omega)$ of a sequence $u_n \in W_0^{1,1}(\Omega)$ could eventually exhibits a nonzero trace on the boundary. Second one, that solutions of (1.3) could be approximated as $p \to 1$ by corresponding solutions to (1.1).

3. Radial solutions

A general view on the nontrivial solutions to (1.1) in a ball is contained in the next statement.

Theorem 3.1 *Assume $1 < p \leq 2$. Then, problem*

$$\begin{cases} -\Delta_p u = \lambda |u|^{p-2}u - |u|^{q-2}u, & x \in B(0,R), \\ u = 0, & x \in \partial B(0,R), \end{cases} \tag{3.1}$$

exhibits the following features.

i) [Range and amplitude] *Nontrivial solutions are only possible when $\lambda > \lambda_{1,p}$ while the normalized amplitude*

$$\alpha := \lambda^{-\frac{1}{q-p}} \|u\|_\infty,$$

satisfies $\alpha < 1$.

ii) [Positive solutions] *There exists a* unique *positive* (*radial*) *solution $u_{\lambda,1}$ for all $\lambda > \lambda_{1,p}$, bifurcating from $u = 0$ at $\lambda = \lambda_{1,p}$ while*:

$$\lambda^{-\frac{1}{q-p}} \|u_{\lambda,1}\|_\infty \to 1 \quad \text{as } \lambda \to \infty.$$

iii) [Existence of branches] *For all $n \geq 2$, a symmetric family $\pm u_{\lambda,n}(r)$ of nontrivial radial solutions,* exactly defined *for all $\lambda > \lambda_{n,p}$, bifurcates from $u = 0$ at $\lambda_{n,p}$ and,*

$$\lambda^{-\frac{1}{q-p}} \|u_{\lambda,n}\|_\infty \to 1 \quad \text{as } \lambda \to \infty.$$

iv) [Nodal properties] *Every $\pm u_{\lambda,n}(r)$ vanishes exactly at $n-1$ values $r_k \in (0,R)$.*

v) [Continuity of the branches] *Bifurcated branches $\pm u_{\lambda,n}$ define a* continuous *curve C_n when parameterized by the normalized amplitude $\alpha = \lambda^{-\frac{1}{q-p}} \|u\|_\infty$, $0 < \alpha < 1$. More precisely, there exist continuous mappings $\alpha \mapsto \lambda_n(\alpha)$, $\alpha \mapsto u_n(\alpha) \in W_0^{1,p}(B(0,R))$, $0 < \alpha < 1$, such that,*

$$\pm u_{\lambda,n} = \pm u_n(\alpha), \qquad \lambda = \lambda_n(\alpha).$$

Proof (Sketch) The scaling $u(r) = \lambda^{\frac{1}{q-p}} v(t)$, $t = \lambda^{\frac{1}{p}} r$, transforms (3.1) into,

$$\begin{cases} -(|v_t|^{p-2}v_t)_t - \dfrac{N-1}{t}|v_t|^{p-2}v_t = |v|^{p-2}v - |v|^{q-2}v, & 0 < t < \lambda^{\frac{1}{p}} R, \\ v(0) = \alpha, \quad v_t(0) = 0, \end{cases} \tag{3.2}$$

where:

$$\max v = \alpha, \qquad 0 < \alpha < 1,$$

and v must satisfies the boundary condition:

$$v(\lambda^{\frac{1}{p}} R) = 0.$$

The initial value problem (3.2) admits a unique C^2 solution $v = v(\cdot, \alpha)$ which is defined in $[0, \infty)$ and satisfies $\lim_{t\to\infty}(v(t), v_t(t)) = (0,0)$. Moreover, v exhibits infinitely many simple zeros,

$$0 < \theta_1(\alpha) < \theta_2(\alpha) < \cdots < \theta_n(\alpha) < \cdots, \qquad \theta_n \to \infty.$$

Functions $\theta_n(\alpha)$ are shown to be continuous in $\alpha \in (0,1)$ and,

$$\lim_{\alpha\to 0+} \theta_n(\alpha) = \omega_n, \qquad \lim_{\alpha\to 1-} \theta_n(\alpha) = \infty,$$

where $\omega_n = \lambda_{n,p}(B(0,1))^{\frac{1}{p}}$.

To solve (3.1) amounts to:

$$\lambda^{\frac{1}{p}} R = \theta_n(\alpha) \qquad \Leftrightarrow \qquad \lambda = R^{-p}\theta_n(\alpha)^p.$$

By setting this value of λ in the expression for u:

$$u(r) = \lambda^{\frac{1}{q-p}} v(\lambda^{\frac{1}{p}} r, \alpha), \qquad 0 \leq r \leq R,$$

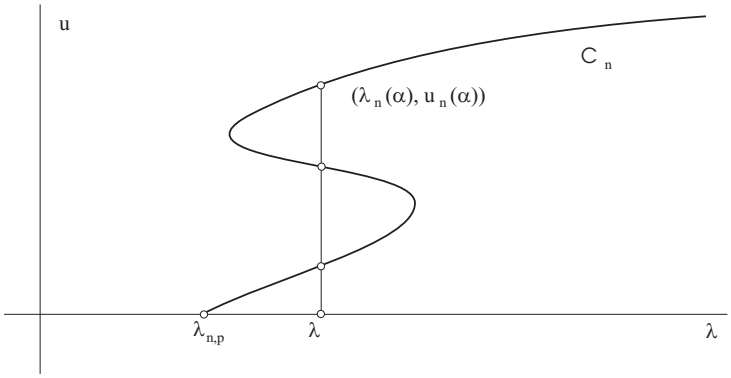


Fig. 1 Family C_n of nontrivial solutions bifurcated from $u = 0$ at $\lambda = \lambda_{n,p}$. Only a half of C_n has been depicted. That one corresponding to $u(0) > 0$. It is stressed that its exact range of existence is $[\lambda_{n,p}, \infty)$.

the family $u_{n,\lambda}$ is obtained. Moreover by defining:

$$\lambda_n(\alpha) = R^{-p}\theta_n(\alpha)^p, \qquad u_n(r,\alpha) = \lambda_n^{\frac{1}{q-p}} v(\lambda_n^{\frac{1}{p}} r, \alpha),$$

$\{u_{\lambda,n}\}$ is alternatively represented as a continuous curve $(\lambda_n(\alpha), u_n(\alpha))$ in $\mathbb{R} \times W_0^{1,p}(B(0,R))$. It should be also observed that $u_n(\cdot,\alpha)$ vanishes at the points,

$$r_k = R\frac{\theta_k(\alpha)}{\theta_n(\alpha)}, \qquad k = 1, \ldots, n.$$

Assertion concerning the existence of the family $u_{\lambda,n}$ exactly at the interval $[\lambda_{n,p}, \infty)$ is a consequence of the estimate:

$$\theta_n(\alpha) > \omega_n, \quad 0 < \alpha < 1.$$

The proof of this fact deserves a delicate proof and it is also omitted (see [22]). □

Remark 3.2 The existence of a global continuum C_n^* bifurcating from zero at $\lambda = \lambda_{n,p}$ was stated in [12] (see also [19]). Theorem 3.1 improves these results in two regards. Firstly, family of solutions $u_{\lambda,n}$ is shown to exists exactly at the range $\lambda > \lambda_{n,p}$. Secondly, ours is not a mere continuum C_n^* but rather a global continuous curve C_n.

4. Limit behavior

The sequence,

$$0 < \bar{\lambda}_1 < \bar{\lambda}_2 < \cdots$$

of radial eigenvalues to $-\Delta_1$,

$$\begin{cases} -\Delta_1 u = \lambda \dfrac{u}{|u|}, & x \in B(0,R), \\ u = 0, & x \in \partial B(0,R), \end{cases} \tag{4.1}$$

has been recently studied in [20]. Among other featured properties it is shown there that,

$$\lim_{p\to 1} \lambda_{n,p} = \bar{\lambda}_n, \qquad \text{for every } n \in \mathbb{N}.$$

Our next result describes a set of *distinguished* nontrivial radial solutions to (1.3). Those ones obtained as the limit of solutions to (1.1) as $p \to 1$. In addition this precise feature of the solutions is characterized by a suitable energy condition. In the forthcoming statement, the reference zeros θ_n introduced in the proof of Theorem 3.1 are involved. It should be remarked that they also depends on $p > 1$ and an important fact to be reported is the existence of their limits $\bar{\theta}_n$ as $p \to 1+$ (see ii) below). Figure 2 depicts this dependence through a simulation.

Theorem 4.1 *The structure of the set of* radial *nontrivial solutions to*

$$\begin{cases} -\text{div}\left(\dfrac{\nabla u}{|\nabla u|}\right) = \lambda \dfrac{u}{|u|} - |u|^{q-2}u, & x \in B(0,R), \\ u = 0, & x \in \partial B(0,R), \end{cases}$$

can be described as follows.

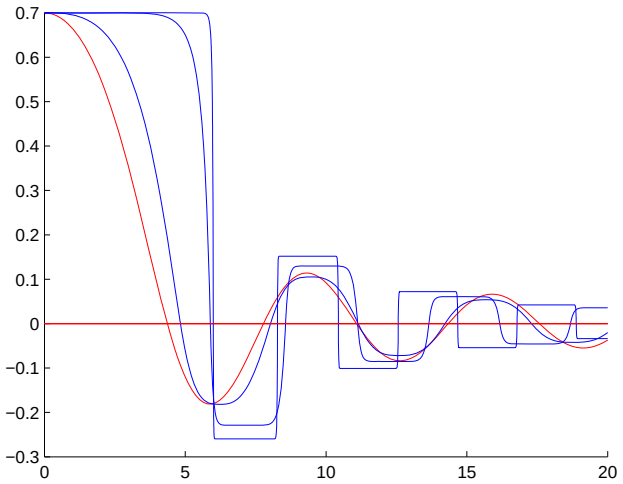


Fig. 2 Profiles of $v(t)$ and corresponding zeros θ_n for varying values of $p > 1$. Simulation has been performed for $N = 3$, $q = 3$, $\alpha = 0.7$. Then, chosen values of p are $p = 2$, $p = 1.5$, $p = 1.1$ and $p = 1.01$. Plots become steeper as p decays to unity.

i) [Normalized amplitude estimate] *Nontrivial solutions are only possible if* $\lambda > \bar{\lambda}_1$. *Moreover, the normalized amplitude* $\alpha := \lambda^{-\frac{1}{q-1}} \|u\|_\infty$ *of such solutions satisfies,*

$$0 < \alpha < 1.$$

ii) [Limits of zeros] *There exists a family of smooth functions* $\bar{\theta}_n(\alpha)$,

$$0 < \bar{\theta}_1(\alpha) < \bar{\theta}_n(\alpha) < \cdots,$$

such that,

$$\lim_{p\to 1} \theta_n(\alpha) = \bar{\theta}_n(\alpha), \qquad 0 < \alpha < 1.$$

iii) [Existence] *To every radial eigenvalue* $\bar{\lambda}_n$ *there corresponds a symmetric family* $\pm\bar{u}_{\lambda,n}$ *of nontrivial solutions which bifurcates from* $u = 0$ *at* $\bar{\lambda}_n$. *In addition, such family is defined for each* $\lambda > \bar{\lambda}_n$ *while the normalized amplitude of its members satisfies,*

$$\lim_{\lambda\to\infty} \lambda^{-\frac{1}{q-1}} \|\bar{u}_{\lambda,n}\|_\infty = 1.$$

iii) [Smoothness] *Family* $\pm\bar{u}_{\lambda,n}$ *constitutes a smooth curve* $\overline{C}_n$ *in* $\mathbb{R} \times BV(B(0,R))$ *when parameterized by the normalized amplitude* $0 < \alpha < 1$. *More precisely, a decreasing family of smooth positive functions* $\alpha \mapsto \bar{\alpha}_n(\alpha)$ *exists such that by setting,*

$$\bar{\lambda}_n(\alpha) = R^{-1}\bar{\theta}_n(\alpha), \qquad \bar{u}_n(\cdot,\alpha) = \bar{\lambda}_n^{\frac{1}{q-1}} \sum_{k=1}^{n} (-1)^{k-1} \bar{\alpha}_{k-1} \chi_{I_k},$$

χ_{I_k} *being the characteristic function of the interval* $I_k = \left(R\dfrac{\bar{\theta}_{k-1}(\alpha)}{\bar{\theta}_n(\alpha)}, R\dfrac{\bar{\theta}_k(\alpha)}{\bar{\theta}_n(\alpha)} \right)$, *then*

$$\pm\bar{u}_{\lambda,n} = \bar{u}_n(\alpha) \qquad \text{for } \lambda = \bar{\lambda}_n(\alpha).$$

iv) [Convergence of branches] *Let* C_n *be the n–th curve of nontrivial solutions introduced in Theorem* 3.1. *Then*

$$C_n \to \overline{C}_n \qquad \text{as } p \to 1+,$$

in the sense that,

$$\lim_{p\to 1} (\lambda_n(\alpha), u_n(\alpha)) = (\bar{\lambda}_n(\alpha), \bar{u}_n(\alpha)) \qquad \text{in } \mathbb{R} \times BV(B(0,R)),$$

for every $0 < \alpha < 1$.

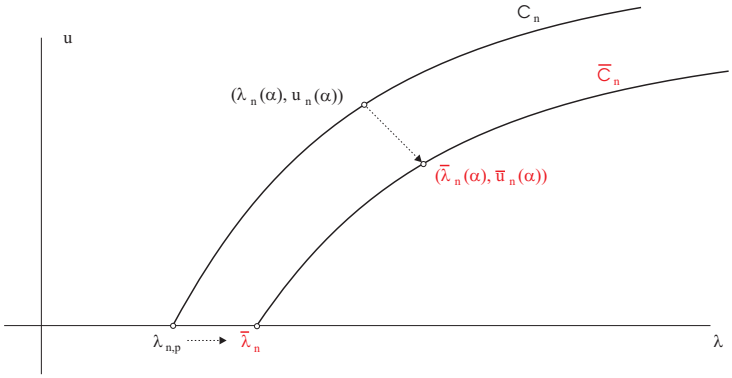


Fig. 3 Convergence of branches as $p \to 1+$.

v) [Uniqueness] Every *nontrivial solution u to* (4.1) *fulfilling the 'energy' condition,*

$$\frac{d}{dr}\left(\lambda|u| - \frac{|u|^q}{q}\right) = -\frac{N-1}{r}|u_r| \qquad \text{in } \mathcal{D}(0,R)'. \tag{4.2}$$

necessarily belongs to some of the previous families $\overline{C}_n = \{\pm\bar{u}_{\lambda,n}\}$.

Proof (Sketch) A first step of compactness nature is the following (subindex p refers to dependence on p). Family $v_p(\cdot,\alpha)$ of solutions to (3.2) admits a subfamily, still denoted v_p, while a function $v_1 \in BV_{loc}(0,\infty)$ exists so that,

$$v_p \rightharpoonup v_1 \qquad \text{weakly in } L^s(0,b;t^{N-1}\,dt) \text{ as } p \to 1,$$

for every $b > 0$ and $1 \le s < \infty$.

A second step consists in proving that $v = v_1(t)$ solves in the sense of Definition 2.1 the initial value problem,

$$\begin{cases} -\left(\dfrac{v_t}{|v_t|}\right)_t - \dfrac{N-1}{t}\dfrac{v_t}{|v_t|} = \dfrac{v}{|v|} - |v|^{q-2}v, & t > 0, \\ v(0+) = \alpha, \qquad v_t(0) = 0, \end{cases} \tag{4.3}$$

together with the energy condition,

$$\left(|v| - \frac{|v|^q}{q}\right)_t = -\frac{N-1}{t}|v_t| \qquad \text{in } \mathcal{D}(0,R)'. \tag{4.4}$$

A third and crucial step is showing that problem (4.3) constrained with condition (4.4) exhibits a unique solution. Moreover, such solution can be expressed in the exact form,

$$v_1(t) = \sum_{n=1}^{\infty} (-1)^{n-1}\bar{\alpha}_{n-1}\chi_{(\bar{\theta}_{n-1},\bar{\theta}_n)}(t),$$

for a precisely computed pair $\bar{\lambda}_n$, $\bar{\theta}_n$, of monotone sequences of positive numbers satisfying $\bar{\lambda}_n \to 0$ and $\bar{\theta}_n \to \infty$.

Final step is checking that family $\bar{u}_{\lambda,n}$ can be defined as,

$$\bar{u}_{\lambda,n}(r) = \lambda^{\frac{1}{q-1}} v_1(\lambda r), \qquad \text{where } \lambda = R^{-1}\bar{\theta}_n.$$

To this purpose suitable candidates for $\mathbf{z}$ and β in Definition 2.1 must be furnished.

A detailed account of the (lengthy) proofs of all these assertions is contained in [22]. □

Remark 4.2

a) Functions $\bar{\theta}_n(\alpha)$ and $\bar{\alpha}_n(\alpha)$ can be recursively computed starting at $n = 0$ with values $\bar{\theta}_0(\alpha) = 0$, $\bar{\alpha}_0(\alpha) = \alpha$.

b) Further families of nontrivial solutions to (4.1) not satisfying the energy condition (4.2) can be found. A characteristic property of such solutions is that they vanish in nonempty interior regions.

Acknowledgements

J. Sabina has been supported by DGI under Grant MTM2014–52822–P; S. Segura has been partially supported by MCIyU & FEDER, under project PGC2018–094775–B–I00.

References

[1] Luigi Ambrosio, Nicola Fusco, and Diego Pallara. *Functions of Bounded Variation and Free Discontinuity Problems*. Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, New York, 2000.

[2] Aomar Anane. *Etude des valeurs propres et de la résonance pour l'opérateur p–Laplacien*. Thése de doctorat. Université Libre de Bruxelles, 1987.

[3] Fuensanta Andreu, Coloma Ballester, Vicent Caselles, and José M. Mazón. The Dirichlet problem for the total variation flow. *J. Funct. Anal.*, 180(2):347–403, 2001.

[4] Fuensanta Andreu-Vaillo, Vicent Caselles, and José M. Mazón. *Parabolic Quasilinear Equations Minimizing Linear Growth Functionals*, volume 223 of *Progress in Mathematics*. Birkhäuser Verlag, Basel, 2004.

[5] Gabriele Anzellotti. Pairings between measures and bounded functions and compensated compactness. *Ann. Mat. Pura Appl. (4)*, 135:293–318 (1984), 1983.

[6] Robert Stephen Cantrell and Chris Cosner. *Spatial ecology via reaction-diffusion equations*. Wiley Series in Mathematical and Computational Biology. John Wiley & Sons, Ltd., Chichester, 2003.

[7] Manuel A. del Pino and Raúl F. Manásevich. Global bifurcation from the eigenvalues of the p-Laplacian. *J. Differential Equations*, 92(2):226–251, 1991.

[8] Paul C. Fife. *Mathematical aspects of reacting and diffusing systems*, volume 28 of *Lecture Notes in Biomathematics*. Springer-Verlag, Berlin-New York, 1979.

[9] J. García-Melián and J. Sabina de Lis. Stationary patterns to diffusion problems. *Math. Methods Appl. Sci.*, 23(16):1467–1489, 2000.

[10] J. García-Melián and J. Sabina de Lis. Stationary profiles of degenerate problems when a parameter is large. *Differential Integral Equations*, 13(10-12):1201–1232, 2000.

[11] J. García Melián and J. Sabina de Lis. Uniqueness to quasilinear problems for the p-Laplacian in radially symmetric domains. *Nonlinear Anal.*, 43(7, Ser. A: Theory Methods):803–835, 2001.

[12] J. García-Melián and J. Sabina de Lis. A local bifurcation theorem for degenerate elliptic equations with radial symmetry. *J. Differential Equations*, 179(1):27–43, 2002.

[13] 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.

[14] Mohammed Guedda and Laurent Véron. Bifurcation phenomena associated to the p-Laplace operator. *Trans. Amer. Math. Soc.*, 310(1):419–431, 1988.

[15] Shoshana Kamin and Laurent Véron. Flat core properties associated to the p-Laplace operator. *Proc. Amer. Math. Soc.*, 118(4):1079–1085, 1993.

[16] Bernhard Kawohl. On a family of torsional creep problems. *J. Reine Angew. Math.*, 410:1–22, 1990.

[17] J. D. Murray. *Mathematical biology. I*, volume 17 of *Interdisciplinary Applied Mathematics*. Springer-Verlag, New York, third edition, 2002. An introduction.

[18] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60:259–268, 1992.

[19] Bryan P. Rynne. Simple bifurcation and global curves of solutions of p-Laplacian problems with radial symmetry. *J. Differential Equations*, 263(6):3611–3626, 2017.

[20] José C. Sabina de Lis and Sergio Segura de León. The limit as $p \to 1$ of the higher eigenvalues of the p-Laplacian operator $-\Delta_p$. *To appear in Indiana Univ. Math. J.*, 2019.

[21] José C. Sabina de Lis and Sergio Segura de León. 1d–logistic reaction and p–laplacian diffusion as p goes to one. *Ricerce di Matematica*, 2021.

[22] José C. Sabina de Lis and Sergio Segura de León. Logistic reaction coupled to p–laplacian diffusion as p goes to 1. *Preprint*, 2021.

[23] Wolfgang Walter. Sturm-Liouville theory for the radial Δ_p-operator. *Math. Z.*, 227(1):175–185, 1998.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 339–343)

Analysis of turbulence models for flow simulation in the aorta

Samuel Santos[1], Jose M. Rojas[2], Pau Romero[1], Miguel Lozano[1], J. Alberto Conejero[3], Ignacio García-Fernández[1]

1. *CoMMLab, Universidad de Valencia, Burjassot, Valencia, Spain,* ignacio.garcia@uv.es
2. *Escuela Técnica Superior de Ingeniería del Diseño, Universidad Politécnica de Valencia, Valencia, Spain,* joroar1@etsid.upv.es
3. *Departmento de Matemática Aplicada, Universidad Politécnica de Valencia, Valencia, Spain,* aconejero@upv.es

Abstract

Computation of Wall Shear Stress (WSS) in the aorta wall is a relevant problem, since it has been related to the appearance of several cardiovascular diseases. In this context, our aim is to solve Navier-Stokes (NS) equations with boundary conditions in the aorta. For an accurate estimation of WSS, a proper election of the turbulence model is of great relevance. We present a study to compare WSS estimation considering three different turbulence models in the thoracic aorta and an analysis of the influence of the aortic valve. The size and properties of the appropriate mesh to use is also discussed. Our simulations are carried out with the Finite Volume Method solver OpenFoam.

1. Introduction

Computational Fluid Dynamics has become an essential tool in the study of blood flow in order to understand genesis of cardiovascular diseases. In this work we focus on the WSS reached in the toracic aorta at peak systolic conditions. Particularly, we know that low values of WSS are related to atherosclerosis, see [5]. To do so, we first need to determine what model is more appropiate to this task. On the one hand we test a model based in the NS equations and on the other hand $k-\epsilon$, $k-\omega$ and SST $k-\omega$ turbulence models which are reformulations of the first one. Once we have chosen our model we will study the influence of three types of aortic valves: a healhty valve and two prosthetic valves. This document is structured as follows: In section 2 we introduce the models of fluid likely to be chosen and the corresponding equations. In section 3 we explain tecnical details of simulations and results obtained. Finally, in section 4 we summarize, conclude and expose some improvements and future works.

2. Models and boundary conditions

Fluid dynamics is governed by NS equations, a system of coupled partial differential equations concerning fluid velocity and pressure. In this work we deal with blood flowing through aorta. In this context we can suppose that blood is an incompressible newtonian fluid. As our aim is to get WSS, we first need to compute blood flow. This requires to solve NS equations with a set of boundary conditions that, in our case, will reproduce peak systolic conditions. We also neglect time derivatives in all model considered so we are calculating instantaneous WSS in the time of maximum blood flow. We consider, in the first place, the usual stationary Navier-Stokes equations for an incompressible Newtonian fluid, given by

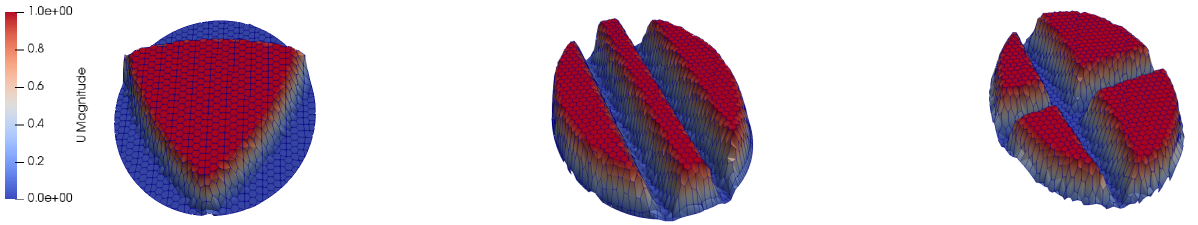
$$\nabla \cdot \vec{v} = 0; \qquad \vec{v} \cdot \nabla \boldsymbol{v} = -\frac{1}{\rho}\nabla p + \frac{1}{\rho}\nabla \cdot \Sigma, \tag{2.1}$$

where $\boldsymbol{v}$ is blood velocity, p is pressure, ρ is the blood density and $\Sigma = \mu\left(\nabla \boldsymbol{v} + \nabla \boldsymbol{v}^T\right)$ is the viscous stress tensor for a Newtonian fluid, with μ the viscosity of blood. We apply a zero-gradient condition in the outlet for the velocity and a non-slip condition in the aortic wall. Besides, velocity profiles in the inlet are settled trying to reproduce a healthy valve and two artificial prosthetic valves as shown in Figure 1. For pressure, we use a zero-gradient condition in the inlet and set a 0 value condition in the outlet.

The other three models are turbulence models (as $Re \sim 6500$) and are included in the context of the Reynolds-Averaged-Navier-Stokes (RANS) equations, consisting in the decomposition of each variable in a time average component and a fluctuating component. For example, $\boldsymbol{v} = \boldsymbol{V} + \boldsymbol{v}'$ where $\boldsymbol{V}$ and $\boldsymbol{v}'$ are the time averaged and the fluctuating component of velocity respectively. The average is taken in a sufficiently high time. The first RANS model considered is the $k-\epsilon$ model where turbulence is stored in the new variables

$$k = \frac{\overline{\boldsymbol{v}' \cdot \boldsymbol{v}'}}{2},$$

 ISBN 978-84-18482-21-2



(a) Natural healthy valve. **(b)** Artificial bileaflet valve. **(c)** Artificial tilting disk valve.

Fig. 1 Vertical profile of some of the inlet boundary conditions considered. The maximum velocity is 1 m/s.

variable	inlet	wall	outlet
ϵ	2	zero gradient	zero gradient
k	10^{-6}	10^{-10}	zero gradient
ω	1	ω-wall-function	zero gradient

Tab. 1 Boundary conditions applied to turbulent variables. All units in SI.

and ϵ. k is called turbulent kinetic energy and ϵ is the rate of dissipation of k. The new equations for these variables are:

$$\boldsymbol{V}\cdot\nabla k = \frac{1}{\rho}\nabla\cdot(\mu_k\nabla k) + \frac{\mu_t}{\rho}(\Sigma' : \nabla\boldsymbol{V}) - \epsilon, \tag{2.2}$$

$$\boldsymbol{V}\cdot\nabla \epsilon = \frac{1}{\rho}\nabla\cdot(\mu_\epsilon\nabla \epsilon) + C_{\epsilon 1}\frac{\epsilon}{\rho k}(\Sigma' : \nabla\boldsymbol{V}) - C_{\epsilon 2}\frac{\epsilon^2}{k}, \tag{2.3}$$

where $\mu_t = \rho C_\mu k^2/\epsilon$ is the tubulent viscosity, $\mu_k = \mu + \mu_t/\sigma_k$ and $\mu_\epsilon = \mu + \mu_t/\sigma_\epsilon$ are effective viscosities and C_μ, $C_{\epsilon 1}$,$C_{\epsilon 2}$, σ_k and σ_ϵ are empirical constants whose values can be consulted in [6]. On the other hand

$$\Sigma' = \rho\overline{\boldsymbol{v}'\otimes\boldsymbol{v}'}$$

is the Reynolds stress tensor.

The second model used is the $k-\omega$ model. We introduce here the variable $\omega = \epsilon/(C_\mu k)$ for which an equation can be derived from (2.3). The new set of equations for the turbulent variables rest

$$\boldsymbol{V}\cdot\nabla k = \frac{1}{\rho}\nabla\cdot(\mu_k\nabla k) + \frac{\mu_t}{\rho}(\Sigma' : \nabla\boldsymbol{V}) - \beta^* k\omega, \tag{2.4}$$

$$\boldsymbol{V}\cdot\nabla \omega = \frac{1}{\rho}\nabla\cdot(\mu_\omega\nabla \omega) + C_{\alpha 1}\frac{\omega}{\rho k}(\Sigma' : \nabla\boldsymbol{V}) - C_{\beta 1}\omega^2, \tag{2.5}$$

This model has the advantage of being more precise near the wall than $k-\epsilon$ model, meanwhile the latter is more precise in the bulk flow (the stream outside the boundary layer). The last turbulence model used is the SST $k-\omega$ model which combines $k-\epsilon$ and $k-\omega$ models through blending functions. The details on the construction of this model are too extensive to be included in this document and can be consulted in [6].

Boundary conditions over the new turbulent variables are shown in table 1. We have taken considerations from [4], for k and ω where also the ω-wall-function can be read. Moreover, we have used formulas and information from [1], to settle boundary conditions on ϵ.

3. Simulation and results

We employ the Finite Volume Method solver OpenFoam to solve the previous models of NS equations in a real aorta acquired from Computerized Tomography. A mesh of about 8.6M elements with maximum spatial resolution of $50\mu m$ in the wall normal direction has been used in order to have accurate computation of the boundary layer. We will not use wall functions in the simulations presented here for pressure and velocity so we compute the entire boundary layer. A representation of the mesh used is shown in figure 2.

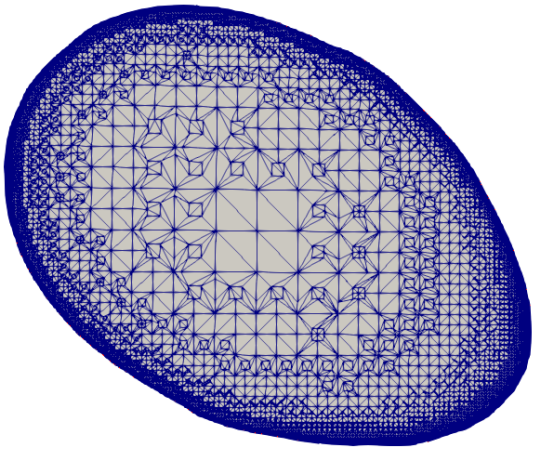
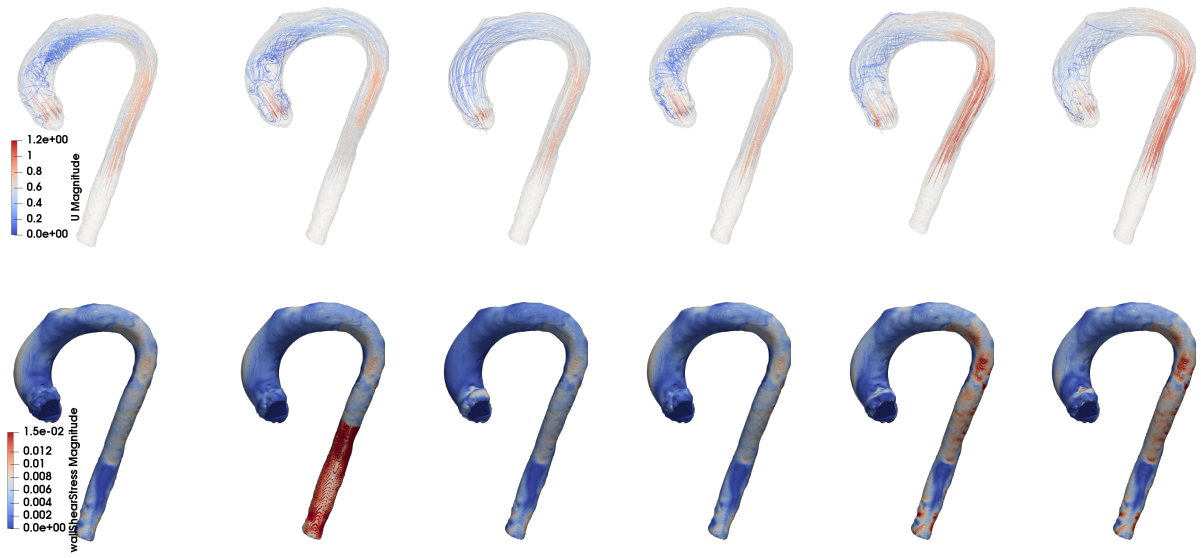


Fig. 2 Section of aorta where the mesh used was outlined.

As it can be seen a 4 refinement level is used combined with 15 extra layers in the aortic wall in order to get good WSS estimation. A Green-Gauss finite volume discretization is employed for the gradient of scalar variables. On the other hand upwind schemes are applied in divergence terms. After discretization a combination of a Gauss-Seidel method, Geometric-algebraic-multi-grid method and a smooth solver are used to solve the linear systems involved. Details can be found in [6, 7]. Also, the resolution algorithm called semi-implicit method for pressure-linked equations (SIMPLE algorithm) is employed in the conservation laws involving pressure and velocity.

Results are shown in figure 3. In the first row blood streamlines are presented, while the corresponding WSS distribution are shown below. Each column corresponds to a different model. In the first four columns a healthy valve has been settled and in the last two columns the bileaflet valve and tilting disk valve for the SST $k-\omega$ model haven been implemented.



(a) No turbulence model and healthy valve.

(b) $k-\epsilon$ model and healthy valve.

(c) $k-\omega$ model and healthy valve.

(d) SST $k-\omega$ model and healthy valve.

(e) SST $k-\omega$ model and bileaflet valve.

(f) SST $k-\omega$ model and tilting disk valve.

Fig. 3 Streamlines (upper row) and WSS/ρ profiles (lower row) obtained with different turbulence models. All units in the SI.

The most remarkable result is the high values of WSS provided by the $k-\epsilon$ in the descending aorta. Since this model does not properly compute turbulence in regions with large pressure gradients (such as the boundary layer) and strong accelerations (when the aortic duct narrows), this model can be assumed to provide a poor approximation.

On the other hand, we know that the $k-\omega$ model is sensitive to boundary conditions of the turbulent variables in the inlet free stream, which does not happen with the $k-\epsilon$ model. This explains the distinct behaviour of streamlines computed with $k-\omega$ model in the cavity of the aorta. Hence, it seems that the most suitable turbulence model, out of the ones used here, is SST $k-\omega$ model. The cavity flow computed with this model looks like the one computed with $k-\epsilon$ model, which is most reliable in this region. Also, the WSS profile computed with SST $k-\omega$ model has more resemblance with the one computed with $k-\omega$ model, the one that behaves well near the wall. Not applying any turbulence model seems in good agreement with SST $k-\omega$ model. Nevertheless k is an esential parameter in the study of diseases like stenosis or coarctation. Then, SST $k-\omega$ model is the one selected from now on.

Streamlines obtained in ascending aorta with the three types of valves are trustable when comparing with experimental and theoretical works, see [2, 8]. Regarding the effect of the valve type, it clearly affects the WSS profile. We consider a WSS critical value of 0.5 Pa below which there is risk of atherosclerosis appearence. Figure 4 shows a detailed analysis of the influence of the type of valve on the WSS values. We analize 7 sections along the aorta, S_i for $1 \leq i \leq 7$. S_1, S_2 and S_3 are placed in the ascending aorta where the results are more trustworthy. We can see that in the case of the bileaflet valve almost 16% of S_1 is in risk, the highest value of the analysis, meanwhile 10% with the tilting disk valve around and 2% with the healthy valve. In S_2 there is no region in risk for any valve and in S_3 we have 6%, 5% and 4% for the healthy, bileaflet and tilting disk valve respectivelly.

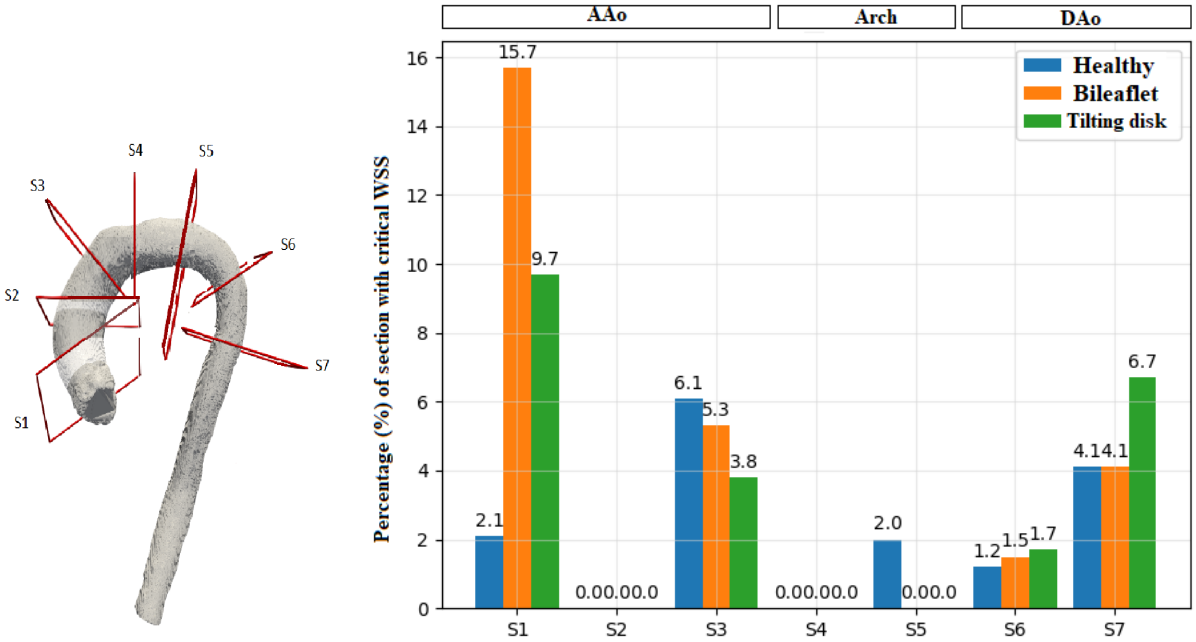


Fig. 4 Locations where sections analyzed were placed (left) and histogram of the percentage of section with critical WSS for the seven sections and for the three types of valve (right).

Values on the descending aorta are not trustable as we are neglecting the supraortic arteries which suppose the 30% of the entire flow. However, we observe an increase of the risk region with artificial valves.

4. Conclusions

First of all we made an analsys about what model of turbulence was more suitable for WSS estimation. We conclude that the SST $k-\omega$ model was the most realiable as it possesses the good properties from both $k-\epsilon$ and $k-\omega$ models and, besides, it calculates the turbulent kinetic energy k which is important in the study of cardiovascular diseases.

Concerning WSS values, it can be observed that both prosthetic valves contribute to increase them and that the bileaflet valve does it in a minor level than the tilting disk valve in the ascending aorta. Hence, we can say that the human biology has made a good work and that whenever a prosthetic valve is needed we support the bileaflet valve above the tilting disk one.

Transitory simulations to take into account the whole cardiac cycle are now taking place, so we will be able to compute other hemodynamic variables of interest as OSI or TAWSS. Also, in the future, a fluid-structure interaction should be considered to get a complete study of genesis of cardiovascular diseases.

References

[1] C. J. Greenshields. OpenFOAM user guide Version 8. The OpenFOAM Foundation, 2020.

[2] U. Gülan, B. Luthi, M. Holzner, A. Liberzon, A. Tsinober & W. Kinzelbach. Experimental study of aortic flow in the ascending aorta via particle tracking velocimetry. Experiments in fuids 53.5, pags. 1469-1485, 2012.

[3] D. A. Jones, M. Chapuis, M. Liefvendahl, D. Norrison and R. Widjaja. RANS Simulations using OpenFOAM Software. No. DST-Group-TR-3204. Defence Science and Technology Group Fishermans Bend Victoria Australia, 2016.

[4] D. Lindblad, A. Jareteg & O. Petit. Implementation and run-time mesh refinement for the $k-\omega$ SST DES turbulence model when applied to airfoils. Project work. Chalmers University of Technology, 2014.

[5] A. M. Malek, S. L. Alper & S. Izumo. Hemodynamic shear stress and its role in atherosclerosis. Jama 282.21, pags. 2035-2042, 1999.

[6] F. Moukalled, L. Mangani and M. Darwish. The finite volume method in computational fluid dynamics. Vol. 113. Berlin, Germany:: Springer, 2016.

[7] J. M. Rojas. Influencia del tipo de válvula en el flujo sanguíneo de la aorta torácica: zonas en riesgo de arterioesclerosis. Final master disertation, Escuela Técnica Superior de Ingeniería del Diseño, Universidad Politécnica de Valencia.

[8] W. Yin, Y. Alemu, K. Affeld, J. Jesty & D. Bluestein. Flowinduced platelet activation in bileaflet and monoleaflet mechanical heart valves. Annals of biomedical engineering 32.8, pags. 1058-1066, 2004.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 344–351)

Overdetermined elliptic problems in onduloid-type domains with general nonlinearities

Jing Wu
jingwulx@correo.ugr.es *Universidad de Granada, Spain*

Abstract

In this paper, we prove the existence of solutions to a general semilinear elliptic problem with overdetermined boundary conditions. The proof uses a local bifurcation argument from the straight cylinder, in analogy with the onduloids and the theory of Constant Mean Curvature surfaces. Such examples have been found already for linear problems or with nonlinearity $f(u) = 1$. In this work we are able to extend this phenomenon for a large class of functions $f(u)$.

Remark: This manuscript, especially the whole proof, is a work in progress in collaboration with David Ruiz and Pieralberto Sicbaldi.

1. Introduction

This paper is devoted to the existence of new solutions of a semilinear overdetermined elliptic problem in the form

$$\begin{cases} \Delta u + f(u) = 0 & \text{in } \Omega \\ u > 0 & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \\ \frac{\partial u}{\partial \nu} = \text{constant} & \text{on } \partial\Omega \end{cases} \tag{1.1}$$

where Ω is a domain of $\mathbb{R}^{n+1}, n \geq 1, f : [0, +\infty) \to \mathbb{R}$ is a $C^{1,\alpha}$ function and ν stands for the exterior normal unit vector about $\partial\Omega$.

A classical result by Serrin [16,23] states that the existence of a positive solution to the overdetermined problem (1.1) yields that the smooth bounded domain Ω must be a ball. This result has applications in various mathematical and physical problems, such as isoperimetric inequalities, spectral geometry and hydrodynamics (see [4,26,27] for the details).

The case when the domain Ω is supposed to be unbounded is also very interesting. Indeed, overdetermined boundary conditions appear in free boundary problems if the variational structure imposes suitable conditions on the separation interface (see [2,6]). In this process, several methods applied to study the regularity of free boundary problems are based on blow-up techniques that lead to the study of an elliptic problem in an unbounded domain. In this framework, Berestycki, Caffarelli and Nirenberg [5] were concerned with the problem (1.1) in unbounded domains and concluded the following conjecture:

BCN Conjecture. Assume that Ω is a smooth domain with $\mathbb{R}^n \backslash \bar{\Omega}$ connected, then the existence of a bounded positive solution to problem (1.1) for some Lipschitz function f implies that Ω is either a ball, a half-space, a generalized cylinder $B^k \times \mathbb{R}^{n-k}$ (B^k is a ball in $\mathbb{R}^k$), or the complement of one of them.

Such conjecture, in the case of exterior domains, is motivated by the works of Reichel [17], Aftalion and Busca [1]. BCN Conjecture actually has motivated various interesting works. For example, Farina and Valdinoci [11] obtained some natural assumptions to conclude that Ω must be a half-space and u is a function only depending on one variable, when Ω is an epigraph for which the problem (1.1) has a solution. Furthermore, in [18] the BCN conjecture is proved for some classes of nonlinearities f; the work [28] gives a complete classification of solutions to harmonic overdetermined problems in the plane; Ros, Ruiz and Sicbaldi in [19] proved that if $\partial\Omega$ is connected and unbounded in dimension 2, then Ω is a half-plane.

The conjecture has been answered with a counterexample for $n \geq 3$ in [25], where the second author constructed a domain by a periodic perturbation of the straight cylinder $B^n \times \mathbb{R}$ for which there exists a periodic solution to the problem (1.1) for $f(u) = \lambda u, \lambda > 0$. More precisely, such domains, as shown in [22], belong to a 1-parameter family $\{\Omega_s\}_{s\in(-\epsilon,\epsilon)}$ and are given by

$$\Omega_s = \left\{ (x,t) \in \mathbb{R}^n \times \mathbb{R} \,:\, |x| < 1 + s\cos\left(\frac{2\pi}{T_s}\, t\right) + O(s^2) \right\}$$

where ϵ is a small constant, $T_s = T_0 + O(s)$ and T_0 depends only on the dimension n. In [10], Fall, Minlend and Weth provided the same kind of work for $f(u) = 1$. In [8] similar solutions are found for the Allen-Cahn nonlinearity $f(u) = u - u^3$, but in domains that are perturbations of a dilated straight cylinder, i.e. perturbations of $(\epsilon^{-1} B^n) \times \mathbb{R}$ for ϵ small. In addition, Ros, Ruiz and Sicbaldi [20] found a perturbation of the complement of a ball B_R that supports a bounded solution to the problem (1.1), when f is a nonlinear function $f(u) = u^p - u$.

The aim of this paper is to perform such a construction under somewhat minimal assumptions on the nonlinearity $f(u)$. For technical reasons, we need the following assumptions:

Assumption 1: There exists a positive radially symmetric solution $\phi_1 \in C^{2,\alpha}(B)$ of the problem

$$\begin{cases} \Delta\phi_1 + f(\phi_1) = 0 & \text{in } B \\ \phi_1 = 0 & \text{on } \partial B \end{cases} \tag{1.2}$$

with $\partial_\nu(x) \neq 0$ for $x \in \partial B$.

Assumption 2: Define the linearized operator $L_D : C^{2,\alpha}_{0,r}(B) \to C^{0,\alpha}_r(B)$ by

$$L_D(\phi) = \Delta\phi + f'(\phi_1)\phi\,, \tag{1.3}$$

where $C^{2,\alpha}_{0,r}(B)$ and $C^{0,\alpha}_r(B)$ denote the spaces of radial functions in $C^{2,\alpha}_0(B)$ and $C^{0,\alpha}(B)$ respectively. We assume that the linearized operator L_D is non-degenerate; in other words, if $L_D(\phi) = 0$ then $\phi = 0$.

Observe that by [12], any solution ϕ_1 of (1.2) needs to be a radially symmetric function.

We are now in position to state our main result:

Theorem 1.1 *If $n \geq 1$, $f : [0, +\infty) \to \mathbb{R}$ is $C^{1,\alpha}$ and assumptions* 1 *and* 2 *hold, then there exists a positive number T_* and a smooth map*

$$\begin{array}{ccc} (-\epsilon, \epsilon) & \to & C^{2,\alpha}(\mathbb{R}/\mathbb{Z}) \times \mathbb{R} \\ s & \mapsto & (v_s, T_s) \end{array}$$

with $v_0 = 0$ and $T_0 = T_*$ such that the overdetermined problem (1.1) *has a solution in the domain*

$$\Omega_s = \left\{ (x,t) \in \mathbb{R}^n \times \mathbb{R} : |x| < 1 + v_s\left(\frac{t}{T_s}\right) \right\}.$$

The solution $u = u_s$ of problem (1.1) *is T_s-periodic in the variable t and hence bounded. Moreover*

$$\int_0^1 v_s(t)\, dt = 0$$

and

$$v_s(t) = s\, \cos(2\pi t) + O(s^2)\,.$$

As a consequence, for all functions f satisfying assumptions 1 and 2 we produce a counterexample to the BCN conjecture diffeomorphic to a cylinder. Assumptions 1 and 2 hold for example in the following cases among many others:

(1) If $f(0) > 0$ and $f'(s) < \lambda_1$ for any $s \in (0, +\infty)$, where λ_1 is the first eigenvalue of the Dirichlet Laplacian in the unit ball of $\mathbb{R}^n$.

(2) If $f(u) = u^p - u$, $1 < p < \frac{n+2}{n-2}$ if $n > 2$, see [15].

(3) If $f(u) = \lambda e^u$ and $\lambda \in (0, \lambda^*)$, $\lambda^* > 0$ receives the name of extremal value, see for instance [9].

Obviously, our theorem covers the result in [10] and is complementary to the results in [22, 25].

2. Some details

The operator L_D defined in Assumption 2 has a diverging sequence of eigenvalues γ_{D_j}, hence there are only a finite number l of them which are negative, i.e.

$$\gamma_{D_1} < \gamma_{D_2} < \cdots < \gamma_{D_l} < 0,\ \gamma_{D_{l+1}} > 0.$$

Actually, these eigenvalues γ_{D_j} are all simple.

Let $z_j \in C^{2,\alpha}_{0,r}(B)$ (normalized by $\|z_j\|_{L^2} = 1$) be the eigenfunctions corresponding to the eigenvalues γ_{D_j}, i.e.

$$\begin{cases} \Delta z_j + f'(\phi_1) z_j + \gamma_{D_j} z_j = 0 & \text{in } B \\ z_j = 0 & \text{on } \partial B \end{cases}. \tag{2.1}$$

As is well known, the operator L_D is related to the quadratic form

$$Q_D : H^1_{0,r}(B) \to \mathbb{R}, \;\; Q_D(\phi) := \int_B \left(|\nabla \phi|^2 - f'(\phi_1)\phi^2 \right).$$

The first eigenvalue of L_D is given by

$$\gamma_{D_1} = \inf \left\{ Q_D(\phi) : \|\phi\|_{L^2(B)} = 1 \right\}.$$

We also define the quadratic form

$$Q : H^1_r(B) \to \mathbb{R}, \;\; Q(\psi) := \int_B \left(|\nabla \psi|^2 - f'(\phi_1)\psi^2 \right) + c\, \omega_n \psi(1)^2,$$

where ω_n is the area of $\mathbb{S}^{n-1}$ and $c = -\phi_1''(1) = n - 1 + \frac{f(0)}{\phi_1'(1)}$.

Observe that,

$$Q|_{H^1_{0,r}(B)} = Q_D.$$

Analogously, we can define

$$\gamma_1 = \inf \left\{ Q(\psi) : \|\psi\|_{L^2(B)} = 1 \right\}. \tag{2.2}$$

It is rather standard to show that γ_1 is achieved by the minimizer ψ_1, and that γ_1 is simple, so ψ_1 is uniquely determined up to a sign. In addition, there holds: $\gamma_1 < \min\{0, \gamma_{D_1}\}$. In fact, it is evident that $\gamma_1 \leq \gamma_{D_1}$ from the variational characterization of the eigenvalues. The strict inequality follows because of the uniqueness of solutions of Initial Vale Problems for ODEs (see [21] for details).

Next, we will consider the Dirichlet problem for the linearized equation in a straight cylinder for periodic functions, namely,

$$\begin{cases} \Delta \psi + f'(\phi_1)\psi = 0 & \text{in } B \times \mathbb{R} \\ \psi(x) = 0 & \text{on } (\partial B) \times \mathbb{R} \end{cases} \tag{2.3}$$

where $\psi(x,t)$ is T-periodic in the variable t.

Define:

$$C_1^T = B \times \mathbb{R}/T\mathbb{Z}.$$

Hence (2.3) is just the linearization of the problem:

$$\begin{cases} \Delta \phi + f(\phi) = 0 & \text{in } C_1^T \\ \phi = 0 & \text{on } \partial C_1^T \end{cases}. \tag{2.4}$$

If ϕ_1 is the solution of Problem (1.2), then the function $\phi_1(x,t) = \phi_1(x)$ (we use a natural abuse of notation) solves (2.4). Define the linearized operator $L_D^T : C^{2,\alpha}_{0,r}(C_1^T) \to C^{\alpha}_{r}(C_1^T)$ (associated to Problem (2.4)) by

$$L_D^T(\phi) = \Delta \phi + f'(\phi_1)\phi,$$

and consider the eigenvalue problem

$$L_D^T(\phi) + \tau \phi = 0.$$

Then the functions $z_j(x,t) = z_j(x)$ from (2.1) solve the problem

$$\begin{cases} \Delta z_j + f'(\phi_1) z_j + \tau_j z_j = 0 & \text{in } C_1^T \\ z_j = 0 & \text{on } \partial C_1^T \end{cases}.$$

Let us define the quadratic form $Q_D^T : H^1_{0,r}(C_1^T) \to \mathbb{R}$ related to L_D^T,

$$Q_D^T(\psi) := \int_{C_1^T} \left(|\nabla \psi|^2 - f'(\phi_1)\psi^2 \right).$$

We will also need to define the quadratic form $Q^T : H^1_r(C^T_1) \to \mathbb{R}$,

$$Q^T(\psi) := \int_{C^T_1} \left(|\nabla\psi|^2 - f'(\phi_1)\psi^2\right) + c\int_{\partial C^T_1} \psi^2.$$

In next proposition we study the behavior of these quadratic forms:

Proposition 2.1 *Define*:

$$\alpha = \inf\left\{Q^T_D(\psi) :\ \psi \in H^1_{0,r}(C^T_1),\ \|\psi\|_{L^2} = 1,\ \int_{C^T_1} \psi\, z_j = 0,\ j = 1, \dots l.\right\},$$

$$\beta = \inf\left\{Q^T(\psi) :\ \psi \in H^1_r(C^T_1),\ \|\psi\|_{L^2} = 1,\ \int_{\partial C^T_1} \psi = 0,\ \int_{C^T_1} \psi\, z_j = 0,\ j = 1, \dots l.\right\},$$

then

$$\alpha = \min\left\{\gamma_{D_{l+1}}, \gamma_{D_1} + \frac{4\pi^2}{T^2}\right\}, \quad \beta = \min\left\{\gamma_{D_{l+1}}, \gamma_1 + \frac{4\pi^2}{T^2}\right\}.$$

Moreover, those infima are achieved. If $\beta = \gamma_1 + \frac{4\pi^2}{T^2}$, the minimizer is equal to

$$\psi_1(x)\cos\left(\frac{2\pi}{T}(t+\delta)\right),$$

where ψ_1 is the minimizer for (2.2) *and $\delta \in [0,1]$.*

Proof Just by defining $\bar{\psi}(x) = \int_0^T \psi(x,t)dt$ and the Poincaré-Wirtinger inequality, see [21] for details. □

Corollary 2.2 *Define $\overline{T}$ as*:

$$\overline{T} = \begin{cases} \frac{2\pi}{\sqrt{-\gamma_{D_1}}} & \text{if } \gamma_{D_1} < 0, \\ +\infty & \text{if } \gamma_{D_1} > 0. \end{cases} \tag{2.5}$$

Then, for $T \in (0, \overline{T})$, we have that $Q^T_D(\psi) > 0$ for any $\psi \in H^1_{0,r}(C^T_1)$ such that $\int_{C^T_1} \psi z_j = 0$, $j = 1, 2, \cdots, l$. As a consequence, L^T_D is nondegenerate.

Defining the cylinder-type domain

$$C^T_{1+v} = \left\{(x,t) \in \mathbb{R}^n \times \mathbb{R}/\mathbb{Z} \ :\ 0 \le |x| < 1 + v\left(\frac{t}{T}\right)\right\},$$

we start with the following result, that allows us to obtain a solution for the Dirichlet problem in the domain C^T_{1+v} and its smooth dependence on T and v.

Proposition 2.3 *Assume that $T < \overline{T}$, where $\overline{T}$ is given by* (2.5). *Then, for all $v \in C^{2,\alpha}_e(\mathbb{R}/\mathbb{Z})$ whose norm is sufficiently small, the problem*

$$\begin{cases} \Delta\phi + f(\phi) = 0 & \text{in } C^T_{1+v} \\ \phi = 0 & \text{on } \partial C^T_{1+v} \end{cases} \tag{2.6}$$

has a unique positive solution $\phi = \phi_{1+v,T} \in C^{2,\alpha}(C^T_{1+v})$. Moreover, ϕ depends smoothly on the function v, and $\phi = \phi_1$ when $v \equiv 0$.

Proof Following the nondegeneracy of the Dirichlet problem, please refer to [21] for details. □

For any $T < \overline{T}$, there exists a neighborhood $\mathcal{U}$ of 0 in $C^{2,\alpha}_{e,m}(\mathbb{R}/\mathbb{Z})$ where the following Dirichlet-to-Neumann operator is well defined and C^1:

$$G : \mathcal{U} \times (0, \overline{T}) \to C^{1,\alpha}_{e,m}(\mathbb{R}/\mathbb{Z}),$$

$$G(v,T)(t) = \left.\frac{\partial\phi_{1+v,T}}{\partial\nu}\right|_{\partial C^T_{1+v}}(T\,t) - \frac{1}{\mathrm{Vol}(\partial C^T_{1+v})}\int_{\partial C^T_{1+v}} \frac{\partial\phi_{1+v,T}}{\partial\nu}, \tag{2.7}$$

where $\phi(v,T)$ is the solution of (2.6) verified by Proposition 2.3.

We will next compute the Fréchet derivative of the operator G. For so, we will need the following lemmas.

Lemma 2.4 *Assume that $T < \overline{T}$, where $\overline{T}$ is given by* (2.5)*. Then for all $v \in C_e^{2,\alpha}(\mathbb{R}/\mathbb{Z})$, there exists a unique solution $\psi_{v,T}$ to the problem*

$$\begin{cases} \Delta\psi_{v,T} + f'(\phi_1)\psi_{v,T} = 0 & \text{in } C_1^T \\ \psi_{v,T} = v(\cdot/T) & \text{on } \partial C_1^T \end{cases}. \tag{2.8}$$

Proof Let $\psi_0(x,t) \in C^{2,\alpha}(C_1^T)$ such that $\psi_0|_{\partial C_1^T} = v(\cdot/T)$. If we set $\omega = \psi_{v,T} - \psi_0$, the problem (2.8) is equivalent to the problem

$$\begin{cases} \Delta\omega + f'(\phi_1)\omega = -\Big(\Delta\psi_0 + f'(\phi_1)\psi_0\Big) & \text{in } C_1^T \\ \omega = 0 & \text{on } \partial C_1^T \end{cases}.$$

Observe that the right hand side of the above equation is in $C_r^{\alpha}(C_1^T)$. Recall the Corollary 2.2, L_D^T is nondegenerate. Hence it is a bijection and the result follows. □

Lemma 2.5 *Let $v \in C_{e,m}^{2,\alpha}(\mathbb{R}/\mathbb{Z})$ and $\psi_v = \psi_{v,T} \in C_r^{2,\alpha}(C_1^T)$ be the solution of* (2.8)*. Then*

$$\int_{C_1^T} \psi_v z_j = 0, \qquad \int_{\partial C_1^T} \frac{\partial \psi_v}{\partial \nu} = 0\,, \qquad j = 1, 2, \cdots, l\,.$$

Proof We can get these results by the straight computation, refer to [21]. □

For $T < \overline{T}$ we can define the linear and continuous operator $H_T : C_{e,m}^{2,\alpha}(\mathbb{R}/\mathbb{Z}) \to C_{e,m}^{1,\alpha}(\mathbb{R}/\mathbb{Z})$ by

$$H_T(v)(t) = \partial_\nu \psi_v(Tt) + c\, v,$$

and $\psi_v = \psi_{v,T}$ as in Lemma 2.4. We present some properties of H_T.

Lemma 2.6 *For any $T < \overline{T}$, the operator*

$$H_T : C_{e,m}^{2,\alpha}(\mathbb{R}/\mathbb{Z}) \to C_{e,m}^{1,\alpha}(\mathbb{R}/\mathbb{Z})$$

is a linear essentially self-adjoint operator and has closed range. Moreover, it is also a Fredholm operator of index zero.

Proof By the straight computation, we can get that the operator H_T is a linear essentially self-adjoint operator. And the rest results follow from [3, 14]. More details refer to [21]. □

We show now that the linearization of the operator G with respect to v at $v = 0$ is given by H_T, up to a constant.

Proposition 2.7 *The map G is C^1, and $D_v(G)|_{v=0} = -\phi_1'(1)\, H_T$.*

Proof By the Proposition 2.3 (the function $\phi(v,T)$ depends smoothly on v), the operator G is C^1. The linear operator obtained by the directional derivative of linearizing G with respect to v, computed at (v,T), is given by

$$G'(w) = \lim_{s\to 0} \frac{G(sw,T) - G(0,T)}{s} = \lim_{s\to 0} \frac{G(sw,T)}{s}.$$

Let $v = sw$, for $y \in \mathbb{R}^n$ and $t \in \mathbb{R}$, we consider the parameterization of C_{1+v}^T given by

$$Y(y,t) := \left(\left(1 + v\left(\frac{t}{T}\right)\right) y, t\right).$$

Let g be the induced metric such that $\hat{\phi} = Y^*\phi$ (smoothly depending on the real parameter s) solves the problem

$$\begin{cases} \Delta_g \hat{\phi} + f(\hat{\phi}) = 0 & \text{in } C_1^T \\ \hat{\phi} = 0 & \text{on } \partial C_1^T \end{cases}.$$

We remark that $\hat{\phi}_1 = Y^*\phi_1$ is the solution of

$$\Delta_g \hat{\phi}_1 + f(\hat{\phi}_1) = 0$$

in C_1^T, and

$$\hat{\phi}_1(y,t) = \phi_1\big((1+sw)y, t\big)$$

on ∂C_1^T. Let $\hat{\phi} = \hat{\phi}_1 + \hat{\psi}$, we can get that

$$\begin{cases} \Delta_g \hat{\psi} + f(\hat{\phi}_1 + \hat{\psi}) - f(\hat{\phi}_1) = 0 & \text{in } C_1^T \\ \hat{\psi} = -\hat{\phi}_1 & \text{on } \partial C_1^T \end{cases}. \tag{2.9}$$

Obviously, $\hat{\psi}$ is a smooth function of s. When $s = 0$, we have $\phi = \phi_1$. Then, $\hat{\psi} = 0$ and $\hat{\phi}_1 = \phi_1$ as $s = 0$. We set

$$\dot{\psi} = \partial_s \hat{\psi}|_{s=0}.$$

Differentiating (2.9) with respect of s and evaluating the result at $s = 0$, we have

$$\begin{cases} \Delta\dot{\psi} + f'(\phi_1)\dot{\psi} = 0 & \text{in } C_1^T \\ \dot{\psi} = -\phi_1'(1)w & \text{on } \partial C_1^T \end{cases}$$

where $r := |y|$. Then $\dot{\psi} = -\phi_1'(1)\,\psi_w$ where ψ_w is as given by Lemma 2.4 (with $\bar{v} = w$). Then, we can write

$$\hat{\phi}(x,t) = \hat{\phi}_1(x,t) + s\dot{\psi}(x,t) + O(s^2).$$

In particular, in a neighborhood of ∂C_1^T we have

$$\begin{aligned} \hat{\phi}(y,t) &= \phi_1\big((1+sw)y, t\big) + s\dot{\psi}(y,t) + O(s^2) \\ &= \phi_1(y,t) + s\big(wr\partial_r\phi_1 + \dot{\psi}(y,t)\big) + O(s^2). \end{aligned}$$

In order to complete the proof of the result, it is enough to calculate the normal derivation of the function $\hat{\phi}$ when the normal is calculated with respect to the metric g. By using cylindrical coordinates $(y,t) = (rz,t)$ where $r > 0$ and $z \in \mathbb{S}^{n-1}$, then the metric g can be expanded in C_1^T as

$$g = (1+sw)^2 dr^2 + 2srw'(1+sw)drdt + \big(1 + s^2r^2(w')^2\big)dt^2 + r^2(1+sw)^2 \overset{\circ}{h}$$

where $\overset{\circ}{h}$ is the metric on $\mathbb{S}^{n-1}$ induced by the Euclidean metric. It follows from this expression that the unit normal vector fields to ∂C_1^T for the metric g is given by

$$\hat{v} = \big((1+sw)^{-1} + O(s^2)\big)\partial_r + O(s)\partial_t.$$

By this, we conclude that

$$g(\nabla\hat{\phi}, \hat{v}) = \partial_r\phi_1 + s\big(w\partial_r^2\phi_1 + \partial_r\dot{\psi}\big) + O(s^2)$$

on ∂C_1^T. From the fact that $\partial_r\phi_1$ is constant and the fact that the term $w\partial_r^2\phi_1 + \partial_r\dot{\psi}$ has mean 0 on ∂C_1^T we obtain

$$G'(w) = \partial_r\dot{\psi} + \phi_1''(1)\,w = -\phi_1'(1)\,\partial_r\psi_w + \phi_1''(1)\,w = -\phi_1'(1)\,H_T(w).$$

This concludes the proof of the result. □

We now define the first eigenvalue of the operator H_T as

$$\sigma(T) = \inf\left\{ \int_0^1 H_T(v)v : v \in C^{2,\alpha}_{e,m}(\mathbb{R}/\mathbb{Z}),\ \int_0^1 v^2 = 1 \right\}.$$

By the Divergence formula, we have

$$Q^T(\psi_v) = T\omega_n \int_0^1 H_T(v)v.$$

Next lemma characterizes the eigenvalue $\sigma(T)$ in terms of the quadratic form Q^T.

Lemma 2.8 *For any $T < \overline{T}$, we have*

$$\sigma(T) = \min\left\{ \frac{1}{T}Q^T(\psi) : \psi \in E,\ \int_{\partial C_1^T} \psi^2 = 1 \right\},$$

where

$$E = \left\{ \psi \in H_r^1(C_1^T) :\ \int_{\partial C_1^T} \psi = 0, \int_{C_1^T} \psi z_j = 0,\ j = 1, \dots l \right\}. \tag{2.10}$$

Moreover, the infimum is attained.

Proof Define $\mu_1 := \inf\left\{Q_D^T(\psi) : \psi \in E, \int_{\partial C_1^T} \psi^2 = 1\right\} \in [-\infty, +\infty)$. We show that μ_1 is achieved by contradiction. Then we can get $\int_0^1 v^2 = \frac{1}{T\omega_n}$, $J_T(v) = \frac{1}{T\omega_n} Q^T(\psi) = \frac{1}{T\omega_n}\mu_1$, refer to [21]. □

We are now in position to prove the following useful result:

Proposition 2.9 *There exists a real positive number $T_* = \frac{2\pi}{\sqrt{-\gamma_1}} < \overline{T}$, then*

(*i*) *if $T < T_*$, then $\sigma(T) > 0$;*

(*ii*) *if $T = T_*$, then $\sigma(T) = 0$;*

(*iii*) *if $T > T_*$, then $\sigma(T) < 0$.*

Moreover, $\mathrm{Ker}(H_{T_*}) = \mathbb{R}\cos(2\pi t)$. *In particular,* $\dim \mathrm{Ker}(H_{T_*}) = 1$.

Proof It follows from Lemma 2.8 and Proposition 2.1, taking into account that $C^{2,\alpha}_{e,m}(\mathbb{R}/\mathbb{Z})$ contains only even functions. □

Now, we are ready to prove that the operator G satisfies the hypotheses of the Crandall-Rabinowitz bifurcation theorem (see [7, 13, 24]). And then, Theorem 1.1 follows immediately from the following proposition and the Crandall-Rabinowitz theorem.

Proposition 2.10 *There exists a real number T_* such that the linearized operator $D_v G(0, T_*)$ has 1-dimensional kernel and can be spanned by the function $v_0 = \cos(2\pi t)$,*

$$\mathrm{Ker}\, D_v G(0, T_*) = \mathbb{R} v_0.$$

The cokernel of $D_v G(0, T_*)$ is also 1-dimensional, and

$$D_T D_v G(0, T_*)(v_0) \notin \mathrm{Im}\, D_v G(0, T_*).$$

Proof Recall from the Proposition 2.7, we know that $D_v G(0, T_*) = -\phi_1'(1)\phi_1 H_{T_*}$. Then we have

$$\mathrm{Im}\, D_v G(0, T_*) = \mathrm{Im}\, H_{T_*}.$$

By the Proposition 2.9, we have that the kernel of the linearized operator $D_v G(0, T_*)$ has dimension 1 and can be spanned by the function $v_0 = \cos(2\pi t)$,

$$\mathrm{Ker}\, D_v G(0, T_*) = \mathbb{R}\, v_0.$$

Then, $\mathrm{codim}\, \mathrm{Im}\, (H_{T_*}) = 1$ follows from the fact that H_T is a Fredholm operator of index zero by Lemma 2.6.

Here, we are ready to prove $D_T D_v G(0, T_*)(v_0) \notin \mathrm{Im}\, D_v G(0, T_*)$. Taking $\xi \in \mathrm{Im}\, D_v G(0, T_*) = \mathrm{Im}\, (H_{T_*})$, $\xi = H_{T_*}(v)$, then we have

$$\int_0^1 \xi v_0 = \int_0^1 H_{T_*}(v) v_0 = \int_0^1 H_{T_*}(v_0) v = 0,$$

because of the fact $H_{T_*}(v_0) = 0$. We have

$$\mathrm{Im}\, (H_{T_*}) = \left\{\xi : \int_0^1 \xi v_0 = 0\right\}.$$

Notice that $D_T D_v G(0, T_*)(v_0) = -\phi_1'(1) D_T|_{T=T_*} H_T(v_0)$, then, in order to prove $D_T D_v G(0, T_*)(v_0) \notin \mathrm{Im}\, D_v G(0, T_*)$, we just need to prove that

$$\int_0^1 \left(D_T|_{T=T_*} H_T(v_0)\right) v_0 \neq 0.$$

Actually,

$$\begin{aligned}\int_0^1 \left(D_T|_{T=T_*} H_T(v_0) v_0\right) &= \frac{d}{dT}\Big|_{T=T_*} \int_0^1 H_T(v_0) v_0 = \frac{1}{\omega_n}\frac{d}{dT}\Big|_{T=T_*} \left(\frac{1}{T} Q^T(\psi_{v_0}, \psi_{v_0})\right) \\ &= \frac{1}{\omega_n}\frac{d}{dT}\Big|_{T=T_*} \left(\frac{1}{2} Q(\psi_1, \psi_1) + \frac{2\pi^2}{T^2}\int_B \psi_1^2\right) = -\frac{4\pi^2}{\omega_n T_*^3}\int_B \psi_1^2 \neq 0,\end{aligned}$$

where the third equality is given by the straight computation of $Q^T(\psi, \psi)$ with the function $\psi_{v_0}(x, t) = \psi_1(x)\cos(\frac{2\pi t}{T})$. □

Acknowledgements

The author was supported by Junta de Andalucía Grant FQM116.

References

[1] Amandine Aftalion, Jérôme Busca. Radial symmetry of overdetermined boundary-value problems in exterior domains. *Arch. Rational Mech. Anal.*, 143:195–206, 1998.

[2] Hans Wilhelm Alt, Luis A. Caffarelli. Existence and regularity for a minimal problem with free boundary. *J. Reine Angew. Math.*, 325:105–144, 1981.

[3] Wolfgang Arendta, Antonious F.M. ter Elst, James B. Kennedya, Manfred Sauterb. The Dirichlet-to-Neumann operator via hidden compactness. *J. Functional Analysis*, 266:1757-1786, 2014.

[4] Catherine Bénéteau, Dmitry Khavinson. The isoperimetric inequality via approximation theory and free boundary problems. *Computational Methods and Function Theory*, 6:253–274, 2006.

[5] Henri Berestycki, Luis A. Caffarelli, Louis Nirenberg. Monotonicity for elliptic equations in undounded Lipschitz domains. *Comm. Pure Appl. Math.*, 50:1089–1111, 1997.

[6] Luis A. Caffarelli, David Jerison, Carlos Kenig. Global energy minimizers for free boundary problems and full regularity in three dimension. *Contemp. Math. Amer. Math. Soc.*, 350:83–97, 2004.

[7] Michael Crandall, Paul Rabinowitz. Bifurcation for simple eigenvalue. *J. Functional Analysis*, 8:321–340, 1983.

[8] Manuel Del Pino, Frank Pacard, Junchen Wei. Serrin's overdetermined problem and constant mean curvature surfaces. *Duke Math. J.*, 64:2643–2722, 2015.

[9] Louis Dupaigne. Stable solutions of elliptic partial differential equations. Chapman-Hall/CRC Monographs and Surveys in Pure and Applied Mathematics, 143, 2011.

[10] Mouhamed Moustapha Fall, Ignace Aristide Minlend, Tobias Weth. Unbounded periodic solutions to Serrin's overdetermined boundary value problem. *Arch. Ration. Mech. Anal.*, 223:737–759, 2017.

[11] Alberto Farina, Enrico Valdinoci. Flattening results for elliptic PDEs in unbounded domains with applications to overdetermined problems. *Arch. Ration. Mech. Anal.*, 195:1025–1058, 2010.

[12] Basilis Gidas, Wei Ming Ni, Louis Nirenberg. Symmetry and related properties via the maximum principle. *Comm. Math. Phys.* 68(3):209–243, 1979.

[13] Hansjörg Kielhöfer. Bifurcation Theory: An Introduction with Applications to PDEs, *Appl. Math. Sci.*, 156, Springer-Verlag, New York 2004.

[14] Carlos S. Kubrusly. Fredholm Theory in Hilbert Space-A Concise Introductory Exposition. *Bull. Belg. Math. Soc. Simon Stevin*, 15:153-177, 2008.

[15] Man Kam Kwong. Uniqueness of positive solutions of $\Delta u - u + u^p = 0$ in $\mathbb{R}^n$. *Arch. Rational Mech. Anal.*, 105(3):243–266, 1989.

[16] Patrizia Pucci, James Serrin. The maximum principle. *Progr. Nonlinear Differential Equations Appl.*, 73, Birkhauser, Basel 2007.

[17] Wolfgang Reichel. Radial symmetry for elliptic boundary-value problems on exterior domains. *Arch. Ration. Mech. Anal.*, 137:381–394, 1997.

[18] Antonio Ros, Pieralberto Sicbaldi. Geometry and Topology of some overdetermined elliptic problem. *J. Differential Equations*, 255(5):951–977, 2013.

[19] Antonio Ros, David Ruiz, Pieralberto Sicbaldi. A rigidity for overdetermined elliptic problems in the plane. *Comm. Pure Appl. Math.*, 70:1223–1252, 2017.

[20] Antonio. Ros, David Ruiz, Pieralberto Sicbaldi. Solutions to overdetermined elliptic problems in nontrivial exterior domains. *J. Eur. Math. Soc.*, 22:253–281, 2020.

[21] David Ruiz, Pieralberto Sicbaldi, Jing Wu. Overdetermined elliptic problems in onduloid-type domains with general nonlinearities. preprint.

[22] Felix Schlenk, Pieralberto Sicbaldi. Bifurcating extremal domains for the first eigenvalue of the Laplacian. *Adv. Math.*, 229:602–632, 2012.

[23] James Serrin. A symmetry problem in potential theory. *Arch. Rational Mech. Anal.*, 43:304–318, 1971.

[24] Joel Smoller. Shock waves and reaction-diffusion equations. Second edition, Grundlehren der Mathematischen Wissenschaften 258, Springer-Verlag, New York, 1994.

[25] Pieralberto Sicbaldi. New extremal domains for the first eigenvalue of the Laplacian in flat tori. *Calc. Var. Partial Differential Equations*, 37:329–344, 2010.

[26] Boyan Sirakov. Overdetermined elliptic problems in physics. Nonlinear PDEs in Condensed Matter and Reactive Flows, Kluwer, 273–295, 2002.

[27] Ivan Stephen Sokolnikoff. Mathematical theory of elasticity. McGraw-Hill, New York, 1956.

[28] Martin Traizet. Classification of the solutions to an overdetermined elliptic problem in the plane. *Geom. Func. An.*, 24(2):690–720, 2014.

XXVI Congreso de Ecuaciones Diferenciales y Aplicaciones
XVI Congreso de Matemática Aplicada
Gijón, 14-18 junio 2021
(pp. 352–358)

A method to construct irreducible totally nonnegative matrices with a given Jordan canonical form

Begoña Cantó[1], Rafael Cantó[1], Ana M. Urbano[1]
Institut de Matemática Multidisciplinar, Universitat Politècnica de València, 46071 València, Spain.

Abstract

Let $A \in \mathbb{R}^{n\times n}$ be an irreducible totally nonnegative matrix (ITN), that is, A is irreducible with all its minors nonnegative. A triple (n, r, p) is called *realizable* if there exists an ITN matrix $A \in \mathbb{R}^{n\times n}$ with $\text{rank}(A) = r$ and $p\text{-rank}(A) = p$ (recall that p-rank(A) is the size of the largest invertible principal submatrix of A). Each ITN matrix A associated with a realizable triple (n, r, p) has p positive and distinct eigenvalues, and for the zero eigenvalue it is verified that $n - r$ and $n - p$ are the geometric and the algebraic multiplicity, respectively. Moreover, since $\text{rank}(A^p) = p$, A has $n - r$ zero Jordan blocks whose sizes are given by the Segre characteristic, $S = (s_1, s_2, \ldots, s_{n-r})$, with $s_i \leq p$, $i = 1, 2, \ldots, n - r$.

We know the number of zero Jordan canonical forms of ITN matrices associated with a realizable triple (n, r, p) and all these zero Jordan canonical forms. The following important question that we present in this talk deals with how to construct an ITN matrix A associated with (n, r, p) and exactly with one of these Segre characteristic S corresponding to the zero eigenvalue.

1. Introduction

A matrix $A \in \mathbb{R}^{n\times n}$ is called totally nonnegative if all its minors are nonnegative and it is abbreviated as TN. The wide study of these matrices is due to the large number of applications in different branches of science, see for instance [1, 7–18]. Now, we recall some basic concepts that we will use throughout the paper:

1. The rank of A, denoted by $\text{rank}(A)$, is the size of the largest invertible square submatrix of A. The principal rank of A, denoted by $p\text{-rank}(A)$, is the size of the largest invertible principal submatrix of A. It is clear that

$$0 \leq p\text{-rank}(A) \leq \text{rank}(A) \leq n$$

2. The characteristic polynomial of a matrix A is given by

$$q_A(\lambda) = \det(\lambda I - A) = \lambda^n + \sum_{k=1}^{n} (-1)^k \left(\sum_{\forall \alpha \in Q(\text{k,n})} \det(A[\alpha]) \right) \lambda^{n-k}$$

where $Q_{(\text{k,n})}$ denotes the set of all increasing sequences of k natural numbers less than or equal to n, for $k, n \in \mathbb{N}$, $1 \leq k \leq n$, see [1]. If $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k) \in Q_{k,n}$ and $\beta = (\beta_1, \beta_2, \ldots, \beta_k) \in Q_{k,n}$, $A[\alpha|\beta]$ denotes the $k \times k$ submatrix of A lying in rows α_i and columns β_i, $i = 1, 2, \ldots, k$. The principal submatrix $A[\alpha|\alpha]$ is abbreviated as $A[\alpha]$.

If A is TN and $p\text{-rank}(A) = p$, the minors of the same order have the same sign or are zero, then there are no cancelations in the summands and then,

$$\begin{aligned} q_A(\lambda) &= \lambda^{n-p} \left(\lambda^p + \textstyle\sum_{k=1}^{p} (-1)^k \left(\sum_{\forall \alpha \in Q(\text{k,n})} \det(A[\alpha]) \right) \lambda^{p-k} \right) \\ &= \lambda^{n-p} \left(\lambda^p - c_1 \lambda^{p-1} + \cdots + (-1)^p c_p \right). \end{aligned}$$

Then, if A is a TN matrix with $\text{rank}(A) = r$ and $p\text{-rank}(A) = p$, has p nonzero eigenvalues and the algebraic and geometric multiplicities of the zero eigenvalue are equal to $n - p$ and $n - r$, respectively.

3. A matrix $A \in \mathbb{R}^{n\times n}$, with $n \geq 2$, is an *irreducible* matrix if there is not a permutation matrix P such that $PAP^T = \begin{bmatrix} B & C \\ O & D \end{bmatrix}$, where O is an $(n - r) \times r$ zero matrix $(1 \leq r \leq n - 1)$. If $n = 1$, $A = (a)$ is irreducible when $a \neq 0$.

Fallat, Gekhtman and Johnson in [8] characterize the irreducible TN matrices as follows: a TN matrix $A = [a_{ij}] \in \mathbb{R}^{n\times n}$ is irreducible if and only if $a_{ij} > 0$ for all i, j such that $|i - j| \leq 1$ and they represent this class of matrices by ITN.

4. If there exists an ITN matrix $A \in \mathbb{R}^{n\times n}$ with $\text{rank}(A) = r$ and p-$\text{rank}(A) = p$, then the triple (n, r, p) is called *realizable* [8, p. 709], and A is considered as an ITN matrix *associated with* the triple (n, r, p). In order to a triple (n, r, p) be realizable it is necessary that $p \leq r \leq n - \left\lceil \frac{n-p}{p} \right\rceil$.

5. If A is an associated matrix with a realizable triple (n, r, p) then, its p nonzero eigenvalues are positive and distinct ([8, Theorem 3.3]). That is, if $\lambda_1, \ldots, \lambda_p, \ldots \lambda_n$ are the eigenvalues of A, we have

$$\lambda_1 > \lambda_2 > \ldots > \lambda_p > 0, \text{ and } \lambda_{p+1} = \lambda_{p+2} = \ldots = \lambda_n = 0, \tag{1.1}$$

   Moreover, since the algebraic multiplicity of the zero eigenvalue is $n - p$ and $\text{rank}(A^p) = p$, the size of the zero Jordan blocks of A is at most p.

Taking into account the above results, given a realizable triple (n, r, p) the following questions arises in a natutal way:

First question: *How many different zero Jordan canonical forms are associated with a realizable triple* (n, r, p)*?*

As we have seen, the ITN matrices associated with a realizable triple (n, r, p) verify that the algebraic multiplicity of the zero eigenvalue is $n-p$, the geometric one is $n-r$ and the size of the Jordan blocks is maximum p. Therefore, this problem is equivalent to the following *in how many ways can we distribute $n-p$ marbles in $n-r$ bags, knowing that all bags must have at least one marble and that at most each bag will fit p marbles.*

In [6], by using Number Theory, the authors calculated this number (represented by $p_{n-r}^{(p)}(n-p)$) and they gived an algorithm to obtain it. For example, if we have the triple realizable $(19, 14, 8)$ applying this algorithm we have that $p_5^{(8)}(11) = 10$.

Second question: Since we konw the number of different zero Jordan canonical forms associated with a realizable triple (n, r, p), then *what are these zero Jordan forms?*

In [6], using properties and the full rank LU factorization of ITN matrices and the Flanders Theorem the authors give and Procedure and the corresponding algorithm to compute the specific different zero Jordan canonical forms. For example, for the realizable triple $(19, 14, 8)$ we have obtained that there are 10 different zero Jordan canonical forms and applying the new algorithm we obtain these specific zero Jordan structures, all of them have 5 zero Jordan blocks of differents sizes. These structures are,

$$\begin{array}{ccccc}
7 & 1 & 1 & 1 & 1 \\
6 & 2 & 1 & 1 & 1 \\
5 & 3 & 1 & 1 & 1 \\
5 & 2 & 2 & 1 & 1 \\
4 & 4 & 1 & 1 & 1 \\
4 & 3 & 2 & 1 & 1 \\
4 & 2 & 2 & 2 & 1 \\
3 & 3 & 3 & 1 & 1 \\
3 & 3 & 2 & 2 & 1 \\
3 & 2 & 2 & 2 & 2
\end{array}$$

Remark 1.1 The sizes of the zero Jordan blocks of a matrix A are known as the Segre characteristic of A relative to its zero eigenvalue. Given an ITN matrix A associated to a realizable triple (n, r, p), if we represent this Segre sequence by $S = (s_1, s_2, \ldots, s_{n-r})$ then, it is satisfied that

$$\begin{array}{ll}
(1) & s_1 \leq \min\{r - p + 1, p\} \\
(2) & s_i \leq s_{i-1}, \ \ i = 2, 3, \ldots, n - r \\
(3) & \sum_{i=1}^{n-r} s_i = n - p
\end{array} \tag{1.2}$$

Associated to the Segre characteristic $S = (s_1, s_2, \ldots, s_{n-r})$ we have the Weyr characteristic of A relative to the zero eigenvalue $W = (w_1, w_2, \ldots, w_{s_1})$, where $w_i = \text{Car}\{k \ : \ s_k \geq i\}$ for $i = 1, 2, \ldots, s_1$ and

$$\begin{array}{ll}
(1) & w_1 = \dim \text{Ker}(A) = n - r \\
(2) & w_i \leq w_{i-1}, \ \ i = 2, 3, \ldots, s_1 \\
(3) & \sum_{j=1}^{i} w_j = \dim \text{Ker}(A^i) \\
(4) & \sum_{j=1}^{s_1} w_j = \dim \text{Ker}(A^{s_1}) = n - p
\end{array} \tag{1.3}$$

Thirth question: Finally, knowing the number of zero Jordan canonical forms and the specific structures associated with a realizable triple (n, r, p), the following question is the main goal of this work, *how to construct an ITN matrix associated with a realizable triple* (n, r, p) *and with* $n - r$ *zero Jordan blocks whose sizes are given by the Segre characteristic* $S = (s_1, s_2, \ldots, s_{n-r})$ *satisfying (1.2).*

To answer this question, in the next section we first described a procedure that allow us to construct an upper block echelon matrix $U \in \mathbb{R}^{n\times n}$, with $\operatorname{rank}(U) = r$, p-$\operatorname{rank}(U) = p$ and $n - r$ zero Jordan blocks whose sizes are given by the Segre characteristic S satisfying (1.2). After that, from U we will obtain the desired ITN matrix A associated with the realizable triple (n, r, p) and with the same zero Jordan structure that U as $A = LU$, where L is a lower triangular matrix with all its nonzero entries equal to 1.

2. Constructing an upper block echelon TN matrix U with a zero Jordan canonical form

In this section we describe a procedure to construct an upper block echelon matrix $U \in \mathbb{R}^{n\times n}$, with $\operatorname{rank}(U) = r$, p-$\operatorname{rank}(U) = p$ and $n - r$ zero Jordan blocks whose sizes are given by the Segre characteristic S satisfaying (1.2).

We recall that a matrix is upper block echelon if each nonzero block, starting from the left, is to the right of the nonzero blocks below and the zero blocks are at the bottom. A matrix is a lower (block) echelon matrix if its transpose is an upper (block) echelon matrix. In the Procedure 1 we use the nonsingular ITN matrix

$$V_q = \begin{bmatrix} 1 & 1 & 1 & \ldots & 1 & 1 \\ 1 & 2 & 2 & \ldots & 2 & 2 \\ 1 & 2 & 3 & \ldots & 3 & 3 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 2 & 3 & \ldots & q-1 & q-1 \\ 1 & 2 & 3 & \ldots & q-1 & q \end{bmatrix} = [\min\{i, j\}]_{q\times q}$$

and the following MatLab notation: $A(i, :)$ denotes the i-th row of A and $A(:, j)$ denotes its j-th column; $\operatorname{ones}(n, m)$ denotes the $n \times m$ matrix of ones; $\operatorname{triu}(\operatorname{ones}(n, m))$ denotes the upper triangular part of $\operatorname{ones}(n, m)$; $\operatorname{zeros}(n, m)$ denotes the $n \times m$ zero matrix.

Note that if $r = p$ the algebraic and geometric multiplicity of the zero eigenvalue Of U is the same, therefore U has $n - r$ zero Jordan blocks of size 1×1. In this case is easy to see that the matrix U can be the following

$$U = \begin{bmatrix} \operatorname{triu}(\operatorname{ones}(p, n)) \\ \operatorname{zeros}(n - p, n) \end{bmatrix}.$$

If $p < r$ we construct a matrix U by blocks as follows

$$U = \begin{bmatrix} U_{11} & U_{12} & U_{13} & U_{14} & \ldots & U_{1,s_1-1} & U_{1,s_1} & U_{1,s_1+1} \\ O & O & U_{23} & U_{24} & \ldots & U_{2,s_1-1} & U_{2,s_1} & U_{2,s_1+1} \\ O & O & O & U_{34} & \ldots & U_{3,s_1-1} & U_{3,s_1} & U_{3,s_1+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ O & O & O & O & \ldots & O & U_{s_1-1,s_1} & U_{s_1-1,s_1+1} \\ O & O & O & O & \ldots & O & O & U_{s_1,s_1+1} \\ O & O & O & O & \ldots & O & O & O \end{bmatrix}.$$

Each block and its size are given in the following procedure.

Procedure 1. Given a realizable triple (n, r, p) and the Segre characteristic $S = (s_1, s_2, \ldots, s_{n-r})$ satisfying (1.2), this procedure obtains an upper block echelon matrix $U \in \mathbb{R}^{n\times n}$, with $\operatorname{rank}(U) = r$, p-$\operatorname{rank}(U) = p$ and $n - r$ zero Jordan blocks whose sizes are given by S.

Step 1. Obtain the conjugated sequence of S, $W = (w_1, w_2, \ldots, w_{s_1})$ and from W define $R = (0, r_2, \ldots, r_{s_1})$, with $r_i = w_i,\ i = 2, 3, \ldots, s_1$.

Step 2. Calculate $n_1 = p + 1 - s_1$ and construct

$$\begin{bmatrix} U_{11} & U_{12} & U_{13} & \ldots & U_{1,s_1} & U_{1,s_1+1} \end{bmatrix} = \operatorname{triu}(\operatorname{ones}(n_1, n))]$$

Step 3. Construct $U_{23} = V_{r_2+1}$ and

$$\begin{bmatrix} O & O & U_{23} & U_{24} & \ldots & U_{2,s_1} & U_{2,s_1+1} \end{bmatrix}$$
$$= [\operatorname{zeros}(r_2 + 1, n_1 + r_2)\ U_{23}\ U_{23}(:, r_2 + 1) * \operatorname{ones}(1, n - n_1 - 2r_2 - 1)]$$

Step 4. For $i = 3, 4, \ldots, s_1$

4.1. If $r_i = r_{i-1}$ construct $U_{i,i+1} = V_{r_i+1}$.

4.2. If $r_i < r_{i-1}$ construct $U_{i,i+1} = \begin{bmatrix} V_{r_i+1} \\ \text{ones}(r_{i-1} - r_i, 1) * V_{r_i+1}(r_i + 1, :) \end{bmatrix}$

After, in both cases,

$$\begin{bmatrix} O & \ldots & O & U_{i,i+1} & U_{i,i+2} & \ldots & U_{i,s_1} & U_{i,s_1+1} \end{bmatrix}$$
$$= [\text{zeros}(r_{i-1} + 1, n_1 + r_2 + \textstyle\sum_{j=2}^{i-1}(r_j + 1)) \;\; U_{i,i+1} \;\; U_{i,i+1}(:, r_i + 1) * \text{ones}(1, n - n_1 - r_2 - \textstyle\sum_{j=2}^{i}(r_j + 1))]$$

Step 5. Finally, the last block is equal to

$$\text{zeros}\left(n - n_1 - (r_2 + 1) - \sum_{i=2}^{s_1-1}(r_i + 1),\; n\right)$$

□

In the answer to the second question we have seen that the realizable triple $(19, 14, 8)$ has associated 10 different zero Jordan canonical forms, being $(4, 3, 2, 1, 1)$ one of them. In the following example we construct an upper block echelon TN matrix U with this Jordan canonical form using Procedure 1.

Example 2.1 Obtain a 19×19 upper block echelon TN matrix U, with $\text{rank}(U) = 14$, $p\text{-rank}(U) = 8$ and with 5 zero Jordan blocks of sizes $S = (4, 3, 2, 1, 1)$.

Since $r \neq p$ and $s_1 = 4$, following Procedure 1 we construct an upper block TN matrix

$$U = \begin{bmatrix} U_{11} & U_{12} & U_{13} & U_{14} & U_{15} \\ O & O & U_{23} & U_{24} & U_{25} \\ O & O & O & U_{34} & U_{35} \\ O & O & O & O & U_{45} \\ O & O & O & O & O \end{bmatrix}.$$

Step 1. The conjugated sequence of S is $W = (5, 3, 2, 1)$ and then, $R = (0, 3, 2, 1)$.

Step 2. $n_1 = p + 1 - s_1 = 5$ and

$$\begin{bmatrix} U_{11} & U_{12} & U_{13} & U_{14} & U_{15} \end{bmatrix} = \text{triu}(\text{ones}(5, 19)) =$$
$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

Step 3. $U_{23} = V_4$ and

$$\begin{bmatrix} O & O & U_{23} & U_{24} & U_{2,5} \end{bmatrix} = [\text{zeros}(4, 8) \;\; U_{23} \;\; U_{23}(:, 4) * \text{ones}(1, 7)] =$$
$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 3 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 \end{bmatrix}.$$

Step 4. For $i = 3$, since $r_3 < r_2$, construct

$$U_{34} = \begin{bmatrix} V_3 \\ \text{ones}(1, 1) * V_3(3, :) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \\ 1 & 2 & 3 \end{bmatrix}.$$

and

$$\begin{bmatrix} O & O & O & U_{34} & U_{35} \end{bmatrix} = [\text{zeros}(4,12)\ U_{34}\ U_{(}34)(:,3) * \text{ones}(1,4)] =$$

$$\begin{bmatrix} 0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&1&1&1 \\ 0&0&0&0&0&0&0&0&0&0&0&0&1&2&2&2&2&2&2 \\ 0&0&0&0&0&0&0&0&0&0&0&0&1&2&3&3&3&3&3 \\ 0&0&0&0&0&0&0&0&0&0&0&0&1&2&3&3&3&3&3 \end{bmatrix}.$$

Now, for $i = 4$ since $r_4 < r_3$, construct

$$U_{45} = \begin{bmatrix} V_2 \\ \text{ones}(1,1) * V_2(2,:) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 2 \end{bmatrix}$$

and

$$\begin{bmatrix} O & O & O & O & U_{45} \end{bmatrix} = [\text{zeros}(3,15)\ U_{45}\ U_{(}45)(:,2) * \text{ones}(1,2)] =$$

$$\begin{bmatrix} 0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1 \\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&2&2&2 \\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&2&2&2 \end{bmatrix}$$

Step 5. The last zero block is

$$\left[\text{zeros}\left(n - n_1 - (r_2+1) - \sum_{i=2}^{s_1-1}(r_i+1),\ n\right)\right] = [\text{zeros}(3,19)]\,.$$

Therefore, the matrix U is

$$\left[\begin{array}{ccccc|ccc|cccc|ccc|cccc}
1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1 \\
0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1 \\
0&0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1 \\
0&0&0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1 \\
0&0&0&0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1 \\
\hline
0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1&1&1&1 \\
0&0&0&0&0&0&0&0&1&2&2&2&2&2&2&2&2&2&2 \\
0&0&0&0&0&0&0&0&1&2&3&3&3&3&3&3&3&3&3 \\
0&0&0&0&0&0&0&0&1&2&3&4&4&4&4&4&4&4&4 \\
\hline
0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&1&1&1 \\
0&0&0&0&0&0&0&0&0&0&0&0&1&2&2&2&2&2&2 \\
0&0&0&0&0&0&0&0&0&0&0&0&1&2&3&3&3&3&3 \\
0&0&0&0&0&0&0&0&0&0&0&0&1&2&3&3&3&3&3 \\
\hline
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1 \\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&2&2&2 \\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&2&2&2 \\
\hline
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0 \\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0 \\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0
\end{array}\right].$$

The following result proves that the matrix U constructed by Procedure 1 verifies the desired properties.

Theorem 2.2 [6, *Theorem* 1]
Consider the matrix U constructed by Procedure 1. *Then the following properties hold*:

1. *U is a TN matrix with rank*$(U) = r$ *and p-rank*$(U) = p$.
2. *U has* $n - r$ *zeros Jordan blocks whose sizes are given by the sequence* $S = (s_1, s_2, \ldots, s_{n-r})$.

3. Construct an ITN matrix with a prescribed zero Jordan structure

In this section we construct an ITN matrix A associated with the realizable triple (n, r, p) and with a zero Jordan canonical form associated with this triple. For that, we use the procedure given in the previous section to construct an upper block echelon TN matrix U of size $n \times n$, with $\text{rank}(U) = r$, $p\text{-rank}(U) = p$ and with a zero Jordan canonical form associated with this triple. Now, we give the following procedure to compute the matrix A.

Procedure 2. Given a realizable triple (n, r, p) and the Segre characteristic $S = (s_1, s_2, \ldots, s_{n-r})$ satisfying (1.2), this procedure obtains an ITN matrix $A \in \mathbb{R}^{n\times n}$, associated with this triple and with $n - r$ zero Jordan blocks whose sizes are given by S.

Step 1. Apply Procedure 1 to construct the upper block matrix U.

Step 2. Construct the lower triangular TN matriz $L = \text{tril}(\text{ones}(n, n))$.

Step 3. Obtain $A = L * U$.

□

The following result proves that the matrix A satisfies the prescribed conditions.

Theorem 3.1 [6, *Proposition* 1, *Theorem* 2]
The matrix A constructed by Procedure 2 *satisfies the following conditions*:

1. *A is a ITN matrix.*
2. *rank*$(A) = r$.
3. *p-rank*$(A) = p$.
4. *Matrices A and U have the same zero Jordan structure.*

Example 3.2 Construct a 19×19 ITN matrix A, associated with the realizable triple $(19, 14, 8)$ and with 5 zero Jordan blocks of sizes $S = (4, 3, 2, 1, 1)$.

Using the matrix U obtained in Example 2.1 and following Procedure 2, we have

$A = \text{tril}(\text{ones}(n, n)) * U =$

$$\left[\begin{array}{ccccc|ccc|cccc|cccc|cccc}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
1 & 2 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 \\
1 & 2 & 3 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 \\
1 & 2 & 3 & 4 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 \\
\hline
1 & 2 & 3 & 4 & 5 & 5 & 5 & 5 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 & 6 \\
1 & 2 & 3 & 4 & 5 & 5 & 5 & 5 & 7 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 \\
1 & 2 & 3 & 4 & 5 & 5 & 5 & 5 & 8 & 10 & 11 & 11 & 11 & 11 & 11 & 11 & 11 & 11 & 11 & 11 \\
1 & 2 & 3 & 4 & 5 & 5 & 5 & 5 & 9 & 12 & 14 & 15 & 15 & 15 & 15 & 15 & 15 & 15 & 15 & 15 \\
\hline
1 & 2 & 3 & 4 & 5 & 5 & 5 & 5 & 9 & 12 & 14 & 15 & 16 & 16 & 16 & 16 & 16 & 16 & 16 & 16 \\
1 & 2 & 3 & 4 & 5 & 5 & 5 & 5 & 9 & 12 & 14 & 15 & 17 & 18 & 18 & 18 & 18 & 18 & 18 & 18 \\
1 & 2 & 3 & 4 & 5 & 5 & 5 & 5 & 9 & 12 & 14 & 15 & 18 & 20 & 21 & 21 & 21 & 21 & 21 & 21 \\
1 & 2 & 3 & 4 & 5 & 5 & 5 & 5 & 9 & 12 & 14 & 15 & 19 & 22 & 24 & 24 & 24 & 24 & 24 & 24 \\
\hline
1 & 2 & 3 & 4 & 5 & 5 & 5 & 5 & 9 & 12 & 14 & 15 & 19 & 22 & 24 & 25 & 25 & 25 & 25 & 25 \\
1 & 2 & 3 & 4 & 5 & 5 & 5 & 5 & 9 & 12 & 14 & 15 & 19 & 22 & 24 & 26 & 27 & 27 & 27 & 27 \\
1 & 2 & 3 & 4 & 5 & 5 & 5 & 5 & 9 & 12 & 14 & 15 & 19 & 22 & 24 & 27 & 29 & 29 & 29 & 29 \\
\hline
1 & 2 & 3 & 4 & 5 & 5 & 5 & 5 & 9 & 12 & 14 & 15 & 19 & 22 & 24 & 27 & 29 & 29 & 29 & 29 \\
1 & 2 & 3 & 4 & 5 & 5 & 5 & 5 & 9 & 12 & 14 & 15 & 19 & 22 & 24 & 27 & 29 & 29 & 29 & 29 \\
1 & 2 & 3 & 4 & 5 & 5 & 5 & 5 & 9 & 12 & 14 & 15 & 19 & 22 & 24 & 27 & 29 & 29 & 29 & 29
\end{array}\right]$$

Acknowledgements

This work has been supported by the Ministerio de Economía y Competitividad under the Spanish DGI grant MTM2017-85669-P-AR.

References

[1] T. Ando. Totally positive matrices. *Linear Algebra and its Applications*, 90:165–219. 1987.

[2] R. Bhatia. *Positive Definite Matrices*. Princeton Series in Applied Mathematics, 2007.

[3] R. Bru, R. Cantó and A. M. Urbano. Eigenstructure of rank one updated matrices. *Linear Algebra and its Applications*, 485:372–391, 2015. https://doi.org/10.1016/j.laa.2015.07.036

[4] R. Cantó, B. Ricarte and A. M. Urbano. Characterizations of rectangular totally and strictly totally positive matrices. *Linear Algebra and its Applications*, 432:2623–2633, 2010. https://doi.org/10.1016/j.laa.2009.12.004

[5] R. Cantó and A. M. Urbano. On the maximum rank of totally nonnegative matrices. *Linear Algebra and its Applications*, 551:125–146, 2018. https://doi.org/10.1016/j.laa.2018.03.045

[6] B. Cantó, R. Cantó and A. M. Urbano. All Jordan canonical forms of irreducible totally nonnegative matrices. *Linear and Multilinear Algebra*, online, 2019. https://doi.org/10.1080/03081087.2019.1676691

[7] S. M. Fallat. Bidiagonal factorizations of totally nonnegative matrices. *Amer. Math. Monthly*, 109:697–712, 2001. DOI: 10.2307/2695613

[8] S. M. Fallat, M. I. Gekhtman and C. R. Johnson. Spectral structures of irreducible totally nonnegative matrices. *SIAM J. Matrix Anal. Appl.*, 22(2):627–645, 2000. https://doi.org/10.1137/S0895479800367014

[9] S. M. Fallat and M. I. Gekhtman. Jordan structure of Totally Nonnegative Matrices. *Canad. Journal of Math.*, 57(1):82–98, 2005. https://doi.org/10.4153/CJM-2005-004-0

[10] S. M. Fallat and C. R. Johnson. *Totally Nonnegative Matrices*. Princenton Series in Applies Mathematics, 2011.

[11] F. R. Gantmacher and M. G. Krein. *Ostsillyatsionye Matritsy i Yadra i Malye Kolebaniya Mekhanicheskikh Sistem*. Gosudarstvenoe Izdatel'stvo, Moskva-Leningrad, 1950. Englishtransl. as "Oscillation Matrices and Kernels and Small Vibrations of Mechanical Systems", USAEC, 1961.

[12] M. Gasca, C. A. Micchelli and J. M. Penã. Almost strictly totally positive matrices. *Numerical Algorithms*, 2:225–236, 1992. DOI: 10.1007/BF02145387

[13] S. Karlin. *Total Positivity*. Volume I, Stanford University Press, Stanford, California, 1968.

[14] P. Koev. *http://math.mit.edu/~plamen/software/STNTool.html*

[15] P. Koev. Accurate eigenvalues and SVDs of totally nonnegative matrices. *SIAM J. Matrix Anal. Appl.*, 27(1):1–23, 2005. DOI. 10.1137/S0895479803438225

[16] A. Pinkus. *Totally Positive Matrices*. Vol. 181 of Cambridge Tracts in Mathematics, Cambridge University Press, UK, 2010.

[17] I. J. Schoenberg. Uber variationsvermindernde lineare Transformationen. *Math. Z.*, 32:321–328, 1930. https://doi.org/10.1007/BF01194637

[18] I. J. Schoenberg and A. Whitney. On Pólya frequency functions III. The positivity of translation determinants with applications to the interpolation problem by spline curves. *Trans. Am. Math. Soc.*, 14:246–259, 1953. https://doi.org/10.1007/978-1-4899-0433-1-16

[19] H. Shapiro. The Weyr characteristic. *The American Mathematical Monthly*, 106:919–929, 1999. DOI: 10.2307/2589746

Author index