SEMA SIMAI Springer Series 4

Fernando Casas
Vicente Martínez *Editors*

Advances in Differential Equations and Applications

SEMA SIMAI SOCIETÀ ITALIANA DI MATEMATICA APPLICATA E INDUSTRIALE
Springer

SEMA SIMAI Springer Series

Volume 4

The SEMA SIMAI Springer Series is a joint series aiming to publish advanced textbooks, research-level monographs and collected works that focus on applications of mathematics to social and industrial problems, including biology, medicine, engineering, environment and finance. Mathematical and numerical modeling is playing a crucial role in the solution of the complex and interrelated problems faced nowadays not only by researchers operating in the field of basic sciences, but also in more directly applied and industrial sectors. This series is meant to host selected contributions focusing on the relevance of mathematics in real life applications and to provide useful reference material to students, academic and industrial researchers at an international level. Interdisciplinary contributions, showing a fruitful collaboration of mathematicians with researchers of other fields to address complex applications, are welcomed in this series.

More information about this series at
http://www.springer.com/series/10532

Fernando Casas • Vicente Martínez
Editors

Advances in Differential Equations and Applications

Springer

Editors
Fernando Casas
Dept. de Matemàtiques and IMAC
Universitat Jaume I
Castelló
Spain

Vicente Martínez
Dept. de Matemàtiques and IMAC
Universitat Jaume I
Castelló
Spain

ISSN 2199-3041 ISSN 2199-305X (electronic)
SEMA SIMAI Springer Series
ISBN 978-3-319-06952-4 ISBN 978-3-319-06953-1 (eBook)
DOI 10.1007/978-3-319-06953-1
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014954349

Printed on acid-free paper

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

Preface

This volume of the SEMA/SIMAI Springer Series arose from the 23rd Congress on Differential Equations and Applications (CEDYA)/13th Congress of Applied Mathematics (CMA). The conference took place at the Universitat Jaume I in Castelló (Spain) on 9–13 September 2013 and was sponsored by Generalitat Valenciana, the Institut de Matemàtiques i Aplicacions de Castelló (IMAC) and the Departament de Matemàtiques of the Universitat Jaume I. It was attended by more than 200 participants, mainly from Spain but also from a further nine countries.

CEDYA has a long tradition in the Spanish applied mathematics community. It was first held in 1978 in El Escorial (Madrid), serving as a meeting point for mathematicians working in different research areas such as differential equations (both ordinary and partial), numerical analysis, control and optimization, and industrial mathematics. Nowadays, CEDYA is renowned as the congress of the Spanish Society of Applied Mathematics (SEMA) and constitutes the main forum and meeting point for applied mathematicians in Spain.

The organizers of the 23rd CEDYA/13th CMA are especially grateful to all members of the Scientific Committee, plenary speakers, organizers of the Special Sessions, and participants for their stimulating contributions, both verbal and written and for providing a lively scientific atmosphere during the conference.

The congress took place at the premises of the Fundació Universitat Empresa (FUE), Universitat Jaume I. The editors wish to acknowledge all the institutions involved in its organization. They are particularly grateful for the assistance and constant support received from the FUE personnel, with a special mention of Begõna Andrés for her outstanding contribution in making the conference the success that (we believe) it was.

The collection of papers in this volume is based on the contributions presented at the conference. The papers were selected after a thorough refereeing process and provide a good summary of the recent activity of the different groups working mainly in Spain on applications of mathematics to various fields of science and technology.

The refereeing and editorial procedures have had to conform to a very specific timetable, and so the editors would like to take this opportunity to thank all the authors and referees for their understanding when coping with such an expedited procedure. Special mention goes to Francesca Bonadei from Springer for her enthusiastic support and encouragement during the different phases of the editorial process.

The papers included in this volume fall into a number of distinct subject areas covered by the conference and thus are arranged in accordance with this subdivision.

The first section is devoted to theoretical aspects of partial differential equations and contains six papers. The second section deals with different aspects relating to ordinary differential equations and dynamical systems, both from a qualitative point of view and also in terms of the design of new numerical techniques for their treatment.

The third section is entitled "Applications and Modeling" and covers topics such as multiresolution, time series, controllability, and models of traffic flow and fire propagation.

Finally, the fourth section, entitled "Numerical Analysis", contains papers presenting new numerical techniques designed to solve specific problems arising in ordinary and partial differential equations as well as numerical linear algebra.

We hope that this volume will appeal to both researchers and practitioners in analytical and numerical aspects of differential equations and numerical analysis as a whole as well as some of their applications but also to the non-experts who wish to gain a taste of the new developments in these areas of current interest.

Castelló, Spain
July 2014

Fernando Casas
Vicente Martínez

Contents

Part I
Partial Differential Equations

A Degenerate Parabolic Logistic Equation

José M. Arrieta, Rosa Pardo, and Aníbal Rodríguez-Bernal

Abstract We analyze the behavior of positive solutions of parabolic equations with a class of degenerate logistic nonlinearity and Dirichlet boundary conditions. Our results concern existence and strong localization in the spatial region in which the logistic nonlinearity cancels. This type of nonlinearity has applications in the nonlinear Schrödinger equation and the study of Bose–Einstein condensates. In this context, our analysis explains the fact that the ground state presents a strong localization in the spatial region in which the nonlinearity cancels.

1 Introduction

In this paper we analyse the behavior of positive solutions of parabolic equations with a degenerate logistic nonlinearity and Dirichlet boundary conditions

$$\begin{cases} u_t - \Delta u = \lambda u - n(x)u^\rho & \text{in } \Omega,\ t > 0, \\ u = 0 & \text{on } \partial\Omega,\ t > 0, \\ u(0) = u_0 \geq 0, \end{cases} \tag{1}$$

Partially supported by Project MTM2012-31298, MINECO, Spain and Grupo de Investigación CADEDIF, UCM.

J.M. Arrieta (✉) • A. Rodríguez-Bernal
Departamento de Matemática Aplicada, Universidad Complutense de Madrid, 28040 Madrid, Spain

Instituto de Ciencias Matemáticas, CSIC-UAM-UC3M-UCM, 28049 Madrid, Spain
e-mail: arrieta@mat.ucm.es; arober@mat.ucm.es

R. Pardo
Departamento de Matemática Aplicada, Universidad Complutense de Madrid, 28040 Madrid, Spain
e-mail: rpardo@mat.ucm.es

F. Casas, V. Martínez (eds.), *Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 4, DOI 10.1007/978-3-319-06953-1_1

where $\Omega \subset \mathbb{R}^N$, $N \geq 1$, is a bounded domain, $\rho > 1$, $\lambda \in \mathbb{R}$ and $n(x) \geq 0$ in Ω. Assume also that $n(x)$ remains strictly positive near the boundary of Ω and therefore

$$K_0 = \{x \in \Omega : n(x) = 0\} \subset \Omega \qquad \text{is a nonempty compact set.} \tag{2}$$

The parabolic problem (1) degenerates into a linear equation on K_0, there the growth rate is exponential and a solution could be expected to be unbounded. In the region where $n(x) > n_0 > 0$, the growth is logistic, and a solution could be expected to be bounded. The question is what kind of behavior could be expected in the whole domain Ω, and how the solution will 'glue' the different behavior in those subregions. Hence K_0 plays a crucial role in the dynamical properties and the asymptotic behavior of solutions of (1), as we will show below.

There is a large amount of mathematical literature in this kind of logistic equations, see below. This type of nonlinearity has also applications in the nonlinear Schrodinger equation and the study of Bose-Einstein condensates. In this context, assumption (2) implies the fact that the *ground state* presents a strong localization in the spatial region K_0, see [12] and references therein.

Throughout this paper we shall assume that the compact set K_0 and the function $n(x)$ satisfy the following hypotheses

(H1) $K_0 = K_1 \cup K_2 \subset \Omega$, where K_1 and K_2 are compact sets and

$$K_1 = \overline{\Omega_0}, \qquad \text{is the closure of a regular connected open set} \quad \Omega_0 \neq \emptyset,$$

$$K_2 \qquad \text{has zero Lebesgue measure.}$$

In some cases (H1) will be strengthened to

(H1′) K_0 satisfies (H1) and

$$K_2 \qquad \text{is a closed regular } d-\text{dimensional manifold, with} \quad d \leq N-1.$$

(H2) $n(x)$ is a Hölder continuous function and

$$n(x) \geq C\big(d_0(x)\big)^{\gamma} \quad \text{for some } \gamma > 0, \qquad \text{where} \quad d_0(x) := \operatorname{dist}(x, K_0).$$

When the set K_0 is empty, that is, if $n(x)$ is strictly bounded away from zero, the parabolic problem (1) is classical and well understood, see e.g. [13] and references therein. Also, when K_0 is "smooth" in the sense that in (H1) we have $K_0 = K_1 = \overline{\Omega_0}$ where Ω_0 is a smooth open set, and $K_2 = \emptyset$, this problem has also been studied in [3–5, 9, 11] and further developments in [6, 8], see also references therein. Therefore here we focus on the effect on the solutions of the presence of the part with empty interior K_2.

Let us consider the stationary associated problem, see [1]. We will denote by $\lambda_1(\omega)$ the first eigenvalue of the Laplace operator defined in an open set ω, with Dirichlet boundary conditions on $\partial\omega$. As λ crosses the value $\lambda_1(\Omega)$, a bifurcation phenomena takes place and a unique positive solution emanates from the trivial one. This solution can be continued in λ up until it reaches a critical value $\lambda_c = \lambda_1(\Omega_0)$, see [1, Theorem 2.3]. Note that this is precisely the same situation as when K_0 is "smooth", i.e. $K_2 = \emptyset$. On the other side, when K_0 is empty, the picture is also as above, with $\lambda_c = \infty$. In [1] we give a detailed description of the behavior of this branch of solutions for $\lambda \in (\lambda_1(\Omega), \lambda_1(\Omega_0))$ and specially as $\lambda \to \lambda_1(\Omega_0)$.

For any $\lambda \in (\lambda_1(\Omega), \lambda_1(\Omega_0))$, there exists a unique classical positive stationary solution, denoted by φ_λ, which is globally asymptotically stable for positive solutions of (1). Moreover, inside Ω_0, the pointwise limit of φ_λ as $\lambda \uparrow \lambda_1(\Omega_0)$ is unbounded, see Theorem 1 for a precise statement and see [1, Theorem 1.1] for a proof. This result is already know in the particular case when $n(x)$ is a smooth function, $K_2 = \emptyset$, and $K_0 = K_1 = \overline{\Omega_0}$, an open set with regular boundary, see [3,4,11].

In K_2 we have two competing mechanisms: on one hand the fact that $n(x) \equiv 0$ in K_2 "pushes" the solution towards $+\infty$ while the fact that K_2 is not "fat" enough means that this effect may not have enough room to force the solution to go to infinity.

Roughly speaking, our main result state that if

$$\gamma + 2 < (\rho - 1)(N - d) \tag{H3}$$

then any positive equilibrium remains bounded on compact sets of $\Omega \setminus K_1$ and, in particular, at each point of $K_2 \setminus K_1$, see Theorem 1 below, see also [1, Theorem 1.1].

We will distinguish two situations for which we will be able to show that the solutions remain bounded in K_2. In case $K_2 \cap K_1 = \emptyset$, any solution will be bounded in K_2, actually it will be so in a neighborhood of K_2. In case $K_2 \cap K_1 \neq \emptyset$, it will turn out that a balance between the geometry of K_2 and the strength of the logistic term, given by the exponent ρ and the behavior of the function $n(x)$ near K_2, will determine the behavior of the solution, see the following theorem.

Theorem 1 *Assume K_0 satisfies (H1) and $n(x)$ satisfies (H2). Then for any $\lambda \in (\lambda_1(\Omega), \lambda_1(\Omega_0))$ there exists a unique classical positive equilibrium, denoted by φ_λ, which is globally asymptotically stable for nonnegative nontrivial solutions of (1), that is, for every $u_0 \gneqq 0$, the solution of (1) satisfy*

$$\lim_{t\to\infty} u(t, x; u_0) = \varphi_\lambda(x).$$

Also we have

$$\lim_{\lambda\uparrow\lambda_1(\Omega_0)} \varphi_\lambda(x) = \infty, \quad \text{for all } x \in \Omega_0, \tag{3}$$

with uniform limit in compact sets of Ω_0. Moreover, we have the following two cases:

(i) If $K_1 \cap K_2 = \emptyset$, then there exists a $\delta > 0$ and $M > 0$ such that

$$|\varphi_\lambda(x)| \leq M, \quad \forall x : d(x, K_2) \leq \delta, \quad \forall \lambda \in (\lambda_1(\Omega), \lambda_1(\Omega_0)).$$

(ii) If $K_1 \cap K_2 \neq \emptyset$, K_0 satisfies (H1′) and hypothesis (H3) holds, then φ_λ remains uniformly bounded on compact sets of $\Omega \setminus K_1$. In particular it remains bounded at each point of $K_2 \setminus K_1$.

Turning back to the parabolic problem, we have the following result:

Theorem 2 *Assume (H1)–(H2) hold. Let $u_0 \geq 0$ be a bounded initial data for (1). Then for any $\lambda > \lambda_1(\Omega_0)$ any positive solution of (1) satisfy*

$$\lim_{t \to \infty} u(x, t) = \infty, \quad \text{for all } x \in \Omega_0, \tag{4}$$

and the limit is uniform in compact sets of Ω_0. Moreover, we have the following:

(i) If $K_1 \cap K_2 = \emptyset$, then there exists a $\delta > 0$ and $M = M(u_0, \lambda, \delta) > 0$ such that

$$|u(x, t; u_0)| \leq M, \quad \forall x : dist(x, K_2) \leq \delta, \quad \forall t > 0.$$

(ii) If $K_1 \cap K_2 \neq \emptyset$, K_0 satisfies (H1′), and hypothesis (H3) holds, then for any $\lambda \geq \lambda_1(\Omega_0)$ any solution of (1) remains uniformly bounded on compact sets of $\Omega \setminus K_1$ as $t \to \infty$. In particular it remains bounded at each point of $K_2 \setminus K_1$.

The proof of this result relies on the following argument. If we denote by u a nonnegative solution of (1), then we obtain first an upper bound of u, independent of λ, in compact sets of $\Omega \setminus K_0$. If $\bar{B}(x_0, a) \subset \Omega \setminus K_0$, where $n(x) \geq \beta$ in this ball, we may compare the solution u with radial solutions of singular Dirichlet problems, posed in $B(x_0, a)$, going to infinity at the boundary, see [5, 7, 10]. By radial symmetry, the minimum of the singular solution is attained at the center of the ball (that is in x_0), and can be estimated in terms of β, a, ρ and the dimension N. Translating this result to our problem, we can move those balls for points in $\Omega \setminus K_0$ next to the boundary of K_0, and state some rate for the upper bounds in terms of some inverse power of the distance to the boundary of K_0. This estimates provide an upper rate at which the solution may diverge to infinity as we approach K_0. See Lemma 2, Proposition 1 and Lemma 3.

Once this estimate is obtained, we realize that the rate obtained with the argument above may imply that the solution u is a solution of a parabolic problem with an L^r trace at the boundary. Parabolic regularity will imply that the solution u is bounded, independent of λ, in compact sets of $\Omega \setminus K_1$. Therefore, we may obtain conditions on ρ, the dimensions N and d and the rate γ at which $n(x)$ approaches to zero, see (H2), which may guarantee that the solution is bounded in $K_2 \setminus K_1$, see Theorem 2.(ii).

This paper is organized as follows. We first show that the solutions are uniformly bounded in compact sets of $\Omega \setminus K_0$ (see Proposition 1 below). Next, we prove that for $\lambda \geq \lambda_1(\Omega_0)$, any solution of the parabolic problem (1) start to grow up in K_1 as $t \to \infty$, see Theorem 2. Also, if the two parts K_1 and K_2 of K_0 are disjoint, then all solutions remain globally bounded on K_2 as $t \to \infty$, see Theorem 2.(i). Finally, when $K_1 \cap K_2 \neq \emptyset$, provides sufficient conditions ensuring that all solutions of (1) remain bounded in $K_2 \setminus K_1$, see Theorem 2.(ii).

2 Boundedness and Unboundedness of Solutions

We analyze where and how solutions of (1) become unbounded. The first thing we can say is that the blow-up is a complete blow-up at every point in Ω_0.

Lemma 1 *Assume K_0 satisfies (H1). Let u be a solution of the parabolic problem (1). If $\lambda > \lambda_1(\Omega_0)$, then*

$$\lim_{t\to\infty} u(x,t) = \infty, \quad \text{for all } x \in \Omega_0.$$

Proof Let $z(x,t)$ be the solution of

$$\begin{cases} z_t - \Delta z = \lambda z, & \text{in } \Omega_0, \quad t > 0, \\ z = 0 & \text{on } \partial\Omega_0, \quad t > 0, \\ z(0) = z_0 \geq 0 & \text{in } \Omega_0 \end{cases}$$

with $z_0 \leq u_0$. Then, by comparison and due to $n(x) \geq 0$ in K_0, $z(x,t) \leq u(x,t)$ for $x \in \Omega_0$. Since $\lambda > \lambda_1(\Omega_0)$ then $z(x,t)$ grows exponentially in Ω_0.

To get upper bounds on the solutions outside Ω_0 we will use the following Lemma, see [5]. This Lemma analyzes the minimum of a radially symmetric solution of a singular logistic equation with constant coefficients and going to infinity at the boundary, see [7, 10].

Lemma 2 *Assume $\rho > 1$ and $\lambda, \beta > 0$ and consider a ball in $\mathbb{R}^N$ of radius $a > 0$ and the following singular Dirichlet problem*

$$-\Delta z = \lambda z - \beta z^\rho \quad \text{in } B(0,a); \qquad z = \infty \quad \text{on } \partial B(0,a).$$

Then, there exists a unique positive radial solution, $z_a(x)$. Moreover z_a satisfies

$$\left(\frac{\lambda}{\beta}\right)^{\frac{1}{\rho-1}} \leq z_a(0) = \inf_{B(0,a)} z_a(x) \leq \left(\frac{\lambda(\rho+1)}{2\beta} + \frac{B}{\beta a^2}\right)^{\frac{1}{\rho-1}}$$

for some constant $B = B(\rho, N) > 0$, B independent of λ.

The above Lemma gives a local upper bound for the parabolic problem, out of K_0.

Proposition 1 *Let $x_0 \in \Omega \setminus K_0$ and let $u_0 \geq 0$ be a bounded initial data for (1). Then for any given $\lambda \geq \lambda_0(K_0)$ there exists $b > 0$ and $M > 0$ such that*

$$0 \leq u(t, x; u_0) \leq M, \quad x \in B(x_0, b), \quad t > 0,$$

where $B(x_0, b)$ denotes the ball centered at x_0 with radius b.

Proof Let $x_0 \in \Omega \setminus K_0$ and let $a > 0$ be such that $B(x_0, a) \subset \Omega \setminus K_0$. Denote $\beta = \inf\{n(x),\ x \in B(x_0, a)\} > 0$ and consider $z(x)$ the translation to $B(x_0, a)$ of the function in Lemma 2.

Given u_0, for a sufficiently small we have that $u_0(x) \leq z(x_0) \leq z(x)$ for $x \in B(x_0, a)$. Hence $z(x)$ is a supersolution for $u(x, t)$ and then

$$u(x, t) \leq z(x), \quad x \in B(x_0, a), \quad t > 0.$$

Now in $B(x_0, a/2)$, $z(x)$ remains bounded and we conclude the proof with $b = a/2$.

Next we discuss the behavior of the solutions in K_0. First we give a universal (and singular) bound.

Lemma 3 *Let $u_0 \geq 0$ be a bounded initial data for (1). Then, there exists a constant $A = A(u_0, \lambda)$ such that the following holds*

$$0 \leq u(t, x; u_0) \leq h(x) = \left(\frac{A}{d_0^2(x)\ \inf_{x \in B_0} n(x)} \right)^{\frac{1}{\rho - 1}},$$

where $B_0 := B\left(x_0, \frac{d_0(x)}{2}\right)$, and $d_0(x) = dist(x, K_0)$.

Proof Let $x_0 \in \Omega \setminus K_0$, hence $B_0 \subset \Omega \setminus K_0$. Denote $\beta(x_0) = \inf\{n(x),\ x \in B_0\} > 0$ and consider $z(x)$ the translation to B_0 of the function in Lemma 2.

Let $u_0 \leq M$ in $\overline{\Omega}$. Using the continuity of $n(x)$, we can assume that $\beta(x_0) \leq \frac{\lambda}{M^{\rho - 1}}$ for all x_0 close enough to K_0. Then, using Lemma 2 we have

$$u_0(x) \leq M \leq \left(\frac{\lambda}{\beta(x_0)} \right)^{\frac{1}{\rho - 1}} \leq z(x_0) \leq z(x), \qquad \forall x \in B_0.$$

Hence $z(x)$ is a supersolution for $u(x, t)$ and then $u(x, t) \leq z(x)$, for all $x \in B_0$, $t > 0$. In particular, for $x = x_0$ we get, from Lemma 2 that for all $t > 0$,

$$u(x_0, t) \leq z(x_0) \leq \left(\frac{\lambda(\rho + 1)}{2\beta(x_0)} + \frac{B}{\beta(x_0) d_0(x_0)^2} \right)^{\frac{1}{\rho - 1}}$$

for some constant $B > 0$. Since x_0 is close enough to K_0 we can assume

$$u(x_0, t) \leq z(x_0) \leq \left(\frac{A}{\beta(x_0) d_0(x_0)^2} \right)^{\frac{1}{\rho - 1}}$$

for all $t > 0$ and some $A > 0$.

From previous results, far from K_0, $u(x, t)$ remains bounded, and for $x_0 \in K_0$ the result is obvious.

Next we want to distinguish the behavior of the solutions in K_1 and on K_2. The following result gives a criteria to check whether a function that is infinity on a compact set of measure zero is integrable. As shown below, this criteria depends on the dimension of the set and on the form the function diverges on the compact set.

Lemma 4 *Assume $K \subset \mathbb{R}^N$ is a compact set with zero Lebesgue measure and dimension $d \leq N - 1$ and consider a function defined on a bounded neighborhood ω of K of the form*

$$f(x) = (dist(x, K))^{-\alpha}, \qquad \text{for some} \quad \alpha > 0, \qquad f\,|_K = \infty.$$

If $r\alpha < N - d$ for some $r \geq 1$, then $f \in L^r(\omega)$.

Proof Note that

$$\int_\omega |f(x)|^r \, dx = \int_0^\infty |A_s| \, ds$$

where $A_s = \{x \in \omega, \ |f(x)|^r \geq s\}$. But

$$|f(x)|^r \geq s \qquad \text{iff} \quad dist(x, K) \leq s^{-\frac{1}{\alpha r}}.$$

Therefore $|A_s| = |\omega_{\delta(s)}|$ where

$$\omega_\delta = \{x \in \omega, \quad dist(x, K) \leq \delta\} \qquad \text{and} \quad \delta(s) = s^{-\frac{1}{\alpha r}}.$$

From the assumption on the dimension of K we get $|\omega_\delta| \leq C\delta^{N-d}$. Moreover, due to $|A_s| \leq |\omega|$, $\int_0^\infty |A_s| \, ds \leq |\omega| + \int_1^\infty |A_s| \, ds$. Therefore,

$$\int_\omega |f(x)|^r \, dx \leq |\omega| + C \int_1^\infty \left(\frac{1}{s} \right)^{\frac{N-d}{\alpha r}} ds < \infty \qquad \text{whenever} \quad 1 < \frac{N-d}{\alpha r},$$

and the result follows.

We prove now Theorem 2.

Proof of Theorem 2 From Lemma 1, any positive solution is unbounded in Ω_0, and so (4) holds. Moreover, with the comparison argument used in the proof of Lemma 1 we get that the limit is uniform in compact sets of Ω_0.

(i) Since $K_1 \cap K_2 = \emptyset$ and $|K_2| = 0$, we can construct a set of the form $V_\delta = \{x \in \Omega : d(x, K_2) < \delta\}$ with $\delta > 0$ small enough so that $K_1 \cap \bar{V}_\delta = \emptyset$ and $\lambda_1(V_\delta)$ is large enough, say $\lambda_1(V_\delta) > \lambda$. Moreover, from Proposition 1, $|u|$ is bounded uniformly in $t > 0$ in ∂V_δ, by a constant, say M.

Hence, the solution U of

$$\begin{cases} U_t - \Delta U = \lambda U & \text{in } V_\delta, \quad t > 0, \\ U = M & \text{on } \partial V_\delta, \quad t > 0, \\ U(0) = u_0 \geq 0 & \text{in } V_\delta \end{cases}$$

becomes a supersolution of $|u(x,t)|$ in V_δ. Since $\lambda < \lambda_1(V_\delta)$ then $U(x,t)$ and therefore $|u(x,t)|$, remains bounded in V_δ.

(ii) From Proposition 1, for any given solution of (1) we have L^∞ bounds on compact sets of $\Omega \setminus K_0$.

Let K be an arbitrary compact set in $\Omega \setminus K_1$, such that $K \cap K_2 \neq \emptyset$. Let B be a "transversal isolating box" for K, that is B is an open bounded set such that $K \subset \overline{B} \subset \Omega \setminus K_1$ and $dim(K_2 \cap \partial B) \leq d - 1$. Then, from Lemmas 3, 4 and condition (H3), we have that there exists a function $h \in L^r(\partial B)$ such that $|u(x,t)| \leq h(x)$ for all $x \in \partial B$. Hence, the solution of

$$\begin{cases} U_t - \Delta U = \lambda U & \text{in } B, \quad t > 0, \\ U = \tilde{h}(x) & \text{on } \partial B, \quad t > 0, \\ U(0) = u_0 \geq 0 & \text{in } B \end{cases}$$

becomes a supersolution of $|u(x,t)|$ in B.

Now, if $\lambda \geq \lambda_1(\Omega_0)$ we can shrink B to be close enough to K_2 such that $\lambda < \lambda_1(B)$. Then, standard parabolic regularity gives L^∞ bounds for $U(x,t)$ for all time, on compact subsets of B. Hence, $u(x,t)$ remains bounded on K_2 as $t \to \infty$. □

Remark 1 It is an interesting open problem to determine whether we always obtain that the solution of the parabolic problem (1) are bounded in compact sets of $\Omega \setminus K_1$ or, in the contrary, that we have cases in which u becomes infinity in K_2 as $t \to \infty$.

Remark 2 This work is still in progress, and we refer to [2] for details and more general results, including more general configurations for the set K_0.

References

1. Arrieta, J.M., Pardo, R., Rodríguez-Bernal, A.: Localization phenomena in degenerate logistic equation. In: Variational and Topological Methods: Theory, Applications, Numerical Simulations, and Open Problems (2012). Electronic Journal of Differential Equations, Conference 21, 1–9 (2014)
2. Arrieta, J.M., Pardo, R., Rodríguez-Bernal, A.: Asymptotic behavior of degenerate logistic equations. Preprint.
3. Fraile, J.M., Medina, P.K., López-Gómez, J., Merino, S.: Elliptic eigenvalue problems and unbounded continua of positive solutions of a semilinear elliptic equation. J. Differ. Equ. **127**(1), 295–319 (1996)
4. Gámez, J.L.: Sub- and super-solutions in bifurcation problems. Nonlinear Anal. **28**(4), 625–632 (1997)
5. García-Melián, J., Gómez-Reñasco, R., López-Gómez, J., Sabina de Lis, J.C.: Pointwise growth and uniqueness of positive solutions for a class of sublinear elliptic problems where bifurcation from infinity occurs. Arch. Ration. Mech. Anal. **145**(3), 261–289 (1998)
6. Gómez-Reñasco, R., López-Gómez, J.: On the existence and numerical computation of classical and non-classical solutions for a family of elliptic boundary value problems Nonlinear Anal-Theor. **48**(4), 567–605 (2002)
7. Keller, J.B.: On solutions of $\Delta u = f(u)$. Commun. Pur. Appl. Math. **10**, 503–510 (1957)
8. López-Gómez, J.: Metasolutions: Malthus versus Verhulst in population dynamics. A dream of Volterra. In: Handbook of Differential Equations: Stationary Partial Differential Equations, vol. II, pp. 211–309. Elsevier/North-Holland, Amsterdam (2005)
9. López-Gómez, J., Sabina de Lis, J.C.: First variations of principal eigenvalues with respect to the domain and point-wise growth of positive solutions for problems where bifurcation from infinity occurs. J. Differ. Equ. **148**(1), 47–64 (1998)
10. Osserman, R.: On the inequality $\Delta u \geq f(u)$. Pac. J. Math. **7**, 1641–1647 (1957)
11. Ouyang, T.: On the positive solutions of semilinear equations $\Delta u + \lambda u - hu^p = 0$ on the compact manifolds. Trans. Am. Math. Soc. **331**(2), 503–527, (1992)
12. Pérez-García, V.M., Pardo, R.: Localization phenomena in nonlinear Schrödinger equations with spatially inhomogeneous nonlinearities: theory and applications to Bose-Einstein condensates. Physica D **238**(15), 1352–1361 (2009)
13. Smoller, J.: Shock waves and reaction-diffusion equations. Grundlehren der Mathematischen Wissenschaften, vol. 258. Springer, New York/Berlin (1983)

Fast and Slow Boundary Oscillations in a Thin Domain

José M. Arrieta and Manuel Villanueva-Pesqueira

Abstract In this work we analyze the behavior of the solutions of the Laplace operator with Neumann boundary conditions in a 2-dimensional thin domain with order of thickness ϵ which presents a high oscillatory behavior at the top and a weak oscillatory behavior at the bottom boundary. We obtain the asymptotic homogenized problem as $\epsilon \to 0$ and we are interested in understanding how the extremely different order of the oscillations affects to the limit.

1 Introduction

We analyze the behavior of the solutions of the Laplace equation with homogeneous Neumann boundary conditions

$$\begin{cases} -\Delta u^\epsilon + u^\epsilon = f^\epsilon & \text{in } R^\epsilon \\ \dfrac{\partial u^\epsilon}{\partial N^\epsilon} = 0 & \text{on } \partial R^\epsilon \end{cases} \tag{1}$$

where $f^\epsilon \in L^2(R^\epsilon)$, N^ϵ is the unit outward normal to ∂R^ϵ and R^ϵ is the thin domain with oscillating boundary,

$$R^\epsilon = \left\{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 \in (0,1),\ -\epsilon\, h(x_1/\epsilon^\beta) < x_2 < \epsilon\, g(x_1/\epsilon^\alpha)\right\}, \tag{2}$$

with $\beta > 1, \alpha < 1$ and $g, h : (0,1) \times \mathbb{R} \to \mathbb{R}$ are C^1 periodic functions with period L_1 and L_2 respectively. Moreover, there exist constants $h_0 = \min_{x\in\mathbb{R}}\{h(x)\}$ and $h_1, g_0, g_1 > 0$ such that $0 \le h_0 \le h(\cdot) \le h_1$, and $0 < g_0 \le g(\cdot) \le g_1$.

J.M. Arrieta (✉) • M. Villanueva-Pesqueira
Departamento de Matemática Aplicada, Universidad Complutense de Madrid, 28040 Madrid, Spain

Instituto de Ciencias Matemáticas, CSIC-UAM-UC3M-UCM, 28049 Madrid, Spain
e-mail: arrieta@mat.ucm.es; manuelvillanueva@mat.ucm.es

F. Casas, V. Martínez (eds.), *Advances in Differential Equations and Applications*,
SEMA SIMAI Springer Series 4, DOI 10.1007/978-3-319-06953-1_2

Observe that the domain R^ϵ shrinks in the vertical direction and it has an oscillatory behavior at the top and bottom boundary. Moreover, the bottom boundary presents a much higher oscillatory behavior than the top boundary. Notice that the period of the oscillations is order ϵ^α at the upper boundary, given by $\epsilon g(x_1/\epsilon^\alpha)$, while the period at the lower boundary is ϵ^β order, given by $\epsilon h(x_1/\epsilon^\beta)$ and they do not coincide with the order ϵ of the height of the domain.

The existence and uniqueness of solutions for problem (1) for each $\epsilon > 0$, is guaranteed by Lax–Milgram Theorem. We will analyze the behavior of solutions as $\epsilon \to 0$. The fact that R^ϵ gets thinner and thinner as $\epsilon \to 0$ suggests that the family of solutions u^ϵ will converge to a function of just one variable and that this function will satisfy certain elliptic equation in one dimension.

The behavior of the solutions for elliptic partial differential equations in thin domains is a subject that has been addressed in different works in the literature. The purely periodic case was investigated in [3, 11] using standard techniques in homogenization theory, as developed in [8, 9, 12]. In [4, 7] the authors treat the problem in locally periodic thin domains. The case where the lower boundary is not oscillatory, say $h(\cdot) \equiv 0$, was treated in [2] and the limit equation is given by

$$\begin{cases} -\dfrac{1}{\mathscr{M}(g)\mathscr{M}(\frac{1}{g})} u_{xx} + u = f, \text{ in } (0,1) \\ \\ u_x(0) = u_x(1) = 0 \end{cases} \tag{3}$$

where $\mathscr{M}(\phi)$ denotes the mean value of a function ϕ which is L-periodic, $\mathscr{M}(\phi) = \frac{1}{L}\int_{(0,L)} \phi \, ds$.

If the domain does not present oscillations in the upper boundary, we assume that $g(\cdot)$, independent of ϵ, defines the upper boundary of the thin domain, and $h_0 = \min_{x\in\mathbb{R}}\{h(x)\}$ then the variational formulation of the limit problem is: $\forall \varphi \in H^1(0,1)$

$$\int_0^1 \left\{ \Big(g(x) + h_0 \Big)\, u_x(x)\, \varphi_x(x) + p(x) \left(u(x) - f(x) \right) \varphi(x) \right\} dx = \int_0^1 p(x) f(x) \varphi \, dx, \tag{4}$$

where $p(x) = g(x) + \frac{1}{L_2}\int_0^{L_2} h(s)\, ds$, for all $x \in (0,1)$. We refer to [5] for details.

Our case is a combination of the two cases described above since the thin domain presents both kind of oscillatory boundary. We are interested in studying how the extremely different order of oscillations affects to the limit problem. Notice that we also consider thin domains with doubly oscillatory boundary in [6] but this case can not be addressed by the same techniques since the difference between the oscillations orders is much larger.

In this paper we will properly combine the techniques used in [5] and the unfolding periodic method, introduced in [10], adapted to this situation. In this way, we will able to pass to the limit and we obtain the following convergence result:

Theorem 1 *Let u^ϵ be the solution of problem (1). Assume to simplify that the non homogeneous term f^ϵ is given by $f^\epsilon(x,y) = f(x)\ \forall\ x \in (0,1)$, with $f \in L^2(0,1)$. Then, there exists $u_0 \in H^1(0,1)$ such that $\epsilon^{-1/2}||u^\epsilon - u_0||_{L^2(R^\epsilon)} \to 0$ and it is the unique solution of the following Neumann problem*

$$\begin{cases} \dfrac{1}{\mathscr{M}\left(\frac{1}{g+h_0}\right)\left(\mathscr{M}(g)+\mathscr{M}(h)\right)} u_{0xx} + u_0 = f, \quad x \in (0,1) \\ u'(0) = u'(1) = 0 \end{cases} \tag{5}$$

Remark 1 Notice that in case $h \equiv 0$ we recover the homogenized limit problem (3). On the other hand, if g is a constant function, $g \equiv g_0$, then the limit equation (5) coincides with the equation obtained in [5, Corollary 2.3].

The paper is organized as follows. In Sect. 2, we fix the notation, introduce the unfolding operator for this case and prove its main properties. In Sect. 3, we provide the proof of the main theorem.

2 Notation and Unfolding Operator

The domain R^ϵ is composed of two parts: one of them, R^ϵ_-, presents high oscillations and the other, R^ϵ_+, is a weakly oscillating domain, that is,

$$R^\epsilon_- = \{(x_1,x_2) \in \mathbb{R}^2 |\, x_1 \in (0,1),\ -\epsilon h(x_1/\epsilon^\beta) < x_2 < -\epsilon h_0\} \tag{6}$$

$$R^\epsilon_+ = \{(x_1,x_2) \in \mathbb{R}^2 |\, x_1 \in (0,1),\ -\epsilon h_0 < x_2 < \epsilon g(x_1/\epsilon^\alpha)\}. \tag{7}$$

We stress the fact that R^ϵ collapses to the interval $(0,1)$ when ϵ goes to 0. Therefore, we will consider the following norms: for $\varphi \in L^2(R^\epsilon)$ and $\psi \in H^1(R^\epsilon)$

$$|||\varphi|||_{L^2(R^\epsilon)} = \epsilon^{-1/2}||\varphi||_{L^2(R^\epsilon)},\ |||\psi|||_{H^1(R^\epsilon)} = \epsilon^{-1/2}||\psi||_{H^1(R^\epsilon)}.$$

An important tool for the analysis below is the unfolding operator for functions defined in R^ϵ_+, thin domains where the order of height is larger than the order of the oscillations. It will be used to interpret integrals over the domain R^ϵ_+ as integrals over a fixed domain. In the sequel we denote:

- $Y^* = \{(y_1,y_2) \in \mathbb{R}^2\ :\ 0 < y_1 < L_1,\ -h_0 < y_2 < g(y_1)\}$.
- $[x_1]_{L_1}$ denotes the unique integer such that $x_1 \in \left([x_1]_{L_1}L_1, ([x_1]_{L_1}+1)L_1\right)$.
- $\Lambda^\epsilon = Int\left\{\bigcup_{k=0}^{N_\epsilon}[\epsilon^\alpha k L_1, \epsilon^\alpha L_1(k+1)]\right\}$ where N_ϵ is the biggest integer such that $\epsilon^\alpha L_1(N_\epsilon+1) \leqslant 1$.
- $H^1_{L_1}(Y^*)$ is the space of functions $\varphi \in H^1(Y^*)$ which are $L_1 - periodic$ in the first variable.

Definition 1 For $\varphi \in L^2(R^\epsilon_+)$, the unfolding operator $\mathscr{T}_\epsilon(\varphi)$ is defined as follows:

$$\mathscr{T}_\epsilon(\varphi)(x_1, y_1, y_2) = \begin{cases} \varphi\Big(\epsilon^\alpha \Big[\frac{x_1}{\epsilon^\alpha}\Big]_{L_1} L_1 + \epsilon^\alpha y_1, \epsilon y_2\Big) & \text{for} \quad (x_1, y_1, y_2) \in \Lambda^\epsilon \times Y^* \\ 0 \qquad \text{for} \quad (x_1, y_1, y_2) \in (0,1)\backslash \Lambda^\epsilon \times Y^*. \end{cases}$$

Remark 2 In order to simplify the proposed method we assume that for all ϵ considered there exists an integer, N_ϵ, such that $\Lambda^\epsilon = (0,1)$.

Proposition 1 *1. Unfolding criterion for integrals (u.c.i.):*

$$\frac{1}{L_1}\int_{(0,1)\times Y^*} \mathscr{T}_\epsilon(\varphi)(x_1, y_1, y_2) dx_1 dy_1 dy_2 = \frac{1}{\epsilon}\int_{R^\epsilon_+} \varphi(x_1, x_2) dx_1 dx_2, \ \forall\, \varphi \in L^2(R^\epsilon_+). \tag{8}$$

2. *For every $\varphi \in L^2(R^\epsilon_+)$, $\mathscr{T}_\epsilon(\varphi) \in L^2\big((0,1)\times Y^*\big)$. In addition, the following relationship exits between their norms: $\|\mathscr{T}_\epsilon(\varphi)\|_{L^2\left((0,1)\times Y^*\right)} = \sqrt{L_1}\, |||\varphi|||_{L^2(R^\epsilon_+)}$.*
3. *For $\varphi \in H^1(R^\epsilon_+)$, one has $\nabla y_1 y_2 \mathscr{T}_\epsilon(\varphi) = \big(\epsilon^\alpha \mathscr{T}_\epsilon\big(\frac{\partial \varphi}{\partial x_1}\big), \epsilon \mathscr{T}_\epsilon\big(\frac{\partial \varphi}{\partial x_2}\big)\big)$.*
4. *Let $\varphi \in L^2(0,1)$. Then considering φ as a function defined in R^ϵ_+ we have $\mathscr{T}_\epsilon(\varphi) \longrightarrow \varphi \quad s - L^2\big((0,1)\times Y^*\big)$.*

Theorem 2 *Let φ^ϵ be in $H^1(R^\epsilon_+)$ for every ϵ, with $|||\varphi|||_{H^1(R^\epsilon)}$ uniformly bounded. Then, there exist a function φ in $H^1(0,1)$ and $\varphi_1 \in L^2\big((0,1); H_{L_1}(Y^*)\big)$ with $\frac{\partial \varphi_1}{\partial y_2} = 0$ such that, up to subsequences:*

$$\mathscr{T}_\epsilon(\varphi^\epsilon) \rightharpoonup \varphi \ w - L^2\big((0,1); H(Y^*)\big), \ \mathscr{T}_\epsilon\Big(\frac{\partial \varphi^\epsilon}{\partial x_1}\Big) \rightharpoonup \frac{\partial \varphi}{\partial x_1} + \frac{\partial \varphi_1}{\partial y_1} \, w - L^2\big((0,1)\times Y^*\big).$$

Proof We will give some ideas on how this result can be proved.

From Property 2 in Proposition 1 we know that there is a subsequence of $\mathscr{T}_\epsilon(\varphi^\epsilon)$, still denoted by $\mathscr{T}_\epsilon(\varphi^\epsilon)$ such that $\mathscr{T}_\epsilon(\varphi^\epsilon) \rightharpoonup \varphi \quad \mathrm{w} - L^2\big((0,1)\times H^1(Y^*)\big)$. Moreover, as a consequence of Property 3 in Proposition 1 we have φ does not depend on y_1 and y_2. In order to show that $\varphi \in H^1(0,1)$ we use similar arguments as in Theorem 2.9 in [1].

We can obtain the other convergence introducing the operator $Z_\epsilon := \frac{1}{\epsilon^\alpha}\Big(\mathscr{T}_\epsilon(\varphi^\epsilon) - \frac{1}{|Y^*|}\int_{Y^*} T_\epsilon(\varphi^\epsilon)\, dy_2 dy_1\Big)$ and arguing in the same way as in the proof of Proposition 3.5 in [10]. Observe that in this case $\frac{\partial \varphi_1}{\partial y_2} = 0$ since $\frac{\partial Z_\epsilon}{\partial y_2} = \epsilon^{1-\alpha}\mathscr{T}_\epsilon\Big(\frac{\partial \varphi^\epsilon}{\partial x_2}\Big)$ and $1 - \alpha > 0$. □

3 Proof of the Main Result

Before we start with the proof of the Theorem 1 we will state a technical result which will be used to define suitable test functions in order to pass to the limit.

Lemma 1 *Let w^ϵ the unique solution of*

$$\begin{cases} \triangle w^\epsilon = 0 \quad \text{in } Q_\epsilon, \\ w^\epsilon(x,0) = w_0(x), \quad \text{on } \Gamma_\epsilon, \\ \dfrac{\partial w^\epsilon}{\partial \nu} = 0, \quad \text{on } \partial Q_\epsilon \setminus \Gamma_\epsilon \end{cases} \tag{9}$$

where the domain Q_ϵ is a rectangle given by $Q_\epsilon = \{(x,y) \in \mathbb{R}^2 \mid -\epsilon^\beta < x < \epsilon^\beta,\ 0 < y < 1\}$, with $\beta > 1$, ν is the outward unit normal to ∂Q_ϵ, Γ_ϵ is the lower boundary of Q_ϵ and w_0 is a function in $H^1(-\epsilon^\beta, \epsilon^\beta)$. Then there exists a constant C, independent of ϵ and w_0, such that

$$\left\| \frac{\partial w^\epsilon}{\partial x} \right\|^2_{L^2(Q_\epsilon)} + \frac{1}{\epsilon^2} \left\| \frac{\partial w^\epsilon}{\partial y} \right\|^2_{L^2(Q_\epsilon)} \leq C\epsilon^{\beta-1} \left\| \frac{\partial w_0}{\partial x} \right\|^2_{L^2(-\epsilon^\beta, \epsilon^\beta)}. \tag{10}$$

Proof See [5] for details. □

Now, we are in conditions to prove the Theorem 1.

Proof The variational formulation of (1) is: find $u^\epsilon \in H^1(R^\epsilon)$ such that

$$\int_{R^\epsilon} \left\{ \frac{\partial u^\epsilon}{\partial x_1} \frac{\partial \varphi}{\partial x_1} + \frac{\partial u^\epsilon}{\partial x_2} \frac{\partial \varphi}{\partial x_2} + u^\epsilon \varphi \right\} dx_1 dx_2 = \int_{R^\epsilon} f\varphi dx_1 dx_2, \quad \forall \varphi \in H^1(R^\epsilon). \tag{11}$$

Taking $\varphi = u^\epsilon$ in (11) and using that $|||f|||_{L^2(R^\epsilon)} \leq C$, with C independent of ϵ, we get that $|||u^\epsilon|||_{L^2(R^\epsilon)} \leq C \quad \forall \epsilon > 0$. Therefore, the compactness Theorem 2 implies that there exist $u_0 \in H^1(0,1)$ and $u_1 \in L^2((0,1); H^1_{L_1}(Y^*))$ with $\frac{\partial u_1}{\partial y_2} = 0$ such that, up to subsequences:

$$\mathscr{T}_\epsilon(u^\epsilon) \rightharpoonup u_0 \, \mathrm{w}\text{-}L^2\big((0,1); H(Y^*)\big), \ \mathscr{T}_\epsilon\Big(\frac{\partial u^\epsilon}{\partial x_1}\Big) \rightharpoonup \frac{\partial u_0}{\partial x_1} + \frac{\partial u_1}{\partial y_1} \, \mathrm{w}\text{-}L^2\big((0,1)\times Y^*\big). \tag{12}$$

Taking into account the convergences above, using the change of variables $(x_1, x_2) \to (x_1, x_2/\epsilon)$ and the same argument as in [5] we obtain the convergence

$$|||u^\epsilon - u_0|||_{L^2(R^\epsilon)} \to 0 \text{ as } \epsilon \to 0. \tag{13}$$

Now we construct appropriate test functions which when used in the variational formulation (11) will allow us to pass to the limit. We begin by the construction of a partition of the interval $[0,1]$ which is essentially related to the oscillations in the lower boundary. Therefore, let us denote by M_ϵ the largest integer such that $M_\epsilon L_2 \epsilon^\beta < 1$, where L_2 is the period of the function h. For a fixed ϵ, we consider the partition $\{\gamma_{0,\epsilon}, \gamma_{1,\epsilon}, \dots, \gamma_{M_\epsilon+1,\epsilon}\}$ where $\gamma_{0,\epsilon} = 0$, $\gamma_{M_\epsilon+1,\epsilon} = 1$ and $\gamma_{n,\epsilon} \in [(n-1)L_2\epsilon^\beta, nL_2\epsilon^\beta]$ a point where the minimum of $h(\cdot/\epsilon^\beta)$ is attained, that is, $h(\gamma_{n,\epsilon}/\epsilon^\beta) = h_0$.

We define the test function as follows. With $\phi \in H^1(0,1)$, we consider $\varphi^\epsilon \in H^1(R^\epsilon)$ defined as

$$\varphi^\epsilon(x_1,x_2) = \begin{cases} X_n^\epsilon(x_1,x_2), & (x_1,x_2) \in R_-^\epsilon \cap Q_n^\epsilon, \quad n = 1,2,\dots \\ \phi(x_1), & (x_1,x_2) \in R_+^\epsilon \end{cases} \tag{14}$$

where Q_n^ϵ is the rectangle $Q_n^\epsilon = \{(x_1,x_2) \,|\, \gamma_{n,\epsilon} < x_1 < \gamma_{n+1,\epsilon},\ -\epsilon h_1 < x_2 < -\epsilon h_0\}$ and the function X_n^ϵ is the solution of the problem

$$\begin{cases} -\triangle X_n^\epsilon = 0, & \text{in } Q_n^\epsilon \\ \dfrac{\partial X_n^\epsilon}{\partial N^\epsilon} = 0, & \text{on } \partial Q_n^\epsilon \backslash \Gamma_n^\epsilon \\ X_n^\epsilon(x_1,x_2) = \phi(x_1), & \text{on } \Gamma_n^\epsilon \end{cases} \tag{15}$$

with $\Gamma_n^\epsilon = \{(x_1, -\epsilon h_0) : \gamma_{n,\epsilon} \le x_1 \le \gamma_{n+1,\epsilon}\}$. It follows from estimate (10) that

$$\left|\left|\left| \frac{\partial X_n^\epsilon}{\partial x_1} \right|\right|\right|^2_{L^2(Q_n^\epsilon)} + \left|\left|\left| \frac{\partial X_n^\epsilon}{\partial x_2} \right|\right|\right|^2_{L^2(Q_n^\epsilon)} \le C\epsilon^{\beta-1} ||\phi'||^2_{L^2(\gamma_{n,\epsilon},\gamma_{n+1,\epsilon})}. \tag{16}$$

Clearly, from the definition of φ^ϵ we have $\varphi^\epsilon(x_1,x_2) - \phi(x_1) = \int_0^{x_2} \frac{\partial \varphi^\epsilon}{\partial x_2}(x_1,s)\,ds$, Thus, taking into account (16) we obtain

$$|||\varphi^\epsilon - \phi|||_{L^2(R^\epsilon)} \to 0 \text{ as } \epsilon \to 0. \tag{17}$$

We now pass to limit in (11) by making use of the test function φ^ϵ defined above. In order to accomplish this, we rewrite the variational formulation as follows and we analyze the convergence of each integral as $\epsilon \to 0$.

$$\epsilon^{-1}\Big(\int_{R_+^\epsilon} \left\{ \nabla u^\epsilon \nabla \varphi^\epsilon \right\} + \int_{R_-^\epsilon} \left\{ \nabla u^\epsilon \nabla \varphi^\epsilon \right\} + \int_{R^\epsilon} u^\epsilon \varphi^\epsilon \Big) = \epsilon^{-1} \int_{R^\epsilon} f^\epsilon \varphi^\epsilon. \tag{18}$$

- First integrand. Using the unfolding criterion for integrals (8) and the convergences (12) we easily get:

$$\epsilon^{-1}\int_{R^\epsilon_+}\left\{\nabla u^\epsilon \nabla\varphi^\epsilon\right\}dx_1dx_2 \to \frac{1}{L_1}\int_{(0,1)\times Y^*}\left(\frac{\partial u_0}{\partial x_1}+\frac{\partial u_1}{\partial y_1}\right)\frac{\partial\phi}{\partial x_1}\,dx_1dy_1dy_2. \quad (19)$$

- Second integrand. From the definition of φ^ϵ, the Cauchy–Schwarz inequality and the inequality (16) we have,

$$\epsilon^{-1}\int_{R^\epsilon_-}\left\{\frac{\partial u^\epsilon}{\partial x_1}\frac{\partial\varphi^\epsilon}{\partial x_1}+\frac{\partial u^\epsilon}{\partial x_2}\frac{\partial\varphi^\epsilon}{\partial x_2}\right\}dx_1dx_2 \to 0. \quad (20)$$

- Third integrand

$$\epsilon^{-1}\int_{R^\epsilon} u^\epsilon\varphi^\epsilon dx_1dx_2 \to \int_0^1 \big(\mathscr{M}(g)+\mathscr{M}(h)\big)u_0\phi\,dx_1. \quad (21)$$

To prove this, observe that

$$\epsilon^{-1}\int_{R^\epsilon} u^\epsilon\varphi^\epsilon dx_1dx_2 = \epsilon^{-1}\int_{R^\epsilon_-}(u^\epsilon-u_0)\varphi^\epsilon\,dx_1dx_2+\epsilon^{-1}\int_{R^\epsilon_-}u_0(\varphi^\epsilon-\phi)\,dx_1dx_2$$
$$+\epsilon^{-1}\int_{R^\epsilon_-}u_0\phi\,dx_1dx_2+\epsilon^{-1}\int_{R^\epsilon_+}u^\epsilon\varphi^\epsilon dx_1dx_2.$$

From (13) and (17) it follows that the first two terms in the right hand side above go to 0. Moreover, the other two terms become

$$\epsilon^{-1}\int_{R^\epsilon_-}u_0\phi\,dx_1dx_2+\epsilon^{-1}\int_{R^\epsilon_+}u^\epsilon\varphi^\epsilon dx_1dx_2$$
$$=\int_0^1 u_0\phi\left(h\left(\frac{x_1}{\epsilon^\alpha}\right)-h_0\right)dx_1+\frac{1}{L_2}\int_{(0,1)\times Y^*}\mathscr{T}_\epsilon(u^\epsilon)\mathscr{T}_\epsilon(\phi)\,dx_1dy_1dy_2.$$

As the limit of $\mathscr{T}_\epsilon(u^\epsilon)$ does not depend on y_1 or y_2 thanks to the Average Convergence for Periodic Functions (see, e.g., [9, p. xvi]) we get (21).

- For the fourth integrand the computations are similar as the third one,

$$\epsilon^{-1}\int_{R^\epsilon} f\varphi^\epsilon dx_1dx_2 \to \int_0^1 \big(\mathscr{M}(g)+\mathscr{M}(h)\big)f\phi\,dx_1 \text{ as } \epsilon\to 0. \quad (22)$$

Therefore, using (19)–(22) we obtain the following equation:

$$\frac{1}{L_1}\int_{(0,1)\times Y^*}\Big(\frac{\partial u_0}{\partial x_1}+\frac{\partial u_1}{\partial y_1}\Big)\frac{\partial \phi}{\partial x_1}\,dx_1dy_1dy_2+\int_0^1\big(\mathscr{M}(g)+\mathscr{M}(h)\big)(u_0-f)\phi\,dx_1=0. \tag{23}$$

Finally, we obtain an explicit expression for $\frac{\partial u_1}{\partial y_1}$. To do this, we take as test function in (11) the function ψ^ϵ given by:

$$\psi^\epsilon(x_1,x_2)=\begin{cases} Y_n^\epsilon(x_1,x_2), & (x_1,x_2)\in R_-^\epsilon\cap Q_n^\epsilon, \quad n=1,2,\dots \\ v^\epsilon(x_1), & (x_1,x_2)\in R_+^\epsilon \end{cases} \tag{24}$$

where $v^\epsilon(x_1,x_2)=\epsilon^\alpha\phi(x_1)\psi(x_1/\epsilon^\alpha)$ with $\phi\in D(0,1)$ and $\psi\in H^1_{L_1}(0,L_1)$, and the function Y_n^ϵ is the solution of the problem

$$\begin{cases} -\triangle Y_n^\epsilon=0, & \text{in } Q_n^\epsilon \\ \dfrac{\partial Y_n^\epsilon}{\partial N^\epsilon}=0, & \text{on } \partial Q_n^\epsilon\backslash\Gamma_n^\epsilon \\ Y_n^\epsilon(x_1,x_2)=v^\epsilon(x_1), & \text{on } \Gamma_n^\epsilon. \end{cases} \tag{25}$$

Observe that by definition Y_n^ϵ satisfies the same estimates as (16). Then, we can argue as in (17) and we obtain

$$|||\psi^\epsilon-v^\epsilon|||_{L^2(R^\epsilon)}\to 0 \text{ as } \epsilon\to 0. \tag{26}$$

Taking ψ^ϵ as a function test in (18) and passing to the limit we get for the first term

$$\epsilon^{-1}\int_{R_+^\epsilon}\Big\{\nabla u^\epsilon\nabla\psi^\epsilon\Big\}dx_1dx_2\to\frac{1}{L_1}\int_{(0,1)\times Y^*}\Big(\frac{\partial u_0}{\partial x_1}+\frac{\partial u_1}{\partial y_1}\Big)\phi\frac{\partial\psi}{\partial x_1}\,dx_1dy_1dy_2. \tag{27}$$

While the other terms go to zero as $\epsilon\to 0$ by the properties of ψ^ϵ

$$\epsilon^{-1}\int_{R_-^\epsilon}\Big\{\frac{\partial u^\epsilon}{\partial x_1}\frac{\partial\psi^\epsilon}{\partial x_1}+\frac{\partial u^\epsilon}{\partial x_2}\frac{\partial\psi^\epsilon}{\partial x_2}u^\epsilon\psi^\epsilon-f\psi^\epsilon\Big\}dx_1dx_2\to 0\to 0 \text{ as } \epsilon\to 0. \tag{28}$$

Due to (27) and (28) we get at the limit

$$\int_{(0,1)\times Y^*}\Big(\frac{\partial u_0}{\partial x_1}(x_1)+\frac{\partial u_1}{\partial y_1}(x_1,y_1)\Big)\phi(x_1)\frac{\partial\psi}{\partial y_1}(y_1)\,dx_1dy_1dy_2=0.$$

By density, this equality holds true for all $\psi \in L^2\big((0,1); H^1_{L_1}(Y^*)\big)$ with $\frac{\partial \psi}{\partial y_2} = 0$. Observe that all functions do not depend on y_2. Then, we can write

$$\int_{(0,1)\times(0,L_1)} \Big(\frac{\partial u_0}{\partial x_1}(x_1) + \frac{\partial u_1}{\partial y_1}(x_1, y_1)\Big)\big(g(y_1) + h_0\big)\frac{\partial \psi}{\partial y_1}(x_1, y_1)\, dx_1 dy_1 = 0.$$

Treating x_1 as a parameter in the above equation we have:

$$-\frac{\partial}{\partial y_1}\Big(\frac{\partial u_1}{\partial y_1}(g + h_0)\Big) = \frac{\partial u_0}{\partial x_1}\frac{\partial g}{\partial y_1}.$$

Consequently, it verifies that $\frac{\partial u_1}{\partial y_1} = -\frac{\partial u_0}{\partial x_1} + \frac{C}{g+h_0}$. Moreover, since u_1 is L_1-periodic

$$0 = \int_{(0,L_1)} \frac{\partial u_1}{\partial y_1}\, dy_1 = -\frac{\partial u_0}{\partial x_1} + C\frac{1}{L_1}\int_{(0,L_1)} \frac{1}{g + h_0}\, dy_1 = -\frac{\partial u_0}{\partial x_1} + C\mathscr{M}\Big(\frac{1}{g + h_0}\Big).$$

Then, we get

$$\frac{\partial u_1}{\partial y_1} = \Big(-1 + \frac{1}{(g + h_0).\mathscr{M}(\frac{1}{g+h_0})}\Big)\frac{\partial u_0}{\partial x_1}.$$

Replacing $\frac{\partial u_1}{\partial y_1}$ by its value in Eq. (23) we obtain:

$$\int_{(0,1)} \frac{1}{\mathscr{M}(\frac{1}{g+h_0})}\frac{\partial u_0}{\partial x_1}\frac{\partial \phi}{\partial x_1}\, dx_1 dy_1 dy_2 + \int_0^1 \big(\mathscr{M}(g) + \mathscr{M}(h)\big)(u_0 - f)\phi\, dx_1 = 0. \tag{29}$$

From Lax-Milgram Theorem we know that u_0 is the unique solution of (29), which is the variational formulation of (5). This complete the proof of Theorem 1. □

Acknowledgements Both authors are partially supported by grant MTM2012-31298, MINECO, Spain and Grupo de Investigación CADEDIF, UCM. The second author, Manuel Villanueva-Pesqueira, also partially supported by a FPU fellowship (AP2010-0786) from the Government of Spain.

References

1. Allaire, G.: Homogenization and two-scale convergence. SIAM J. Math. Anal. **32**, 1482–1518 (1992)
2. Arrieta, J.M.: Spectral properties of Schrödinger operators under perturbations of the domain. Ph.D. thesis, Georgia Institute of Technology (1991)

3. Arrieta, J.M., Carvalho, A.N., Pereira, M.C., Da Silva, R.P.: Semilinear parabolic problems in thin domains with a highly oscillatory boundary. Nonlinear Anal-Theor. **74**(15), 5111–5132 (2011)
4. Arrieta, J.M., Pereira, M.C.: Homogenization in a thin domain with an oscillatory boundary. J. Math. Pures Appl. **96**(1), 29–57 (2011)
5. Arrieta, J.M., Pereira, M.C.: The Neumann problem in thin domains with very highly oscillatory boundaries. J. Math. Anal. Appl. **444**(1), 86–104 (2013)
6. Arrieta, J.M., Villanueva-Pesqueira, M.: Thin domains with doubly oscillatory boundary. Math. Method. Appl. Sci. **37**, 158–166 (2014). doi:10.1002/mma.2875
7. Arrieta, J.M., Villanueva-Pesqueira, M.: Locally periodic thin domains with varying period. C.R. Acad. Sci. Paris, Ser. I **352**, 397–403 (2014)
8. Bensoussan, A., Lions, J.L., Papanicolaou, G.: Asymptotic Analysis for Periodic Structures. North-Holland, Amsterdam (1978)
9. Cioranescu, D., Saint Jean Paulin, J.: Homogenization of Reticulated Structures. Springer, Berlin (1999)
10. Cioranescu, D., Damlamian, A., Griso, G.: The periodic unfolding method in homogenization. SIAM J. Math. Anal. **40**(4), 1585–1620 (2008)
11. Mel'nyk, T.A., Popov, A.V.: Asymptotic analysis of boundary-value problems in thin perforated domains with rapidly varying thickness. Nonlinear Oscil. **13**(1), 57–84 (2010)
12. Sánchez-Palencia, E.: Non-Homogeneous Media and Vibration Theory. Lecture Notes in Physics, vol. 127. Springer, Berlin (1980)

A Corrector Result for the Wave Equation with High Oscillating Periodic Coefficients

Juan Casado-Díaz, Julio Couce-Calvo, Faustino Maestre, and José Domingo Martín-Gómez

Abstract In the homogenization of a wave problem with oscillating coefficients in the diffusion term it is well known that the corresponding limit equation has the same structure with a diffusion term which agrees with the elliptic homogenized limit. Thus one can think that the oscillations of the solution of the wave equation are similar to the ones of the corresponding elliptic problem and then that the corrector for the elliptic problem is still a corrector for the wave problem. However in a paper by Brahim-Otsmane, Francfort and Murat, 1992, it was proved that this only holds if the initial data are "well posed". In general, it is necessary to add to the elliptic corrector another term depending on the initial data. In this paper we obtain this term in the case of a wave problem posed in $\mathbb{R}^N$ with periodic coefficients. This term is obtained using the two-scale convergence theory. It oscillates periodically in the space variable but almost periodically in the time one.

1 Introduction

In the present paper we consider a wave problem in $\mathbb{R}^N$ with periodic oscillating coefficients (see (1) below). The limit equation for this problem is well known [4, 5, 9, 11, 12] and consists in replacing the oscillating term which multiplies the second time derivative by its weak limit and the oscillating diffusion term by its elliptic homogenization limit [1,4,14–17]. However, as it is proved in [5], the elliptic corrector does not provides a corrector (i.e. an approximation of the solution in the strong topology of H^1) for the wave problem. Namely, it is necessary to add to the elliptic corrector another term which depends non locally on the initial data. However the structure of this new term is not know in general. The aim of this paper is to characterize it for problem (1) (the results in [12] hold for more general problems). This has been carried out in [10] for a more general problem where the coefficients of the equation also oscillate in the time variable and contain a first order

J. Casado-Díaz (✉) • J. Couce-Calvo • F. Maestre • J.D. Martín-Gómez
Facultad de Matemáticas, Departamento de Ecuaciones Diferenciales y Análisis Numérico, Universidad de Sevilla, C/ Tarfia s/n, 41012 Sevilla, Spain
e-mail: jcasadod@us.es; couce@us.es; fmaestre@us.es; jdmartin@us.es

F. Casas, V. Martínez (eds.), *Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 4, DOI 10.1007/978-3-319-06953-1_3

term which introduces a non-local term in the limit. In the case of (1), we show in Theorem 2 that this term can be obtained as a Fourier series of the form

$$\varepsilon \sum_{j\in\mathbb{Z}\setminus\{0\}} \sum_{k=1}^{k_j} z_j^k(t,x)\Phi_j^k(\frac{x}{\varepsilon})e^{i\lambda_j\frac{t}{\varepsilon}},$$

where, for every $j \in \mathbb{N}$, $\lambda_j = -\lambda_{-j}$ are the squared root of the eigenvalues of problem (3) and the functions Φ_j^k, $k = 1,\cdots,k_j$, a basis of the corresponding eigenfunctions space. For every $j \in \mathbb{Z}\setminus\{0\}$, the functions z_j^k are obtained as the solutions of a first order hyperbolic system. The proof of this result is obtained using the two-scale convergence theory [1, 7, 8, 15]. Some related results are also obtained in [6, 13]. We also refer to [2, 3], where the authors consider some other problems relative to the obtention of correctors for a wave equation with oscillating coefficients.

2 Homogenization and Corrector Results

The present section is devoted to the homogenization of a wave equation with oscillating coefficients in $\mathbb{R}^N$. Namely, for a real function $\rho \in L^\infty_\sharp(Y)$ and an Hermitian matrix $A \in L^\infty_\sharp(Y)^{N\times N}$ (the index $\sharp$ means periodicity and Y is the unitary cube $Y = (0,1)^N$) such that there exists $\alpha > 0$ satisfying

$$\rho(y) \geq \alpha, \quad A(y)\xi\cdot\bar{\xi} \geq \alpha|\xi|^2, \quad \forall\,\xi\in\mathbb{C}^N, \ \text{a.e. } y\in\mathbb{R}^N,$$

and functions $f \in L^1(0,T;L^2(\mathbb{R}^N))$, with $T > 0$, $u^0 \in H^1(\mathbb{R}^N)$, $u^1 \in C^1_\sharp(Y;H^1(\mathbb{R}^N))$, $v \in C^0_\sharp(Y;L^2(\mathbb{R}^N))$. Let us consider the wave problem in $Q_T = (0,T)\times\mathbb{R}^N$:

$$\begin{cases} \rho(\dfrac{x}{\varepsilon})\partial_{tt}^2 u_\varepsilon - \operatorname{div}_x\left(A(\dfrac{x}{\varepsilon})\nabla_x u_\varepsilon\right) = f \ \text{ in } Q_T \\ u_{\varepsilon|t=0} = u^0(x) + \varepsilon u^1(x,\dfrac{x}{\varepsilon}), \quad \partial_t u_{\varepsilon|t=0} = v(x,\dfrac{x}{\varepsilon}) \qquad \text{in } \mathbb{R}^N \\ u_\varepsilon \in L^\infty(0,T;H^1(\mathbb{R}^N)),\ \partial_t u_\varepsilon \in L^\infty(0,T;L^2(\mathbb{R}^N)). \end{cases} \tag{1}$$

Our work consists in describing the asymptotic behavior of the solutions of this problem when ε tends to zero. For this aim, we introduce the mean value

$$\rho_m = \int_Y \rho(y)\,dy,$$

and the homogenized matrix A_h of A (see e.g. [1,4,15]) by

$$A_h e_j = \int_Y A(y)(e_j + \nabla_y w_i)\, dy, \quad 1 \le j \le N,$$

with $e_1, \ldots, e_N$ the canonical basis in $\mathbb{R}^N$ and $w_1, \ldots, w_N$ the solutions of

$$\begin{cases} -\mathrm{div}_y A(e_j + \nabla_y w_j) = 0 \ \text{ in } \mathbb{R}^N \\ w_j \ \text{periodic of period } Y. \end{cases} \tag{2}$$

Moreover, we consider $\mu_0 = 0 < \mu_1 < \mu_2 < \cdots$ the eigenvalues of the problem

$$\begin{cases} -\mathrm{div}_y\,(A\nabla_y \Phi) = \mu_j \rho \Phi \ \text{ in } \mathbb{R}^N \\ \Phi \ \text{periodic of period } Y, \end{cases} \tag{3}$$

and W_j the corresponding eigenfunctions space, with dimension k_j.

For each W_j, we consider an ortonormal basis

$$\Phi_j^k \in W_j, \quad 1 \le k \le k_j, \qquad \int_Y \rho \Phi_j^k \Phi_j^l\, dy = \delta_{jl}, \quad 1 \le k, l \le k_j.$$

Moreover, given an eigenvalue μ_j, we denote

$$\lambda_j = \sqrt{\mu_j}, \quad \lambda_{-j} = -\sqrt{\mu_j}, \quad \Phi_{-j}^k = \Phi_j^k,$$

and

$$a_j^{lk} = \int_Y (\bar{A}\nabla_y \Phi_j^l \Phi_j^k - A\nabla_y \Phi_j^k \Phi_j^l)\, dy, \quad 1 \le l, k \le k_j,\ j \in \mathbb{Z} \setminus \{0\}, \tag{4}$$

where $\bar{A}$ is the conjugated matrix of A.

The limit problem corresponding to (1) is a classical result which we recall in the following theorem (see e.g. [5,9,11]).

Theorem 1 *The sequence of solutions u_ε of problem (1) satisfies*

$$u_\varepsilon \overset{*}{\rightharpoonup} u_0 \ \text{ in } L^\infty(0,T;H^1(\mathbb{R}^N)),$$

with u_0 the unique solution of the wave problem

$$\begin{cases} \rho_m \partial_{tt}^2 u_0 - div_x\big(A_h \nabla_x u_0\big) = f \ \text{ in } Q_T \\ u_{0|t=0} = u^0, \quad \partial_t u_{0|t=0} = \dfrac{1}{\rho_m} \displaystyle\int_Y \rho(y) v(x,y)\, dy \qquad \text{in } \mathbb{R}^N \\ u_0 \in L^\infty(0,T;H^1(\mathbb{R}^N)),\ \partial_t u_0 \in L^\infty(0,T;L^2(\mathbb{R}^N)). \end{cases} \tag{5}$$

We observe that in the previous theorem, the matrix A_h is the same matrix which appears in the homogenization of the elliptic problem

$$\begin{cases} -\text{div}\, A(\frac{x}{\varepsilon})\nabla u_\varepsilon = f \ \text{ in } \Omega \\ u_\varepsilon = 0 \ \text{ on } \partial\Omega, \end{cases}$$

where Ω is a bounded domain of $\mathbb{R}^N$. For this problem it is well known [1,4,14,15, 17] that for every $f \in H^{-1}(\Omega)$, the solution u_ε of (6) converges weakly in $H_0^1(\Omega)$ to the unique solution of

$$\begin{cases} -\text{div}\, A_h \nabla u_0 = f \ \text{ in } \Omega \\ u_0 = 0 \ \text{ on } \partial\Omega. \end{cases}$$

Moreover, if u is smooth enough (assuming f is smoother) and defining $u_1 : \Omega \times Y \to \mathbb{R}$ by

$$u_1(x, y) = \sum_{j=1}^{N} \partial_j u_0(x) w_j(y),$$

with the functions w_j given in (2), we have

$$u_\varepsilon - u_0 - \varepsilon u_1(x, \frac{x}{\varepsilon}) \to 0 \ \text{ in } H^1(\Omega).$$

This is a corrector result, i.e. an approximation of u_ε in the strong topology of $H^1(\Omega)$ and not only in the weak one. Since in problem (5) the matrix A_h is the same that in the elliptic problem, one can think that the elliptic corrector is also a corrector for the wave problem, i.e. that defining

$$u_1(t, x, y) = \sum_{j=1}^{N} \partial_{x_j} u_0(t, x) w_j(y), \ \text{ a.e. } (t, x, y) \in Q_T \times \mathbb{R}^N$$

with u_0 the solution of (5), and assuming u_0 smooth enough, we also have

$$u_\varepsilon - u_0 - \varepsilon u_1(t, x, \frac{x}{\varepsilon}) \to 0 \ \text{ in } H^1(Q_T), \tag{6}$$

where u_ε is the solution of (1). However, as it was observed in [5], this only holds if the initial data are "well chosen". In the case of problem (1), this means that the functions u^0, u^1, v must satisfy

$$\begin{cases} -\text{div}_y \left(A(\nabla_x u^0 + \nabla_y u^1) \right) = 0 \ \text{ in } \mathbb{R}^N \\ v \ \text{ is independent of } y. \end{cases} \tag{7}$$

In general (see [5]), a strong approximation in $H^1(Q_T)$ of the solutions for problem (1) can be constructed by introducing the solution z_ε of the wave problem

$$\begin{cases} \rho(\frac{x}{\varepsilon})\partial_{tt}^2 z_\varepsilon - \operatorname{div}_x\left(A(\frac{x}{\varepsilon})\nabla_x z_\varepsilon\right) = 0 \ \text{ in } Q_T \\ z_{\varepsilon|t=0} = \varepsilon\left(u^1(x,\frac{x}{\varepsilon}) - \sum_{j=1}^{N}\partial_j u^0(x) w_j(\frac{x}{\varepsilon})\right) \quad \text{in } \mathbb{R}^N \\ \partial_t z_{\varepsilon|t=0} = v(x,\frac{x}{\varepsilon}) - \int_Y v(x,y)\,dy \quad \text{in } \mathbb{R}^N \\ z_\varepsilon \in L^\infty(0,T;H^1(\mathbb{R}^N)),\ \partial_t z_\varepsilon \in L^\infty(0,T;L^2(\mathbb{R}^N)). \end{cases} \tag{8}$$

Then, we have

$$u_\varepsilon - u_0 - \varepsilon u_1(t,x,\frac{x}{\varepsilon}) - z_\varepsilon \to 0 \ \text{ in } H^1(Q_T). \tag{9}$$

However (9) cannot be considered as a corrector result because the structure of z_ε is not explicit. For each ε we need to solve a partial differential equation to obtain z_ε. It is interesting to remark that z_ε only depends on the initial conditions and it converges strongly to zero in $H^1(Q_T)$ when (7) is satisfied. This is in fact the proof that (6) holds if and only if (7) holds such as we said above.

Our aim in the present paper is to obtain an explicit corrector for (8) and then, thanks to (9) we get a corrector result for (1). More generally, we consider the problem

$$\begin{cases} \rho(\frac{x}{\varepsilon})\partial_{tt}^2 z_\varepsilon - \operatorname{div}_x\left(A(\frac{x}{\varepsilon})\nabla_x z_\varepsilon\right) = 0 \ \text{ in } Q_T \\ z_{\varepsilon|t=0} = \varepsilon\zeta(x,\frac{x}{\varepsilon}),\ \partial_t z_{\varepsilon|t=0} = \eta(x,\frac{x}{\varepsilon}) \quad \text{in } \mathbb{R}^N \\ z_\varepsilon \in L^\infty(0,T;H^1(\mathbb{R}^N)),\ \partial_t z_\varepsilon \in L^\infty(0,T;L^2(\mathbb{R}^N)), \end{cases} \tag{10}$$

where $\zeta \in C^1_\sharp(Y;H^1(\mathbb{R}^N))$, $\eta \in C^0_\sharp(Y;L^2(\mathbb{R}^N))$ and

$$\int_Y \eta(x,y)\,dy = 0, \ \text{ a.e. } x \in \mathbb{R}^N.$$

Using that the functions Φ_j^k, with $1 \le k \le k_j$, $j > 0$ are a basis of $H^1_\sharp(Y)/\mathbb{R}$, we can decompose

$$\zeta(x,y) = \sum_{j=1}^{\infty}\sum_{k=1}^{k_j}\zeta_j^k(x)\Phi_j^k(y) \ \text{ in } L^2(\mathbb{R}^N;H^1_\sharp(Y)/\mathbb{R})$$

$$\eta(x,y) = \sum_{j=1}^{\infty}\sum_{k=1}^{k_j}\eta_j^k(x)\Phi_j^k(y) \ \text{ in } L^2(\mathbb{R}^N;L^2_\sharp(Y)).$$

We have

Theorem 2 *Under the above conditions, we define $z : Q_T \times \mathbb{R} \times Y \to \mathbb{C}$ by*

$$z(t,x,s,y) = \sum_{j\in\mathbb{Z}\setminus\{0\}} \sum_{k=1}^{k_j} z_j^k(t,x)\Phi_j^k(y)e^{i\lambda_j s}, \tag{11}$$

where for every $j \in \mathbb{Z}\setminus\{0\}$, the coefficients z_j^k are the solution of the first order hyperbolic system

$$2i\lambda_j\delta_{kl}\partial_t z_j^k - \sum_{k=1}^{k_j} \mathrm{div}_x(a_j^{lk} z_j^k) = 0 \ \text{ in } Q_T, \tag{12}$$

with a_j^{lk} defined by (4), combined to the initial conditions

$$\begin{cases} z_{j\,|t=0}^k = \dfrac{1}{2}\left(\zeta_j^k - \dfrac{i}{\lambda_j}\eta_j^k\right) \ \text{ in } \mathbb{R}^N,\ 1 \le k \le k_j, & \text{if } j > 0 \\ z_{j\,|t=0}^k = \dfrac{1}{2}\left(\zeta_{-j}^k + \dfrac{i}{\lambda_{-j}}\eta_j^k\right) \ \text{ in } \mathbb{R}^N,\ 1 \le k \le k_j, & \text{if } j < 0. \end{cases}$$

Then, if ζ and η are smooth enough, we have

$$z_\varepsilon - \varepsilon z\left(t,x,\frac{t}{\varepsilon},\frac{x}{\varepsilon}\right) \to 0 \ \text{ in } H^1(Q_T). \tag{13}$$

Proof Since the initial conditions in (10) for z_ε and $\partial_t z_\varepsilon$ are bounded in $H^1(\mathbb{R}^N)$ and $L^2(\mathbb{R}^N)$ respectively, we deduce that z_ε is bounded in $L^\infty(0,T;H^1(\mathbb{R}^N)) \cap W^{1,\infty}(0,T;L^2(\mathbb{R}^N))$. Defining then $L^2_\sharp(Y)$ as the space of functions in $L^2_{loc}(\mathbb{R}^N)$, which are periodic of period Y and

$$H^1_\sharp(\mathbb{R}\times Y) = \Big\{ \sum_{p\in\mathbb{R}\setminus\{0\}} h_p(y)e^{ips} :\ h_p \in H^1_\sharp(Y), \\ \sum_{p\in\mathbb{R}\setminus\{0\}} \left(\|\nabla h_p\|^2_{L^2_\sharp(Y)^N} + |p|^2\|h_p\|^2_{L^2_\sharp(Y)}\right) < \infty\Big\},$$

we deduce that up to a subsequence there exists a function $z_0 \in L^\infty(0,T;H^1(\mathbb{R}^N)) \cap W^{1,\infty}(0,T;L^2(\mathbb{R}^N))$ and a function $z \in L^2(Q_T;H^1_\sharp(\mathbb{R}\times Y))$, such that

$$z_\varepsilon \overset{*}{\rightharpoonup} z \ \text{ in } L^\infty(0,T;H^1(\mathbb{R}^N)) \cap W^{1,\infty}(0,T;L^2(\mathbb{R}^N))$$

$$\partial_t z_\varepsilon \overset{2-s}{\rightharpoonup} \partial_t z_0 + \partial_s z \tag{14}$$

$$\nabla_x z_\varepsilon \overset{2-s}{\rightharpoonup} \nabla_x z_0 + \nabla_y z, \tag{15}$$

where we recall [1, 7, 8, 15] that a bounded sequence g_ε in $L^2(Q_T)$ is said that two-scale converges to a function $g \in L^2(Q_T; L^2_\sharp(\mathbb{R} \times Y))$, and it is noted by $g_\varepsilon \overset{2-s}{\rightharpoonup} g$, if for every function $\varphi \in C_c^\infty(Q_T)$, every $\Phi \in L^2_\sharp(Y)$ and every $p \in \mathbb{R}$, we have

$$\int_{Q_T} g_\varepsilon(t,x)\varphi(t,x)\Phi(\frac{x}{\varepsilon})e^{ip\frac{t}{\varepsilon}}\,dxdt \to \int_{Q_T\times Y} M_s(g(t,x,y,s)e^{ips})\varphi(t,x)\Phi(y)\,dtdxdy,$$

where

$$M_s(g(t,x,y,s)e^{ips}) = \lim_{r\to\infty}\frac{2}{r}\int_{-r}^{r} g(t,x,y,s)e^{ips}\,ds.$$

The problem is to characterize these functions z_0 and z. For this purpose, we use in (10) two types of test functions (see [10] for the details).

Firstly, we take a test function of the form $\varphi_0(t,x) + \varepsilon\varphi_1(t,x)\Phi(\frac{x}{\varepsilon})e^{ip\frac{t}{\varepsilon}}$, with $\varphi_0, \varphi_1 \in C_c^1(Q_T)$, to deduce

$$z_0 = 0 \ \text{ a.e. in } Q_T$$

$$\rho\partial^2_{ss}z - \mathrm{div}_y(A\nabla_y z) = 0 \ \text{ in } Q_T \times \mathbb{R} \times \mathbb{R}^N.$$

In particular, the second equation implies that z is of the form given by (11). The problem is to characterize the functions z_j^k. For this purpose, we use a second class of test functions, which are of the form $\varphi(t,x)\Phi_j^k(\frac{x}{\varepsilon})e^{i\lambda_j\frac{t}{\varepsilon}}$, with $\varphi \in C_c^1(Q_T)$. Denoting $\psi(t,x,s,y) = \varphi(t,x)\Phi_j^k(y)e^{i\lambda s}$, we get

$$\int_{Q_T}\int_Y M_s\big(-2\rho\partial_s u_1\partial_t\psi + A\nabla_y z\cdot\nabla_x\psi - \mathrm{div}_x(\overline{A}\nabla_y\psi)\,z\big)\,dydtdx = 0.$$

From this equation, we conclude that the functions z_j^k satisfy (12).

On the other hand, we can also show that z satisfies the initial conditions

$$z_{|t=0} = \zeta \ \text{ in } L^2(\mathbb{R}^N; H^1_\sharp(Y)/\mathbb{R}), \quad \partial_s z_{|t=0} = \eta \ \text{ in } L^2(\mathbb{R}^N; L^2_\sharp(Y)/\mathbb{R}),$$

which gives the initial conditions (13) for the functions z_j^k.

We have then proved that the function z which appears in convergences (14), (15) agrees with the function z defined in the statement of Theorem 2. From these convergences and assuming ζ and η (and then z) smooth enough, we can now pass to the limit in the energy identity corresponding to Eq. (10) to prove (13). □

Acknowledgements The authors have been partially supported by the project MTM2011-24457 of the "Ministerio de Economía y Competitividad" of Spain.

References

1. Allaire, G.: Homogenization and two-scale convergence. SIAM J. Math. Anal. **23**, 1482–1518 (1992)
2. Allaire, G., Friz, L.: Localization of high-frequence waves propagating in a locally periodic medium. Proc. R. Soc. Edinb. A. **140**, 897–926 (2010)
3. Allaire, G., Palombaro, M., Rauch, J.: Diffractive behavior of the wave equation in periodic media: weak convergence analysis. Annali di Matematica. **188**, 561–589 (2009)
4. Bensoussan, A., Lions, J.L., Papanicolau, G.: Asymptotic Analysis for Periodic Structures. North Holland, Amsterdam/New York (1978)
5. Brahim-Otsmane, S., Francfort, G.A., Murat, F.: Correctors for the homogenization of the wave and heat equations. J. Math. Pures Appl. **71**, 197–231 (1992)
6. Brassart, M., Lenczner, M.: A two scale model for the periodic homogenization of the wave equation. J. Math. Pures Appl. **93**, 474–517 (2010)
7. Casado-Díaz, J., Gayte, I.: A general compactness result and its application to the two-scale convergence of almost periodic functions. C. R. Acad. Sci. Paris I. **323**, 329–334 (1996)
8. Casado-Díaz, J., Gayte, I.: The two-scale convergence method applied to generalized Besicovitch spaces. Proc. R. Soc. Lond. A Mat. **458**, 2925–2946 (2002)
9. Casado-Díaz, J., Couce-Calvo, J., Maestre, F., Martín-Gómez, J.D.: Homogenization and corrector for the wave equation with discontinuos coefficients in time. J. Math. Anal. Appl. **379**, 664–681 (2011)
10. Casado-Díaz, J., Couce-Calvo, J., Maestre, F., Martín-Gómez, J.D.: Homogenization and correctors for the wave equation with periodic coefficients. Math. Models Methods Appl. Sci. (2014). doi:10.1142/S0218202514500031
11. Colombini, F., Spagnolo, S.: On the convergence of solutions of hyperbolic equations. Commun. Part. Diff. Equ. **3**, 77–103 (1978)
12. Francfort, G.A., Murat, F.: Oscillations and energy densities in the wave equation. Commun. Part. Diff. Equ. **17**, 1785–1865 (1992)
13. Lenczner, M., Kader, M., Perrier, P.: Two-scale model of the wave equation with oscillating coefficients. C. R. Acad. Sci. II, Mech. Phys. Astron. **328**, 335–340 (2000)
14. Murat, F.: H-convergence. Séminaire d'Analyse Fonctionnelle et Numérique, 1977–78 (Université d'Alger, multicopied, 34pp.) English translation: Murat, F., Tartar, L.: H-convergenc. In: Cherkaev, L., Kohn, R.V. (eds.) Topics in the Mathematical Modelling of Composite Materials. Progress in Nonlinear Differential Equations and Their Applications, vol. 31, pp. 21–43. Birkaüser, Boston (1998)
15. Nguetseng, G.: A general convergence result for a functional related to the theory of homogenization. SIAM J. Math. Anal. **20**, 608–623 (1989)
16. Spagnolo, S.: Sulla convergenza di soluzioni di equazioni paraboliche ed ellittiche. Ann. Sc. Norm. Sup. Pisa Cl. Sci. **22**, 571–597 (1968)
17. Tartar, L.: The general theory of the homogenization: a personalized introduction. Springer, Heidelberg/New York (2009)

Weak Solutions to a Nonuniformly Elliptic PDE System in the Harmonic Regime

María Teresa González Montesinos and Francisco Ortegón Gallego

Abstract We study the existence of weak solutions to a nonlinear strongly coupled parabolic–elliptic PDEs arising in the heating induction-conduction process of steel hardening. In this setting, our major concern is to consider the case when the electric conductivity is nonuniformly elliptic which, together with a right hand side in L^1 in the energy balance equation, yields to a difficult theoretical situation. The existence result gives a weak solution to a similar PDEs system where the energy balance equation has been perturbed by a measure term.

1 Introduction

The aim of this work is to analyze the existence of weak solutions to a nonlinear PDEs system arising in the heating induction-conduction process of a steel workpiece [7, 8, 10, 11, 13]. Since we are dealing with high oscillating sinusoidal in time for both electric potential and magnetic vector potential, we introduce a change of variables separating the two time scales. This leads us to a new PDEs system, the so-called harmonic regime, namely

$$-\nabla\cdot(\sigma(\theta)\nabla\varphi) = i\lambda\omega\nabla\cdot(\sigma(\theta)\boldsymbol{A}) + \nabla\cdot\left(\sigma(\theta)\nabla\varphi^0\right) \text{ in } \Omega_T = \Omega\times(0,T), \quad (1)$$

$$i\omega\sigma(\theta)\boldsymbol{A} + L(\boldsymbol{A}) = -\sigma(\theta)\nabla\varphi \text{ in } D_T = D\times(0,T), \quad (2)$$

$$\varphi = 0 \text{ on } \Gamma_0\times(0,T), \quad \frac{\partial\varphi}{\partial n} = -i\lambda\omega\boldsymbol{A}\cdot n \text{ on } \Gamma_1\times(0,T), \ + \text{b.c. on } \boldsymbol{A}, \quad (3)$$

M.T. González Montesinos (✉)
Departamento de Matemática Aplicada I, Universidad de Sevilla, Avda. Reina Mercedes s/n, 41012 Sevilla, Spain
e-mail: mategon@us.es

F. Ortegón Gallego
Facultad de Ciencias, Departamento de Matemáticas, Universidad de Cádiz, Campus del Río San Pedro, 11510 Puerto Real, Spain
e-mail: francisco.ortegon@uca.es

F. Casas, V. Martínez (eds.), *Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 4, DOI 10.1007/978-3-319-06953-1_4

$$\rho c_\varepsilon \frac{\mathrm{d}\theta}{\mathrm{d}t} - \nabla\cdot(\kappa(\theta)\nabla\theta) = \frac{\sigma(\theta)}{2}|i\,\omega\boldsymbol{A} + \nabla\varphi|^2 + G \text{ in } \Omega_T, \tag{4}$$

$$\frac{\partial\theta}{\partial n} = 0 \text{ on } \partial\Omega\times(0,T), \quad \theta(\cdot,0) = \theta_0 \text{ in } \Omega. \tag{5}$$

In this context, $\Omega,\ D\subset\mathbb{R}^3$ are open, bounded, connected and Lipschitz-continuous sets such that $\bar{\Omega}\subset D$, $\partial\Omega = \Gamma_0\cup\Gamma_1$ is a smooth partition of the boundary of Ω. The unknowns are the electric potential, φ, the magnetic vector potential, $\boldsymbol{A}$, and the temperature, θ; σ and κ stand for the electric and thermal conductivities, respectively, ω is the frequency, θ_0 the initial temperature and $L\in\mathscr{L}(\mathbb{X},\mathbb{X}')$ is an elliptic operator defined on a certain Hilbert space $\mathbb{X}$ with values on its dual space $\mathbb{X}'$. Also, ρ is the density and c_ε is the specific heat at constant pressure. Finally, $\varphi^0\in L^2(H^1(\Omega))$ is a given function with zero flux gradient on Γ_1 and i is the imaginary unity.

In this work we have included in (1) the divergence term $i\,\lambda\omega\nabla\cdot(\sigma(\theta)\boldsymbol{A})$, where $\lambda\in\left[0,1-\frac{1}{\omega}\right)$ is a parameter. Usually, this term is not taken into account, that is $\lambda = 0$. Notice that in the original model we have $\lambda = 1$ (cf. [2, 3]).

This work is organized as follows. In Sect. 2, we describe the notation used along this paper, introduce some functional spaces, enumerate the hypotheses on data and give the main result. In Sect. 3 we sketch the proof of the main result by introducing approximate problems, deriving the necessary a priori estimates and, finally, passing the limit.

2 Notation, Assumptions and Main Result

Let $\Omega_1,\Omega_2\subset\mathbb{R}^3$ be two open bounded, connected and Lipschitz-continuous sets such that $S = \bar{\Omega}_1\cap\bar{\Omega}_2\neq\emptyset$ is a smooth surface. We then consider the set of conductors $\Omega = \Omega_1\cup\Omega_2\cup\mathrm{int}(S)$ where $\mathrm{int}(S)$ means the interior of S within the induced topology. Ω_1 is the steel workpiece whereas Ω_2 is the copper inductor; since $S\neq\emptyset$, the workpiece and the inductor are put in contact so that Ω itself becomes the coil. Let $\Gamma_0\subset\partial\Omega_2$ be a smooth surface.

For a normed linear space V, we put $\boldsymbol{V} = (V)^3$. Also, if X is a Banach space, we write $L^p(X) = L^p(0,T;X)$ and $W^{1,p}(X) = W^{1,p}(0,T;X)$, where p' is the conjugate exponent of p. Let V be the complex valued Hilbert space $V = \{\phi\in H^1(\Omega)\,/\,\phi = 0 \text{ on } \Gamma_0\}$ provided with the norm $\|\phi\|_V = \left(\int_\Omega|\nabla\phi|^2\right)^{1/2}$, which is equivalent to the standard norm in $H^1(\Omega)$ on V.

We also consider a complex valued Hilbert space $\mathbb{X}$ such that $\boldsymbol{H}_0^1(D)\subset\mathbb{X}\subset\boldsymbol{H}^1(D)$ where lies the magnetic vector potential $\boldsymbol{A}$. Obviously, the space $\mathbb{X}$ is related to the boundary conditions of $\boldsymbol{A}$. For instance, it may take the form

$$\mathbb{X} = \{\boldsymbol{v}\in\boldsymbol{H}^1(D)\,/\,\boldsymbol{v} = \boldsymbol{0} \text{ on } \partial D\}, \text{ or}$$

$$\mathbb{X} = \{\boldsymbol{v}\in\boldsymbol{H}^1(D)\,/\,\nabla\cdot\boldsymbol{v} = \boldsymbol{0} \text{ in } D,\ \boldsymbol{v}\times n{=}0 \text{ on } \partial D\} \text{ where } \partial D{\in}C^{1,1} \text{ in this case.}$$

On the other hand, the elliptic operator $L \in \mathscr{L}(\mathbb{X}, \mathbb{X}')$ is given by

$$L(\boldsymbol{v}) = \nabla \times \left(\frac{1}{\mu}\nabla \times \boldsymbol{v}\right) - \delta\nabla(\nabla \cdot \boldsymbol{v}),$$

where μ is the magnetic permeability (a positive bounded function) and $\delta > 0$ a constant value.

In the analysis of parabolic problems with right hand side in L^1 it is useful the next result (see [14])

Lemma 1 *Let X, B and Y be three Banach spaces such that $X \hookrightarrow B \hookrightarrow Y$, all embeddings being continuous and the injection $X \hookrightarrow B$ compact. For $1 \leq p, q < +\infty$ define $\mathscr{W}$ to be the Banach space $\mathscr{W} = \left\{v \in L^p(X) \,/\, \frac{\mathrm{d}v}{\mathrm{d}t} \in L^q(Y)\right\}$. Then, the embedding $\mathscr{W} \hookrightarrow L^p(B)$ holds and is compact.*

The assumptions on data now follows.

(H.1) $\sigma : D \times \mathbb{R} \to \mathbb{R}$ is given by

$$\sigma(x,s) = \begin{cases} \sigma^{(1)}(s) \text{ if } x \in \Omega_1, s \in \mathbb{R} \\ \sigma^{(2)}(s) \text{ if } x \in \Omega_2, s \in \mathbb{R}, \\ 0 \qquad \text{if } x \in D \setminus \bar{\Omega}, s \in \mathbb{R}, \end{cases}$$

where $\sigma^{(1)},\ \sigma^{(2)} \in C(\mathbb{R})$, and there exist some constant $C_1,\ C_2,\ K_1,\ K_2 > 0$ and $0 < \alpha < 5/3$ such that for all $s \in \mathbb{R}$ we have

$$0 < \frac{C_1}{1 + |s|^\alpha} \leq \sigma^{(1)}(s) \leq C_2, \quad K_1 \leq \sigma^{(2)}(s) \leq K_2.$$

(H.2) $\rho = \rho_i$ and $c_\varepsilon = c_\varepsilon^i$ in Ω_i, $i = 1, 2$ where $\rho_1, \rho_2, c_\varepsilon^1, c_\varepsilon^2 \in \mathbb{R}$ are positive constant values.

(H.3) $\kappa : \Omega \times \mathbb{R} \mapsto \mathbb{R}$ is a Carathéodory function and there exist two constant values κ_1 and κ_2 such that, almost everywhere $x \in \Omega$ and for all $s \in \mathbb{R}$, we have $0 < \kappa_1 \leq \kappa(x,s) \leq \kappa_2$.

(H.4) $L \in \mathscr{L}(\mathbb{X}, \mathbb{X}')$ and there exists a constant value $\alpha > 0$ such that, for all $\boldsymbol{v} \in \mathbb{X}$,

$$\langle L(\boldsymbol{v}), \bar{\boldsymbol{v}}\rangle_{\mathbb{X}',\mathbb{X}} \geq \alpha \|\boldsymbol{v}\|_{\mathbb{X}}^2.$$

(H.5) $\lambda \in \left[0, 1 - \frac{1}{\omega}\right)$.

(H.6) $\varphi^0 \in L^2(H^1(\Omega))$ and $\frac{\partial \varphi^0}{\partial n} = 0$ on $\Gamma_1 \times (0,T)$.

(H.7) $G \in L^1(\Omega_T)$.

(H.8) $\theta_0 \in L^1(\Omega)$.

The main result of this paper is the next

Theorem 1 *Under the assumptions (H.1)–(H.8) there exist three measurable functions* φ, $\theta : \Omega_T \mapsto \mathbb{R}$, $\boldsymbol{A} : D_T \mapsto \mathbb{R}^3$, *and a Radon measure* $\boldsymbol{\mu} \in \mathscr{M}(\Omega_T)$ *such that*

$$\varphi \in L^r(W^{1,r}(\Omega)), \text{ for all } r \in [1, 10/(5+3\alpha)); , \varphi = 0 \text{ on } \Gamma_0, \tag{6}$$

$$\sigma(\theta)^{1/2}\nabla\varphi \in L^2(\boldsymbol{L}^2(\Omega)), \boldsymbol{A} \in L^2(\mathbb{X}), \tag{7}$$

$$\int_{\Omega_T} \sigma(\theta)\nabla\varphi \cdot \nabla\bar{\phi} = -i\omega\lambda \int_{\Omega_T} \sigma(\theta)\boldsymbol{A} \cdot \nabla\bar{\phi} + \int_{\Omega_T} \sigma(\theta)\nabla\varphi^0\nabla\bar{\phi}, \phi \in L^2(V), \tag{8}$$

$$i\omega \int_{\Omega_T} \sigma(\theta)\boldsymbol{A} \cdot \bar{\boldsymbol{v}} + \int_0^T \langle L(\boldsymbol{A}), \bar{\boldsymbol{v}}\rangle_{\mathbb{X}',\mathbb{X}} = -\int_{\Omega_T} \sigma(\theta)\nabla\varphi \cdot \bar{\boldsymbol{v}}, \ \boldsymbol{v} \in L^2(\mathbb{X}), \tag{9}$$

$$\theta \in L^p(W^{1,p}(\Omega)) \cap C([0,T]; (W^{1,p'}(\Omega))'), \text{ for all } p \in [1, 5/4), \tag{10}$$

$$\theta(\cdot, 0) = \theta_0 \text{ in } \Omega, \tag{11}$$

$$\begin{aligned} -\int_{\Omega_T} \rho c_\varepsilon \theta\zeta_{,t} + \int_{\Omega_T} \kappa(\theta)\nabla\theta\nabla\zeta &= \int_{\Omega_T} \left[\frac{\sigma(\theta)}{2}|i\omega\boldsymbol{A} + \nabla\varphi|^2 + G\right]\zeta \\ &+ \int_{\Omega_T} \zeta \mathrm{d}\boldsymbol{\mu} + \int_\Omega \theta_0(x)\zeta(x,0), \\ &\text{for all } \zeta \in \mathscr{D}(\bar{\Omega}_T) \text{ such that } \zeta(\cdot, T) = 0 \text{ in } \Omega. \end{aligned} \tag{12}$$

Remark 1 Due to (H.1), the function σ is not uniformly elliptic. In particular, we cannot derive the regularity $\varphi \in L^2(V)$. This is also related with the "strange term" $\boldsymbol{\mu}$ appearing in the equation for the temperature.

3 Proof of the Main Result

In order to prove the Theorem 1 we first introduce a sequence of approximate problems then deduce some a priori estimates. The approximate problems regularize the solution in three different ways: (1) introduction of a time derivative term in the equations of φ and $\boldsymbol{A}$ to assure the measurability of both functions when passing to the limit; (2) modification of the electric conductivity in order to deal with uniformly elliptic operators; and (3) truncation of the L^1 terms in the energy equation.

3.1 Approximate Problems

For $k \in \mathbb{N}$ we introduce the approximate the function σ as follows

$$\sigma_k(x,s) = \begin{cases} \sigma^{(1)}(s) + \dfrac{1}{k} & \text{if } x \in \Omega_1,\ s \in \mathbb{R}, \\ \sigma^{(2)}(s) & \text{if } x \in \Omega_2,\ s \in \mathbb{R}, \\ 0 & \text{if } x \in D \setminus \bar{\Omega},\ s \in \mathbb{R}. \end{cases}$$

We also use the truncation function T_k at height $k > 0$, that is

$$T_k(s) = \begin{cases} -k, \text{ if } s < -k, \\ s, \text{ if } |s| \le k, \\ k, \text{ if } s > k. \end{cases}$$

The approximate problems of (1)–(5) are given by

$$\varphi_k \in L^2(V),\ \boldsymbol{A}_k \in L^2(\mathbb{X}),\ \theta_k \in L^2(H^1(\Omega)) \cap C\left([0,T]; L^2(\Omega)\right), \tag{13}$$

$$\frac{1}{k}\frac{\mathrm{d}\varphi_k}{\mathrm{d}t} - \nabla\cdot(\sigma_k(\theta_k)\nabla\varphi_k) = i\lambda\omega\nabla\cdot(\sigma_k(\theta_k)\boldsymbol{A}_k) + \nabla\cdot\left(\sigma_k(\theta_k)\nabla\varphi^0\right) \text{ in } \Omega_T = \Omega\times(0,T), \tag{14}$$

$$\frac{1}{k}\frac{\mathrm{d}\boldsymbol{A}_k}{\mathrm{d}t} + \omega(i+\omega)\sigma_k(\theta_k)\boldsymbol{A}_k + (1-i\omega)L(\boldsymbol{A}_k) = -(1-i\omega)\sigma_k(\theta_k)\nabla\varphi_k \text{ in } D_T, \tag{15}$$

$$\varphi_k = 0 \text{ on } \Gamma_0\times(0,T)\ ,\ \frac{\partial\varphi_k}{\partial n} = -i\lambda\omega\boldsymbol{A}_k\cdot n \text{ on } \Gamma_1\times(0,T), \tag{16}$$

$$\boldsymbol{A}_k = 0 \text{ on } \partial D\times(0,T), \tag{17}$$

$$\varphi_k(\cdot,0) = 0 \text{ in } \Omega, \quad \boldsymbol{A}_k(\cdot,0) = 0 \text{ in } D, \tag{18}$$

$$\rho c_\varepsilon\frac{\mathrm{d}\theta_k}{\mathrm{d}t} - \nabla\cdot(\kappa(\theta_k)\nabla\theta_k) = F_k \text{ in } \Omega_T, \tag{19}$$

$$\frac{\partial\theta_k}{\partial n} = 0 \text{ on } \partial\Omega\times(0,T), \quad \theta_k(\cdot,0) = T_k(\theta_0) \text{ in } \Omega, \tag{20}$$

where $F_k = \dfrac{\sigma_k(\theta_k)}{2}T_k\left(|i\omega\boldsymbol{A}_k + \nabla\varphi_k|^2\right) + T_k(G)$ and $D_T = D\times(0,T)$.

For the system (13)–(20) it can be shown the following existence result [12].

Lemma 2 *For every $k \geq 1$, there exists a weak solution $(\varphi_k, \boldsymbol{A}_k, \theta_k)$ to problem (13)–(20).*

Remark 2 Since we are dealing with complex valued function spaces, the key point is to define the right bilinear elliptic form related to the system for $(\varphi_k, \boldsymbol{A}_k)$ for a given θ_k. From that point on, the proof of Lemma 2 is a straightforward application of J. L. Lions' theorem together with Schauder's fixed point theorem.

3.2 A Priori Estimates

For the solution of (13)–(20) it is easy to obtain the following estimates

$$\int_{\Omega_T} \sigma_k(\theta_k)|\boldsymbol{A}_k|^2 \leq \frac{C_2}{\omega^2} \int_{\Omega_T} \sigma_k(\theta_k)|\nabla\varphi_k|^2. \tag{21}$$

$$\int_0^T \|\boldsymbol{A}_k\|_{\mathbb{X}}^2 \leq \frac{C_2}{\alpha\omega} \int_{\Omega_T} \sigma_k(\theta_k)|\nabla\varphi_k|^2. \tag{22}$$

$$\int_{\Omega_T} \sigma_k(\theta_k)|\nabla\varphi_k|^2 \leq C_\lambda \|\varphi^0\|_{L^2(H^1(\Omega))}^2, \tag{23}$$

where

$$\lim_{\lambda \to (1-1/\omega)^-} C_\lambda = +\infty.$$

From these estimates we deduce

$$\left(\sigma_k(\theta_k)^{1/2}\boldsymbol{A}_k\right) \text{ is bounded in } L^2(\boldsymbol{L}^2(\Omega)), \tag{24}$$

$$(\boldsymbol{A}_k) \text{ is bounded in } L^2(\mathbb{X}).$$

On the other hand, since $\mathbb{X} \hookrightarrow \boldsymbol{L}^2(D)$ there exists a constant $C > 0$ such that $\|\boldsymbol{v}\|_{\boldsymbol{L}^2(\Omega)} \leq \|\boldsymbol{v}\|_{\boldsymbol{L}^2(D)} \leq C\|\boldsymbol{v}\|_{\mathbb{X}}$, for all $\boldsymbol{v} \in \mathbb{X}$. Thus,

$$(\boldsymbol{A}_k) \text{ is bounded in } L^2(\boldsymbol{L}^2(\Omega)).$$

From (23) and (24) it yields

$$(F_k) \text{ is bounded in } L^1(\Omega_T),$$

and thus, owing to (H.7), we obtain

$$(\theta_k) \text{ is bounded in } L^p(W^{1,p}(\Omega)), \text{ for all } 1 \leq p < 5/4, \tag{25}$$

Remark 3 In [4] it was shown that (25) holds true when dealing with homogeneous Dirichlet boundary conditions. In the case of homogeneous Neumann boundary conditions, this result was shown by Clain in [6].

According to (H.1) and (25) we obtain that $(\kappa(\theta_k)\nabla\theta_k)$ is bounded in $L^p(\boldsymbol{L}^p(\Omega))$. Therefore $(\nabla\cdot(\kappa(\theta_k)\nabla\theta_k))$ is bounded in $L^1((W^{1,p'}(\Omega))')$. Since $1 \leq p < 5/4$, Sobolev's embedding implies in particular that

$$L^1(\Omega) \hookrightarrow (W^{1,p'}(\Omega))'$$

and, in conclusion,

$$\left(\frac{\mathrm{d}\theta_k}{\mathrm{d}t}\right) \text{ is bounded in } L^1((W^{1,p'}(\Omega))'), \text{ for all } 1 \leq p < 5/4. \tag{26}$$

3.3 Passing to the Limit

Choosing $1 \leq q < p^* = 3p/(3-p)$, $X = W^{1,p}(\Omega)$, $B = L^q(\Omega)$ and $Y = (W^{1,p'}(\Omega))'$, and since the embeddings $X \hookrightarrow B$ and $B \hookrightarrow Y$ are continuous and compact, respectively, from Lemma 1 it yields that the space

$$\mathscr{W} = \left\{ v \in L^p(W^{1,p}(\Omega)) \,/\, \frac{\mathrm{d}v}{\mathrm{d}t} \in L^1((W^{1,p'}(\Omega))') \right\}$$

is compactly embedded in $L^p(L^q(\Omega))$. Moreover, since $1 \leq p < 5/4$ and $1 \leq q < 15/7$, and thanks to (25) and (26), we deduce that the sequence (θ_k) is relatively compact in $L^p(L^q(\Omega))$, for $1 \leq p < \frac{5}{4}$ and $1 \leq q < \frac{15}{7}$. Therefore, we may extract a subsequence, still denoted in the same way, such that $\theta_k \to \theta$ strongly in $L^p(L^q(\Omega))$ and almost everywhere in Ω_T. Consequently $\sigma_k(\theta_k) \rightharpoonup \sigma(\theta)$ in $L^\infty(\Omega_T)$-weak–$*$ and almost everywhere in Ω_T.

Since (θ_k) is bounded in $L^r(\Omega_T)$, for $1 \leq r < 5/3$, and according to (H.1) it yields that $\left(\sigma_k(\theta_k)^{-1}\right)$ is bounded in $L^r(\Omega_T)$, for $1 \leq r < 5/(3\alpha)$. Thus $(\nabla\varphi_k)$ is bounded in $L^r(\Omega_T)$, for $1 \leq r < 10/(5+3\alpha)$, and, up to a subsequence, $\nabla\varphi_k \rightharpoonup \nabla\varphi$ in $L^r(\Omega_T)$, $\Phi = \sigma(\theta)^{1/2}\nabla\varphi$ in $L^2(\Omega_T)$.

As to $(\boldsymbol{A}_k)$, we deduce the existence of an element $\boldsymbol{A} \in L^2(\mathbb{X})$ such that, up to a subsequence, $\boldsymbol{A}_k \rightharpoonup \boldsymbol{A}$ weakly in $L^2(\boldsymbol{L}^2(\Omega))$, $\boldsymbol{A}_k \rightharpoonup \boldsymbol{A}$ weakly in $L^2(\mathbb{X})$, and thus $\sigma_k(\theta_k)^{1/2}\boldsymbol{A}_k \rightharpoonup \sigma(\theta)^{1/2}\boldsymbol{A}$ weakly in $L^2(\boldsymbol{L}^2(\Omega))$. Finally, by making $k \to \infty$ in (14) and (15) we obtain (8) and (9).

All the properties deduced up till now are not enough in order to assure the strong convergence of (F_k) in $L^1(\Omega_T)$. Nevertheless, there exists a Radon measure $\boldsymbol{\mu} \in \mathscr{M}(\Omega_T)$ such that $F_k \rightharpoonup \frac{1}{2}\sigma(\theta)|i\omega\boldsymbol{A} + \nabla\varphi|^2 + G + \boldsymbol{\mu}$ in $\mathscr{M}(\Omega_T)$-weak–$*$. We can pass to the limit in (19) to obtain (12).

Remark 4 Our future work consists in establishing under what conditions on σ can we assure that $\boldsymbol{\mu} = 0$ or, in other words, how can one derive the strong convergence $\sigma_k(\theta_k)^{1/2}\nabla\varphi_k \to \sigma(\theta)^{1/2}\nabla\varphi$ in $L^2(\Omega_T)$.

Remark 5 The analysis of the uniqueness of a solution to (6)–(12) is a very complex task even if we already know that $\boldsymbol{\mu} = 0$. This is related to the low regularity of the unknowns obtained in our existence result. Indeed, a system like (1)–(5) is a generalization of the so-called thermistor problem [1, 5, 9] which involves only two unknowns, namely, the electric potential and the temperature.

Acknowledgements This research was partially supported by Ministerio de Economía y Competitividad of the Spanish government under grant MTM2010-16401 with the participation of FEDER, and Consejería de Educación y Ciencia of Junta de Andalucía, research group FQM-315.

The authors wish to thank the referee for his comments and suggestions have led to improve this presentation.

References

1. Antontsev, S.N., Chipot, M.: The thermistor problem: existence, smoothness, uniqueness, blowup. SIAM J. Math. Anal. **25**(4), 1128–1156 (1994)
2. Bermúdez, A., Bullón, J., Pena, F., Salgado, P.: A numerical method for transient simulation of metallurgical compound electrodes. Finite Elem. Anal. Des. **39**, 283–299 (2003)
3. Bermúdez, A., Gómez, D., Muñiz, M.C., Salgado, P.: Transient numerical simulation of a thermoelectrical problem in cylindrical induction heating furnaces. Adv. Comput. Math. **26**, 39–62 (2007)
4. Boccardo, L., Gallouët, T.: Non-linear elliptic and parabolic equations involving measure data. J. Funct. Anal. **87**, 149–169 (1989)
5. Cimatti, G.: Remark on existence and uniqueness for the thermistor problem under mixed boundary conditions. Quatr. Appl. Math. **47**, 117–121 (1989)
6. Clain, S.: Analyse mathématique et numérique d'un modèle de chauffage par induction. Ph.D. thesis, EPFL, Laussan (1994)
7. Díaz Moreno, J.M., García Vázquez, C., González Montesinos, M.T., Ortegón Gallego, F.: Analysis and numerical simulation of an induction–conduction model arising in steel heat treating. J. Comput. Appl. Math. **236**, 3007–3015 (2012)
8. Díaz Moreno, J.M., García Vázquez, C., González Montesinos, M.T., Ortegón Gallego, F., Viglialoro, G.: Mathematical modeling of heat treatment for a steering rack including mechanical effects. J. Numer. Math. **20**(3–4), 215–231 (2012). doi:10.1515/jnum-2012-0011
9. González Montesinos, M.T., Ortegón Gallego, F.: Renormalized solutions to a nonlinear parabolic-elliptic system. SIAM J. Math. Anal. **36**(6), 1991–2003 (2005)
10. González Montesinos, M.T., Ortegón Gallego, F.: Analysis of a nonuniformly elliptic and nonlinear coupled parabolic-elliptic system arising in steel hardening. Int. J. Comput. Math. **90**(10), 2079–2091 (2013). doi:10.1080/00207160.2013.771837

11. González Montesinos, M.T., Ortegón Gallego, F.: On an induction–conduction PDE system in harmonic regime. Nonlinear Anal. RWA. **15**, 58–66 (2014)
12. González Montesinos, M.T., Ortegón Gallego, F.: On a parabolic nonlinear coupled system in the harmonic regime. (To appear)
13. Hömberg, D.: A mathematical model for induction hardening including mechanical effects. Nonlinear Anal. RWA. **5**, 55–90 (2004)
14. Simon, J.: Compact sets in $L^p(0, T; B)$. Ann. Mat. Pur. Appl. sér. IV, **146**, 65–96 (1987)

Perturbation of Analytic Semigroups in Uniform Spaces in $\mathbb{R}^N$

Carlos Quesada and Aníbal Rodríguez-Bernal

Abstract We solve some linear parabolic equations obtained from perturbations of parabolic equations given by operators defining analytic semigroups. We consider several classes of initial data, in particular in low regularity spaces taken from the uniform Bessel-Lebesgue scale of spaces. We make special focus on smoothing estimates of the solution. Robustness and convergence with respect to the perturbation are also obtained.

1 Introduction

In this paper we address the solvability of some second and fourth order linear parabolic equations in $\mathbb{R}^N$. In particular, we study the problems

$$\begin{cases} u_t - \Delta u + \sum_{j=1}^N b_j(x)\partial_j u + c(x)u = 0 & x \in \mathbb{R}^N, \quad t > 0 \\ u(0,x) = u_0(x) & x \in \mathbb{R}^N, \end{cases} \tag{1}$$

Partially supported by Project MTM2012-31298, MICINN and GR58/08 Grupo 920894, UCM, Spain.

C. Quesada (✉)
Universidad Complutense de Madrid, Madrid, Spain

Departamento de Matemática Aplicada, Facultad de Ciencias Matemáticas, Plaza de Ciencias 3, Ciudad Universitaria, 28040 Madrid, Spain
e-mail: carlosqu@ucm.es

A. Rodríguez-Bernal
Departamento de Matemática Aplicada, Universidad Complutense de Madrid, 28040 Madrid, Spain

Instituto de Ciencias Matemáticas, CSIC-UAM-UC3M-UCM, 28049 Madrid, Spain
e-mail: arober@ucm.es

F. Casas, V. Martínez (eds.), *Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 4, DOI 10.1007/978-3-319-06953-1_5

and

$$\begin{cases} u_t + \Delta^2 u + d(x) D^a u = 0, \ x \in \mathbb{R}^N, \ t > 0 \\ u(0) = u_0 \qquad\qquad\qquad\qquad \text{in } \mathbb{R}^N \end{cases} \tag{2}$$

where in both cases, the lower order coefficients are assumed to have some local integrability properties and no asymptotic decay as $|x| \to \infty$ whatsoever. More precisely they are assumed to belong to some locally uniform Lebesgue spaces. To be more precise, let $L^p_U(\mathbb{R}^N)$ denote the locally uniform space composed of the functions $f \in L^p_{loc}(\mathbb{R}^N)$ such that there exists $C > 0$ such that for all $x_0 \in \mathbb{R}^N$

$$\int_{B(x_0,1)} |f|^p \le C \tag{3}$$

endowed with the norm

$$\|f\|_{L^p_U(\mathbb{R}^N)} = \sup_{x_0 \in \mathbb{R}^N} \|f\|_{L^p(B(x_0,1))}$$

(for $p = \infty$, $L^\infty_U(\mathbb{R}^N) = L^\infty(\mathbb{R}^N)$).

The initial data in (1) and (2) will be assumed to belong to some uniform Bessel space $\dot{H}^{2\gamma,q}_U(\mathbb{R}^N)$ and $\dot{H}^{4\gamma,q}_U(\mathbb{R}^N)$ respectively; see Sect. 2 for further details. Our goal in this paper is to show that both problems are well posed, with a unique solution given by the analytic semigroup defined by the equation.

Furthermore, we study the smoothing properties of the solution, and the continuous dependence with respect to changes in the coefficients.

Concerning (1) a similar result, without the continuity with respect of perturbations in the coefficients, was proved in [3] and later recovered in Theorem 5.3 in [4], assuming additionally that

$$p_j \ge q > 1, \quad \text{for } j = 0, \dots, N.$$

That result was later recovered in [8] with different techniques. The result in [4, 8] just allowed for $\gamma \ge 0$ in (15). Here, we remove such restrictions allowing in particular a larger class of initial data, since in (15), γ can be even negative. Also, with the additional assumptions above, Theorem 1 recovers Theorem 5.3 in [4].

The paper is organized as follows. In Sect. 2 we briefly recall the main properties of uniform Lebesgue and Bessel spaces that will be needed hereafter. Then in Sect. 3 we study the semigroup generated by Δ^2 in the uniform spaces. Finally, Sects. 4 and 5 are devoted to the study of Eqs. (1) and (2).

2 Uniform Spaces

We will study parabolic equations in large spaces which contain the Sobolev-Bessel scale of spaces, namely the locally uniform spaces.

For this, consider the locally uniform space $L^q_U(\mathbb{R}^N)$, composed of the functions $f \in L^p_{loc}(\mathbb{R}^N)$ such that there exists $C > 0$ such that for all $x_0 \in \mathbb{R}^N$

$$\int_{B(x_0,1)} |f|^p \leq C \tag{4}$$

endowed with the norm

$$\|f\|_{L^p_U(\mathbb{R}^N)} = \sup_{x_0 \in \mathbb{R}^N} \|f\|_{L^p(B(x_0,1))}$$

(for $p = \infty$, $L^\infty_U(\mathbb{R}^N) = L^\infty(\mathbb{R}^N)$).

Now, for $1 \leq q \leq \infty$ defined as in (4) denote by $\dot{L}^q_U(\mathbb{R}^N)$ the closed subspace of $L^q_U(\mathbb{R}^N)$ consisting of all elements which are translation continuous with respect to $\|\cdot\|_{L^q_U(\mathbb{R}^N)}$, that is

$$\|\tau_y \phi - \phi\|_{L^q_U(\mathbb{R}^N)} \to 0 \text{ as } |y| \to 0,$$

where $\{\tau_y, y \in \mathbb{R}^N\}$ denotes the group of translations. Note that $L^q(\mathbb{R}^N) \subset \dot{L}^q_U(\mathbb{R}^N)$ for $1 \leq q < \infty$ and for $q = \infty$ we get $L^\infty_U(\mathbb{R}^N) = L^\infty(\mathbb{R}^N)$ and $\dot{L}^\infty_U(\mathbb{R}^N) = BUC(\mathbb{R}^N)$.

Thus we introduce the *uniform Bessel-Sobolev spaces* $H^{k,q}_U(\mathbb{R}^N)$, with $k \in \mathbb{N}$, as the set of functions $\phi \in H^{k,q}_{loc}(\mathbb{R}^N)$ such that

$$\|\phi\|_{H^{k,q}_U(\mathbb{R}^N)} = \sup_{x \in \mathbb{R}^N} \|\phi\|_{H^{k,q}(B(x,1))} < \infty$$

for $k \in \mathbb{N}$. Then denote by $\dot{H}^{k,q}_U(\mathbb{R}^N)$ a subspace of $H^{k,q}_U(\mathbb{R}^N)$ consisting of all elements which are translation continuous with respect to $\|\cdot\|_{H^{k,q}_U(\mathbb{R}^N)}$, that is

$$\|\tau_y \phi - \phi\|_{H^{k,q}_U(\mathbb{R}^N)} \to 0 \text{ as } |y| \to 0$$

where $\{\tau_y, y \in \mathbb{R}^N\}$ denotes the group of translations.

Consider the complex interpolation functor denoted by $[\cdot,\cdot]_\theta$, for $\theta \in (0,1)$, see [9] for details. Then for $1 \leq q < \infty$, $k \in \mathbb{N} \cup \{0\}$ and $s \in (k, k+1)$ we define $\theta \in (0,1)$ such that $s = \theta(1+k) + (1-\theta)k$, that is $\theta = s - k$. Then one can define the intermediate spaces by interpolation as

$$H^{s,q}_U(\mathbb{R}^N) = [H^{k+1,q}_U(\mathbb{R}^N), H^{k,q}_U(\mathbb{R}^N)]_\theta,$$

and

$$\dot{H}_U^{s,q}(\mathbb{R}^N) = [\dot{H}_U^{k+1,q}(\mathbb{R}^N), \dot{H}_U^{k,q}(\mathbb{R}^N)]_\theta.$$

For details on the construction of the interpolation scale, see [2].

Using Proposition 4.2 in [4] it is easy to see that the sharp embeddings of Bessel spaces translate into

$$\dot{H}_U^{s,q}(\mathbb{R}^N) \subset \begin{cases} \dot{L}_U^r(\mathbb{R}^N),\ s - \frac{N}{q} \geq -\frac{N}{r},\ 1 \leq r < \infty & if\ s - \frac{N}{q} < 0 \\ \dot{L}_U^r(\mathbb{R}^N),\ 1 \leq r < \infty & if\ s - \frac{N}{q} = 0 \\ C_b^\eta(\mathbb{R}^N) & if\ s - \frac{N}{q} > \eta \geq 0. \end{cases} \tag{5}$$

The uniform Bessel spaces can be extended to negative indexes by a general extrapolation procedure as in [2]. In this way one can define the extrapolated space $\dot{H}_U^{-k}(\mathbb{R}^N)$ as the completion of $\dot{L}_U^q(\mathbb{R}^N)$ with the norm $\|(-\Delta + I)^{-k/2}u\|_{\dot{L}_U^q(\mathbb{R}^N)}$. Using complex interpolation, for $0 < s < k$, $k \in \mathbb{N}$, the intermediate spaces are given by

$$\dot{H}_U^{-s,q}(\mathbb{R}^N) = [\dot{L}_U^q(\mathbb{R}^N), \dot{H}_U^{-k,q}(\mathbb{R}^N)]_\theta, \qquad with\ \theta = \frac{s}{k}.$$

For the negative side of the scale, the following embedding holds

$$\dot{L}_U^p(\mathbb{R}^N) \hookrightarrow \dot{H}_U^{-s,q}(\mathbb{R}^N) \qquad \text{if} \quad s - \frac{N}{q'} \geq -\frac{N}{p'}, \qquad s > 0, \tag{6}$$

see Proposition 3.1 in [7].

The heat equation has been studied in these spaces. In [4], the Laplace operator was considered in the scale of spaces $H_U^{s,q}(\mathbb{R}^N)$, $s \geq 0$ and $\dot{H}_U^{s,q}(\mathbb{R}^N)$, and it was proved that $-\Delta$ defines an analytic semigroup $S_{-\Delta}(t)$. However in the "undotted" spaces the semigroup generated by $-\Delta$ is analytic but not strongly continuous. These spaces are less convenient to use because smooth functions are not dense in them; see [4].

3 Elliptic Estimates in Uniform Spaces

We now want to study the semigroup defined by $\Delta^2 := (-\Delta)(-\Delta)$ in those spaces. For it, we need a deeper knowledge of $-\Delta$, thus we start with the following Proposition.

Proposition 1 *(i) For $1 < q < \infty$, in the space $\dot{L}^q_U(\mathbb{R}^N)$ the operator $-\Delta$ with domain $D(-\Delta) = \dot{H}^{2,q}_U(\mathbb{R}^N)$, satisfies the estimate*

$$\|(-\Delta - \lambda)^{-1}\|_{\mathscr{L}(\dot{L}^q_U(\mathbb{R}^N))} \leq M|\lambda|^{-1}$$

for all λ in a sector $S_{0,\phi}$ for $\phi > 0$ arbitrarily small where

$$S_{a,\phi} = \{z \in \mathbb{C} : \phi \leq |arg(z-a)| \leq \pi,\ z \neq a\}. \tag{7}$$

Furthermore, $\sigma(-\Delta) = [0, \infty)$.

(ii) For $1 < q < \infty$, in the space $\dot{L}^q_U(\mathbb{R}^N)$ the operator Δ^2 with domain $D(\Delta^2) = \dot{H}^{4,q}_U(\mathbb{R}^N)$, satisfies the estimate

$$\|(\Delta^2 - \lambda)^{-1}\|_{\mathscr{L}(\dot{L}^q_U(\mathbb{R}^N))} \leq M|\lambda|^{-1}$$

for all λ in a sector $S_{0,2\phi}$ for $\phi > 0$ arbitrarily small.
Furthermore, $\sigma(\Delta^2) = [0, \infty)$.

Sketch of the proof First recall from Theorem 2.1 in [4] that $D(-\Delta) = \dot{H}^{2,q}_U(\mathbb{R}^N)$. To prove part (i), observe that, as in page 32–33 in [5], we can obtain an expression for the operator $(-\Delta + \mu I)^{-1}$, provided $Re(\sqrt{\mu}) > 0$, as a convolution operator.

The expression is

$$u = (-\Delta + \mu)^{-1} f = \Gamma_\mu * f, \qquad Re(\sqrt{\mu}) > 0$$

with

$$\Gamma_\mu(x) = \sqrt{\mu}^{N-2} G_2(\sqrt{\mu}x), \qquad x \in \mathbb{R}^N, \qquad Re(\sqrt{\mu}) > 0$$

where G_2 is the Green's function for $(-\Delta + I)$.

Now observe that if $\lambda \in S_{0,\phi}$ with $\phi > 0$ then for $\mu = -\lambda \in \mathbb{C} \setminus (-\infty, 0]$ we can choose $Re(\sqrt{\mu}) > 0$. For such λ we are going to check that for $f \in \dot{L}^q_U(\mathbb{R}^N)$ we have the following estimate for $u = \Gamma_\mu * f$,

$$\|u\|_{L^q_U(\mathbb{R}^N)} \leq C \frac{1}{|\lambda|} \|f\|_{L^q_U(\mathbb{R}^N)}, \qquad \lambda \in S_{0,\phi} \qquad \phi > 0.$$

Let $\{Q_i\}$, $i \in \mathbb{Z}^N$, be a partition of $\mathbb{R}^N$ in open disjoint cubes centered in $i \in \mathbb{Z}^N$ with edges of length 1, parallel to the axes. Thus $Q_i \cap Q_j = \emptyset$ for $i \neq j$ and $\mathbb{R}^N = \cup_i \overline{Q_i}$.

Then we fix $i \in \mathbb{Z}^N$ and decompose $f \in \dot{L}^q_U(\mathbb{R}^N)$ in a *far* and a *near* region as in Proposition 2.1 in [4]. For this we denote by $N(i)$ the set for indices j such that $\overline{Q_i} \cap \overline{Q_j} \neq 0$. That is, the set for which

$$d_{ij} := \inf\{dist(x, y), x \in Q_i, y \in Q_j\}$$

satisfies that $d_{ij} = 0$. Thus we can define, for each $i \in \mathbb{Z}^N$ fixed

$$Q_i^{near} = \cup_{j \in N(i)} Q_j \qquad and \qquad Q_i^{far} = \mathbb{R}^N \setminus Q_i^{near}.$$

Hence, we decompose $f := f_i^{near} + f_i^{far} := f\chi_{Q_i^{near}} + f\chi_{Q_i^{far}}$, where χ denotes the characteristic function and $u := u_i^{near} + u_i^{far}$ with

$$u_i^{near} := \Gamma_\mu * f_i^{near} \qquad u_i^{far} := \Gamma_\mu * f_i^{far}.$$

The resolvent estimate will follow from the following estimates of the two terms of the decomposition. For λ as above, we have first,

$$\|u_i^{near}\|_{L^q(Q_i)} \leq \|u_i^{near}\|_{L^q(\mathbb{R}^N)} \leq \frac{C}{|\lambda|}\|f_i^{near}\|_{L^q(\mathbb{R}^N)} = \frac{C(N)}{|\lambda|}\|f\|_{L^q(Q_i^{near})} \tag{8}$$

for all $\lambda \in S_{0,\phi}$ where we have used the resolvent estimate for $-\Delta$ in $L^q(\mathbb{R}^N)$.

It can also be proved, see [7]

$$\|u_i^{far}\|_{L^\infty(Q_i)} \leq \frac{C}{|\lambda|}\|f\|_{L^1_U(Q_i^{far})}, \qquad \lambda \in S_{0,\phi} \tag{9}$$

for some C independent if $i \in \mathbb{Z}^N$.

Using (8) and (9), since the constants for the embedding $L^\infty(Q_i) \hookrightarrow L^q(Q_i)$ and the restrictions $L^q_U(\mathbb{R}^N) \hookrightarrow L^q(Q_i^{near})$, $L^q_U(\mathbb{R}^N) \hookrightarrow L^1_U(Q_i^{near})$ depend on N but can be chosen independent of p, q and i, (8) and (9) imply

$$\|u\|_{L^q(Q_i)} \leq \frac{C}{|\lambda|}\|f\|_{L^q_U(\mathbb{R}^N)}, \qquad \lambda \in S_{0,\phi}$$

for each $i \in \mathbb{Z}^N$ with C independent of i and $\lambda \in S_{0,\phi}$, which gives the result.

For part (ii), we use (i) and [6, 10.5] and we get that Δ^2 is sectorial with sector $S_{0,2\phi}$. Note that $\sigma(\Delta^2) \subset [0, \infty)$ because $\phi > 0$ is arbitrarily small.

Also, note that $u(x) = e^{i\omega x}$, $\omega \in \mathbb{R}^N$ satisfies $u \in \dot{L}^q_U(\mathbb{R}^N)$ and

$$-\Delta u = \lambda u \qquad \Delta^2 u = \lambda^2 u$$

for $\lambda = |\omega|^2 \subset [0, \infty)$.

■

4 Parabolic Equations in Uniform Spaces

It is known from [4] that $-\Delta$ defines an analytic semigroup $S_{-\Delta}(t)$ which gives the solution $u(t) = S_{-\Delta}(t)u_0$ for the parabolic Laplacian problem

$$\begin{cases} u_t - \Delta u = 0, \ x \in \mathbb{R}^N, t > 0 \\ u(0) = u_0, \quad \text{in } \mathbb{R}^N \end{cases} \tag{10}$$

in the uniform Bessel spaces $\{\dot{H}^{2\alpha,q}_U(\mathbb{R}^N)\}_{\alpha\in\mathbb{R}}$ that satisfies the smoothing estimates

$$\|S_{-\Delta}(t)u_0\|_{\dot{H}^{2\alpha,q}_U(\mathbb{R}^N)} \leq \frac{M_{\alpha,\beta}e^{\mu_0 t}}{t^{\alpha-\beta}}\|u_0\|_{\dot{H}^{2\beta,q}_U(\mathbb{R}^N)}, \qquad t > 0, \quad u_0 \in \dot{H}^{2\beta,q}_U(\mathbb{R}^N)$$

for $1 < q < \infty$, $\alpha, \beta \in \mathbb{R}$, $\alpha \geq \beta$. In the Lebesgue spaces, $\dot{L}^q_U(\mathbb{R}^N)$, $1 < q < \infty$, the semigroup satisfies

$$\|S_{-\Delta}(t)u_0\|_{\dot{L}^r_U(\mathbb{R}^N)} \leq \frac{M_{q,r}e^{\mu_0 t}}{t^{\frac{N}{2}(\frac{1}{q}-\frac{1}{r})}}\|u_0\|_{\dot{L}^q_U(\mathbb{R}^N)}, \qquad t > 0, \quad u_0 \in \dot{L}^q_U(\mathbb{R}^N)$$

for any $\mu_0 > 0$ and $1 < q \leq r \leq \infty$ and some $M_{q,r} > 0$. Notice that this also follows from the estimate in Proposition 1, (i).

Furthermore, the semigroup is order preserving. Recall from [1] that if $u_0 \geq 0$ then $S_0(t)u_0 \geq 0$ for all $t \geq 0$. Now, for $u_0 \in \dot{H}^{\beta,q}_U(\mathbb{R}^N)$ take $\{u^n_0\}_{n\in\mathbb{N}}$ regular such that $u^n_0 \to u_0$ then $S_0(t)u^n_0 \to S_0(t)u_0$ and since $S_0(t)u^n_0 \geq 0$ for all $n \in \mathbb{N}$ then $S_0(t)u_0 \geq 0$. Note that this can be done because we are using the "dotted" spaces, where regular functions are dense.

Using the resolvent estimates in Proposition 1 and semigroup theory as in [5] we get the following result for the parabolic bi-Laplacian equation.

Lemma 1 *Consider the problem*

$$\begin{cases} u_t + \Delta^2 u = 0 \ x \in \mathbb{R}^N, \ t > 0 \\ u(0) = u_0 \qquad in \ \mathbb{R}^N. \end{cases} \tag{11}$$

(i) Then for each $1 < q < \infty$, *(11) defines an analytic semigroup,* $S_{\Delta^2}(t)$, *in* $\dot{H}_U^{4\beta,q}(\mathbb{R}^N)$, $\beta \in \mathbb{R}$, *such that for any* $\mu_0 > 0$ *there exists C such that*

$$\|S_{\Delta^2}(t)u_0\|_{\dot{H}_U^{4\alpha,q}(\mathbb{R}^N)} \leq \frac{M_{\alpha,\beta}e^{\mu t}}{t^{\alpha-\beta}}\|u_0\|_{\dot{H}_U^{4\beta,q}(\mathbb{R}^N)}, \qquad t > 0, \quad u_0 \in \dot{H}_U^{4\beta,q}(\mathbb{R}^N)$$

with $\alpha, \beta \in \mathbb{R}$, $\alpha \geq \beta$.

(ii) The analytic semigroup $S_{\Delta^2}(t)$, *in* $\dot{L}_U^q(\mathbb{R}^N)$, $1 < q < \infty$, *satisfies*

$$\|S_{\Delta^2}(t)u_0\|_{\dot{L}_U^r(\mathbb{R}^N)} \leq \frac{M_{q,r}e^{\mu_0 t}}{t^{\frac{N}{4}(\frac{1}{q}-\frac{1}{r})}}\|u_0\|_{\dot{L}_U^q(\mathbb{R}^N)}, \qquad t > 0, \quad u_0 \in \dot{L}_U^q(\mathbb{R}^N)$$

for any $\mu_0 > 0$ *and* $1 < q \leq r \leq \infty$ *and some* $M_{q,r} > 0$.

5 Perturbed Equations with Low Regularity Initial Data

We now study the parabolic equations with main operator $-\Delta$ and Δ^2, adding perturbations $P_a \in \mathscr{L}(\dot{H}_U^{s,q}(\mathbb{R}^N), \dot{H}_U^{-\sigma,q}(\mathbb{R}^N))$ to be specified below, where $s, \sigma \geq 0$ and $s + \sigma < m$ where m is the order of the main operator.

For such function consider the problem

$$u(t; u_0) = S(t)u_0 + \int_0^t S(t-\tau)P(\tau)\,d\tau, \qquad t > 0, \tag{12}$$

with u_0 to be chosen below, and where $S(\cdot)$ can be $S_{-\Delta}(\cdot)$ or $S_{\Delta^2}(\cdot)$.

In this situation, estimating (12), it can be proved that for $u_0 \in \dot{H}_U^{m\gamma,q}(\mathbb{R}^N)$ with $\gamma \in (\frac{s}{m}-1, \frac{s}{m}]$, the unique map $u(t; u_0)$ satisfying (12) defines an analytic semigroup $S_P(t)u_0 := u(t, u_0)$ which furthermore satisfies

$$\|S_P(t)u_0\|_{\dot{H}_U^{m\gamma',q}(\mathbb{R}^N)} \leq Ct^{-(\gamma-\gamma')}\|u_0\|_{\dot{H}_U^{m\gamma,q}(\mathbb{R}^N)}$$

where $\gamma' \geq \gamma$ and $\gamma' \in [-\frac{\sigma}{m}, 1 - \frac{\sigma}{m})$. See [8] for details.

We now define in particular $P_a u = d(x)D^a u$ with $d \in \dot{L}_U^p(\mathbb{R}^N)$ where D^a denotes any derivative of order a, which satisfies the above assumptions for some s and σ. We now calculate the set of admissible pairs (s, σ)

Proposition 2 *Let* $P_a u = d(x)D^a u$ *with* $d \in \dot{L}_U^p(\mathbb{R}^N)$, $a \in \{0, 1, 2, 3\}$. *Let* $s \geq a$, $\sigma \geq 0$. *Then for* $1 < q < \infty$, *if*

$$(s - a - \frac{N}{q})_- + (\sigma - \frac{N}{q'})_- > -\frac{N}{p'} \tag{13}$$

we have

$$P_a \in \mathscr{L}(\dot{H}_U^{s,q}(\mathbb{R}^N), \dot{H}_U^{-\sigma,q}(\mathbb{R}^N)), \qquad \|P_a\|_{\mathscr{L}(\dot{H}_U^{s,q}(\mathbb{R}^N),\dot{H}_U^{-\sigma,q}(\mathbb{R}^N))} \le C\|d\|_{\dot{L}_U^p(\mathbb{R}^N)}.$$

Proof First note that $u \in \dot{H}_U^{s,q}(\mathbb{R}^N)$, thus $D^a u \in \dot{H}_U^{s-a,q}(\mathbb{R}^N)$. Because of (13) we can choose $r, \rho \ge 1$ such that $(s - a - \frac{N}{q})_- > -\frac{N}{r}$ and $(\sigma - \frac{N}{q'})_- > -\frac{N}{\rho'}$ with $\frac{1}{\rho} = \frac{1}{r} + \frac{1}{p}$ (and so $r \ge p'$).

Therefore we can use the inclusion $\dot{H}_U^{s-a,q}(\mathbb{R}^N) \hookrightarrow \dot{L}_U^r(\mathbb{R}^N)$ and then $P_a u \in \dot{L}_U^\rho(\mathbb{R}^N)$. We now use the inclusion $\dot{L}_U^\rho(\mathbb{R}^N) \hookrightarrow \dot{H}_U^{-\sigma,q}(\mathbb{R}^N)$ from (6) to get the result. □

Using the above proposition we can regard such lower order derivatives with space dependence as a perturbations satisfying the framework in [8], so we can use the perturbation techniques there and in [7]. In the case when the main operator is $-\Delta$ we get

Theorem 1 *Assume for* $j = 1, \dots, N$,

$$\|b_j\|_{\dot{L}_U^{p_j}(\mathbb{R}^N)} \le R_j \quad \text{and} \quad \|c\|_{\dot{L}_U^{p_0}(\mathbb{R}^N)} \le R_0$$

where $p_j > N$ *and* $p_0 > \frac{N}{2}$. *Define* $a_0 = 0$, $a_j = 1$ *and for* $j = 1, \dots, N$ *and,* $\tilde{p} = \min\{p_j,\ j = 1, \dots, N\} > N$. *If* $q' < \tilde{p}$ *and* $q > p_0$, *we will also assume* $p_0 > \frac{Nq}{N+q}$.

Then for any $1 < q < \infty$ *there exists non-empty interval* $I(q) \subset (-\frac{1}{2}, 1)$ *containing* $(-1 + \max_j\{\frac{a_j}{2} + \frac{N}{2p_j}\}, 1 - \max_j\{\frac{N}{2p_j}\})$, *such that for any* $\gamma \in I(q)$, *we have a strongly continuous, order preserving, analytic semigroup* $S(t)$ *in the space* $\dot{H}_U^{2\gamma,q}(\mathbb{R}^N)$, *for the problem*

$$\begin{cases} u_t - \Delta u + \sum_{j=1}^N b_j(x)\partial_j u + c(x)u = 0 & x \in \mathbb{R}^N, \quad t > 0 \\ u(0,x) = u_0(x) & x \in \mathbb{R}^N \end{cases} \tag{14}$$

with $u(t; u_0) = S(t)u_0$, $t \ge 0$.

Moreover the semigroup has the smoothing estimate

$$\|S(t)u_0\|_{\dot{H}_U^{2\gamma',q}(\mathbb{R}^N)} \le \frac{M_{\gamma',\gamma} e^{\mu t}}{t^{\gamma'-\gamma}} \|u_0\|_{\dot{H}_U^{2\gamma,q}(\mathbb{R}^N)}, \quad t > 0,\ u_0 \in \dot{H}_U^{2\gamma}(\mathbb{R}^N) \tag{15}$$

for every $\gamma, \gamma' \in I(q)$ *with* $\gamma' \ge \gamma$, *and*

$$\|S(t)u_0\|_{\dot{L}_U^r(\mathbb{R}^N)} \le \frac{M_{q,r} e^{\mu t}}{t^{\frac{N}{2}(\frac{1}{q}-\frac{1}{r})}} \|u_0\|_{\dot{L}_U^q(\mathbb{R}^N)}, \quad t > 0,\ u_0 \in \dot{L}_U^q(\mathbb{R}^N) \tag{16}$$

for $1 < q \le r \le \infty$ *with some* $M_{\gamma',\gamma}$, $M_{q,r}$ *and* $\mu \in \mathbb{R}$ *depending on* R_j *and* R_0.
Furthermore,

$$I(q) = (-1 + \max_{j=0,\dots,N}\{\frac{a_j}{2} + \frac{N}{2}(\frac{1}{p_j} - \frac{1}{q})_+\}, 1 - \frac{N}{2}(\frac{1}{\min_{j=0,\dots,N}\{p_j\}} - \frac{1}{q})_+). \tag{17}$$

Finally, if, as $\epsilon \to 0$

$$b_j^\epsilon \to b_j \quad \text{in } \dot{L}_U^{p_j}(\mathbb{R}^N), \quad p_j > N, \quad j = 1, \dots, N,$$

$$c^\epsilon \to c \quad \text{in } \dot{L}_U^{p_0}(\mathbb{R}^N), \quad p_0 > N/2$$

then for every $T > 0$ *there exists* $C(\epsilon) \to 0$ *as* $\epsilon \to 0$, *such that*

$$\|S_\epsilon(t) - S(t)\|_{\mathscr{L}(\dot{H}_U^{2\gamma,q}(\mathbb{R}^N), \dot{H}_U^{2\gamma',q}(\mathbb{R}^N))} \le \frac{C(\epsilon)}{t^{\gamma'-\gamma}}, \qquad \forall\ 0 < t \le T$$

for all $\gamma, \gamma' \in I(q)$, $\gamma' \ge \gamma$ *and for all* $1 < q \le r \le \infty$,

$$\|S_\epsilon(t) - S(t)\|_{\mathscr{L}(\dot{L}_U^q(\mathbb{R}^N), \dot{L}_U^r(\mathbb{R}^N))} \le \frac{C(\epsilon)}{t^{\frac{N}{2}(\frac{1}{q}-\frac{1}{r})}}, \qquad \forall\ 0 < t \le T.$$

Finally, when the main operator is Δ^2 we get

Theorem 2 *Let* $a \in \{0, 1, 2, 3\}$, $d \in \dot{L}_U^p(\mathbb{R}^N)$ *such that* $\|d\|_{\dot{L}_U^p(\mathbb{R}^N)} \le R_0$ *with* $p > \frac{N}{4-a}$. *Then for any* $1 < q < \infty$ *and any* P_a *as in Proposition 2 there exists an interval* $I(q,a) \subset (-1 + \frac{a}{4}, 1)$ *containing* $(-1 + \frac{a}{4} + \frac{N}{4p}, 1 - \frac{N}{4p})$, *such that for any* $\gamma \in I(q,a)$, *we have a continuous, analytic semigroup,* $S(t)$ *in the space* $\dot{H}_U^{4\gamma,q}(\mathbb{R}^N)$, *for the problem*

$$\begin{cases} u_t + \Delta^2 u + d(x) D^a u = 0, \ x \in \mathbb{R}^N, \ t > 0 \\ u(0) = u_0 \qquad\qquad\qquad\qquad\quad \text{in } \mathbb{R}^N. \end{cases}$$

Moreover the semigroup has the smoothing estimate

$$\|S(t)u_0\|_{\dot{H}_U^{4\gamma',q}(\mathbb{R}^N)} \le \frac{M_{\gamma',\gamma} e^{\mu t}}{t^{\gamma'-\gamma}} \|u_0\|_{\dot{H}_U^{4\gamma,q}(\mathbb{R}^N)}, \quad t > 0, \ u_0 \in \dot{H}_U^{4\gamma}(\mathbb{R}^N)$$

for every $\gamma, \gamma' \in I(q,a)$ *with* $\gamma' \ge \gamma$, *and*

$$\|S(t)u_0\|_{\dot{L}_U^r(\mathbb{R}^N)} \le \frac{M_{q,r} e^{\mu t}}{t^{\frac{N}{4}(\frac{1}{q}-\frac{1}{r})}} \|u_0\|_{\dot{L}_U^q(\mathbb{R}^N)}, \quad t > 0, \ u_0 \in \dot{L}_U^q(\mathbb{R}^N)$$

for $1 < q \le r \le \infty$ *with some* $M_{\gamma',\gamma}$, $M_{q,r}$ *and* $\mu \in \mathbb{R}$ *depending on* d *only through* R_0.

For each P_a, *the interval* $I(q,a)$ *is given by*

$$I(q,a) = (-1 + \frac{a}{4} + \frac{N}{4}(\frac{1}{p} - \frac{1}{q'})_+, 1 - \frac{N}{4}(\frac{1}{p} - \frac{1}{q})_+) \subset (-1 + \frac{a}{4}, 1).$$

Finally, if, as $\epsilon \to 0$

$$d_\epsilon \to d \quad \text{in } \dot{L}^p_U(\mathbb{R}^N), \quad p > \frac{N}{4-k}$$

then for every $T > 0$ *there exists* $C(\epsilon) \to 0$ *as* $\epsilon \to 0$, *such that*

$$\|S_\epsilon(t) - S(t)\|_{\mathscr{L}(\dot{H}^{4\gamma,q}_U(\mathbb{R}^N), \dot{H}^{4\gamma',q}_U(\mathbb{R}^N))} \le \frac{C(\epsilon)}{t^{\gamma'-\gamma}}, \qquad \forall\ 0 < t \le T$$

for all $\gamma, \gamma' \in I(q,a,b)$, $\gamma' \ge \gamma$ *and for all* $1 < q \le r \le \infty$,

$$\|S_\epsilon(t) - S(t)\|_{\mathscr{L}(\dot{L}^q_U(\mathbb{R}^N), \dot{L}^r_U(\mathbb{R}^N))} \le \frac{C(\epsilon)}{t^{\frac{N}{4}(\frac{1}{q}-\frac{1}{r})}}, \qquad \forall\ 0 < t \le T.$$

Note that different perturbations can be combined together, although not all combinations are allowed. Consider a finite family of perturbations $P_i := P_{a_i}$ with $\|d_i\|_{L^{p_i}_U(\mathbb{R}^N)} \le R_0$, with $p_i > \frac{N}{4-a_i}$, $i = 1, \dots, J$.

Denote $P := \sum_i P_i$, then for any $1 < q < \infty$, if

$$\max_i\{a_i + (\frac{N}{p_i} - \frac{N}{q'})_+\} + \max_i\{(\frac{N}{p_i} - \frac{N}{q})_+\} < 4 \tag{18}$$

then the results in Theorem 2 hold for an interval $I(q,P) \subset (-1 + \frac{\max_i\{a_i\}}{4}, 1)$ containing $(-1 + \max_i\{\frac{a_i}{4} + \frac{N}{4p_i}\}, 1 - \max_i\{\frac{N}{4p_i}\})$, instead of $I(q,a)$.

In particular, if $p_i = p$ for all i, all possible perturbations can be combined, since all P_i satisfy (18) and so does P as well.

References

1. Amann, H.: Nonhomogeneous Linear and Quasilinear Elliptic and Parabolic Boundary Value Problems. Teubner, Stuttgart (1993)
2. Amann, H.: Linear and Quasilinear Parabolic Problems, vol. I. Birkhäuser, Boston (1995)
3. Amann, H., Hieber, M., Simonett, G.: Bounded H_∞-calculus for elliptic operators. Differ. Integral Equ. **7**(3–4), 613–653 (1994)
4. Arrieta, J.M., Cholewa, J W., Dlotko, T., Rodríguez-Bernal, A.: Linear parabolic equations in locally uniform spaces. Math. Models Methods Appl. Sci. **14**(2), 253–293 (2004)

5. Henry, D.: Geometric Theory of Semilinear Parabolic Equations. Springer, Berlin/New York (1981)
6. Komatsu, H.: Fractional powers of operators. Pac. J. Math. **19**(2), 285–346 (1966)
7. Quesada, C., Rodríguez-Bernal, A.: Smoothing and perturbation for some fourth order linear parabolic equations in $\mathbb{R}^n$. J. Math. Anal. Appl. **412**(2), 1105–1134 (2014)
8. Rodríguez-Bernal, A.: Perturbation of analytic semigroups in scales of Banach spaces and applications to parabolic equations with low regularity data. SEMA J. **53**, 3–54 (2011)
9. Triebel, H.: Interpolation Theory, Function Spaces, Differential Operators. Ambrosius Barth, Heidelberg (1995)

Nonlinear Nonlocal Reaction-Diffusion Equations

Aníbal Rodríguez-Bernal and Silvia Sastre-Gómez

Abstract Let $\Omega \subset \mathbb{R}^N$, and J be a nonnegative function defined in $\Omega \times \Omega$. We consider the problem

$$\begin{cases} u_t(x,t) = \displaystyle\int_\Omega J(x,y)u(y,t)dy - h(x)u(x,t) + f(x,u(x,t)),\ x \in \Omega,\ t > 0 \\ u(x,0) = u_0(x), \quad x \in \Omega, \end{cases} \tag{1}$$

with $h \in L^\infty(\Omega)$, $u_0 \in L^p(\Omega)$ and the function f defined as $f : \Omega \times \mathbb{R} \to \mathbb{R}$, that maps (x,s) into $f(x,s)$. We assume f globally Lipschitz or f locally Lipschitz in the variable $s \in \mathbb{R}$, uniformly with respect to $x \in \Omega$, and f satisfies that there exist $C \in \mathbb{R}$ and $D \geq 0$ such that

$$f(\cdot,s)s \leq Cs^2 + D|s|,\ \forall s \in \mathbb{R}.$$

The aim is to study the existence and uniqueness and we give some asymptotic estimates of the norm $L^\infty(\Omega)$ of the solution u of the problem (1), following the ideas of [2], and we prove the existence of two ordered extremal equilibria, like in [6], which give some information about the set that attracts the dynamics of the solution of (1), for all u_0 in $L^\infty(\Omega)$.

A. Rodríguez-Bernal (✉)
Departamento de Matemática Aplicada, Universidad Complutense de Madrid, 28040 Madrid, Spain

Instituto de Ciencias Matemáticas, CSIC-UAM-UC3M-UCM, Madrid, Spain
e-mail: arober@mat.ucm.es

S. Sastre-Gómez
Departamento de Matemática Aplicada, Universidad Complutense de Madrid, 28040 Madrid, Spain
e-mail: silviasastre@mat.ucm.es

F. Casas, V. Martínez (eds.), *Advances in Differential Equations and Applications*,
SEMA SIMAI Springer Series 4, DOI 10.1007/978-3-319-06953-1_6

1 Introduction

Let $\Omega \subset \mathbb{R}^N$ be an open, bounded set. The problem we are going to work with is the following

$$\begin{cases} u_t(x,t) = (K-hI)(u)(x,t) + f(x,u(x,t)) = L(u)(x,t) + f(x,u(x,t)), x \in \Omega, t > 0 \\ u(x,0) = u_0(x), \ x \in \Omega, \end{cases} \tag{2}$$

with $u_0 \in L^p(\Omega)$, $h \in L^\infty(\Omega)$, and $K(u)(x,t) = \int_\Omega J(x,y)u(y,t)dy$, where $J(x,y)$ is a nonnegative function defined as $J : \Omega \times \overline{\Omega} \to \mathbb{R}$. We assume that J satisfies that $J \in C(\overline{\Omega} \times \overline{\Omega})$, $J(x,y) = J(y,x)$, and

$$J(x,y) > 0, \ \forall x, \ y \in \Omega \ : dist(x,y) < R, \text{ with } 0 < R \in \mathbb{R}.$$

Under these conditions, we have that, for $1 \leq p \leq \infty$, the linear operator

$$L = K - hI \in \mathscr{L}(L^p(\Omega), L^p(\Omega)).$$

The function f is defined as $f : \Omega \times \mathbb{R} \to \mathbb{R}$, that maps (x,s) into $f(x,s)$.

The equations like (2) have been mainly used to model diffusion processes, as we can see in [1, 3–5]. In particular if $u(x,t)$ is thought of as the density of a single population at the point x per unit time t, and $J(x,y)$ is thought of as the density of probability of jumping from location y to location x, then the operator $K(u)(x,t)$ is the rate at which individuals are arriving to position x from all other places, and f is the rate of local reaction.

We will start giving a result of existence and uniqueness of the solution associated to (2), with f globally Lipschitz. We will also give some comparison results for the solution with the f globally Lipschitz. After that, we will be able to prove the existence and uniqueness of the solution associated to (2), with f locally Lipschitz in the variable $s \in \mathbb{R}$, uniformly with respect to $x \in \Omega$, plus other extra hypothesis.

Finally, we state some asymptotic estimates of the solution, and we will finish proving under some hypotheses on f, the existence of two extrema equilibria φ_m and φ_M. This means that all the solutions enter between the two extremal equilibria when time goes to infinity.

2 Existence, Uniqueness, Positiveness and Comparison Results for Lipschitz Nonlinearities

For $1 \leq p \leq \infty$, we have that $L = K - hI \in \mathscr{L}(L^p(\Omega), L^p(\Omega))$ is a linear operator that generates a group

$$e^{Lt} \in \mathscr{L}(L^p(\Omega), L^p(\Omega)) \quad \forall t \in \mathbb{R}$$

which does not regularize. The solutions of the initial value problem

$$u_t(\cdot,t) = L(u)(\cdot,t), \qquad u(0) = u_0 \in L^p(\Omega)$$

are given by $e^{Lt}u_0$, and they satisfy comparison results.

In this section we focus on the existence and uniqueness of solution of the problem (2) with f globally Lipschitz, and the solution will be denoted as $u(x,t,u_0)$. Furthermore, the solution associated to Eq. (2) will be given by the Variation of Constants Formula,

$$u(\cdot,t,u_0) = e^{Lt}u_0 + \int_0^t e^{L(t-s)} f(\cdot,u(s))\,ds. \tag{3}$$

To be able to prove the existence and uniqueness of solution of (2), we need first to give the following definition.

Definition 1 For $1 \le p \le \infty$, the Nemitcky operator associated to f, is defined as an operator

$$F : L^p(\Omega) \to L^p(\Omega), \text{ such that } F(u)(x) = f(x,u(x)).$$

We introduce the following problem, that is equal to (2) substituting the globally Lipschitz function f with the associated Nemitcky operator F,

$$\begin{cases} u_t(t) = (K - hI)(u)(t) + F(u(t)) \; t > 0 \\ u(0) = u_0. \end{cases} \tag{4}$$

In the proposition below, we give the existence and uniqueness of the solution to (4), with F globally Lipschitz.

Proposition 1 *For $1 \le p \le \infty$, if F is globally Lipschitz then the problem (4) has a unique global solution u for every $u_0 \in L^p(\Omega)$, that is given by the Variation of Constants Formula (3), and $u \in C^1(\mathbb{R}, L^p(\Omega))$ is a strong solution in $L^p(\Omega)$.*

The solutions of the problem (4) have some monotonocity properties:

- Given two ordered initial data, the corresponding solutions of (4) remain ordered.
- If $F(u) \ge 0, \;\; \forall u \ge 0$. Given a nonnegative initial data, the corresponding solution of (4) is nonnegative.
- Given F and G globally Lipschitz such that $F \ge G$. If we denote by $u_F(\cdot,t,u_0)$ and $u_G(\cdot,t,u_0)$ the solution to (4) with nonlinear terms F, and G respectively. Then

$$u_F(x,t,u_0) \ge u_G(x,t,u_0).$$

Definition 2 $\overline{u}$ is a **supersolution** to (4) if for $t \geq s$

$$\overline{u}(\cdot,t) \geq e^{L(t-s)}\overline{u}(s) + \int_s^t e^{L(t-r)} F(\overline{u})(\cdot,r)dr.$$

We say that $\underline{u}$ is a **subsolution** if the reverse inequality holds.

Proposition 2 *For $1 \leq p \leq \infty$. Let $\overline{u}(\cdot,t)$ be a supersolution to (4), and $u(\cdot,t,u_0)$ be a solution to (4). If F is globally Lipschitz and $\overline{u}(0) \geq u_0$, then*

$$\overline{u}(\cdot,t) \geq u(\cdot,t,u_0),$$

as long as both exist. The same is true for subsolutions if the reverse inequality holds.

3 Existence and Uniqueness of Solutions, with f Locally Lipchitz

In this section we prove the existence and uniqueness of solutions of the problem

$$\begin{cases} u_t(x,t) = (K-hI)(u)(x,t) + f(x,u(x,t)) = L(u)(x,t) + f(x,u(x,t)), \ x \in \Omega, \ t>0 \\ u(x,0) = u_0(x), \ x \in \Omega, \end{cases} \tag{5}$$

with f locally Lipschitz satisfying the increasing property (8), needed to prove the existence of solution.

Since the group e^{Lt} does not regularize, and f is locally Lipschitz we can not prove with fixed-point argument the existence and uniqueness of the solutions of the problem (5). To be able to prove the existence and uniqueness of solutions of (5), with f locally Lipschitz, we introduce a truncated globally Lipschitz function, f_k, associated to f, with $k > 0$ such that

$$f_k(x,u) = f(x,u), \quad \text{for} \quad |u| \leq k, \quad \text{for all} \ \ x \in \Omega. \tag{6}$$

We introduce the following problem, that is equal to (5) substituting the locally Lipschitz function f with the associated truncated globally Lipschitz function f_k,

$$\begin{cases} u_t(t) = (K-hI)(u)(t) + f_k(\cdot,u(t)) = L(u)(t) + F_k(u(t)), \\ u(0) = u_0. \end{cases} \tag{7}$$

Since f_k is globally Lipschitz, the associated Nemitcky operator is also globally Lipschitz. Thus, thanks to Proposition 1, the problem (7) has a unique solution in $C^1(\mathbb{R}, L^p(\Omega))$ for any initial data $u_0 \in L^p(\Omega)$.

In the following proposition we prove the existence and uniqueness of solution of (5), with f locally Lipschitz and initial data bounded.

Proposition 3 *If $f : \Omega \times \mathbb{R} \to \mathbb{R}$, sends (x, s) to $f(x, s)$, is locally Lipschitz in the variable $s \in \mathbb{R}$, uniformly with respect to $x \in \Omega$, and f satisfies that there exist $C \in \mathbb{R}$ and $D \geq 0$ such that*

$$f(\cdot, s)s \leq Cs^2 + D|s|, \ \forall s \in \mathbb{R}, \tag{8}$$

then the problem (5) with initial data $u_0 \in L^\infty(\Omega)$ has a unique global solution $u \in C^1([0, T], L^\infty(\Omega))$, for all $T > 0$.

Proof We set

$$h_0(\cdot) = \int_\Omega J(\cdot, y)dy \in L^\infty(\Omega).$$

First of all, let us prove that $(h_0 - h)s + f(\cdot, s)$ satisfies the hypothesis (8). Since $f(\cdot, s)s$ satisfies (8), then

$$\begin{aligned}(h_0 - h)s^2 + f(\cdot, s)s &\leq \sup_{x\in\Omega} |h_0(x) - h(x)|\, s^2 + Cs^2 + D|s| \\ &\leq C_1 s^2 + D|s|.\end{aligned} \tag{9}$$

We denote $C_1 = C$ to simplify the notation.

Fix $0 < M \in \mathbb{R}$. We introduce the auxiliary problem

$$\begin{cases} \dot{z}(t) = Cz(t) + D \\ z(0) = M. \end{cases} \tag{10}$$

There exists a unique solution to (10), $z \in C(\mathbb{R})$, given by

$$z(t) = -\frac{D}{C} + e^{Ct}C_2, \quad \text{with } C_2 = M + \frac{D}{C}.$$

Let $T > 0$ be an arbitrary time, then

$$|z(t)| < \max\left\{-\frac{D}{C} + e^{|C|T}C_2,\ M\right\} \quad \forall t \in [0, T]. \tag{11}$$

Let f_k be the globally Lipschitz function associated to f. We denote by $u_k(t, u_0)$ the solution to the problem

$$\begin{cases} (u_k)_t(x, t) = (K - hI)(u_k)(x, t) + f_k(x, u_k(x, t)), \ x \in \Omega \subset \mathbb{R}^N, \ t \in \mathbb{R} \\ u_k(x, 0) = u_0(x), \quad x \in \Omega. \end{cases} \tag{12}$$

Thanks to Proposition 1, we have the existence and uniqueness of solutions of (12). Moreover, the solution $u_k(t, u_0)$ is in $C^1(\mathbb{R}, L^\infty(\Omega))$.

Given $T > 0$ and $M > 0$, we choose

$$k = \max\left\{-\frac{D}{C} + e^{|C|T} C_2,\ M\right\}.$$

Thanks to the definition of f_k, (6), and (11) we have that

$$f_k(z(t)) = f(z(t)),\ \forall t \in [0, T]. \tag{13}$$

In particular, since $h,\ h_0 \in L^\infty(\Omega)$, from (9) and (13) we have that

$$(h_0 - h)z(t)^2 + f_k(\cdot, z(t))z(t) \le \tilde{C} z(t)^2 + D|z(t)| \quad . \tag{14}$$

We denote $\tilde{C} = C$ to simplify the notation.

We prove below that z is a supersolution of (12) for every $t \in [0, T]$.

The solution $z(t)$ is nonnegative for all $t \in [0, T]$. Then, thanks to (14), and since $z(t)$ is independent of the variable x, we have that $K(z(t)) = h_0 z(t)$. Thus,

$$\begin{aligned} K(z)(t) - hz(t) + f_k(\cdot, z(t)) &= (h_0 - h)z(t) + f_k(\cdot, z(t)) \\ &\le Cz(t) + D = \dot{z}(t), \quad \text{for all } t \in [0, T]. \end{aligned}$$

Let us consider the auxiliary problem

$$\begin{cases} \dot{w}(t) = Cw(t) + D \\ w(0) = -M. \end{cases} \tag{15}$$

The solution associated to (15) satisfies that

$$|w(t)| < \max\left\{\left|\frac{D}{C} + e^{|C|T} C_3\right|,\ M\right\} \quad \forall t \in [0, T]. \tag{16}$$

Analogously to z, we can prove below that w is a subsolution of (12) for every $t \in [0, T]$. Choosing now,

$$k = \max\left\{-\frac{D}{C} + e^{|C|T} C_2, M, \left|\frac{D}{C} + e^{|C|T} C_3\right|\right\}$$

and thanks to Proposition 2, for all $u_0 \in L^\infty(\Omega)$, such that $\|u_0\|_{L^\infty(\Omega)} \le M$, we obtain that

$$w(t, -\|u_0\|_{L^\infty(\Omega)}) \le u_k(t, u_0) \le z(t, \|u_0\|_{L^\infty(\Omega)}), \quad \forall t \in [0, T]. \tag{17}$$

Thanks to (11), (16) and (17) we have that

$$|u_k(t,u_0)| \leq \max\left\{-\frac{D}{C}+e^{CT}C_2, M, \left|-\frac{D}{C}+e^{CT}C_3\right|\right\} = k \quad \text{for all } t\in[0,T].$$

Therefore, $|u_k(t,u_0)| \leq k$. Thanks to the definition of f_k, we obtain that $f_k(u_k(t,u_0)) = f(u_k(t,u_0))$. Thus,

$$u_k(x,t,u_0) = u(x,t,u_0), \qquad \text{for all } t\in[0,T] \text{ and } x\in\Omega.$$

Thanks to Proposition 1, we have that $u_k(t,u_0)$ exists for all $t\in\mathbb{R}$. Hence, we have proved the existence of a strong solution of (5) for all $t\in[0,T]$, that satisfies the Variation of Constants Formula. A continuation argument, leads us to the existence of solution $u(t,u_0)$ for all $t\geq 0$ for the problem (5).

Now, let us prove the uniqueness of solution. We consider a solution $u\in C([0,T],L^\infty(\Omega))$ of the problem (5) with initial data $u_0\in L^\infty(\Omega)$. Then considering $\sup\limits_{t>0}\sup\limits_{x\in\Omega}|u(x,t,u_0)| \leq \tilde{C} < \infty$. Hence if we choose $k\geq\tilde{C}$, then $f_k(\cdot,u_k(t)) = f(\cdot,u_k(t))$ for all $t>0$. Thus, u_k and u coincide. Furthermore, from Proposition 1, we have the uniqueness of the solution u of (5) with initial data bounded, for all $t>0$. Thus, the result. □

Remark 1 In the previous Proposition 3, we have proved that the solution u of the problem (5), with initial data $u_0\in L^\infty(\Omega)$ is in fact the solution of the problem (7), with nonlinear term f_k that is the globally Lipschitz function associated to f. Then the solution u of (5) satisfies all the monotonocity properties that the solution of the problem (7) verify.

In the following proposition we state the existence and uniqueness of solution of (5) with initial data $u_0\in L^p(\Omega)$, and we state that the solution is strong in $L^1(\Omega)$.

Proposition 4 *For $1\leq p<\infty$, we assume that $f:\Omega\times\mathbb{R}\to\mathbb{R}$, sends (x,s) to $f(x,s)$, is locally Lipschitz in the variable $s\in\mathbb{R}$, uniformly with respect to $x\in\Omega$.*

If f satisfies that $f(\cdot,0)\in L^\infty(\Omega)$, and

$$\frac{\partial f}{\partial u}(\cdot,u)\leq\beta(\cdot)\in L^\infty(\Omega) \tag{18}$$

and

$$\left|\frac{\partial f}{\partial u}(\cdot,u)\right|\leq C(1+|u|^{p-1}), \quad 1<p<\infty, \tag{19}$$

then Eq. (5) with initial data $u_0\in L^p(\Omega)$ has a global unique solution $u\in C([0,T],L^p(\Omega))\cap C^1([0,T],L^1(\Omega))$, $\forall T>0$, and it is a strong solution in $L^1(\Omega)$.

4 Asymptotic Estimates

In this section, we study the asymptotic estimates of the norm $L^\infty(\Omega)$ of the solution u of the problem

$$\begin{cases} u_t(x,t) = (K-hI)(u)(x,t) + f(x,u(x,t)) = L(u)(x,t) + f(x,u(x,t)),\ x \in \Omega,\ t>0 \\ u(x,0) = u_0(x),\ x \in \Omega, \end{cases} \tag{20}$$

with f locally Lipschitz satisfying some extra-conditions we will need to add. Our aim in this section is to prove the existence of two ordered extremal equilibria, φ_m, φ_M, one minimal and another maximal, respectively, and we prove that all the solutions with initial data in $L^\infty(\Omega)$ enter between the two equilibria φ_m and φ_M, as time goes to infinity.

In this section we will consider that f in problem (20) satisfies that

$$f(x,u)u \le C(x)|u|^2 + D(x)|u|, \quad x \in \Omega,\ u \in \mathbb{R}$$

with $C \in L^\infty(\Omega)$ and $0 \le D \in L^\infty(\Omega)$. In the following proposition we give bounds of $|u(t)|$, where u is the solution to (20).

Proposition 5 *For $1 \le p \le \infty$. We assume that $f : \Omega \times \mathbb{R} \to \mathbb{R}$ is locally Lipschitz and f satisfies that there exist $C \in L^\infty(\Omega)$ and $0 \le D \in L^\infty(\Omega)$ such that*

$$f(x,u)u \le C(x)|u|^2 + D(x)|u|, \quad x \in \Omega,\ u \in \mathbb{R}. \tag{21}$$

Let $\mathscr{U}(t)$ be the solution of

$$\begin{cases} \mathscr{U}_t(t) = L(\mathscr{U}(t)) + C\,\mathscr{U}(t) + D,\ t>0 \\ \mathscr{U}(0) = |u_0|. \end{cases} \tag{22}$$

Then the solution, u, of (20), with initial data in $L^p(\Omega)$ satisfies that

$$|u(t)| \le \mathscr{U}(t), \quad \text{for all } t \ge 0.$$

In the proposition below, we give an asymptotic estimate of the norm $L^\infty(\Omega)$ of the solution of (20), that is given in terms of the norm of the equilibrium of the problem (22).

Proposition 6 *Let Φ be the equilibrium solution of (22), satisfying*

$$L(\Phi) + C(\cdot)\Phi + D(\cdot) = 0, \tag{23}$$

with $C \in L^\infty(\Omega)$ and $0 \le D \in L^\infty(\Omega)$.

If

$$\inf \sigma_{L^\infty(\Omega)}(-L - C) \geq \delta > 0, \tag{24}$$

then $\Phi \in L^\infty(\Omega)$, *and* $\Phi \geq 0$.

If $u_0 \in L^\infty(\Omega)$, *the solution* u *of the problem* (20) *satisfies that*

$$\overline{\lim_{t\to\infty}} \|u(t, u_0)\|_{L^\infty(\Omega)} \leq \|\Phi\|_{L^\infty(\Omega)}.$$

In the following proposition, we state the existence of two ordered extremal equilibria, which gives information about the set that attracts the dynamics of the solution of (20), with $u_0 \in L^\infty(\Omega)$.

Proposition 7 *If the hypothesis of Proposition 6 are satisfied, then there exist two ordered extremal equilibria,* $\varphi_m \leq \varphi_M$, *of* (20), *such that any other equilibria* ψ *of* (20) *satisfies* $\varphi_m \leq \psi \leq \varphi_M$. *Furthermore, the set*

$$\{v \in L^p(\Omega) : \varphi_m \leq v \leq \varphi_M\}$$

attracts the dynamics of the system, i.e., there exist $\underline{u}(t)$ *and* $\overline{u}(t)$ *such that* $\underline{u}(t) \leq u(t, u_0) \leq \overline{u}(t)$ *for all* $u_0 \in L^\infty(\Omega)$, *and*

$$\lim_{t\to\infty} \underline{u}(t) = \varphi_m \quad \textit{and} \quad \lim_{t\to\infty} \underline{u}(t) = \varphi_M$$

in $L^p(\Omega)$ *for all* $1 \leq p < \infty$.

References

1. Andreu, F., Mazon, J.M., Rossi, J.D., Toledo, J.: The Neumann problem for nonlocal nonlinear diffusion equations. J. Evol. Equ. **8**(1), 189–215 (2008)
2. Arrieta, J.M., Carvalho, A.N., Rodriguez-Bernal, A.: Attractors of parabolic problems with nonlinear boundary conditions. Uniform Bounds. Commun. Part. Differ. Equ. **25**(1–2), 1–37 (2000)
3. Bates, P., Chmaj, A.: An integrodifferential model for phase transitions: stationary solutions in higher dimensions. J. Stat. Phys. **95**, 1119–1139 (1999)
4. Chasseigne, E., Chaves, M., Rossi, J.D.: Asymptotic behavior for nonlocal diffusion equations. J. Math. Pures Appl. **86**, 271–291 (2006)
5. Hutson, V., Martinez, S., Mischaikow, K., Vickers, G.T.: The evolution of dispersal. J. Math. Biol. **47**(6), 483–517 (2003)
6. Rodriguez-Bernal, A., Vidal-Lopez, A.: Extremal equilibria for nonlinear parabolic equations in bounded domains and applications. J. Differ. Equ. **244**, 2983–3030 (2008)

Part II
Ordinary Differential Equations and Dynamical Systems

Analytic Approximations for Linear Differential Equations with Periodic or Quasi-periodic Coefficients

Ana Arnal and Cristina Chiralt

Abstract A perturbative procedure is proposed to compute analytic approximations to the fundamental matrix of linear differential equations with periodic or quasi-periodic coefficients. The algorithm allows one to construct high-order analytic approximations to the characteristic exponents and thus analyze the stability of the system. In addition, the approximate matrix solutions preserve by construction qualitative properties of the exact solution.

1 Introduction

The linear system of differential equations

$$\dot{Y} \equiv \frac{dY}{dt} = A(t)Y, \qquad Y(0) = I, \tag{1}$$

with $A(t)$ a T-periodic matrix, is an example of a reducible system: by means of the transformation $Y = P(t)Z$, with $P(t)$ a non singular periodic matrix, a new system $\dot{Z} = KZ$ is obtained, where now the coefficient matrix $K = P^{-1}(t)A(t)P(t) - P^{-1}(t)\dot{P}(t)$ is constant. This is the so-called Lyapunov transformation [1]. As a consequence, the solution of the original system can be written globally as $Y(t) = P(t)\exp(tK)$. This is just a rephrasing of the well known Floquet theorem for linear periodic differential equations [9].

From this result it is clear that the stability conditions of the solution $Y(t)$ only depend on the matrix K, specifically on its eigenvalues (the characteristic exponents of the system), whose real parts are uniquely determined. Thus, the trivial solution of (1) is asymptotically stable if and only if the real part of the characteristic exponents is negative, and it is stable if and only if all the characteristic exponents have non positive real part, with the vanishing or purely imaginary characteristic exponents being simple elementary divisors of the matrix $K - \lambda I$, $\lambda \in \mathbb{C}$ [9]. From

A. Arnal (✉) • C. Chiralt
Departament de Matemàtiques, Institut de Matemàtiques i Aplicacions de Castelló (IMAC), Universitat Jaume I, 12071 Castellón, Spain
e-mail: parnal@uji.es; chiralt@uji.es

F. Casas, V. Martínez (eds.), *Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 4, DOI 10.1007/978-3-319-06953-1_7

these properties, it is clear that computing the matrix K or the monodromy matrix $Y(T) = \exp(TK)$ is extremely useful. Unfortunately, although the Floquet theorem gives us information about the structure of solution of the system (1), it does not provide any practical method to get K and/or the transformation matrix $P(t)$.

Here we propose an algorithmic procedure to get approximations to both K and $P(t)$, and therefore to the solution $Y(t)$ in the form prescribed by the Floquet theorem when $A(t) = A_0 + \varepsilon A_1(t) + \varepsilon^2 A_2(t) + \cdots$ in terms of the parameter $\varepsilon > 0$. The algorithm is recursive and determines the periodic transformation $P(t)$ as the exponential of a certain matrix $\Omega(t)$. This property guarantees by construction that the approximations preserve certain qualitative properties of the exact solution. In addition the algorithm can be easily implemented with a symbolic algebra package.

If, on the other hand, the coefficient matrix $A(t)$ is quasi-periodic, the problem of reducing (1) to a system with constant coefficients is far more difficult. When the terms $A_1(t), A_2(t), \ldots$ are sufficiently small, Shtokalo [8] constructed asymptotic expansions for the solution which allowed him to examine the stability of the system. It turns out that the procedure we have developed for periodic systems can also be generalized to this setting with only minor modifications.

2 Algorithm

Let us consider the $d \times d$ system

$$\frac{\partial}{\partial t} Y(t,\varepsilon) = A(t,\varepsilon)\, Y(t,\varepsilon), \qquad Y(t_0 = 0, \varepsilon) = I \tag{2}$$

with

$$A(t,\varepsilon) = A_0 + \sum_{j \geq 1} \varepsilon^j A_j(t) = A_0 + \varepsilon A_1(t) + \varepsilon^2 A_2(t) + \cdots \tag{3}$$

and $A_j(t+T) = A_j(t)$, $j \geq 1$, for a certain $T > 0$. The goal is then to construct a transformation $P(t,\varepsilon)$ with inverse

$$Y(t,\varepsilon) \xrightarrow{P(t,\varepsilon)} Z(t,\varepsilon) = P^{-1}(t,\varepsilon)\, Y(t,\varepsilon)\, P(0,\varepsilon) \tag{4}$$

such that for the system in the new coordinates one has

$$\frac{\partial}{\partial t} Z(t,\varepsilon) = K(\varepsilon)\, Z(t,\varepsilon), \qquad Z(0,\varepsilon) = I, \tag{5}$$

with a constant coefficient matrix given by

$$K(\varepsilon) = P^{-1}(t,\varepsilon) A(t,\varepsilon) P(t,\varepsilon) + \frac{\partial P^{-1}(t,\varepsilon)}{\partial t}\, P(t,\varepsilon). \tag{6}$$

We construct $P(t,\varepsilon)$ as a near-identity transformation, i.e., $P(t,\varepsilon) = I + \mathscr{O}(\varepsilon)$, in such a way that it satisfies an equation similar to (2) but now with respect to ε. More specifically, in view of Eq. (4), we impose

$$\frac{\partial}{\partial\varepsilon} P^{-1}(t,\varepsilon) = L(t,\varepsilon) P^{-1}(t,\varepsilon), \qquad P^{-1}(t,0) = I \tag{7}$$

in terms of a (still unknown) generator $L(t,\varepsilon)$. Alternatively,

$$\frac{\partial}{\partial\varepsilon} P(t,\varepsilon) = -P(t,\varepsilon) L(t,\varepsilon), \qquad P(t,0) = I. \tag{8}$$

Once $L(t,\varepsilon)$ has been determined, it is possible to obtain $P(t,\varepsilon)$ by formally applying the Magnus expansion [3,6] to the linear equation (7), so that

$$P^{-1}(t,\varepsilon) = \exp \Omega(t,\varepsilon), \tag{9}$$

where $\Omega(t,\varepsilon)$ is an infinite series depending $L(t,\varepsilon)$ and its nested commutators.

To determine the generator $L(t,\varepsilon)$, we differentiate Eq. (6) with respect to ε and use (7)–(8) to get

$$\frac{\partial K}{\partial\varepsilon} = [L,K] + P^{-1}\frac{\partial A}{\partial\varepsilon} P + \frac{\partial L}{\partial t}, \tag{10}$$

that is,

$$\frac{\partial K}{\partial\varepsilon} = [L,K] + \mathrm{e}^{\mathrm{ad}_\Omega}\frac{\partial A}{\partial\varepsilon} + \frac{\partial L}{\partial t}, \tag{11}$$

with

$$\mathrm{e}^{\Omega}\,\frac{\partial A}{\partial\varepsilon}\,\mathrm{e}^{-\Omega} = \mathrm{e}^{\mathrm{ad}_\Omega}\frac{\partial A}{\partial\varepsilon} = \sum_{n\geq 0}\frac{1}{n!}\mathrm{ad}_\Omega^n\frac{\partial A}{\partial\varepsilon} \tag{12}$$

in terms of the adjoint operator ad: $\mathrm{ad}_\Omega B \equiv [\Omega, B] = \Omega\, B - B\,\Omega$ and $\mathrm{ad}_\Omega^n B \equiv [\Omega, \mathrm{ad}_\Omega^{n-1} B]$.

Since $A(t,\varepsilon)$ is given as a series in powers of ε, (see Eq. (3)), we determine both the generator $L(t,\varepsilon)$ and the new coefficient matrix $K(\varepsilon)$ also as formal series in ε:

$$K(\varepsilon) = \sum_{n=0}^{\infty}\varepsilon^n K_n, \qquad L(t,\varepsilon) = \sum_{n=0}^{\infty}\varepsilon^n L_{n+1}(t). \tag{13}$$

The successive terms $K_n, L_n(t)$ in (13) can be obtained from Eq. (11) by applying the following procedure:

1. Insert the series $L(t,\varepsilon)$ into Eq. (7) and compute the Magnus expansion of $\Omega(t,\varepsilon)$,

$$\Omega(t,\varepsilon) = \sum_{n=1}^{\infty} \varepsilon^n v_n(t), \tag{14}$$

in terms of $L_k(t)$. This step has been thoroughly analyzed in [4], where in particular a recursive algorithm for the computation of $v_n(t)$ is given. The first terms in the series (14) read

$$\begin{aligned} v_1 &= L_1, \\ v_2 &= \frac{1}{2}L_2, \\ v_3 &= \frac{1}{3}L_3 - \frac{1}{12}[L_1, L_2] \\ v_4 &= \frac{1}{4}L_4 - \frac{1}{12}[L_1, L_3]. \end{aligned} \tag{15}$$

2. Insert the series (14) into Eq. (12) to express $\mathrm{e}^{\mathrm{ad}_\Omega}\dfrac{\partial A}{\partial \varepsilon}$ as a power series in ε,

$$\mathrm{e}^{\mathrm{ad}_\Omega}\frac{\partial A}{\partial \varepsilon} = \sum_{n=0}^{\infty} \varepsilon^n w_n(t). \tag{16}$$

In particular,

$$\begin{aligned} w_0 &= A_1, \\ w_1 &= 2A_2 + [L_1, A_1], \\ w_2 &= 3A_3 + 2[L_1, A_2] + \frac{1}{2}[L_2, A_1] + \frac{1}{2}[L_1, [L_1, A_1]]. \end{aligned} \tag{17}$$

Again, a recursive procedure for the computation of $w_n(t)$ in (16) can be found in [4]. In general, w_n $(n \geq 1)$ depends on A_k and L_m, with $1 \leq k \leq n+1$, $1 \leq m \leq n$.

3. Finally, insert the series (13) and (16) into Eq. (11), and equate terms of the same power in ε. In this way we arrive at

$$\begin{aligned} &K_0 = A_0 \\ &\frac{dL_n}{dt} + [L_n, A_0] = nK_n - F_n, \qquad n \geq 1 \end{aligned} \tag{18}$$

with

$$F_1 \equiv w_0 = A_1 \tag{19}$$

$$F_n \equiv \sum_{j=1}^{n-1} [L_{n-j}, K_j] + w_{n-1}, \qquad n > 1. \tag{20}$$

For the first terms we have explicitly

$$\frac{dL_1}{dt} + [L_1, A_0] = K_1 - A_1$$

$$\frac{dL_2}{dt} + [L_2, A_0] = 2K_2 - 2A_2 - [L_1, K_1 + A_1]$$

$$\frac{dL_3}{dt} + [L_3, A_0] = 3K_3 - 3A_3 - [L_2, K_1 + \frac{1}{2}A_1] - [L_1, K_2 + 2A_2 + \frac{1}{2}[L_1, A_1]].$$

These equations allow us to get K_n and $L_n(t)$ recursively once K_m and $L_m(t)$ with $m = 1, \ldots, n-1$ have been previously determined.

For later use, we notice that Eq. (18) can also be written as

$$\frac{dL_n}{dt} = \mathrm{ad}_{A_0} L_n + nK_n - F_n \tag{21}$$

in terms of the linear operator ad_{A_0}.

3 The Lyapunov Transformation in Periodic Systems

Since our goal is to construct approximations to the solution of (2) according with the Floquet theorem, we choose $K(\varepsilon)$ as a constant matrix and obtain the successive terms $L_n(t)$ as periodic matrices in t: $L_n(t+T) = L_n(t)$ for all $n \geq 1$. In this way, $\Omega(t+T, \varepsilon) = \Omega(t, \varepsilon)$ and $Z(t, \varepsilon) = \exp(tK(\varepsilon))$.

To begin with, we integrate Eq. (18) over the period and divide by T:

$$\frac{L_n(T) - L_n(0)}{T} = [A_0, \frac{1}{T}\int_0^T L_n(t)dt] + nK_n - \frac{1}{T}\int_0^T F_n(t)dt. \tag{22}$$

Since L is periodic, then $L_n(T) - L_n(0) = 0$, so that

$$nK_n = \langle F_n \rangle - [A_0, \langle L_n \rangle], \tag{23}$$

where $\langle F_n \rangle$ and $\langle L_n \rangle$ denote the average of F_n and L_n over the interval $[0, T]$, respectively:

$$\langle F_n \rangle \equiv \frac{1}{T} \int_0^T F_n(t) dt, \qquad \langle L_n \rangle \equiv \frac{1}{T} \int_0^T L_n(t) dt. \tag{24}$$

On the other hand, the formal solution of Eq. (21) reads

$$L_n(t) = \mathrm{e}^{t\,\mathrm{ad}_{A_0}} L_n(0) + \mathrm{e}^{t\,\mathrm{ad}_{A_0}} \int_0^t \mathrm{e}^{-s\,\mathrm{ad}_{A_0}} (nK_n - F_n(s))\, ds. \tag{25}$$

Now, inserting (23) into this expression we get

$$L_n(t) = \mathrm{e}^{t\,\mathrm{ad}_{A_0}} L_n(0) + (I - \mathrm{e}^{t\,\mathrm{ad}_{A_0}}) \langle L_n \rangle + \mathrm{e}^{t\,\mathrm{ad}_{A_0}} \int_0^t \mathrm{e}^{-s\,\mathrm{ad}_{A_0}} (\langle F_n \rangle - F_n(s))\, ds,$$

where we have used the formal identity

$$\int_0^t \mathrm{e}^{-s\,\mathrm{ad}_{A_0}} (-\mathrm{ad}_{A_0} \langle L_n \rangle) = (\mathrm{e}^{-t\,\mathrm{ad}_{A_0}} - I) \langle L_n \rangle.$$

If we denote by $G_n(s)$ the antiderivative of $\mathrm{e}^{-s\,\mathrm{ad}_{A_0}} (\langle F_n \rangle - F_n(s))$, i.e., $G_n(t)$ is such that

$$\frac{dG_n(t)}{dt} = \mathrm{e}^{-t\,\mathrm{ad}_{A_0}} (\langle F_n \rangle - F_n(t)),$$

then clearly

$$L_n(t) = \mathrm{e}^{t\,\mathrm{ad}_{A_0}} L_n(0) + (I - \mathrm{e}^{t\,\mathrm{ad}_{A_0}}) \langle L_n \rangle + \mathrm{e}^{t\,\mathrm{ad}_{A_0}} (G_n(t) - G_n(0)). \tag{26}$$

In summary, the new constant coefficient matrix and the generator of the transformation are given recursively by

$$\begin{aligned} nK_n &= \langle F_n \rangle - [A_0, \langle L_n \rangle] \\ L_n(t) &= \langle L_n \rangle + \mathrm{e}^{t\,\mathrm{ad}_{A_0}} (L_n(0) - \langle L_n \rangle + G_n(t) - G_n(0)), \end{aligned} \tag{27}$$

for $n \geq 1$, starting with $K_0 = A_0$. Notice that there are two undetermined parameters at each step in these expressions, both related with the generator: its initial value $L_n(0)$ and the average $\langle L_n \rangle$. To construct explicitly the transformation we have to fix these values. The problem then admits infinite solutions. Next we consider just two different possibilities:

1. We fix the initial condition $L_n(0) = 0$. Then, $L_n(T) = 0$ by periodicity and (26) evaluated at $t = T$ leads to

$$0 = (I - \mathrm{e}^{T\,\mathrm{ad}_{A_0}})\langle L_n\rangle + \mathrm{e}^{T\,\mathrm{ad}_{A_0}}(G_n(T) - G_n(0)). \tag{28}$$

In other words, we can choose $\langle L_n\rangle$ as an arbitrary solution of the matrix equation (28) or alternatively,

$$\int_0^T \mathrm{e}^{-s\,\mathrm{ad}_{A_0}}[A_0, C_n]\,ds = G_n(T) - G_n(0) = \int_0^T \mathrm{e}^{-s\,\mathrm{ad}_{A_0}}(\langle F_n\rangle - F_n(s))ds, \tag{29}$$

where C_n denotes the unknown matrix. In this way, the problem is solved if we take

$$\begin{aligned} nK_n &= \langle F_n\rangle - [A_0, C_n] \\ L_n(t) &= C_n + \mathrm{e}^{t\,\mathrm{ad}_{A_0}}\,(G_n(T) - G_n(0) - C_n)\,, \end{aligned} \tag{30}$$

with C_n any particular solution of Eq. (29). As a matter of fact, this is a non-homogeneous system of d^2 linear equations with d^2 unknowns (the elements of C_n) that has a unique solution C_n if and only if $\lambda_k - \lambda_l \neq 0 \bmod \frac{2\pi i}{T}$, $k \neq l$, where λ_k, λ_l are distinct eigenvalues of A_0. Otherwise, some preliminary transformations lead the matrix A_0 to this situation [7].

In summary, if we impose the initial condition $L_n(0) = 0$ and periodicity for $L_n(t)$, then we can build explicitly the series $\Omega(t + T, \varepsilon) = \Omega(t, \varepsilon)$, with $\Omega(0, \varepsilon) = 0$, so that the solution is given by

$$Y(t,\varepsilon) = P(t,\varepsilon)\,\mathrm{e}^{tK(\varepsilon)} = \mathrm{e}^{-\Omega(t,\varepsilon)}\,\mathrm{e}^{tK(\varepsilon)} = \exp\left(-\sum_{n\geq 1}\varepsilon^n v_n(t)\right)\exp\left(t\sum_{n\geq 0}\varepsilon^n K_n\right) \tag{31}$$

where $K_0 = A_0$ and K_n, $n \geq 1$, are constant matrices. In addition, the series obtained for $K(\varepsilon)$ and $P(t, \varepsilon)$ are convergent for sufficiently small values of ε [5].

2. As a second option, we construct L_n such that its average $\langle L_n\rangle = 0$. In that case, from (27),

$$K_n = \frac{1}{n}\langle F_n\rangle. \tag{32}$$

Then we determinate the value of $L_n(0)$ so that $L_n(t)$ in (27) is T-periodic, in particular $L_n(T) = L_n(0)$. From (27) we get

$$L_n(0) = \mathrm{e}^{T\,\mathrm{ad}_{A_0}}(L_n(0) + G_n(T) - G_n(0))$$

or

$$\int_0^T \frac{d}{ds}\left(\mathrm{e}^{-s\,\mathrm{ad}_{A_0}} L_n(0)\right) ds = G_n(T) - G_n(0).$$

Since

$$\frac{d}{ds}\left(\mathrm{e}^{-s\,\mathrm{ad}_{A_0}} L_n(0)\right) = \frac{d}{ds}\left(\mathrm{e}^{-sA_0} L_n(0) \mathrm{e}^{sA_0}\right) = \mathrm{e}^{-sA_0}\left(L_n(0)A_0 - A_0 L_n(0)\right)\mathrm{e}^{sA_0},$$

it turns out that $L_n(0)$ has to satisfy Eq. (29). Therefore, the new coefficient matrix and the corresponding generator are given by

$$\begin{aligned} K_n &= \frac{1}{n}\langle F_n \rangle \\ L_n(t) &= \mathrm{e}^{t\,\mathrm{ad}_{A_0}}(C_n + G_n(t) - G_n(0)), \end{aligned} \tag{33}$$

where $C_n = L_n(0)$ is any solution of (29). In general $L_n(0) \neq 0$ and therefore $\Omega(0,\varepsilon) \neq 0$, so that the solution of (2) reads

$$Y(t,\varepsilon) = \mathrm{e}^{-\Omega(t,\varepsilon)}\,\mathrm{e}^{tK(\varepsilon)}\,\mathrm{e}^{\Omega(0,\varepsilon)}. \tag{34}$$

Here $\Omega(t+T,\varepsilon) = \Omega(t,\varepsilon)$ is computed with the generators L_n. In consequence

$$Y(t+T,\varepsilon) = Y(t,\varepsilon)\,\mathrm{e}^{-\Omega(0,\varepsilon)}\,\mathrm{e}^{TK(\varepsilon)}\,\mathrm{e}^{\Omega(0,\varepsilon)}.$$

We notice that, although the structure prescribed by Floquet's theorem is no longer reproduced, $M \equiv \mathrm{e}^{-\Omega(0,\varepsilon)}\,\mathrm{e}^{TK(\varepsilon)}\,\mathrm{e}^{\Omega(0,\varepsilon)}$ *is* a monodromy matrix, with the same eigenvalues as $\mathrm{e}^{TK(\varepsilon)}$. In other words, the eigenvalues of the new matrix $K(\varepsilon)$ given by (33) are also the characteristic exponents of the system.

4 Generalization to the Quasi-periodic Case

Let us consider now Eq. (3) in the quasi-periodic case, i.e., when the matrices $A_j(t)$, $j = 1, 2, \ldots$, in (3) are of the form

$$A_j(t) = \sum_{l=1}^{r} C_{j,l}\, e^{i\mu_l t}. \tag{35}$$

Here $C_{j,l}$ are constant matrices, and μ_l are real numbers, so that the elements of the matrices $A_j(t)$ are trigonometric polynomials with arbitrary frequencies μ_l. The algorithm proposed by Shtokalo [8] for analyzing the stability of the trivial solution

of system (2) consists essentially in constructing a change of variables that transform Eq. (2) into (5),

$$\frac{\partial}{\partial t} Z(t,\varepsilon) = \left(A_0 + \sum_{j\geq 1} \varepsilon^j K_j \right) Z(t,\varepsilon), \tag{36}$$

where K_j are constant matrices. In Shtokalo's procedure, the change of variables and the matrix $K(\varepsilon)$ are constructed perturbatively, as power series of ε, without paying much attention to the approximations of the solution of (3) and the preservation of the main qualitative properties if may possess [5, 8].

It turns out that the procedure developed in the previous sections for constructing the Lyapunov transformation for periodic linear systems can also be applied in this setting with only minor changes. To proceed, let us first recall that for a quasi-periodic function $f(t)$, there exists the limit

$$\lim_{T\to\infty} \frac{1}{T} \int_a^{a+T} f(t)dt = \langle f \rangle, \tag{37}$$

uniformly with respect to a. The number $\langle f \rangle$ is called the *mean value* of the quasi-periodic function $f(t)$. In addition, this mean value defined for quasi-periodic functions coincides with the usual mean value over the period for periodic functions. Moreover, if $f(t)$ is a trigonometric polynomial,

$$f(t) = C_0 + \sum_{l=1}^{r} C_l\, e^{i\mu_l t},$$

where $\mu_l \neq 0$, $l = 1, \ldots, r$, the mean value $\langle f \rangle = C_0$.

Again, the starting point is Eq. (18). Integrating over the interval $t \in [0, T]$, for an arbitrary $T > 0$, and dividing by T, we get Eq. (22). Taking the limit $T \to \infty$ results in

$$\lim_{T\to\infty} \frac{L_n(T)}{T} = [A_0, \langle L_n \rangle] + nK_n - \langle F_n \rangle.$$

Since we aim to construct the terms of the generator as trigonometric polynomials we impose

$$\lim_{T\to\infty} \frac{L_n(T)}{T} = 0,$$

so that we recover in this setting the expressions (27) for K_n and L_n, where now $\langle \cdot \rangle$ denotes the mean value (37).

At this point, at least two alternatives are possible:

1. Choose $L_n(0) = 0$. Then, a trigonometric polynomial for $L_n(t)$ results as long as $\langle L_n \rangle = -G_n(0)$. In other words,

$$K_n = \frac{1}{n}\langle F_n \rangle + \frac{1}{n}[A_0, G_n(0)] \tag{38}$$
$$L_n(t) = -G_n(0) + \mathrm{e}^{t\,\mathrm{ad}_{A_0}} G_n(t).$$

2. Determine L_n as a trigonometric polynomial with zero mean value, $\langle L_n \rangle = 0$. This can be achieved by taking $L_n(0) = G_n(0)$, and thus

$$K_n = \frac{1}{n}\langle F_n \rangle \tag{39}$$
$$L_n(t) = \mathrm{e}^{t\,\mathrm{ad}_{A_0}} G_n(t).$$

A detailed treatment of this case will be the subject of subsequent work [2].

5 Illustrative Example

We next illustrate the algorithm on a simple periodic example. In particular, we consider the system

$$\begin{aligned} \dot{y}_1 &= \varepsilon(-1 + 2\sin t)y_1 + \varepsilon\, y_2 \\ \dot{y}_2 &= -y_2 + \varepsilon\, y_1 \end{aligned} \tag{40}$$

worked out by Malkin [1]. Here ε is a real parameter and the period $T = 2\pi$. Using the method of small parameters, he showed that the characteristic exponents of the system are negative at least for $\varepsilon < 1/9$, whereas in [9] the domain of values of ε that ensure asymptotic stability is extended up to $\varepsilon < 2/3$.

The fundamental matrix $Y(t, \varepsilon)$ corresponding to system (40) verifies Eq. (4) with

$$A(t, \varepsilon) = A_0 + \varepsilon A_1(t) \equiv \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix} + \varepsilon \begin{pmatrix} -1 + 2\sin t & 1 \\ 1 & 0 \end{pmatrix}. \tag{41}$$

First we carry out the first procedure by fixing $L_n(0) = 0$, i.e., we determine K_n and L_n by Eq. (30), up to $n = 10$ and compute the solution matrix (31). In Fig. 1 we plot the difference between the Frobenius norm of our approximation, $Y(t, \varepsilon)$, and the exact result (as determined by numerical integration) when $n = 5$ and $n = 10$ terms are taken in the series.

Next we compute the eigenvalues of $K(\varepsilon)$ as a function of ε by applying the second alternative, i.e., by means of (33), and compare with the exact result (as

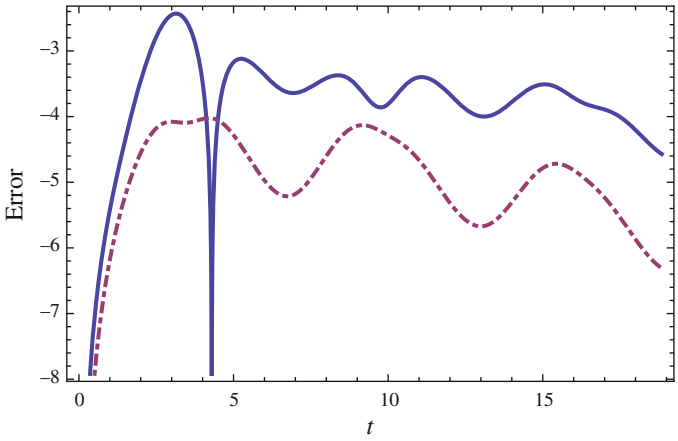


Fig. 1 Error in the approximation (in logarithmic scale) between the approximation of order ε^5 (*solid line*) and order ε^{10} (*dashed line*) with respect to the exact solution

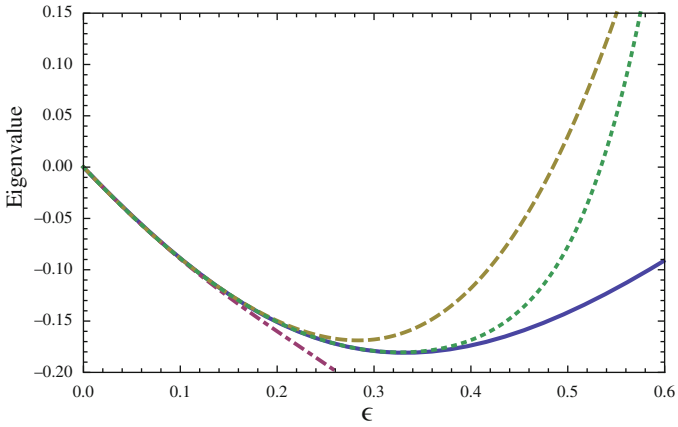


Fig. 2 One of the characteristic exponents of system (40), obtained by direct numerical integration (*solid line*), and by the perturbative algorithm of order ε^2 (*dot-dashed line*), ε^4 (*dashed line*) and ε^{10} (*dotted line*), as a function of ε

determined by the numerical integration of Eq. (40) with 25 digits of accuracy). One of the eigenvalues turns out to be always negative, whereas the second one is negative only for $\varepsilon < 0.745023$, so that it is this value which determines the stability region of the system.

In Fig. 2, we represent this exact eigenvalue (solid line) together with the results rendered by the perturbative algorithm of order ε^2 (dot-dashed line), ε^4 (dashed line) and ε^{10} (dotted line).

Notice that higher order approximations provide results that are indistinguishable from the exact value for increasingly larger values of the perturbation parameter ε.

Acknowledgements This work has been partially supported by project MTM2010-18246-C03-02 from Ministerio de Ciencia e Innovación (Spain).

References

1. Adrianova, L.Ya.: Introduction to Linear Systems of Differential Equations. AMS, Providence (1995)
2. Arnal, A., Chiralt, C., Casas, F: On a practical algorithm to analyze the reducibility of quasi-periodic linear systems. (2014, in progress)
3. Blanes, S., Casas, F., Oteo, J.A., Ros, J.: The Magnus expansion and some of its applications. Phys. Rep. **470**, 151–238 (2009)
4. Casas, F., Chiralt, C.: A Lie–Deprit perturbation algorithm for linear differential equations with periodic coefficients. Discret. Cont. Dyn-A. **34**, 959–975 (2014)
5. Erugin, N.P.: Linear Systems of Differential Equations. Academic Press/Elsevier, New York (1966)
6. Magnus, W.: On the exponential solution of differential equations for a linear operator. Commun. Pure Appl. Math. **7**, 649–673 (1954)
7. Roseau, M.: Vibrations non linéaires et théorie de la stabilité. Springer, Berlin (1966)
8. Shtokalo, I.Z.: Linear Differential Equations with Variable Coefficients. Hindustan, Delhi (1961)
9. Yakubovich, V.A., Starzhinskii, V.M.: Linear Differential Equations with Periodic Coefficients. Wiley, New York (1975)

Building Non Singular Morse-Smale Flows on 3-Dimensional Lens Spaces

Beatriz Campos and Pura Vindel

Abstract Fat handles are flow manifolds diffeomorphic to tori; therefore, each attractive (repulsive) fat handle can be identified along its boundary with a solid torus with one repulsive (attractive) orbit in its core in such a way that a NMS flow on a lens space $L(p, q)$ is obtained.

1 Introduction

Morse-Smale flows are the structurally stable flows on 2-dimensional manifolds, forming a dense open subset of C^∞ -vector fields. For the three dimensional case, Morse-Smale flows are not dense but they define an open set in the set of C^1-vector fields; in most cases they can be studied from Non-Singular Morse-Smale flows.

Non Singular Morse Smale flows (NMS for short) have been widely studied. D. Asimov [1] showed that every manifold with Euler characteristic zero admits this type of flows unless its dimension is three. J.W. Morgan [7] proved that a 3-dimensional manifold prime to $S^2 \times S^1$ admits such flows if and only if it is a graph manifold.

This kind of flows are characterized by their non-wandering set consisting of a finite number of closed hyperbolic orbits and the transversal intersections of their stable and unstable manifolds.

It is not easy to find a complete characterization of flows defined on three-dimensional manifolds. Some achievements in this direction has been made by M. Wada [8], who obtains the topological characterization of the links of periodic orbits in S^3. This characterization for NMS flows in the space $S^2 \times S^1$ have been obtained by A. Cordero et al. in [6].

Despite these important results, different NMS flows can be characterized by the same link. Therefore, to obtain the complete description of a flow it is necessary to reproduce its phase space. To obtain the phase portrait is very hard for 3-manifolds,

B. Campos (✉) • P. Vindel
Departament de Matemàtiques, Institut de Matemàtiques i Aplicacions de Castelló, Universitat Jaume I, 12071 Castelló, Spain
e-mail: campos@uji.es; vindel@uji.es

F. Casas, V. Martínez (eds.), *Advances in Differential Equations and Applications*,
SEMA SIMAI Springer Series 4, DOI 10.1007/978-3-319-06953-1_8

especially when the number of periodic orbits increases. In a previous paper [2], we obtain some NMS flows on lens spaces, with only one saddle orbit.

We prove in [5] that flows with unknotted and unlinked saddle orbits, denoted as $\mathscr{F}_A\left(S^3\right)$, can be obtained from the identification of fat handles along their boundaries.

From these results, in this paper we obtain flows on lens spaces by identifying fat handles (Proposition 4): one attractive (repulsive) fat handle is identified with one solid torus with one repulsive (attractive) orbit in its core in such a way that a flow on a lens space $L(p,q)$ is obtained.

Now, we recall some definitions and results on which our work is based.

A non singular Morse-Smale flow (or NMS for short) is a flow without fixed points, consisting of a finite number of hyperbolic periodic orbits where the intersections of stable and unstable manifolds of the saddle orbits are transversal.

D. Asimov [1] and J.W. Morgan [7] established a correspondence between NMS flows and round handle decompositions of the corresponding manifold. These flows are defined on flow manifolds. A flow manifold is a pair $(M, \partial_- M)$ where a nonsingular vector field on M exists, pointing inwards on $\partial_- M$ and outwards on $\partial_+ M$ and satisfying $\partial M = \partial_- M \cup \partial_+ M$, $\partial_- M \cap \partial_+ M = \emptyset$.

Proposition 1 (Morgan) *Given a flow manifold (X,∂_X) with a NMS flow, then (X,∂_X) has a round handle decomposition whose core circles are the closed orbits of the flow.*

For the case of dimension 3, the round handles are diffeomorphic to tori and correspond to 0-handles when there is a repulsive periodic orbit in the core, to 2-handles if there is an attractive periodic orbit in the core and to 1-handles if the orbit is a saddle; 0, 1 and 2 are the indices of the periodic orbits. A set of indexed periodic orbits is called an indexed link.

The round handle decomposition for a compact, orientable 3-manifold M was modified by Morgan:

$$\emptyset = M_0 \subset M_1 \subset \cdots \subset M_i \subset M_{i+1} \subset \cdots \subset M_N = M \tag{1}$$

where each manifold M_i, called fat round handle, is obtained from M_{i-1} by attaching a round 1-handle by means of one or two attaching circles (see Fig. 1).

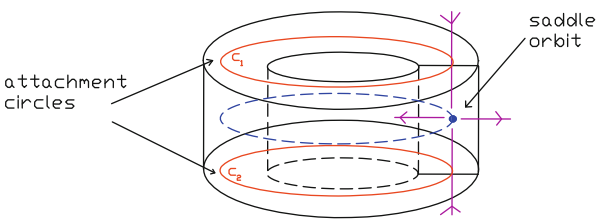


Fig. 1 Round 1-handle

The round handle decomposition of a NMS flow gives the sequence of 1-handle attachments. Each attachment on a fat round handle yields to a new fat round handle. In the following section we show in detail these fat handles and we see the different types of fat handles obtained when the number of saddle orbits increases.

2 Fat Handles

The 3-sphere S^3 is obtained joining two solid tori, by identifying transversal circles of one torus with longitudinal circles of the other.

Therefore, a polar NMS flow on S^3 can be obtained by identifying properly one repulsive and one attractive tori along their boundaries.

In [5] we obtain NMS flows on S^3 with unknotted and unlinked saddle orbits, denoted by $\mathscr{F}_A(S^3)$, by identifying one repulsive and one attractive tori along their boundaries. These tori, with a flow going inwards or outwards, correspond to the fat round handles.

Given a flow $\varphi \in \mathscr{F}_A(S^3)$, a repulsive fat handle is obtained by removing one attractive orbit and an attractive fat handle is obtained by removing one repulsive orbit. The *basic fat round handles* are the fat handles with one saddle orbit (see Figs. 2 and 3) and we denote them by describing the periodic orbits that contain. If the fat handle has an attractive or repulsive orbit in its core we refer to it as thick torus; if it has no orbit in its core, we refer to it as solid torus.

In the following, let h denote the Hopf link, let d denote a trivial separated knot corresponding to an attractive or repulsive orbit, let u denote a trivial separated knot corresponding to a saddle orbit and let $\cdot$ denote the separated sum of links.

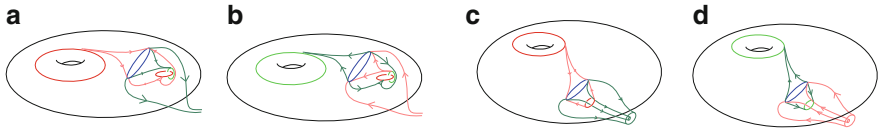


Fig. 2 Basic fat handles of class $[I]$. (**a**) Repulsive fat round handle: (h.d.u); (**b**) Attractive fat round handle: (h.d.u); (**c**) Repulsive fat round handle: (d.d.u); (**d**) Attractive fat round handle: (d.d.u)

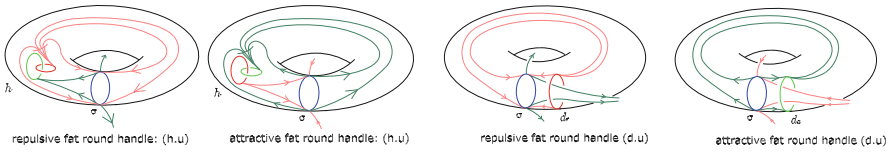


Fig. 3 Basic fat handles of class $[II]$ and $[III]$

We classify the fat handles obtained from $\mathscr{F}_A(S^3)$ flows in the following way:

- A repulsive (attractive) fat handle belongs to class $[I]$ if it corresponds to a thick torus with a repulsive (attractive) orbit filling the essential hole of the torus and the invariant manifolds of the saddles orbits go outwards (inwards) the torus by means of inessential circles.
- A repulsive (attractive) fat handle belongs to class $[II]$ if it corresponds to a solid torus, the invariant manifolds of the saddles orbits go outwards (inwards) the torus by means of essential circles and there is not any attractive or repulsive orbit in the canonical region of the identification.
- A repulsive (attractive) fat handle belongs to class $[III]$ if it corresponds to a solid torus, the invariant manifolds of the saddles orbits go outwards (inwards) the torus by means of essential and inessential circles and there is one attractive or repulsive orbit, filling a non essential hole in the torus, in the canonical region of the identification.

From the identification of one attractive and one repulsive basic fat handles we obtain the $\mathscr{F}_A(S^3)$ flows with two saddle orbits; by removing one repulsive (attractive) orbit in these flows we obtain iterated fat handles with two saddles. Following this process we obtain fat handles with n saddle orbits (see [5]) and we prove that they can be classified in one of these three classes defined above.

Proposition 2 *For $\mathscr{F}_A\left(S^3\right)$ –flows, a fat handle with n saddle orbits belongs to class $[I]$, $[II]$ or $[III]$.*

When there are not heteroclinic trajectories connecting saddles, a fat handle with n saddle orbits is obtained by the iterated connected sum of tori [3]. Two examples are showed in Fig. 4.

Let us remark that the identification along their boundaries of two fat handles without any orbit in their cores yields to a transversal intersection of two invariant manifolds of saddle orbits (see [4]).

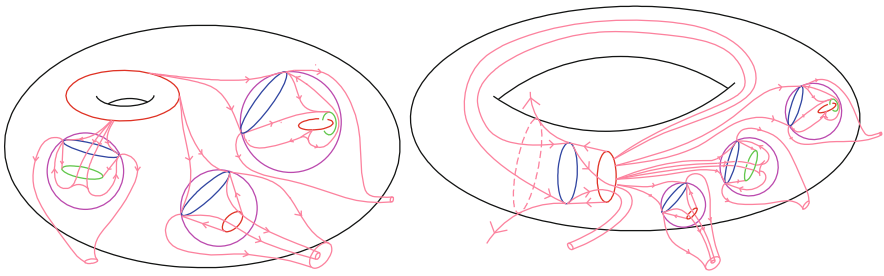


Fig. 4 Fat handles of class $[I]$ and [III]

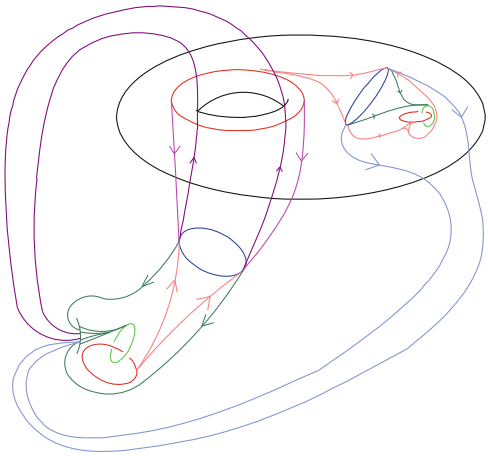


Fig. 5 Identification of one fat handle of class $[I]$ and one of class $[II]$

Along the proof of these results, in [4], all the flows φ with n unlinked saddles are obtained. Therefore,

Theorem 1 *A flow $\varphi \in \mathscr{F}_A\left(S^3\right)$ with n saddle orbits can be obtained by identifying fat handles along their boundaries.*

This result enables to build any flow on the 3-sphere with unlinked and unknotted saddle orbits. We reproduce the complete phase space of these NMS flows on the 3-sphere when two fat handles are identified along their boundaries in such a way that S^3 is obtained; that is, by identifying longitudinal circles of one of the tori with the transversal circles of the other torus. Let us remark that the fat handles must be properly identified in order to reach a flow on S^3. For example, a fat handle of class $[I]$ can not be identified with fat handle of class $[II]$ because a bitorus is obtained (see Fig. 5) and the boundaries of round handles embedded in S^3 must be tori.

This method also permits to obtain NMS flows on different 3-manifolds if the circles of the tori are identified in another way. Following the same process used on S^3 we build NMS flows on other lens spaces. Let us notice that, for the 3-sphere all the orbits are local and we can obtain a flow by identifying the fat handles along their boundaries. But for any $L(p,q)$ lens space there can be local and global orbits and a very complicated picture may appear. So, we only can assure a NMS flow if we identify one of the previous fat handles with one torus with one attractive (or repulsive) orbit in its core.

3 Building Flows in Lens Spaces

The lens space $L(p,q)$ is the 3-manifold of Heegaard genus 1 whose Heegaard diagram consists of a $(p,q)-$torus knot on the surface of a solid torus; namely, $L(p,q)$ is the result of joining two solid tori τ_1 and τ_2 via a homeomorphism

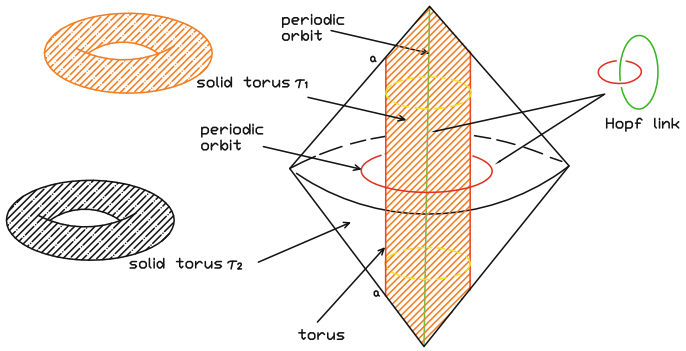


Fig. 6 S^3

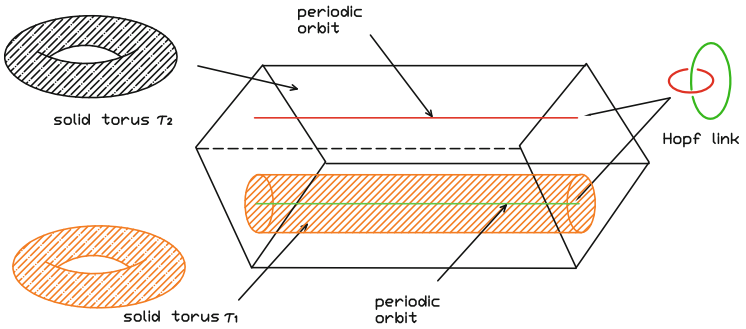


Fig. 7 $S^2 \times S^1$

$h : \partial\tau_1 \to \partial\tau_2$ where h takes a meridian m on $\partial\tau_1$ to a (p, q)-torus knot on $\partial\tau_2$. The 3-manifold resulting directly from this identification is difficult to visualize except, perhaps, $L\,(1, 0)$ that corresponds to the 3-sphere and $L\,(0, 1)$, equivalent to $S^2 \times S^1$. As we have seen before S^3 is obtained by identifying latitudes on $\partial\tau_1$ with meridians on $\partial\tau_2$ and vice versa. Similarly, $S^2 \times S^1$ is obtained by identifying longitudinal circles on $\partial\tau_1$ with longitudinal circles on $\partial\tau_2$ and transversal circles with transversal circles. We can see these spaces in Figs. 6 and 7. In both cases, one attractive and one repulsive orbit are in the core of each solid torus, and they are linked. So, the easier flow in both spaces is a polar flow corresponding to the Hopf link.

Proposition 3 *A polar flow always can be obtained on a lens space.*

Proof The polar flow is obtained by the identification along their boundaries of one torus with an attractive orbit (a 2-handle) in its core and another torus with a repulsive orbit in its core (a 0-handle). This identification is made via a homeomorphism $h : \partial\tau_1 \to \partial\tau_2$ where h takes a meridian m on $\partial\tau_1$ to a (p, q)-torus knot on $\partial\tau_2$. We have the simplest round handle decomposition and

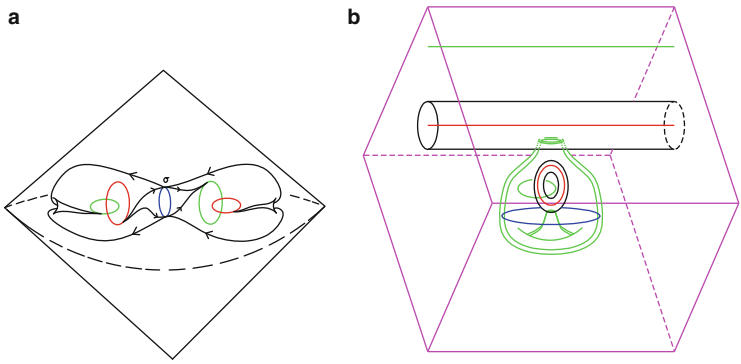


Fig. 8 NMS flows on different lens spaces. (**a**) NMS flow defined on S^3; (**b**) NMS flow defined on $S^2 \times S^1$

its corresponding NMS flow on a lens space $L\,(p, q)$ is obtained. This flow is called polar flow and the set of its periodic orbits is the Hopf link. □

The fat handles showed in Sect. 2 are flow manifolds diffeomorphic to tori; therefore, each of these attractive (repulsive) fat handles can be identified with one solid torus with one repulsive (attractive) orbit in its core in such a way that a flow on a lens space $L(p, q)$ is obtained.

For example, we show in Fig. 8a the flow on S^3 obtained by identifying longitudinal circles of the repulsive basic fat handle $h \cdot d \cdot u$ with transversal circles of one attractive solid torus and we show in Fig. 8b the flow on $S^2 \times S^1$ obtained by identifying longitudinal circles of the repulsive basic fat handle $h \cdot d \cdot u$ with longitudinal circles of one attractive solid torus.

Let us observe that the link of periodic orbits is the same for both flows, $h \cdot h \cdot u$, but the topology of the orbits is different. All the periodic orbits embedded in S^3 are local whereas in $S^2 \times S^1$, some of them can be global. Recall that a local orbit is an orbit that can be isolated in a three-dimensional disk D^3 and an global orbit can not be isolated in a D^3.

Similarly, from the flows on S^3 we can obtain NMS flows on different lens spaces depending on the way the tori are identified.

Proposition 4 *Given a flow on one attractive (repulsive) fat handle τ_1, a NMS flow on a lens space $L(p, q)$ can be obtained by identifying meridians on $\partial\tau_1$ with (p, q)-torus knots on the boundary of a repulsive (attractive) solid torus τ_2.*

Proof As we said before, $L(p, q)$ can be formed by the identification of two solid tori τ_1 and τ_2, via a homeomorphism $h : \partial\tau_1 \to \partial\tau_2$ where h takes a meridian m on $\partial\tau_1$ to a (p, q)-torus knot on $\partial\tau_2$.

We proved in [5] that the fat handles with n unknotted and unlinked saddles orbits are tori. So, we can consider τ_1 as one repulsive (attractive) fat handle with n saddle orbits and τ_2 as one solid torus with one attractive (repulsive) orbit in its core, i.e., τ_2 is a 2-handle (0-handle).

These two tori are identified by means the homeomorphism previously defined from $\partial\tau_1$ onto $\partial\tau_2$.

As τ_2 has only one attractive (repulsive) orbit in its core, the whole flow going inwards through $\partial\tau_2$ is collected by this orbit. The result is a NMS flow on the 3-dimensional lens space $L(p, q)$. □

Acknowledgements Supported by Ministerio de Ciencia y Tecnología MTM2011-28636-C02-02 and by Universitat Jaume I P11B2011-30

References

1. Asimov, D.: Round handles and non-singular Morse-Smale flows. Ann. Math. **102**, 41–54 (1975)
2. Campos, B., Cordero, A., Martínez Alfaro, J., Vindel, P.: NMS flows on three-dimensional manifolds with one Saddle periodic orbit. Acta Math. Sin. (Engl. Ser.) **20**(1), 47–56 (2004)
3. Campos, B., Vindel, P.: NMS flows on S^3 with no heteroclinic trajectories connecting saddle orbits. J. Dyn. Differ. Equ. **24**(2), 181–196 (2012). doi:10.1007/s10884-012-9247-4
4. Campos, B., Vindel, P.: Transversal intersections of invariant manifold of NMS flows on S^3. Discret. Contin. Dyn. Syst. A **32**(1), 41–56 (2012). doi:10.3934/dcds.2012.32.41
5. Campos, B., Vindel, P.: Fat handles and phase portraits of non singular Morse-Smale flows on S^3 with unknotted saddle orbits. Adv. NonLinear Stud. **14**, 605–617 (2014). arXiv:1403.5174
6. Cordero, A., Martínez Alfaro, J., Vindel, P.: Round handle decomposition of $S^2 \times S^1$. Dyn. Syst. **22**(2), 179–202 (2007)
7. Morgan, J.W.: Non-singular Morse-Smale flows on 3-dimensional manifolds. Topology **18**(1), 41–53 (1979)
8. Wada, M.: Closed orbits of non-singular Morse-Smale flows on S^3. J. Math. Soc. Jpn. **41**(3), 405–413 (1989)

Parameterization Method for Computing Quasi-periodic Reducible Normally Hyperbolic Invariant Tori

Marta Canadell and Àlex Haro

Abstract We consider the problem of numerically computing quasi-periodic normally hyperbolic invariant tori (NHIT) with fixed frequency as well as their invariant bundles. The algorithm is based on a KAM scheme to find the parameterization of a torus with fixed Diophantine frequency (by adjusting parameters of the model), and suitable Floquet transformations that reduce the linearized dynamics to constant coefficients. We apply this method to continue curves of quasi-periodic NHIT of a perturbed dynamical system and to explore the mechanism of breakdown of these invariant tori. We observe in these continuations that the invariant bundles may collide even if the Lyapunov multipliers remain separated.

1 Introduction

It is well known the importance of the study of dynamical systems, and how the invariant objects help to describe the global dynamics. The goal of this work is to present an efficient algorithm to compute one particular kind of invariant objects: normally hyperbolic invariant tori with internal dynamics conjugated to a (Diophantine) rotation. The convergence of the algorithm enters in the realm of KAM theory [7].

The problem of finding quasi periodic normally hyperbolic invariant tori with an specific set of frequencies has been already considered in the literature. The necessity of external parameters to adjust the frequencies is considered, e.g., in [1]. However, these rigorous results have a perturbative nature and are hardly applicable in numerical computations far from perturbative regime.

The usual numerical methods for computing invariant tori with quasi periodic dynamics are based on solving a system of non-linear equations arising from a

M. Canadell (✉) • À. Haro
Departament de Matemática Aplicada i Anàlisi, Universitat de Barcelona, Gran Via 585, 08007 Barcelona, Spain

School of Mathematics, Georgia Institute of Technology, 686 Cherry Street, 30332 Atlanta, Georgia, USA
e-mail: marta.canadell@math.gatech.edu; alex@maia.ub.es

F. Casas, V. Martínez (eds.), *Advances in Differential Equations and Applications*,
SEMA SIMAI Springer Series 4, DOI 10.1007/978-3-319-06953-1_9

Fourier discretization of the invariance equation [8]. In this particular problem, [15] use continuation methods without adjusting parameters (but having the frequency as an unknown). However, the method fails when it crosses strong resonances. On the other side, [16] uses the adjust of parameters, but the use of a large matrix method cannot be go ahead because of the colossal computation time and the memory needed to store the large matrices arising from the straightforward application of a Newton method.

In this work, we describe a Newton-like method to compute an invariant torus with fixed frequency based on reducibility of the normal dynamics, and the adjust of parameters, which allows us to obtain accurate results up to parameter values even very close to the torus breakdown.

2 The Setting

Let $F_a : \mathbb{T}^d \times \mathbb{R}^n \to \mathbb{T}^d \times \mathbb{R}^n$, $m = d + n$, be a family of diffeomorphisms for the parameter $a \in \mathbb{R}^d$. Let $\mathscr{K}$ be a d-dimensional (parameterized) torus, that is $\mathscr{K} = K(\mathbb{T}^d)$ where $K : \mathbb{T}^d \to \mathbb{T}^d \times \mathbb{R}^n$ is an injective immersion.

We say that the torus parameterized by K, $\mathscr{K}$, is F_a-**invariant** with a *quasi-periodic* motion given by the frequency $\omega \in \mathbb{R}^d$, if K satisfies the invariance equation:

$$F_a(K(\theta)) - K(\theta + \omega) = 0. \tag{1}$$

Note that (1) is an equation for K and a given the family F_a. The frequency of the motion $\omega \in \mathbb{R}^d$ satisfies the Diophantine condition if $|\omega \cdot q - p| \geq \gamma |q|_1^{-\tau}$, $q \in \mathbb{Z}^d \setminus \{0\}$, $p \in \mathbb{Z}$.

Heuristically, we say that $\mathscr{K}$ is a **Normally Hyperbolic Invariant Torus** (NHIT for short) of F_a if it is F_a-invariant and the tangent bundle of $\mathbb{T}^d \times \mathbb{R}^n$ restricted to $\mathscr{K}$, $\mathrm{T}_{\mathscr{K}}(\mathbb{T}^d \times \mathbb{R}^n)$, splits into three continuous subbundles

$$\mathrm{T}_{\mathscr{K}}(\mathbb{T}^d \times \mathbb{R}^n) = \mathrm{N}^S\mathscr{K} \oplus \mathrm{T}\mathscr{K} \oplus \mathrm{N}^U\mathscr{K} \tag{2}$$

such that $\mathrm{D}F_a$ contracts $\mathrm{N}^S\mathscr{K}$ more sharply than $\mathrm{T}\mathscr{K}$ and $\mathrm{D}F_a$ expands $\mathrm{N}^U\mathscr{K}$ more sharply than $\mathrm{T}\mathscr{K}$ [9, 14]. Bundles $\mathrm{N}^S\mathscr{K}$ and $\mathrm{N}^U\mathscr{K}$ are referred to as the stable and the unstable subbundles of $\mathscr{K}$, respectively.

Here, we consider the problem of numerically computing NHIT with a fixed rotation of (Diophantine) frequency ω, where it is needed to adjust parameters a to keep the frequency fixed (as in [3]). Following [12], we find very useful to compute the torus and the bundles at the same time by using a Newton method.

In particular, we look for invariant tori $\mathscr{K}$ parametrized by K that are homotopic to the zero-section of $\mathbb{T}^d \times \mathbb{R}^n$, $\mathbb{T}^d \times \{0\}$, that is

$$K(\theta) = \begin{pmatrix} \theta \\ 0 \end{pmatrix} + K_p(\theta),$$

where $K_p : \mathbb{T}^d \to \mathbb{R}^d \times \mathbb{R}^n$ is 1-periodic in the θ-variables. For each $\theta \in \mathbb{T}^d$, the d column vectors of the $m \times d$ matrix $\mathrm{D}K(\theta)$ provide a basis of the fiber $\mathrm{T}_{K(\theta)}\mathscr{K}$ of the tangent bundle. Hence, the matrix-valued map $L : \mathbb{T}^d \to \mathbb{R}^{m\times d}$ defined as $L(\theta) = \mathrm{D}K(\theta)$, provides a global frame for the tangent bundle. Also, $\mathrm{N}\mathscr{K}$ is defined by a matrix-valued map $N : \mathbb{T}^d \to \mathbb{R}^{m\times n}$ generated by n vectors linearly independents to $L(\theta)$ for each $\theta \in \mathbb{T}^d$, so that the column vectors of $L(\theta)$ joined with the column vectors of $N(\theta)$ form a basis of $\mathrm{T}_{K(\theta)}\mathbb{T}^d \times \mathbb{R}^n \simeq \mathbb{R}^m$. In other words, the matrix valued map $P : \mathbb{T}^d \to \mathbb{R}^{m\times m}$, obtained by juxtaposing L and N so that $P(\theta) = (L(\theta)\ \ N(\theta))$, provides an adapted frame around the torus.

From the definition of the invariance equation we get the tangent bundle $\mathrm{T}\mathscr{K}$ invariant, so that

$$\mathrm{D}F_a(K(\theta))L(\theta) - L(\theta + \omega) = 0. \tag{3}$$

It is also desirable to work with a normal bundle $\mathrm{N}\mathscr{K}$ which is also invariant. Using global frames, this reads as

$$\mathrm{D}F_a(K(\theta))N(\theta) - N(\theta + \omega)\Lambda_N(\theta) = 0, \tag{4}$$

for a suitable dynamics on the normal bundles $\Lambda_N : \mathbb{T}^d \to \mathbb{R}^{n\times n}$. In such a case, the adapted frame P introduced above reduces the linearized dynamics to a block diagonal matrix $\Lambda = \mathrm{blockdiag}(\mathrm{Id}, \Lambda_N) : \mathbb{T}^d \to \mathbb{R}^{m\times m}$

$$P(\theta + \omega)^{-1}\mathrm{D}F_a(K(\theta))P(\theta) - \Lambda(\theta) = 0. \tag{5}$$

Under normal hyperbolicity properties, the invariant normal bundle decomposes into stable and unstable subbundles. Indeed, this is the case if $N(\theta) = (N^S(\theta)\ N^U(\theta))$ and $\Lambda_N(\theta) = \mathrm{blockdiag}(\Lambda_S(\theta), \Lambda_U(\theta))$, with Λ_S contracting and Λ_U expanding.

Moreover, under Diophantine conditions on ω, if the invariant normal bundle decomposes into n one dimensional subbundles, then the normal dynamics is reduced to a diagonal constant matrix $\Lambda_N = \mathrm{diag}(\lambda_{d+1}, \ldots, \lambda_{d+n})$, with real *eigenvalues* $|\lambda_j| \neq 1$. In such a case, the torus is said to be **reducible**. For the sake of simplicity, we will consider in this paper that the eigenvalues have different moduli. More general cases are considered in [5, 7].

Remark 1 If the bundles are non-orientable, we can consider P defined from $\tilde{\mathbb{T}}^d = (\mathbb{R}/2\mathbb{Z})^d$ instead of $\mathbb{T}^d$ by using a double covering trick [13].

Remark 2 (NHIT regardless the internal dynamics) There is the case when we do not know which is the internal dynamics. Then, we have to proceed by taking the internal dynamics as another unknown of the invariance equation instead of adjusting parameters to fix the frequency. In other words, Eqs. (1) and (4) becomes:

$$F_a(K(\theta)) - K(f(\theta)) = 0, \tag{6}$$

$$\mathrm{D}F_a(K(\theta))N(\theta) - N(f(\theta))\Lambda_N(\theta) = 0, \tag{7}$$

and they have to be solved for K, f, N and Λ_N. For more details on computations of NHIT regardless the internal dynamics see [5, 6].

3 Specification of One Step of a Newton-Like Method

In the following, we explain how to perform one step of a Newton-like method to solve invariance equations (1) and (4) above, in the reducible case. Starting with an approximate parameterization of a NHIT K for a F_a, an approximate invariant normal bundle N and its linearized dynamics Λ_N, the aim of one step of Newton method is to compute their corresponding corrections. To do so, the parameter a is adjusted in order to keep fixed the frequency ω. When it is possible to reduce the system, each step in the Newton-like method becomes very fast [13]. The procedure is repeated until we achieve the desired error-tolerance. We will not consider here the rigorous results on the convergence of the algorithm that hold under suitable *Melnikov conditions*.

Remark 3 (Validation theorem) The theoretical framework of this algorithm is based on KAM techniques that we are not going to detail here. To give a brief idea of it, under normal hyperbolicity and additional non-degeneracy conditions on an approximate invariant torus and on the adjusting parameter a, if the error estimates are small enough (in suitable Banach spaces of real-analytic periodic functions), the theorem ensures that there is a true invariant torus and an adapting parameter nearby. Remarkably, the method of proving the theorem is similar to the algorithm presented here. In [7] we obtain a validation theorem of existence of NHIT with fixed Diophantine frequency ω, based on KAM techniques, that proves the convergence of the procedure.

Since we are dealing with periodic functions to represent the torus K and the adapted frame P, it is natural to represent them in Fourier series. For a periodic function f,

$$f(\theta) = \sum_{k \in \mathbb{Z}^d} f_k e^{2\pi \mathbf{i} k \theta} \tag{8}$$

is its Fourier series. The average of f is $< f >= f_0$.

3.1 Substep 1: Correction of the Torus K and Parameter a

Let $R : \mathbb{T}^d \to \mathbb{R}^m$ and $S^N : \mathbb{T}^d \to \mathbb{R}^{m\times n}$ be the errors in the invariance equation of the torus and in the invariance equation of the normal bundle respectively, that is

$$R(\theta) = F_a(K(\theta)) - K(\theta + \omega), \tag{9}$$

$$S^N(\theta) = \mathrm{D}F_a(K(\theta))N(\theta) - N(\theta + \omega)\Lambda_N. \tag{10}$$

We assume that both R and S^N are "small".

Then, the adapted frame P is approximately invariant, since $S : \mathbb{T}^d \to \mathbb{R}^{m\times m}$ defined by

$$S(\theta) = \mathrm{D}F_a(K(\theta))P(\theta) - P(\theta + \omega)\Lambda \tag{11}$$

is in fact $S(\theta) = \begin{pmatrix}\mathrm{D}R(\theta) & S^N(\theta)\end{pmatrix}$, which is also "small".

We consider the correction of the torus $\bar{K} = K + \Delta K$ of the form $\Delta K(\theta) = P(\theta)\xi(\theta)$, being $\xi : \mathbb{T}^d \to \mathbb{R}^m$ a periodic function. The adjustment of a is given by δ, $\bar{a} = a + \delta$. Then, by substituting new approximations in (1) and using first order Taylor expansion, we obtain

$$-R(\theta) = \frac{\partial F_a}{\partial a}(K(\theta))\delta + P(\theta + \omega)\Lambda\xi(\theta) - P(\theta + \omega)\xi(\theta + \omega) + \mathscr{O}_2, \tag{12}$$

where we apply definitions (9) and (11) above, and $\mathscr{O}_2$ collect the quadratically small terms. Multiplying (12) by $P(\theta + \omega)^{-1}$ and neglecting quadratically small terms, we obtain the cohomological equation

$$-\tilde{R}(\theta) = \Lambda\xi(\theta) - \xi(\theta + \omega) + B(\theta)\delta, \tag{13}$$

where $\tilde{R}(\theta) = P(\theta + \omega)^{-1}R(\theta)$ is the error of the approximate solution in the adapted frame and $B(\theta) = P(\theta + \omega)^{-1}\frac{\partial F_a}{\partial a}(K(\theta))$. Splitting (13) into tangent and normal components, a Newton step reduces our equation to the block diagonal system

$$-\tilde{R}^L(\theta) = \xi^L(\theta) - \xi^L(\theta + \omega) + B^L(\theta)\delta, \tag{14}$$

$$-\tilde{R}^N(\theta) = \Lambda_N\xi^N(\theta) - \xi^N(\theta + \omega) + B^N(\theta)\delta. \tag{15}$$

Tangent component. Let $r(\theta) := -\tilde{R}(\theta) - B(\theta)\delta$. We have to solve the cohomological equation

$$\xi^L(\theta) - \xi^L(\theta + \omega) = r^L(\theta). \tag{16}$$

We chose δ as

$$\delta = - < B^L >^{-1} < \tilde{R}^L >, \tag{17}$$

to ensure r^L has zero average, provided that $< B^L(\theta) >$ invertible. Since ω satisfies Diophantine conditions and r^L has zero average, we can solve (16), and its solution is obtained by solving order by order in terms of Fourier modes:

$$\xi_k^L = \frac{r_k^L}{1 - e^{2\pi \mathbf{i} k\omega}}, \ k \neq 0. \tag{18}$$

Notice that ξ_0^L is free. In particular, we choose $\xi_0^L = 0$.

Normal component. As Λ_N is diagonal, Eq. (15) splits into n equations, corresponding to their n normal components. For each $i = d+1, \ldots, d+n$, we solve the equation term by term in Fourier modes:

$$r^i(\theta) = \lambda_i \xi^i(\theta) - \xi^i(\theta + \omega) \quad \rightarrow \quad \xi_k^i = \frac{r_k^i}{\lambda_i - e^{2\pi \mathbf{i} k\omega}}, \ k \in \mathbb{Z}^d \tag{19}$$

Since, by assumption, $|\lambda_i| \neq 1$, there are no resonances in (19).

Remark 4 The absence of resonances in (19) is known as *first Melnikov condition*. This is also important for dealing with complex eigenvalues of modulus 1, i.e. elliptic eigenvalues.

3.2 Substep 2: Correction of the Floquet Transformations

We redefine the error in the invariance equation of the adapted frame for the new $\bar{K}$ and $\bar{a}$ as $S(\theta) = \mathrm{D}F_{\bar{a}}(\bar{K}(\theta))P(\theta) - P(\theta + \omega)\Lambda$, which is close to the previous $S(\theta)$. We consider the corrections of the normal bundle, $\bar{N} = N + \Delta N$, and its linearized dynamics, $\bar{\Lambda}_N = \Lambda_N + \Delta\Lambda_N$, of the form:

$$\Delta N(\theta) = P(\theta) Q^N(\theta), \qquad \Delta\Lambda_N = \mathrm{diag}(\delta_{d+1}, \ldots, \delta_{d+n}), \tag{20}$$

where $Q^N : \mathbb{T}^d \to \mathbb{R}^{m \times n}$ is a periodic matrix map. Doing similar computations as in the substep 1, we obtain the cohomological equation:

$$-\tilde{S}^N(\theta) = \Lambda Q^N(\theta) - Q^N(\theta + \omega)\Lambda_N - \begin{pmatrix} O \\ \Delta\Lambda_N \end{pmatrix}, \tag{21}$$

where $\tilde{S}^N(\theta) = P(\theta + \omega)^{-1} S^N(\theta)$.

Using the matrix notation $Q^N(\theta) = (Q^{i,j}(\theta))$, $i = 1, \ldots, n+d$, $j = d+1, \ldots, d+n$, Eq. (21) splits into $m \times n$ equations and we solve them in terms of Fourier modes:

$$i \le d,\ i \ne j:\ -\tilde{S}^{i,j}(\theta) = Q^{i,j}(\theta) - Q^{i,j}(\theta+\omega)\lambda_j$$
$$\to Q_k^{i,j} = \frac{\tilde{S}_k^{i,j}}{\lambda_j e^{2\pi \mathbf{i} k\omega} - 1},\ \forall k. \tag{22a}$$

$$i > d,\ i \ne j:\ -\tilde{S}^{i,j}(\theta) = \lambda_i Q^{i,j}(\theta) - Q^{i,j}(\theta+\omega)\lambda_j$$
$$\to Q_k^{i,j} = \frac{\tilde{S}_k^{i,j}}{\lambda_j e^{2\pi \mathbf{i} k\omega} - \lambda_i},\ \forall k. \tag{22b}$$

$$i > d,\ i = j:\ -\tilde{S}^{i,i}(\theta) = \lambda_i Q^{i,i}(\theta) - Q^{i,i}(\theta+\omega)\lambda_i - \delta^i$$
$$\to \begin{array}{l} Q_k^{i,i} = \begin{cases} \frac{\tilde{S}_k^{i,i}}{\lambda_i (e^{2\pi \mathbf{i} k\omega} - 1)}, & k \ne 0, \\ 0, & k = 0, \end{cases} \\ \delta_i = \tilde{S}_0^{i,i}. \end{array} \tag{22c}$$

Note that, since we assume $|\lambda_i| \ne |\lambda_j| \ne 1$, there are no resonances in (22a) and (22b).

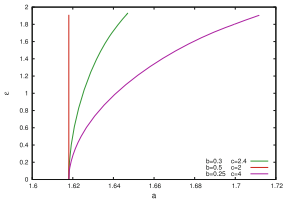
Summarizing, we have obtained new better approximations $\bar{K} = K + P\xi$, $\bar{a} = a + \delta$ and $\bar{N} = N + PQ^N$ and $\bar{\Lambda}_N = \Lambda_N + \Delta\Lambda_N$.

4 Implementation

We consider as a toy example the 3D-*F*attened Arnold Family [2], given by the diffeomorphism $F_{a,\epsilon} : \mathbb{R}/2\pi\mathbb{Z} \times \mathbb{R}^2 \to \mathbb{R}/2\pi\mathbb{Z} \times \mathbb{R}^2$ defined as:

$$F_{a,\epsilon}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x + a + \epsilon(\sin(x) + y + z/2) \\ b(\sin(x) + y) \\ c(\sin(x) + y + z) \end{pmatrix} \tag{23}$$

where $b < 1$, $c > 1$ are fixed parameters, $a \in \mathbb{R}$ is the adjusting parameter and $\epsilon \in \mathbb{R}$ is the perturbation parameter. This system has a constant determinant of the Jacobian $\det(\mathrm{D}F_{a,\epsilon}) = bc$. In these implementations, we will continue curves of saddle NHIT with respect to ϵ with the same fixed frequency $\omega = (\sqrt{5}+1)/2$ up to a critical value, for which the torus seams to be destroyed. Our computations are done with an error-tolerance $||R|| < 10^{-10}$ and using as much as Fourier modes, $\mathrm{N_F}$, we need to get the torus well approximated with this error tolerance (see last column on the table of Fig. 1). We emphasize that reaching such an accuracy with such a hight number of Fourier modes (here $\mathrm{N_F} = 1{,}048{,}576 = 2^{20}$) is much



Example 1: dissipative case (b=0.3, c=2.4)					
a	ϵ	Λ_L	Λ_S	Λ_U	N_F
1.6180339887	0.0000000000	1.00	0.3000000000	2.4000000000	64
1.6226299607	0.8000000000	1.00	0.2937671574	2.4509206762	64
1.6364204820	1.6000000000	1.00	0.2742024426	2.6257971784	128
1.6424191999	1.8000000000	1.00	0.2655688776	2.7111610611	512
1.6470212201	1.9281689453	1.00	0.2584528482	2.7858079528	1048576

Example 2: conservative case (b=0.25, c=4)					
a	ϵ	Λ_L	Λ_S	Λ_U	N_F
1.6180339887	0.0000000000	1.00	0.2500000000	4.0000000000	64
1.6331584739	0.8000000000	1.00	0.2467877231	4.0520654241	64
1.6795149735	1.6000000000	1.00	0.2372308102	4.2153040710	128
1.6994712606	1.8000000000	1.00	0.2336399759	4.2800894669	512
1.7118185512	1.9052001953	1.00	0.2314804290	4.3200196418	1048576

Example 3: conservative-reversible case (b=0.5, c=2)					
a	ϵ	Λ_L	Λ_S	Λ_U	N_F
1.6180339887	0.0000000000	1.00	0.5000000000	2.0000000000	64
1.6180339887	0.8000000000	1.00	0.4899204837	2.0411475603	64
1.6180339887	1.6000000000	1.00	0.4527757852	2.2085986764	256
1.6180339887	1.8000000000	1.00	0.4317848576	2.3159682014	1024
1.6180339887	1.9102784824	1.00	0.4144623759	2.4127642417	1048576

Fig. 1 *Left*: Curves of quasi-periodic NHIT in the parameter plane (a, ϵ). *Right*: Λ values, which corresponds to Lyapunov multipliers, for the three different implementations

beyond the limits of large matrix methods based on full discretization of invariance equations, which already suffer with, say, $N_F = 1{,}024 = 2^{10}$.

We examine three examples, to be described below (see Fig. 1). For a better understanding of the breakdown of the invariant torus, we consider as observables the Lyapunov multipliers (the absolute values of the eigenvalues of the reduced matrix Λ) and the minimum angles between bundles. These observables measure the quality of normal hyperbolicity properties. The Lyapunov multipliers $\Lambda_S = |\lambda_2| < \Lambda_L = 1 < \Lambda_U = |\lambda_3|$ are shown in the table of Fig. 1. The maximal Lyapunov multiplier Λ_U and the minimum angles between bundles, $\widehat{L - N^S}$, $\widehat{L - N^U}$ and $\widehat{N^S - N^U}$, appear in Fig. 2 along the continuation with respect to ϵ. In Fig. 3 we show the last computed torus and the angles between their invariant bundles as a function of θ. In previous works, different breakdowns due to collision of bundles have been observed [4, 10, 11, 13].

Example 1 For parameters $b = 0.3$ and $c = 2.4$, the system is *dissipative* ($bc = 0.72$). During the continuation, the tangent and stable bundles approach till finally collide and the torus is destroyed. Near the breakdown, there is a lineal decay to zero of the angle $\widehat{L - N^S}$. Despite that, the Lyapunov multipliers Λ_L and Λ_S are moving away from each other. Notice that N^U remains far from L and N^S.

Example 2 For parameters $b = 0.25$ and $c = 4$, the system is *conservative* ($bc = 1$). As in Example 1, the breakdown of the torus is due to a bundle collision of the tangent and stable bundle, while Λ_L and Λ_S do not collide even though their product value is constant to 1, $\mathrm{D}F_{a,\epsilon} = 1 = \Lambda_L \Lambda_S \Lambda_U$.

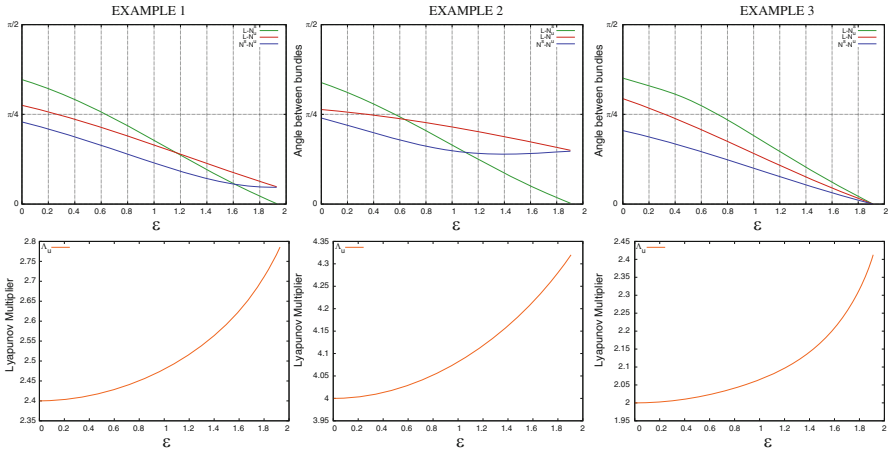


Fig. 2 Results for all ϵ values of the continuation process. *Top*: Minimum angles between bundles. *Bottom*: maximal Lyapunov multiplier Λ_U, where Λ_S is just $\Lambda_S = \frac{|bc|}{\Lambda_U}$

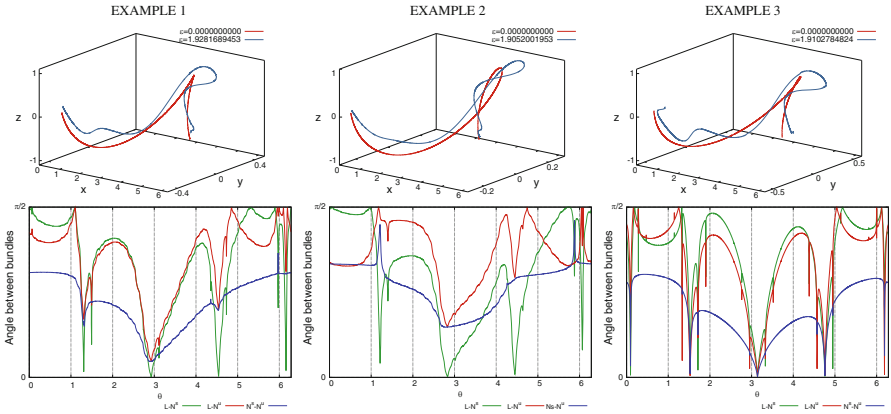


Fig. 3 Results for the last torus we can compute. *Top*: invariant torus. *Bottom*: angles between bundles

Example 3 For parameters $b = 0.5$ and $c = 2$, the system is *conservative* ($bc = 1$) and *reversible*. That is, $F_{a,\epsilon} = I_1 I_0$ with involutions I_0 and I_1 given by:

$$I_1 \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -x & +2\pi \\ \sin(x) & + y + \frac{3}{4}z \\ & - z \end{pmatrix}, \; I_0 \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -x + \epsilon y + \frac{\epsilon}{4}z + a + 2\pi \\ \frac{y}{2} + \frac{3}{8}z \\ 2y - \frac{z}{2} \end{pmatrix}.$$

Moreover, by the symmetries of the system, the breakdown of the torus is due to a triple collision, all bundles collide together. All values $\widehat{L - N^S}$, $\widehat{L - N^U}$, $\widehat{N^S - N^U}$

tend to zero, linearly, as we approach the breakdown. Notice also that, in this example, the adjusting parameter a is fixed to the frequency, $a = \omega$, along the continuation.

Acknowledgements M.C. and A.H. have been funded by the Spanish grants MTM2009-09723 and MTM2012-32541. M.C. has also been funded by the FPI grant BES-2010-039663 and A.H. by the Catalan grant 2009-SGR-67.

References

1. Broer, H.W., Huitema, G.B., Sevryuk, M.B.: Quasi-periodic Motions in Families of Dynamical Systems. Order Amidst Chaos. Lecture Notes in Mathematics, vol. 1645. Springer, Berlin (1996)
2. Broer, H., Osinga, H., Vegter, G.: Algorithms for computing normally hyperbolic invariant manifolds. Z. Angew. Math. Phys. **48**(3), 480–524 (1997)
3. Calleja, R., Celletti, A., de la Llave, R.: KAM theory for conformally symplectic systems: efficient algorithms and their validation. J. Differ. Equ. **255**(5), 978–1049 (2013)
4. Calleja, R., Figueras, J.-L.: Collision of invariant bundles of quasi-periodic attractors in the dissipative standard map. Chaos **22**(3), 033114 (2012)
5. Canadell, M.: Computation of normally hyperbolic invariant manifolds. PhD thesis, Universitat de Barcelona (July 2014)
6. Canadell, M., Haro, A.: A Newton-like method for computing normally hyperbolic invariant tori. Chapter 5 of the parameterization method for invariant manifolds: from rigorous results to effective computations (2014, in progress)
7. Canadell, M., Haro, A.: A KAM-like theorem for Quasi-Periodic Normally Hyperbolic Invariant Tori (2014, in progress)
8. Castellà, E., Jorba, A.: On the vertical families of two-dimensional tori near the triangular points of the bicircular problem. Celest. Mech. Dyn. Astron. **76**(1), 35–54 (2000)
9. Fenichel, N.: Persistence and smoothness of invariant manifolds for flows. Indiana Univ. Math. J. **21**, 193–226 (1971)
10. Figueras, J.L.: Fiberwise hyperbolic invariant Tori in quasiperiodically skew product systems. PhD. thesis, Universitat de Barcelona (2011)
11. Haro, A., de la Llave, R.: Manifolds on the verge of a hyperbolicity breakdown. Chaos **16**(1), 013120 (2006)
12. Haro, A., de la Llave, R.: A parameterization method for the computation of invariant tori and their whiskers in quasi-periodic maps: numerical algorithms. Discret. Contin. Dyn-B. **6**, 1261–1300 (2006)
13. Haro, A., de la Llave, R.: A parameterization method for the computation of invariant tori and their whiskers in quasi-periodic maps: explorations and mechanisms for the breakdown of hyperbolicity. SIAM J. Appl. Dyn. Syst. **6**(1), 142–207 (2007)
14. Hirsch, M.W., Pugh, C.C., Shub, M.: Invariant Manifolds. Lecture Notes in Mathematics, vol. 583. Springer, Berlin (1977)
15. Osinga, H., Schilder, F., Vogt, W.: Continuation of quasi-periodic invariant tori. SIAM J. Appl. Dyn. Syst. **4**(3), 459–488 (2005)
16. Peckham, B.B., Schilder, F.: Computing Arnol'd tongue scenarios. J. Comput. Phys. **220**(2), 932–951 (2007)

Existence of Homoclinic and Heteroclinic Connections in Continuous Piecewise Linear Systems

Victoriano Carmona, Fernando Fernández-Sánchez, and Elisabeth García-Medina

Abstract In the present work, the existence of global connections in a continuous piecewise linear system is analytically proven. Concretely, by using a common technique we prove the existence of a pair of homoclinic connections and a reversible T-point heteroclinic cycle. The main ideas of this proof can be extended to other piecewise linear systems.

1 Introduction and Statements of Main Results

The proof of the existence of a global connection in differential systems is generally a difficult task, even in the case of continuous piecewise linear systems. Regarding global connections in $\mathbb{R}^3$, in [3] the authors prove the existence of homoclinic connections to saddle focus equilibria in the three-parameter unfolding of a nilpotent singularity of codimension three. Some recents works [6, 7] have been devoted to a different approach, which consists on the derivation of computer-assisted proofs for the existence of global connections.

In [1, 2] the authors studied some global connections of system

$$\begin{cases} \dot{x} = y, \\ \dot{y} = z, \\ \dot{z} = 1 - y - \lambda(1 + \lambda^2)|x|, \end{cases} \tag{1}$$

where the parameter λ is strictly positive. In the present work, we give a common proof for the existence of a pair of homoclinic connections and a reversible T-point heteroclinic cycle.

V. Carmona (✉) • F. Fernández-Sánchez • E. García-Medina
Departamento Matemática Aplicada II, Universidad de Sevilla, E.T.S. de Ingeniería, Camino de los Descubrimientos s/n, 41092 Sevilla, Spain
e-mail: vcarmona@us.es; fefesan@us.es; egarme@us.es

F. Casas, V. Martínez (eds.), *Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 4, DOI 10.1007/978-3-319-06953-1_10

System (1) is volume-preserving, time-reversible with respect to the involution $\mathbf{R}(x, y, z) = (-x, y, -z)$ and it can be written in matrix form as

$$\dot{\mathbf{x}} = \begin{cases} A^{-}\mathbf{x} + \mathbf{e}_3 & \text{if } x \leq 0, \\ A^{+}\mathbf{x} + \mathbf{e}_3 & \text{if } x \geq 0, \end{cases} \quad \text{with} \quad A^{\pm} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \mp\lambda(1+\lambda^2) & -1 & 0 \end{pmatrix},$$

$\mathbf{x} = (x, y, z)^T$ and $\mathbf{e}_3 = (0, 0, 1)^T$. It is formed by two linear systems separated by the plane $\{x = 0\}$, called the separation plane, and it can be considered as a piecewise linear version of the Michelson system [5].

The global connections of this system can be classified attending to the number of intersections of the global connection and the separation plane. A homoclinic connection of system (1) is direct if it intersects the separation plane at exactly two points. On the other hand, a reversible T-point heteroclinic cycle is direct if the heteroclinic connection corresponding to the one-dimensional invariant manifolds has exactly three intersections with the separation plane while the heteroclinic connection corresponding to the two-dimensional invariant manifolds has only one intersection.

The main aim of the present work is to prove analytically the existence of these global connections in system (1). The following result, which is the core of the paper, establishes their existence.

Theorem 1 *There exist two real values* $\lambda_h, \lambda_T \in \left(1/2, \sqrt{3}\right)$ *such that for* $\lambda = \lambda_h$ *the piecewise linear system (1) has two direct homoclinic connections (which are symmetric with respect to the involution* $\mathbf{R}$*) and for* $\lambda = \lambda_T$ *the system has a direct T-point heteroclinic cycle.*

Some numerical computations allow to obtain $\lambda_h \simeq 0.66076$ and $\lambda_T \simeq 0.65154$. In Fig. 1, the projections onto the (x, y)-plane of the global connections given by Theorem 1 are shown.

The rest of the paper is organized as follows. In Sect. 2, some properties of the piecewise linear system (1) are presented. After that, a set of conditions for the existence of a direct homoclinic connection and a direct reversible T-point

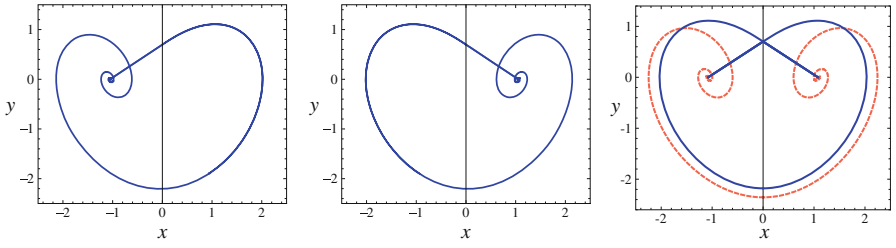


Fig. 1 Projections onto the (x, y)-plane of the two direct homoclinic connections and the direct T-point heteroclinic cycle

heteroclinic cycle is introduced. In Sect. 3, we show some results that allow us to prove the existence of both global connections. Finally, Sect. 4 is devoted to prove Theorem 1.

2 Geometric Elements and Conditions for the Existence of Global Connections

In this section, some properties of system (1) are shown. These properties has been described previously in [1, 2] but, for the sake of completeness, it is convenient to remind them. After that, a set of conditions for the existence of a direct homoclinic connection and a direct reversible T-point heteroclinic cycle is introduced.

To analyze the dynamical behavior of system (1) we will use some Poincaré half-maps associated to this system. By means of the flow of the system $\dot{\mathbf{x}} = A^{-}\mathbf{x} + \mathbf{e}_3$ with $x \leq 0$, some points $\mathbf{p}_0$, belonging to the separation plane, can be transformed into points $\mathbf{q}_0$ of this plane, so a Poincaré half-map Π_- in the half-space $\{x \leq 0\}$ can be defined as $\mathbf{q}_0 = \Pi_-(\mathbf{p}_0)$. Analogously, we can define a Poincaré half-map Π_+ in the half-space $\{x \geq 0\}$, so that a Poincaré map for system (1) is defined as $\Pi = \Pi_+ \circ \Pi_-$.

Since $\lambda > 0$, the eigenvalues of A^- are $\lambda, \alpha \pm \mathrm{i}\beta$, with

$$\alpha = -\lambda/2 \quad \text{and} \quad \beta = \sqrt{4 + 3\lambda^2}/2. \tag{2}$$

By the reversibility with respect to $\mathbf{R}$, the eigenvalues of A^+ are $-\lambda$ and $-\alpha \pm \mathrm{i}\beta$. Therefore, there exist two saddle-focus equilibria $\mathbf{p}^{\pm} = (\pm 1/(\lambda + \lambda^3), 0, 0)^T$.

The unstable invariant manifold $W^u(\mathbf{p}^-)$ contains the straight half-line

$$L^- = \left\{\mathbf{p}^- - \mu\,(1, \lambda, \lambda^2)^T : \; -1/(\lambda^3 + \lambda) \leq \mu < +\infty\right\},$$

that intersects the separation plane at $\mathbf{m}^- = \left(0, 1/(\lambda^2 + 1), \lambda/(\lambda^2 + 1)\right)^T$. The stable invariant manifold $W^s(\mathbf{p}^-)$ is locally contained in the half-plane $P^- = \left\{\lambda(\lambda^2 + 1)x + \lambda^2 y + \lambda z = -1, \; x \leq 0\right\}$, that intersects the separation plane along the straight-line $D^- = \{\lambda^2 y + \lambda z = -1, \; x = 0\}$. Note that not every point in D^- belongs to $W^s(\mathbf{p}^-)$. The straight-line D^- and the z-axis intersect at the point $\mathbf{q}^- = (0, 0, -1/\lambda)^T$, where the orbit is tangent to the separation plane. Hence, the segment $S^- \subset D^-$ with end points $\mathbf{q}^-$ and $\Pi_-^{-1}(\mathbf{q}^-)$ is contained in $W^s(\mathbf{p}^-)$.

By the reversibility, the geometric elements in the half-space $\{x \geq 0\}$ can be obtained. Thus, the invariant manifold $W^s(\mathbf{p}^+)$ contains the straight half-line $L^+ = \left\{\mathbf{p}^+ + \mu\,(1, -\lambda, \lambda^2)^T : \; -1/(\lambda^3 + \lambda) \leq \mu < +\infty\right\}$, that intersects the separation plane at $\mathbf{m}^+ = \left(0, 1/(\lambda^2 + 1), -\lambda/(\lambda^2 + 1)\right)^T$. On the other hand, the invariant manifold $W^u(\mathbf{p}^+)$ is locally contained in $P^+ = \left\{\lambda(\lambda^2 + 1)x - \lambda^2 y + \lambda z = 1, \; x \geq 0\right\}$, that intersects the separation plane along the straight-line $D^+ = \{-\lambda^2 y + \lambda z = 1, \; x = 0\}$. This straight-line and the z-axis intersect at $\mathbf{q}^+ = (0, 0, 1/\lambda)^T$. The straight-lines D^+ and D^- intersect

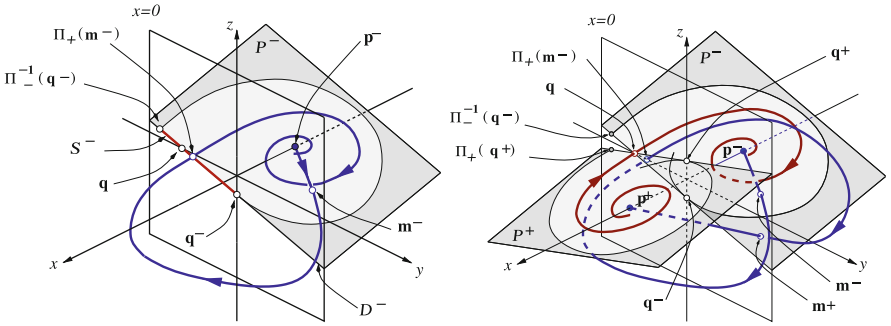


Fig. 2 Schematic picture of a direct homoclinic connection to the equilibrium $\mathbf{p}^-$ and a direct reversible T-point heteroclinic cycle

at $\mathbf{q} = \left(0, -1/\lambda^2, 0\right)$. In Fig. 2, a schematic picture of these global connections together with the geometric elements that have been described above is shown.

At this point, we are able to introduce a set of conditions that characterize two global connections: a direct homoclinic connection and a direct reversible T-point heteroclinic cycle.

For every point $\mathbf{p}_0 = (x_0, y_0, z_0)^T$, we denote by $\mathbf{x}^-(t; \lambda, \mathbf{p}_0)$ (resp. $\mathbf{x}^+(t; \lambda, \mathbf{p}_0)$) the solution of linear system $\dot{\mathbf{x}} = A^-\mathbf{x} + \mathbf{e}_3$ (resp. $\dot{\mathbf{x}} = A^+\mathbf{x} + \mathbf{e}_3$) with parameter λ and initial condition $\mathbf{x}(0; \lambda) = \mathbf{p}_0$.

A direct homoclinic connection to $\mathbf{p}^-$ has to intersect the separation plane at point $\mathbf{m}^-$. On the other hand, this connection exists if the condition $\Pi_+(\mathbf{m}^-) \in S^-$ holds. Therefore, system (1) has a direct homoclinic connection to $\mathbf{p}^-$ if and only if there exist two real values $t_h, \lambda_h > 0$ such that

(H1) $\mathbf{x}^+(t_h; \lambda_h, \mathbf{m}^-) \in D^-$,
(H2) $x^+(t; \lambda_h, \mathbf{m}^-) > 0$ for every $t \in (0, t_h)$,
(H3) $\mathbf{x}^+(t_h; \lambda_h, \mathbf{m}^-) \in S^-$.

By integrating system (1) taking into account the condition (H2), it is obvious that the condition (H1) holds if and only if there exist two real values $t_h, \lambda_h > 0$ such that the pair $(t, \lambda) = (t_h, \lambda_h)$ is a solution of the system

$$\begin{cases} E_1(t, \lambda) = 0, \\ E_2(t, \lambda) = 0, \end{cases} \tag{3}$$

where

$$E_1(t, \lambda) = 2\lambda^2 e^{\frac{3t\lambda}{2}} (2\beta \cos(\beta t) - 3\lambda \sin(\beta t)) + 2\beta(\lambda^2 - (3\lambda^2 + 1)e^{t\lambda} + 1), \tag{4}$$

$$E_2(t, \lambda) = 2\lambda^2 e^{\frac{t\lambda}{2}} \left(2\beta \cos(\beta t) + \lambda \sin(\beta t)\right) + 2\beta(\lambda^2 + 1) \tag{5}$$

and β is given in (2).

Let us introduce a set of conditions for the existence of a direct reversible T-point heteroclinic cycle. The heteroclinic connection corresponding to the one-dimensional manifolds intersects necessarily the separation plane at $\mathbf{m}^-$ and $\mathbf{m}^+$. This connection is direct if the relationship $\Pi_+(\mathbf{m}^-) = \Pi_-^{-1}(\mathbf{m}^+)$ holds. Due to the reversibility, this fact occurs if $\Pi_+(\mathbf{m}^-) \in \{x = 0, z = 0\}$. Therefore, system (1) has a one-dimensional heteroclinic connection if and only if there exist two real values $t_T, \lambda_T > 0$ such that the following conditions hold:

(T1) $\mathbf{x}^+(t_T; \lambda_T, \mathbf{m}^-) \in \{x = 0, z = 0\}$,
(T2) $x^+(t; \lambda_T, \mathbf{m}^-) > 0$ for every $t \in (0, t_T)$.

By integrating system (1) taking into account the condition (T2), it is obvious that the condition (T1) is satisfied if and only if there exist two real values $t_T, \lambda_T > 0$ such that the pair $(t, \lambda) = (t_T, \lambda_T)$ is a solution of the system

$$\begin{cases} E_1(t, \lambda) = 0, \\ F_1(t, \lambda) = 0, \end{cases} \tag{6}$$

where the expression of E_1 is given in (4),

$$F_1(t, \lambda) = 2e^{\frac{3t\lambda}{2}} \left(2(2\lambda^2 + 1)\beta \cos(\beta t) - \lambda \sin(\beta t)\right) - 2(\lambda^2 + 1)\beta$$

and β is given in (2).

The heteroclinic connection corresponding to the two-dimensional manifolds must intersect the separation plane at the point $\mathbf{q}$. Therefore, a condition for the existence of a direct two-dimensional heteroclinic connection is

(T3) $\mathbf{q} \in S^-$.

Note that the condition (T3) involves the condition (H3). From Proposition 3.3 in [2], it follows that there exists a unique real value $\lambda_0 \in (0, 1/2)$ such that if $\lambda \geq \lambda_0$, then the condition (T3) holds. Some numerical computations allow to obtain $\lambda_0 \simeq 0.41527$.

Observe that the set of conditions for the existence of a pair of direct homoclinic connections and of a direct reversible T-point heteroclinic cycle are similar. Concretely, the conditions (H1) and (T1) lead to systems (3) and (6), that have a similar structure. A generic system that includes both systems is

$$A \cdot \mathbf{X} = C, \tag{7}$$

where $\mathbf{X} = (\cos(\beta t), \sin(\beta t))^T$, $C = (c_1(t, \lambda), c_2(t, \lambda))^T$ and

$$A = \begin{pmatrix} a(t, \lambda) & b(t, \lambda) \\ c(t, \lambda) & d(t, \lambda) \end{pmatrix}.$$

3 Preliminary Results

This section is devoted to analyzing the existence of solutions of system (7) and a result that allows us to prove the conditions (H2) and (T2).

In the following lemma, whose proof is direct, system (7) is reduced to other equivalent system with two equations and one inequality, in which it is possible to apply the Poincaré-Miranda Theorem [4]. One of these equations is a linear combination of the equations of system (7) and the other is obtained by considering $X = \cos(\beta t)$ and $Y = \sin(\beta t)$ and taking into account the trigonometrical identity $X^2 + Y^2 - 1 = 0$.

Lemma 1 *Let A_i, for $i = 1, 2$, be the matrix obtained by replacing the ith-column of A by the column matrix C. If $\det(A_2) \cdot \det(A) < 0$ for $t, \lambda > 0$, then system (7) is equivalent to the system*

$$\begin{cases} E(t,\lambda) = 0, \\ p(t,\lambda) = 0, \\ \sin(\beta t) < 0 \end{cases} \tag{8}$$

where $E(t,\lambda) = m(t,\lambda)\cos(\beta t) + n(t,\lambda)\sin(\beta t)$ and

$$p(t,\lambda) = \left((m(t,\lambda))^2 + (n(t,\lambda))^2\right)/\left(\det(A)\right)^2 - 1,$$

with $m(t,\lambda) = \det(A_2)$ and $n(t,\lambda) = -\det(A_1)$.

In the following proposition, we give conditions for the existence of solution of system (8).

Proposition 1 *For $k \in \mathbb{N}$ and $\lambda > 0$, let us consider the interval $I_k = \left[l_k(\lambda), u_k(\lambda)\right]$, where $l_k(\lambda) = 2(2k-1)\pi/\sqrt{3\lambda^2+4}$, $u_k(\lambda) = 4k\pi/\sqrt{3\lambda^2+4}$. If there exist two real values $\lambda_1, \lambda_2 > 0$ such that $m(l_k(\lambda),\lambda) \cdot m(u_k(\lambda),\lambda) > 0$ for all $\lambda \in [\lambda_1, \lambda_2]$ and $p(t,\lambda_1) \cdot p(t,\lambda_2) < 0$ for every $t \in I_k$, then system (7) has a solution in the open set $\Omega_k = \left\{(t,\lambda) \in \mathbb{R}^2 : l_k(\lambda) < t < u_k(\lambda), \lambda_1 < \lambda < \lambda_2\right\}$.*

Proof From Lemma 1, for $t, \lambda > 0$ systems (7) and (8) are equivalent if and only if the inequality $\det(A_2) \cdot \det(A) < 0$ holds. Note that for every $(t,\lambda) \in \Omega_k$ this inequality is satisfied and so the inequality $\sin(\beta t) < 0$ holds.

Now, we are going to see that system (8) has solution in Ω_k. The change of variables $\mu = \lambda^2$, $\tau = \sqrt{4+3\lambda^2}\, t/2$ transforms this system into the system

$$\begin{cases} \tilde{E}(\tau,\mu) := E\left(2\tau/\sqrt{4+3\mu}, \sqrt{\mu}\right) = 0, \\ \tilde{p}(\tau,\mu) := p\left(2\tau/\sqrt{4+3\mu}, \sqrt{\mu}\right) = 0, \\ \sin(\tau) < 0, \end{cases} \tag{9}$$

and the open set Ω_k into $\tilde{\Omega}_k = ((2k-1)\pi, 2k\pi) \times \left(\lambda_1^2, \lambda_2^2\right)$.

From hypothesis of this proposition, it is obvious that the function $\tilde{E}$ takes different signs at the vertical sides of the boundary of $\tilde{\Omega}_k$. On the other hand, the function $\tilde{p}$ takes different signs at the horizontal sides of the boundary of $\tilde{\Omega}_k$. The conclusion of this proposition is followed by applying the Poincaré-Miranda Theorem. □

Now, we present a result that will allow us to verify that the conditions (H2) and (T2) are fulfilled.

Proposition 2 *Let $f(t)$ be a function such that its first derivative is given by*

$$f'(t) = c_1 e^{-t\lambda} + \big(c_2 \cos(\beta t) + c_3 \sin(\beta t)\big) e^{\frac{t\lambda}{2}},$$

where $\lambda > 0$ and $c_1 \cdot c_2 > 0$. If there exists $t_* \in (0, 2\pi/\beta)$ such that $f(0) = f(t_*) = 0$, $f'(0) > 0$ and $f'(t_*) < 0$, then $f(t) > 0$ for every $t \in (0, t_*)$.

Proof In order to prove this result, it is enough to show that $f(t) \neq 0$ in $(0, t_*)$, since the conditions $f(0) = 0$ and $f'(0) > 0$ holds.

Let us assume that there exists a real value $\hat{t} \in (0, t_*)$ such that $f(\hat{t}) = 0$. Then, $f'(t)$ must vanish in at least three values in $(0, t_*)$. The change of variables $\tau = \beta t$ transforms the equation $f'(t) = 0$ in $h(\tau) = -1$, with

$$h(\tau) = \big(c_2 \cos(\tau) + c_3 \sin(\tau)\big) e^{\frac{3\tau\lambda}{2\beta}} {c_1}^{-1},$$

which must vanish in $(0, 2\pi)$ at least in three values. Since $h(0) = c_2 c_1^{-1} > 0$, equation $h(\tau) = 0$ must have at least three solutions in $(0, 2\pi)$, what is not possible and the proof is concluded. □

4 Existence of a Direct Homoclinic Connection and a Direct Reversible T-Point Heteroclinic Cycle

In this section, we prove the main theorem of this work by using the results given in the previous section.

Proof of Theorem 1 First, from Lemma 1 and Proposition 1, we prove that systems (3) and (6) has at least a solution and so the conditions (H1) and (T1) are fulfilled. After that, by using Proposition 2, we check that the conditions (H2) and (T2) are satisfied. As has been mentioned above, if there exists a unique real value $\lambda_0 \in (0, 1/2)$ such that $\lambda \geq \lambda_0$ then the condition (T3) follows from Proposition 3.3 in [2]. Note that this interval will contain the parameter values for which the conditions (H1)–(H2) and (T1)–(T2) are satisfied.

Let us prove that systems (3) and (6) have a solution in the open set

$$\Omega = \left\{(t, \lambda) \in \mathbb{R}^2 : 2\pi/\sqrt{3\lambda^2 + 4} < t < 4\pi/\sqrt{3\lambda^2 + 4}, 1/2 < \lambda < \sqrt{3}\right\}.$$

Both systems can be written as $A \cdot \mathbf{X} = C$, with

$$A = \begin{pmatrix} 4\lambda^2\beta e^{\frac{3t\lambda}{2}} & -6\lambda^3 e^{\frac{3t\lambda}{2}} \\ c(t,\lambda) & d(t,\lambda) \end{pmatrix} \text{ and } C = \begin{pmatrix} -2\beta\left(\lambda^2 - (3\lambda^2+1)e^{t\lambda} + 1\right) \\ -2\beta(\lambda^2+1) \end{pmatrix}, \tag{10}$$

where $c(t,\lambda) = c_h(t,\lambda) := 4\lambda^2\beta e^{\frac{t\lambda}{2}}$ and $d(t,\lambda) = d_h(t,\lambda) := 2\lambda^3 e^{\frac{t\lambda}{2}}$ in system (3) and $c(t,\lambda) = -(2\lambda^2+1)\,\lambda^{-2}c_h(t,\lambda)e^{t\lambda}$ and $d(t,\lambda) = \lambda^{-2}d_h(t,\lambda)$ in system (6). From Lemma 1, system (10) is equivalent to system (8) for every $t, \lambda > 0$.

In both system the functions $m(t,\lambda)$, defined in Lemma 1, can be written as

$$m(t,\lambda) = -4w_1(\lambda^2)\big((\lambda^2+1)(k_1e^{t\lambda}-1) + k_1\lambda^2 e^{t\lambda}\big)e^{k_2 t\lambda}$$

where the two real values k_1, k_2 and the function w_1 change depending on the system considered and it holds that $k_1, k_2 > 0$ and w_1 is strictly positive at $(0, +\infty)$. For every $\lambda > 0$, it is easy to see that $m(t,\lambda) < 0$ and so the inequality $m(\pi/\beta,\lambda) \cdot m(2\pi/\beta,\lambda) > 0$ is fulfilled.

On other hand, in both systems the sign of function $p(t,\lambda)$, defined in Lemma 1, coincides with the sign of the function

$$z(t,\lambda) = -4^{k_1}\lambda^6 e^{3t\lambda} + k_3(\lambda^2+1)^2(3\lambda^2+1)e^{2t\lambda} + (\lambda^2+1)w_2(\lambda^4)e^{t\lambda} + (\lambda^2+1)^3$$

where $k_3 > 0$ and the function w_2 changes depending on the system considered.

The change of variables $\mu = \lambda^2$, $s = \exp\left(\sqrt{\mu}t\right)$ transforms the function z into

$$\tilde{z}(s,\mu) = -4^{k_1}\mu^3 s^3 + k_3(\mu+1)^2(3\mu+1)s^2 + (\mu+1)^2 w_2(\mu^2)s + (\mu+1)^3,$$

defined for $s \geq 1$ and $\mu > 0$. Since the derivative of $\tilde{z}(s,3)$ is negative in $\mathbb{R}$ and $\tilde{z}(1,3) < 0$, we get $\tilde{z}(s,3) < 0$ for $s \geq 1$. On the other hand, we are going to check that $\tilde{z}(s,1/4) > 0$ for certain values of the variable s. It is easy to see that the derivative of $\tilde{z}(s,1/4)$ is positive in $[1,27]$. Taking into account that $\tilde{z}(1,1/4) > 0$ it follows that $\tilde{z}(s,1/4) > 0$ for every $s \in [1,27]$. Thus, we conclude that function $z(t,\sqrt{3}) < 0$ for every $t > 0$ and that $z(t,1/2) > 0$ for every $t \in I_1 = \left[4\pi/\sqrt{19}, 8\pi/\sqrt{19}\right]$, since this interval is contained in $[1,27]$. From Proposition 1, systems (3) and (6) have a solution in the open set Ω.

Now, we prove that the conditions (H2) and (T2) are satisfied. Let us see that if $(t_*,\lambda_*) \in \Omega$ is a solution of system (10), then $x(t;\lambda_*,\mathbf{m}^-) > 0$ for every $t \in (0,t_*)$. In order to do it, we check that the function $f(t) = x(t;\lambda_*,\mathbf{m}^-)$ satisfies the hypothesis of Proposition 2. According to the equations of system (1), the equality $f'(t) = y(t;\lambda_*,\mathbf{m}^-)$ holds, where

$$y(t;\lambda_*,\mathbf{m}^-) = c_1e^{-t\lambda_*} + (c_2\cos(\beta_* t) + c_3\sin(\beta_* t))\,e^{\frac{t\lambda_*}{2}}, \tag{11}$$

with

$$\beta_* = \frac{\sqrt{4+3\lambda_*^2}}{2}, \quad c_1 = \frac{1}{3\lambda_*^2+1}, \quad c_2 = \frac{2\lambda_*^2 c_1}{\lambda_*^2+1}, \quad c_3 = \frac{\left(3\lambda_*^2+2\right)c_2}{2\lambda_*\beta_*}.$$

On the one hand, $f(0) = 0$ and $f'(0) = c_1 + c_2 > 0$, since $c_1, c_2 > 0$. On the other hand, since (t_*, λ_*) is a solution of system (10), we can obtain the expressions of the trigonometric functions in terms of t_* and λ_*. By substituting these expressions in (11), we obtain that $f'(t_*) = (e^{-t_*\lambda_*} - k_1)/(k_1{\lambda_*}^2) < 0$, with $k_1 > 0$. By applying Lemma 2 we conclude that $f(t) > 0$ for every $t \in (0, t_*)$. Therefore, the conditions (H2) and (T2) are fulfilled and the proof is concluded. □

Acknowledgements This work has been partially supported by the *Ministerio de Economía y Competitividad*, *Plan Nacional I+D+I* cofinanced with FEDER funds, in the frame of the projects MTM2009-07849, MTM2010-20907-C02-01 and MTM2012-31821 and by the *Consejería de Innovación y Ciencia de la Junta de Andalucía* (TIC-0130, P08-FQM-03770, P12-FQM-1658).

References

1. Carmona, V., Fernández-Sánchez, F., García-Medina, E., Teruel, A.E.: Existence of homoclinic connections in continuous piecewise linear systems. Chaos **20**(1), 013124 (2010)
2. Carmona, V., Fernández-Sánchez, F., Teruel, A.E.: Existence of a reversible T-point heteroclinic cycle in a piecewise linear version of the Michelson system. SIAM J. Appl. Dyn. Syst. **7**, 1032–1048 (2008)
3. Ibáñez, S., Rodríguez, J.A.: Shil'nikov configurations in any generic unfolding of the nilpotent singularity of codimension three on R^3. J. Differ. Equ. **208**, 147–175 (2005)
4. Kulpa, W.: The Poincaré–Miranda theorem. Am. Math. Mon. **6**, 545–550 (1997)
5. Michelson, D.: Steady solutions of the Kuramoto–Sivashinsky equation. Physica D **19**, 89–111 (1986)
6. Wilczak, D.: Symmetric heteroclinic connections in the Michelson system: a computer assisted proof. SIAM J. Appl. Dyn. Syst. **4**(3), 489–514 (2005)
7. Wilczak, D.: The existence of Shilnikov homoclinic orbits in the Michelson system: a computer assisted proof. Found. Comput. Math. **6**(4), 495–535 (2006)

Study of Errors in the Integration of the Two Body Problem Using Generalized Sundman's Anomalies

José Antonio López Ortí, Francisco José Marco Castillo, and María José Martínez Usó

Abstract As is well known, the numerical integration of the two body problem with constant step presents problems depending on the type of coordinates chosen. It is usual that errors in Runge–Lenz's vector cause an artificial and secular precession of the periaster although the form remains symplectic, theoretically, even when using symplectic methods. Provided that it is impossible to preserve the exact form and all the constants of the problem using a numerical method, a possible option is to make a change in the variable of integration, enabling the errors in the position of the periaster and in the speed in the apoaster to be minimized for any eccentricity value between 0 and 1.

The present work considers this casuistry. We provide the errors in norm infinite, of different quantities such as the Energy, the module of the Angular Moment vector and the components of Runge–Lenz's vector, for a large enough number of orbital revolutions.

1 Introduction

One of the principal problems present in spatial mechanics is the integration of the equations of motion of an artificial satellite in orbit around the Earth. This motion can be approached in a geocentric system of coordinates by means of the equations [2]

$$\frac{d^2\mathbf{r}}{dx^2} = -\frac{Gm}{r^3}\mathbf{r} - \nabla U + \mathbf{F} \tag{1}$$

J.A. López Ortí (✉) • F.J. Marco Castillo
Departamento de Matemáticas, Universidad Jaume I de Castellón, Av Sos Baynat s/n, 12071 Castellón, Spain
e-mail: lopez@mat.uji.es; marco@mat.uji.es

M.J. Martínez Usó
Departamento de Matemática Aplicada, Universidad Politécnica de Valencia, Camino Vera s/n, Valencia, Spain
e-mail: mjmartin@mat.upv.es

F. Casas, V. Martínez (eds.), *Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 4, DOI 10.1007/978-3-319-06953-1_11

where $\mathbf{r}$ is the vector of geocentric position of the satellite, G is the constant of universal gravitation, m is the terrestrial mass, U is the generating potential of the conservative perturbing forces, such as the luni-solar and other planets attraction as well as the forces due to the not sphericity of the Earth and $\mathbf{F}$ represents the non conservative perturbing forces such as the friction with the higher layers of the atmosphere, etc.

The integration of the previous problem can be carried out by means of analytical techniques from Gauss's planetary equations [8] or using numerical techniques, with the choice of an appropriate numerical method together with a convenient step. The perturbing forces acting on a satellite are usually small, so a common procedure in the construction of integrators adapted for this problem is the development of efficient integrators for the two body problem (this is the not disturbed problem) and then using them for the resolution of the general problem. The study here presented is focused on the elliptic motion. In this case, the two major problems appear when the eccentricity is high. First, the temporary distribution of points on the orbit is very unequal depending on the region. Second, the orbit has zones with very different curvatures. To settle these problems and obtain an acceptable precision, we can use several techniques [14] such as the choice of a very small uniform step, a variable step, and finally a change of the temporal variable so that a better temporal distribution of positions is obtained on the orbit near the perigee where the speed of the satellite is faster without reducing excessively the concentration of positions in the perigee, where the curvature, as well as in the perigee, is maximum. The problem of reparameterization of the temporal variable has been studied by several author [2,4,7,10,11] using several kinds of anomalies.

In this work we follow this third way, we study especially a family of transformations derived from $dt = K_\alpha r^\alpha d\tau$ [7], called Sundman's generalized transformations.

In this section we briefly explain the terminology associated to the two body problem. A more detailed version can be seen in [1,13]. The two body problem is a classic celestial mechanics problem regarding to the problem of the motion of two punctual bodies under the action of their gravitational forces. One of the most usual ways of studying this problem is by means of the study of the relative motion: the motion of a body, generally the one of smaller mass, called secondary, with respect to that of higher mass, called primarily. If $\mathbf{r}$ is the vector of position of the secondary with respect the primary, the motion follows the equation

$$\ddot{\mathbf{r}} = -\mu \frac{\mathbf{r}}{r^3},\ \mu = G(m + m'),\ \mathbf{r}(0) = \mathbf{r}_0,\ \dot{\mathbf{r}}(0) = \mathbf{v}_0 \tag{2}$$

It is known that the two body problem satisfies Kepler's laws. The orbits of the secondary with respect to the primary are conical, with the primary in the principal focus, the area swept by the radius vector that links the primary with the secondary is proportional to the time; and the reason between the cube of the major semiaxis and the square of the period is constant for an elliptical orbit.

In the two body problem appear several important magnitudes such as the integral of the areas $\mathbf{C} = \mathbf{r} \times \mathbf{v}$ whose meaning is the double of the areolar speed. On the other hand, we have also that the vector A, called the Laplace–Runge–Lenz's vector defined as

$$\mathbf{A} = \mathbf{v} \times \mathbf{C} - \mu \frac{\mathbf{r}}{r} \tag{3}$$

is constant. It is usual to represent the vector as $\mathbf{A} = \mu \mathbf{e}$.

The equation of the relative orbit is obtained computing the scalar multiplication of $\mathbf{r}$ and $\mu \mathbf{e}$ providing

$$r = \frac{p}{1 + e \cos V} \tag{4}$$

where $p = \frac{C^2}{\mu}$ is the parameter of the conic, e the eccentricity and V the angle between $\mathbf{A}$ and $\mathbf{r}$, known as true anomaly. This angle is measured from $\mathbf{A}$. The $\mathbf{A}$ vector determines the direction of the periaster and its norm is directly related to the eccentricity e.

In addition, h is constant too. h is the integral of the energy and its value is

$$h = \frac{1}{2} v^2 - \frac{\mu}{r} \tag{5}$$

where $v^2 = \mathbf{v} \cdot \mathbf{v}$.

In the case of the elliptical motion ($0 \leq e < 1$) the value of the parameter is given by $p = a(1 - e^2)$ and the period P by $P = \frac{2\pi a^2 \sqrt{1-e^2}}{C}$, where a it is the major semi axis of the ellipse. In this case, we also define the mean motion n as $n = \frac{2\pi}{P}$ and the mean anomaly as $M = n(t - T_0)$ where T_0 is the epoch of the closest approach.

Finally, in the elliptical motion it is also of great interest the so called eccentric anomaly E related to the mean anomaly M through Kepler's equation $E - e \sin E = M$.

If the orbital system of coordinates (x, y) is considered, with O placed in the primary focus, OX in the direction of the periaster and OY perpendicular to OX so that the motion takes place in direct sense, it turns out that $r = a(1 - e \cos E)$, $x = r \cos V = a(\cos E - e)$, $y = r \sin V = a\sqrt{1 - e^2} \sin E$.

In Sect. 2 we briefly study the generalized family of Sundman's anomalies, that we use as temporal variables in the numerical integration of the two body problem.

In Sect. 3 we compute the numerical integration of different examples of the simple two body problem along 100,000 revolutions. We study the effect in the integral of the areas and the energy of the α value for different eccentricities. We also study the dependence of α on the eccentricity and the numerical precession considering Runge-Lenz's vector along long periods of time.

In section "Conclusions" we give the main conclusions of the work.

2 Study of the Family of Sundman Generalized Anomalies

In the year 1912 Sundman [12], introduced the change of temporary variable $dt = Crd\tau$ in order to regularize the problem of three bodies. Later Nacozy [6, 11] extended this transformation to a more general $dt = C_\alpha r^\alpha d\tau$, such family of transformations includes the mean, eccentric and true anomalies for values of $\alpha = 0, 1, 2$ and appropriate values of C_α [9]. From the above mentioned family Lopez [9] introduces the concept of Sundman's generalized anomaly Ψ_α as a function $\Psi_\alpha(M)$ so that

- $dM = K_\alpha(e) r^\alpha d\tau$, $K_\alpha(e) = nC_\alpha$ being n is the mean motion.
- $\Psi_\alpha(\pi) = \pi$, $\Psi_\alpha(2\pi) = 2\pi$.
- $\Psi_\alpha(M + 2\pi) = \Psi_\alpha(M)$, $\Psi_\alpha(-M) = -\Psi_\alpha(M)$.

To this aim, it is sufficient that

$$K_\alpha(e) = \frac{1}{2\pi} \int_0^{2\pi} (1 - e\cos E)^{1-\alpha} dE \tag{6}$$

whose value is given by [9]

$$K_\alpha(e) = a^{-\alpha} \left\{ (1-e)^{1-\alpha} F\left(\frac{1}{2}, \alpha - 1, 1; \frac{2e}{e-1}\right) + (1+e)^{1-\alpha} F\left(\frac{1}{2}, \alpha - 1, 1; \frac{2e}{1+e}\right) \right\} \tag{7}$$

where $F(a, b, c; z)$ is the hypergeometrical function.

The function $\Psi_\alpha - M$ can be developed as Fourier series depending on M and as Fourier series of Ψ_α [9], where the development of $\frac{1}{r}$, $r \sin V$ y $r \cos V$ is also obtained as Fourier series depending on Ψ_α. So, a set of developments sufficient for the analytical treatment of the problem is provided.

With regard to the concerned numerical methods, the differential equations of motion depend on the t variable; in this way, we have

$$\frac{d}{dt} = n\frac{d}{dM} = \frac{n}{K_\alpha(e)} r^{-\alpha} \frac{d}{d\Psi_\alpha}. \tag{8}$$

Thus, the equations of motion of the two body problem in the orbital coordinates (x, y) are

$$\frac{dx}{d\Psi_\alpha} = \frac{K_\alpha(e)}{n} r^\alpha v_x, \quad \frac{dv_x}{d\Psi_\alpha} = -\frac{K_\alpha(e)}{n} r^\alpha GM \frac{x}{r^3} \tag{9}$$

$$\frac{dy}{d\Psi_\alpha} = \frac{K_\alpha(e)}{n} r^\alpha v_y, \quad \frac{dv_y}{d\Psi_\alpha} = -\frac{K_\alpha(e)}{n} r^\alpha GM \frac{y}{r^3}. \tag{10}$$

The use of an appropriate value for the α parameter improves the efficiency of the integration in the two body problem. The optimal value for α for each value of eccentricity can be approached by

$$\alpha(e) = 1.5541e^4 - 1.94142e^3 + 0.582338e^2 + 0.252954e + 1.54422. \quad (11)$$

The robustness of these value has been tested using a Runge–Kutta of eight order [3] and Gragg-Bulirsch-Stoer integrators [5].

3 Numerical Results

In the present section we study the motion on the orbital plane of a fictitious satellite with the same orbital elements that the old HEOSII satellite except for the eccentricity, that is modified to simulate different cases. As one would expect, low values of the eccentricity do not provide significantly different results. To test the efficiency of the integrators for higher values of the eccentricity we consider a high value $e = 0.5$ and an extreme value $e = 0.95$.

First we carry out the integration of a satellite with a high eccentricity $e = 0.5$, considering 10,000 orbits, using a classic Runge–Kutta method of fourth order with 1,000 uniform steps. Firstly we employ the mean anomaly $\alpha = 0$ and secondly the Nacozy's intermediate anomaly $\alpha = 1.5$. In this last case, the results are improved. Figure 1 shows the magnitude of the variations in the quantities C, H, e, ω, that are constants at the perigee in the analytical solution of two body problem, depending on the used anomaly. In each subfigure the OX axis represents the number of revolutions and the OY axis the value of the quantities C, H, e is the eccentricity and ω is in radians. For the initial epoch $t = 0$ $C = 188{,}109.144$ and $H = -1.68376245$

The integration is repeated for a case with extreme eccentricity $e = 0.95$. In this case, the use of the mean anomaly as variable of integration provides absolutely inadmissible results. We obtain considerably improved results in the case $\alpha = 1.9$, shown in Fig. 2. For $t = 0$ $C = 67{,}823.519$.

The long time error in the quantities of the energy H, the areas integral C, the eccentricity $e(t)$ and the numerical precession of the perigee ω can be improved using an appropriate value of α. If $e = 0.5$ and $e = 0.95$ the dependence of the results on the value of the chosen parameter α is evident. In order to test the robustness of the method, these results have been compared with the ones obtained using a Bulirsch-Stoer method. In both cases the results are similar.

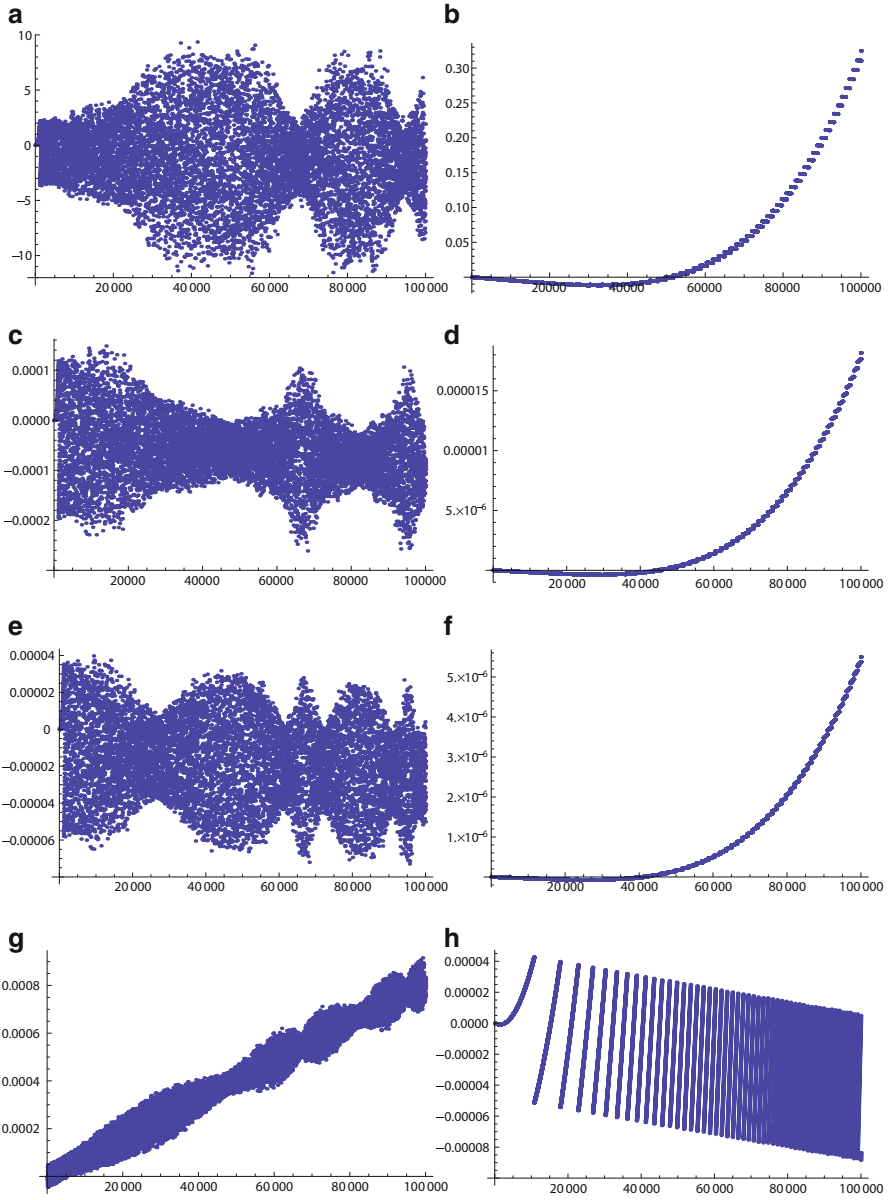


Fig. 1 Time evolution of ΔC, ΔH, Δe and $\Delta\omega$. (**a**) ΔC : $e = 0.5$ $\alpha = 0.0$. (**b**) ΔC : e=0.5. $\alpha = 1.5$. (**c**) ΔH : $e = 0.5$ $\alpha = 0.0$. (**d**) ΔH : $e = 0.95$ $\alpha = 1.9$. (**e**) $\Delta e(t)$: $e = 0.5$ $\alpha = 0.0$. (**f**) $\Delta e(t)$: $e = 0.5$ $\alpha = 1.5$. (**g**) $\Delta\omega$: $e = 0.5$ $\alpha = 0.0$. (**h**) $\Delta\omega$: $e = 0.5$ $\alpha = 1.5$

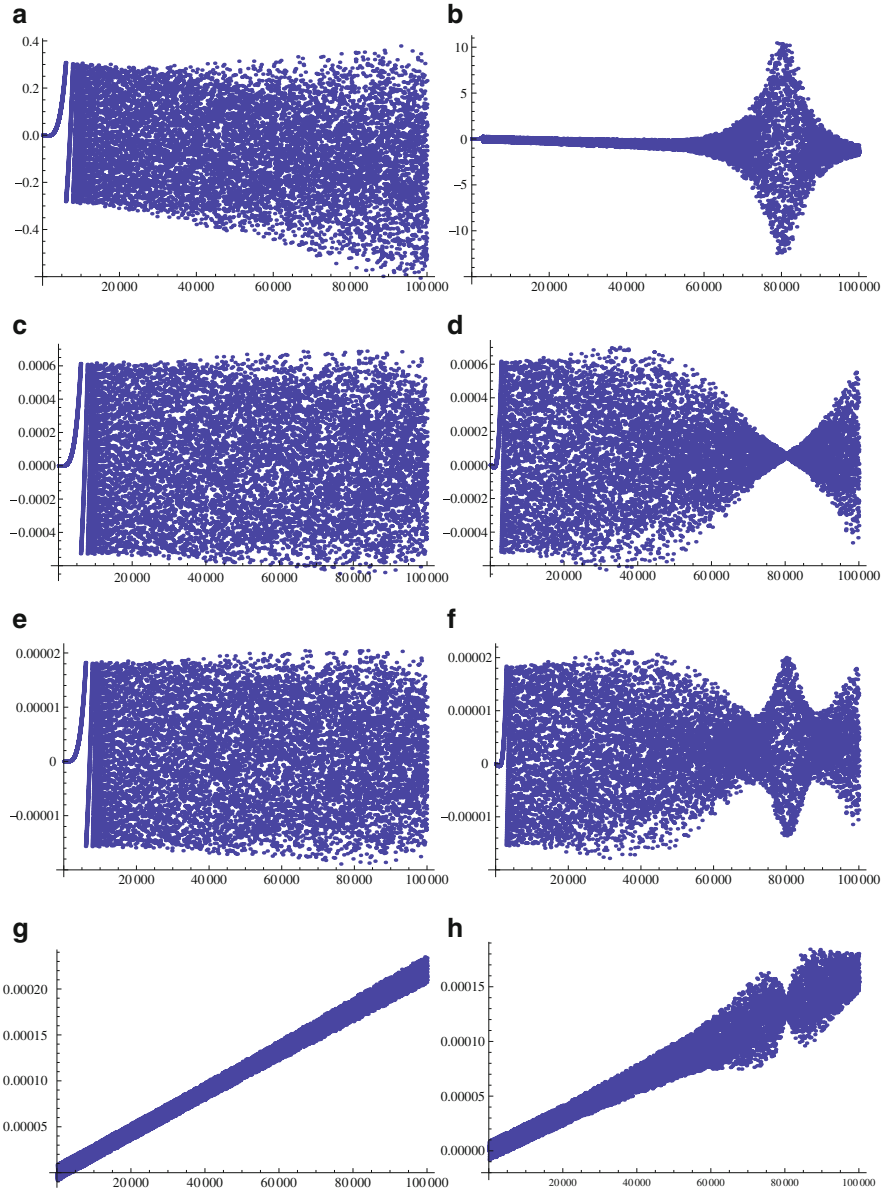


Fig. 2 Time evolution of C, H, e and ω. (**a**) $\Delta C \ : \ e = 0.95 \ \alpha = 1.5$. (**b**) $\Delta C \ : \ e = 0.95 \ \alpha = 1.9$. (**c**) $\Delta H \ : \ e = 0.95 \ \alpha = 1.5$. (**d**) $\Delta H \ : \ e = 0.95 \ \alpha = 1.9$. (**e**) $\Delta e(t) \ : \ e = 0.95 \ \alpha = 1.5$. (**f**) $\Delta e(t) \ : \ e = 0.95 \ \alpha = 1.9$. (**g**) $\Delta \omega \ : \ e = 0.95 \ \alpha = 1.5$. (**h**) $\Delta \omega \ : \ e = 0.95 \ \alpha = 1.9$

Conclusions
In this work it is remarked that it is of great importance the temporary variable chosen in the numerical integration of the orbital motion. The use of an appropriate anomaly from the family of Sundman's generalized anomalies improves the preservation along long temporal periods of quantities that must remain invariant in the two body problem. The first integrals given by the constant of the areas, the energy, the direction of apoaster and the Laplace–Runge–Lenz's vector, which determines the value of the eccentricity, are slightly sensitive to the value of α for small eccentricities. When the value of the eccentricity increases, the conservation of these quantities is a much more delicate problem. For extreme values of the eccentricity $e = 0.95$ the results obtained for low values of α are inadmissible. In these cases, the most adequate values for α are between 1.5 and 1.9.

Acknowledgements This work has been partially supported by a grant P1.1B2012-47 of the Universidad Jaume I.

References

1. Brower, D., Clemence, G.M.: Celestial Mechanics. Academic, New York (1965)
2. Brumberg, E.V.: Length of arc as independent argument for highly eccentric orbits. Celest. Mech. **53**, 323–328 (1992)
3. Fehlberg, E., Marsall, G.C.: Classical fifth, sixth, seventh and eighth Runge–Kutta formulas with stepsize control. Technical report, NASA, R-287 (1968)
4. Ferrándiz, J.M., Ferrer, S., Sein-Echaluce, M.L.: Generalized elliptic anomalies. Celest. Mech. **40**, 315–328 (1987)
5. Gragg, W.B.: Repeated extrapolation to the limit in the numerical solution of ordinary differential equations. SIAM J. Numer. Anal. **2**, 384–403 (1965)
6. Janin, G.: Accurate computation of highly eccentric satellite orbits. Celest. Mech. **10**, 451–467 (1974)
7. Janin, G., Bond, V.R.: The elliptic anomaly. Technical memorandum, NASA, n. 58228 (1980)
8. Levallois, J.J., Kovalevsky, J.: Géodésie Générale, vol. 4. Eyrolles, Paris (1971)
9. López, J.A., Agost, V., Barreda, M.: A note on the use of the generalized Sundman transformations as temporal variables in celestial mechanics. Int. J. Comput. Math. **89,** 433–442 (2012)
10. López, J.A., Marco, F.J., Martínez, M.J.: A study about the integration of the elliptical orbital motion based on a special one-parametric family of anomalies. Abstr. Appl. Anal. **2014**, ID 162060, 1–11 (2014)
11. Nacozy, P.: The intermediate anomaly. Celest. Mech. **16**, 309–313 (1977)
12. Sundman, K.: Memoire sur le probleme des trois corps. Acta Math. **36**, 105–179 (1912)
13. Tisserand, F.F.: Traité de Mecanique Celeste. Gauthier-Villars, Paris (1896)
14. Velez, C.E., Hilinski, S.: Time transformation and Cowell's method. Celest. Mech. **17**, 83–99 (1978)

Piecewise Linear Analogue of Hopf-Zero Bifurcation in an Extended BVP Oscillator

Enrique Ponce, Javier Ros, and Elísabet Vela

Abstract In this work, we consider a family of symmetric piecewise linear systems in three dimensions. By using the analysis done in Ponce et al. (Physica D 250:34–46, 2013), in which the authors detect and characterize the appearance of limit cycles for the bifurcation parameter passing through its critical value, we show that Hopf-zero bifurcation takes place for a certain range of parameters in an extended version of Bonhoeffer-van der Pol oscillator.

1 Introduction and Preliminary Results

Hopf-zero bifurcation, also called fold-Hopf or Hopf-pitchfork bifurcation, is a specific bifurcation in 3D vector fields. This bifurcation is characterized by the simultaneous crossing of three eigenvalues at the imaginary axis of the complex plane. In a recent paper [6], the authors give for the first time, information about the general unfolding of such bifurcation in the framework of piecewise linear systems.

Here, we apply the achieved theoretical results to a physically and biologically interesting oscillator system, the Bonhoeffer-van der Pol (BVP for short) oscillator, which can be considered as a generalization of both the Duffing oscillator and the well-known van der Pol oscillator, see [3]. It is pointed out that apart from the usual period-doubling bifurcations leading to chaotic dynamics, the system also exhibits resonance or phase-locking phenomena when external constant and periodic forces are applied.

Thus, we consider the same family of piecewise linear differential systems studied in [6] written in the Luré form,

$$\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x}) = A_R\mathbf{x} + \mathbf{b}\,\mathrm{sat}(x), \tag{1}$$

E. Ponce (✉) • J. Ros • E. Vela
Departamento Matemática Aplicada II, Camino Descubrimientos, E.T.S. Ingeniería, 41092 Sevilla, Spain
e-mail: eponcem@us.es; javieros@us.es; elivela@us.es

F. Casas, V. Martínez (eds.), *Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 4, DOI 10.1007/978-3-319-06953-1_12

where $\mathbf{x} = (x, y, z)^T \in \mathbb{R}^3$ and the dot represents the derivative with respect to the time τ. Without loss of generality, we can assume that the matrix A_R and the vector $\mathbf{b}$ have the following expressions

$$A_R = \begin{pmatrix} t & -1 & 0 \\ m & 0 & -1 \\ d & 0 & 0 \end{pmatrix} \text{ and } \mathbf{b} = \begin{pmatrix} T - t \\ M - m \\ D - d \end{pmatrix}, \tag{2}$$

where t, m, d, T, M and D are the basic parameters of the model, to be later identified. The scalar function sat stands for the normalized saturation given by

$$\text{sat}(u) = \begin{cases} 1 \text{ if } u > 1, \\ u \text{ if } |u| \leqslant 1, \\ -1 \text{ if } u < -1. \end{cases}$$

System (1)–(2) has the following properties:

(a) It is symmetric with respect to the origin, i.e. $\mathbf{F}(\mathbf{x}) = -\mathbf{F}(-\mathbf{x})$.
(b) In the region with $|x| \leqslant 1$ it becomes the homogeneous system

$$\dot{\mathbf{x}} = A_C \mathbf{x} = \begin{pmatrix} T & -1 & 0 \\ M & 0 & -1 \\ D & 0 & 0 \end{pmatrix} \mathbf{x}(\tau). \tag{3}$$

(c) The coefficients t, m, d and T, M, D are the linear invariants (trace, sum of principal minors of order two and determinant) of the matrices A_R and A_C, respectively.

Note that $A_C = A_R + \mathbf{b}\mathbf{e}_1^T$, where $\mathbf{e}_1 = (1, 0, 0)^T$, and that the considered family of systems is in the so-called generalized Liénard form, see [1], which is in fact equivalent to the observable canonical form in control theory [2]. Thus, under generic conditions for every system of the form (1), after some change of variables, we can get the matrices in the form given in (2) and (3).

As done in [6], we introduce ε as the main initial bifurcation parameter such that the three eigenvalues of the matrix A_C are $-\varepsilon$ and $\rho\varepsilon \pm \omega i$, where $\rho \in \mathbb{R}$ and $\omega \in \mathbb{R}^+$ are auxiliary fixed parameters. Thus for $\varepsilon = 0$ the three eigenvalues are 0 and $\pm\omega i$, which are located on the imaginary axis of the complex plane, reproducing so the critical situation associated to the Hopf-zero bifurcation in differentiable dynamics. Accordingly we choose

$$T(\varepsilon) = (2\rho - 1)\varepsilon,\ M(\varepsilon) = \omega^2 + \rho\varepsilon^2(\rho - 2),\ D(\varepsilon) = -\varepsilon(\rho^2\varepsilon^2 + \omega^2), \tag{4}$$

to be assumed hereinafter.

Note that for $\varepsilon = 0$, the solutions of system (3) give rise to orbits that, if they are completely contained in the central zone, define planar ellipses forming a bounded set foliated by periodic orbits. This periodic set has the shape of two solid cones sharing the elliptic disc $\omega^2 x^2 + y^2 \leqslant \omega^2$ in the plane $z = 0$ as their common basis, see Fig. 1.

Regarding this piecewise linear version of the Hopf-zero bifurcation, the following results come directly from theorems 3 and 6 in [6], where much more qualitative and quantitative information can be found. We need first to define the parameter $\delta = d - t\omega^2$, which characterizes the criticality of the bifurcation.

Theorem 1 *Let us consider system* (1)–(2) *under conditions* (4) *where it is assumed* $\rho \neq 0$ *and* $\delta = d - t\omega^2 \neq 0$ *and fixed.*

For $\varepsilon = 0$ *the system* (1)–(2) *undergoes a tri-zonal limit cycle bifurcation, that is, from the configuration of periodic orbits that exists in the central zone for* $\varepsilon = 0$, *one limit cycle appears for* $\rho\delta\varepsilon > 0$ *and* $|\varepsilon|$ *sufficiently small. It is symmetric with respect to the origin and bifurcates from the ellipse*

$$\Gamma = \{(x, y, z)^T \in \mathbb{R}^3 : \omega^2 x^2 + y^2 = \omega^2, z = 0\}.$$

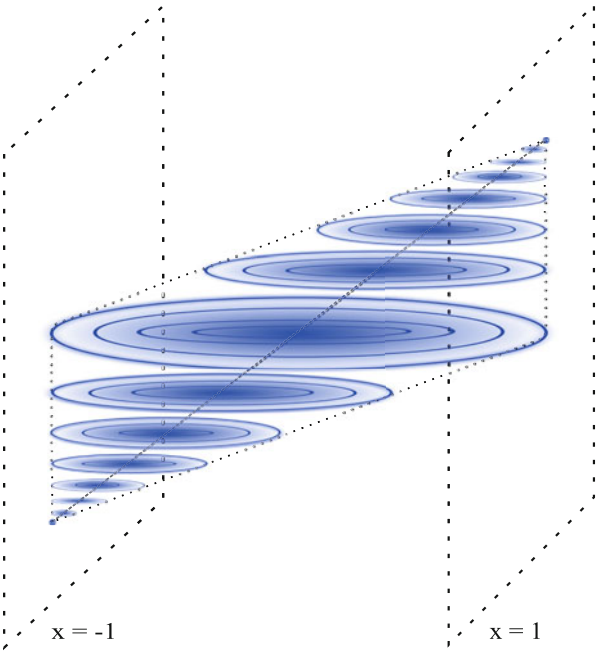


Fig. 1 Structure of the periodic orbits for $\varepsilon = 0$ in the central zone. The two solid cones are completely foliated by periodic orbits surrounding the segment of equilibrium points $\{(x, 0, x\omega^2)^T : |x| \leqslant 1\}$

Furthermore, the bifurcating limit cycle is stable if and only if $t < 0$, $\delta > 0$ *and* $d < 0$.

In the case $\delta = d - t\omega^2 = 0$ with $d \neq 0$, a similar result but requiring a special consideration, appeared in [6]; furthermore, such a case is analyzed in a biparametric context in [5].

From the quoted analysis in [6], we must also emphasize the following result about the possible simultaneous bifurcation of bizonal limit cycles. Note that, due to the symmetry of the system (1)–(2), bizonal limit cycles will always appear in pairs, each one crossing one of the planes $x = 1$ and $x = -1$, respectively. Thus, from the periodic orbits tangent to the planes $x = \pm 1$ when $\varepsilon = 0$, considering the two ones contained in the planes $z = \pm\hat{z}$, where

$$\hat{z} = \frac{d\rho\omega^2}{d\rho + \delta},$$

a limit cycle bifurcate under appropriate hypotheses, as follows.

Theorem 2 *Consider system* (1)–(2) *under conditions* (4) *and assuming that* $\rho \neq 0$, $\delta = d - t\omega^2 \neq 0$, $d\rho + \delta \neq 0$,

$$0 < \hat{z} = \frac{d\rho\omega^2}{d\rho + \delta} < \omega^2,$$

and fixed. Under these hypotheses a bizonal limit cycle bifurcation takes place for the critical value $\varepsilon = 0$. *Thus, a symmetrical pair of limit cycles appears when* $\rho\delta\varepsilon > 0$ *and* $|\varepsilon|$ *is sufficiently small. They are stable if and only if* $t < 0$ *and* $\rho > 0$, *or* $t = 0$, $\rho > 0$ *and* $d(2\rho - 1) < 0$.

2 Realization in an Extended BVP Oscillator

In this section we consider an extended BVP oscillator, which is consisted of two capacitors, an inductor, a linear resistor and a nonlinear conductance, as shown in Fig. 2. To obtain more information about this circuit, see [4], where a smooth nonlinearity is assumed for the conductance and a rich variety of dynamical behaviors is found. The circuit equations are as follows:

$$C\frac{dv_1}{dt} = -i - g(v_1), \quad C\frac{dv_2}{dt} = i - \frac{v_2}{r}, \quad L\frac{di}{dt} = v_1 - v_2,$$

where v_1 and v_2 are the voltages across the capacitors, the symbol i stands for the current through the inductance L, and the $v - i$ characteristics of the nonlinear resistor is written as $g(v) = -av - b\,\mathrm{sat}(cv)$, where $a, b, c > 0$. Note that here we adopt a PWL version of the nonlinearity considered in [4].

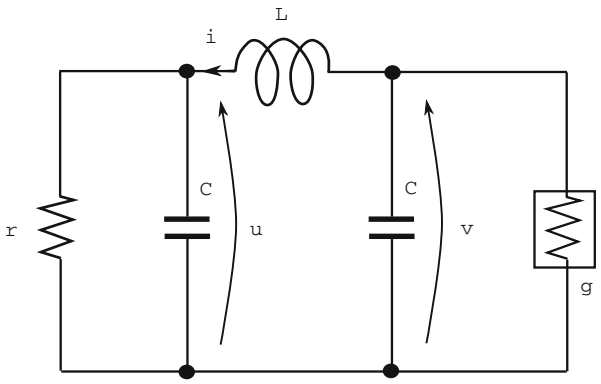


Fig. 2 The extended BPV oscillator proposed in [4]

After some standard manipulations, the normalized equations of the extended BVP oscillator become

$$\begin{cases} \dot{x} = -z + \alpha x + \text{sat}(\beta x), \\ \dot{y} = z - \gamma y, \\ \dot{z} = x - y, \end{cases}$$

where the dot represents derivative with respect to the new time τ, and

$$\tau = \frac{1}{\sqrt{LC}}t, \quad \alpha = a\sqrt{\frac{L}{C}}, \quad \beta = bc\sqrt{\frac{L}{C}}, \quad \gamma = \frac{1}{r}\sqrt{\frac{L}{C}},$$

$$x = \frac{v_1}{b}\sqrt{\frac{C}{L}}, \quad y = \frac{v_2}{b}\sqrt{\frac{C}{L}}, \quad z = \frac{i}{b}.$$

Making now the change of variables $X = \beta x$, we obtain the system in its Luré form,

$$\dot{\mathbf{x}} = \begin{pmatrix} \alpha & 0 & -\beta \\ 0 & -\gamma & 1 \\ 1/\beta & -1 & 0 \end{pmatrix} \mathbf{x} + \begin{pmatrix} \beta \\ 0 \\ 0 \end{pmatrix} \text{sat}(\mathbf{e}_1^T \mathbf{x}), \tag{5}$$

and we will rename X as x in the sequel, for convenience. It is easy to see that system (5) is observable if and only if $\beta \neq 0$; in particular, since $\beta > 0$, it can be

written in the form (1)–(2), and so we can apply both Theorems 1 and 2. Effectively, with a linear change of variables given by the matrix

$$P = \frac{1}{\beta}\begin{pmatrix} \beta & 0 & 0 \\ \gamma^2 - 1 & \gamma & 1 \\ \gamma & 1 & 0 \end{pmatrix},$$

we can write system (5) in its Liénard form as

$$\dot{\mathbf{x}} = \begin{pmatrix} \alpha - \gamma & -1 & 0 \\ 2 - \alpha\gamma & 0 & -1 \\ \alpha - \gamma & 0 & 0 \end{pmatrix}\mathbf{x} + \begin{pmatrix} \beta \\ -\beta\gamma \\ \beta \end{pmatrix}\mathrm{sat}(x), \tag{6}$$

where now the trace, the sum of second order principal minors and the determinant in the different zones are evident, namely

$$\begin{array}{lll} T = \alpha + \beta - \gamma, & t = \alpha - \gamma, \\ M = 2 - \gamma(\alpha + \beta), & m = 2 - \alpha\gamma, \\ D = \alpha + \beta - \gamma, & d = \alpha - \gamma. \end{array} \tag{7}$$

Note that the origin is always an equilibrium point and that from the last component of (6) we have the equilibria condition $(\gamma - \alpha)x = \beta\,\mathrm{sat}(x)$, so that we have an extra symmetric pair of equilibria whenever

$$0 < \gamma - \alpha < \beta. \tag{8}$$

Following a similar procedure to the one done for the Chua's circuit in [6], and looking for the Hopf-pitchfork bifurcation in this model, we need to check not only the hypotheses of different theorems of [6] but also the feasibility of conditions (4). That is, we need to impose

$$\begin{array}{l} E_1 := (2\rho - 1)\varepsilon - \alpha - \beta + \gamma = 0, \\ E_2 := \omega^2 + \rho\varepsilon^2(\rho - 2) - 2 + \gamma(\alpha + \beta) = 0, \\ E_3 := -\varepsilon(\rho^2\varepsilon^2 + \omega^2) - \alpha - \beta + \gamma = 0. \end{array} \tag{9}$$

From (7) we observe that T and D are identically equal, what implies that we cannot move the position of eigenvalues at will. Thus, for instance, the parameter ρ cannot be fixed a priori; it must depend instead on ε in order to satisfy (9). More precisely, from (9) by substracting E_3 from E_1 we have

$$\varepsilon(\rho^2\varepsilon^2 + \omega^2) + (2\rho - 1)\varepsilon = 0,$$

leading for $\varepsilon \neq 0$ to the condition

$$\varepsilon^2\rho^2 + 2\rho + \omega^2 - 1 = 0. \tag{10}$$

Therefore the value of ρ cannot be arbitrarily chosen, nor constant (as we supposed before). Indeed, as the only solution of (10) that becomes regular at $\varepsilon = 0$, we have

$$\rho = \rho(\varepsilon) = \frac{1-\omega^2}{1+\sqrt{1-\varepsilon^2(\omega^2-1)}}, \quad \text{with } \rho(0) = \frac{1-\omega^2}{2}, \tag{11}$$

and consequently we have the condition $2\rho(0) < 1$. We assume in the sequel the above choice for $\rho(\varepsilon)$ and neglect the third equation of (9), so to be automatically fulfilled. We also rewrite the second equation by using the above relation, namely

$$\begin{aligned} E_1 &:= [2\rho(\varepsilon) - 1]\varepsilon - \alpha - \beta + \gamma = 0, \\ E_2 &:= -1 - 2\rho(\varepsilon)(1+\varepsilon^2) + \gamma(\alpha+\beta) = 0. \end{aligned} \tag{12}$$

In looking for the Hopf-pitchfork bifurcation to take place at $\varepsilon = 0$, we need

$$\begin{aligned} E_1^0 &:= -\alpha - \beta + \gamma = 0, \\ E_2^0 &:= \omega^2 - 2 + \gamma(\alpha+\beta) = 0, \end{aligned} \tag{13}$$

leading to the necessary condition

$$\omega^2 + (\alpha+\beta)^2 = 2, \tag{14}$$

that is, we must assume both $\omega < \sqrt{2}$ and $\alpha + \beta < \sqrt{2}$, so that from (11) we also obtain $-1 < 2\rho(0) < 1$.

In what follows, we assume $\beta > 0$ fixed and we allow α and γ to be moved, writing $\alpha(\varepsilon)$ and $\gamma(\varepsilon)$ for the functions satisfying (12). From Eqs. (13) we obtain the following equalities,

$$\begin{aligned} \alpha_0 &= \alpha(0) = -\beta + \sqrt{2-\omega^2}, \\ \gamma_0 &= \gamma(0) = \sqrt{2-\omega^2}. \end{aligned} \tag{15}$$

From (12) and using the equalities of (15), it is easy to check that the required condition to reproduce the eigenvalues transition is

$$\det\left(\frac{\partial(E_1, E_2)}{\partial(\alpha,\gamma)}\right)_{\varepsilon=0} = \begin{pmatrix} -1 & 1 \\ \gamma_0 & \alpha_0+\beta \end{pmatrix} = -\alpha_0 - \beta - \gamma_0 = -2\gamma_0 \neq 0.$$

Under this last condition, the Implicit Function Theorem assures, for $|\varepsilon|$ sufficiently small, the existence of a branch of solutions $(\rho(\varepsilon), \alpha(\varepsilon), \gamma(\varepsilon))$ of (9), with β a fixed parameter, leading to the eigenvalue transition corresponding to the Hopf-pitchfork bifurcation. From (7), when ε vanishes, we obtain $t = d = -\beta$, and

$m = \omega^2 + \beta\sqrt{2-\omega^2}$. Thus, for this set of parameters, it is easy to check that the non-degeneracy condition

$$\delta = d - t\omega^2 = -\beta\left(1-\omega^2\right) \neq 0 \tag{16}$$

holds if and only if $\omega \neq 1$, and then we have necessarily $\rho(0) \neq 0$.

Then, our PWL Hopf-pitchfork bifurcation at $\varepsilon = 0$ in system (6) is guaranteed by Theorem 1 in two cases:

(a) $0 < \omega < 1$, which requires $\alpha + \beta > 1$ and leads to $\rho(0) > 0$, and
(b) $1 < \omega < \sqrt{2}$, which requires $\alpha + \beta < 1$ and gives $\rho(0) < 0$.

We know then that the bifurcating tri-zonal limit cycle appears for $\rho\delta\varepsilon > 0$, that is, for $-\beta(1-\omega^2)^2\varepsilon > 0$; in short, for $\varepsilon < 0$. It is stable if and only if $t < 0$, $d < 0$ and $\delta > 0$, that is, if $\beta > 0$ and $\omega > 1$. Thus we have a tri-zonal unstable limit cycle in case (a) and a stable limit cycle in the case (b), appearing for $\varepsilon < 0$ in both cases.

To apply Theorem 2, we now compute the value of $d\rho(0) + \delta$, obtaining

$$-\beta\frac{1-\omega^2}{2} - \beta\left(1-\omega^2\right) = -3\beta\frac{1-\omega^2}{2},$$

and the value

$$\frac{d\rho(0)}{d\rho(0)+\delta} = \frac{1}{3}.$$

Thus, the hypotheses of Theorem 2 are fulfilled both in case (a) and (b), obtaining that two bizonal limit cycles also bifurcate for $\varepsilon < 0$. These bizonal limit cycles are stable for $t < 0$ and $\rho > 0$, being so stable in case (a) and unstable in case (b).

Note that we obtain the simultaneous bifurcation of three limit cycles for $\varepsilon < 0$, and since from (11) we have

$$T(\varepsilon) = [2\rho(\varepsilon) - 1]\varepsilon = -\omega^2\varepsilon + O\left(\varepsilon^2\right),$$

it must be concluded that the bifurcation occurs for $T > 0$, that is for $\gamma < \alpha + \beta$. We must also remark from (8) that then there also appear two isolated equilibrium points and they are stable if $t, d < 0$, that is $\beta > 0$, and $mt - d < 0$, which from (7) leads to $m - 1 > 0$. This last inequality is equivalent to $\gamma < 1/\alpha$; at the bifurcation values we have $\gamma_0 = \alpha_0 + \beta$ and so it is fulfilled in the case (b), where $\alpha_0 < 1$ is guaranteed.

From the above analysis, we can summarize our results using γ as the main bifurcation parameter and stating what we have proved.

Theorem 3 *Considering system (5) or equivalently system (6) with $\alpha > 0$, $\beta > 0$ and $\alpha + \beta < \sqrt{2}$, the following statements hold.*

(a) For $\alpha + \beta - \gamma < 0$ the origin is the only equilibrium of the system. Furthermore, if $\gamma(\alpha + \beta) < 1$ then the origin is asymptotically stable.

(b) For $\alpha + \beta - \gamma = 0$ *the system undergo a PWL analogue of the Hopf-zero bifurcation; from the periodic set existing at such critical situation, for* $\alpha + \beta - \gamma > 0$ *and sufficiently small the bifurcation leads to the simultaneous appearance of three limit cycles (one tri-zonal and two bizonal ones) along with two additional equilibrium points, being the origin not stable any longer.*

Furthermore, if $\alpha + \beta < 1$ $(1 < \alpha + \beta < \sqrt{2})$*, then the bifurcating tri-zonal limit cycle is stable (unstable) while the bifurcation bizonal limit cycles are unstable (stable). The bifurcating equilibrium points are stable whenever* $\alpha + \beta < 1$ *and, in the case* $1 < \alpha + \beta < \sqrt{2}$*, when* $\gamma < 1/\alpha$.

The unfolding of the degeneration appearing for $\alpha + \beta = 1$ needs a special treatment to be done elsewhere.

Acknowledgements Authors are partially supported by the *Ministerio de Ciencia y Tecnología, Plan Nacional I+D+I*, in the frame of projects MTM2010-20907 and MTM2012-31821, and by the *Consejería de Educación y Ciencia de la Junta de Andalucía* under grants TIC-0130.

References

1. Carmona, V., Freire, E., Ponce, E., Torres, F.: On simplifying and classifying piecewise-linear systems. IEEE Trans. Circuits Syst. **49**, 609–620 (2002)
2. Chen, C.-T.: Linear System Theory and Design, 3rd edn. OUP, New York (1998)
3. Lakshmanan, M., Murali, K.: Chaos in Nonlinear Oscillators: Controlling and Synchronization. Series on Nonlinear Science, vol. 13. World Scientific, Singapore (1996)
4. Nishiuchi, Y., Ueta, T., Kawakami, H.: Stable torus and its bifurcation phenomena in a simple three-dimensional autonomous circuit. Chaos Solitons Fractals **27**, 941–951 (2006)
5. Ponce, E., Ros, J., Vela, E.: A Hopf-zero degenerated case in symmetric piecewise linear systems. In: Progress and Challenges in Dynamical Systems. Springer Proceedings in Mathematics & Statistics, vol. 54. Springer, Berlin Heidelberg (2013)
6. Ponce, E., Ros, J., Vela, E.: Unfolding the fold-Hopf bifurcation in piecewise linear continuous differential systems with symmetry. Physica D **250**, 34–46 (2013)

Part III
Applications and Modeling

On Multiresolution Transforms Based on Weighted-Least Squares

Francesc Aràndiga and Dionisio F. Yáñez

Abstract This work is devoted to construct Harten's multiresolution transforms using Weighted-Least squares for different discretizations. We establish a relation between the filters obtained using some decimation operators. Some properties and examples of filters are presented.

1 Introduction: Harten's Framework for Multiresolution

In the last years different multiresolution (MR) transforms have been developed in order to design compression algorithms (see [1–3, 6, 7]).

In Harten's framework for MR is defined by two operators: discretization operator $\mathscr{D}_k$ and reconstruction operator $\mathscr{R}_k$ (k implies more the resolution level).

Let be $\mathscr{F}$ a functions space and let be V^k a vectorial space of discrete signals. The discretization function $\mathscr{D}_k : \mathscr{F} \to V^k$ is a linear operator that discretizes a function $f \in \mathscr{F}$ to a signal $f^k = \mathscr{D}_k f$. The reconstruction function $\mathscr{R}_k : V^k \to \mathscr{F}$ is a linear or **non** linear operator that maps a discrete signal to a continuous function.

The principal relation to these operators is the consistence. Therefore,

$$\mathscr{D}_k \mathscr{R}_k = I_{V^k}, \tag{1}$$

where I_{V^k} is the identity function.

In order to construct a MR scheme we define the decimation operator as $\mathscr{D}_k^{k-1} = \mathscr{D}_{k-1}\mathscr{R}_k$ and the prediction operator as $\mathscr{P}_{k-1}^k = \mathscr{D}_k \mathscr{R}_{k-1}$.

F. Aràndiga
Departament de Matemática Aplicada, Universitat de València, C/Doctor Moliner, 46100 Burjassot, Valencia, Spain
e-mail: arandiga@uv.es

D.F. Yáñez (✉)
Departamento de Matemáticas, Campus Capacitas, CC. NN. y CC. SS. aplicadas a la Educación, Universidad Católica de Valencia, C/Sagrado Corazón, 46110 Godella, Valencia, Spain
e-mail: dionisiofelix.yanez@ucv.es

F. Casas, V. Martínez (eds.), *Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 4, DOI 10.1007/978-3-319-06953-1_13

It is easy to prove that the operators satisfy the consistence property (1) then:

$$\mathscr{D}_k^{k-1}\mathscr{P}_{k-1}^k = I_{V^{k-1}}. \tag{2}$$

However $\mathscr{P}_{k-1}^k\mathscr{D}_k^{k-1}f^k$ is an approximation of f^k, therefore we can define the prediction error as:

$$e^k = f^k - \mathscr{P}_{k-1}^k f^{k-1}. \tag{3}$$

The operators $\mathscr{D}_k^{k-1}$ and $\mathscr{P}_{k-1}^k$ constitute a pyramid MR scheme [5]. If $N_k = \dim(V^k)$ then the error e^k have the same dimension that f^k, therefore the pair (f^{k-1}, e^k) represent redundant information that we can eliminate. For this, we apply the linear operator $\mathscr{D}_k^{k-1}$ (linear) to Eq. (3) and obtain:

$$\mathscr{D}_k^{k-1}e^k = \mathscr{D}_k^{k-1}f^k - \mathscr{D}_k^{k-1}\mathscr{P}_{k-1}^k f^{k-1} = f^{k-1} - f^{k-1} = 0, \tag{4}$$

then $e^k \in \mathscr{N}(\mathscr{D}_k^{k-1}) = \{v | v \in V^k, \mathscr{D}_k^{k-1}v = 0\}$. We can eliminate by selecting a set of basis functions, $\{\mu_j^k\}$, in $\mathscr{N}(\mathscr{D}_k^{k-1})$ and defining a function $G_k : \mathscr{N}(\mathscr{D}_k^{k-1}) \to \mathscr{G}^k$, which assigns to any $e^k \in \mathscr{N}(\mathscr{D}_k^{k-1})$ the sequence d^k in the basis, $d^k = G_k e^k$, and let $\tilde{G}_k$ be the canonical injection $\mathscr{N}(\mathscr{D}_k^{k-1}) \hookrightarrow V^k$. Then $G_k\tilde{G}_k = I_{\mathscr{G}_k}, \tilde{G}_k G_k = I_{\mathscr{N}(\mathscr{D}_k^{k-1})}$. There is a one to one correspondence between f^k and (f^{k-1}, d^k): Given f^k we evaluate

$$f^{k-1} = \mathscr{D}_k^{k-1}f^k, \quad d^k = G_k(f^k - \mathscr{P}_{k-1}^k\mathscr{D}_k^{k-1}f^k), \tag{5}$$

and given $f^{k-1} = \mathscr{D}_k^{k-1}f^k$ and d^k we reconstruct f^k by

$$\mathscr{P}_{k-1}^k f^{k-1} + \tilde{G}_k d^k = \mathscr{P}_{k-1}^k\mathscr{D}_k^{k-1}f^k + \tilde{G}_k G_k(f^k - \mathscr{P}_{k-1}^k\mathscr{D}_k^{k-1}f^k) = f^k. \tag{6}$$

This single stage is iterated on the decimated signal for a MR representation $f^N \equiv (f^0, d^1, \ldots, d^N)$.

The Algorithms to obtain the direct and inverse multiscale transformations are the following:

Algorithm 1 *Encoding* $f^N \to Mf^N = (f^0, d^1, \ldots, d^N)$

for $k = N, \ldots, 1$
$f^{k-1} = \mathscr{D}_k^{k-1}f^k$
$d^k = G_k(f^k - \mathscr{P}_{k-1}^k f^{k-1})$
end

Algorithm 2 *Decoding* $Mf^N \to M^{-1}Mf^N$

$$\begin{aligned} &for \quad k = 1, \ldots, N \\ &\quad f^k = \mathscr{P}^k_{k-1} f^{k-1} + \tilde{G}_k d^k \\ &end \end{aligned}$$

The Algorithms 1 and 2 have the same structure as Mallat's decomposition and reconstruction operators (see [8]). Also, Harten's framework for MR is related to the one-step *lifting scheme* scheme developed by Sweldens et al. (see [9]).

We design a MR scheme using Weighted-Least squares for different discretizations: point-value, cell-average and hat-based. The stability is ensured by the linearity of the operator. We present some examples of filters.

1.1 Discretization Operators: Point-Value, Cell-Average and Hat-Based Discretizations

In this section we define the most used examples of discretization operators, therefore we explicit the decimation operators. Also, the functions G_k and $\tilde{G}_k$ are presented. The choice of the discretization operator allow us to explain the nature of the data, i.e. how the data have been obtained. We begin with the point-value discretization in $[0, 1]$ (in a bounded set Ω is similar).

Let X^N be a uniform partition of $[0, 1]$, then

$$X^N = \{x_j^N\}_{j=0}^{J_N}, \quad x_j^N = jh_N, \quad h_N = 1/J_N, \quad J_N = 2^N J_0,$$

where J_0 is some integer ($J_N < +\infty$). We consider a set of nested dyadic grids $X^k = \{x_j^k\}_{j=0}^{J_k}, k = N-1, \ldots, 0$, as

$$x_j^{k-1} = x_{2j}^k, \quad j = 0, \ldots, J_{k-1} := J_k/2.$$

Therefore, we define the discretization operator as the evaluation of the function in a point of the grid, i.e.

$$\mathscr{D}_k : \mathscr{F} = \mathscr{B}[0,1] \to V^k \quad f_j^k = (\mathscr{D}_k f)_j = f(x_j^k), \quad 0 \le j \le J_k,$$

where $\mathscr{B}[0, 1]$ is the set of bounded real functions in $[0, 1]$ and V^k is the vector space of dimension $J_k + 1$.

As $f_j^{k-1} = f(x_j^{k-1}) = f(x_{2j}^k) = f_{2j}^k$, the decimation operator is defined by:

$$(\mathscr{D}_k^{k-1} f^k)_j = f_{2j}^k, \quad j = 0, \ldots, J_{k-1}.$$

As $e^k \in \mathscr{N}(\mathscr{D}_k^{k-1})$ we can define $(G_k)_{i,j} = \delta_{2i-1,j}$ and $(\tilde{G}_k)_{i,j} = \delta_{i,2j-1}$.

The second discretization presented is based in cell-average. This is used in image compression. Let $f \in \mathscr{L}^1$ be a function discretized by:

$$(\mathscr{D}_k f)_j = \int 2^k \phi(j - 2^k x) f(x) dx, \tag{7}$$

where ϕ is a compactly weighting function. Similarity, $(\mathscr{D}_k f)_j = (\phi_k * f)(x_j^k)$, with $\phi_k(x) = 2^k \phi(2^k x)$. For this example, ϕ is the Haar scaling function

$$\omega_0(x) = \begin{cases} 1, -\frac{1}{2} \le x \le \frac{1}{2}; \\ 0, \text{ in other case.} \end{cases} \tag{8}$$

Therefore the discretization $(\mathscr{D}_k f)_j$ is the average over the cell $c_j^k = (2^{-k}(j - 1), 2^{-k} j)$,

$$(\mathscr{D}_k f)_j = 2^k \int_{c_j^k} f(x) dx. \tag{9}$$

It is easy to prove that $\omega_0(x) = \omega_0(2x) + \omega_0(2x - 1)$, therefore $f_j^{k-1} = \frac{1}{2} f_{2j}^k + \frac{1}{2} f_{2j-1}^k$. Thus, the decimation operator is defined by

$$(\mathscr{D}_k^{k-1} f^k)_j = \frac{1}{2} f_{2j}^k + \frac{1}{2} f_{2j-1}^k. \tag{10}$$

In order to construct the operators G_k and $\tilde{G}_k$ we observe that $\mathscr{D}_k^{k-1} e^k = 0$, then $e_{2j}^k = -e_{2j-1}^k$ and

$$d_j^k = (G_k e^k)_j = e_{2j-1}^k, \quad \begin{cases} e_{2j-1}^k = (\tilde{G}_k d^k)_{2j-1} = d_j^k, \\ e_{2j}^k = (\tilde{G}_k d^k)_{2j} = -d_j^k. \end{cases} \tag{11}$$

Finally, if ϕ is the *hat* function $\omega_0 * \omega_0 = \omega_0^2$, i. e.

$$\omega_0^2(x) = \begin{cases} 1 + x, -1 \le x \le 0, \\ 1 - x, \ 0 \le x \le 1, \\ 0, \qquad \text{in other case,} \end{cases} \tag{12}$$

then the decimation operator $\mathscr{D}_k^{k-1}$ is:

$$\omega_0^2(x) = \frac{1}{2}\omega_0^2(2x - 1) + \omega_0^2(2x) + \frac{1}{2}\omega_0^2(2x + 1)$$

$$(\mathscr{D}_k^{k-1} f^k)_j = \frac{1}{4} f_{2j-1}^k + \frac{1}{2} f_{2j}^k + \frac{1}{4} f_{2j+1}^k.$$

$$
\begin{array}{ccccccccccccc}
 & & & & & & \mathbf{1}f^k_{2j} & & & & & & \\
 & & & & \mathbf{1}f^k_{2j-1} & & & & \mathbf{1}f^k_{2j} & & & & \\
 & & & \mathbf{1}f^k_{2j-1} & & & \mathbf{2}f^k_{2j} & & & \mathbf{1}f^k_{2j+1} & & & \\
 & & \mathbf{1}f^k_{2j-2} & & & \mathbf{3}f^k_{2j-1} & & & \mathbf{3}f^k_{2j} & & & \mathbf{1}f^k_{2j+1} & \\
 & \mathbf{1}f^k_{2j-2} & & & \mathbf{4}f^k_{2j-1} & & & \mathbf{6}f^k_{2j} & & & \mathbf{4}f^k_{2j+1} & & \mathbf{1}f^k_{2j+2} \\
\mathbf{1}f^k_{2j-3} & & & \mathbf{5}f^k_{2j-2} & & & \mathbf{10}f^k_{2j-1} & & & \mathbf{10}f^k_{2j} & & & \mathbf{5}f^k_{2j+1} & & \mathbf{1}f^k_{2j+2}
\end{array}
$$

Fig. 1 Each line $n = 0, \ldots, 5$, represents the decimation, $2^n(\mathscr{D}_k^{k-1} f^k)_j$, when $n = 0$ is point-value, $n = 1$ cell-average and $n = 2$ *hat*-based discretization

Following the same strategy than the before examples, G_k and $\tilde{G}_k$ are defined by:

$$
d_j^k = (G_k e^k)_j = e^k_{2j-1}, \quad \begin{cases} e^k_{2j-1} = (\tilde{G}_k d^k)_{2j-1} = d_j^k, \\ e^k_{2j} = (\tilde{G}_k d^k)_{2j} = -\frac{1}{2}(d_j^k + d^k_{j+1}). \end{cases} \tag{13}
$$

We can generalize these examples using as function ϕ in Eq. (7):

$$
\omega_0^{n+1} = \omega_0^n * \chi_{[-\frac{1}{2},\frac{1}{2}]},\ \omega_0^0 = \delta, \tag{14}
$$

the it is easy to prove that the coefficients, $\{\alpha_l^n\}$, can be computed by the recursive relation:

$$
\alpha_l^{n+1} = \frac{1}{2}(\alpha_l^n + \alpha_{l-1}^n),\ \alpha_l^0 = \delta_{l,0}. \tag{15}
$$

In Fig. 1 the coefficients are calculated. For $n = 0$ we have point-value, $n = 1$ cell-average and $n = 2$ *hat*-based discretizations.

2 Prediction Operators Based on Weighted-Least Squares

We will use the framework developed in [4]. If $f^{k-1} = \{f_j^{k-1}\}_{j=0}^{J_{k-1}}$ are the values of the function in the nodes $x_j^{k-1} = jh_{k-1}$, therefore $f_j^{k-1} = f(x_j^{k-1})$.

For a fitting point x^k_{2j-1}, we estimate a curve based only on the nearest neighborhood determined by a kern function $K_s(x^k_{2j-1}, x_l^{k-1})$. This function assigns a weight to each f_l^{k-1} and this depends on the distance between x_l^{k-1} and x^k_{2j-1}.

The kern functions K_s are indexed by the parameter $s \geq 1$ which is not necessary integer and it indicates the number of data taken in the approximation. In order to simplify the notation we define $\tilde{h}_k = (1 + \epsilon)h_k$, with $0 < \epsilon < 1$. We introduce the parameter ϵ because we want to amplify the bandwidth. Therefore, when s is

an integer we will use $2s$ points. If we do not introduce this term we would use $2(s-1)$, without loss of generality we take $\epsilon = 2 \cdot 10^{-3}$. This determines the bandwidth $(x^k_{2j-1} - (2s-1)\tilde{h}_k, x^k_{2j-1} + (2s-1)\tilde{h}_k)$. We use the data points which are within this band in the approximation (see more details in [4]). Therefore,

$$K_s(x^k_{2j-1}, x) = \omega\left(\frac{x^k_{2j-1} - x}{(2s-1)\tilde{h}_k}\right) \tag{16}$$

where $\omega(u) \geq 0$ is a weight function that assigns largest weights to observations close to x^k_{2j-1}.

We denote as $z(x)$ a polynomial of degree r:

$$z(x) = \sum_{i=0}^{r} \gamma_i x^i = A_r(x)^T \bar{\gamma}_r \tag{17}$$

where $A_r(x) = (1, x, \ldots, x^r)^T$ and $\bar{\gamma}_r = (\gamma_0, \ldots, \gamma_r)^T$. We use the function $L(x, y) = (x-y)^2$ as a loss-function and weight functions such that $\omega(u) = 0, |u| > 1$, we have that

$$K_s(x^k_{2j-1}, x^{k-1}_{j+l}) \neq 0 \text{ if } -\lfloor s \rfloor \leq l \leq \lfloor s \rfloor - 1,$$

where $\lfloor \cdot \rfloor$ is the function that rounds a number to the nearest integer less than or equal to it.

Our problem would be the following:

$$\begin{aligned}
\hat{z}(x) &= \underset{z(x)\in \Pi^r_1(\mathbb{R})}{\arg\min} \sum_{l=0}^{J_{k-1}} K_s(x^k_{2j-1}, x^{k-1}_l) L(f^{k-1}_l, z(x^{k-1}_l)) \\
\hat{\gamma}_r = (\tilde{\gamma}_0, \ldots, \tilde{\gamma}_r) &= \underset{\gamma_i\in\mathbb{R}, i=0,\ldots,r}{\arg\min} \sum_{l=0}^{J_{k-1}} K_s(x^k_{2j-1}, x^{k-1}_l)\left(f^{k-1}_l - \sum_{i=0}^{r} \gamma_i (x^{k-1}_l)^i\right)^2 \\
&= \underset{\gamma_i\in\mathbb{R}, i=0,\ldots,r}{\arg\min} \sum_{l=-\lfloor s \rfloor}^{\lfloor s \rfloor - 1} K_s(x^k_{2j-1}, x^{k-1}_{j+l})\left(f^{k-1}_{j+l} - \sum_{i=0}^{r} \gamma_i (x^{k-1}_{j+l})^i\right)^2
\end{aligned} \tag{18}$$

and we calculate $\hat{z}(x^k_{2j-1}) = \sum_{i=0}^{r} \tilde{\gamma}_i (x^k_{2j-1})^i$, $j = 1, \ldots, J_{k-1}$.

Fixed r, j and s, the problem can be rewritten as:

We denote as $\bar{f}^{k-1} = (f^{k-1}_{j-\lfloor s \rfloor}, \ldots, f^{k-1}_{j+\lfloor s \rfloor - 1})^T$, the matrix $2\lfloor s \rfloor \times (r+1)$

$$\mathbb{X} = \begin{pmatrix} 1 & x_{j-\lfloor s\rfloor}^{k-1} & \dots & (x_{j-\lfloor s\rfloor}^{k-1})^r \\ 1 & x_{j-\lfloor s\rfloor+1}^{k-1} & \dots & (x_{j-\lfloor s\rfloor+1}^{k-1})^r \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{j+\lfloor s\rfloor-1}^{k-1} & \dots & (x_{j+\lfloor s\rfloor-1}^{k-1})^r \end{pmatrix}; \tag{19}$$

and the matrix $2\lfloor s\rfloor \times 2\lfloor s\rfloor$

$$\mathbb{W} = \begin{pmatrix} \omega\big(\frac{x_{2j-1}^k - x_{j-\lfloor s\rfloor}^{k-1}}{(2s-1)\tilde{h}_k}\big) & 0 & \dots & 0 \\ 0 & \omega\big(\frac{x_{2j-1}^k - x_{j-\lfloor s\rfloor+1}^{k-1}}{(2s-1)\tilde{h}_k}\big) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \omega\big(\frac{x_{2j-1}^k - x_{j+\lfloor s\rfloor-1}^{k-1}}{(2s-1)\tilde{h}_k}\big) \end{pmatrix}. \tag{20}$$

Then, our problem is to calculate $\bar{\gamma}_r = (\gamma_0, \dots, \gamma_r)^T$ such that

$$\mathbb{X}^T \mathbb{W} \mathbb{X} \bar{\gamma}_r = \mathbb{X}^T \mathbb{W} \bar{f}^{k-1}, \tag{21}$$

whose solution is

$$\hat{\gamma}_r = (\mathbb{X}^T \mathbb{W} \mathbb{X})^{-1} \mathbb{X}^T \mathbb{W} \bar{f}^{k-1}.$$

Therefore, we have that $\hat{z}(x) = A_r(x)^T \widehat{\gamma}_r$ and

$$\begin{aligned} (\mathscr{P}_{k-1}^k f^{k-1})_{2j-1} = \hat{z}(x_{2j-1}^k) &= A_r(x_{2j-1}^k)^T (\mathbb{X}^T \mathbb{W} \mathbb{X})^{-1} \mathbb{X}^T \mathbb{W} \bar{f}^{k-1} \\ &= \sum_{l=-\lfloor s\rfloor}^{\lfloor s\rfloor-1} L_l(x_{2j-1}^k) f_{j+l}^{k-1} = \sum_{l=1}^{\lfloor s\rfloor} \beta_l (f_{j+l-1}^{k-1} + f_{j-l}^{k-1}) \end{aligned} \tag{22}$$

where $L_l(x_{2j-1}^k) = A_r(x_{2j-1}^k)^T (\mathbb{X}^T \mathbb{W} \mathbb{X})^{-1} \mathbb{X}^T \mathbb{W} e_l$, with $e_l = (\delta_{l,i})_{i=-\lfloor s\rfloor}^{\lfloor s\rfloor-1}$ and $\beta_l = L_{|l|}(x_{2j-1}^k)$, $l = 1, \dots, \lfloor s\rfloor$.

Observe that the prediction operator is linear independently of the weight function chosen, therefore the MR scheme is stable. Also, as the data points are equally spaced, the filters obtained in each MR level k depending on the weight function $\omega(u)$, the bandwidth s (using $2\lfloor s\rfloor$ nodes) and the degree of the polynomials chosen, r. It is not necessary to solve the problem in each point.

In Table 1 the filters obtained using the weight function $\omega(u) = (1 - |u|^2)^3$, $|u| \le 1$ are showed. We denote it as `trwt`.

Table 1 1D filters obtained in MR based on weighted-least squares

trwt	$r = 3$			
	2.5	3.5	4.5	5.5
β_5				−0.0052364
β_4			−0.0111931	−0.0393571
β_3		−0.0279061	−0.0457477	+0.0131481
β_2	−1/16	+0.0212183	+0.1419025	+0.1865625
β_1	+9/16	+0.5066877	+0.4150384	+0.3448828

2.1 Prediction Operator Based on Weight-Least Squares for Cell-Average and Hat-Based Discretizations

We obtain the filters for cell-average MR setting via primitive function (see also [3, 6, 7]). We define it by:

$$F(x) = \int_0^x f(y)dy, \quad f(x) = \frac{d}{dx}F(x).$$

Then, the relation between $\{f_j^k\}$ and $\{F_j^k\}$ is the following:

$$F_j^k = 2^{-k}\sum_{n=0}^{j} f_n^k, \quad f_j^k = 2^k(F_j^k - F_{j-1}^k), \quad j = 1, \ldots, J_k. \tag{23}$$

Therefore we can define the prediction operator for cell-average discretization using the operator defined for point-values of the function $F(x)$ as

$$(\mathscr{P}_{k-1}^k f^{k-1})_j = 2^k\left(\mathscr{A}(x_j^k, F^{k-1}) - \mathscr{A}(x_{j-1}^k, F^{k-1})\right), \tag{24}$$

where $\mathscr{A}(x_\gamma^k, F^{k-1})$ is an approximation in the point x_γ^k, with $\gamma = j-1, j$ using the values $\{F^k\}$.

If we have calculated an approximation using Weighted-Least square in point-values context:

$$\begin{cases} \mathscr{A}(x_{2j-1}^k, F^{k-1}) = \sum_{l=-\lfloor s\rfloor}^{\lfloor s\rfloor-1} \lambda_l F_{j+l}^{k-1}, \\ \mathscr{A}(x_{2j-2}^k, F^{k-1}) = F_{j-1}^{k-1}. \end{cases}$$

where

$$\lambda_{-l} = \lambda_{l-1} = \beta_l, \quad \sum_{l=-\lfloor s\rfloor}^{\lfloor s\rfloor-1} \lambda_l = 1 \quad l = 1, \ldots, \lfloor s\rfloor \tag{25}$$

Then

$$
\begin{aligned}
(\mathscr{P}_{k-1}^{k} f^{k-1})_{2j-1} =& 2^k \left(\mathscr{A}(x_{2j-1}^k, F^{k-1}) - \mathscr{A}(x_{2j-2}^k, F^{k-1}) \right) \\
=& 2^k \Big(\sum_{l=-\lfloor s \rfloor}^{\lfloor s \rfloor - 1} \lambda_l F_{j+l}^{k-1} - F_{j-1}^{k-1} \Big) \\
=& 2 \Bigg(\sum_{l=-\lfloor s \rfloor}^{\lfloor s \rfloor - 1} \lambda_l - 1 \Bigg) \sum_{n=0}^{j-\lfloor s \rfloor} f_n^k + \\
& + 2 \sum_{m=1}^{\lfloor s \rfloor - 1} \Bigg(\sum_{l=-\lfloor s \rfloor + m}^{\lfloor s \rfloor - 1} \lambda_l - 1 \Bigg) f_{j-\lfloor s \rfloor + m}^k + 2 \sum_{m=0}^{\lfloor s \rfloor - 1} \Bigg(\sum_{l=m}^{\lfloor s \rfloor - 1} \lambda_l \Bigg) f_{j+m}^k \\
=& \sum_{l=-\lfloor s \rfloor}^{\lfloor s \rfloor - 1} \hat{\lambda}_l f_{j+l}^k .
\end{aligned}
$$

In Table 2 we can see the relation between filters, with $\lambda_{-l} = -\lambda_l$, with $l = \lfloor s \rfloor - 1$.

In Table 3 the correspondent filters of Table 1 for cell-average discretization are showed.

Table 2 Relation between the filters using point-value and cell-average discretizations

	$r = 4$	$r = 6$	$r = 8$	$r = 10$
$\hat{\lambda}_4$				$2\lambda_4$
$\hat{\lambda}_3$			$2\lambda_3$	$2(\lambda_3 + \lambda_4)$
$\hat{\lambda}_2$		$2\lambda_2$	$2(\lambda_2 + \lambda_3)$	$2(\lambda_2 + \lambda_3 + \lambda_4)$
$\hat{\lambda}_1$	$2\lambda_1$	$2(\lambda_1 + \lambda_2)$	$2(\lambda_1 + \lambda_2 + \lambda_3)$	$2(\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4)$
$\hat{\lambda}_0$	$2(\lambda_0 + \lambda_1)$	$2(\lambda_0 + \lambda_1 + \lambda_2)$	$2(\lambda_0 + \lambda_1 + \lambda_2 + \lambda_3)$	$2(\lambda_0 + \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4)$

Table 3 1D cell-average filters calculated via primitive function

`trwt`	$r = 3$			
	2.5	3.5	4.5	5.5
$\hat{\lambda}_4$				−0.0104728
$\hat{\lambda}_3$			−0.0223863	−0.0891871
$\hat{\lambda}_2$		−0.0558122	−0.1138819	−0.0628907
$\hat{\lambda}_1$	−1/8	−0.0133755	+0.1699231	+0.3102343
$\hat{\lambda}_0$	1	1	1	1

Table 4 Relation between the filters using point-value and hat-based discretizations

	$r = 4$	$r = 6$	$r = 8$	$r = 10$
$\hat{\lambda}_4$				$-8\lambda_4$
$\hat{\lambda}_3$			$-8\lambda_3$	$4(-2\lambda_3 - 4\lambda_4)$
$\hat{\lambda}_2$		$-8\lambda_2$	$4(-2\lambda_2 - 4\lambda_3)$	$4(-2\lambda_2 - 4\lambda_3 - 6\lambda_4)$
$\hat{\lambda}_1$	$-8\lambda_1$	$4(-2\lambda_1 - 4\lambda_2)$	$4(-2\lambda_1 - 4\lambda_2 - 6\lambda_3)$	$4(-2\lambda_1 - 4\lambda_2 - 6\lambda_3 - 8\lambda_4)$

Finally, using the same reasoning, we design the filters for hat-based discretization. For this, we use the following equations:

$$H(x) = \int_0^x \int_0^y f(z)dzdy, \quad f(x) = \frac{d^2}{dx^2} H(x), \tag{26}$$

and relations

$$H_j^k = 4^{-k} \sum_{m=0}^{j-1} \sum_{n=0}^{m} f_n^k, \quad f_j^k = 4^k (H_{j-1}^k - 2H_j^k + H_{j+1}^k), \quad j = 1, \ldots, J_k; \tag{27}$$

The filters showed in Table 4 are calculated using that $(\mathscr{P}_{k-1}^k f^{k-1})_{2j} = 2f_j^{k-1} - \frac{1}{2}(f_{2j-1}^k + f_{2j+1}^k)$ and Eq. (25).

In both cases, cell-average and hat-based discretizations, it is easy to calculate the order of the scheme. Also the MR schemes are stable because of the linearity of the prediction operators obtained.

Acknowledgements This research was partially supported by Spanish MCINN MTM 2011-22741.

References

1. Amat, S., Donat, R., Liandrat, J., Trillo, J.C.: A fully adaptive PPH multi-resolution scheme for image processing. Math. Comput. Model. **46**, 2–11 (2001)
2. Aràndiga, F., Cohen, A., Yáñez, D.F.: Learning-based multiresolution transforms with application to image compression. Signal Process. **93**, 2474–2484 (2012)
3. Aràndiga, F., Donat, R.: Nonlinear multiscale descompositions: the approach of A. Harten. Numer. Algorithms **23**, 175–216 (2000)
4. Aràndiga, F., Yáñez, D.F.: Generalized wavelets design using Kernel methods. Application to signal processing. J. Comput. Appl. Math. **250**, 1–15 (2013)
5. Burt, P.J., Adelson, E.H.: The Laplacian pyramid as a compact image code. IEEE Trans. Commun. **31**, 532–540 (1983)
6. Getreuer, P., Meyer, F.: ENO multiresolution schemes with general discretizations. SIAM J. Numer. Anal. **46**, 2953–2977 (2008)

7. Harten, A.: Multiresolution representation of data: general framework. SIAM J. Numer. Anal. **33**, 1205–1256 (1996)
8. Mallat, S.: A Wavelet Tour of Signal Processing. Academic, New York (1999)
9. Sweldens, W.: The lifting scheme: a construction of second generation wavelets. SIAM J. Numer. Anal. **29**, 511–546 (1998)

Signal Denoising with Harten's Multiresolution Using Interpolation and Least Squares Fitting

Francesc Aràndiga and José Jaime Noguera

Abstract Harten's multiresolution has been successfully applied to the signal compression using interpolatory reconstructions with nonlinear techniques. Here we study the applicability of these techniques to remove noise to piecewise smooth signals. We use two reconstruction types: interpolatory and least squares, and we introduce ENO and SR nonlinear techniques. The standard methods adaptation to noisy signals and the comparative of the different schemes are the subject of this paper.

1 Introduction: Harten's Multiresolution

Multiscale decompositions are efficient tools for analyzing the information contained in a signal, providing various applications such as signal compression and denoising. If f^L represents a sampling of a signal, $f(x)$, in the finest resolution level L, the multiresolution schemes rearrange this information leading to the decomposition $\{f^0, e^1, e^2, \ldots, e^L\}$, where f^0 corresponds to the sampling at the coarsest resolution level and each sequence e^k represents the information which is necessary to recover f^k from f^{k-1}. If $f(x)$ is smooth the details e^k have small magnitude and we can remove them without a great loss of information, providing excellent compression capabilities. Harten introduces its notion of multiresolution in [5] and later generalizes it in [6, 7]. We start revising the basic aspects of this formulation.

Let us consider V^{k+1} to be a $k+1$ dimension linear space. The *discretization* operator, $\mathrm{D}_{k+1} : \mathrm{F} \to V^{k+1}$, allows to obtain the discrete values of a function, $f \in \mathrm{F}$, while the *reconstruction* operator, $\mathrm{R}_{k+1} : V^{k+1} \to \mathrm{F}$, performs the reverse operation, and it must satisfy the following property:

$$\mathrm{D}_{k+1}\mathrm{R}_{k+1} = I_{V^{k+1}}. \tag{1}$$

F. Aràndiga (✉) • J.J. Noguera
Departamento de Matemática Aplicada, Universitat de València, C/ Doctor Moliner, 46100 Burjassot, Valencia, Spain
e-mail: arandiga@uv.es; jnoguera373b@cv.gva.es

F. Casas, V. Martínez (eds.), *Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 4, DOI 10.1007/978-3-319-06953-1_14

The reconstruction operator can be nonlinear. This allows the introduction of techniques that improve the approximation in the presence of discontinuities (see [1]), being this a fundamental difference from linear multiscale decompositions, such as the *wavelet transform* (see [3]).

The connection between two levels of resolution (larger k, higher resolution) is given by two operators: *decimation*, $D_{k+1}^k : V^{k+1} \to V^k$ and *prediction*, $P_k^{k+1} : V^k \to V^{k+1}$, that must satisfy the *consistency* requirement: $D_{k+1}^k P_k^{k+1} = I_{V^k}$. However, for the inverse composition $P_k^{k+1} D_{k+1}^k \neq I_{V^{k+1}}$. Then, we define the *prediction error*: $e^{k+1} = (I_{V^{k+1}} - P_k^{k+1} D_{k+1}^k) v^{k+1}$, being $v^{k+1} \in V^{k+1}$. If $\{\mu_i^{k+1}\}$ is a basis of the null space of D_{k+1}^k, we can express $e^{k+1} = \sum_i d_i^{k+1} \mu_i^{k+1}$ (see [1]). We call d_i^{k+1} the scale coefficients at level k.

Finally, decimation operators can be constructed from a sequence of discretization operators, provided they are nested (see [1]):

$$\mathrm{D}_{k+1} f = 0 \Rightarrow \mathrm{D}_k f = 0, \qquad \forall k \in \mathbb{N}, \quad \forall f \in \mathrm{F}. \tag{2}$$

The most commonly used discretizations are point-value and cell average, [1]. Since our goal is noise removal and it is eliminated naturally by cell average decimation, this will be used in what follows.

1.1 Cell Average Discretization in [0,1]

Consider a set of nested dyadic grids defined in [0,1]:

$$X^k = \{x_i^k\}_{i=0}^{N_k}, \quad N_k = 2^k N_0, \quad x_i^k = i h_k, \quad h_k = \tfrac{1}{N_k}, \quad k = 0, \ldots, L, \tag{3}$$

where $N_0 \in \mathbb{N}$. If $\mathrm{F} = L^1([0,1])$, the cell average discretization operator $\mathrm{D}_{k+1} : \mathrm{F} \to V^{k+1}$, is defined in [7] as:

$$\bar{f}_i^{k+1} := (\mathrm{D}_{k+1} f)_i = \frac{1}{h_{k+1}} \int_{x_{i-1}^{k+1}}^{x_i^{k+1}} f(x) dx, \quad 1 \le i \le N_{k+1}. \tag{4}$$

By integral properties is easy to see that $\bar{f}_i^k = \frac{1}{2}(\bar{f}_{2i}^{k+1} + \bar{f}_{2i-1}^{k+1})$, defining this way the decimation operator and satisfying (2).

2 Interpolatory Reconstruction for Cell Averages

We define the r-th order interpolatory reconstruction as:

$$\overline{\mathrm{IC}}_{nl,nr}^{r-1}(x, \bar{f}^k) = g_i(x), \quad x \in [x_{i-1}^k, x_i^k], \quad i = 1, \ldots, N_k, \tag{5}$$

where $g_i(x)$ is the polynomial of degree $r-1 = nl + nr$ such that:

$$\frac{1}{h_k}\int_{x^k_{i+s-1}}^{x^k_{i+s}} g_i(x)dx = \bar{f}^k_{i+s}, \qquad s = -nl, \ldots, nr, \quad nl, nr \in \mathbb{N}. \tag{6}$$

Prediction operator is calculated as follows:

$$(P_k^{k+1}\bar{f}^k)_{2i-j} = \left(\mathrm{D}_{k+1}(\overline{\mathrm{IC}}^{r-1}_{nl,nr}(x;\bar{f}^k))\right)_{2i-j} = \frac{1}{h_{k+1}}\int_{x^{k+1}_{2i-j-1}}^{x^{k+1}_{2i-j}} \overline{\mathrm{IC}}^{r-1}_{nl,nr}(x;\bar{f}^k)dx,$$

where $j = 0, 1$. Therefore $\frac{1}{2}((P_k^{k+1}\bar{f}^k)_{2i-1} + (P_k^{k+1}\bar{f}^k)_{2i}) = \bar{f}^k_i$, satisfying (1). Also $e^{k+1}_{2i-1} + e^{k+1}_{2i} = 0$ and we can define $d^{k+1}_i = e^{k+1}_{2i-1}$, (see [1]).

Then, the multiresolution scheme is:

Codification, $\bar{f}^L \to \{\bar{f}^0, d^1, d^2, \ldots, d^L\}$ (Direct Transformation):

$$\begin{cases} \textbf{For } k = L-1, \ldots, 0 \\ \quad \bar{f}^k_i = \frac{1}{2}(\bar{f}^{k+1}_{2i-1} + \bar{f}^{k+1}_{2i}), & i = 1, \ldots, N_k, \\ \quad d^{k+1}_i = \bar{f}^{k+1}_{2i-1} - (P_k^{k+1}\bar{f}^k)_{2i-1}, & i = 1, \ldots, N_k. \end{cases} \tag{7}$$

Decodification, $\{\bar{f}^0, d^1, d^2, \ldots, d^L\} \to \bar{f}^L$ (Inverse Transformation):

$$\begin{cases} \textbf{For } k = 0, \ldots, L-1 \\ \quad \bar{f}^{k+1}_{2i-1} = (P_k^{k+1}\bar{f}^k)_{2i-1} + d^{k+1}_i, & i = 1, \ldots, N_k, \\ \quad \bar{f}^{k+1}_{2i} = 2\bar{f}^k_i - \bar{f}^{k+1}_{2i-1} \equiv (P_k^{k+1}\bar{f}^k)_{2i} - d^{k+1}_i, & i = 1, \ldots, N_k. \end{cases} \tag{8}$$

2.1 Nonlinear Techniques

Nonlinear techniques help us to improve the reconstructions in the presence of discontinuities. Here, we will apply ENO and SR techniques.

2.1.1 ENO Technique

The ENO (*Essentially Non-Oscillatory*, [8]) interpolation technique consists in choose for each interval, a stencil that do not cross a discontinuity. This is possible if the working interval contains no discontinuities. Typically stencil is selected according to the magnitude of the divided differences. However, if the function is contaminated by noise, the information provided by divided differences is unreliable

and we must seek an alternative. Inspired by [9] we use a choice that is not affected by the presence of noise. If $m = nl + nr + 1$, we define the measure:

$$\bar{E}_2(x_i, m, l) = \sum_{j=1}^{m} \left(\bar{q}_i^{\overline{\mathrm{LS}}_{m-l,l-1}}(x^k_{i-(m-l)+j-1}; \bar{f}^k, s) - \bar{f}^k_{i-(m-l)+j-1} \right)^2, \tag{9}$$

where $\bar{q}_i^{\overline{\mathrm{LS}}_{nl,nr}}(x; \bar{f}^k, s)$ is the cell averages least squares polynomial of degree $s - 1 < nl + nr$ constructed from the stencil $\{x^k_{i-nl}, \ldots, x^k_{i+nr}\}$.

Now, we take:

$$\bar{E}_2(x_i, m, l^*) = min\left\{\bar{E}_2(x_i, m, 1), \bar{E}_2(x_i, m, 2), \ldots, \bar{E}_2(x_i, m, m)\right\}, \tag{10}$$

and the ENO stencil for $I^k_i = (x^k_{i-1}, x^k_i)$ is $\left\{x^k_{i-nl_i}, \ldots, x^k_{i+nr_i}\right\}$, with:

$$nl_i := m - l^*, \qquad nr_i := l^* - 1. \tag{11}$$

2.1.2 SR Technique

With the technique SR (*Subcell Resolution*, [2,4]) we can improve the approximation even in the interval containing the discontinuity. The idea is to properly extend the adjacent interpolating polynomials to the point of discontinuity.

First, we define some useful concepts. If $f(x)$ has a jump in $[x^k_{i-1}, x^k_i]$ the primitive function of f, $F(x) = \int_0^x f(y)dy \in \mathrm{C}([0,1])$ has a corner (a discontinuity in the derivative) there. Note that the sets $\{\bar{f}^k_i\}_{i=1}^{N_k}$ and $F^k = \{F^k_i\}_{i=0}^{N_k}$ are equivalents due to the relations $F^k_i = F(x^k_i) = \int_0^{x^k_i} f(y)dy = h_k \sum_{j=1}^{i} \bar{f}^k_j$ and $\bar{f}^k_j = \frac{1}{h_k}(F^k_j - F^k_{j-1})$.

If $m = nl + nr + 1$, the SR technique is summarized as follows:

1. Taking stencils with m nodes, we calculate the ENO stencils by (11).
2. If $nl_{i-1} = m - 1$ and $nl_{i+1} = 0$ the stencils for the cells I^k_{i-1} and I^k_{i+1} are disjoint. We label the cell I^k_i as *suspect* of containing a discontinuity.
3. For each suspicious cell we define the function

$$G_i^{\overline{\mathrm{IC}}}(x) = q^{\overline{\mathrm{IP}}}_{i+1,0,m-1}(x; F^k, r) - q^{\overline{\mathrm{IP}}}_{i-1,m-1,0}(x; F^k, r), \tag{12}$$

where $q^{\overline{\mathrm{IP}}}_{j,nl,nr}(x; F^k, s)$ is the polynomial of degree s that interpolates the point values (x^k_l, F^k_l), $j - nl - 1 \le l \le j + nr$.

If $G_i^{\overline{\mathrm{IC}}}(x^k_{i-1}) \cdot G_i^{\overline{\mathrm{IC}}}(x^k_i) < 0$ we label the cell I^k_i as *singular*.

4. If $G_i^{\overline{\mathrm{IC}}}(x_{i-1}^k)\cdot G_i^{\overline{\mathrm{IC}}}(x_{2i-1}^{k+1}) < 0$, the node x_{2i-1}^{k+1} lies at the right of the discontinuity. Then the predicted values are obtained as follows:

$$(P_k^{k+1}\bar{f}^k)_{2i} = \frac{q_{i+1,0,m-1}^{\overline{\mathrm{IP}}}(x_{2i}^{k+1};F^k,r) - q_{i+1,0,m-1}^{\overline{\mathrm{IP}}}(x_{2i-1}^{k+1};F^k,r)}{h_{k+1}}, \tag{13}$$

$$(P_k^{k+1}\bar{f}^k)_{2i-1} = 2\bar{f}_i^k - (P_k^{k+1}\bar{f}^k)_{2i}. \tag{14}$$

In the other case, x_{2i-1}^{k+1} is located at the left of the discontinuity and:

$$(P_k^{k+1}\bar{f}^k)_{2i-1} = \frac{q_{i-1,m-1,0}^{\overline{\mathrm{IP}}}(x_{2i-1}^{k+1};F^k,r) - q_{i-1,m-1,0}^{\overline{\mathrm{IP}}}(x_{2i-2}^{k+1};F^k,r)}{h_{k+1}}, \tag{15}$$

$$(P_k^{k+1}\bar{f}^k)_{2i} = 2\bar{f}_i^k - (P_k^{k+1}\bar{f}^k)_{2i-1}. \tag{16}$$

3 Least Squares Reconstruction for Cell Averages

Schemes in this case are similar to those discussed in Sect. 2 but now we use least squares fitting instead interpolation fitting. We define the r-th order least squares reconstruction for cell averages as:

$$\overline{\mathrm{LSC}}_{nl,nr}^{r-1}(x,\bar{f}^k) = g_i(x), \quad x\in[x_{i-1}^k, x_i^k], \quad i=1,\ldots,N_k, \tag{17}$$

where $g_i(x)$ is the polynomial of degree $r-1 < nl+nr$ such that:

$$\frac{1}{h_k}\int_{x_{i+s-1}^k}^{x_{i+s}^k} g_i(x)dx = \bar{f}_{i+s}^k, \qquad s=-nl,\ldots,nr, \quad nl,nr\in\mathbb{N}. \tag{18}$$

The prediction operator is calculated as follows:

$$(P_k^{k+1}\bar{f}^k)_{2i-j} = \left(\mathrm{D}_{k+1}(\overline{\mathrm{LSC}}_{nl,nr}^{r-1}(x;\bar{f}^k))\right)_{2i-j} = \frac{1}{h_{k+1}}\int_{x_{2i-j-1}^{k+1}}^{x_{2i-j}^{k+1}} \overline{\mathrm{LSC}}_{nl,nr}^{r-1}(x;\bar{f}^k)dx,$$

with $j=0,1$. Note that we don't have any interpolation condition and (1) is not fulfilled because of $\frac{1}{2}((P_k^{k+1}\bar{f}^k)_{2i-1} + (P_k^{k+1}\bar{f}^k)_{2i}) \neq \bar{f}_i^k$. At this point we suggest two options for (8):

- Forcing consistency: $\bar{f}_{2i}^{k+1} = 2\bar{f}_i^k - \bar{f}_{2i-1}^{k+1}$. We denote it as $\overline{\mathrm{LSC}-C}$.
- Losing consistency: $\bar{f}_{2i}^{k+1} = (P_k^{k+1}\bar{f}^k)_{2i} - d_i^{k+1}$. We denote it as $\overline{\mathrm{LSC}-NC}$.

3.1 Nonlinear Techniques

We can apply the nonlinear techniques similarly to the exposed in Sect. 2.1, but considering the following adaptations:

- In (9), $s \leq r$ is allowed.
- The G function, (12), in this case is defined as follows:

$$G_i^{\overline{\mathrm{LSC}}}(x) = q_{i+1,0,m-1}^{\overline{\mathrm{LSP}}}(x; F^k, r) - q_{i-1,m-1,0}^{\overline{\mathrm{LSP}}}(x; F^k, r), \tag{19}$$

where $q_{j,nl,nr}^{\overline{\mathrm{LSP}}}(x; F^k, s)$ is the s-th degree polynomial that approximates in the least squares sense to the point values (x_l^k, F_l^k) for $j - nl - 1 \leq l \leq j + nr$. Since there are no interpolatory conditions, we can't express $G_i^{\overline{\mathrm{LSC}}}$ in terms of $\{\bar{f}_i^k\}$. To apply the SR technique for this reconstruction, we use the function $G_i^{\overline{\mathrm{IC}}}(x)$ to decide whether we are facing a singular cell. Obviously we need to use in $G_i^{\overline{\mathrm{IC}}}(x)$ polynomials with the same length that $\overline{\mathrm{LSC}}$.

4 Numerical Experiments

In this section we present some numerical experiments for denoising applying the reconstructions studied in this paper.

We define the function:

$$g(x) = \begin{cases} -\frac{4x-3}{5} sin(\frac{3}{2}\pi(\frac{4x-3}{5})^2) & \text{if } 0 \leq x < \frac{3\pi}{29}, \\ |sin2\pi(\frac{4x-3}{5}) + \frac{\pi}{1000}| & \text{if } \frac{3\pi}{29} \leq x \leq 1. \end{cases} \tag{20}$$

The work function is of the following type: $f(x) = g(x) + n(x)$, where $g(x)$ is defined in (20) and $n(x)$ is some white Gaussian noise.[1] To measure the noise of the signal, we consider the *Signal-to-Noise Ratio*, expressed in dB: $SNR(g, f) := 10log_{10}(\sum_{i=1}^{N} g_i^2 / \sum_{i=1}^{N}(g_i - f_i)^2)$, where N is the signal length.

The experiment consists of: we fix SNR = 25 dB, and we consider a discretization with $N_L = 2^{6+L}$ nodes, obtaining $\{\bar{f}_i^L\}_{i=1}^{N_L}$. First we decimate L levels for cell averages to get $\{\bar{f}_i^0\}_{i=1}^{64}$. Then we apply L levels of an inverse transform (with $d_i^k = 0\ \forall i, k$) using different reconstructions, obtaining $\{\hat{f}_i^L\}_{i=1}^{N_L}$. For evaluate the denoising goodness we use the *Root Mean Squared Error*:

$$RMSE(\hat{f}^L, \bar{g}^L) = \sqrt{\frac{1}{N_L}\sum_{i=1}^{N_L}(\hat{f}_i^L - \bar{g}_i^L)^2}. \tag{21}$$

[1] Generated using the function *awgn* of MATLAB ®.

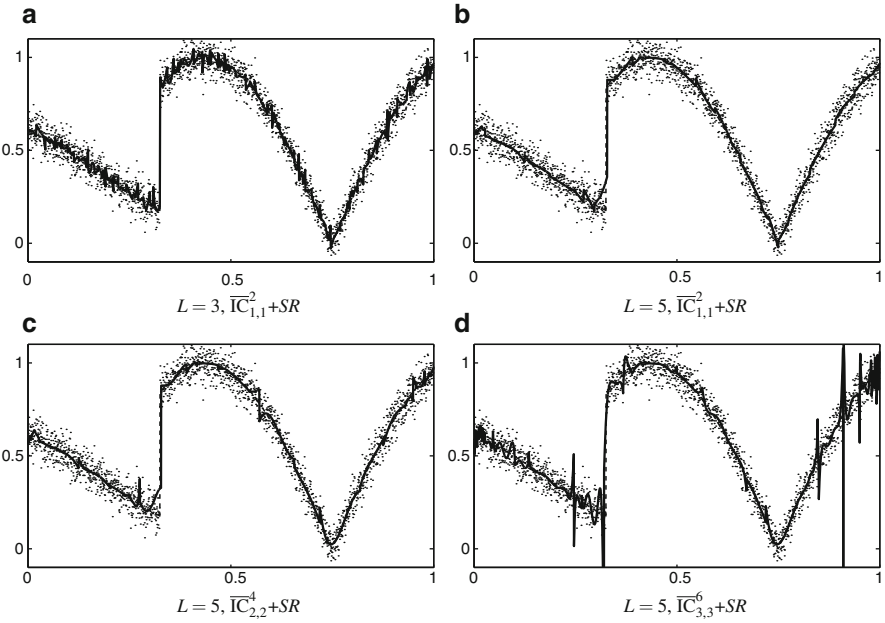


Fig. 1 Denoising with $\overline{\mathrm{IC}}$. Errors: (**a**) RMSE = 0.02196; (**b**) RMSE = 0.03206; (**c**) RMSE = 0.03526; (**d**) RMSE = 0.07537

In Fig. 1 we can see the results that we obtain with $\overline{\mathrm{IC}} + SR$ reconstruction. In (a) ($L = 3$) we see that the noise removal is poor, due to we use few levels. *Gibbs phenomenon* does not appear, thereby nonlinear techniques achieve their objective. In (b) we increase the number of levels, $L = 5$, obtaining an efficient noise removal. In Fig. 1c, d we show what happens if we raise the degree of $\overline{\mathrm{IC}}$. We can see that the results are worse because we are using high degree polynomials with noisy data and we obtain values with lower smoothness. The oscillations are amplified by raising levels. Then we conclude that the degree of $\overline{\mathrm{IC}}$ must be low.

In Fig. 2 we use least squares reconstruction for cell averages with SR technique. In (a) we use $\overline{\mathrm{LSC} - C}$ and, as expected, we do not get good results because by forcing consistency we create oscillations which are transmitted and extended to higher levels. However, if we use $\overline{\mathrm{LSC} - NC}$, with the same parameters, we obtain smoother results (Fig. 2b). There is a problem with the non consistency: we lose the connection between two consecutive levels and we could lose the correct location of discontinuities, as we can see in Fig. 2c. Nevertheless there is an advantage with respect to $\overline{\mathrm{IC}}$ reconstruction: we can see in (d), only with three levels, that we achieve remove noise efficiently. This fact is confirmed by the RMSE (shown in the figure legend). For that, we need to use longer stencils and also we improve the reliability of discontinuity detection. Remember that we can't do it for $\overline{\mathrm{IC}}$ because it involves an increase of the degree and therefore worst reconstructions. Also using few levels we reduce the computational cost and the effects of non consistency.

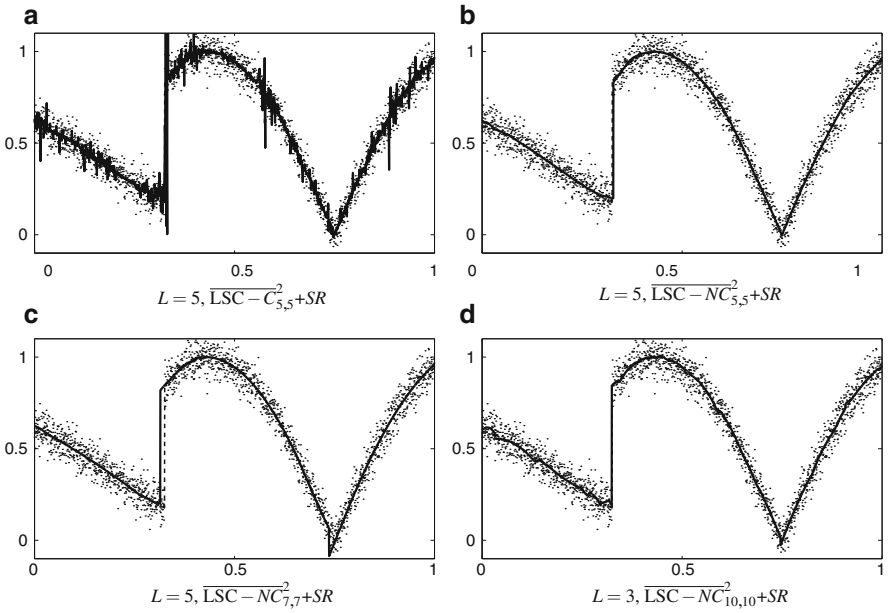


Fig. 2 Denoising with $\overline{\mathrm{LS}}$. Errors: (**a**) RMSE = 0.07252; (**b**) RMSE = 0.03560; (**c**) RMSE = 0.06847; (**d**) RMSE = 0.02271

Conclusions

We have studied the applicability of the Harten's multiresolution with non-linear techniques (ENO and SR) to the signal denoise, obtaining adaptations to the standard schemes (using $\bar{E}_2$ instead of divided differences for locating discontinuities). We have used two reconstructions types: interpolatory and least squares, and the latter with some adaptations (consistent and non consistent) to improve the denoise.

Based on our numerical experiments we can conclude that with the $\overline{\mathrm{IC}}$ reconstruction we can remove efficiently noise and for it we must use low degrees and high levels. If we use $\overline{\mathrm{LSC}}$ reconstruction we must use a non consistent version, causing that we lose the exact discontinuity position. However, in some cases this may be advantageous over the interpolatory reconstruction. For example, if there are insufficient number of initial data to apply a large number of levels we can eliminate a significant amount of noise using few levels of $\overline{\mathrm{LSC}}$ reconstruction.

As future work, we plan to design consistent reconstructions combining interpolation and least squares, in order to take advantage of both reconstructions.

References

1. Aràndiga, F., Donat, R.: Nonlinear multiscale decompositions: the approach of A. Harten. Numer. Algorithms. **23**, 175–216 (2000)
2. Aràndiga, F., Cohen, A., Donat, R., Dyn, N.: Interpolation and approximation of piecewise smooth functions. SIAM J. Numer. Anal. **43**(1), 41–57 (2005)
3. Daubechies, I.: Ten lectures on Wavelets. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia (1993)
4. Harten, A.: ENO schemes with subcell resolution. J. Comput. Phys. **83**, 148–184 (1989)
5. Harten, A.: Discrete multiresolution analysis and generalized wavelets. J. Appl. Numer. Math. **12**, 153–192 (1993)
6. Harten, A.: Multiresolution representation of data II. Technical report, UCLA CAM report, n. 93–13 (1993)
7. Harten, A.: Multiresolution representation of data: a general framework. SIAM J. Numer. Anal. **33**(3), 1205–1256 (1996)
8. Harten, A., Engquist, B., Osher, S., Chakravarthy, S.: Uniformly high order accurate essentially non-oscillatory schemes III. J. Comput. Phys. **71**, 231–303 (1987)
9. Mizrachi, D.: Remoivong noise from discontinous data. PhD. thesis, School of Mathematical Sciences, Tel-Aviv University (1991)

The Wavelet Scalogram in the Study of Time Series

Vicente J. Bolós and Rafael Benítez

Abstract Wavelet theory has been proved to be a useful tool in the study of time series. Specifically, the scalogram allows the detection of the most representative scales (or frequencies) of a signal. In this work, we present the scalogram as a tool for studying some aspects of a given signal. Firstly, we introduce a parameter called *scale index*, interpreted as a measure of the degree of the signal's non-periodicity. In this way, it can complement the maximal Lyapunov exponent method for determining chaos transitions of a given dynamical system. Secondly, we introduce a method for comparing different scalograms. This can be applied for determining if two time series follow similar patterns.

1 Introduction

A *wavelet function* (or wavelet, for short), is a function $\psi \in L^2(\mathbb{R})$ with zero average (i.e. $\int_{\mathbb{R}} \psi = 0$), normalized (i.e. $\|\psi\| = 1$), and *centered* in the neighborhood of $t = 0$ (see [1] for other properties). Scaling ψ by a positive quantity s, and translating it by $u \in \mathbb{R}$, we define a family of *time-frequency atoms*, $\psi_{u,s}$, as

$$\psi_{u,s}(t) := \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right), \qquad u \in \mathbb{R}, s > 0. \tag{1}$$

V.J. Bolós (✉)
Facultad de Economía, Departamento de Matemáticas para la Economía y la Empresa, Universidad de Valencia, Avda. Tarongers s/n, 46022 Valencia, Spain
e-mail: vbolos@uv.es

R. Benítez
Departamento de Matemáticas, Centro Universitario de Plasencia, Universidad de Extremadura, Avda. Virgen del Puerto 2, 10600 Plasencia, Spain
e-mail: rbenitez@unex.es

F. Casas, V. Martínez (eds.), *Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 4, DOI 10.1007/978-3-319-06953-1_15

Given $f \in L^2(\mathbb{R})$, the *continuous wavelet transform* (CWT) of f at time u and scale s is defined as

$$Wf(u,s) := \langle f, \psi_{u,s} \rangle = \int_{-\infty}^{+\infty} f(t)\psi^*_{u,s}(t)\mathrm{d}t, \tag{2}$$

and it provides the frequency component (or *details*) of f corresponding to the scale s and time location u.

The revolution of wavelet theory comes precisely from this fact: the two parameters (time u and scale s) of the CWT in (2) make possible the study of a signal in both domains (time and frequency) simultaneously, with a resolution that depends on the scale of interest. According to these considerations, the CWT provides a time-frequency decomposition of f in the so called *time-frequency plane* [2, Figure 1]. This method, as it is discussed in [3], is more accurate and efficient than other techniques such as the *windowed Fourier transform* (WFT).

The *scalogram* of f is defined by the function

$$\mathscr{S}(s) := \|Wf(s,u)\| = \left(\int_{-\infty}^{+\infty} |Wf(s,u)|^2 \mathrm{d}u\right)^{\frac{1}{2}}, \tag{3}$$

representing the *energy* of Wf at a scale s. Obviously, $\mathscr{S}(s) \geq 0$ for all scale s, and if $\mathscr{S}(s) > 0$ we will say that the signal f has details at scale s. Thus, the scalogram allows the detection of the most representative scales (or frequencies) of a signal, that is, the scales that contribute the most to the total energy of the signal.

If we are only interested in a given time interval $[t_0, t_1]$, we can define the corresponding *windowed scalogram* by

$$\mathscr{S}_{[t_0,t_1]}(s) := \|Wf(s,u)\|_{[t_0,t_1]} = \left(\int_{t_0}^{t_1} |Wf(s,u)|^2 \mathrm{d}u\right)^{\frac{1}{2}}. \tag{4}$$

2 Analysis of Compactly Supported Discrete Signals

In practice, to make a signal f suitable for a numerical study, we have to

(i) Consider that it is defined over a finite time interval $I = [a, b]$, and
(ii) Sample it to get a discrete set of data.

Regarding the first point, boundary problems arise if the support of $\psi_{u,s}$ overlaps $t = a$ or $t = b$. There are several methods for avoiding these problems, like using *periodic wavelets*, *folded wavelets* or *boundary wavelets* (see [1]); however, these methods either produce large amplitude coefficients at the boundary or complicate the calculations. So, if the wavelet function ψ is compactly supported and the

interval I is big enough, the simplest solution is to study only those wavelet coefficients that are not affected by boundary effects.

Taking into account the considerations mentioned above, the *inner scalogram* of f at a scale s is defined by

$$\mathscr{S}^{\text{inner}}(s) := \mathscr{S}_{J(s)}(s) = \|Wf(s,u)\|_{J(s)} = \left(\int_{c(s)}^{d(s)} |Wf(s,u)|^2 \mathrm{d}u\right)^{\frac{1}{2}}, \tag{5}$$

where $J(s) = [c(s), d(s)] \subseteq I$ is the maximal subinterval in I for which the support of $\psi_{u,s}$ is included in I for all $u \in J(s)$. Obviously, the length of I must be big enough for $J(s)$ not to be empty or too small, i.e. $b - a \gg sl$, where l is the length of the support of ψ.

Since the length of $J(s)$ depends on the scale s, the values of the inner scalogram at different scales cannot be compared. To avoid this problem, we can *normalize* the inner scalogram:

$$\overline{\mathscr{S}}^{\text{inner}}(s) = \frac{\mathscr{S}^{\text{inner}}(s)}{(d(s) - c(s))^{\frac{1}{2}}}. \tag{6}$$

With respect to the sampling of the signal, any discrete signal can be analyzed in a *continuous way* using a piecewise constant interpolation. In this way, the CWT provides a scalogram with a better resolution than the discrete wavelet transform (DWT), that considers dyadic levels instead of continuous scales (see [1]).

3 The Scale Index

Although there is no universally accepted definition of chaos, a bounded signal is considered chaotic if (see [4])

(a) It shows sensitive dependence on the initial conditions, and
(b1) It is non-periodic, or
(b2) It does not *converge* to a periodic orbit.

Usually, chaos transitions in bifurcation diagrams are numerically detected by means of the Maximal Lyapunov Exponent (MLE). Roughly speaking, Lyapunov exponents characterize the rate of separation of initially nearby orbits and a system is thus considered chaotic if the MLE is positive. Therefore, the MLE technique is focused on the sensitivity to initial conditions, i.e. on criterion (a).

As to criteria (b1) and (b2), Fourier analysis can be used for studying non-periodicity. However chaotic signals may be highly non-stationary, which makes wavelets more suitable (as it is discussed in [5]).

We are going to introduce a new parameter, the *scale index*, that will give us information about the degree of non-periodicity of a signal. To this end we are going

to state first some results for the wavelet analysis of periodic functions (for further reading please refer to [1] and references therein).

The next theorem gives us a criterion for distinguishing between periodic and non-periodic signals. It ensures that if a signal f has details at every scale (i.e. the scalogram of f does not vanish at any scale), then it is non-periodic.

Theorem 1 *Let $f : \mathbb{R} \to \mathbb{C}$ be a T-periodic function in $L^2([0,T])$, and let ψ be a compactly supported wavelet. Then $Wf(u,2T) = 0$ for all $u \in \mathbb{R}$.*

Note that if $f : \mathbb{R} \to \mathbb{C}$ is a T-periodic function in $L^2([0,T])$, and ψ is a compactly supported wavelet, then $Wf(u,s)$ is well-defined for $u \in \mathbb{R}$ and $s \in \mathbb{R}^+$, although f is not in $L^2(\mathbb{R})$. For a detailed proof see [2].

From this result we obtain the following corollary.

Corollary 1 *Let $f : I = [a,b] \to \mathbb{C}$ a T-periodic function in $L^2([a,a+T])$. If ψ is a compactly supported wavelet, then the (normalized) inner scalogram of f at scale $2T$ is zero.*

These results constitute a valuable tool for detecting periodic and non-periodic signals, because a signal with details at every scale must be non-periodic (see [2, Figure 2]). Note that in order to detect numerically wether a signal *tends to be periodic*, we have to analyze its scalogram throughout a relatively wide time range.

Moreover, since the scalogram of a T-periodic signal vanishes at all $2kT$ scales (for all $k \in \mathbb{N}$), it is sufficient to analyze only scales greater than a fundamental scale s_0. Thus, a signal which has details at an arbitrarily large scale is non-periodic.

In practice, we shall only study the scalogram on a finite interval $[s_0, s_1]$. The most representative scale of a signal f will be the scale $s_{\max}$ for which the scalogram reaches its maximum value. If the scalogram $\mathscr{S}(s)$ never becomes too small compared to $\mathscr{S}(s_{\max})$ for $s > s_{\max}$, then the signal is "numerically non-periodic" in $[s_0, s_1]$.

Taking into account these considerations, we will define the *scale index* of f in the scale interval $[s_0, s_1]$ as the quotient

$$i_{\text{scale}} := \frac{\mathscr{S}(s_{\min})}{\mathscr{S}(s_{\max})}, \tag{7}$$

where $s_{\max}$ is the smallest scale such that $\mathscr{S}(s) \leq \mathscr{S}(s_{\max})$ for all $s \in [s_0, s_1]$, and $s_{\min}$ the smallest scale such that $\mathscr{S}(s_{\min}) \leq \mathscr{S}(s)$ for all $s \in [s_{max}, s_1]$. Note that for compactly supported signals only the normalized inner scalogram will be considered.

From its definition, the scale index i_{scale} is such that $0 \leq i_{\text{scale}} \leq 1$ and it can be interpreted as a measure of the degree of non-periodicity of the signal: the scale index will be zero (or numerically close to zero) for periodic signals and close to one for highly non-periodic signals.

The selection of the scale interval $[s_0, s_1]$ is an important issue in the scalogram analysis. Since the non-periodic character of a signal is given by its behavior at large scales, there is no need for s_0 to be very small. In general, we can choose s_0 such that $s_{\max} = s_0 + \epsilon$ where ϵ is positive and close to zero. On the other hand, s_1 should be large enough for detecting significant periodicities. But as s_1 increases, so does the computational cost. In fact, the larger s_1 is, the wider the time span should be where the signal is analyzed, in order to maintain the accuracy of the normalized inner scalogram.

Scales $s_{\min}$ and $s_{\max}$ determine the pattern that the scalogram follows. For example, in non-periodic signals $s_{\min}$ can be regarded as the "least non-periodic scale". Moreover, if $s_{\min} \simeq s_1$, then the scalogram decreases at large scales and s_1 should be increased in order to distinguish between a non-periodic signal and a periodic signal with a very large period.

In [2] there is a complete study of the scale index versus MLE and bifurcation diagrams in some particular chaotic systems: the forced Bonhoeffer-van der Pol (BvP) oscillator, the Henon map, and the logistic map. It is shown that there is a correspondence between the chaotic regions of the bifurcation diagram, the regions where the MLE is positive, and the regions where i_{scale} is positive. Moreover, the scale index detects sudden expansions or contractions of the size of the attractor in BvP that are not detected by the MLE.

In spite of being a relatively new-developed method, the scale index has been already applied in a wide range of different areas such as the analysis of precipitation time series in meteorology [6], the study of cardiac dynamics in Electrical Engineering [7], speech signals [8] and pseudo random number generators in Complexity Theory and Cryptography [9].

4 Scalogram Comparisons

In this section, we are going to introduce a method for comparing the scalograms of two time series. This can be a complement to some other tools as the *cross wavelet power* or the *wavelet coherence* introduced in [10].

First, we are going to make some considerations. Any function $f \in L^2(\mathbb{R})$ can be written as

$$f = \sum_{k,z\in\mathbb{Z}} d_{k,z}\psi_{k,z}, \tag{8}$$

where $d_{k,z} := \langle f, \psi_{k,z}\rangle$ (that are called *wavelet* or *detail coefficients*) and $\psi_{k,z}$ is the dyadic version of (1), i.e.

$$\psi_{k,z}(t) := \frac{1}{\sqrt{2^z}}\psi\left(\frac{t - 2^z k}{2^z}\right), \tag{9}$$

for all $k, z \in \mathbb{Z}$. So, in order to make fair comparisons, it is convenient to re-distribute homogeneously the scales that contribute to the decomposition of f. Hence, we re-scale the scalogram in a dyadic way

$$\hat{\mathscr{S}}(z) := \mathscr{S}(2^z), \tag{10}$$

where $z \in \mathbb{Z}$.

Given two signals f, f', we can compare their re-scaled scalograms $\hat{\mathscr{S}}, \hat{\mathscr{S}}'$ in order to know if they follow similar patterns. We can make an *absolute* comparison, but it only has sense if both signals use the same measure units or the scalograms have been normalized in some manner. So, if we work with finite time series from t_0 to t_1, then it is recommended to study only the finite scale interval $[s_0, s_1]$ where s_0 is two times the time step and s_1 is the quotient of the length of the time series and the size of the original wavelet ψ; then we can normalize the scalograms and make a *relative* comparison given by

$$\left\| \frac{\hat{\mathscr{S}}}{\|\hat{\mathscr{S}}\|} - \frac{\hat{\mathscr{S}}'}{\|\hat{\mathscr{S}}'\|} \right\|, \tag{11}$$

where

$$\|\hat{\mathscr{S}}\| = \|\hat{\mathscr{S}}\|_{[z_0,z_1]} = \left(\int_{z_0}^{z_1} |\hat{\mathscr{S}}(z)|^2 \mathrm{d}z \right)^{\frac{1}{2}},$$

with $[z_0, z_1]$ is the corresponding dyadic scale interval in which we make the study (i.e. $s_0 = 2^{z_0}$ and $s_1 = 2^{z_1}$); analogously for $\hat{\mathscr{S}}'$.

Finally, we can also compare re-scaled windowed scalograms in a given time subinterval of $[t_0, t_1]$ (in order to locate the study in time) and only for a given scale subinterval of $[s_0, s_1]$. Using this technique, we can compute the scalogram difference centered in a determined time and scale, and so it could be an alternative to the *cross wavelet power* or the *wavelet coherence* introduced in [10].

This method for comparing scalograms can be interpreted as a measure of the similarity between the patterns of two signals, because two signals with similar scalograms follow "similar patterns". But, what does it mean "similar patterns"? The next result clarifies the matter.

Proposition 1 *Given a signal $f \in L^2(\mathbb{R})$, we have that $\pm f(t + c_1) + c_2$ has the same scalogram as f, where $c_1, c_2 \in \mathbb{R}$. Moreover, if the wavelet is antisymmetric, then we can affirm that $\pm f(\pm t + c_1) + c_2$ has the same scalogram as f.*

Proof The square of the scalogram of $\pm f(t + c_1) + c_2$ is given by

$$\int_{-\infty}^{+\infty} \left| \int_{-\infty}^{+\infty} (\pm f(t + c_1) + c_2)\, \psi^*_{u,s}(t) \mathrm{d}t \right|^2 \mathrm{d}u = \int_{-\infty}^{+\infty} \left| \int_{-\infty}^{+\infty} f(t + c_1) \psi^*_{u,s}(t) \mathrm{d}t \right|^2 \mathrm{d}u, \tag{12}$$

because $\int_{-\infty}^{+\infty} c_2 \psi_{u,s}^*(t)\mathrm{d}t = 0$. Then, making the change of variable $t' = t + c_1$ in (12) and renaming t' as t, we have

$$\int_{-\infty}^{+\infty} \left| \int_{-\infty}^{+\infty} f(t)\psi_{u,s}^*(t - c_1)\mathrm{d}t \right|^2 \mathrm{d}u. \tag{13}$$

Since $\psi_{u,s}(t - c_1) = \psi_{u+c_1,s}(t)$, making the change of variable $u' = u + c_1$ in (13) and renaming u' as u, we obtain the expression of the square of the scalogram of f.

On the other hand, if ψ is antisymmetric, then the square of the scalogram of $f(-t)$ is given by

$$\int_{-\infty}^{+\infty} \left| \int_{-\infty}^{+\infty} f(-t)\psi_{u,s}^*(t)\mathrm{d}t \right|^2 \mathrm{d}u. \tag{14}$$

Making the change of variable $t' = -t$ in (14) and renaming t' as t, we have

$$\int_{-\infty}^{+\infty} \left| \int_{-\infty}^{+\infty} f(t)\psi_{u,s}^*(-t)\mathrm{d}t \right|^2 \mathrm{d}u. \tag{15}$$

Moreover, since ψ is antisymmetric, it is easy to prove that $\psi_{u,s}(-t) = -\psi_{-u,s}(t)$; then, making the change of variable $u' = -u$ in (15) and renaming u' as u, we obtain the expression of the square of the scalogram of f. □

Note that, in general, a wavelet is "more or less" antisymmetric; so, we can conclude that the scalogram of $\pm f(\pm t + c_1) + c_2$ is very similar to the scalogram of f. Moreover, it is easy to prove that if we consider all the possible windowed scalograms, only a signal of the form

$$\pm f(t) + c, \qquad c \in \mathbb{R}$$

has the same windowed scalograms (all of them) as f.

In conclusion, this method for comparing scalograms measures the similarity between the patterns of two signals taking into account that f and $\pm f(t) + c$ follow the same patterns. This also happens with some other tools, e.g. the *wavelet coherence*.

A work related with scalograms comparisons, their interpretations, applications and the relations with other tools like the *wavelet coherence* is currently being developed.

References

1. Mallat, S.: A Wavelet Tour of Signal Processing. Academic, London (1999)
2. Benítez, R., Bolós, V.J., Ramírez, M.E.: A wavelet-based tool for studying non-periodicity. Comput. Math. Appl. **60**(3), 634–641 (2010)
3. Kaiser, G.: A Friendly Guide to Wavelets. Birkhäuser, Boston (1994)
4. Strogatz, S.H.: Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering. Perseus, Reading (1994)
5. Chandre, C., Wiggins, S., Uzer, T.: Time-frequency analysis of chaotic systems. Physica D **181**(3–4), 171–196 (2003)
6. Fan, Q., Wang, Y., Zhu, L.: Complexity analysis of spatial–temporal precipitation system by PCA and SDLE. Appl. Math. Model. **37**(6), 4059–4066 (2013)
7. Behnia, S., Ziaei, J., Ghiassi, M.: New Approach to the Study of Heartbeat Dynamics Based on Mathematical Model. Paper presented at the 21st Iranian Conference on Electrical Engineering, ICEE, pp. 1–5. Ferdowsi University of Mashhad, Mashhad, 4–16 May, 2013
8. Hesham, M.: Wavelet-scalogram based study of non-periodicity in speech signals as a complementary measure of chaotic content. Int. J. Speech Tech. **16**, 353–361 (2013)
9. Akhshani, A., Akhavan, A., Mobaraki, A., Lim, S.C., Hassan, Z.: Pseudo random number generator based on quantum chaotic map. Commun. Nonlinear. Sci. **19**, 101–111 (2014)
10. Torrence, C., Compo, G.P.: A practical guide to wavelet analysis. Bull. Am. Math. Soc. **79**, 61–78 (1998)

A Simplified Wildland Fire Model Applied to a Real Case

Luis Ferragut, María Isabel Asensio, José Manuel Cascón, and Diego Prieto

Abstract We present a simplified 2D wildland fire model with some 3D effects, which takes into account convection and radiation. The topography, the fuel load and type and the meteorological data required by the model (temperature, humidity and wind) are provided via GIS. The wind conditions can be considered a given data in all domain, or can be computed by the wind model developed by authors. Given the fire ignition location and time the model provide the state of landscape for several time steps, allowing to establish the perimeter of the fire at different instants. By modifying the fuel load and type raster files, fire suppression tactics can be incorporated in order to adapt the simulation to the real situation, since the simplified model and its numerical solution allow a computational time much less than real time.

1 Introduction

Wildland fire is the complicated combination of energy (heat) released due to chemical reactions in the process of combustion and the transport of that energy to surrounding unburnt fuel and its subsequent ignition, and occurs on scales ranging from millimeters up to kilometers, that makes the modelling of wildland fire behavior a real complex problem.

L. Ferragut (✉) • M.I. Asensio
Instituto Universitario de Física Fundamental y Matemáticas, Universidad de Salamanca, C/ Parque s.n., casa n°2, 37008 Salamanca, Spain
e-mail: ferragut@usal.es; mas@usal.es

J.M. Cascón
Departamento de Economía e Historia Económica, Universidad de Salamanca, Edificio FES, Campus Miguel de Unamuno, 37007 Salamanca, Spain
e-mail: casbar@usal.es

D. Prieto
Departamento de Matemática Aplicada, Universidad de Salamanca, C/ Parque s.n., casa n°2, 37008 Salamanca, Spain
e-mail: dpriher@usal.es

F. Casas, V. Martínez (eds.), *Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 4, DOI 10.1007/978-3-319-06953-1_16

In recent years, the advances in computational power and the increase in the capabilities of spatial information technologies (remote sensing and geographic information systems) offer great potential for the effective simulation of wildland fire behavior. This has re-intensified the interest in the fire behavior modelling as can be appreciated in the diverse and interesting reviews that have appeared recently on wildland fire modelling [10, 13–15]. Most of these reviews classify fire models according to the nature of their construction as being: physical, semi-physical or empirical, although the nomenclature varies. Some authors [13] define the physical (theoretical) models as those that attempt to represent both the physics and the chemistry of fire spread; and quasi-physical models as those that attempt to represent only the physics.

The model we present in this paper is a simplified quasi-physical model based on the fundamental physics of combustion and fire spread. The model takes into account the three generally accepted forms of heat transfer: conduction, convection and radiation. The most important physical processes driving the heat transfer in a wildland fire are convection and radiation. In low wind conditions, the dominating mechanism is radiation [16], but in conditions where wind is not insignificant, it is convection that dominates [17]. However, it is not reasonable to assume one works without the other and thus both processes must be considered, though it is possible to avoid diffusion.

Based on previous fire models developed by the present authors, [1, 6, 7] we present a simplified two-dimensional wildland fire model with some three-dimensional effects. The model takes into account: the moisture content by using an enthalpy multivalued operator, the energy convected by the gas pyrolyzed through the elementary control volume, the energy lost in the vertical direction and the radiation.

The model also takes into account the topography of the surface where the fire takes place and the fuel load and type. This information is provided by a GIS as well as the meteorological data required by the model: temperature, humidity and wind (direction and velocity) from any meteorological service. The local adjustment wind field model developed by the present authors [2, 8] provides a wind field adjusted to the given meteorological data (wind direction and velocity at several points).

Given the fire ignition location and time, the fire model provides the state of landscape (burning, burnt and unburnt area) for several time steps, allowing to establish the perimeter of the fire at different instants. By modifying the fuel load and type raster files, fire suppression tactics can be incorporated in the location and time desired in order to adapt the simulation to the real situation, since the simplified model and its numerical solution allow a computational time much less than real time. The numerical solution proposed for this model tries to reduce the computational time defining the actives nodes and making use of parallel computation.

The outlines of the paper are as follows: In Sect. 2 we describe the fire model, in Sect. 3 we establish the numerical scheme and in Sect. 4 we report on numerical experiments to simulate a real fire that happened in an area covered by shrub and grass near Salamanca (Spain) in September, 2012. Finally some conclusions and ideas that are being developed at present are given in section "Conclusions".

2 The Fire Model

Following the models in [5] and [9], and previous models developed by the present authors as the model with local radiation in [1], the model with a multivalued operator for the enthalpy in [6] and the model with no local radiation in [7], we present here a simplified two-dimensional model with some three-dimensional effects. This model takes into account: the moisture content by using an enthalpy multivalued operator, the energy convected by the gas pyrolyzed through the elementary control volume, the energy lost in the vertical direction and the radiation from the flames above the surface where the fire takes place.

Let $d = [0, l_x] \times [0, l_y] \subset \mathbb{R}^2$ be a rectangle representing the projection of the surface S where the fire takes place, defined by the mapping

$$\begin{aligned} S : \ d &\longmapsto \mathbb{R}^3 \\ (x, y) &\longmapsto (x, y, h(x, y)). \end{aligned}$$

We will assume that vegetation can be represented by a given fuel load M, ($\mathrm{kg\,m^{-2}}$) together with a moisture content M_v, (kg of water/kg of dry fuel). M and M_v are scalar functions defined on d. Additionally we will assume that the height F of the flames in a particular fire is known and bounded by δ.

In order to take into account some three-dimensional effects, and particularly the radiation from the flames above the surface S, we will consider the following three-dimensional domain,

$$D = \{(x, y, z) : x, y \in d, h(x, y) < z < h(x, y) + \delta\}.$$

The governing equations for the fire model are based on the energy and mass conservation equation in the surface S, and the radiation equation in D. The non-dimensional simplified equations for the fire model are,

$$\partial_\tau e + \beta \mathbf{v} \cdot \nabla e + \alpha u = r \quad \text{in } S \quad \tau \in (0, \tau_{max}), \tag{1}$$

$$e \in G(u) \quad \text{in } S \quad \tau \in (0, \tau_{max}), \tag{2}$$

$$\partial_\tau c = -g(u)c \quad \text{in } S \quad \tau \in (0, \tau_{max}). \tag{3}$$

We complete the problem with homogeneous Dirichlet boundary conditions and the following initial conditions,

$$u(x, y, 0) = u_0(x, y) \quad \text{in } S, \tag{4}$$

$$c(x, y, 0) = c_0(x, y) \quad \text{in } S. \tag{5}$$

The unknowns $e = \frac{E}{MCT_\infty}$, non-dimensional enthalpy, $u = \frac{T-T_\infty}{T_\infty}$, non-dimensional temperature of the solid fuel and $c = \frac{M}{M_0}$, mass fraction of solid fuel, are bidimensional variables defined in $S \times (0, \tau_{max})$. The physical quantities E, T and M are enthalpy, temperature of solid fuel and fuel load respectively, besides the heat capacity of solid fuel C, a reference temperature T_∞ and M_0 the initial fuel load.

The non-dimensional enthalpy e is an element of a multivalued maximal monotone operator G [4], given by:

$$G(u) = \begin{cases} u & if \quad u < u_v \\ [u_v\ ,\ u_v + \lambda_v] & if \quad u = u_v \\ u + \lambda_v & if \ u_v < u < u_p \\ [u_p + \lambda_v\ ,\ \infty] & if \quad u = u_p \end{cases}$$

where u_v and u_p are the non-dimensional evaporation temperature of the water and the non-dimensional pyrolysis temperature of the solid fuel, respectively. The quantity λ_v is the non-dimensional evaporation heat related to the evaporation latent heat Λ_v

$$\lambda_v = \frac{M_v \Lambda_v}{CT_\infty}.$$

It should be noticed that in the burnt zone the multivalued operator does not exactly represent the physical phenomena as the water vapor is no longer in the porous medium. This drawback can be circumvented setting $\lambda_v = 0$ in the burnt area. For more details about the multivalued operator see [6].

The convective term $\beta \mathbf{v} \cdot \nabla e$ represents the energy convected by the gas pyrolyzed through the elementary control volume, where the wind velocity $\mathbf{v}$ is re-scaled by a correction factor $\beta \propto \frac{(MCT)_g}{(MCT)_s}$. The surface wind velocity can be considered given data or can be computed by means of the wind model developed by authors in [2,8]. The coupling of this convection model with a fire model was detailed by authors in [6].

The term αu represents the energy loss by natural convection in the vertical direction. The parameter α is related to physical quantities by $\alpha = \frac{H[t]}{MC}$, where H is the natural convection coefficient.

The right hand side of Eq. (3) represents the lost of solid fuel due to combustion, so $g(u) = 0$ when $u < u_p$, and $g(u)$ is constant when $u = u_p$ where the constant is inverse proportional to the half life time of combustion of each type of fuel.

The right hand side of Eq. (1) describes the thermal radiation reaching the surface S from the flame above the layer. The intensity of radiation is defined as the radiation energy passing through an area per unit time, per unit of projected area and per unit of solid angle. The projected area is formed by taking the area that the energy is passing through and projecting its normal to the direction of travel. The unit elemental solid angle is centered about the direction of travel and has its origin at the area element.

After non-dimensionalization, the radiation equations in the direction $\boldsymbol{\Omega}$ can be written as

$$\boldsymbol{\Omega} \cdot \nabla i + a^* i = \epsilon(1 + u_g)^4 \quad \text{in } D, \tag{6}$$

$$i = 0 \quad \text{on } \partial D \cap \{\mathbf{x};\ \boldsymbol{\Omega} \cdot \mathbf{N} < 0\}, \tag{7}$$

where $i = \frac{I[t]}{MCT_\infty}$ is the nondimensional radiation intensity, $a^* = [l]a$ is the nondimensional absorption coefficient, $u_g = \frac{T_g - T_\infty}{T_\infty}$ is the nondimensional flame temperature, $\epsilon = \frac{[l][t]a\sigma[T]^3}{MC\pi}$ depends on the Stefan-Boltzmann constant $\sigma = 5.6699 \times 10^{-8}\ \mathrm{Wm^{-2}K^{-4}}$ and $\mathbf{N}$ is the outer unit normal vector field to ∂D. In a first approximation we have considered a gray body and neglected the scattering. Here, $a(\mathbf{x})$ is the mean absorption coefficient of the gray body and is a function of the point $\mathbf{x} = (x, y, z) \in D$. The right hand side represents the total emissive power of a blackbody. The incident energy at a point $\mathbf{x} = (x, y, h(x, y))$ of the surface S due to radiation from the flame above the surface per unit time and per unit area will be obtained summing up the contribution of all directions $\boldsymbol{\Omega}$, that is

$$r(\mathbf{x}) = \int_{\omega=0}^{2\pi} i(\mathbf{x}, \boldsymbol{\Omega})\boldsymbol{\Omega} \cdot \mathbf{N}\, d\omega \tag{8}$$

where we have only considered the hemisphere above the fuel layer, and each contribution depends on the flame height. For more details about how to derive the radiation term of this model see [7].

3 Numerical Method

Model simulation must be achieved much faster than real time to be useful in decision support. In order to reduce the computational time we propose to solve the equations of the model only in an environment of the fire front, defining for each time step the set of active nodes. We define a uniform and fine mesh at the beginning of the numerical process, and we solve the corresponding equations only

in the set of active nodes formed by the nodes placed inside the fire front and their environment. This reduces the computational time since we do not have to solve the equations of the model where the solution does not change at all.

3.1 Time Integration

Let $\Delta\tau = \tau^{n+1} - \tau^n$ be a time step and let c^n, e^n and u^n denote approximations at time step τ^n to the exact solution c, e and u, respectively.

We consider an implicit scheme by discretizing the total derivative, see [11],

$$\partial_\tau e + \beta \mathbf{v} \cdot \nabla e \approx \frac{1}{\Delta\tau}(e^{n+1} - \bar{e}^n),$$

where $\bar{e}^n = e^n \circ X^n$, and $X^n(\mathbf{x}) = X(\mathbf{x}, \tau^{n+1}, \tau^n) \approx \mathbf{x} - \beta\mathbf{v}\Delta\tau$ is the position at time τ^n of the particle which is at position $\mathbf{x}$ at time τ^{n+1}. At each time step, we solve,

$$\frac{e^{n+1} - \bar{e}^n}{\Delta\tau} + \alpha u^{n+1} = r^n, \tag{9}$$

$$e^{n+1} \in G(u^{n+1}), \tag{10}$$

$$\frac{c^{n+1} - c^n}{\Delta\tau} = -g(u^{n+1})c^{n+1}. \tag{11}$$

The basic idea is to treat implicitly the positive terms. The non local radiation term r depends strongly on the temperature u and on the fuel mass c, therefore, it will be evaluated explicitly at time τ^n. Once the radiation r^n is known, the problem given by Eqs. (9)–(11) is non linear due to the multivalued operator G. However, the solution of this problem can be reduced to explicit calculations.

3.2 Numerical Solution of the Multivalued Equation

The multivalued operator in Eq. (10) is maximal monotone, then its resolvent $J_\mu = (Id + \mu G)^{-1}$ for any $\mu > 0$ is a well defined univalued operator. Moreover the Yosida approximation of G, $G_\mu = \frac{Id - J_\mu}{\mu}$ is a Lipschitz operator and the inclusion Eq. (10) is equivalent [3] for all $\mu > 0$ to the equation

$$e^{n+1} = G_\mu(u^{n+1} + \mu e^{n+1}), \tag{12}$$

or

$$u^{n+1} = J_\mu(u^{n+1} + \mu e^{n+1}). \tag{13}$$

On the other hand, rearranging Eq. (9) we have

$$u^{n+1} + \frac{1}{\alpha\Delta\tau}e^{n+1} = \frac{1}{\alpha\Delta\tau}\bar{e}^n + \frac{1}{\alpha}r^n. \tag{14}$$

Taking $\mu = 1/(\alpha\Delta\tau)$ by substitution in Eq. (13) we obtain

$$u^{n+1} = J_{1/\alpha\Delta\tau}(\frac{1}{\alpha\Delta\tau}\bar{e}^n + \frac{1}{\alpha}r^n). \tag{15}$$

Once u^{n+1} has been obtained by solving Eq. (15), we calculate e^{n+1} and c^{n+1} explicitly

$$e^{n+1} = \bar{e}^n - \alpha\Delta\tau u^{n+1} + \Delta\tau r^n, \tag{16}$$

$$c^{n+1} = \frac{c^n}{1 + \Delta\tau g(u^{n+1})}. \tag{17}$$

It remains to explain how to calculate u^{n+1} in Eq. (15). That is, for a given $b = \frac{1}{\alpha\Delta\tau}\bar{e}^n + \frac{1}{\alpha}r^n$, compute $s = J_{1/\alpha\Delta\tau}(b)$ is equivalent to solve

$$(\alpha\Delta\tau\ Id + G)s \ni \bar{b} = \alpha\Delta\tau\ b, \tag{18}$$

then s is given by

$$\begin{array}{lll}
\text{if } \bar{b} < (1+\alpha\Delta\tau)u_v & \text{then } s = \frac{\bar{b}}{1+\alpha\Delta\tau} \\
\text{if } (1+\alpha\Delta\tau)u_v < \bar{b} < (1+\alpha\Delta\tau)u_v + \mu_v & \text{then } s = u_v \\
\text{if } (1+\alpha\Delta\tau)u_v + \mu_v < \bar{b} < (1+\alpha\Delta\tau)u_p + \mu_v & \text{then } s = \frac{\bar{b}-\lambda_v}{1+\alpha\Delta\tau} \\
\text{if } (1+\alpha\Delta\tau)u_p + \mu_v < \bar{b} < \infty & \text{then } s = u_p.
\end{array}$$

Notice that Eqs. (12), (16) and (17) can be solved simultaneously in all the active nodes, then parallel computation can be used to improve the computational time.

3.3 Numerical Solution of the Radiation Equation

The radiation term r in Eq. (1) is computed by numerical integration of Eq. (8). We propose Gauss-Legendre nodes for the polar angle and Gauss-Chebyshev nodes for the azimuthal angle. See [7] for more details.

To compute the incident radiation in the direction $\boldsymbol{\Omega}$ on a point $\mathbf{x} = (\bar{x}, \bar{y}, \bar{z}) \in S$, with $\bar{z} = h(\bar{x}, \bar{y})$, we consider the characteristic line expressed in cartesian coordinates

$$[0, \xi] \longmapsto \mathscr{R}^3$$
$$\xi \longrightarrow (x(\xi) = \bar{x} + \xi\Omega_1,\ y(\xi) = \bar{y} + \xi\Omega_2,\ z(\xi) = \bar{z} + \xi\Omega_3).$$

On the characteristic, Eq. (6) becomes

$$\frac{di}{d\xi} + a^* i = \delta(1 + u_g)^4, \tag{19}$$

which can be solved together with the condition

$$\lim_{\xi \to \infty} i(\xi) = 0. \tag{20}$$

Equation (19) is solved by a backward finite difference method of order two based on an interpolation for the variables a and u_g. It should be noted that the three-dimensional mesh does not need to be explicitly computed. To obtain this scheme, we consider u_g and a given by a non-null input value firstly defined over the surface S where the nondimensional fuel temperature u, given by the solution of the energy and fuel equations, Eqs. (1)–(3), reaches the pyrolysis temperature. The non-null values of the non-dimensional flame temperature u_g and the mean absorption coefficient a depends on the type of fuel.

To expand the non-dimensional flame temperature u_g (respectively the mean absorption coefficient a) in D we proceed as follows: If there is no wind, we extend the temperature vertically, that is, we define the extension $\tilde{u}$ by $\tilde{u}(x, y, z) = u(x, y, h(x, y))$ for all points $(x, y, z) \in D$ and $h(x, y) < z < h(x, y) + \delta$. In the case of wind conditions, we compute the extended field assuming a convective transport, that is, $\tilde{u}(x, y, z) = u(x - (z - h(x, y))\frac{v_x}{v_z}, y - (z - h(x, y))\frac{v_y}{v_z}, h(x, y))$. Here (v_x, v_y, v_z) stands for the velocity field which we suppose to be known. More precisely, (v_x, v_y) is a horizontal meteorological velocity field and v_z is computed by the rule $v_z \sim \sqrt{gH}$. Otherwise a three-dimensional velocity field can be computed involving only two-dimensional computations using the model developed by authors in [2, 8].

4 A Real Case

In order to test the model we simulate a real fire that happened in Serradilla del Llano (a municipality located in the south of Salamanca, Spain) in September, 2012. The fire began on the 14th of September at 18 : 20, remained controlled by the

firefighters at 21 : 30 The fire re-ignited the following day being controlled on the 15th of September at 20:30, and completely extinguished on the 16th of September at 20:00 The fire burned 93.98 Ha of grass and 128.19 Ha of shrubs.

Weather variables related to wildland fires: wind (direction and velocity), temperature, relative humidity, and precipitation, were provided for the fire area by MeteoLogica S.A. Although the meteorological information indicates a maximum wind of 24 km/h, the firefighters present indicated wind gusts of 45 km/h near the river where the slope was higher.

The simulation area is a rectangle of 3.5×3.26 km, where the minimum height is 712 m and the maximum is 955 m. The mesh size is 700×652 nodes. The orthoimages and the topographic relief are provided by the National Topographic Base 1:25.000 (BTN25) and the Numerical Cartographic Base 1:25.000 (BCN25) [18]. The topographic data file used for the simulation is 1:5.000, obtained by resampling the corresponding data from the BSN25. The fuel data are provided by the Spanish Forest Map 1:50.000 (MFE50) [19] which includes the following fuel types, where the wooded areas differ depending on the fraction of content covered (FCC) for the total of the woodland (percentage of soil covered by the horizontal projection of the top of the trees):

1. Herbaceous cultures.
2. Pastures.
3. Shrubs.
4. Disperse woodland (FCC: 5–20 %).
5. Spread woodland (FCC: 20–50 %).
6. Opened forest (FCC: 50–70 %).
7. Closed forest (FCC: >70 %).

There are parameters depending on the fuel type such as the fuel load M, moisture content M_v, the height of the flame F and the half life time of combustion. The following table gathers the values of these parameters used in the simulation, depending on each type of fuel.

Fuel type	M (kgm^{-2})	M_v	F(m)	$t_{1/2}$(s)
1	0.2	0.06	1.5	60
2	0.2	0.06	2	60
3	0.5	0.06	2.5	180
4	0.5	0.06	3	180
5	0.5	0.06	4	180
6	3.5	0.06	4	180
7	10	0.06	4	180

In Fig. 1 we represent the simulation area, the perimeter of the real fire that took place on September the 14 and the areas where the firefighters worked. The perimeter of the simulated fire each hour during 3 h are detailed in Fig. 2 where the

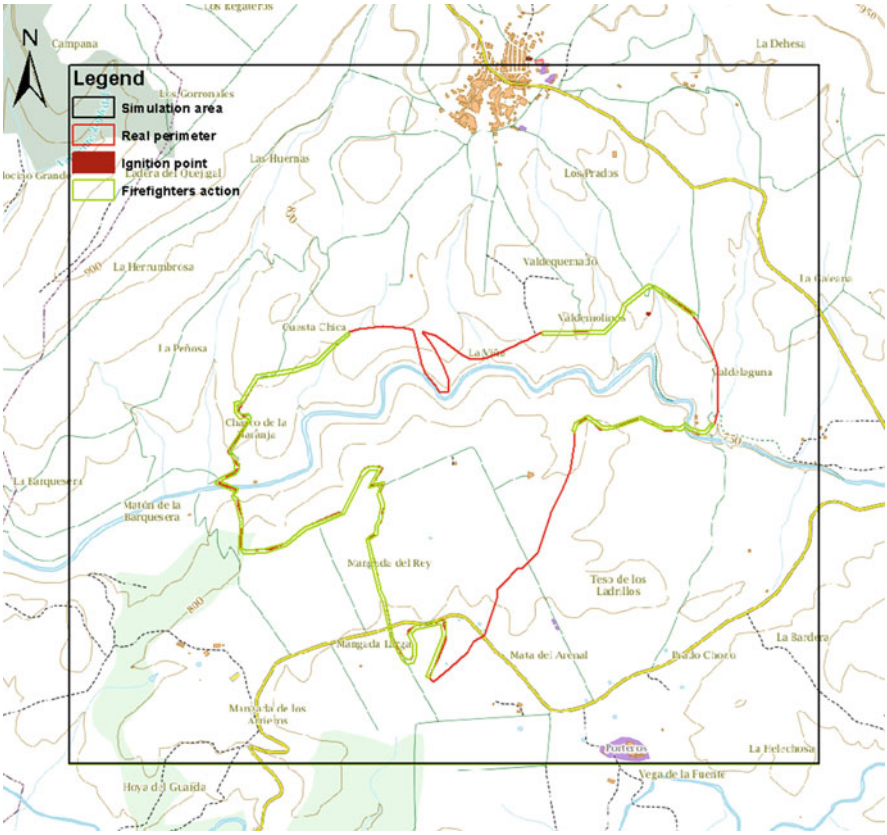


Fig. 1 Perimeter of the Serradilla del Llano fire (09/14/2012), simulation area and Firefighters actions

perimeter of the fire after 3 h is similar to the real perimeter. The differences between the real perimeter and the simulated one owe to the uncertainty in the parameters, the meteorological data and the information on the firefighters actions. The efficiency of the model might get improved using data assimilation-based parameters estimation methods.

These simulations have been computed on a Dell Precision T7500 workstation, equipped with two processors Intel Xeon X5650 (6 cores each one working at a frequency of 2, 66 GHz, 12 cores altogether) and 24 GB RAM. The 3 h of simulation involved 385 s, where we have used the parallel computation and active nodes, which has allowed to simulate 3 h in a few minutes.

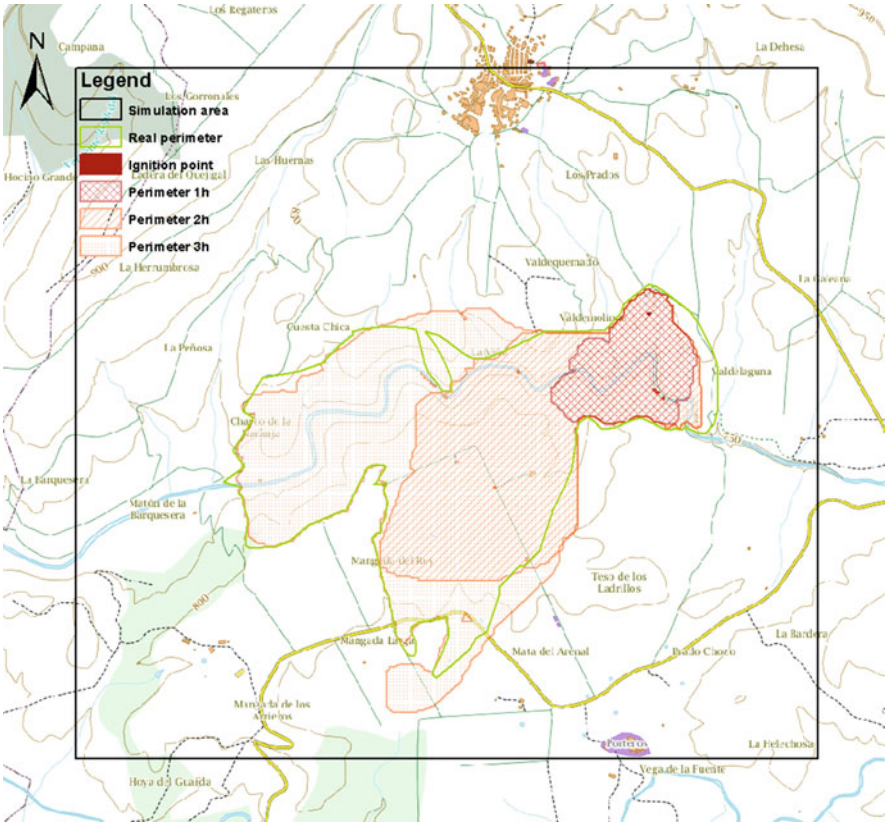


Fig. 2 Simulation perimeters after 1, 2 and 3 h with firefighters actions

Conclusions

We present a simplified quasi-physical fire model taking into account the topography of the surface, the fuel load and type, temperature, humidity and wind, and the most important mechanisms of heat transfer: convection and radiation. The model has been applied to a real case simulating 3 h of a real fire in a few minutes and allowing to incorporate the fire fighting actions.

The model is sufficiently simple as to depend on a few parameters that will allow data assimilation-based parameters estimation, but sufficiently precise as to reflect important phenomena for the evolution of a fire.

The simplicity of the model and the numerical techniques proposed allow to achieve real time simulations in very competitive computational times.

(continued)

The model includes its own wind model which allows to compute a hight definition wind field over the simulation domain that takes into account topography, temperature and meteorological wind data in a few points, in contrast to other models that consider winds to be constant in space, that is, with no orographic or thermal effects on winds. Therefore, the wind model could be solved with a reduced basis scheme [12]. In this way, the thermal effects could be included in the simulation without ruining the computational time.

The model is being integrated into a GIS system as a module which allows the necessary data to be obtained from the GIS system and provide friendly results for the GIS system.

Acknowledgements This work has been partially supported by *Secretaría de Estado de Investigación, Desarrollo e Innovación* and *Centro para el Desarrollo Tecnológico Industrial* of the *Ministerio de Economía y Competitividad* of the Spanish Government, Grant contract: CGL2011-29396-C03-02 and CEN-20101010 and by *Consejería de Educación* of the *Junta de Castilla y León*, Grant contract: SA266A12-2. The authors are also grateful to Ignacio Juárez Relaño chief of the *Sección de Protección de la Naturaleza* of the *Servicio Territorial de Medio Ambiente* of Salamanca, for his technical support providing all the necessary information about the *Serradilla del Llano* fire.

References

1. Asensio, M.I., Ferragut, L.: On a wildland fire model with radiation. Int. J. Numer. Methods Eng. **54**, 137–157 (2002)
2. Asensio, M.I., Ferragut, L., Simon, J.: A convection model for fire spread simulation. Appl. Math. Lett. **18**, 673–677 (2005)
3. Bermúdez, A., Moreno, C.: Duality methods for solving variational inequalities. Comput. Math. Appl. **7**, 43–58 (1981)
4. Brézis, H.: Operateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert. North-Holland, Amsterdam (1973)
5. Cox, G.: Combustion fundamentals of fire. Academic, London (1995)
6. Ferragut, L., Asensio, M.I., Monedero, S.: A numerical method for solving convection-reaction-diffusion multivalued equations in fire spread modelling. Adv. Eng. Softw. **18**, 366–371 (2006)
7. Ferragut, L., Asensio, M.I., Monedero, S.: Modelling radiation and moisture content in fire spread. Commun. Numer. Methods Eng. **23**, 819–833 (2007)
8. Ferragut, L., Asensio, M.I., Simon, J.: High definition local adjustment model for 3D wind fields performing only 2D computations. Int. J. Numer. Methods Biomed. Eng. **27**, 510–523 (2011)
9. Margerit, J., Séro-Guillaume, O.: Modelling forest fires. Part II: reduction to two-dimensional models and simulation of propagation. Int. J. Heat Mass Transf. **45**, 1723–1737 (2002)
10. Pastor, E., Zarate, L., Planas, E., Arnaldos, J.: Mathematical models and calculation systems for the study of wildland fire behaviour. Prog. Energy Combust. **29**(2), 139–153 (2003)
11. Pironneau, O.: On the transport-difussion algorithm and its applications to the Navier-Stokes equations. Numer. Math. **38** 309–332 (1982)

12. Rozza, G.: Reduced basis approximation and error bounds for potential flows inparametrized geometrics. Commun. Comput. Phys. **9**(1), 1–48 (2011)
13. Sullivan, A.L.: Wildland surface fire spread modelling, 1990–2007. 1: physical and quasi-physical models. Int. J. Wildland Fire **18**, 349–368 (2009)
14. Sullivan, A.L.: Wildland surface fire spread modelling, 1990–2007. 2: empirical and quasi-empirical models. Int. J. Wildland Fire **18**, 369–386 (2009)
15. Sullivan, A.L.: Wildland surface fire spread modelling, 1990–2007. 3: simulation and mathematical analogue models. Int. J. Wildland Fire **18**, 387–403 (2009)
16. Weber, R.O.: Analytical models of fire spread due to radiation. Combust. Flame **78**, 398–408 (1989)
17. Weber, R.O.: Modelling fire spread through fuel beds. Prog. Energy Combust. **17**(1), 67–82 (1991)
18. Instituto Geográfico Nacional (2014). Centro nacional de información geográfica. http://centrodedescargas.cnig.es/CentroDescargas. Accessed 18 Sep 2014
19. Ministerio de Agricultura, Alimentación y Medio Ambiente of the Spanish Goverment (2006). Spanish Forest Map 1:50.000(MFE50). http://www.magrama.gob.es/es/biodiversidad/servicios/banco-datos-naturaleza/informacion-disponible/mfe50.aspx. Accessed 18 Sep 2014

Functional Output-Controllability of Time-Invariant Singular Linear Systems

María Isabel García-Planas and Sonia Tarragona

Abstract In the space of finite-dimensional singular linear continuous-time-invariant systems described in the form

$$\left.\begin{aligned} E\dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) \end{aligned}\right\} \tag{1}$$

where $E, A \in M = M_n(\mathbb{C})$, $B \in M_{n\times m}(\mathbb{C})$, $C \in M_{p\times n}(\mathbb{C})$, functional output-controllability character is considered. A simple test based in the computation of the rank of a certain constant matrix that can be associated to the system is presented.

1 Introduction

A great many physical problems as for example electrical networks, multibody systems, chemical engineering, Economics, semidiscretized Stokes equations, Convolutional codes among others, use state space representation as (1) for description.

This linear system can be described with a input-output relation called transfer function obtained by applying Laplace transformation to Eq. (1)

$$\left.\begin{aligned} sEX - x(0) &= AX + BU \\ Y &= CX, \end{aligned}\right\},$$

obtaining the following relation

$$H(s)U(s) = C(sE - A)^{-1}x(0) + C(sE - A)^{-1}BU(s). \tag{2}$$

M.I. García-Planas (✉)
Universitat Politècnica de Catalunya, Barcelona, Spain
e-mail: maria.isabel.garcia@upc.edu

S. Tarragona
Universidad de León, León, Spain
e-mail: sonia.tarragona@unileon.es

F. Casas, V. Martínez (eds.), *Advances in Differential Equations and Applications*,
SEMA SIMAI Springer Series 4, DOI 10.1007/978-3-319-06953-1_17

If the system is relaxed (that is to say if the initial state is $x(0) = 0$), Eq. (2) is reduced to

$$H(s) = C(sE - A)^{-1}B. \tag{3}$$

The controllability concept of a dynamical standard system is largely studied by several authors and under many different points of view, (see [1–3,14] for example). Nevertheless, functional controllability for the output vector of a system has been less treated for the standard case and even less for the singular case, (see [7, 10, 12, 13] for example).

The functional output-controllability generally means, that the system can steer output of dynamical system along the arbitrarily given curve over any interval of time, independently of its state vector. A similar but least essentially restrictive condition is the pointwise output-controllability.

J.L. Domínguez in [6] examine the functional output controllability of a linear system describing a fixed speed wind turbine formed by a squirrel cage generator connected directly to the grid. Working over finite fields, Fragouli and Wessel [9] analyze the minimality among strictly equivalent encoders using the functional output controllability character. The authors use the term output observable instead of functional output controllable, it is the same concept but working in discrete variable.

In this paper functional output-controllability for singular systems is analyzed generalizing the study realized for standard systems and a simple test based on computing the ranks of certain matrices in order to study this property is presented. Notice that, in [11], the authors present a test for the study of functional output-controllability of regular singular systems, which result is therefore a particular case of the one presented in this article where regularizable systems are considered.

2 Preliminaries

In this paper, it is considered the singular state space system introduced in Eq. (1)

$$\left.\begin{aligned} E\dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) \end{aligned}\right\},$$

where x is the state vector, y is the output vector, u is the input (or control) vector, $A \in M_n(C)$ is the state matrix, $B \in M_{n\times m}(C)$ is the input matrix and $C \in M_{p\times n}(C)$ is the output matrix.

For simplicity, we will write the systems by quadruples of matrices (E, A, B, C).

In particular we will are interested in systems (called regular) which are those that satisfy the relation $\det(\lambda E + \mu A) \neq 0$ for some $(\lambda, \mu) \in \mathbb{C}^2$, or those systems (called regularizable), which through a feedback proportional and/or derivative and/or an output injection proportional and/or derivative become regular. More

concretely (E, A, B, C) is regularizable if and only if there exist matrices $F_E^B, F_A^B \in M_{m\times n}(\mathbb{C})$, $F_E^C, F_A^C \in M_{n\times p}(\mathbb{C})$, such that the system $(E+BF_E^B+F_E^C C, A+BF_A+F_A^C, B, C)$ is regular.

Remark 1 If a singular system is regular there exists a unique solution for any consistent initial condition.

Remember that an initial condition is called consistent with the system, if the associated initial value problem has at least one solution.

A manner to understand the properties of the system is using algebraic techniques. One of the main aspects of this approach is defining an equivalence relation preserving these properties.

The equivalence relation considered is such that is derived after to make the following elementary transformations: basis change in the state space, basis change in the input space, basis change in the output space, proportional feedback, derivative feedback, proportional output injection, derivative output injection and a premultiplication by an invertible matrix.

More concretely.

Definition 1 Two systems (E_i, A_i, B_i, C_i), $i = 1, 2$, are equivalent if and only if there exist matrices $P, Q \in Gl(n;\mathbb{C})$, $R \in Gl(m;\mathbb{C})$, $S \in Gl(p;\mathbb{C})$, $F_E^B, F_A^B \in M_{m\times n}(\mathbb{C})$, $F_E^C, F_A^C \in M_{n\times p}(\mathbb{C})$ such that

$$\begin{aligned} E_2 &= QE_1P + QB_1F_E^B + F_E^C C_1 P, \\ A_2 &= QA_1P + QB_1F_A^B + F_A^C C_1 P, \\ B_2 &= QB_1R, \\ C_2 &= SC_1P. \end{aligned} \tag{4}$$

That can be written in the following constant matrix form:

$$\begin{pmatrix} E_2 & B_2 & \\ C_2 & & \\ & A_2 & B_2 \\ & C_2 & \end{pmatrix} = \begin{pmatrix} Q & F_E^C & \\ & S & \\ & & Q \quad F_A^C \\ & & S \end{pmatrix} \begin{pmatrix} E_1 & B_1 & \\ C_1 & & \\ & A_1 & B_1 \\ & C_1 & \end{pmatrix} \begin{pmatrix} P & & \\ F_E^B & R & \\ & P & \\ & F_A^B & R \end{pmatrix}. \tag{5}$$

Or in the following polynomial matrix form:

$$\begin{pmatrix} sE_2 - A_2 & B_2 \\ C_2 & \end{pmatrix} = \begin{pmatrix} Q & sF_E^C - F_A^C \\ & S \end{pmatrix} \begin{pmatrix} sE_1 - A_1 & B_1 \\ C_1 & \end{pmatrix} \begin{pmatrix} P & \\ sF_E^B - F_A^B & R \end{pmatrix}. \tag{6}$$

Remark 2 The null terms are not written in these matrices. From now on, if there confusion is not possible, zeroes also omit in the matrices defined in blocks.

Having defined an equivalence relation, the standard procedure then is to look for a canonical form, that is to say to look for a quadruple of matrices which is equivalent to a given quadruple and which has a simple form from which we can

directly read off the properties and invariants of the corresponding singular system. For a better understanding, we will give the following notations: I_ℓ denotes the ℓ-order identity matrix, $N_i = \operatorname{diag}(N_{i_1}, \ldots, N_{i_t}) \in M_{n_i}(\mathbb{C})$, $i = 1, 2, 3, 4$, $N_{i_j} = \begin{pmatrix} 0 & I_{n_{i_j}-1} \\ 0 & 0 \end{pmatrix} \in M_{n_{i_j}}(C)$, $J = \operatorname{diag}(J_1, \ldots, J_t) \in M_{n_5}(\mathbb{C})$, $J_i = \operatorname{diag}(J_{i_1}, \ldots, J_{i_s})$, $J_{i_j} = \lambda_i I_{i_j} + N$.

Proposition 1 *A system (E, A, B, C) is regularizable if and only if it can be reduced to (E_r, A_r, B_r, C_r) where:*

$$E_r = \begin{pmatrix} I_1 & & & & \\ & I_2 & & & \\ & & I_3 & & \\ & & & I_4 & \\ & & & & N_1 \end{pmatrix},\ A_r = \begin{pmatrix} N_2 & & & & \\ & N_3 & & & \\ & & N_4 & & \\ & & & J & \\ & & & & I_5 \end{pmatrix},\ B_r = \begin{pmatrix} B_1 & 0 & 0 \\ 0 & B_2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \text{ and}$$

$$C_r = \begin{pmatrix} C_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & C_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Remark 3 1. The standard part of the system is maximal among all possible reductions of the system.

2. Not all parts (i),..., (v), necessarily appears in the decomposition of the system.

The reduced form can be obtained from the following complete set of invariants

$$\{\sigma, \rho_i^J, \rho_i^K, \rho_i^L, \rho_i^M\}$$

where

(i) $\sigma = \{\lambda \in \mathbb{C} \mid \operatorname{rank} \begin{pmatrix} \lambda E - A & B \\ C & 0 \end{pmatrix} < \operatorname{rank} \begin{pmatrix} sE - A & B \\ C & 0 \end{pmatrix}\}$.

(ii) $\rho_i^{J(\lambda)} = \operatorname{rank}(J_i(\lambda))$ with

$$J_1(\lambda) = \begin{pmatrix} \lambda E - A & B \\ C & 0 \end{pmatrix},\ J_2(\lambda) = \begin{pmatrix} \lambda E - A & B & & \\ C & 0 & & \\ E & 0 & \lambda E - A & B \\ & & C & 0 \end{pmatrix}, \ldots,$$

$$J_i(\lambda) = \begin{pmatrix} \lambda E - A & B & & & & & \\ C & 0 & & & & & \\ E & 0 & \lambda E - A & B & & & \\ & & C & 0 & & & \\ & & E & 0 & \ddots & & \\ & & & & & \lambda E - A & B \\ & & & & & C & 0 \end{pmatrix} \in M_{i(n+p)\times i(n+m)}(\mathbb{C})$$
$$\forall \lambda \in \sigma \tag{7}$$

(iii) $\rho_i^K = \operatorname{rank}(K_i)$ with

$$K_1 = \begin{pmatrix} E & B \\ C & 0 \end{pmatrix},\ K_2 = \begin{pmatrix} E & B & & \\ C & 0 & & \\ A & 0 & E & B \\ & & C & 0 \end{pmatrix}, \ldots,$$
$$K_i = \begin{pmatrix} E & B & & & & \\ C & 0 & & & & \\ A & 0 & E & B & & \\ & & C & 0 & & \\ & & A & 0 & & \\ & & & & \ddots & \\ & & & & E & B \\ & & & & C & 0 \end{pmatrix} \in M_{i(n+p)\times i(n+m)}(\mathbb{C}). \tag{8}$$

(iv) $\rho_i^L = \operatorname{rank}(L_i)$ with

$$L_0 = (B), L_1 = \begin{pmatrix} E & B & 0 \\ C & 0 & 0 \\ A & 0 & B \end{pmatrix},\ L_2 = \begin{pmatrix} E & B & & & \\ C & 0 & & & \\ A & 0 & E & B & 0 \\ & & C & 0 & 0 \\ & & A & 0 & B \end{pmatrix}, \ldots,$$
$$L_i = \begin{pmatrix} E & B & & & & & \\ C & 0 & & & & & \\ A & 0 & E & B & & & \\ & & C & 0 & & & \\ & & A & 0 & \ddots & & \\ & & & & & E & B & 0 \\ & & & & & C & 0 & 0 \\ & & & & & A & 0 & B \end{pmatrix} \in M_{(i+1)n+ip)\times in+(o+1)m)}(\mathbb{C}). \tag{9}$$

(v) $\rho_i^M = \operatorname{rank}(M_i)$ with

$$M_0 = (C), M_1 = \begin{pmatrix} A & B & E \\ C & 0 & 0 \\ 0 & 0 & C \end{pmatrix}, \; M_2 = \begin{pmatrix} A & B & E & & \\ C & 0 & 0 & & \\ & & A & B & E \\ & & C & 0 & 0 \\ & & 0 & 0 & C \end{pmatrix}, \ldots,$$

$$M_i = \begin{pmatrix} A & B & E & & & & & \\ C & 0 & 0 & & & & & \\ & & A & B & E & & & \\ & & C & 0 & 0 & & & \\ & & & & \cdot & & & \\ & & & & & \ddots & & \\ & & & & & & A & B & E \\ & & & & & & C & 0 & 0 \\ & & & & & & 0 & 0 & C \end{pmatrix} \in M_{in+(i+1)p \times (i+1)n+im}(\mathbb{C}) \tag{10}$$

(For more details see [4, 5]).

3 Functional Output-Controllability for Standard Systems

In order to make more comprehensible this work, we begin by recalling the concept of functional output-controllability for standard systems

The output-controllability means, that the system can steer output of the dynamical system independently of its state vector.

Definition 2 A standard system is functional output-controllable if and only if its output can be steered along the arbitrarily given curve over any interval of time. It means that if it is given any output $y_d(t)$, $t \geq 0$, there exists t_1 and a control u_t, $t \geq 0$, such that for any $t \geq t_1$, $y(t) = y_d(t)$.

Proposition 2 ([2]) *A system is functional output-controllable if and only*

$$\operatorname{rank} C(sI - A)^{-1} B = p$$

in the field of rational functions.

A necessary and sufficient condition for functional output-controllability is

Proposition 3 ([2, 8])

$$\operatorname{rank} \begin{pmatrix} sI - A & B \\ C & 0 \end{pmatrix} = n + p.$$

3.1 Test for Functional Output-Controllability for Standard Systems

The functional output-controllability can be computed by means of the rank of a constant matrix in the following manner

Theorem 1 ([10]) *The system* (A, B, C) *is functional output-controllable if and only if*

$$\operatorname{rank} oC_{\mathrm{f}}(A, B, C) = \operatorname{rank} \begin{pmatrix} C & & & \\ CA & CB & & \\ CA^2 & CAB & CB & \\ \vdots & & & \ddots \\ CA^n & CA^{n-1}B & \ldots & CAB\ CB \end{pmatrix} = (n+1)p\,.$$

Remark 4 We call

$$oC_i = \begin{pmatrix} C & & & \\ CA & CB & & \\ CA^2 & CAB & CB & \\ \vdots & & & \ddots \\ CA^i & CA^{i-1}B & \ldots & CAB\ CB \end{pmatrix},\ \forall i \geq 1.$$

(i) If the system (A, B, C) is functional output-controllable, then the matrices oC_i have full row rank for all $0 \leq i \leq n$.
(ii) If the matrix oC_{n-1} has full row rank, it does not necessarily the matrix oC_n has full row rank.

4 Functional Output-Controllability for Singular Systems

We begin by considering singular systems that are regular.

The output-controllability character can be generalized to regular singular systems in the following manner.

Definition 3 A regular singular system is functional output-controllable if and only if its output can be steered along the arbitrarily given curve over any interval of time. It means that if it is given any output $y_d(t)$, $t \geq 0$, there exists t_1 and a control u_t, $t \geq 0$, such that for any $t \geq t_1$, $y(t) = y_d(t)$.

Proposition 4 *A relaxed regular singular system is functional output-controllable if and only*

$$\operatorname{rank} H(s) = p$$

in the field of rational functions.

Proof According to Eq. (3), $H(s) = C(sE - A)^{-1}B$.

If $\operatorname{rank} H(s) = p$, then $H(s)H(s)^*$ is invertible, then it suffices to consider

$$U(s) = H(s)^*(H(s)H(s)^*)^{-1}Y(s)$$

If $\operatorname{rank} H(s) < p$, we can obtain a $Y(s)$ with $Y(s) \notin \operatorname{Im} H(s)$. □

A necessary and sufficient condition for functional output-controllability is

Proposition 5

$$\operatorname{rank}\begin{pmatrix} sE - A & B \\ C & 0 \end{pmatrix} = n + p.$$

Proof

$$\operatorname{rank}\begin{pmatrix} sE - A & B \\ C & 0 \end{pmatrix} = \operatorname{rank}\begin{pmatrix} I & 0 \\ 0 & C(sE - A)^{-1}B \end{pmatrix}.$$

□

Remark 5 Notice that for $E = I$ the proposition coincides with Proposition 3.

Remark 6 If $\operatorname{rank} C < p$ the system is not functional output-controllable. Then, henceforth without lost of generality, we suppose that $\operatorname{rank} C = p$.

In order to obtain more properties we make use of the equivalence relation defined in Definition 1. It permits us to consider an equivalent simple reduced form for the system.

Proposition 6 *The functional output-controllability character is invariant under equivalence relation.*

Proof

$$\operatorname{rank}\begin{pmatrix} Q & sF_E^C - F_A^C \\ 0 & S \end{pmatrix}\begin{pmatrix} sE - A & B \\ C & 0 \end{pmatrix}\begin{pmatrix} P & 0 \\ sF_E^B - F_A^B & R \end{pmatrix} = \operatorname{rank}\begin{pmatrix} sE - A & B \\ C & 0 \end{pmatrix}.$$

□

Propositions 5 and 6 permit us generalize the definition of functional output-controllability to regularizable singular systems.

Definition 4 A regularizable singular linear system is functional output-controllable under proportional and derivative feedback and proportional and derivative output injection, if and only if all equivalent regular singular systems are functional output-controllable.

Corollary 1 *A regularizable singular linear system* (E, A, B, C) *is functional output-controllable under proportional and derivative feedback and proportional and derivative output injection, if and only if*

$$\operatorname{rank}\begin{pmatrix} sE - A & B \\ C & 0 \end{pmatrix} = n + p.$$

Remark 7 If the singular system is not regularizable it is not functional output-controllable.

4.1 Test for Functional Output-Controllability for Singular Systems

The functional output-controllability can be computed by means of the rank of a certain constant matrix defined in the following manner.

For each system (E, A, B, C) we consider the collection of matrices M_i considered in (10).

Proposition 7 *The system* (E, A, B, C) *is functional output-controllable if and only if all matrices* M_i *have full row rank.*

Proposition 8 *For all* $\ell \geq n$ *we have that*

$$\operatorname{rank} M_{\ell+1} - \operatorname{rank} M_{\ell} = \operatorname{rank} M_{\ell+2} - \operatorname{rank} M_{\ell+1}.$$

Calling now $M_n = oC_{\mathrm{f}}(E, A, B, C)$, and taking into account Propositions 7 and 8, we have the following result.

Theorem 2 *The system* (E, A, B, C) *is functional output-controllable if and only if*

$$\operatorname{rank} oC_{\mathrm{f}}(E, A, B, C) = \operatorname{rank}\begin{pmatrix} A & B & E & 0 & 0 & 0 & \dots & & 0 \\ C & 0 & 0 & 0 & 0 & 0 & & & \\ 0 & 0 & A & B & E & 0 & & & \\ 0 & 0 & C & 0 & 0 & 0 & & & \\ \vdots & & & & & & \ddots & & \\ & & & \dots & & & A & B & E \\ & & & \dots & & & C & 0 & 0 \\ & & & \dots & & & 0 & 0 & C \end{pmatrix} = (n+1)p + n^2 .$$

Remark 8 For $E = I$, the theorem coincides with the theorem for standard systems. It suffices to make block elementary row and columns transformations to the matrix $oC_f(I, A, B, C)$:

$$\operatorname{rank}\begin{pmatrix} A & B & I & 0 & 0 & 0 & 0 & \dots & 0 \\ C & 0 & 0 & 0 & 0 & 0 & & & \\ 0 & 0 & A & B & I & 0 & & & \\ 0 & 0 & C & 0 & 0 & 0 & & & \\ \vdots & & & & & & & & \\ & & \dots & & & A & B & I & \\ & & \dots & & & C & 0 & 0 & \\ & & \dots & & & 0 & 0 & C & \end{pmatrix} = \operatorname{rank}\begin{pmatrix} I & & & & & & \\ & \ddots & & & & & \\ & & I & & & & \\ & & & C & & & \\ & & & CA & CB & & \\ & & & CA^2 & CAB & CB & \\ & & & \vdots & & \ddots & \\ & & & CA^n & CA^{n-1}B & & CB \end{pmatrix}.$$

Proof of the Theorem Proposition 6 permit us to consider the system in its reduced form

$$\begin{aligned} &\operatorname{rank}\begin{pmatrix} sE - A & B \\ C & 0 \end{pmatrix} = \operatorname{rank}\begin{pmatrix} sI_1 - N_1 & B_1 \\ C_1 & 0 \end{pmatrix} + \operatorname{rank}\begin{pmatrix} sI_2 - N_2 & B_2 \end{pmatrix} \\ &+\operatorname{rank}\begin{pmatrix} sI_3 - N_3 \\ C_2 \end{pmatrix} + \operatorname{rank}\begin{pmatrix} sI_4 - J \end{pmatrix} + \operatorname{rank}\begin{pmatrix} sN_1 - I_{n_1} \end{pmatrix} = \\ &n_2 + p_2 + n_3 + n_4 + n_5 + n_1 = n + p_2. \end{aligned}$$

The rank is $n + p$ if and only if $p = p_2$. In order to obtain p_2, it suffices to compute the $r_i^{\mathcal{O}}$ numbers associated to the system [4].

An alternative proof can be obtained considering the pencil $\mathbf{A} + s\mathbf{B}$ with

$$\mathbf{A} = \begin{pmatrix} -A & B & 0 \\ C & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \text{ and } \mathbf{B} = \begin{pmatrix} E & 0 & B \\ 0 & 0 & 0 \\ C & 0 & 0 \end{pmatrix}$$

and compute the ranks of the matrices

$$\mathbb{M}_i = \begin{pmatrix} \mathbf{A}^t & 0 & \dots & 0 \\ \mathbf{B}^t & \mathbf{A}^t & & \vdots \\ \vdots & \vdots & \ddots & \mathbf{A}^t \\ 0 & 0 & \dots & \mathbf{B}^t \end{pmatrix}.$$

These matrices appear when one tries to obtain the elements of the Ker of $\mathbf{A}^t + s\mathbf{B}^t$.

Example 1 Let (E, A, B, C) be a system with $E = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, $A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, $B = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ and $C = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$

$$oC_{\mathrm{f}}(E, A, B, C) = \begin{pmatrix}
0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0
\end{pmatrix}.$$

Using Matlab, it is easy to compute the rank of this matrix, we have

$$\operatorname{rank} oC_{\mathrm{f}}(E, A, B, C) = 13.$$

Then, the system is functional output-controllable.

But if we consider the system (E_1, A_1, B_1, C_1) with $E_1 = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, $A_1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, $B_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$ and $C_1 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$

$$oC_{\mathrm{f}}(E_1, A_1, B_1, C_1) = \begin{pmatrix}
0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0
\end{pmatrix}.$$

As before, using Matlab, it is easy to computing the rank of this matrix,

$$\operatorname{rank} oC_{\mathrm{f}}(E_1, A_1, B_1, C_1) = 11.$$

Then the system it is not functional output-controllable.

Remark 9 (i) If the singular system (E, A, B, C) is functional output-controllable, then the matrices M_i has full row rank for all $0 \leq i \leq n$.

(ii) If the matrix M_{n-1} has full row rank, the matrix M_n does not necessarily has full row rank, as it can be seen in the following example.

Example 2 Let (E, A, B, C) with $E = -I$, $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, $B = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $C = \begin{pmatrix} 1 & 0 \end{pmatrix}$.

$$\operatorname{rank} \begin{pmatrix} A & B & -I \\ C & 0 & 0 \\ 0 & 0 & C \end{pmatrix} = \operatorname{rank} \begin{pmatrix} 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} = 4 = n + 2p,$$

but

$$\operatorname{rank} \begin{pmatrix} A & B & -I & & \\ C & 0 & 0 & & \\ 0 & 0 & A & B & -I \\ 0 & 0 & C & 0 & 0 \\ 0 & 0 & 0 & 0 & C \end{pmatrix} = \operatorname{rank} \begin{pmatrix} 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} = 6 < 7.$$

Theorem 2 provides an iterative method to compute functional output-controllability in the following manner.

Step 1: Compute rank M_0. If rank $< p$ the system is not functional output-controllable.
If rank $= p$, then

Step 2: Compute rank M_ℓ. If rank $< (\ell + 1)p + \ell n$ the system is not output-controllable.
If rank $= (\ell + 1)p + \ell n$ and $\ell = n$ the system is functional output-controllable, and if $\ell < n$ go to step 2.

Finally we want to highlight the following: Partitioning matrices $B = \begin{pmatrix} B_1 & \dots & B_m \end{pmatrix}$ and $C = \begin{pmatrix} C_1^t & \dots & C_p^t \end{pmatrix}^t$ by columns and rows respectively, we can compute if there is any SISO subsystem, that is functional output-controllable, in the following manner.

Corollary 2 *Let* (E, A, B, C) *be any system. The subsystem SISO system* E, A, B_i, C_j *for some* $1 \leq i \leq m$, $1 \leq j \leq p$ *is functional output-controllable, if and only if*

$$\operatorname{rank} \begin{pmatrix} A & B_i & -E & 0 & 0 & 0 & \dots & & 0 \\ C_j & 0 & 0 & 0 & 0 & 0 & & & \\ 0 & 0 & A & B_i & -E & 0 & & & \\ 0 & 0 & C_j & 0 & 0 & 0 & & & \\ \vdots & & & & & \ddots & & & \\ & & & \dots & & & A & B_i & -E \\ & & & \dots & & & C_j & 0 & 0 \\ & & & \dots & & & 0 & 0 & C_j \end{pmatrix} = (n+1) + n^2.$$

From this result, it is easy to prove the following proposition.

Proposition 9 *Let* (E, A, B, C) *be a functional output-controllable system. Then, for all* $1 \leq j \leq p$, *there is at least one* i, $1 \leq i \leq m$ *such that the SISO system* (E, A, B_i, C_j) *is functional output-controllable.*

Remark 10 Notice that not necessarily all SISO subsystems are functional output-controllable, and in the case that all SISO subsystems are functional output-controllable, the complete system is not necessarily functional output-controllable.

References

1. Cardetti, F., Gordina, M.: A note on local controllability on lie groups. Syst. Control Lett. **57**, 978–979 (2008)
2. Chen, C.: Introduction to Linear System Theory. Holt, New York (1970)
3. Dai, L.: Singular Control Systems. Springer, New York (1989)
4. Díaz, A.: Sistemas Singulares. Invariantes y Formas Canónicas. PhD. Thesis, Universitat Politècnica de Catalunya (2006)
5. Diaz, A., Garcia-Planas, M.I.: An alternative collection of structural invariants for matrix pencils under strict equivalence. WSEAS Trans. Syst. Control **4**(10), 487–496 (2009)
6. Domínguez-García, J.L.: Computing bounds for the distance of functional output-controllable systems representing fixed speed wind turbine. Cybern. Phys. **2**(2), 77–83 (2013)
7. Domínguez-García, J.L., García-Planas, M.I.: Output controllability analysis of fixed speed wind turbine. In: Proceedings of the 5th International Conference on Physics and Control, León (2011)
8. Ferreira, P.: On degenerate systems. Int. J. Control **24**(4), 585–588 (1876)
9. Fragouli, Ch., Wesel, R.D.: Convolutional Codes and Matrix Control Theory. In: Proceedings of the 7th International Conference on Advances in Communications and Control, Athens (1999)
10. García-Planas, M.I., Domínguez-García, J.: Alternative tests for functional and pointwise output-controllability of linear time-invariant systems. Syst. Control Lett. **62**(5), 382–387 (2013)
11. García-Planas, M.I., Tarragona, S.: Testing functional output-controllability of time-invariant singular linear systems. Cybern. Phys. **2**(2), 957–965 (2013)

12. García-Planas, M.I., Tarragona, S.: Analysis of functional output-controllability of time-invariant singular linear systems. In: Proceedings of the XXIII Congreso de Ecuaciones Diferenciales y Aplicaciones / XIII Congreso de Matemática Aplicada, Castellón 2013. e-Treballs d'Informàtica i Tecnologia, N. 15, pp. 957–965, Publicacions de la Universitat Jaume I (2014).
13. Germani, A., Monaco, S.: Functional output-controllability for linear systems on hilbert. Syst. Control Lett. **2**(5), 313–320 (1983)
14. Kundur, K.: Power System Stability and Control. McGraw-Hill, New York (1994)

Optimising the Welding Process in the Manufacture of Offshore Mooring Chains

Carlos Gorria, Mikel Lezaun, David Pardo, Eduardo Sáinz de la Maza, D. Bilbao, Igor Gutiérrez, and Mariano Lueches

Abstract Vicinay Cadenas S.A is a world leader in the manufacture of mooring chains for the offshore industry. Welding is a key part of chain manufacturing. This study seeks to determine how the manufacturing parameters of welding machines influence the appearance of inhomogeneities. The idea is to optimise current manufacturing processes and acquire knowledge that will enable the firm to develop new products with diameters in excess of those produced to date. To that end, multivariate analysis techniques are used to study manufacturing data on various chains and an algorithm is designed to select the spreads of the adjustable variables that contain the lowest (highest) percentage of links with inhomogeneities. The application of this algorithm to a number of chains with different diameters manufactured on the same machine provides an estimate of the table of settings that should be used to make chains with dimensions larger than those currently made.

C. Gorria (✉) • M. Lezaun • E. Sáinz de la Maza
Departamento de Matemática Aplicada y Estadística e Investigación Operativa, Universidad del País Vasco - UPV/EHU, Facultad de Ciencia y Tecnologt'ia (Leioa), Apdo. 644, E-48080 Bilbao, Spain
e-mail: carlos.gorria@ehu.es; mikel.lezaun@ehu.es; eduardo.sainzdelamaza@ehu.es

D. Pardo
Universidad del País Vasco - UPV/EHU and Ikerbasque, Basque Foundation for Science, Alameda Urquijo 36-5, Plaza Bizkaia, E-48011 Bilbao, Spain
e-mail: dzubiaur@gmail.com

D. Bilbao • I. Gutiérrez
Vicinay Marine Innnovación A.E.I., C/ Particular de Sagarduy 5, E-48015 Bilbao, Spain
e-mail: dbilbao@vicinayinnovacion.com; igutierrez@vicinayicadeenas.com; igutierrez@vicinayinnovacion.com

M. Lueches
Vicinay Cadenas S.A., C/ Particular de Sagarduy s/n, E-48015 Bilbao (Bizkaia), Spain
e-mail: mlueches@vicinaycadenas.com

F. Casas, V. Martínez (eds.), *Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 4, DOI 10.1007/978-3-319-06953-1_18

1 Introduction

Vicinay Cadenas S.A. (VCSA) is a steel processing company that specialises in the manufacture of chains and accessories for mooring lines used in the offshore industry, mainly for oil and gas. Its main customers are major oil companies such as Exxon, Shell and BP, which require mooring systems for their offshore rigs. VCSA manufactures chains of various types according to the preferences of its customers, and is a world leader in its field.

This paper focuses on the welding process, which is a key part of chain manufacturing. Section 2 outlines the whole process of chain manufacturing. Section 3 explains the welding process used at VCSA. Section 4 describes the databases drawn up by VCSA for all its products. Section 5 presents a 40-variable multivariate analysis of the welding process. Section 6 focuses on the 7 adjustable variables that most influence weld quality. An algorithm is designed to select spreads of these variables that contain a preset number of links and the lowest (highest) percentage of links containing inhomogeneities. This algorithm has been selected up by VCSA to improve the tables of settings for its welding machines. The last section of the paper describes the results when the designed algorithm is used on production runs of chains with different diameters on the same machine. This provides an estimated table of settings that can be applied when embarking on the manufacture of larger chains.

2 The Production Process at Vicinay Cadenas S.A.

Manufacturing a mooring line is a long, complex process. The raw material used comprises a round steel bar stock whose diameter and grade are set specifically in each order. The first step is to saw the bars at a set length, determined by the design of the link.

2.1 Link Manufacturing

Chains are manufactured link by link, with each link being attached to the previous one. The entire process comprises five consecutive stages.

Heating The bars are heated to 800–900 °C. At VCSA three forms of heating are used: gas furnaces, Joule effect heating and induction heating.

Bending Once they are heated, the bars are placed in a bar stock bender that shapes them into links. First, one end is bent and hooked onto the last link formed. Then, the other end is bent to close the link and give it its final shape. The ends to be welded are located in the middle of one of the straight sides of the link.

Welding The length of chain with the link blank is then taken to the welding station. VCSA uses a technique called Flash-Butt Welding (FBW) [1–4], which is explained below.

Trimming During welding sticks, part of the material is expelled to the welded area in the form of trims. The trimming machine cuts off this excess material and leaves the welded area straight.

Pressing Finally, the link is taken to the press, where it is compressed on the two straight sides to give it the required design width.

2.2 Heat Treatment

Once the chain is complete it is treated with heat in continuous furnaces to consolidate the welds and give the steel the mechanical properties required for its working life. Heat treatment comprises three stages: solubilisation, quenching and tempering.

2.3 Tensile Load Testing

All links are subjected to a tensile load test section by section. This is performed due to three reasons: to ensure that the chain meets the dimensional requirements, to ensure that the welding is correct and to give the chain tensile strength.

2.4 Inspection

After load testing, all links are subjected to non-destructive testings using dye penetrants, magnetic particles and ultrasounds. If any inhomogeneities are detected, the link in question may be reworked or rejected and removed. Although it is not compulsory, many customers – especially in the case of large-diameter chains – require VCSA to perform a similar inspection prior to heat treatment.

2.5 Destructive Testing

The regulations in place require destructive tests to be carried out on simple links before the product can be accepted. These links are manufactured and attached provisionally to the rest of the chain via connectors and then removed following the final inspection. They are used to conduct tensile breaking strength tests to ensure that the link does not break until after it has borne a preset load for a given time.

Test pieces taken from the link are also submitted to laboratory tensile strength and resilience or Charpy tests. The breaking strength of the test pieces must exceed a predetermined figure.

2.6 Product Certification

The production process ends with certification by a classification society such as the American Bureau of Shipping, Det Norske Veritas or Lloyd's Register. This is an essential prerequisite for delivery of the chain to the customer.

3 Flash Butt Welding

FBW is a resistance welding technique that requires no external filler material. The link is placed on the welding bench and gripped on each side of the welding area with copper clamps, which secure it tightly. These clamps are attached to the bench and connected to an electrical circuit, so that there is a voltage difference between the two opposing surfaces to be welded. One side of the bench is fixed in position and the other is moveable, which is known as the carriage. The welding sequence is as follows (cf. Fig. 1).

Preheating By approaching and retracting the carriage, the two surfaces to be welded are brought closer and then separated several times so that short-circuits

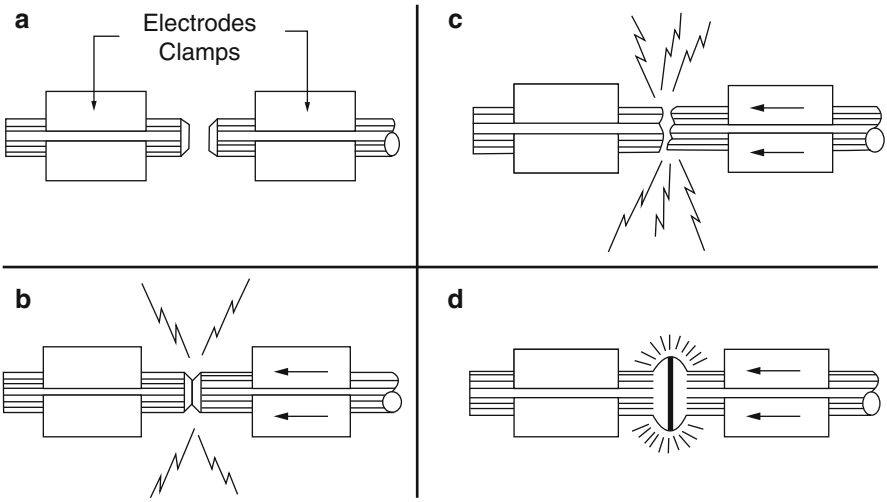


Fig. 1 FBW process: (**a**) Commencement of welding; (**b**) Preheating; (**c**) Flash; (**d**) Upsetting

are generated. Once this heats them to a preset temperature, the resistance of the material leads the voltage difference in the discharge dropping to the point where the sensor no longer detects it. The machine then moves on to the flash stage.

Flash The carriage is moved forward to keep the two surfaces in continuous contact and material is gradually consumed in the form of sparks. When the number of millimetres set via the numerical control unit has been consumed, the machine goes on to the forging or upsetting stage.

Upsetting The carriage makes a high-pressure strike that lasts 1 or 2 s. As a result, the area to be welded becomes forged and consolidated.

4 Database for the Whole Production Process

Vicinay Cadenas S.A. keeps a complete record of the manufacturing of each link, including the date and time of each stage of the process, which operator and machine performed them and any detected incidents. It also records data about the raw material (pour number & grade of steel), the cutting (bar length), the welding process (temperatures, times, travels, feed rates, pressures, current, electrical voltage, etc.), the heat treatment (temperature, feed rate in each furnace, hardness at the exit from the furnace, etc.), the load tests (maximum load applied, link dimensions before and after testing), the inspections (codes for any inhomogeneities and corrective actions if any) and the destructive tests (links tested, numerical results of breaking strength tests, maximum load applied, energy absorbed by impact, etc.) .

5 Study of Manufacturing Variables

The FBW process at VCSA is automatic. Once the machine is set up, the link blank is inserted, the start command is initialized, and the machine performs the whole weld unaided. VCSA records measurement data on 40 manufacturing parameters that are considered to influence weld quality [5]. Seven of these parameters are directly adjustable by the production staff, and the rest are not adjustable but are affected by the adjustable ones, the ambient conditions, the state of the machine, the raw material, etc. The company has the recommended figures and tolerances for these parameters set down in tables of settings for each link diameter. As indicated, this study seeks mainly to determine the extent to which the selected settings affect weld quality, with a view to optimising the settings tables when appropriate.

To study all the manufacturing variables together, two manufacturing runs on the same machine but with different chain characteristics are used. Seven thousand and two hundred and forty seven links from manufacturing run “A” and 8855 from run “B” are considered. The study divides into four parts.

Data filtering This is performed for two reasons: the first is to exclude those links whose data are unrealistic, incorrect, or are isolated to the extent that they would contaminate a statistical analysis. The second is to help to detect influential atypical links.

Correlation of variables Determining whether or not there is correlation between variables helps us to decide what variables need to be handled to cause a variation in other significant variables and bring their readings closer to the desired figures. This method makes it easier to interpret the various multivariate analyses conducted, the principal component analysis and the discriminant analysis.

Discriminant analysis This is used to determine which variables are most influential in the appearance of inhomogeneities, and to estimate the probability of a link having an inhomogeneity. In this case, there are too few links with inhomogeneities for the results to be significant enough to be useful.

Multivariate analysis. Principal components Principal component analysis represents a set of observed variables in a group of individuals or elements by a smaller number of new variables constructed by means of linear combinations of the original ones [6]. The link between principal components (PCs) and inhomogeneities is studied here. In manufacturing run "A", the first principal component (PC1) discriminates links with inhomogeneities and without inhomogeneities. Thus, links with $PC1 > -1{,}065$ (34.5 % of total) do not contain inhomogeneities. In manufacturing run "B", PC1 and PC3 together divide links in groups with high or low concentration of inhomogeneities. The main difficulty in making use of these results lies in translating the sets defined by the PCs into Cartesian product of intervals in terms of the original variables, as this is what is needed in practice to implement the regulation and rank of variation of the manufacturing variables.

6 Study of Seven Adjustable Variables: Algorithm for Locating Optimal Regions

From fabrication data of all links in a chain, the objective here is to design an algorithm that finds a region such that the number of inhomogeneities is minimum (or maximum). Out of the existing 40 variables that are recorded during the fabrication process, for this algorithm we have restricted ourselves to just seven of them. This subset of variables has been selected by Vicinay's experts based on the criteria of being the most critical during the fabrication process and at the same time can be manually adjusted.

An individual analysis of each variable has shown to provide no valuable information in terms of identifying regions with a low number of inhomogeneities. Thus, we will treat the seven variables simultaneously. Mathematically, we will identify each link with a point $\mathbf{x} \in \mathbf{R}^7$. For the algorithm to be useful, the following requirements hold:

1. The search of optimal regions in $\mathbf{R}^7$ should be restricted to tensor product of 1D intervals.
2. The length of each 1D interval should be bounded below by the minimum range of variability that limits the precision that can be achieved during fabrication.
3. The optimal region should have a minimum preset number of links, either expressed in terms of an absolute number (e.g., 1,000 or 2,000) or a relative percentage (e.g., 20 or 30 %).

The proposed algorithm consists of the following steps:

(a) For each link $\mathbf{x} \in \mathbf{R}^7$, we compute the distance to all remaining links.
(b) For each link $\mathbf{x} \in \mathbf{R}^7$, we compute the region of closest links whose size is that given by the minimum preset number of links established by condition 3 (above).
(c) We increase the size of each of the above regions as much as possible without introducing any link with inhomogeneities.
(d) Out of all the computed regions, we select the one with minimum number of inhomogeneities.

A critical point in the above algorithm is the choice of a distance. Since a distance can be obtained from a norm, and in view of the above condition 1, we select a (weighted) L-infinity norm in $\mathbf{R}^7$. This norm has the property that equal distance points can be expressed as a tensor product of 1D intervals. Thus, and algorithm based on such norm ensures that the above first condition is satisfied. To satisfy the above condition 2, we select the following weights for each variable:

$$a_1 = 0.034,\ a_2 = 0.2,\ a_3 = 0.66,\ a_4 = 0.66,\ a_5 = 0.66,\ a_6, = 13.33,\ a_7 = 0.143.$$

The above weights have been selected based on the range of variability of each variable and the expertise of senior members of Vicinay.

Table 1 shows the full ranges of the 7 variables in manufacturing run "B" and the best spread for those variables with at least 1,000 links.

Table 1 Manufacturing run "B". Full range of the adjustable variables and spreads of those variables containing 1,000 links and the lowest concentration of inhomogeneities

	Full range	Best spreads
Variable x_1	773,14–878,32	804,800–852,210
Variable x_2	163,303–172,157	166,920–172,050
Variable x_3	40,045–43,14	40,81–42,75
Variable x_4	15,162–20,558	15,65–18,03
Variable x_5	44,798–47,95	44,800–47,040
Variable x_6	0,807–0,944	0,862–0,944
Variable x_7	172,578–186,032	178,82–186,03
N^o of links	8.855	1.000
N^o of inhomogeneities		2

The designed method enables to obtain tables for setting the adjustable welding variables to be improved, so that the best spreads for those variables are selected. In addition, the best and worst "rectangular" areas of a preset size and with a minimum number of links are selected. The results are similar to those indicated above.

7 Selecting the Optimal Method

Working with technicians from Vicinay, we have assessed the results of the different methods used with a view to applying them to improve the tables of settings. We concluded that all methods of analysis used are complementary and should continue to be developed in parallel. However, the one that locates the optimal regions for the seven adjustable manufacturing variables is selected as the optimal one, since it is the method that best suits the specific needs of the industrial process in question, the know-how of the company and the physics of welding.

8 Study of Nine Manufacturing Runs from the Same Machine

The method designed for locating the best and worst spreads of the welding variables was then applied to nine manufacturing runs of chains with different diameters on the same machine, with a view to drawing up a table of settings for it. Results show that the nominal figures for the variables increase when the diameter of the links to be welded increases, in a way that is consistent with the experience of the company. This has led to a modification of the whole table of settings for the machine. The new table is an improvement on the one previously used at Vicinay Cadenas S.A. The results also provide an estimate for the table or settings that needs to be applied to tackle the manufacture of new, larger chains. At a later stage of the project, the modelling of FBW will be studied (see for instance [7]).

Acknowledgements The project has been funded entirely by Vicinay Cadenas S.A.

References

1. Dent, P., et al.: Flash, upset, and percussion welding. In Welding Handbook, American Weiding Society, pp. 581–609 (1997)
2. Ichikawa, M., Nishi, T., Saito, T.: Flash butt welding. Patent US 4506134 A, 1985
3. Ichiyama, Y., Kodama, S.: Flash-Butt Welding of High Strength Steel. Technical Report Nippon Steel, n. 95 (1997)
4. Johnson, R.: Solid state joining techniques. Mater. World **7**(11), 684–685 (1999)

5. Kim, D.C., So, W.J., Kang. M.J.: Effect of flash butt welding parameters on weld quality of mooring chain. Arch. Mater. Sci. Eng. **28**, 112–117 (2009)
6. Peña, D.: Análisis de datos multivariantes. McGraw-Hill, Madrid (2002)
7. Wang, W., Shi, Y., Lei, Y., Tian, Z.: FEM simulation on microstructure of DC flash butt welding for an ultra-fine grain steel. J. Mater. Process. Technol. **161**, 497–503 (2005)

A Model of Traffic Flow in a Network

Ángela Jiménez-Casas and Aníbal Rodríguez-Bernal

Abstract We obtain a mathematical model which governs the traffic flow of material objects in a net to generalize several previous models like air traffic (Sridhar, Menon (2005) Comparison of linear dynamic models for air traffic flow management. IFAC, Prague; Sun, Strub, Bayen (2007) Netw Heterog Media 2(4):569–594). We analyze the existence and uniqueness of solutions for some particular case.

1 Introduction

The goal of this work is to obtain a model to describe the air traffic flow when we consider discontinuous functions in time to generalize the previous model [3, 4], and such that if we assume additional hypothesis of regularity we get a system of differential equations with delay (see [1, 2]).

We consider a net given by a set of points **nodes** connected by **edges** such that the material objects is moving from node to node through edges, satisfying the followings rules:

Principles of traffic flow of material objects in a net

1. The traffic flow goes from one *node* to another *node*.
2. The nodes are connected by *edges* and the traffic goes only by *edges*.
3. On the edge connecting the node i with j only goes the material object that previously belonged to i.

Á. Jiménez-Casas (✉)
Departamento de Matemática Aplicada, Escuela Técnica Superior de Ingeniería (ICAI), Universidad Pontificia Comillas de Madrid, Calle de Alberto Aguilera 25, E-28015 Madrid, Spain
e-mail: ajimenez@upcomillas.es

A. Rodríguez-Bernal
Departamento de Matemática Aplicada, Universidad Complutense de Madrid, E-28040 Madrid, Spain

Instituto de Ciencias Matemáticas, CSIC-UAM-UC3M-UCM, E-28049 Madrid, Spain
e-mail: arober@ucm.es

F. Casas, V. Martínez (eds.), *Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 4, DOI 10.1007/978-3-319-06953-1_19

2 Formulation of the Equations of Traffic in a Network

2.1 Definitions and Notations

We consider a net given by $\mathcal{G} = (V, E)$ where $V = \{v_1, v_2, \ldots, v_N\}$ is the sets of nodes, and E, is the sets of edges (ordered pairs of different nodes of V). Hereafter we identify each node with the place, this is v_i is the node i. Thus the graph $\mathcal{G}$ is in turn represented by the adjacency matrix $G = (a_{ij})_{ij} \in \mathcal{M}_{N\times N}$, such that $a_{ij} = 1 \Leftrightarrow$ the node i is connected with the node j and $a_{ij} = 0$ if the node i is not connected with j. We agree that $a_{ii} = 1$.

If $i \neq j$ we note that

$$i \to j \quad \text{if } a_{ij} = 1 \quad \text{and} \quad i \nrightarrow j \quad \text{if } a_{ij} = 0.$$

We note that i is connected with j implies that j is connected with i for every i, j, then the graph is given by a symmetric matrix.

We denote by $N^*(i) = \{j,\ a_{ij} = 1\}$ the set of nodes connecting with the node i, given by the elements row i of the matrix associated to the graph, G, which are different to zero and $N(i) = \{j, \quad a_{ij} = 1, j \neq i\}$ the set of nodes connecting with the node i, different to i, that is $N^*(i) = N(i) \cup \{i\}$.

For $i, j = 1, 2, .., N$

- If $i \to j$, we denote by $f_{ij} \geq 0$ the number of material objects that goes over the edge (i, j). If $i \nrightarrow j$, then $f_{ij} = 0$.
- $f_{ii} \geq 0$, is the number of material objects belonging to the node i.
- If $i \to j$, we denote by $\tau_{ij} \geq 0$ the time used by the material objects to arrive in the node j from the node i, we assume that it is the same for all material objects in the net and if $i \nrightarrow j$, then $\tau_{ij} = 0$.
- $\mathcal{M}_{\mathcal{G}} = \{A * G, A \in \mathcal{M}_{N\times N}\}$ where $C := A * B$ iff $c_{ij} = a_{ij}.b_{ij}$, is the Hadamard's product. If we define $P_{\mathcal{G}}(A) = A * G$ then $P_{\mathcal{G}} = P_{\mathcal{G}}^2$ is a projection and $\mathcal{M}_{\mathcal{G}} = P_{\mathcal{G}}(\mathcal{M}_{N\times N})$.
- $\mathcal{M}_{\mathcal{G}}^+$ is a subject of $\mathcal{M}_{\mathcal{G}}$ given by $A = (a_{ij}) \in \mathcal{M}_{\mathcal{G}}^+$ iff $A \in \mathcal{M}_{\mathcal{G}}$, $a_{ij} \geq 0$ for all i, j.
- $\mathcal{M}_{\mathcal{G}}^{+,*}$ is the subject of $\mathcal{M}_{\mathcal{G}}^+$ given by $A = (a_{ij}) \in \mathcal{M}_{\mathcal{G}}^{+,*}$ iff $A \in \mathcal{M}_{\mathcal{G}}^+$, $a_{ii} = 0$ for all i.

Thus, the traffic flow in the net, or number of materials objects in the net, associated to $\mathcal{G}$ is given by a function on time

$$t \mapsto f(t) = (f_{ij}(t))_{ij} \in \mathcal{M}_{\mathcal{G}}{}^+. \tag{1}$$

In order to study the evolution of the number of material objects in the net $f(t)$, we defined the following rates.

For $i, j = 1, 2, .., N$

- If $i \to j$, we denote by " rate of takeoffs", $T_{ij}(t) \geq 0$, the rate of take offs from the node i to the node j by unit time. If $i \nrightarrow j$, then $T_{ij}(t) = 0$ and we assume also that $T_{ii}(t) = 0$.
- $T_i(t) \geq 0$ denote the total rate of take offs from the node i by unit time.
- If $i \to j$ we denote by "rate of landings", $L_{ij}(t) \geq 0$, the rate of landings over the node i belonging to the node j by unit time. If $i \nrightarrow j$ then $L_{ij}(t) = 0$ and we also assume that $L_{ii}(t) = 0$
- $L_i(t)$ denote the total rate of landings over the node i by unit time.

2.2 Constitutive Equations in a Network Traffic

Under the above notations the **Principles of traffic flow** are given by:
for $i, j \in \{1, 2, .., N\}$, if $i \to j$ then,

$$L_{ij}(t) = T_{ji}(t - \tau_{ji}).(PI). \tag{2}$$

Therefore, the number of material objects in the edge, f_{ij}, verifies:

$$f_{ij}(t) = f_{ij}(s) + \int_s^t T_{ij}(r)dr - \int_s^t L_{ji}(r)dr, \quad t \geq s$$

and using (PI)(2), $\int_s^t L_{ji}(r)dr = \int_s^t T_{ij}(r - \tau_{ij})dr = \int_{s-\tau_{ij}}^{t-\tau_{ij}} T_{ij}(r)dr$, thus

$$f_{ij}(t) = f_{ij}(s) + \int_{t-\tau_{ij}}^t T_{ij}(r)dr - \int_{s-\tau_{ij}}^s T_{ij}(r)dr.$$

This is, for $t \geq s$ the number of objects in this edge verifies:

$$f_{ij}(t) - \int_{t-\tau_{ij}}^t T_{ij}(r)dr = f_{ij}(s) - \int_{s-\tau_{ij}}^s T_{ij}(r)dr. \tag{3}$$

Now, taking into account (PI)(2) together with τ_{ij} we get

$$f_{ij}(t) = \int_{t-\tau_{ij}}^t T_{ij}(r)dr, \text{ for every } t.(PII). \tag{4}$$

By the other hand, for every $i = 1, 2, .., N$ and $t \geq s$ the number of material objects belonging to the node i verifies that

$$f_{ii}(t) = f_{ii}(s) + \int_s^t L_i(r)dr - \int_s^t T_i(r)dr(PIII). \tag{5}$$

Using (PI)(2) together with $N(i)$ we get

$$L_i(t) = \sum_{j \in N(i)} L_{ij}(t) = \sum_{j \in N(i)} T_{ji}(t - \tau_{ji}) \geq 0, T_i(t) = \sum_{j \in N(i)} T_{ij}(t) \geq 0.$$

Then, from (PIII)(5) for every $t \geq s$ we have:

$$f_{ii}(t) = f_{ii}(s) + \sum_{j \in N(i)} \int_s^t L_{ij}(r) - \sum_{j \in N(i)} \int_s^t T_{ij}(r) dr$$

with $\int_s^t L_{ij}(r) = \int_s^t T_{ji}(r - \tau_{ji}) = \int_{s-\tau_{ji}}^{t-\tau_{ji}} T_{ji}(r)$ and for every $t \geq s$ in an isolated network

$$f_{ii}(t) = f_{ii}(s) + \sum_{j \in N(i)} \int_{s-\tau_{ji}}^{t-\tau_{ji}} T_{ji}(r) - \sum_{j \in N(i)} \int_s^t T_{ij}(r) dr. \tag{6}$$

Therefore, Eqs. (4) and (6), are the constitutive equations in the network.

2.3 Initial Value Problem

To determine the traffic flow, $f(t)$, when $t \geq 0$, we note the initial value on the edge is given by $T_{ij}(t)$ in $[t_0 - \tau_{ij}, t_0]$, since from (PII)(4) we have

$$f_{ij}(t_0) = \int_{t_0-\tau_{ij}}^{t_0} T_{ij}(r) dr, \quad i \to j. \tag{7}$$

Although on the node from (PIII) (6) we need to know the initial condition

$$f_{ii}(t_0) = \theta_i^0 \in \mathbb{R}^+, \quad \forall i \in \{1, 2, \ldots, N\}. \tag{8}$$

We define

$$\tau_* = \max\{\tau_{ij}\}_{ij}$$

and fix $\theta^0 \in (\mathbb{R}^+)^N$, then given $T(t) \in \mathscr{M}_{\mathscr{G}}^{+,*}$ with $t \geq t_0 - \tau_*$, the traffic flow $f(t) \in \mathscr{M}_{\mathscr{G}}^{+}$ (1), satisfies for $i, j \in \{1, 2, \ldots, N\}$ and $t \geq t_0$

$$\begin{cases} f_{ij}(t) = \displaystyle\int_{t-\tau_{ij}}^{t} T_{ij}(r) dr, \qquad i \to j, \quad (f_{ij}(t) = 0 \quad \text{for all} \quad i, j / \quad i \nrightarrow j), \\ f_{ii}(t) = \theta_i^0 + \displaystyle\sum_{j \in N(i)} \int_{t_0-\tau_{ji}}^{t-\tau_{ji}} T_{ji}(r) dr - \sum_{j \in N(i)} \int_{t_0}^{t} T_{ij}(r) dr \end{cases} \tag{9}$$

in particular satisfy the initial data on edges and nodes.

We note, if the functions T_{ij} are continuous, then (9) is equivalent to the following ordinary differential system of retarded type [1]:

$$\begin{cases} f_{ij}(t) = 0 \quad \text{for all} \quad i, j/ \quad i \nrightarrow j, \\ f'_{ij}(t) = T_{ij}(t) - T_{ij}(t - \tau_{ij}), \qquad i \to j, \\ f'_{ii}(t) = \sum_{j \in N(i)} T_{ji}(t - \tau_{ji}) - \sum_{j \in N(i)} T_{ij}(t), \end{cases} \tag{10}$$

with $t \geq t_0$, and satisfying the initial data on the edge and on the nodes.

2.4 Functional Setting

We note that (4) together with (6) define a functional

$$f(t) = \mathscr{F}(T, s, \theta)(t), \quad \theta_i = f_{ii}(s), \quad t \geq s, \tag{11}$$

with T defined on $[s - \tau_*, t]$ with values on $\mathscr{M}_{\mathscr{G}}^{+,*}$ where $\tau_* = \max\{\tau_{ij}\}_{ij}$.

The right hand side of (4) and (6) define a linear and continuous operator $\mathscr{F}^*(T, s, \theta)$, in (T, θ), satisfying that, for every $\tau > 0$ and $s \in \mathbb{R}$,

$$\mathscr{F}^*(\cdot, s, \cdot) : L^\infty([s - \tau_*, s + \tau], \mathscr{M}_{\mathscr{G}}) \times \mathbb{R}^N \longrightarrow \mathscr{C}([s, s + \tau], \mathscr{M}_{\mathscr{G}}). \tag{12}$$

We note that (11), (12) implies **causality**, this is: in order to know $\mathscr{F}^*(T, s, \theta)(t)$ we have to know T prior to t. Thus, $\mathscr{F}$ in (11) is the restriction of the operator $\mathscr{F}^*$ on the subset $X_{s,\tau} \times (\mathbb{R}^+)^N$, i.e. $\mathscr{F} = \mathscr{F}^*|_{X_{s,\tau} \times (\mathbb{R}^+)^N}$

$$\begin{array}{rcl} \mathscr{F}(\cdot, s, \cdot) : X_{s,\tau} \times (\mathbb{R}^+)^N & \mapsto & \mathscr{C}([s, s + \tau], \mathscr{M}_{\mathscr{G}}^+) \\ (T, \theta) & \mapsto & f = \mathscr{F}(T, s, \theta) \end{array} \tag{13}$$

and is Lipschitz-continuous where

$$X_{s,\tau} = \left\{ T \in L^\infty([s - \tau_*, s + \tau], \mathscr{M}_{\mathscr{G}}^{+,*}), \mathscr{F}^*(T, s, \theta) \in \mathscr{C}\left([s, s + \tau], \mathscr{M}_{\mathscr{G}}^+\right) \right\}. \tag{14}$$

3 A Decision Operator on the Nodes

We consider now the rate of takeoff as a function depending on the material objects on the nodes of the net. This is:

$$T = \mathscr{D}(x), \text{ with } x = (f_{11}, \ldots, f_{NN}) \in (\mathbb{R}^+)^N$$

where $\mathscr{D}(x)$ is the "decision operator" and we note that now f_{ij} are given by x, since

$$f_{ij}(t) = \int_{t-\tau_{ij}}^{t} D_{ij}(x)(r)\, dr, \quad i \to j. \tag{15}$$

By this way, we have the traffic flow in this case is given by the function which describes the objects at the nodes,

$$t \mapsto x(t) \in (\mathbb{R}^+)^N$$

such that for every $i \in \{1, 2, \ldots, N\}$, x_i verifies that

$$x_i(t) = \theta_i^0 + \sum_{j \in N(i)} \int_{t_0-\tau_{ji}}^{t-\tau_{ji}} D_{ji}(x)(r)dr - \sum_{j \in N(i)} \int_{t_0}^{t} D_{ij}(x)(r)dr, \quad t \geq t_0, \tag{16}$$

where the operator $\mathscr{D}$ has to satisfy suitable properties that allow us to prove the existence and uniqueness of solution from de initial value problem (16). One of this is the Principle of causality: in order to know $\mathscr{D}(x)(t)$ we have to know x prior to t.

We assume that there exists $\tau_0 \geq 0$ with $\tau_0 \leq \tau_*$ where $\tau_* = \max\{\tau_{ij}\}$, such that for every interval $I \subset \mathbb{R}$ with $l(I) \geq \tau_0$, we have

$$\mathscr{D} : L^\infty(I, (\mathbb{R}^+)^N) \mapsto X_I \subset L^\infty(I, \mathscr{M}_{\mathscr{G}}^+) \tag{17}$$

where, as in (14),

$$X_I = \{T \in L^\infty(I, \mathscr{M}_{\mathscr{G}}^{+,*}), \quad \text{with} \quad \mathscr{F}^*(T, s, \theta) \in L^\infty(I, \mathscr{M}_{\mathscr{G}}^+)\}. \tag{18}$$

In particular, for every $s \in \mathbb{R}$ and $\tau > 0$,

$$\mathscr{D} : L^\infty([s-\tau_*, s+\tau], (\mathbb{R}^+))^N \to X_{s,\tau} \subset L^\infty([s-\tau_*, s+\tau], \mathscr{M}_{\mathscr{G}}^{+,*}) \tag{19}$$

where $X_{s,\tau} = X_I$ with $I = [s-\tau_*, s+\tau]$ is as (14); see (12).

Thus, from (11) we have that (16) is given by

$$x(t) = \mathscr{H}(x, t_0, \theta^0)(t) = \mathscr{F}_{diag}(\mathscr{D}(x), t_0, \theta^0)(t), \quad t \geq t_0. \tag{20}$$

And thus (9) becomes

$$\begin{cases} f_{ij}(t) = 0 \quad \text{for all} \quad i, j / \quad i \nrightarrow j, \\ f_{ij}(t) = \displaystyle\int_{t-\tau_{ij}}^{t} D_{ij}(x)(r)dr, \qquad i \to j, \\ x_i = \theta_i^0 + \displaystyle\sum_{j \in N(i)} \int_{t_0-\tau_{ji}}^{t-\tau_{ji}} D_{ji}(x)(r)dr - \sum_{j \in N(i)} \int_{t_0}^{t} D_{ij}(x)(r)dr \end{cases} \tag{21}$$

where $f_{ii}(t) = x_i$.

From (13) the right hand side of (21) is a function on $[t_0, t_0+\tau]$, although in (19) x is defined in $[t_0-\tau_*, t_0+\tau]$. Therefore, to determine x using the operator $\mathscr{D}$, we need to know x in $[t_0-\tau_*, t_0]$ and this is now the initial condition for (20).

By this way, given the vector $\theta^0 \in (\mathbb{R}^+)^N$ and the function $y \in L^\infty([t_0-\tau_*, t_0], (\mathbb{R}^+)^N)$, to solve (21) allow us to find $x \in L^\infty([t_0-\tau_*, t_0+h], (\mathbb{R}^+)^N)$ such that $x = y$ in $[t_0-\tau_*, t_0]$ and $x = \mathscr{F}_{diag}(\mathscr{D}(x), t_0, \theta_0)$ in $[t_0, t_0+\tau]$. In particular, from (13) we have that $x \in \mathscr{C}([t_0, t_0+\tau], (\mathbb{R}^+)^N)$ and verifies the initial conditions on the edges and on the nodes, this is:

$$f_{ij}(t_0) = \int_{t_0-\tau_{ij}}^{t_0} D_{ij}(y)(r)dr \quad i \to j,$$

$$f_{ii}(t_0) = x_i(t_0) = \theta_i^0 \in \mathbb{R}^+, \quad \text{for all } i \in \{1, 2, \ldots, N\}.$$

Theorem 1 *Under above notations and hypotheses, we consider the initial data $\theta^0 \in (\mathbb{R}^+)^N$ together with non negative function $y \in L^\infty\left([t_0-\tau_*, t_0], (\mathbb{R}^N)^+\right)$, where $\tau_* = \max\{\tau_{ij}\}_{ij}$. We assume that for some $0 < \tau < \infty$ the decision operator $\mathscr{D}$ satisfies (17)*

$$\mathscr{D} : L^\infty([t_0-\tau_*, t_0+\tau], (\mathbb{R}^N)^+) \to X_{t_0,\tau} \subset L^\infty([t_0-\tau_*, t_0+\tau], \mathscr{M}_{\mathscr{G}}^{+,*})$$

and $\mathscr{D}$ is Lipschitz with constant $L_{\mathscr{D}}$ such that

$$\tau N_0 L_{\mathscr{D}} < 1$$

where $N_0 \leq N(N-1)$ is the number or edges in the network.

Then, there exists a unique solution x of (20), (21) with $x \in L^\infty([t_0-\tau_*, t_0+\tau], (\mathbb{R}^N)^+) \cap \mathscr{C}([t_0, t_0+\tau], (\mathbb{R}^N)^+)$ such that $x = y$ in $[t_0-\tau_*, t_0]$.

We assume also that $y \in \mathscr{C}([t_0-\tau_*, t_0], (\mathbb{R}^N)^+)$.

(i) If, for some $i \in \{1, \ldots, N\}$, $y_i(t_0) \neq \theta_i^0$, then $x_i \in \mathscr{C}([t_0-\tau_*, t_0) \cup [t_0, t_0+\tau], \mathbb{R}^+)$ and at $t = t_0$ has a finite jump.

(ii) If $y_i(t_0) = \theta_i^0$, for some $i \in \{1, \ldots, N\}$, then $x_i \in \mathscr{C}([t_0-\tau_*, t_0+\tau], \mathbb{R}^+)$.

(iii) If $y(t_0) = \theta^0$ then $x \in \mathscr{C}([t_0-\tau_*, t_0+\tau], (\mathbb{R}^N)^+)$.

Proof Identifying $L^\infty\left([t_0-\tau_*, t_0+\tau], (\mathbb{R}^N)^+\right)$ with the space $L^\infty\left([t_0-\tau_*, t_0], (\mathbb{R}^N)^+\right) \times L^\infty\left([t_0, t_0+\tau], (\mathbb{R}^N)^+\right)$, we consider the operator H given by

$$\begin{aligned} H : Y &\mapsto Y \\ x &\mapsto H(x) = \mathscr{F}_{diag}(\mathscr{D}(y, x), t_0, \theta^0) \end{aligned} \tag{22}$$

where $Y = L^\infty\left([t_0, t_0+\tau], (\mathbb{R}^N)^+\right)$ such that (20) is equivalent to $x = H(x)$, $x \in Y$.

From the properties of $\mathscr{F}_{diag}$ and $\mathscr{D}$ we get H is Lipschitz, with constant $L_H \leq \tau N_0 L_{\mathscr{D}}$, so we have the existence and uniqueness of the fixed point since $\tau N_0 L_{\mathscr{D}} < 1$.

Indeed, for every $x_1, x_2 \in Y = L^\infty([t_0, t_0+\tau], (\mathbb{R}^N)^+)$ from (17) we have $T^1 = \mathscr{D}(y, x_1)$, $T^2 = \mathscr{D}(y, x_2)$, verify $T^1 = T^2$ en $[t_0 - \tau_*, t_0]$.

Thus,

$$|\mathscr{F}_{ii}(T^1, \theta^0)(t) - \mathscr{F}_{ii}(T^2, \theta^0)(t)| \leq \sum_{j \in N(i)} \int_{t_0-\tau_{ji}}^{t-\tau_{ji}} |T^1_{ji}(r) - T^2_{ji}(r)| dr +$$

$$+ \sum_{j \in N(i)} \int_{t_0}^{t} |T^1_{ij}(r) - T^2_{ij}(r)| dr$$

and using again $T^1 = T^2$ in $[t_0 - \tau_*, t_0]$, we get

$$\|\mathscr{F}_{ii}(T^1, \theta^0) - \mathscr{F}_{ii}(T^2, \theta^0)\|_{L^\infty[t_0, t_0+\tau]} \leq \tau N_0 \|T^1 - T^2\|_Y. \tag{23}$$

Therefore

$$\|\mathscr{F}(T^1, \theta^0) - \mathscr{F}(T^2, \theta^0)\|_Y \leq \tau N_0 \|T^1 - T^2\|_Y$$

and from $T^1 = \mathscr{D}(y, x_1)$, $T^2 = \mathscr{D}(y, x_2)$,

$$\|T^1 - T^2\|_Y \leq L_{\mathscr{D}} \|x_1 - x_2\|_Y$$

and we conclude.

In particular, there exists a unique solution of (20), (21) $x \in Y = L^\infty([t_0, t_0 + \tau], (\mathbb{R}^N)^+)$ such that $(y, x) \in L^\infty([t_0 - \tau_*, t_0 + \tau], (\mathbb{R}^N)^+)$ and in particular $x_i(t_0) = \theta_i^0$. Moreover, from (13) we have $x \in \mathscr{C}([t_0, t_0 + \tau], (\mathbb{R}^N)^+)$.

If we assume also $y \in \mathscr{C}([t_0 - \tau_*, t_0], (\mathbb{R}^N)^+)$ and $y_i(t_0) = \theta_i^0 = x_i(t_0)$ we have x_i is continuous function in $[t_0 - \tau_*, t_0 + \tau]$ and has a finite jump at t_0 si $y_i(t_0) \neq \theta_i = x_i(t_0)$. Finally, if $y(t_0) = \theta^0$ then $x \in \mathscr{C}([t_0 - \tau_*, t_0 + \tau], (\mathbb{R}^N)^+)$. □

Acknowledgements Partially supported by grant MTM2012-31298 from Ministerio de Economia y Competitividad, Spain, GR58/08 Grupo 920894 BSCH-UCM, Grupo de Investigación CADEDIF and by Project FIS2009-12964-C05-03, SPAIN.

References

1. Arino, O., Sanchez, E.: Linear theory of abstract functional differential equations of retarded type. J. Math. Anal. Appl. **191**, 547–571 (1995)

2. Hale, J.K., Magalhaes, L.T., Oliva, W.M.: Dynamics in Infinite Dimensions. Springer, New York (2002)
3. Sridhar, B., Menon, P.K.: Comparison of Linear Dynamic Models for Air Traffic Flow Management. Proceedings of the 16th IFAC World Congress, 1962–1968 (2005)
4. Sun, D., Strub, I.S., Bayen, A.M.: Comparison of the performance of four Eulerian network flow models for strategic air traffic network flow models for strategic air traffic management. Netw. Heterog. Media. **2**(4), 569–594 (2007)

Fire Spotting Effects in Wildland Fire Propagation

Gianni Pagnini

Abstract Wildland fire propagation is affected by events with random character. Two of them are turbulence, due to the Atmospheric Boundary Layer and to the fire-induced flow, and fire spotting, when sparks or embers are carried by convection and they start new fires when they land. Fire front position gets therefore a random character, too. A formulation which includes random effects due to both turbulence and fire spotting is discussed. It generalizes the level-set method for tracking random fronts. Under the assumption that fire spotting is a downwind-phenomenon, differences between fire propagation in the windward and in the leeward sectors are analyzed. In particular it emerges that the variability in time of the average ember jump-length and of the mean wind direction push fire advancement.

1 Introduction

Wildland fire propagation is a complex multi-scale, as well as a multi-physics and multi-discipline process, strongly influenced by the atmospheric wind. The wildland fire is fed by the fuel on the ground and displaced, beside meteorological and orographical factors, also by the hot air that pre-heats the fuel and aids the fire propagation. Heat transfer is turbulent due to the Atmospheric Boundary Layer and the fire-induced flow. In general, fire-atmosphere coupling has an important role in fire front propagation [5, 12]. Moreover, fire generates firebrands that when land on the ground are further sources of fire. Both turbulence and jump-length of firebrands are random processes that affect the fireline propagation. Hence, the fire front propagation becomes a random process, too. Accounting for the effects of turbulence and fire spotting improves the usefulness of the operational models and thereby increases the firefighting safety and in general the efficiency of efforts for fire suppression and nature preservation.

G. Pagnini (✉)
BCAM, Basque Center for Applied Mathematics, Alameda de Mazarredo 14, 48009 Bilbao, Spain

IKERBASQUE, Basque Foundation for Science, Alameda Urquijo 36–5, Plaza Bizkaia, 48011 Bilbao, Spain
e-mail: gpagnini@bcamath.org

F. Casas, V. Martínez (eds.), *Advances in Differential Equations and Applications*,
SEMA SIMAI Springer Series 4, DOI 10.1007/978-3-319-06953-1_20

Fire propagation has been mainly modelled by using reaction-diffusion type equations, see e.g. [1, 2, 10], and the level-set method, see e.g. [4, 9, 11, 12]. Here, an approach that generalizes the level-set method [19] to track random fronts is proposed to model the global random effects on fire front propagation due to turbulence and fire spotting. Actually, the reaction-diffusion equation associated to the level-set method is derived. Such approach, based on the statistical distribution of the level-set contour, has been previously introduced to account for turbulent heat transfer only [14] and here its extension to include also fire spotting is studied. Turbulent heat transport acts in both windward and leeward sector but, excluding particular situations, fire spotting can be assumed to be a downwind-phenomenon acting only in the leeward direction. Within the modelling approach proposed, differences between *windward* and *leeward* sectors are discussed. In particular, it is emerged that variability in time of the average jump-length of embers and of the direction of the mean wind enhance the propagation of the fire.

2 Model Formulation

Let $\Gamma(t)$ be the fire line contour, then in a two dimensional domain it can be represented as an isoline of an auxiliary function $\gamma(\boldsymbol{x},t)$, i.e. $\Gamma(t) = \{\boldsymbol{x},t : \gamma(\boldsymbol{x},t) = \gamma_0 = constant\}$. The evolution equation of the isoline γ_0 is given by

$$\frac{D\gamma}{Dt} = \frac{\partial\gamma}{\partial t} + \frac{d\boldsymbol{x}}{dt}\cdot\nabla\gamma = \frac{D\gamma_0}{Dt} = 0\,. \tag{1}$$

When the motion of the surface points is directed towards the normal direction it holds

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{V}(\boldsymbol{x},t) = \mathscr{V}(\boldsymbol{x},t)\,\hat{\boldsymbol{n}}\,, \quad \hat{\boldsymbol{n}} = -\frac{\nabla\gamma}{||\nabla\gamma||}\,, \tag{2}$$

and (1) becomes

$$\frac{\partial\gamma}{\partial t} = \mathscr{V}(\boldsymbol{x},t)\,||\nabla\gamma||\,, \tag{3}$$

which is the *ordinary* level-set equation. Let $\varphi(\gamma(\boldsymbol{x},t))$ be an indicator function such that

$$\varphi(\gamma(\boldsymbol{x},t)) = \begin{cases} 0\,,\ \gamma(\boldsymbol{x},t) \le \gamma_0\,, \\ \\ 1\,,\ \gamma(\boldsymbol{x},t) > \gamma_0\,, \end{cases} \tag{4}$$

then, in wildland fire propagation, it is assumed that $\varphi(\boldsymbol{x},t) = 1$ marks the burned area $\Omega(t)$, i.e. $\Omega(t) = \{\boldsymbol{x},t : \varphi(\boldsymbol{x},t) = 1\}$, and $\varphi(\boldsymbol{x},t) = 0$ marks the unburned

area, i.e. $\boldsymbol{x} \notin \Omega(t)$. The boundary of $\Omega(t)$ is $\Gamma(t)$, that is the front line contour of the wildland fire. In literature models [4, 9, 11, 12], quantity $\mathscr{V}(\boldsymbol{x}, t)$ is identified with the so-called Rate Of Spread (ROS). Several determinations of the ROS have been proposed, some of them are based on experimental data and others on physical insight, see e.g. [3, 6, 7, 9, 17]. The present formulation holds for any determination of the ROS.

Let the burning fireline be embodied by a large number of *active* flame holders. Let the motion of each *active* flame holder belonging to the fireline be random due to turbulence and fire spotting effects. For any realization indexed by ω, the random trajectory of each *active* flame holder is stated to be $\boldsymbol{X}^{\omega}(t, \overline{\boldsymbol{x}}_0) = \overline{\boldsymbol{x}}_{ROS}(t, \overline{\boldsymbol{x}}_0) + \chi^{\omega} + \xi^{\omega}$, where χ and ξ are two random noises that reproduce the randomness of turbulence and fire spotting. The deterministic component $\overline{\boldsymbol{x}}_{ROS}$ corresponds to the motion obtained by literature determination of the ROS [3, 6, 7, 9, 17]. By using statistical mechanics formalism [8], the trajectory of a single *active* flame holder is marked out by the one-particle density function $f^{\omega}(\boldsymbol{x}; t) = \delta(\boldsymbol{x} - \boldsymbol{X}^{\omega}(t, \overline{\boldsymbol{x}}_0))$, where $\delta(\boldsymbol{x})$ is the Dirac-delta function. The random trajectory $\boldsymbol{X}(t, \overline{\boldsymbol{x}}_0)$ has the same fixed initial condition $\boldsymbol{X}^{\omega}(0, \overline{\boldsymbol{x}}_0) = \overline{\boldsymbol{x}}_{ROS}(0, \overline{\boldsymbol{x}}_0) = \overline{\boldsymbol{x}}_0$ in all realizations. Let $\gamma(\overline{\boldsymbol{x}}_0, 0)$ be the initial fixed fireline contour, the evolution in time of the fireline according to the ω-realization of the trajectories of the *active* flame holders follows to be

$$\gamma^{\omega}(\boldsymbol{x}(t)) = \int_{\Gamma_0} \gamma(\overline{\boldsymbol{x}}_0, 0)\, \delta(\boldsymbol{x} - \boldsymbol{X}^{\omega}(t, \overline{\boldsymbol{x}}_0))\, d\overline{\boldsymbol{x}}_0\,, \tag{5}$$

where $\Gamma_0 = \{\boldsymbol{x} : \gamma(\overline{\boldsymbol{x}}, 0) = \gamma_0\}$.

Denoting by $\langle\cdot\rangle$ the ensemble average, the average trajectory $\langle \boldsymbol{X}(t; \overline{\boldsymbol{x}}_0)\rangle = \overline{\boldsymbol{x}}(t, \overline{\boldsymbol{x}}_0)$ is driven by the deterministic velocity field $d\overline{\boldsymbol{x}}/dt = \boldsymbol{V}(\overline{\boldsymbol{x}}, t)$. Then, trajectory $\overline{\boldsymbol{x}}(t, \overline{\boldsymbol{x}}_0)$ emerges to be time-reversible and the Jacobian of the transformation follows to be $J = d\overline{\boldsymbol{x}}_0/d\overline{\boldsymbol{x}} \neq 0$. When the fireline length $\mathscr{L}(t)$ grows, the number $\mathscr{N}(t)$ of the *active* flame holders composing the fireline grows as well. Then the growing ratio of the fireline, i.e. $\mathscr{L}(t)/\mathscr{L}(0)$, and that of the number of the *active* flame holders, i.e. $\mathscr{N}(t)/\mathscr{N}(0)$, are equal. Hence, to each *active* flame holder it can be associated an *action length* d stated as $d = \mathscr{L}(t)/\mathscr{N}(t) = \mathscr{L}(0)/\mathscr{N}(0) = constant$. As a consequence of this reasoning, a condition of incompressibility type follows: $J = 1$. Finally, by time inversion and ensemble averaging, from (5) the effective fire front contour emerges to be in terms of the indicator function $\varphi(\boldsymbol{x}, t)$ as follows

$$\begin{aligned}\langle\varphi^{\omega}(\boldsymbol{x}(t))\rangle &= \langle \int_{R^2} \varphi(\overline{\boldsymbol{x}}, t)\delta(\boldsymbol{x} - \boldsymbol{X}^{\omega}(t, \overline{\boldsymbol{x}}))d\overline{\boldsymbol{x}}\rangle = \int_{R^2} \varphi(\overline{\boldsymbol{x}}, t)\langle\delta(\boldsymbol{x} - \boldsymbol{X}^{\omega}(t, \overline{\boldsymbol{x}}))\rangle d\overline{\boldsymbol{x}} \\ &= \int_{R^2} \varphi(\overline{\boldsymbol{x}}, t)\, f(\boldsymbol{x}; t|\overline{\boldsymbol{x}})\, d\overline{\boldsymbol{x}} = \varphi_e(\boldsymbol{x}, t)\,,\end{aligned} \tag{6}$$

where $f(\boldsymbol{x};t|\overline{\boldsymbol{x}}) = \langle\delta(\boldsymbol{x}-\boldsymbol{X}^{\omega}(t,\overline{\boldsymbol{x}}))\rangle$ is the probability density function (PDF) of the distribution of the particles of the fireline contour around the average front location $\overline{\boldsymbol{x}}$ and the definition of $\varphi(\overline{\boldsymbol{x}},t)$ stated in (4) has been used.

Field variable $\varphi_e(\boldsymbol{x},t)$ is computed from formula (6) where indicator function $\varphi(\overline{\boldsymbol{x}},t)$ follows from solving the level-set equation driven by the average front velocity as discussed in Sect. 3.2, see Eqs. (19) and (22). The pure deterministic motion governed by the level-set equation (3) is recovered when $f(\boldsymbol{x};t|\overline{\boldsymbol{x}}) \to \delta(\boldsymbol{x}-\overline{\boldsymbol{x}})$.

Since the effective fireline contour $\varphi_e(\boldsymbol{x},t)$ is a smooth function continuously ranging from 0 to 1, a criterion to mark burned points have to be stated. For example, points $\boldsymbol{x}$ such that $\varphi_e(\boldsymbol{x},t) > 0.5$ are marked as burned and the effective burned area emerges to be $\Omega_e(t) = \{\boldsymbol{x}, t : \varphi_e(\boldsymbol{x},t) > 0.5\}$. However, beside this criterion, a further criterion associated to an ignition delay due to the pre-heating action of the hot air or to the landing of firebrands is introduced. Hence, in the proposed modelling approach, an unburned point $\boldsymbol{x}$ will be marked as burned when one of these two criteria is met.

This ignition delay, due to a certain *heating-before-burning mechanism*, can be depicted as an accumulation in time of heat [14], i.e.

$$\psi(\boldsymbol{x},t) = \int_0^t \varphi_e(\boldsymbol{x},\eta)\,\frac{d\eta}{\tau}\,, \tag{7}$$

where $\psi(\boldsymbol{x},0) = 0$ corresponds to the unburned initial condition and τ is a characteristic ignition delay that can be understood as an electrical resistance. Since the fuel can burn because of two pathways, i.e. hot-air heating and firebrand landing, the resistance analogy suggests that τ can be approximatively computed as resistances acting in parallel, i.e.

$$\frac{1}{\tau} = \frac{1}{\tau_h} + \frac{1}{\tau_f} = \frac{\tau_f + \tau_h}{\tau_h \tau_f}\,, \tag{8}$$

where τ_h and τ_f are the ignition delays due to hot air and firebrands, respectively.

The amount of heat is proportional to the increasing of the fuel temperature $T(\boldsymbol{x},t)$, then

$$\psi(\boldsymbol{x},t) \propto \frac{T(\boldsymbol{x},t) - T(\boldsymbol{x},0)}{T_{ign} - T(\boldsymbol{x},0)}\,, \quad T(\boldsymbol{x},t) \le T_{ign}\,, \tag{9}$$

where T_{ign} is the ignition temperature. Finally, when $\psi(\boldsymbol{x},t) = 1$ the ignition temperature is assumed to be reached, so that a new ignition occurs in $(\boldsymbol{x},t)$ and, with reference to (6), the modelled fire goes on by setting $\varphi(\boldsymbol{x},t) = 1$, see the numerical algorithm in Sect. 4.

3 Windward and Leeward Differences

The windward sector of the fireline propagation is assumed to be affected solely by turbulence while the leeward is affected by both turbulence and fire spotting. In this Section the differences between the two sectors are analyzed within the modelling approach described above. The indices w and ℓ refer to *windward* and *leeward* quantities, respectively.

3.1 Particle Probability Density Function

Since $\varphi(\boldsymbol{x}, t)$ is an indicator function, formula (6) turns out to be

$$\varphi_e(\boldsymbol{x}, t) = \int_{\Omega(t)} f(\boldsymbol{x}; t|\overline{\boldsymbol{x}})\, d\overline{\boldsymbol{x}}\,, \tag{10}$$

that was originally proposed to model the burned mass fraction in turbulent premixed combustion [13]. By applying the Reynolds transport theorem to (10), the evolution equation of the effective fire front $\varphi_e(\boldsymbol{x}, t)$ is [13]

$$\frac{\partial \varphi_e}{\partial t} = \int_{\Omega(t)} \frac{\partial f}{\partial t}\, d\overline{\boldsymbol{x}} + \int_{\Omega(t)} \nabla_{\overline{\boldsymbol{x}}} \cdot [\boldsymbol{V}(\overline{\boldsymbol{x}}, t)\, f(\boldsymbol{x}; t|\overline{\boldsymbol{x}})]\, d\overline{\boldsymbol{x}}\,. \tag{11}$$

Equation (11) is the reaction-diffusion equation associated to the level-set equation (3).

An important property of the proposed approach is the possibility to manage real world situations when a fire overcomes a zone without fuel, e.g. roads, firebreak lines, rivers. On the contrary, by using the classical level-set method this issue can not be solved, because when there is no fuel the ROS is null and the fire front stops, see Eq. (3). Indeed, in the present formulation (11), when the ROS is null the fireline spreading is driven by the joint action of the turbulent motion of the hot air and fire spotting in terms of the particle PDF $f(\boldsymbol{x}; t|\overline{\boldsymbol{x}})$ according to the equation

$$\frac{\partial \varphi_e}{\partial t} = \int_{\Omega(t)} \frac{\partial f}{\partial t}\, d\overline{\boldsymbol{x}}\,. \tag{12}$$

Hence, in this approach modelling of random processes is embodied by the PDF $f(\boldsymbol{x}; t|\overline{\boldsymbol{x}})$.

This PDF results from the sum of two independent random variables, i.e. $(\overline{\boldsymbol{x}}_{ROS} + \chi)$ and ξ, regarding turbulence and fire spotting. This means that $f(\boldsymbol{x}; t|\overline{\boldsymbol{x}})$ is determined by the convolution between the PDF corresponding to $(\overline{\boldsymbol{x}}_{ROS} + \chi)$, hereinafter labeled as G, and the PDF corresponding to ξ, hereinafter labeled as q. Embers are pushed by the atmospheric mean wind $\boldsymbol{U}$ and land to a certain

distance $\ell^\omega = ||\xi^\omega||$ from the fireline. The mean wind $\boldsymbol{U}$ is assumed to be the same in all realizations and with the same direction all over the domain involved, i.e. $\hat{\boldsymbol{n}}_U = \hat{\boldsymbol{n}}_U(t)$. Then, the effect of ξ emerges to be positive and aligned with the mean wind direction, i.e. $\xi^\omega = \ell^\omega \, \hat{\boldsymbol{n}}_U$. Turbulent noise χ is a zero-mean noise, i.e. $\langle \chi \rangle = 0$, while fire-spotting noise ξ has a positive mean value, i.e. $\langle \ell \rangle > 0$. Finally

$$f(\boldsymbol{x}; t|\overline{\boldsymbol{x}}) = \int_0^{+\infty} G(\boldsymbol{x} - \overline{\boldsymbol{x}}_{ROS} - \ell \, \hat{\boldsymbol{n}}_U; t) \, q(\ell; t) \, d\ell \,, \tag{13}$$

and the average front position is $\overline{\boldsymbol{x}}(t) = \int_{R^2} \boldsymbol{x} \; f(\boldsymbol{x}; t|\overline{\boldsymbol{x}}) \, d\boldsymbol{x} = \overline{\boldsymbol{x}}_{ROS} + \langle \ell \rangle$. Let α be the angle between $\hat{\boldsymbol{n}}$ and $\hat{\boldsymbol{n}}_U$, then in the *leeward* fireline sector, i.e. $0 \le \alpha < \pi/2$,

$$f_\ell(\boldsymbol{x}; t|\overline{\boldsymbol{x}}) = \int_0^{\infty} G(\boldsymbol{x} - \overline{\boldsymbol{x}}_{ROS} - \ell \, \hat{\boldsymbol{n}}_U; t) \, q(\ell; t) \, d\ell \,, \quad \overline{\boldsymbol{x}}(t) = \overline{\boldsymbol{x}}_{ROS} + \langle \ell \rangle \, \hat{\boldsymbol{n}}_U \,, \tag{14}$$

otherwise in the *windward* sector $f_w(\boldsymbol{x}; t|\overline{\boldsymbol{x}}) = G(\boldsymbol{x} - \overline{\boldsymbol{x}}_{ROS}; t)$ with $\overline{\boldsymbol{x}}(t) = \overline{\boldsymbol{x}}_{ROS}$. The presence of fire spotting enlarges the burned area because $||\overline{\boldsymbol{x}}|| > ||\overline{\boldsymbol{x}}_{ROS}||$.

If turbulent heat transfer is modelled by a Gaussian PDF with diffusion coefficient $\mathscr{D}$, then the spreading around the fireline position $\overline{\boldsymbol{x}}_{ROS}$ is described by

$$\frac{\partial G}{\partial t} = \mathscr{D} \, \nabla^2 G \,, \quad G(\boldsymbol{x}; 0|\overline{\boldsymbol{x}}_{ROS}) = \delta(\boldsymbol{x} - \overline{\boldsymbol{x}}_{ROS}) \,. \tag{15}$$

Hence, from (13), the evolution equation for $f(\boldsymbol{x}; t|\overline{\boldsymbol{x}})$ follows to be

$$\begin{aligned} \frac{\partial f}{\partial t} &= \mathscr{D} \, \nabla^2 f + \int_0^{\infty} G(\boldsymbol{x} - \overline{\boldsymbol{x}}_{ROS} - \ell \, \hat{\boldsymbol{n}}_U; t) \, \frac{\partial q(\ell; t)}{\partial t} \, d\ell \\ &\quad + \nabla_{\overline{\boldsymbol{x}}} \cdot \frac{d\hat{\boldsymbol{n}}_U}{dt} \int_0^{\infty} G(\boldsymbol{x} - \overline{\boldsymbol{x}}_{ROS} - \ell \, \hat{\boldsymbol{n}}_U; t) \, \ell \, q(\ell; t) \, d\ell \,, \end{aligned} \tag{16}$$

where the property $\nabla_{\boldsymbol{x}} \, G = -\nabla_{\overline{\boldsymbol{x}}} \, G$ has been used. Finally, inserting (16) into (11), it follows

$$\begin{aligned} \frac{\partial \varphi_e}{\partial t} &= \mathscr{D} \, \nabla^2 \varphi_e + \int_{\Omega(t)} \left\{ \int_0^{\infty} G(\boldsymbol{x} - \overline{\boldsymbol{x}}_{ROS} - \ell \, \hat{\boldsymbol{n}}_U; t) \, \frac{\partial q(\ell; t)}{\partial t} \, d\ell \right\} d\overline{\boldsymbol{x}} \\ &\quad + \int_{\Omega(t)} \nabla_{\overline{\boldsymbol{x}}} \cdot \frac{d\hat{\boldsymbol{n}}_U}{dt} \left[\int_0^{\infty} G(\boldsymbol{x} - \overline{\boldsymbol{x}}_{ROS} - \ell \, \hat{\boldsymbol{n}}_U; t) \, \ell \, q(\ell; t) \, d\ell \right] d\overline{\boldsymbol{x}} \\ &\quad + \int_{\Omega(t)} \nabla_{\overline{\boldsymbol{x}}} \cdot [\boldsymbol{V}(\overline{\boldsymbol{x}}, t) \, f(\boldsymbol{x}; t|\overline{\boldsymbol{x}})] \, d\overline{\boldsymbol{x}} \,, \end{aligned} \tag{17}$$

in which clearly the variability in time of heat transfer generates a diffusive behaviour (the first term in RHS) while the variability in time of the fire spotting generates two source terms (the second and the third term in RHS) that sum to the source term driven by the velocity field $\boldsymbol{V}$ (the fourth term in RHS). Then, the variability in time of fire spotting enhances the ROS as it will be shown in the next Section.

Equation (17) is more difficult to be solved than other reaction-diffusion equations in literature [1,2,10]. The only case exactly solved is the plane fire front limit, see Sect. 3.2. However, it is here reminded that $\varphi_e(\boldsymbol{x},t)$ is practically computed by using (6) and Eq. (17) has been derived with the aim to understood the role of each involved process.

3.2 Average Front Velocity

In the *windward* sector it holds $\overline{\boldsymbol{x}} = \overline{\boldsymbol{x}}_{ROS}$ and then the average front velocity is

$$\frac{d\overline{\boldsymbol{x}}}{dt} = \frac{d\overline{\boldsymbol{x}}_{ROS}}{dt} = \mathscr{V}_{ROS}(\overline{\boldsymbol{x}},t)\,\hat{\boldsymbol{n}}\,, \tag{18}$$

such that its modulus equals the value of the ROS as derived in literature [3,6,7,9,17]. Then, in the *windward* sector, the level-set equation reads

$$\frac{\partial \gamma_w}{\partial t} = \mathscr{V}_{ROS}(\boldsymbol{x},t)\,||\nabla\gamma_w||\,. \tag{19}$$

What concerns the *leeward* sector, it holds $\overline{\boldsymbol{x}}(t) = \overline{\boldsymbol{x}}_{ROS}(t) + \langle\ell(t)\rangle\,\hat{\boldsymbol{n}}_U$ and then

$$\begin{aligned}\frac{d\overline{\boldsymbol{x}}}{dt} = \frac{d}{dt}(\overline{\boldsymbol{x}}_{ROS}(t) + \langle\ell(t)\rangle\,\hat{\boldsymbol{n}}_U) &= \mathscr{V}_0(\overline{\boldsymbol{x}},t)\,\hat{\boldsymbol{n}} + \frac{d\langle\ell\rangle}{dt}\,\hat{\boldsymbol{n}}_U + \langle\ell\rangle\,\frac{d\hat{\boldsymbol{n}}_U}{dt} \\ &= \mathscr{V}_0(\overline{\boldsymbol{x}},t)\,\hat{\boldsymbol{n}} + \boldsymbol{V}_f + \boldsymbol{V}_U\,,\end{aligned} \tag{20}$$

where $\boldsymbol{V}_f$ and $\boldsymbol{V}_U$ represent the components of the average front velocity due to the time variability of the average ember jump-length $\langle\ell\rangle$ and of the mean wind direction $\hat{\boldsymbol{n}}_U$, respectively.

Formula (20) is a key result of the present approach. In fact, when fire spotting occurs, the *effective* fire spreading expressed by the modulus of the average front velocity is

$$\mathscr{V}(\boldsymbol{x},t) = \mathscr{V}_{ROS}(\boldsymbol{x},t) + \boldsymbol{V}_f\cdot\hat{\boldsymbol{n}} + \boldsymbol{V}_U\cdot\hat{\boldsymbol{n}}\,. \tag{21}$$

Finally, in the *leeward* sector the level-set equation turns out to be

$$\frac{\partial \gamma_\ell}{\partial t} = (\mathscr{V}_0(\boldsymbol{x},t) + \boldsymbol{V}_f \cdot \hat{\boldsymbol{n}} + \boldsymbol{V}_U \cdot \hat{\boldsymbol{n}})\, ||\nabla \gamma_\ell|| \,. \tag{22}$$

In particular, the computation of $\boldsymbol{V}_f$ follows from the chosen parameterization of $\langle \ell \rangle$ and the computation of $\boldsymbol{V}_U$ from the behaviour of the mean wind direction $\hat{\boldsymbol{n}}_U(t)$. Examples of fire spotting parameterization can be found in [15, 16, 18].

3.3 Plane Fire Front Limit: Exact Solution

When only turbulence is acting, the effective front contour is determined by

$$\varphi_{eT}(\boldsymbol{x},t) = \int_{\Omega(t)} G(\boldsymbol{x} - \overline{\boldsymbol{x}};t)\, d\overline{\boldsymbol{x}} \,, \quad \Omega = \Omega_w \, U \, \Omega_\ell \,, \tag{23}$$

then, in the *windward* sector, i.e. $\boldsymbol{x} \in \Omega_w$, $\varphi_w(\boldsymbol{x},t) = \varphi_{eT}(\boldsymbol{x},t)$ and in the *leeward* sector, i.e. $\boldsymbol{x} \in \Omega_\ell$, it holds

$$\varphi_\ell(\boldsymbol{x},t) = \int_0^\infty \varphi_{eT}(\boldsymbol{x} - \ell\, \hat{\boldsymbol{n}}_U;t)\, q(\ell,t)\, d\ell \,. \tag{24}$$

For a plane front with Gaussian turbulence, the exact solution of the pure-turbulent process is [13]

$$\varphi^G_{eT}(x,t) = \frac{1}{2}\left\{ \mathrm{Erfc}\left[\frac{x - \mathscr{L}_R(t)}{2\sqrt{\mathscr{D}\, t}} \right] - \mathrm{Erfc}\left[\frac{x - \mathscr{L}_L(t)}{2\sqrt{\mathscr{D}\, t}} \right] \right\} \,, \tag{25}$$

and then in the *leeward* sector the exact solution turns out to be

$$\varphi_\ell(x,t) = \int_0^\infty \varphi^G_{eT}(x - \ell, t)\, q(\ell,t)\, d\ell \,, \tag{26}$$

where Erfc is the complementary Error function and $\mathscr{L}_R$ and $\mathscr{L}_L$ are the right and left deterministic fronts, i.e. $\Omega(t) = [\mathscr{L}_L(t); \mathscr{L}_R(t)]$. These results are important for comparison with the ordinary level-set method in the "cold" case and in the "hot" case, i.e. without and with taking into account the heating-before-burning mechanism (7).

In fact, consider exact solution (25) for the "cold" case. If $\mathscr{V} = constant$, the right front position of the level-set contour is $\mathscr{L}_R = \mathscr{L}_0 + \mathscr{V} t$ and for the randomized level-set it holds for $0 < t < \infty$

$$\varphi^G_{eT}(\mathscr{L}_R,t) = \frac{1}{2}\left\{ 1 - \mathrm{Erfc}\left[\frac{\mathscr{L}_R - \mathscr{L}_L}{2\sqrt{\mathscr{D}\, t}} \right] \right\} < \frac{1}{2} \,. \tag{27}$$

Then setting the right front position $\mathscr{L}_R$ in the leeward sector, by using (26), it follows

$$\varphi_\ell(\mathscr{L}_R, t) = \frac{1}{2}\left\{1 - \int_0^\infty \mathrm{Erfc}\left[\frac{\mathscr{L}_R - \mathscr{L}_L - \ell}{2\sqrt{\mathscr{D}\,t}}\right] q(\ell, t)\, d\ell\right\} < \frac{1}{2}\,, \tag{28}$$

because functions inside integral are positive. Hence both the "cold" isolines $\varphi_{eT}^G = 1/2$ and $\varphi_\ell(\boldsymbol{x}, t) = 1/2$ in the windward and leeward sector, respectively, are slower than the ordinary level-set contour.

What concerns the "hot" case, if only turbulence is considered, when $\psi(\boldsymbol{x}, \Delta t) = 1$ then $\tau = \tau_h = \int_0^{\Delta t} \varphi_{eT}(\boldsymbol{x}, t)\, dt$. Since $\varphi_{eT}(\boldsymbol{x}, t) = \int_{\Omega(t)} G(\boldsymbol{x}; t|\overline{\boldsymbol{x}})\, d\overline{\boldsymbol{x}} = \int_{R^2} \varphi(\overline{\boldsymbol{x}}, t) G(\boldsymbol{x}; t|\overline{\boldsymbol{x}})\, d\overline{\boldsymbol{x}}$ and $G(\boldsymbol{x}; t|\overline{\boldsymbol{x}}) = \delta(\boldsymbol{x} - \overline{\boldsymbol{x}}(0)) + \mathscr{D}\int_0^t \nabla^2 G(\boldsymbol{x}; s|\overline{\boldsymbol{x}})\, ds$, by substitution it follows

$$\tau_h = \Delta t + \mathscr{D}\int_0^{\Delta t} \left\{\int_{R^2}\left[\int_0^t \nabla^2 G(\boldsymbol{x}; s|\overline{\boldsymbol{x}})\, ds\right] \varphi(\overline{\boldsymbol{x}}, t)\, d\overline{\boldsymbol{x}}\right\}\, dt\,. \tag{29}$$

In the deterministic case $\mathscr{D} = 0$ then $\tau_h = \Delta t$. Hence when $\nabla^2 G > 0$, i.e. $||\boldsymbol{x}|| > ||\overline{\boldsymbol{x}}|| + \sqrt{2\mathscr{D}\Delta t}$, the "hot" front is faster because for fixed τ_h it holds $\Delta t < \tau_h$.

What concerns the leeward sector, if $q(\ell, t) = q(\ell)$, at the distance r from the main perimeter in the outward direction $\hat{\boldsymbol{n}}$ it holds

$$\psi_e(r, \Delta t) = \int_0^\infty \psi_{eT}(r - \ell, t)\, q(\ell)\, d\ell\,. \tag{30}$$

Since $\psi(r_a, \Delta t) > \psi(r_b, \Delta t)$ when $r_a < r_b$, then

$$\psi_e(r, \Delta t) = \int_0^\infty \psi_{eT}(r - \ell, t)\, q(\ell)\, d\ell > \psi_{eT}(r, \Delta t)\,, \tag{31}$$

and $\psi(r, \Delta_a) > \psi(r, \Delta_b)$, when $\Delta_a > \Delta_b$. Finally $\psi_e(r, \Delta t) = 1 = \psi_{eT}(r, \Delta t')$ such that it holds $\Delta t < \Delta t' < \tau$.

4 Numerical Results

A wildland fire propagating in a flat terrain covered by an idealized Pinus ponderosa ecosystem is simulated as a simple case study. Simulation details are reported in the caption of Fig. 1. The study of further cases can be found in [15].

The adopted numerical algorithm is the following:

1. The central difference approximation of the gradient of the level-set function is calculated and the gradient is normalized to obtain the unit normal to the front.

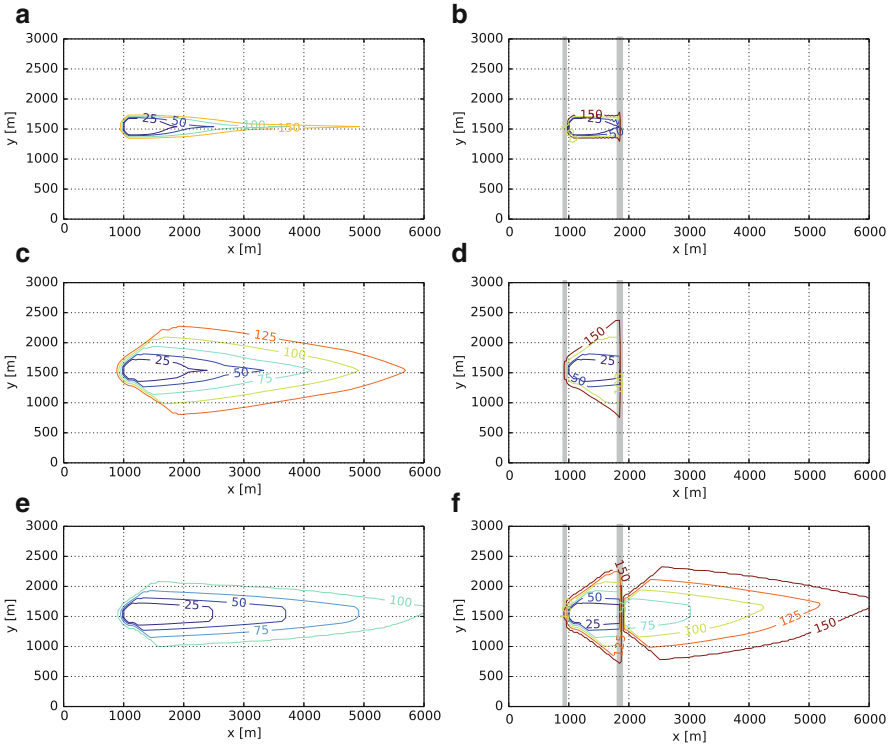


Fig. 1 Plots show the level-set method (*top row*), the present modelling approach when only turbulence (*middle row*) and when both turbulence and fire spotting are considered (*bottom row*). The labels on the contour lines indicate the elapsed time in minutes. Fire spotting has been parameterized according to [16, 18]. In particular, a stationary log-normal distribution for jump-length of embers is chosen with mean $\mu = \langle \ln \ell \rangle$ and standard deviation $s = \langle (\ln \ell - \mu)^2 \rangle$ stated equal to: $\mu = 1.32\, I_f^{0.26}\, U_t^{0.11} - 0.02$ and $s = 4.95\, I_f^{-0.01}\, U_t^{-0.02} - 3.48$, where U_t is the modulus of the mean wind as measured at the top of the tree canopy (10 m [18]) and assumed constant both in value (17.88 m s^{-1}) and direction (x-axis), and $I_f = I + I_t$ where $I = 20{,}000\,\mathrm{kW\,m^{-1}}$ is the fire intensity and $I_t = 0.015\,\mathrm{kW\,m^{-1}}$ is the tree torching intensity. Since temporal constancy of fire spotting statistics and mean wind direction, it holds $\boldsymbol{V}_f = \boldsymbol{V}_U = 0$. Other simulation parameters are: $V_{ROS} = I/(H w_0)$ where $H = 22{,}000\,\mathrm{kJ\,kg^{-1}}$ is the fuel low heat of combustion and $w_0 = 2.243\,\mathrm{kg\,m^{-2}}$ is the oven-dry mass of fuel, $\mathscr{D} = 0.04\,\mathrm{m^2\,s^{-1}}$, $\tau_h = 600\,\mathrm{s}$, $\tau_f = 60\,\mathrm{s}$ and the width of firebreaks is 60 m in the windward sector and 90 m in the leeward sector

2. The modulus of the average front velocity $\mathscr{V}(\boldsymbol{x}, t)$ is calculated in each point of the Cartesian grid.
3. The first stage of the Total Variation Diminishing (TVD) Runge–Kutta scheme is completed to obtain an approximation of the new value of the level-set function for the next time step.
4. Steps from 1 to 3 are repeated using the new value of the level-set function.
5. The second stage of the TVD Runge–Kutta scheme gives the new value of the level-set function.

6. The new value of $\varphi_e(\boldsymbol{x},t)$ is calculated through numerical integration of the product of $\varphi(\boldsymbol{x},t)$ times the PDF $f(\boldsymbol{x};t|\overline{\boldsymbol{x}})$ as stated in (6). If $\varphi_e(\boldsymbol{x},t) > 0.5$ then point $\boldsymbol{x}$ is marked as burned.
7. Function $\psi(\boldsymbol{x},t)$ is updated for each point by integration in time with the current value of $\varphi_e(\boldsymbol{x},t)$. In any point with $\psi(\boldsymbol{x},t) > 1$, the ignition is possible so point $\boldsymbol{x}$ is marked as burned and $\varphi(\boldsymbol{x},t) = 1$ is stated to allow ignition and simulated fire goes on.
8. Current time is updated as well as the level-set function and the operations are repeated for a new time step.

The present analysis constitutes a proof-of-concept and it needs to be subjected to a future validation. Hence, numerical results are understood as explorative exercises to investigate the potentialities of the approach. From comparison of the level-set method against the proposed model when only turbulence and when both turbulence and fire spotting are taken into account, it emerges the suitability of the proposed approach to simulate a fire that overcomes a firebreak zone, in contrast to the level-set method. Moreover, it emerges also that the inclusion of turbulence allows for simulating fire flank and backing fire and the inclusion of fire spotting strongly enhances the frontline propagation. This richness of model behaviours supports the proposed formulation as a promising approach to simulate the complex phenomenology of real wildland fire propagation.

Conclusions

In this paper an approach to model the effects of random processes occurring in wildland fire propagation is presented. The random processes considered are turbulence and fire spotting. The fireline propagation is modelled as a particle random trajectory problem. The resulting governing equation emerges to be a reaction-diffusion equation associated to the level-set method. Random processes have been inserted into the level-set approach by randomizing the position of the contour points. This formulation emerges to be suitable to manage real world dangerous situations such as the faster propagation and the overcoming of a break-fire because of the diffusion of the hot air and embers jumping. An important obtained result is the determination of the effective fire spread which includes fire spotting.

Acknowledgements T. Chacón, C. Parés and M.E. Vázquez are acknowledged for the invitation to give a talk at the special session *Matemáticas del Planeta Tierra* during the XXIII CEDYA / XIII CMA in Castellón, Spain, September 9–13 2013. Moreover, A. Mentrelli is also acknowledge for pictures and many useful discussions. This research is supported by the Basque Government through the BERC 2014–2017 program and by the Spanish Ministry of Economy and Competitiveness MINECO: BCAM Severo Ochoa accreditation SEV-2013-0323.

References

1. Asensio, M.I., Ferragut, L.: On a wildland fire model with radiation. Int. J. Numer. Methods Eng. **54**, 137–157 (2002)
2. Babak, P., Bourlioux, A., Hillen, T.: The effect of wind on the propagation of an idealized forest fire. SIAM J. Appl. Math. **70**, 1364–1388 (2009)
3. Balbi, J.H., Morandini, F., Silvani, X., Filippi, J.B., Rinieri, F.: A physical model for wildland fires. Combust. Flame **156**, 2217–2230 (2009)
4. Beezley, J.D., Chakraborty, S., Coen, J.L., Douglas, C.C., Mandel, J., Vodacek, A., Wang, Z.: Real-time data driven wildland fire modeling. In: Bubak, M., van Albada, G.D., Dongarra, J., Sloot, P.A.M. (eds.) Proceedings of ICCS 2008, Kraków, June 2008. Lecture Notes in Computer Science, vol. 5103, pp. 46–53. Springer, Heidelberg (2008)
5. Filippi, J.B., Bosseur, F., Mari, C., Lac, C., Lemoigne, P., Cuneot, B., Veynante, D., Cariolle, D., Balbi, J.H.: Coupled atmosphere-wildland fire modelling. J. Adv. Model. Earth Syst. **1**, 1–9 (2009)
6. Finney, M.: Fire growth using minimum travel time methods. Can. J. Forest Res. **32**, 1420–1424 (2002)
7. Finney, M.: Calculation of fire spread rates across random landscapes. Int. J. Wildland Fire. **12**, 167–174 (2003)
8. Klimontovich, Yu.L.: Nonlinear Brownian motion. Phys.-Usp. **37**, 737–767 (1994)
9. Mallet, V., Keyes, D.E., Fendell, F.E.: Modeling wildland fire propagation with level set methods. Comput. Math. Appl. **57**, 1089–1101 (2009)
10. Mandel, J., Bennethum, L.S., Beezley, J.D., Coen, J.L., Douglas, C.C., Kim, M., Vodacek, A.: A wildland fire model with data assimilation. Math. Comput. Simul. **79**, 584–606 (2008)
11. Mandel, J., Beezley, J.D., Coen, J.L., Kim, M.: Data assimilation for wildland fires: ensemble Kalman filters in coupled atmosphere-surface models. IEEE Contr. Syst. Mag. **29**, 47–65 (2009)
12. Mandel, J.; Beezley, J.D., Kochanski, A.K.: Coupled atmosphere-wildland fire modeling with WRF 3.3 and SFIRE 2011. Geosci. Model. Dev. **4**, 591–610 (2011)
13. Pagnini, G., Bonomi, B.: Lagrangian formulation of turbulent premixed combustion. Phys. Rev. Lett. **107**, 044–503 (2011)
14. Pagnini, G., Massidda, L.: The randomized level-set method to model turbulence effects in wildland fire propagation. In: Spano, D., Bacciu, V., Salis, M., Sirca, C. (eds.) Modelling Fire Behaviour and Risk, Conference on Fire Behaviour and Risk, Alghero, October 2011. Proceedings of ICFBR 2011, pp. 126–131 (2012)
15. Pagnini, G., Mentrelli, A.: Modelling wildland fire propagation by tracking random fronts. Nat. Hazards Earth Syst. Sci. **14**, 2249–2263 (2014)
16. Perryman, H.A., Dugaw, C.J., Varner, J.M., Johnson, D.L.: A cellular automata model to link surface fires to firebrand lift-off and dispersal. Int. J. Wildland Fire. **22**, 428–439 (2013)
17. Rothermel, R.C.: A mathematical model for predicting fire spread in wildland fires. Technical Report INT-115, USDA Forest Service (1972)
18. Sardoy, N., Consalvi, J.L., Kaiss, A., Fernandez-Pello, A.C., Porterie, B.: Numerical study of ground-level distribution of firebrands generated by line fires. Combust. Flame **154**, 478–488 (2008)
19. Sethian, J.A., Smereka, P.: Level set methods for fluid interfaces. Annu. Rev. Fluid Mech. **35**, 341–372 (2003)

Part IV
Numerical Analysis

Solving the Pertubed Quantum Harmonic Oscillator in Imaginary Time Using Splitting Methods with Complex Coefficients

Philipp Bader and Sergio Blanes

Abstract Efficient splitting algorithms for the Schrödinger eigenvalue problem with perturbed harmonic oscillator potentials in higher dimensions are considered. The separability of the Hamiltonian makes the problem suitable for the application of splitting methods. Using algebraic techniques, we show how to apply Fourier spectral methods to propagate higher dimensional quantum harmonic oscillators, thus retaining the near integrable structure and fast computability. This methods is then used to solve the eigenvalue problem by imaginary time propagation. High order fractional time steps of order greater than two necessarily have negative steps and can not be used for this class of diffusive problems. However, the use of fractional complex time steps with positive real parts does not negatively impact on stability and only moderately increases the computational cost. We analyze the performance of this class of schemes and propose new highly optimized sixth-order schemes for near integrable systems which outperform the existing ones in most cases.

1 Introduction

We consider the eigenvalue problem for the stationary Schrödinger equation (SE) ($\hbar = m = 1$),

$$H\phi_i(x) = E_i\phi_i(x), \qquad i = 0, 1, 2, \ldots \tag{1}$$

where

$$H = T + V(x) = \frac{1}{2}p^T \Lambda p + V(x), \quad x \in \mathbb{R}^d, \tag{2}$$

P. Bader (✉) • S. Blanes
Instituto de Matemática Multidisciplinar, Universitat Politècnica de València,
46022 Valencia, Spain
e-mail: phiba@imm.upv.es; serblaza@imm.upv.es

F. Casas, V. Martínez (eds.), *Advances in Differential Equations and Applications*,
SEMA SIMAI Springer Series 4, DOI 10.1007/978-3-319-06953-1_21

$V(x)$ denotes the interaction potential and $p = -i\nabla$ is the momentum operator. In this work, we will focus on special techniques for the important case of the perturbed harmonic oscillator in higher dimensions,

$$V(x) = V_0 + \varepsilon V_\varepsilon(x) = \frac{1}{2} x^T \Omega x + \varepsilon V_\varepsilon(x), \quad \varepsilon \ll 1,$$

which is relevant, for example, if one is interested in the lower excited states, which evolve near the minimum of the potential. This particular problem has attracted great interest among theorists and practitioners [2, 12, 16, 17, 19] due to its relevance for the understanding of the atomic and molecular structure of matter.

An efficient solution is given by the imaginary time propagation method (ITP), i.e., the evolution of the time-dependent Schrödinger equation in imaginary time ($t = -i\tau$), whose formal evolution operator is $\exp(-\tau H)$. In practice, any initial condition converges under the action of $\exp(-\tau H)$ asymptotically to the ground state solution when $\tau \to \infty$.

We will show how to compute the operators $\mathrm{e}^{-\tau(T+V_0)}$ and $\mathrm{e}^{-\tau\varepsilon V_\varepsilon}$ exactly in the coordinate and momentum space, respectively, in order to apply the operator splitting technique which combines these exponential operators with appropriate coefficients to yield an approximation of $\mathrm{e}^{-\tau H}$. The computational cost depends on the number of changes between these coordinates which are cheaply performed by Fast Fourier transforms (FFT). However, splitting methods of order $p > 2$ require negative time-steps [20, 21] and the instabilities caused thereof are analogous to the ones for the integration of a diffusion equation backwards in time.

In this paper, we propose new splitting methods that take into account the near integrable structure of the problem by solving the harmonic part exactly by means of Fourier transforms and overcome the order barrier by using complex time-steps. The obtained methods outperform the existing splitting schemes when high accuracy is desired and could be appropriate for elaborating a variable order algorithm. The conducted numerical experiments illustrate the efficiency of the new methods.

2 Imaginary Time Integration for the Schrödinger Equation

An important property of the Hermitian operator H is that (choosing properly the origin of the potential) its eigenvalues $0 \leq E_0 \leq E_1 \leq \ldots$ are real and non-negative, and the corresponding eigenfunctions ϕ_i can be chosen to form a real orthonormal basis on its domain. The imaginary time Schrödinger equation reads

$$-\frac{\partial}{\partial\tau}\psi(x,\tau) = H\psi(x,\tau), \qquad \psi(x,0) = \psi_0(x), \tag{3}$$

with formal solution $\psi(x,\tau) = e^{-\tau H}\psi(x,0)$. After expanding the initial condition ψ_0 in the basis of eigenfunctions ϕ_i, the time evolution of (3) is given by

$$\psi(x,\tau) = e^{-\tau H}\sum_i \langle\phi_i(x)\,|\,\psi(x,0)\rangle\,\phi_i(x) = \sum_i e^{-\tau E_i}\ \langle\phi_i(x)\,|\,\psi(x,0)\rangle\ \ \phi_i(x), \tag{4}$$

where $\langle\cdot\,|\,\cdot\rangle$ is the usual L^2 scalar product. Asymptotically, for a sufficiently long time integration, we get $\psi(x,\tau) \to e^{-\tau E_0}\,c_0\phi_0$ since the other exponentials decay more rapidly. The convergence rate depends of course on the separation of the eigenvalues. For simplicity, we restrict ourselves to the non-degenerate case $E_0 < E_1$.

Normalization of the asymptotic value yields the eigenfunction ϕ_0 and the corresponding eigenvalue is computed via $E_0 = \langle\phi_0|H\phi_0\rangle$. Excited states can be obtained by propagating different wave functions simultaneously (or successively) in time and using, for example, the Gram-Schmidt orthonormalization or diagonalizing the overlap matrix [1].

The problem is further simplified by truncating the spatial domain and assuming periodic boundary conditions, which is justified since all bounded eigenstates vanish at a sufficiently large distance from the origin.

The potential V is represented in this grid by a diagonal matrix and the periodicity of the system allows for the use of spectral methods (in space) for the calculation of T, namely the Fast Fourier Transform, after which the matrix representation of T also becomes diagonal on each dimension, and the computational costs for the application of V and T to a vector are thus proportional to N^d and $N^d \log N$, respectively.

3 Splitting Methods for the Schrödinger Equation

To approximate the time evolution (4), i.e., the computation of $\mathrm{e}^{-\tau H}$ acting on a vector, standard splittings consist of compositions of the operators $\mathrm{e}^{-\tau V}$ and $\mathrm{e}^{-\tau T}$ evaluated at different times. A first example is provided by the well-known Strang splitting

$$\Psi_h^{[2]} \equiv \mathrm{e}^{-\frac{h}{2}V}\,\mathrm{e}^{-hT}\,\mathrm{e}^{-\frac{h}{2}V}, \tag{5}$$

verifying $\Psi_h^{[2]} = \mathrm{e}^{-hH} + \mathscr{O}(h^3)$ with $h \equiv \Delta\tau$. Higher order approximations can be obtained by a more general composition

$$\Psi_h^{[p]} \equiv \prod_{i=1}^{m} \mathrm{e}^{-a_i hT}\,\mathrm{e}^{-b_i hV}, \tag{6}$$

where $\Psi_h^{[p]} = \mathrm{e}^{-hH} + \mathcal{O}(h^{p+1})$ if the coefficients a_i, b_i are chosen such that they satisfy a number of order conditions (with m sufficiently large). It is well-known, however, that methods of order greater than two ($p > 2$) necessarily have negative coefficients [7,13,20,21]. While this is usually not a problem for the coefficients b_i, having negative a_i coefficients makes the algorithm badly conditioned (in the limit $N \to \infty$).

Composition methods with coefficients b_i positive are also convenient for the present case of unbounded potentials since negative values of b_i can generate large round-off errors in the exponential $\mathrm{e}^{-b_i V}$ at the boundaries if the interval-size of the spatial discretization is not appropriately chosen and the potential takes exceedingly large values.

Splitting methods are particularly appropriate for the numerical integration of this problem since the choice of the time step, h, is not affected by the mesh size.

Before we concentrate on order conditions to obtain higher order methods, we point out that it is beneficial to separate the quadratic part and to treat the remainder as a perturbation since the harmonic oscillator has a simple and fast solution using FFTs [3,11]. The central result of this work will give rise to such a splitting approach for higher dimensional problems,

Theorem 1 *For symmetric positive definite matrices $\Lambda, \Omega \in \mathbb{R}^{d\times d}$ and functions*

$$f(h,\Lambda,\Omega) = \sqrt{\Omega\Lambda}\tan\left(\frac{h}{2}\sqrt{\Omega\Lambda}\right)\Lambda^{-1}, \qquad g(h,\Lambda,\Omega) = \Omega^{-1}\sqrt{\Omega\Lambda}\sin\left(h\sqrt{\Omega\Lambda}\right), \tag{7}$$

the following decomposition is satisfied for $|h\lambda_{\max}(\sqrt{\Omega\Lambda})| < \pi$:

$$e^{-ih\frac{1}{2}(p^T\Lambda p + q^T\Omega q)} = e^{-i\frac{1}{2}q^T f(h,\Lambda,\Omega)q}\, e^{-i\frac{1}{2}p^T g(h,\Lambda,\Omega)p}\, e^{-i\frac{1}{2}q^T f(h,\Lambda,\Omega)q}, \tag{8}$$

where $q^T = (x_1, x_2, \ldots, x_d)$ and $p^T = -i(\partial_{x_1}, \partial_{x_2}, \ldots, \partial_{x_d})$ and the stepsize is restricted by the largest eigenvalue $\lambda_{\max}$ of $\sqrt{\Omega\Lambda}$.

Remark 1 From a formal point of view, the symmetry of Λ and Ω as well as the positivity are not necessary and could be replaced by invertibility. These conditions only play a role when we express the formal series in terms of trigonometric functions with the help of the (positive) matrix square root. For quantum mechanics, however, we require Hermitian operators, e.g., for $p^T\Lambda p$, $\Lambda^* = \Lambda$ is implied. Furthermore, since the corresponding operators commute, there is a degree of freedom in the representation of the matrices in the Hamiltonian and by choosing them real and Hermitian, Λ, Ω are uniquely determined.

Proof The proof is a generalization of the result for the one-dimensional case [3]. Given the functions f, g, we can prove the lemma directly by recalling that two operators are identical on a sufficiently small time interval if they satisfy the same

first order differential equation with the same initial conditions [22]. We thus verify that the right-hand side of (8) also solves the propagator equation

$$i\dot{U} = (A_1 + B_1)U, \qquad U(0) = I, \tag{9}$$

with $A_1 = \frac{1}{2}p^T \Lambda p$ and $B_1 = \frac{1}{2}q^T \Omega q$ and is therefore identical to the propagator on the left-hand side. Now set

$$\tilde{U}(t) = e^{-if(t)A_1} e^{-ig(t)B_1} e^{-if(t)A_1}$$

and plugging it into (9) yields

$$(A_1 + B_1)\tilde{U} \stackrel{!}{=} \Big(\dot{f} A_1 + e^{-ifA_1}\, \dot{g} B_1\, e^{ifA_1} + e^{-ifA_1}\, e^{-igB_1}\, \dot{f} A_1\, e^{igB_1}\, e^{ifA_1} \Big) \tilde{U}.$$

After algebraic manipulation, we obtain two independent non-linear differential equations for $f(t)$ and $g(t)$ with initial condition $f(0) = g(0) = 0$ in order to satisfy $\tilde{U}(0) = I$. It is then easy to check that f, g given in (7) solve these equations. As a result, we have that $\tilde{U}(t) = U(t)$ locally in a neighborhood of the origin and (8) is proved identically.

The functions f, g can be found by exploiting that the operators A_1, B_1 generate the same algebra as their corresponding classical mechanical operators and the computations can be carried out in a matrix setting as follows: The classical mechanical system is

$$\frac{d}{dt}\begin{pmatrix} q \\ p \end{pmatrix} = \begin{pmatrix} 0 & \Lambda \\ -\Omega & 0 \end{pmatrix} \begin{pmatrix} q \\ p \end{pmatrix}$$

and the evolution operator is computed to

$$\exp\left(t \begin{pmatrix} 0 & \Lambda \\ -\Omega & 0 \end{pmatrix} \right) = \begin{pmatrix} \cos(t\sqrt{\Lambda\Omega}) & \Lambda(\sqrt{\Omega\Lambda})^{-1} \sin(t\sqrt{\Omega\Lambda}) \\ -\Omega(\sqrt{\Lambda\Omega})^{-1} \sin(t\sqrt{\Lambda\Omega}) & \cos(t\sqrt{\Omega\Lambda}) \end{pmatrix}. \tag{10}$$

It is known that products of real psd matrices $\Lambda\Omega, \Omega\Lambda$ can be diagonalized and have positive eigenvalues [15, Corr. 7.6.2], which means that the positive square root in (10) exists and is unique. We further note that the product $\Lambda\Omega$ is symmetric iff. $[\Lambda, \Omega] = 0$. For the decomposition, we compose the exponentials

$$\begin{pmatrix} 1 & 0 \\ -B & 1 \end{pmatrix} \begin{pmatrix} 1 & A \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -B & 1 \end{pmatrix} = \begin{pmatrix} 1 - AB & A \\ -2B + BAB & 1 - BA \end{pmatrix}. \tag{11}$$

Equating (11) and (10), we easily obtain the expressions for $f,\ g$ in (7). □

With the identity (8) at hand, we propose to replace the generic splitting (6) by

$$\Psi_h^{[p]} \equiv \prod_{i=1}^{m} \mathrm{e}^{-a_i h H_0}\, \mathrm{e}^{-b_i h \varepsilon V_\varepsilon}, \tag{12}$$

which comes at the same computational cost of two FFTs per stage and has the important benefit of error terms proportional to powers of the small parameter ε. Applying the Baker-Campbell-Hausdorff formula to the composition (12), the order conditions can be obtained and for details, we refer to [4].

Since ε is assumed to be small, the error expansion for a consistent method Ψ_h can be asymptotically expressed as

$$\Psi_h - \mathrm{e}^{-hH} = \sum_{i\geq 1} \sum_{k\geq s_i} e_{i,k}\, \varepsilon^i h^{k+1}, \text{ as } (h,\varepsilon) \to (0,0),$$

where the s_i start from the first non-vanishing error coefficient $e_{s_i,k}$. We say that Ψ_h is of generalized order $(s_1, s_2, \ldots, s_m)$ (where $s_1 \geq s_2 \geq \cdots \geq s_m$) if the local error satisfies that

$$\Psi_h - \mathrm{e}^{-hH} = \mathscr{O}(\varepsilon h^{s_1+1} + \varepsilon^2 h^{s_2+1} + \cdots + \varepsilon^m h^{s_m+1}).$$

4 New Splitting Methods for the ITP Problem

Due to the aforementioned necessity of negative coefficients a_i, b_j for splitting methods of order higher than two, an alternative strategy has to be pursued to circumvent this order barrier. A successful remedy comes from considering complex coefficients in the composition (12). In other problems where the presence of negative real coefficients is unacceptable, the use of high-order splitting methods with complex coefficients having positive real part has shown to possess some advantages. In recent years a systematic search for new methods with complex coefficients has been carried out and the resulting schemes have been tested in different settings: Hamiltonian systems in celestial mechanics [10], the time-dependent Schrödinger equation in quantum mechanics [5, 6] and also in the more abstract setting of evolution equations with unbounded operators generating analytic semigroups [9,14]. Many of the existing splitting methods with complex coefficients have been constructed by applying the composition technique to the symmetric second-order leapfrog scheme (5). For example, a fourth-order integrator can be obtained with the symmetric composition

$$\Psi_h^{[4]} = \Psi_{\alpha h}^{[2]}\, \Psi_{\beta h}^{[2]}\, \Psi_{\alpha h}^{[2]}, \qquad \alpha = 1/(2 - 2^{1/3}\mathrm{e}^{2ik\pi/3}), \quad \beta = 2^{1/3}\mathrm{e}^{2ik\pi/3}\alpha \tag{13}$$

and $k = 1, 2$. In both cases, one has $\mathrm{Re}(\alpha), \mathrm{Re}(\beta) > 0$. Higher order composition methods with complex coefficients and positive real part can be found in Refs. [8,9, 14], where several numerical examples are also reported.

We present the best methods that have been obtained by a systematic search among symmetric compositions. Since H_0 and V_ε have qualitatively different properties, we analyze both TVT-and VTV-type compositions, defined as

$$\text{TVT: } \Psi_h^{[p]} = \mathrm{e}^{-a_1 h H_0}\mathrm{e}^{-b_1 h\varepsilon V_\varepsilon}\mathrm{e}^{-a_2 h H_0}\cdots\mathrm{e}^{-a_2 h H_0}\mathrm{e}^{-b_1 h\varepsilon V_\varepsilon}\mathrm{e}^{-a_1 h H_0},$$

$$\text{VTV: } \Psi_h^{[p]} = \mathrm{e}^{-b_1 h\varepsilon V_\varepsilon}\mathrm{e}^{-a_1 h H_0}\mathrm{e}^{-b_2 h\varepsilon V_\varepsilon}\cdots\mathrm{e}^{-b_2 h\varepsilon V_\varepsilon}\mathrm{e}^{-a_1 h H_0}\mathrm{e}^{-b_1 h\varepsilon V_\varepsilon}.$$

In principle, both compositions have the same computational cost for the same number of exponentials. Nevertheless, due to a projection step to the real part after each full time-step, only in the VTV composition we can concatenate the last map in the current step with the first stage in the next one. The TVT compositions thus require two additional FFTs in comparison with the VTV composition, and this is accounted for in the numerical experiments.

We have explored both TVT and VTV compositions of order six with different number of stages. Among the solutions that minimize $\sum_i (|a_i| + |b_i|)$ and/or the absolute value of the real part of the coefficients appearing at the leading error terms, we choose the ones that give the best performance on a series of numerical examples.

The best methods for our purpose have nine stages and the two free parameters are used to achieve generalized order (8,6) and are denoted by $\mathrm{T86}_9$ and $\mathrm{V86}_9$ in Table 1 (Murua and Makazaga, Private communication, 2012).

Table 1 Splitting methods of order (8,6)

$\mathrm{V86}_9$	$\mathrm{T86}_9$
$b_1 = 0.0324977060374 + 0.0106413103804i$	$a_1 = 0.0422578972998 - 0.0142157802241i$
$a_1 = 0.0878956804412 + 0.0360525761828i$	$b_1 = 0.0948948693677 - 0.0379638064725i$
$b_2 = 0.0941809234226 + 0.0238668753626i$	$a_2 = 0.0952603984718 + 0.0045187258914i$
$a_2 = 0.0953518553990 - 0.0651283760351i$	$b_2 = 0.0973746603817 + 0.0885188779317i$
$b_3 = 0.1011329530972 - 0.1122017573370i$	$a_3 = 0.0999605789447 + 0.0902719950713i$
$a_3 = 0.1218655755949 - 0.0549740024714i$	$b_3 = 0.1185847935200 + 0.0383562506084i$
$b_4 = 0.1609413821194 - 0.0161276438969i$	$a_4 = 0.1486955304026 + 0.0114381171876i$
$a_4 = 0.1415068827184 + 0.0246072290465i$	$b_4 = 0.1368651197603 - 0.0235874049695i$
$b_5 = 1/2 - (b_1 + b_2 + b_3 + b_4)$	$a_5 = 1/2 - (a_1 + a_2 + a_3 + a_4)$
$a_5 = 1 - 2(a_1 + a_2 + a_3 + a_4)$	$b_5 = 1 - 2(b_1 + b_2 + b_3 + b_4)$

5 Numerical Examples

As test bench for the numerical methods, we consider an anisotropically perturbed harmonic oscillator

$$H = \frac{1}{2}\left(\partial_x^2 + \partial_y^2\right) + \frac{1}{2}(x, y)\begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \frac{1}{100}x^4. \tag{14}$$

The numerical integration proceeds as follows: starting from random initial data, we iterate with fixed time-step until the sufficiently large final time $T = 50$ and compare the result with the exact solution, $u_{ex}(T)$, which has been obtained by integrating with a much smaller time step. The spatial interval is fixed for all experiments to $[-10, 10]^2$ and is discretized with $N = 128$ equidistant mesh points in each dimension. At each step, we project the obtained vector to its real part and normalize it to one in $\ell_2(\mathbb{R}^2)$, i.e., given the method $\Psi_h^{[p]}$ and initial conditions, $u_n \in \mathbb{R}^{N^2}$, we compute u_{n+1} as

$$\tilde{u}_{n+1} = \mathrm{Re}(\Psi_h^{[p]}\, u_n);$$

and then normalize the solution $u_{n+1} = \tilde{u}_{n+1} / \|\tilde{u}_{n+1}\|$, where the norm is given by

$$\|w\|^2 \equiv \Delta x^2 \sum_{j=1}^{N^2} w_j^2, \quad w = (w_1, \ldots, w_{N^2}) \in \mathbb{R}^{N^2}.$$

We take as the computational cost the number of Fourier transforms necessary until the final time. In addition, the methods using complex coefficients are penalized by a factor 2 in the computational cost, which comes from the use of complex Fourier transforms instead of real FFT. We repeat the numerical integrations for different values of the time step, i.e., $h = T/M$ for different values of M. We take as the approximate solution, $u_a(T)_M$, in each case and measure the error as error $= \|u_{ex}(T) - u_a(T)\|$. The reference methods will be the second order method optimized for near integrable systems, V82 [18] and the fourth order complex triple-jump scheme (13), referenced as Yoshida 4. All methods are computed with either a splitting in harmonic part plus perturbation, indicated by subscript $_H$, or by splitting into kinetic and potential part, subscript $_F$, neglecting the near-integrable structure. After the substitution $\delta = -ih$, the stability condition in theorem becomes $|\mathrm{Im}(h)\lambda_{max}| < \pi$ and $\mathrm{Re}(h) > 0$ and the perturbation part is easily propagated after discretization by the exponential of a diagonal matrix. In this setting, the higher order in the small parameter is amplified and the efficiency plots in Fig. 1 indicate that the new methods outperform the existing ones when high precision is sought.

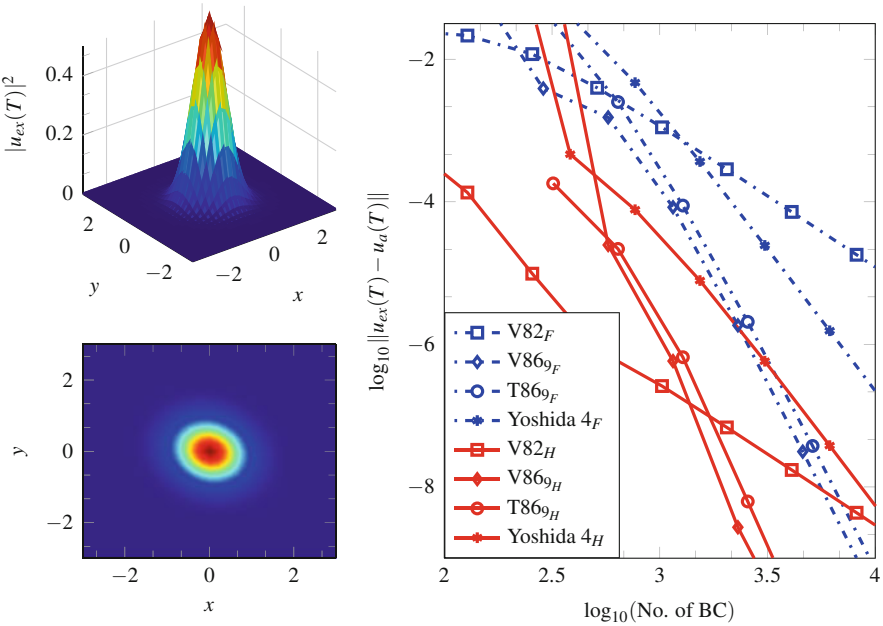


Fig. 1 In the left column, the squared absolute value of the solution at $T = 50$, in 3D (*top panel*) and from above (*bottom panel*), is displayed. The *right column* shows the efficiency curves (accuracy versus computational cost) for the 2D perturbed harmonic oscillator (14) integrated using $N_x = N_y = 128$ equidistant grid points on $[-10, 10]^2$. The standard splitting in kinetic and potential parts (*dashed blue lines*) is clearly dominated by the near-integrable splitting (*red solid lines*)

Conclusions

We have studied the quantum harmonic oscillator and derived efficient decompositions that allow the use of Fourier spectral methods and go in hand with perturbations of the potential. The Schrödinger eigenvalue problem in imaginary time often has such a structure and we have designed splitting schemes using complex coefficients that can overcome the order barrier for parabolic problems since the coefficients have only positive real parts. The obtained sixth order methods are clearly superior to any classical ones for high precisions.

An efficient implementation should take into account, for example, a preliminary time integration on a coarse mesh using simple precision arithmetic in order to get, as fast as possible, a smooth and relatively accurate solution from a random initial guess, and next consider a refined mesh using arithmetic in double precision. For simple precision arithmetic and low accuracies, it

(continued)

suffices to consider only low order methods, and when higher accuracies are desired we turn to double precision, variable time step and variable order methods. The best algorithm could depend on the class of problems to solve.

Acknowledgements We wish to acknowledge Ander Murua and Joseba Makazaga for providing the methods $T86_9$ and $V86_9$. This work has been partially supported by Ministerio de Ciencia e Innovación (Spain) under project MTM2010-18246-C03. P.B. also acknowledges the support through the FPU fellowship AP2009-1892.

References

1. Aichinger, M., Krotscheck, E.: A fast configuration space method for solving local Kohn–Sham equations. Comput. Mater. Sci. **34**, 183–212 (2005)
2. Auer, J., Krotscheck, E., Chin, S.A.: A fourth-order real-space algorithm for solving local Schrödinger equations. J. Chem. Phys. **115**, 6841–6846 (2001)
3. Bader, P., Blanes, S.: Fourier methods for the perturbed harmonic oscillator in linear and nonlinear Schrödinger equations. Phys. Rev. E. **83**, 046711 (2011)
4. Bader, P., Blanes, S., Casas, F.: Solving the Schrödinger eigenvalue problem by the imaginary time propagation technique using splitting methods with complex coefficients. J. Chem. Phys. **139**, 124117 (2013)
5. Bandrauk, A.D., Dehghanian, E., Lu, H.: Complex integration steps in decomposition of quantum evolution operators. Chem. Phys. Lett. **419**, 346–350 (2006)
6. Bandrauk, A.D., Lu, H.: Exponential propagators (integrators) for the time-dependent Schrödinger equation. J. Theor. Comput. Chem. **12**, 1340001 (2013)
7. Blanes, S., Casas, F.: On the necessity of negative coefficients for operator splitting schemes of order higher than two. Appl. Numer. Math. **54**, 23–37 (2005)
8. Blanes, S., Casas, F., Chartier, P., Murua, A.: Optimized high-order splitting methods for some classes of parabolic equations. Math. Comput. **82**, 1559–1576 (2013)
9. Castella, F., Chartier, P., Descombes, S., Vilmart, G.: Splitting methods with complex times for parabolic equations. BIT **49**, 486–508 (2009)
10. Chambers, J.E.: Symplectic integrators with complex time steps. Astron. J. **126**, 1119–1126 (2003)
11. Chin, S.A., Krotscheck, E.: Fourth-order algorithms for solving imaginary-time Gross-Pitaevskii equation in a rotating anisotropic trap. Phys. Rev. E. **72**, 036705 (2005)
12. Chin, S. A., Janecek, S., Krotscheck, E.: Any order imaginary time propagation method for solving the Schrödinger equation. Chem. Phys. Lett. **470**, 342–346 (2009)
13. Goldman, D., Kaper, T.J.: nth-order operator splitting schemes and nonreversible systems. SIAM J. Numer. Anal. **33**, 349–367 (1996)
14. Hansen, E., Ostermann, A.: High order splitting methods for analytic semigroups exist. BIT **49**, 527–542 (2009)
15. Horn, R.A., Johnson, C.R.: Matrix Analysis, 2nd edn. Cambridge University Press, Cambridge (2013)
16. Janecek S., Krotscheck, E.: A fast and simple program for solving local Schrödinger equations in two and three dimensions. Comput. Phys. Commun. **178**, 835–842 (2008)
17. Lehtovaara, L., Toivanen, J., Eloranta, J.: Solution of time-independent Schrödinger equation by the imaginary time propagation method. J. Comput. Phys. **221**, 148–157 (2007)

18. McLachlan, R.I.: Composition methods in the presence of small parameters. BIT **35**, 258–268 (1995)
19. Roy, A.K., Gupta, N., Deb, B.M.: Time-dependent quantum-mechanical calculation of ground and excited states of anharmonic and double-well oscillators. Phys. Rev. A. **65**, 012109 (2001)
20. Sheng, Q.: Solving linear partial differential equations by exponential splitting. IMA J. Numer. Anal. **9**, 199–212 (1989)
21. Suzuki, M.: General theory of fractal path integrals with applications to many-body theories and statistical physics. J. Math. Phys. **32**, 400–407 (1991)
22. Wilcox, R.M.: Exponential operators and parameter differentiation in quantum physics. J. Math. Phys. **8**, 962–982 (1967)

A High-Order Well-Balanced Central Scheme for the Shallow Water Equations in Channels with Irregular Geometry

Ángel Balaguer-Beser, María Teresa Capilla, Beatriz Nácher-Rodríguez, Francisco José Vallés-Morán, and Ignacio Andrés-Doménech

Abstract This paper presents a new numerical scheme based on the finite volume method to solve the shallow water equations in channels with rectangular section and variable width. Time integration is carried out by means of a Runge-Kutta scheme with a natural continuous extension, using a new temporary forward flow at the midpoint of each cell which considers the physical flow and the source term primitive of the shallow water model. That term takes into account the gradient of bed height, channel width and friction energy loss model. Spatial integration is based on a central scheme in which flows only have to be evaluated on the midpoint of the cells where the solution is reconstructed. In this way, it is not necessary to know the structure of the partial differential equations to be solved. A centered three degree reconstruction polynomial is applied, using a slope correction to the midpoint of each cell to prevent the occurrence of spurious numerical oscillations. Some benchmark examples show the non-oscillatory behavior of numerical solutions in channels with a variable width. A comparison between numerical results and those obtained experimentally on a laboratory flume is also carried out.

1 Introduction

This paper describes a new central numerical scheme that solves the shallow water equations considering a rectangular section and variable width. The system of

Á. Balaguer-Beser (✉) • M.T. Capilla Romá
Departamento de Matemática Aplicada, Universitat Politècnica de València, Cno. de Vera s/n, E-46022 Valencia, Spain
e-mail: abalague@mat.upv.es; tcapilla@mat.upv.es

B. Nácher-Rodríguez • F.J. Vallés-Morán • I. Andrés-Doménech
Instituto Universitario de Investigación de Ingeniería del Agua y Medio Ambiente (IIAMA), Universitat Politècnica de València, Cno. de Vera s/n, E-46022 Valencia, Spain
e-mail: beanacro@cam.upv.es; fvalmo@hma.upv.es; igando@hma.upv.es

F. Casas, V. Martínez (eds.), *Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 4, DOI 10.1007/978-3-319-06953-1_22

partial differential equations which solves this problem (continuity and momentum equations) can be expressed in this way (see [10]):

$$\begin{cases} h_t + (q)_x = -q\frac{B'(x)}{B(x)} \\ (q)_t + \left(\frac{q^2}{h} + \frac{1}{2}gh^2\right)_x = -gh(Z_b)_x - \frac{q^2}{h}\frac{B'(x)}{B(x)} - g\,n^2\,q\,\left|\frac{q}{h}\right| R_h^{-4/3} \end{cases} \tag{1}$$

where $h(x,t)$ is the height of the fluid above the bottom of the channel (water depth), $q(x,t)$ is the specific discharge (flow rate per unit width), which is related to the average horizontal velocity $v(x,t)$ by the expression $q(x,t) = h(x,t)v(x,t)$, $Z_b(x)$ is the function which describes the bed height, g is the acceleration of gravity ($g = 9.8\,\mathrm{m/s^2}$), $B(x)$ is the channel width at each point x, n is Manning's roughness coefficient and $R_h(x,t)$ represents the hydraulic radius which is expressed as:

$$R_h(x,t) = \frac{h(x,t)B(x)}{B(x) + 2h(x,t)}. \tag{2}$$

System of equations (1) can be rewritten using $\eta(x,t) = h(x,t) + Z_b(x,t)$ instead of $h(x,t)$, as suggested in [2]. This gives the following system of equations:

$$\begin{pmatrix} \eta \\ q \end{pmatrix}_t + \begin{pmatrix} q \\ \frac{q^2}{\eta - Z_b} + \frac{1}{2}g(\eta - Z_b)^2 \end{pmatrix}_x = \begin{pmatrix} -q\frac{B'(x)}{B(x)} \\ -g(\eta - Z_b)(Z_b)_x - \frac{q^2}{\eta - Z_b}\frac{B'(x)}{B(x)} - g\,S_f \end{pmatrix} \tag{3}$$

where S_f models the friction term by means of Manning's formula, so that:

$$S_f = n^2\,q\left|\frac{q}{\eta - Z_b}\right|(R_h)^{-4/3}. \tag{4}$$

Introducing the vector of variables, $u = (\eta, q)^T$, system (3) can be expressed as:

$$\frac{\partial u(x,t)}{\partial t} + \frac{\partial f(u(x,t))}{\partial x} = \sum_{k=1}^{3} s_k(x, u(x,t)) \Leftrightarrow u_t + f_x = s = s_1 + s_2 + s_3 \tag{5}$$

where $f(u)$ is the flux vector, s_1 is the source term related to the bed slope, s_2 depends on the channel width and s_3 is the friction term:

$$s_1 = \begin{pmatrix} 0 \\ -g(\eta - Z_b)(Z_b)_x \end{pmatrix},\ s_2 = \begin{pmatrix} -q\frac{B'(x)}{B(x)} \\ -\frac{q^2}{\eta - Z_b}\frac{B'(x)}{B(x)} \end{pmatrix},\ s_3 = \begin{pmatrix} 0 \\ -g\,S_f \end{pmatrix}. \tag{6}$$

2 A Central Numerical Scheme

The spatial domain $[0, L]$ is divided into N equally spaced nodes x_j. Furthermore, 3 additional nodes are considered on the left of $x = 0$, and also on the right of $x = L$ to avoid loss of accuracy in the domain boundaries. Thus, initially $\Delta x = \frac{L}{N}$ is defined and thereby so are points $x_{j+\frac{1}{2}} = -3\Delta x + j\Delta x,\ j = 0, 1, \ldots, (N + 6)$. Then the nodes $x_j = \frac{x_{j-\frac{1}{2}} + x_{j+\frac{1}{2}}}{2},\ j = 1, \ldots (N + 6)$ are considered. Time interval is discretized by means of: $t^0 = 0,\ t^n = t^{n-1} + \Delta t^n, n \geq 1$. Time step Δt^n is computed taking into account the CFL criteria, so for stability, the following condition has to be satisfied:

$$\Delta t^n = CFL \frac{\Delta x}{\max_j \left(\sqrt{g\hat{h}_j^n} + |\hat{v}_j^n| \right)},\quad CFL \leq 0.35 \tag{7}$$

where $\hat{h}_j^n$ and $\hat{v}_j^n$ are the point-values at time t^n of the water depth, h and the water velocity v, respectively, considering $x = x_j$ when n is even and $x = x_{j+\frac{1}{2}}$ when n is odd. The central scheme integrates equation (5) in the control volume: $[x_j, x_{j+1}] \times [t^n, t^{n+1}]$ when n is even so (see [4,7]):

$$\overline{u}_{j+\frac{1}{2}}^{n+1} = \overline{u}_{j+\frac{1}{2}}^{n} - \frac{1}{\Delta x} \left[\int_{t^n}^{t^{n+1}} f_{j+1}(\tau)\, d\tau - \int_{t^n}^{t^{n+1}} f_j(\tau)\, d\tau \right] + \int_{t^n}^{t^{n+1}} \overline{s}_{j+\frac{1}{2}}(\tau)\, d\tau \tag{8}$$

being:

$$\overline{u}_{j+\frac{1}{2}}^{n} = \frac{1}{\Delta x} \int_{x_j}^{x_{j+1}} u(x, t^n)\, dx, \qquad f_j(t) = f(u(x_j, t)),$$

$$\overline{s}_{j+\frac{1}{2}}(t) = \frac{1}{\Delta x} \int_{x_j}^{x_{j+1}} \sum_{k=1}^{3} s_k(x, u(x, t))\, dx. \tag{9}$$

The ideas described in [4,5] are extended for developing the numerical scheme of this paper, considering the source term, s, which also includes the channel width, s_2, and the friction term, s_3. Thus, numerical algorithm takes into account the following steps:

1. Reconstruction, at time t^n, of averages, $\overline{u}_{j+\frac{1}{2}}^n$ and point-values, $\hat{u}_j^n$. Those point-values will be used in the numerical approximation of source term and fluxes.
2. Reconstruction of Runge-Kutta fluxes, $k_j^{(i)}$. Time integration will be developed by means of a fourth-order Runge-Kutta scheme coupled with a Natural Continuous Extension (NCE) [4]. That procedure uses the Runge-Kutta fluxes, $k_j^{(i)},\ 1 \leq i \leq 4$, which coincide with a numerical evaluation of $(-f_x + s)$

in Eq. (5), computed starting from the point values $\hat{u}_j^n$. For that the following discrete values are defined:

$$K_j(x_k;\hat{u}^{(i)}) = -\left[f_k^{(i)} - f_j^{(i)}\right] + \left(\int_{x_j}^{x_k} (s_1 + s_2 + s_3)\, dx\right)^{(i)}.$$

Using a non-oscillatory interpolating polynomial it holds that:

$$\left.\frac{dK_j^{(i)}}{dx}\right|_{x=x_j} = k_j^{(i)} = (-f_x + s_1 + s_2 + s_3)_j^{(i)}, \quad \forall\, 1 \le i \le 4.$$

3. Evaluation of the point-value solution, $\hat{u}_j^{n+\beta_k}$, for time flux integrals. A Gaussian quadrature rule with two nodes of integration is selected in order to achieve fourth-order accuracy in time. Thus, the point-values of the solution: $\hat{u}_j^{n+\beta_k} = u(x_j, t^n + \beta_k \Delta t),\ k \in \{0, 1\}$ have to be computed.
4. Source term integration. Source term integrals are also evaluated using a Gauss quadrature rule with two integration nodes so that the resulting scheme is well balanced (see [3,5,6]).

Centered fourth-order non-oscillatory polynomials described in [4] are used in the reconstruction process of the first two steps.

3 Numerical Results

Test 3.1: Steady flows in channels with contractions Simulations reproduce a series of steady flows in a domain with variations in both width and topography. This example was represented for example in [1]. A frictionless channel of length $L = 3$ m is considered, which presents a converging and diverging geometry and a topographic small hill. Bed elevation is defined as:

$$Z_b(x) = \begin{cases} 0.1\cos^2[\pi(x-1.5)] & \text{if } |x-1.5| < 0.5 \\ 0 & \text{otherwise} \end{cases}$$

and the channel width is given by:

$$B(x) = \begin{cases} 1 - 0.1\cos^2[\pi(x-1.5)] & \text{if } |x-1.5| < 0.5 \\ 1 & \text{otherwise} \end{cases}.$$

Simulations involving subcritical and supercritical flow, as well as a combination of both, were performed until the steady-state solution was reached. During the simulations, the domain is discretized by a uniform grid with 100 cells, and CFL = 0.35. The numerical results are compared against the analytical solution

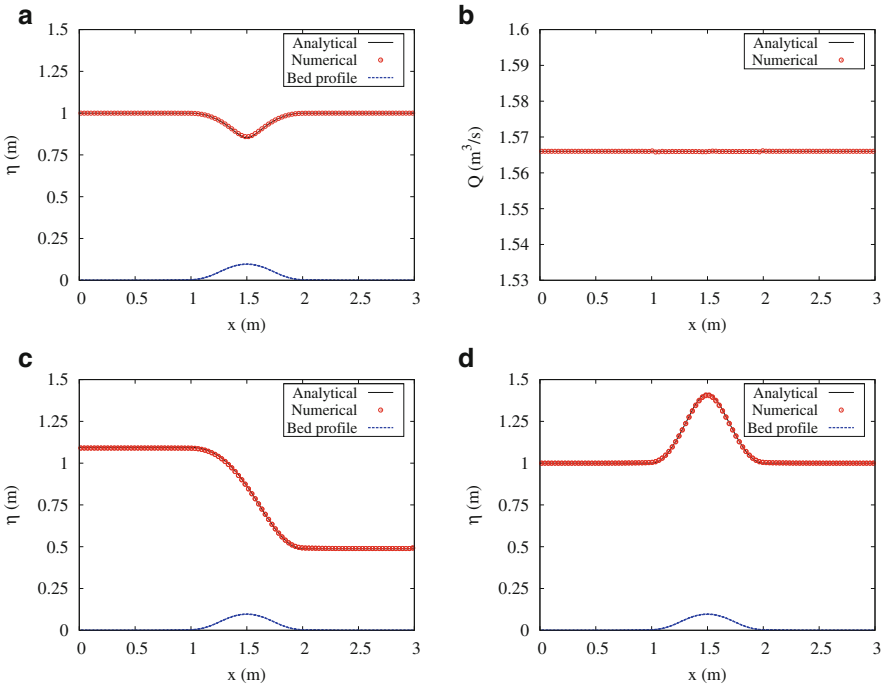


Fig. 1 Test 3.1: Solution to subcritical, supercritical and flow regime combination cases. η is the free surface level and $Q = B(x)q(x)$ is the total discharge. (**a**) η in subcritical flow problem; (**b**) Q in subcritical flow problem; (**c**) η in flow regime combination problem; (**d**) η in supercritical flow problem

obtained by assuming that there are two quantities that remain constant throughout the channel: total discharge $Q(x) = B(x)q(x)$ and total energy $E = h(x) + Z_b(x) + q(x)^2/(2gh(x))$, which is valid in absence of a hydraulic jump and friction losses [8].

In the subcritical flow problem (Fig. 1a, b), a discharge of $q = 1.566\,\mathrm{m}^2/\mathrm{s}$ is imposed at the inflow and a depth of $h = 1$ m is fixed at the outflow. In the case with a change on flow regime (Fig. 1c) an inflow discharge condition of $q = 1.879\,\mathrm{m}^2/\mathrm{s}$ is imposed. Then the flow becomes critical at the channel throat and continues onto a supercritical flow. Finally, Fig. 1d shows the supercritical flow that is produced by setting the inflow depth to $h = 1$ m and the inflow discharge to $q = 5.325\,\mathrm{m}^2/\mathrm{s}$. For all these cases, numerical results for the water surface shows good agreement with the exact solutions, with an acceptable level of accuracy for the steady discharge conditions. Total discharge Q also shows a non-oscillatory behavior in the last two cases, being similar to the subcritical case and for this reason it has not been shown in this paper.

Test 3.2: Tidal wave in a short channel with variable depth and width Benchmark test proposed in [10] is considered which describes the propagation of

a tidal wave into a channel with $L = 1{,}500$ m. This is an important benchmark for checking well-balancing of the numerical schemes (see [9]). Initial conditions are $\eta(x,0) = 12$ m and $q(x,0) = 0\,\mathrm{m}^2/\mathrm{s}$ and boundary conditions are:

$$\eta(x,t) = 12{+}4{+}4\sin\left(\pi\left(\frac{4t}{86{,}400} - \frac{1}{2}\right)\right)\ \mathrm{m}\ \forall x \le 0 \text{ and } q(x,t) = 0\ \mathrm{m}^2/s\ \forall x \ge L$$

The width functions and the bed profile are represented in Fig. 2a, b. Large variations in width and depth take place throughout the channel. The friction term is included by setting Manning's coefficient to $n = 0.1$. Numerical and asymptotic (see [10]) solutions are computed in a mesh with $N = 300$ nodes. Time step is defined using (7) and considering that CFL = 0.35.

Free-surface elevation and discharge per unit-width at times $t = 10{,}800$ s and $t = 32{,}400$ s are presented in Fig. 2b–d. Asymptotic solution for the free-surface elevation is horizontal and varies in time according to $\eta(0,t)$. Free-surface elevation at time $t = 32{,}400$ s is identical to that presented in Fig. 2b because $\eta(0, 10{,}800) = \eta(0, 32{,}400)$. The agreement between asymptotic and computed solution is always remarkable, both for free surface and discharge, without spurious numerical oscillations.

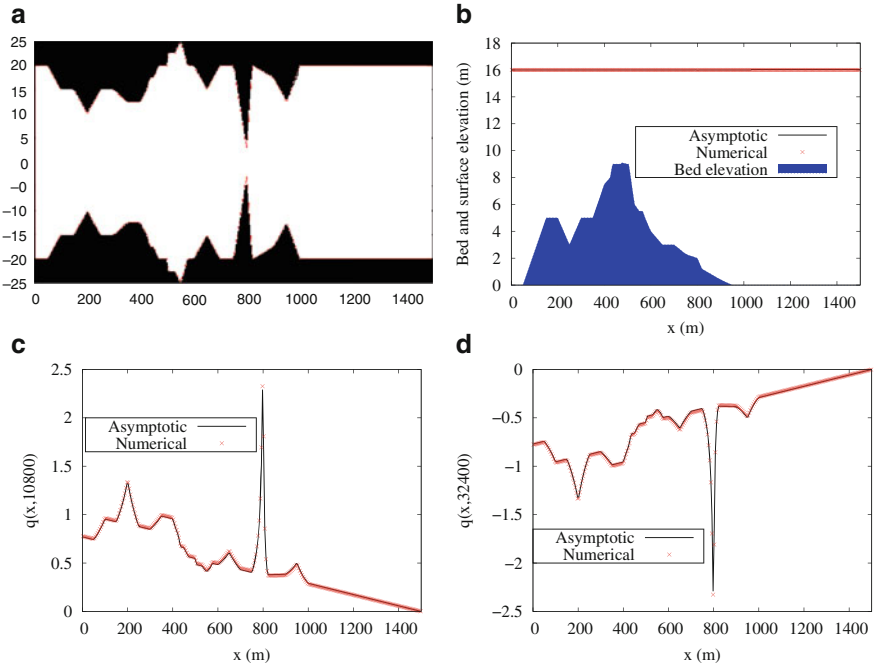


Fig. 2 Test 3.2: Geometry, numerical results and asymptotic solution for tide in a short channel. (**a**) Channel width (**b**) Bed and free-surface elevation; (**c**) Specific discharge at $t = 10{,}800$ s; (**d**) Specific discharge at $t = 32{,}400$ s

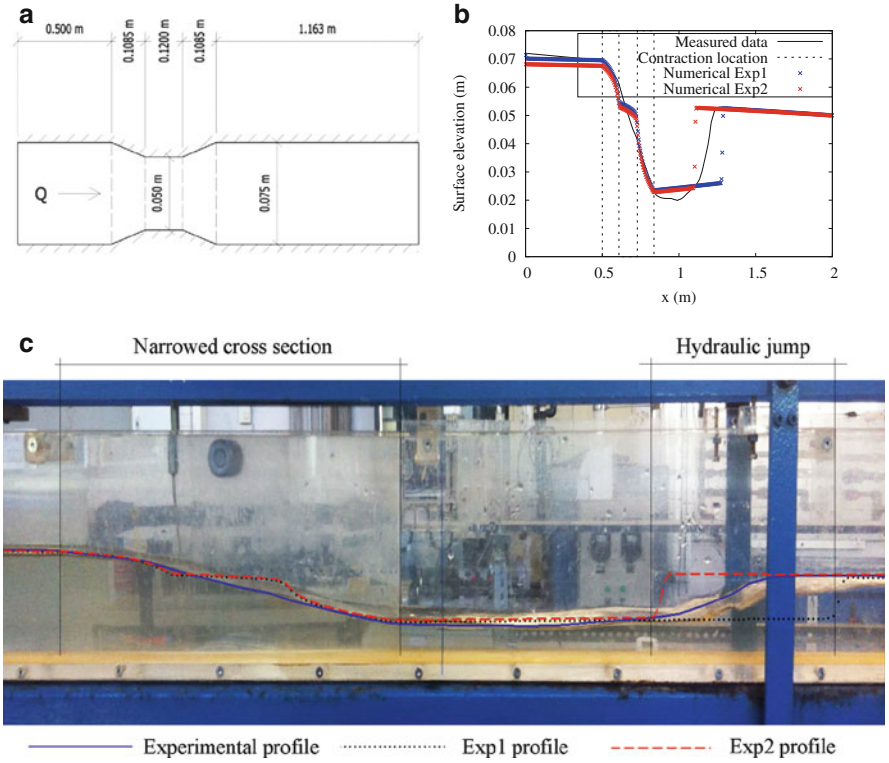


Fig. 3 Test 3.3: Numerical profiles plotted against experimental profile. Laboratory test is shown in the background. (**a**) Sketch of the laboratory experiment; (**b**) Free-surface elevation at $t = 60$ s; (**c**) Laboratory image together with numerical profiles

Test 3.3: Comparison with experimental results This test is considered in order to evaluate the capacity of the model to represent changes in open channel flow regimes, from subcritical to supercritical, and viceversa, including a hydraulic jump. A comparison is performed between numerical and experimental results. A channel of $L = 2.5$ m length is considered, with a horizontal bed ($Z_b(x) = 0, \ \forall x$). Figure 3a shows a plan view sketch of the laboratory flume, showing the variation in the width of the channel. Manning's coefficient is equal to 0.008, which is an appropriate value for methacrylate flumes, according to technical literature. Initial condition is set to: $h(x,0) = 0.05\,\text{m}$ and $Q(x,0) = 0\,\text{m}^3/\text{s}, \ \forall x \in]0, 2.5[$. Boundary conditions are the following:

$$Q(x,t) = 1.638 \cdot 10^{-3}\ \text{m}^3/\text{s}, \forall x \leq 0, \ \forall t, \ h(x,t) = 0.05\ \text{m}, \forall x \geq 2.5, \ \forall t \tag{10}$$

Numerical solutions are computed on a mesh with $N = 400$ points and at $t = 60$ s with a time step computed with CFL $= 0.35$. For the numerical solution,

besides the defined hydrograph and the downstream depth condition, two upstream conditions were tested. In the first case (named Exp1), water depth was set to 0.072 m, which is the value measured in the test. In the second case (named Exp2), water surface derivative at $x = 0$ is set equal to zero. Thus, the influence of knowing the upstream water depth value for subcritical flow can be assessed.

Figure 3c shows different water surface profiles plotted over a photograph of the laboratory test. Numerical and experimental profiles are also shown in Fig. 3b. As those figures illustrate, both cases show quite similar profiles, being the hydraulic jump location the only significant difference. Where the cross section is narrowed, the change from subcritical to supercritical regime is properly represented, and upstream and downstream numerically calculated depths are close to those observed during the test. With regard to depths upstream and downstream the hydraulic jump (conjugated depths), Exp2 provides the values obtained in the test, whereas Exp1 values differ in less than 5 %. When it comes to the water depth in the uppermost section of Exp2, the calculated value is 0.068 m, just 5.5 % lower than the measured one. The length of the hydraulic jump is not properly calculated, as might be expected from a one-dimensional model, which does not take into account the effect of turbulence. Nevertheless, it is possible to implement some equations that modify the solution at the hydraulic jump location, provided it occurs, by defining a certain length according to empirical relations widely used in the field of hydraulics.

Conclusions

The numerical scheme presented in this paper, which works with variable channel widths and friction energy losses, has shown its potential to represent different flow regimes in cases with strong irregularities in bed channel geometry and cross section, providing a remarkable agreement with well known benchmark solutions. It has also been proved to give reliable water surfaces when compared to experimental results, even when major discontinuities appear, as is the case of the hydraulic jump.

Acknowledgements This work was supported by the "Programa de Apoyo a la Investigación y Desarrollo" (PAID-05-12) of the Universitat Politècnica de València.

References

1. Alias, N.A., Liang, Q., Kesserwani, G.: A Godunov-type scheme for modelling 1D channel flow with varying width and topography. Comput. Fluids. **46**, 88–93 (2011)
2. Caleffi, V., Valiani, A., Bernini, A.: High-order balanced CWENO scheme for movable bed shallow water equations. Adv. Water Resour. **30**, 730–741 (2007)
3. Canestrelli, A., Siviglia, A., Dumbser, M., Toro, E.F.: Well-balanced high-order centered schemes for non-conservative hyperbolic systems. Applications to shallow water equations with fixed and mobile bed. Adv. Water Resour. **32**(6), 834–844 (2009)

4. Capilla, M.T., Balaguer-Beser, A.: A well-balanced high-resolution shape-preserving central scheme to solve one-dimensional sediment transport equations. Adv. Eng. Softw. **50**, 19–28 (2012)
5. Capilla, M.T., Balaguer-Beser, A.: A new well-balanced non-oscillatory central scheme for the shallow water equations on rectangular meshes. J. Comput. Appl. Math. **252**, 62–74 (2013)
6. Castro, M.J., Pardo, A., Parés, C., Toro, E.F.: On some fast well-balanced first order solvers for nonconservative systems. Math. Comput. **79**, 1427–1472 (2010)
7. Castro, M.J., Parés, C., Puppo, G., Russo, G.: Central schemes for nonconservative hyperbolic systems. SIAM J. Sci. Comput. **34**, B523–B558 (2012)
8. Hubbard, M.E.: On the accuracy of one-dimensional models of steady converging-diverging open channel flows. Int. J. Numer. Methods Fluids. **35**, 785–808 (2001)
9. Rosatti, G., Bonaventura, L., Deponti, A., Garegnani, G.: An accurate and efficient semi-implicit method for section-averaged free-surface flow modelling. Int. J. Numer. Methods Fluids. **65**, 448–473 (2011)
10. Vázquez-Cendón, M.E.: Improved treatment of source terms in upwind schemes for the shallow water equations in channels with irregular geometry. J. Comput. Phys. **148**, 497–526 (1999)

On Tridiagonal Sign Regular Matrices and Generalizations

Álvaro Barreras and Juan Manuel Peña

Abstract In this paper, some characterizations of strictly k–banded nonsingular sign regular matrices are presented and new sufficient conditions for the total positivity of tridiagonal matrices are provided.

1 Introduction

Totally positive, TP, matrices are matrices with all their minors nonnegative and have important applications to many fields (see [8]). Given the bidiagonal decomposition of a nonsingular totally positive matrix, many computations can be performed with high relative accuracy, including the calculation of their singular values, eigenvalues or their inverses (see [6]).

Totally positive matrices belong to the more general class of sign regular matrices. In [2], nonsingular tridiagonal sign regular matrices were characterized and a stable test to check if a matrix belongs to this class was presented. By Theorem 4 of this paper we prove that, if A is a tridiagonal nonsingular sign regular matrix, then either A or $-A$ is TP or A is a strictly tridiagonal sign regular matrix. Thus, in Sect. 3 we analyze the extension of the last previous concept: strictly banded nonsingular sign regular matrices. These matrices are characterized in several ways (see Theorems 1 and 2). Section 4 provides new sufficient conditions for the total positivity of a tridiagonal matrix, improving previous results. In the following Sect. 2, we include some basic notations and results.

Á. Barreras (✉) • J.M. Peña
Department of Applied Mathematics/IUMA, Universidad de Zaragoza, Pedro Cerbuna12, E-50009 Zaragoza, Spain
e-mail: albarrer@unizar.es; jmpena@unizar.es

F. Casas, V. Martínez (eds.), *Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 4, DOI 10.1007/978-3-319-06953-1_23

2 Basic Concepts and Results

Let us introduce some basic notations. Given $k, l \in \{1, \ldots, n\}$, let α (resp., β) be any increasing sequence of k (resp., l) positive integers less than or equal to n. Their set is denoted by $Q_{k,n}$. Let A be a real $n \times n$ matrix. Then we denote by $A[\alpha \mid \beta]$ the $k \times l$ submatrix of A containing rows numbered by α and columns numbered by β. Furthermore, if $\alpha = \beta$, we denote $A[\alpha] := A[\alpha \mid \alpha]$. Let us recall that the minors of the form $\det A[1, \ldots, k]$ ($k \leq n$) are called *leading principal minors* of A.

A vector of signs $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ with $\varepsilon_j \in \{\pm 1\}$ for $j \leq n$ is called a *signature*. An $n \times n$ matrix A is *sign regular* of order k (SR_k) with signature ε if, for each $j = 1, \ldots, k$, all minors of order j have the same sign ε_j or are zero. If A is sign regular of order k for $k = 1, \ldots, n$, then we say that A is *sign regular* (SR). If all minors of A of order less than or equal to k are nonnegative, then we say that A is *totally positive* of order k (TP_k). If A is totally positive of order k for $k = 1, \ldots, n$, then A is called *totally positive* (TP). Observe that totally positive matrices form a subclass of sign regular matrices.

As the following result shows (see Theorem 3.1 of [1]) the product of SR matrices is a SR matrix.

Proposition 1 *Let A and B be two $n \times n$ SR matrices with signature $\varepsilon^{(A)}$ and $\varepsilon^{(B)}$ respectively. Then the product AB is a SR matrix with signature $\varepsilon^{(A)}\varepsilon^{(B)}$.*

Let us consider the $n \times n$ matrix

$$P = \begin{pmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{pmatrix}, \tag{1}$$

which is the reverse of the identity matrix. Let us observe that P is a SR matrix with signature $\varepsilon = (\varepsilon_i)_{1 \leq i \leq n}$ given by $\varepsilon_i = (-1)^{\lfloor \frac{i}{2} \rfloor}$ for $i = 1, \ldots, n$, where $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x, a positive real number.

The following result corresponds to the well-known Shadow's Lemma for TP matrices extended to $n \times n$ SR matrices with signature $\varepsilon = (\varepsilon_i)_{1 \leq i \leq n}$ satisfying $\varepsilon_2 = +1$ (see Lemma 2.2 of [7]).

Lemma 1 *Let $A = (a_{ij})_{1 \leq i,j \leq n}$ be a SR matrix with signature $\varepsilon = (\varepsilon_i)_{1 \leq i \leq n}$ with $\varepsilon_2 = +1$. If $a_{ij} = 0$ for any $i, j \in \{1, \ldots, n\}$ then one of the following conditions holds:*

(1) $a_{kj} = 0$ for all $k \leq n$,
(2) $a_{ik} = 0$ for all $k \leq n$,
(3) $a_{kl} = 0$ for all $k \geq i$ and $l \leq j$,
(4) $a_{kl} = 0$ for all $k \leq i$ and $l \geq j$.

The following result forms part of the Theorem 3 of [3] and presents an interesting property of SR matrices with a negative entry in their signature. But

first, let us present some cases $(a), (b), (c)$ that are not included in the following proposition and arise when $-A$, PA or $-PA$ is a totally positive matrix respectively. Given the signature $\varepsilon = (\varepsilon_i)_{1\leq i\leq n}$ of an $n \times n$ sign regular matrix, consider the following cases:

(a) $\varepsilon_i = (-1)^i$ for all $i \leq n$.
(b) $\varepsilon_i = (-1)^{\lfloor \frac{i}{2} \rfloor}$ for all $i \leq n$.
(c) $\varepsilon_i = (-1)^{\lfloor \frac{i}{2}+i \rfloor}$ for all $i \leq n$.

Proposition 2 *Let $A = (a_{ij})_{1\leq i,j\leq n}$ be a nonsingular SR matrix with signature $\varepsilon = (\varepsilon_i)_{1\leq i\leq n}$ such that (a)–(c) do not hold. If there exists a positive integer k with $2 < k \leq n$ such that $\varepsilon_i = +1$ for all $i < k$ and $\varepsilon_k = -1$, then $a_{ij} \neq 0$ whenever $|i - j| \leq n - k + 1$.*

3 Strictly k–Banded Nonsingular SR Matrices

Let us introduce the following classes of matrices: k–banded matrices and strictly k–banded matrices. k–banded SR matrices have been considered in [5].

Definition 1 Given a matrix $A = (a_{ij})_{1\leq i,j\leq n}$ let us consider an integer $k < n$. We say that A is a *k–banded* matrix if $a_{ij} = 0$ when $|i - j| > k$. If, in addition $a_{ij} \neq 0$ when $|i - j| = k$, we say that A is an *strictly k–banded* matrix.

Observe that (strictly) 0–banded matrices are (strictly) diagonal matrices, (strictly) 1–banded matrices are (strictly) *tridiagonal* or Jacobi matrices and (strictly) 2–banded matrices are also called (strictly) *pentadiagonal* matrices.

The reverse matrices of k–banded matrices are also studied in this paper and so we present the following definition.

Definition 2 Given a matrix $A = (a_{ij})_{1\leq i,j\leq n}$ let us consider an integer $k < n$. We say that A is an *anti–k–banded* matrix if $a_{ij} = 0$ when $i + j < n - k + 1$ and when $i + j > n + k + 1$. If, in addition $a_{ij} \neq 0$ when $i + j = n - k + 1$ and when $i + j = n + k + 1$, we say that A is *strictly anti–k–banded*.

Observe that, if A is a (strictly) k–banded matrix, then AP and PA are (strictly) anti–k–banded matrices.

The following result characterizes SR strictly k–banded matrices in terms of their nonzero entries.

Theorem 1 *Let $A = (a_{ij})_{1\leq i,j\leq n}$ be a nonsingular SR matrix and let k be an integer such that $k < n$. Then the following conditions are equivalent:*

(i) A is strictly k–banded.
(ii) $a_{ij} \neq 0$ if and only if $|i - j| \leq k$.

Proof $(i) \Rightarrow (ii)$. Since A is strictly k–banded, we can take an index $i \in \{k + 1, \ldots, n - 1\}$ such that $a_{i,i-k}$ and $a_{i+1,i-k+1}$ are nonzero and $a_{i+1,i-k} = 0$. Thus,

the minor $\det A[i, i+1 \mid i-k, i-k+1] = a_{i,i-k}a_{i+1,i-k+1}$ is strictly positive and since A is SR by hypothesis, we have that $\varepsilon_2 = 1$.

Let us suppose that $a_{ij} = 0$ for some i, j such that $|i-j| \leq k$. Then we can apply Lemma 1 and we have that one of the conditions (1)–(4) holds. Observe that, since A is nonsingular, there are no columns or rows of zeros in A, and so neither (1) nor (2) holds. If (3) holds, then $a_{hl} = 0$ for all $h \geq i$ and $l \leq j$, in particular $a_{i,i-k} = 0$, but this contradicts the fact that A is strictly k–banded. Analogously, it can be checked that (4) does not hold. Then we conclude that $a_{ij} \neq 0$ for all i, j satisfying $|i-j| \leq k$.

$(ii) \Rightarrow (i)$. It is trivial by Definition 1. □

The following result relates total positivity with strictly k–banded matrices.

Theorem 2 *Let A be an $n \times n$ k–banded nonsingular SR nonnegative matrix with signature $\varepsilon = (\varepsilon_i)_{1 \leq i \leq n}$ and $\varepsilon_{n-k+1} = -1$. Then A is strictly k–banded if and only if A is TP_{n-k}.*

Proof To prove the direct implication it suffices to consider the minors $\det A[k+1, \ldots, l \mid 1, \ldots, l-k] = a_{k+1,1}a_{k+2,2} \cdots a_{l,l-k} > 0$ for $l = k+1, \ldots, n$. Thus, we have positive minors of orders 1 to $n-k$ and we conclude that A is TP_{n-k}.

For the converse, let us recall that $\varepsilon_i = 1$ for $i \leq n-k$ because A is TP_{n-k} and that $\varepsilon_{n-k+1} = -1$ by hypothesis. So we can apply Proposition 2 to conclude that $a_{ij} \neq 0$ whenever $|i-j| \leq k$. Finally, by Theorem 1, we have that A is strictly k–banded. □

The next result corresponds to Theorem 3.3 of [2] and characterizes tridiagonal nonsingular SR matrices.

Theorem 3 *Let A be an $n \times n$ ($n \geq 2$) tridiagonal nonsingular nonnegative matrix. Then A is SR if and only if A is TP_{n-1}.*

The following result completes the classification of tridiagonal nonsingular SR matrices and shows that there are only the following cases: either A or $-A$ is TP or A is strictly tridiagonal.

Theorem 4 *Let A be an $n \times n$ tridiagonal nonsingular matrix. Then A is SR if and only if either A or $-A$ is TP or A is strictly tridiagonal SR.*

Proof As we have seen in Theorem 3, there are only two possible signatures for a tridiagonal nonsingular SR nonnegative matrix: either $(1, \ldots, 1)$ (TP case) or $(1, \ldots, 1, -1)$. Since a SR matrix is either nonnegative or opposite to a nonnegative matrix, if A is tridiagonal nonsingular SR, then either A is TP or has signature $(1, \ldots, 1, -1)$ or it is opposite to a matrix satisfying one of the previous properties. Since the opposite to a strictly tridiagonal matrix is also strictly tridiagonal, it only remains to see that A is strictly tridiagonal when A is nonnegative and has signature $(1, \ldots, 1, -1)$. Since $\varepsilon_n = -1$, we can derive from Proposition 2 (for $k = n$) and Theorem 1 that A is strictly tridiagonal.

The converse is trivially satisfied. □

Let us now focus on SR pentadiagonal matrices. We can derive the following corollary from Theorem 2.

Corollary 1 *Let A be a an $n \times n$ pentadiagonal nonsingular SR nonnegative matrix with signature $\varepsilon = (\varepsilon_i)_{1 \le i \le n}$ and $\varepsilon_{n-1} = -1$. Then A is strictly pentadiagonal if and only if A is TP_{n-2}.*

We can also extend Theorem 2 to anti–k–banded matrices, as the following result shows.

Theorem 5 *Let A be an $n \times n$ anti–k–banded nonsingular SR nonnegative matrix with signature $\varepsilon = (\varepsilon_i)_{1 \le i \le n}$ and $\varepsilon_{n-k+1} = (-1)^{\lfloor \frac{n-k+1}{2} \rfloor + 1}$. Then A is strictly anti–k–banded if and only if A has signature $\varepsilon_i = (-1)^{\lfloor \frac{i}{2} \rfloor}$ for $i \le n-k$.*

Proof Let us consider the matrix P presented in (1). Recall that P is a SR matrix with signature $\hat{\varepsilon} = (\hat{\varepsilon}_i)_{1 \le i \le n}$ given by $\hat{\varepsilon}_i = (-1)^{\lfloor \frac{i}{2} \rfloor}$ for $i = 1, \ldots, n$.

Now, by Proposition 1, we have that AP is a SR matrix and the $(n-k+1)$–th entry of the signature of AP is $\varepsilon_{n-k+1}\hat{\varepsilon}_{n-k+1} = (-1)^{2\lfloor \frac{n-k+1}{2} \rfloor + 1} = -1$. Besides, AP is nonsingular, nonnegative and k–banded. Thus, by Theorem 2, we have that AP is strictly k–banded if and only if AP is TP_{n-k}. Let us observe that $P^{-1} = P$ and then $(AP)P = A$. Note that if AP is TP_{n-k}, we have, by Proposition 1, that A has signature $\varepsilon_i = (-1)^{\lfloor \frac{i}{2} \rfloor}$ for $i \le n-k$. Finally, we can conclude that A is strictly k–banded if and only if A has signature $\varepsilon_i = (-1)^{\lfloor \frac{i}{2} \rfloor}$ for $i \le n-k$. □

4 Sufficient Conditions for the Total Positivity of Tridiagonal Matrices

In this section we present several sufficient conditions for the total positivity of tridiagonal matrices. We start with a consequence of Theorem 4.

Theorem 6 *Let $A = (a_{ij})_{1 \le i,j \le n}$ be a tridiagonal nonsingular SR nonnegative matrix. If $a_{ij} = 0$ for some i, j with $|i-j| = 1$, then A is TP.*

Now, we are going to present a formula for the determinant of a tridiagonal matrix which will be used to derive a new sufficient condition for the total positivity of a tridiagonal matrix. Let us observe that the expression for the determinant of a matrix is reduced if we consider tridiagonal matrices. In fact, if A is tridiagonal, its determinant can be expressed as $\det A = a_{11} \det A[2, \ldots, n] - a_{21} \det A[1, 3, \ldots, n \mid 2, \ldots, n]$. Expanding the previous formula we have that

$$\det A = a_{11} \det A[2, \ldots, n] - a_{21}a_{12} \det A[3, \ldots, n]. \tag{2}$$

The following result generalizes the previous formula. Let us mention that if α is the empty set, then we say that $\det A[\alpha] = 1$. For instance, if A is an $n \times n$ matrix, we have that $\det A[1, 0] = 1$ and $\det A[n+1, n] = 1$.

Let us prove the following formula to compute the determinant of a tridiagonal matrix. This formula was presented in p. 99 of [8] without proof. We include a proof in the following result for the sake of completeness.

Lemma 2 *Let A be a tridiagonal $n \times n$ matrix ($n > 3$), then*

$$\begin{aligned} \det A = &\det A[1,\dots,i]\det A[i+1,\dots,n] \\ &-a_{i,i+1}a_{i+1,i}\det A[1,\dots,i-1]\det A[i+2,\dots,n] \end{aligned} \tag{3}$$

for $i = 1,\dots,n-1$.

Proof If $i = 1$, then the formula is the trivial case (2). We proceed analogously if $i = n-1$.

For $i \in \{2,\dots,n-2\}$, we proceed by induction on i. If $i = 2$, then we expand the determinant of A and, since A is tridiagonal, we have that

$$\begin{aligned} \det A &= a_{11}\det A[2,\dots,n] - a_{12}a_{21}\det A[3,\dots,n] \\ &= a_{11}(a_{22}\det A[3,\dots,n] - a_{23}a_{32}\det A[4,\dots,n]) - a_{12}a_{21}\det A[3,\dots,n] \\ &= (a_{11}a_{22} - a_{12}a_{21})\det A[3,\dots,n] - a_{11}a_{23}a_{32}\det A[4,\dots,n] \\ &= \det A[1,2]\det A[3,\dots,n] - a_{23}a_{32}\det A[1]\det A[4,\dots,n]. \end{aligned}$$

Thus, formula (3) holds for $i = 2$.

Suppose that the formula is valid for $i = k-1$. Let us prove that it is also valid for $i = k$. By the induction hypothesis we know that

$$\begin{aligned} \det A = &\det A[1,\dots,k-1]\det A[k,\dots,n] \\ &-a_{k-1,k}a_{k,k-1}\det A[1,\dots,k-2]\det A[k+1,\dots,n]. \end{aligned} \tag{4}$$

If we expand $\det A[k,\dots,n]$ we have that

$$\det A[k,\dots,n] = a_{kk}\det A[k+1,\dots,n] - a_{k,k+1}a_{k+1,k}\det A[k+2,\dots,n],$$

and then, we can write the product $\det A[1,\dots,k-1]\det A[k,\dots,n]$ as

$$\begin{aligned} &\det A[1,\dots,k-1]a_{kk}\det A[k+1,\dots,n] \\ &-\det A[1,\dots,k-1]a_{k,k+1}a_{k+1,k}\det A[k+2,\dots,n]. \end{aligned} \tag{5}$$

So, by (4) and (5), we have that

$$\begin{aligned} \det A = &(\det A[1,\dots,k-1]a_{kk} \\ &-a_{k-1,k}a_{k,k-1}\det A[1,\dots,k-2])\det A[k+1,\dots,n] \\ &-a_{k,k+1}a_{k+1,k}\det A[1,\dots,k-1]\det A[k+2,\dots,n]. \end{aligned}$$

Since $A[1,\ldots,k]$ is tridiagonal and taking into account the definition of determinant, we have that

$$\det A[1,\ldots,k] = a_{kk} \det A[1,\ldots,k-1] - a_{k-1,k}a_{k,k-1} \det A[1,\ldots,k-2].$$

Then, replacing this in the previous formula, we can conclude that formula (3) holds for all $i \leq n-1$. □

It is well–known (see p. 100 of [8]) that given an $n \times n$ tridiagonal nonnegative matrix A, it is sufficient to verify that $\det A[1,\ldots,k] > 0$ for $k = 1,\ldots,n$ to conclude that A is TP. That characterization involves n minors. Theorem 7 can be used to reduce to $n-1$ the number of minors that guarantee that a tridiagonal nonnegative matrix is TP.

Theorem 7 *Let A be an $n \times n$ ($n \geq 3$) tridiagonal nonnegative matrix. If $\det A[1,\ldots,k] > 0$ for $k \leq n-2$ and $\det A > 0$, then A is TP.*

Proof Since the positivity of the leading principal minors of A implies that A is TP (see p. 100 of [8]), it remains to prove that $\det A[1,\ldots,n-1] > 0$. Let us start seeing that $a_{nn} \neq 0$. If $a_{nn} = 0$, then by formula (3) for $i = n-1$, we have that $\det A = -a_{n-1,n}a_{n,n-1} \det A[1,\ldots,n-2]$, where $a_{n-1,n}a_{n,n-1} \geq 0$ and $\det A[1,\ldots,n-2] > 0$ by hypothesis. Then $\det A \leq 0$, which contradict the fact that A has positive determinant. So we have that $a_{nn} > 0$.

Now, if $a_{n-1,n}a_{n,n-1} = 0$, then $(0 <) \det A = a_{nn} \det A[1,\ldots,n-1]$. Thus, we conclude that $\det A[1,\ldots,n-1] > 0$.

If $a_{n-1,n}a_{n,n-1} \neq 0$ we have, by (3), that

$$\det A = a_{nn} \det A[1,\ldots,n-1] - a_{n-1,n}a_{n,n-1} \det A[1,\ldots,n-2],$$

where $\det A > 0$, $a_{n-1,n}a_{n,n-1} \geq 0$ and $\det A[1,\ldots,n-2] > 0$ by hypothesis and $a_{nn} > 0$. So we have that $\det A[1,\ldots,n-1] > 0$. □

Example 1 shows that Theorem 7 cannot be extended to k–banded matrices in general.

Finally, we shall derive a last sufficient condition for total positivity using a condition on diagonal dominance. Let us recall that a matrix $A = (a_{ij})_{1\leq i,j\leq n}$ satisfying that $|a_{ii}| > \sum_{j=1,j\neq i}^{n} |a_{ij}|$, for all $i = 1,\ldots,n$ is called *strictly diagonally dominant (by rows)*.

The following result can be found in Theorem 6.1.10 of [4] and it contains the Levy–Desplanques theorem.

Theorem 8 *Let $A = (a_{ij})_{1\leq i,j\leq n}$ be a strictly diagonally dominant matrix. Then*

(i) A is nonsingular.
(ii) Besides, if $a_{ii} > 0$ for all $i \leq n$, then all eigenvalues of A have positive real part.

From the previous theorem we can derive more properties of strictly diagonally dominant matrices, as the following results shows.

Proposition 3 *Let* $A = (a_{ij})_{1\leq i,j\leq n}$ *be a strictly diagonally dominant matrix. If* $a_{ii} > 0$ *for all* $i \leq n$, *then* $\det A[1,\ldots k] > 0$ *for all* $k \leq n$.

Proof Since the determinant is the product of the eigenvalues, we know by Theorem 8 that A has positive determinant. Furthermore, since every submatrix of A of the form $A[1,\ldots,k]$ for $k = 1,\ldots,n$ is also strictly diagonally dominant and it has positive diagonal entries, then, again by Theorem 8, $\det A[1,\ldots,k] > 0$ for all $k \leq n$. □

The following result shows that strict diagonal dominance is a sufficient condition for total positivity for tridiagonal nonnegative matrices.

Theorem 9 *Let* $A = (a_{ij})_{1\leq i,j\leq n}$ *be a tridiagonal nonnegative matrix. If* A *is strictly diagonally dominant, then* A *is TP.*

Proof Observe that, since A is strictly diagonally dominant and nonnegative, then $a_{ii} > 0$ for all $i = 1,\ldots,n$. Thus we can apply Proposition 3 to conclude that $\det A[1,\ldots,k] > 0$ for all $k \leq n$. So, by Theorem 7, A is TP. □

The following example shows that neither Theorem 7 nor Theorem 9 can be extended to k–banded matrices for $k \geq 2$.

Example 1 Let us consider the matrix

$$A = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 3 \end{pmatrix}.$$

This matrix is pentadiagonal nonnegative and strictly diagonally dominant but A has a negative minor $\det A[2,3|1,2] = -2$, and so A is not TP. So, the matrix A shows that Theorem 9 cannot be extended to k–banded matrices in general. Analogously, we observe that Theorem 7 cannot be extended to k–banded matrices in general: $\det A[1] = 3 > 0$ and $\det A = 20 > 0$ but A is not TP.

References

1. Ando, T.: Totally positive matrices. Linear Algebra Appl. **90**, 165–219 (1987)
2. Barreras, A., Peña, J.M.: Characterizations of Jacobi sign regular matrices. Linear Algebra Appl. **436**, 381–388 (2012)
3. Cortés, V., Peña, J.M.: Required nonzero patterns for nonsingular sign regular matrices. Linear Algebra Appl. **432**, 1990–1994 (2010)
4. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge (1985)
5. Huang, R.: Sign structure preserves for m-banded factorizations of sign regular matrices. Linear Algebra Appl. **436**, 1990–2000 (2012)

6. Koev, P.: Accurate computations with totally nonnegative matrices. SIAM J. Matrix Anal. Appl. **29**, 731–751 (2007)
7. Peña, J.M.: Sign regular matrices of order two. Linear Multilinear A. **50**, 91–97 (2002)
8. Pinkus, A.: Totally Positive Matrices. Cambridge Tracts in Mathematics, vol. 181. Cambridge University Press, Cambridge (2010)

High Order Variational Integrators: A Polynomial Approach

Cédric M. Campos

Abstract We reconsider the variational derivation of symplectic partitioned Runge-Kutta schemes. Such type of variational integrators are of great importance since they integrate mechanical systems with high order accuracy while preserving the structural properties of these systems, like the symplectic form, the evolution of the momentum maps or the energy behaviour. Also they are easily applicable to optimal control problems based on mechanical systems as proposed in Ober-Blöbaum et al. (ESAIM Control Optim Calc Var 17(2):322–352, 2011).

Following the same approach, we develop a family of variational integrators to which we refer as symplectic Galerkin schemes in contrast to symplectic partitioned Runge-Kutta. These two families of integrators are, in principle and by construction, different one from the other. Furthermore, the symplectic Galerkin family can as easily be applied to optimal control problems, for which Campos et al. (Higher order variational time discretization of optimal control problems. In: Proceedings of the 20th international symposium on mathematical theory of networks and systems, Melbourne, 2012) is a particular case.

1 Introduction

In recent years, much effort in designing numerical methods for the time integration of (ordinary) differential equations has been put into schemes which are *structure preserving* in the sense that important *qualitative* features of the original dynamics are preserved in its time discretization, cf. the recent monograph [6]. A particularly elegant way to, e.g. derive symplectic integrators, is by discretizing Hamilton's principle as suggested by [14, 15], see also [11, 13].

However most part of the theory and examples rely on second order schemes, hence some effort must still be put into the development of accurate high order schemes that, in long term simulations, can drastically reduce the overall computational cost. A clear example are the so called symplectic partitioned Runge-Kutta

C.M. Campos (✉)
Instituto de Ciencias Matemáticas, Campus de Cantoblanco, Calle Nicolás Cabrera 15, 28049 Madrid, Spain
e-mail: cedricmc@icmat.es

F. Casas, V. Martínez (eds.), *Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 4, DOI 10.1007/978-3-319-06953-1_24

methods that integrate mechanical systems driven by a Lagrangian $L\colon (q,\dot{q}) \in \mathbb{R}^n \times \mathbb{R}^n \mapsto L(q,\dot{q}) \in \mathbb{R}$ and, possibly, by a force $f\colon (q,\dot{q}) \in \mathbb{R}^n \times \mathbb{R}^n \mapsto p = f(q,\dot{q}) \in \mathbb{R}^n$. A detailed study of such methods can be found in [6].

The paper is structured as follows: Sect. 2 is a short introduction to Discrete Mechanics and Sect. 3 describes the variational derivation of high order schemes using polynomial collocation. Finally we briefly enumerate the relations and differences between symplectic partitioned Runge-Kutta schemes and symplectic Galerkin ones in Sect. 4 and conclude by outlining future research directions in the last section.

2 Discrete Mechanics and Variational Integrators

One of the main subjects of Geometric Mechanics is the study of dynamical systems governed by a Lagrangian. Typically one considers a mechanical system with *configuration manifold* Q together with a *Lagrangian function* $L\colon TQ \to \mathbb{R}$, where the associated *state space* TQ describes the position and velocity of a particle moving in the system. A consequence of the *principle of least action*, also known as *Hamilton's principle*, establishes that the natural motions $q\colon [0,T] \to Q$ of the system are characterized by the celebrated *Euler-Lagrange equation* (refer to [1]),

$$\frac{\partial L}{\partial q} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}} = 0\,. \tag{1}$$

Different preservation laws are present in these systems. For instance the Hamiltonian flow preserves the natural symplectic structure of T^*Q and the total energy of the system. Also, if the Lagrangian possess Lie group symmetries, then *Noether's theorem* asserts that some quantities are conserved, like for instance the linear momentum and/or the angular momentum.

Discrete Mechanics is, roughly speaking, a discretization of Geometric Mechanics theory. As a result, one obtains a set of discrete equations equivalent to the Euler-Lagrange equation (1) above but, instead of a direct discretization of the ODE, the latter are derived from a discretization of the base objects of the theory, the state space TQ, the Lagrangian L, etc. In fact, one seeks for a sequence $\{(t_0,q_0),(t_1,q_1),\ldots,(t_n,q_n)\}$ that approximates the actual trajectory $q(t)$ of the system ($q_k \approx q(t_k)$), for a constant time-step $h = t_{k+1} - t_k > 0$.

A *variational integrator* is an iterative rule that outputs this sequence and it is derived in an analogous manner to the continuous framework. Given a discrete Lagrangian $L_d\colon Q \times Q \to \mathbb{R}$, which is in principle thought to approximate the continuous Lagrangian action over a short time

$$L_d(q_k,q_{k+1}) \approx \int_{t_k}^{t_{k+1}} L(q(t),\dot{q}(t))\mathrm{d}t\,,$$

one applies a variational principle to derive the well-known discrete Euler-Lagrange (DEL) equation,

$$D_1 L_d(q_k, q_{k+1}) + D_2 L_d(q_{k-1}, q_k) = 0\,, \tag{2}$$

where D_i stands for the partial derivative with respect to the i-th component. The equation defines an integration rule of the type $(q_{k-1}, q_k) \mapsto (q_k, q_{k+1})$, however if we define the pre- and post-momenta

$$p_k^- := -D_1 L_d(q_k, q_{k+1}) \qquad \text{and} \qquad p_k^+ := D_2 L_d(q_{k-1}, q_k)\,, \tag{3}$$

the Euler-Lagrange equation (2) is read as the momentum matching $p_k^- = p_k^+ =: p_k$ and defines an integration rule of the type $(q_k, p_k) \mapsto (q_{k+1}, p_{k+1})$.

The nice part of the story is that the integrators derived in this way naturally preserve (or nearly preserve) the quantities that are preserved in the continuous framework, the symplectic form, the total energy and, in presence of symmetries, the linear and/or angular momentum (for more details, see [11]). Furthermore, other aspects of the continuous theory can be "easily" adapted, symmetry reduction [2,4,7], constraints [8,9], control forces [3,12], etc.

3 High Order Variational Integrators

High order variational integrators for time dependent or independent systems (HOVI) are a class of integrators that, by using a multi-stage approach, aim at a high order accuracy on the computation of the natural trajectories of a mechanical system while preserving some intrinsic properties of such systems. In particular, symplectic-partitioned Runge-Kutta methods (spRK) and, what we call here, symplectic Galerkin methods (sG) are s-stage variational integrators of order up to $2s$.

The derivation of these methods follows a general scheme. For a fixed time step h, one considers a series of points q_k, refereed as macro-nodes. Between each couple of macro-nodes (q_k, q_{k+1}), one also considers a set of micro-data, the s stages: For the particular cases of sG and spRK methods, micro-nodes $Q_1, \ldots, Q_s$ and micro-velocities $\dot{Q}_1, \ldots, \dot{Q}_s$, respectively. Both macro-nodes and micro-data (micro-nodes or micro-velocities) are required to satisfy a variational principle, giving rise to a set of equations, which properly combined, define the final integrator.

In what follows, we will use the following notation: Let $0 \le c_1 < \ldots < c_s \le 1$ denote a set of collocation points and consider the associated Lagrange polynomials and nodal weights, that is,

$$l^j(t) := \prod_{i \neq j} \frac{t - c_i}{c_j - c_i} \qquad \text{and} \qquad b_j := \int_0^1 l^j(t)\mathrm{d}t\,,$$

respectively. Note that the pair of (c_i, b_i)'s define a quadrature rule and that, for appropriate c_i's, this rule may be a Gaussian-like quadrature, for instance, Gauss-Legendre, Gauss-Lobatto, Radau or Chebyshev.

Now, for the sake of simplicity and independently on the method, we will use the same notation for the nodal coefficients. We define for spRK and sG, respectively,

$$a_{ij} := \int_0^{c_i} l^j(t)\mathrm{d}t \qquad \text{and} \qquad a_{ij} := \left.\frac{\mathrm{d}l^j}{\mathrm{d}t}\right|_{c_i} .$$

Moreover, for spRK, we will also use the nodal weights and coefficients $(\bar{b}_j, \bar{a}_{ij})$ given by Eq. (5) and, for sG, the source and target coefficients

$$\alpha^j := l^j(0) \qquad \text{and} \qquad \beta^j := l^j(1) .$$

Finally, we assume that L denotes a Lagrangian from $\mathbb{R}^n \times \mathbb{R}^n$ to $\mathbb{R}$, from which we define

$$P_i := \left.\frac{\partial L}{\partial \dot{q}}\right|_i = \left.\frac{\partial L}{\partial \dot{q}}\right|_{(Q_i, \dot{Q}_i)} \qquad \text{and} \qquad \dot{P}_i := \left.\frac{\partial L}{\partial q}\right|_i = \left.\frac{\partial L}{\partial q}\right|_{(Q_i, \dot{Q}_i)} ,$$

where $(Q_i, \dot{Q}_i)$ are couples of micro-nodes and micro-velocities given by each method. Besides, D_i will stand for the partial derivative with respect to the i-th component.

3.1 Symplectic-Partitioned Runge-Kutta Methods

Although the variational derivation of spRK methods in the framework of Geometric Mechanics is already known (see [11] for an "intrinsic" derivation, or [6] for a "constrained" one), it would be interesting to present it here again in order to ease the comprehension of and the comparison with sG methods below. However, due to (literal) space constraints, we will only point out the different ingredients with respect to the sG recipe. An "extended" version of this paper may be found on arXiv, with the detailed recipe.

A partitioned Runge-Kutta method is an s-stage integrator $(q_0, p_0) \mapsto (q_1, p_1)$ given by the equations

$$q_1 = q_0 + h \sum_{j=1}^{s} b_j \dot{Q}_j , \qquad p_1 = p_0 + h \sum_{j=1}^{s} \bar{b}_j \dot{P}_j , \tag{4a}$$

$$Q_i = q_0 + h \sum_{j=1}^{s} a_{ij} \dot{Q}_j , \qquad P_i = p_0 + h \sum_{j=1}^{s} \bar{a}_{ij} \dot{P}_j , \tag{4b}$$

$$P_i = \frac{\partial L}{\partial \dot{q}}(Q_i, \dot{Q}_i) , \qquad \dot{P}_i = \frac{\partial L}{\partial q}(Q_i, \dot{Q}_i) , \tag{4c}$$

where (b_j, a_{ij}) and $(\bar{b}_j, \bar{a}_{ij})$ are two different Runge-Kutta methods associated to collocation points $0 \leq c_1 < \ldots < c_s \leq 1$ and time step $h > 0$.

It is shown that the previous integrator is *symplectic* whenever the two sets of coefficients satisfy the relations

$$b_i \bar{a}_{ij} + \bar{b}_j a_{ji} = b_i \bar{b}_j \,, \qquad b_i = \bar{b}_i \,. \tag{5}$$

Moreover, it can be derived as a geometric variational integrator, a fact that was firstly noticed by Sanz-Serna [13] and Suris [14].

The main ingredient in spRK that differs from sG is the characterization of the space of polynomials where we extremize the Lagrangian. Here we first consider an initial point $q_0 \in \mathbb{R}^n$ and inner vectors $\{\dot{Q}_i\}_{i=1,\ldots,s} \subset \mathbb{R}^n$, in order to define the polynomial curves

$$\dot{Q}(t) := \sum_{j=1}^{s} l^j(t/h)\dot{Q}_j \qquad \text{and} \qquad Q(t) := q_0 + h\sum_{j=1}^{s}\int_0^{t/h} l^j(\tau)\mathrm{d}\tau\,\dot{Q}_j \,,$$

that give the target point and micro node equations (4a.1) and (4b.1).

3.2 Symplectic Galerkin Methods

Galerkin methods are a class of methods to transform a problem given by a continuous operator (such as a differential operator) to a discrete problem. As such, spRK methods falls into the scope of this technique and could be also classified as "symplectic Galerkin" methods. However, we want to stress on the difference between what is called spRK in the literature and what we here refer as sG. The wording should not be confused by the one used in [11].

Given points $\{Q_i\}_{i=1,\ldots,s} \subset \mathbb{R}^n$, we define the polynomial curves

$$Q(t) := \sum_{j=1}^{s} l^j(t/h)Q_j \qquad \text{and} \qquad \dot{Q}(t) := \frac{1}{h}\sum_{j=1}^{s} \dot{l}^j(t/h)Q_j \,.$$

We have

$$Q_i = Q(h\cdot c_i) \quad \text{and} \quad \dot{Q}_i := \dot{Q}(h\cdot c_i) = \frac{1}{h}\sum_{j=1}^{s} a_{ij}Q_j \,.$$

Note that the polynomial curve Q is uniquely determined by the points $\{Q_i\}_{i=1,\ldots,s}$. In fact, it is the unique polynomial curve Q of degree $s-1$ such that $Q(h\cdot c_i) = Q_i$. However, if we define the configuration points

$$q_0 := Q(h\cdot 0) = \sum_{j=1}^{s} \alpha^j Q_j \qquad \text{and} \qquad q_1 := Q(h\cdot 1) = \sum_{j=1}^{s} \beta^j Q_j \tag{6}$$

and consider them fixed, then Q is uniquely determined by q_0, q_1 and the Q_i's but a couple. For instance, we may consider Q_1 and Q_s as functions of the others, since the relations (6) define a system of linear equations where the coefficient matrix has determinant $\gamma := \alpha^1\beta^s - \alpha^s\beta^1 \neq 0$ (if and only if $c_1 \neq c_s$). More precisely,

$$\begin{pmatrix} Q_1 \\ Q_s \end{pmatrix} = \frac{1}{\gamma}\begin{pmatrix} \beta^s & -\alpha^s \\ -\beta^1 & \alpha^1 \end{pmatrix}\begin{pmatrix} q_0 - \sum_{j=2}^{s-1}\alpha^j Q_j \\ q_1 - \sum_{j=2}^{s-1}\beta^j Q_j \end{pmatrix}.$$

We now define the two-point discrete Lagrangian

$$L_d(q_0, q_1) := \operatorname*{ext}_{\mathscr{P}^{s-1}} L_d(Q_1, \ldots, Q_s) = \operatorname*{ext}_{\mathscr{P}^{s-1}} h\sum_{i=1}^{s} b_i L(Q_i, \dot{Q}_i)$$

where $\mathscr{P}^{s-1} = \mathscr{P}^{s-1}([0,h], \mathbb{R}^n, q_0, q_1)$ is the space of polynomials Q of order $s-1$ from $[0,1]$ to $\mathbb{R}^n$ such that the points Q_i's determine such polynomials as discussed above, where $L_d(Q_1, \ldots, Q_s) = h\sum_{i=1}^{s} b_i L(Q_i, \dot{Q}_i)$ is the multi-point discrete Lagrangian and where L is the continuous Lagrangian. The extremal is realized by a polynomial $Q \in \mathscr{P}^{s-1}([0,h], \mathbb{R}^n, q_0, q_1)$ such that

$$\delta L_d(Q_1, \ldots, Q_s) \cdot (\delta Q_1, \ldots, \delta Q_s) = 0 \tag{7}$$

for any variations $(\delta Q_1, \ldots, \delta Q_s)$, taking into account that

$$\delta q_0 = \delta q_1 = 0 \qquad \text{and} \qquad \delta Q_i = \sum_{j=2}^{s-1} \frac{\partial Q_i}{\partial Q_j}\delta Q_j\,,\ i = 1, s.$$

For convenience, the previous Eq. (7) is developed afterwards.

By the momenta-matching rule (3), we have that

$$\begin{aligned} p_0 &= -D_1 L_d(q_0, q_1) \\ &= -\beta^s/\gamma \cdot D_1 L_d(Q_1, \ldots, Q_s) \\ &\quad + \beta^1/\gamma \cdot D_s L_d(Q_1, \ldots, Q_s) \end{aligned} \qquad \begin{aligned} p_1 &= D_2 L_d(q_0, q_1) \\ &= -\alpha^s/\gamma \cdot D_1 L_d(Q_1, \ldots, Q_s) \\ &\quad + \alpha^1/\gamma \cdot D_s L_d(Q_1, \ldots, Q_s). \end{aligned}$$

Coming back to Eq. (7), we have that

$$\begin{aligned} &\delta L_d(Q_1, \ldots, Q_s) \cdot (\delta Q_1, \ldots, \delta Q_s) \\ &\quad = \sum_{j=2}^{s-1}\Bigg[D_1 L_d(Q_1, \ldots, Q_s)\frac{\partial Q_1}{\partial Q_j} + D_j L_d(Q_1, \ldots, Q_s) \\ &\quad + D_s L_d(Q_1, \ldots, Q_s)\frac{\partial Q_s}{\partial Q_j}\Bigg]\delta Q_j. \end{aligned}$$

Combining these, we obtain that for any $j = 1, \dots, s$

$$D_j L_d(Q_1, \dots, Q_s) = -\alpha^j p_0 + \beta^j p_1 . \tag{8}$$

The integrator is defined by

$$D_j L_d(Q_1, \dots, Q_s) = -\alpha^j p_0 + \beta^j p_1 , \ j = 1, \dots, s; \tag{9a}$$

$$q_0 = \sum_{j=1}^{s} \alpha^j Q_j \qquad \text{and} \qquad q_1 = \sum_{j=1}^{s} \beta^j Q_j . \tag{9b}$$

Finally, using the definition of the discrete Lagrangian, we may write the equations that define the sG integrator (without forces) in a pRK fashion, that is

$$q_0 = \sum_{j=1}^{s} \alpha^j Q_j , \qquad q_1 = \sum_{j=1}^{s} \beta^j Q_j , \tag{10a}$$

$$\dot{Q}_i = \frac{1}{h} \sum_{j=1}^{s} a_{ij} Q_j , \qquad \dot{P}_i = \frac{\beta^i p_1 - \alpha^i p_0}{h \bar{b}_i} + \frac{1}{h} \sum_{j=1}^{s} \bar{a}_{ij} P_j , \tag{10b}$$

$$P_i = \frac{\partial L}{\partial \dot{q}}(Q_i, \dot{Q}_i) , \qquad \dot{P}_i = \frac{\partial L}{\partial q}(Q_i, \dot{Q}_i) , \tag{10c}$$

where $b_i a_{ij} + \bar{b}_j \bar{a}_{ji} = 0$ and $b_i = \bar{b}_j$.

We remark that Eq. (8) generalizes the ones obtained in [3, 10], where the collocation points are chosen such that $c_1 = 0$ and $c_s = 1$, which is a rather particular case.

4 Relations Between spRK and sG

First of all, it is worth to say that, with a little bit of extra technicalities, one can easily include forces into both schemes. As a result, one would only need to redefine in (4) and (10)

$$\dot{P}_i = \frac{\partial L}{\partial q}(Q_i, \dot{Q}_i) + f(Q_i, \dot{Q}_i) ,$$

where $f : (q, \dot{q}) \in \mathbb{R}^n \times \mathbb{R}^n \mapsto p = f(q, \dot{q}) \in \mathbb{R}^n$ is the external force.

As already mentioned, both methods can be considered of Galerkin type. In this sense, spRK and sG could be refereed as a symplectic Galerkin integrators of 1st and 0th kind, respectively, since spRK is derived from the 1st derivative of an extremal polynomial and sG from the polynomial itself. At this point, a very natural

question could arise: Are spRK and sG actually two different integrator schemes? Even though the derivations of both methods are quite similar, they are in general different (although they could coincide for particular choices of the Lagrangian, the collocation points and the integral quadrature). A weak but still fair argument to support this is that, at each step, spRK relies on the determination of the micro-velocities $\dot{Q}_i$, while sG does so on the micro-nodes Q_i. All the other "unknowns" are then computed from the determined micro-data.

Example 1 Here we are going to consider the perhaps simplest case of all, that is, a Lagrangian of the form kinetic minus potential energy, $L(q,\dot{q}) = \frac{1}{2}\dot{q}^T M \dot{q} - U(q)$, with M a constant mass matrix; $s = 2$ micro-nodes (inner-stages); and Lobatto's quadrature, $c_1 = 0$, $c_2 = 1$. Under such assumptions and after some simple computations, both schemes, spRK (4) and sG (10), will reduce to the well-known *leap-frog or Verlet method*:

$$\begin{aligned} p_{1/2} &= p_0 - \tfrac{h}{2}\nabla U_0\,, \\ q_1 &= q_0 + hM^{-1}p_{1/2}\,, \\ p_1 &= p_{1/2} - \tfrac{h}{2}\nabla U_1\,. \end{aligned}$$

Example 2 In the previous example, two of the assumptions are crucial so the spRK and sG schemes coincide. To counter it, we resume Example 1, this time considering a Lagrangian with a scalar mass matrix dependent on the configuration, that is, a Lagrangian of the form $L(q,\dot{q}) = \frac{1}{2}\lambda(q)\|\dot{q}\| - U(q)$, with $\lambda\colon Q \to \mathbb{R}$. Under this modified assumption and noting $\lambda_{1/2} := \frac{\lambda_0+\lambda_1}{2}$, $(\nabla)\lambda_i := (\nabla)\lambda(q_i)$, $(\nabla)U_i := (\nabla)U(q_i)$, $i = 0, 1$, the spRK scheme (4) as well as the sG scheme (10) reduce to

$$\begin{aligned} p_{1/2} &= p_0 + \frac{h}{2}\left(\frac{\nabla\lambda_0}{2\lambda_{\mathsf{a}}^2}\|p_{1/2}\|^2 - \nabla U_0\right)\,, \\ q_1 &= q_0 + \frac{h}{2}\left(\frac{1}{\lambda_{\mathsf{a}}} + \frac{1}{\lambda_{\mathsf{b}}}\right)p_{1/2}\,, \\ p_1 &= p_{1/2} + \frac{h}{2}\left(\frac{\nabla\lambda_1}{2\lambda_{\mathsf{b}}^2}\|p_{1/2}\|^2 - \nabla U_1\right)\,, \end{aligned}$$

with a slight difference. While for the spRK scheme the subindexes are $\mathsf{a} = 0$ and $\mathsf{b} = 1$, in the sG scheme they are $\mathsf{a} = \mathsf{b} = 1/2$. It is important to note that, even though the difference is small, it makes both schemes certainly different. Besides and in any case, one may notice that they reduce to the Verlet method for a constant λ and that in the general case the first two equations define $p_{1/2}$ and q_1 implicitly.

Example 3 We counter again Example 1, but this time considering an alternate quadrature rule, Legendre's one, that is, $c_1 = \frac{1}{2} - \frac{1}{\sqrt{3}}$, $c_2 = \frac{1}{2} + \frac{1}{\sqrt{3}}$. Under this modified assumption and noting $q_{1/2} = \frac{Q_1+Q_2}{2}$, $\dot{q}_{1/2} = \sqrt{3}\frac{Q_2-Q_1}{h}$, the spRK and sG schemes (10) again reduce to a similar one

$$q_0 = q_{1/2} - \tfrac{h}{2}\dot{q}_{1/2} + \boxed{\tfrac{h^2}{24}M^{-1}(\dot{P}_1 + \dot{P}_2)}, \qquad q_1 = q_0 + h\dot{q}_{1/2},$$

$$p_0 = M\dot{q}_{1/2} - \tfrac{h}{2}\left[\left(\tfrac{1}{2} + \tfrac{\sqrt{3}}{6}\right)\dot{P}_1 + \left(\tfrac{1}{2} - \tfrac{\sqrt{3}}{6}\right)\dot{P}_2\right], \qquad p_1 = p_0 + \tfrac{h}{2}(\dot{P}_1 + \dot{P}_2),$$

where the framed term appears only in the spRK case. Note that the left equations define implicitly Q_1 and Q_2, which are of use in the right ones.

With respect to the accuracy of the schemes, for any Gaussian quadrature (Gauss-Legendre, Gauss-Lobatto, Radau and Chebyshev) and any method (spRK and sG), the schemes have convergence order $2s - 2$, except for the combination of Gauss-Lobatto together with spRK which is $2s$, being s the number of internal stages.

Let's finish underlying that sG, as spRK, is inherently symplectic.

Conclusions

In this work, by revisiting the variational derivation of spRK methods [6, 11], we have presented a new class of high order variational integrators within the family of Galerkin schemes. These integrators are symplectic per construction and, therefore, well suited for long term simulations, where the high order accuracy of the schemes can be exploited to reduce the overall computational cost. Also, they can be easily adapted to implement constraints or symmetry reduction [2, 7, 11] or, together with a non-linear programming (NLP) solver, to integrate optimal control problems [3, 12].

For the future, the sG schemes deserve a proper analysis to establish the actual differences with respect to spRK schemes and results on the convergence rates. And to further take advantage of the high accuracy of the methods, we envisage to design time adaptive algorithms. Joint work with O. Junge (TUM), S. Ober-Blöbaum (UPB) and E. Trélat (CNRS-UPMC) on the control direction has already started in order to generalize [3] and clarify some aspects of [5]. As well, we have begun digging along the lines of constrained systems and higher order Lagrangians.

Acknowledgements This work has been partially supported by MEC (Spain) Grants MTM2010-21186-C02-01, MTM2011-15725E, the ICMAT Severo Ochoa project SEV-2011-0087 and the European project IRSES-project "Geomech-246981". Besides, the author wants to specially thank professors Sina Ober-Blöbaum, from the University of Paderborn, and David Martín de Diego, from the Instituto de Ciencias Matemáticas, for fruitful conversations on the subject and their constant support.

References

1. Abraham, R., Marsden, J.E.: Foundations of Mechanics. Benjamin/Cummings, Reading (1978)
2. Campos, C.M., Cendra, H., Díaz, V.A., Martín de Diego, D.: Discrete Lagrange-d'Alembert-Poincaré equations for Euler's disk. Rev. R. Acad. Cienc. Exactas Fís. Nat. Ser. A Math. RACSAM. **106**(1), 225–234 (2012)
3. Campos, C.M., Junge, O., Ober-Blöbaum, S.: Higher order variational time discretization of optimal control problems. In: Proceedings of the 20th International Symposium on Mathematical Theory of Networks and Systems, Melbourne (2012)
4. Colombo, L., Jiménez, F., Martín de Diego, D.: Discrete second-order Euler-Poincaré equations. Applications to optimal control. Int. J. Geom. Methods Mod. Phys. **9**(4), 1250037 (2012)
5. Hager, W.W.: Runge-Kutta methods in optimal control and the transformed adjoint system. Numer. Math. **87**(2), 247–282 (2000)
6. Hairer, E., Lubich, C., Wanner, G.: Geometric Numerical Integration. Springer Series in Computational Mathematics, vol. 31. Springer, Heidelberg (2010)
7. Iglesias, D., Marrero, J.C., de Diego, D.M., Martínez, E.: Discrete nonholonomic Lagrangian systems on Lie groupoids. J. Nonlinear Sci. **18**(3), 221–276 (2008)
8. Johnson, E.R., Murphey, T.D.: Dangers of two-point holonomic constraints for variational integrators. In: Proceedings of American Control Conference, St. Louis, Missouri, USA, June 10–12, 2009. IEEE, pp. 4723–4728. Piscataway (2009). http://www.a2c2.org/conferences/acc2009/
9. Kobilarov, M., Marsden, J.E., Sukhatme, G.S.: Geometric discretization of nonholonomic systems with symmetries. Discret. Cont. Dyn. Syst. **3**(1), 61–84 (2010)
10. Leok, M.: Foundations of computational geometric mechanics. Ph.D. thesis, California Institute of Technology (2004)
11. Marsden, J.E., West, M.: Discrete mechanics and variational integrators. Acta Numer. **10**, 357–514 (2001)
12. Ober-Blöbaum, S., Junge, O., Marsden, J.: Discrete mechanics and optimal control: an analysis. ESAIM Control Optim. Calc. Var. **17**(2), 322–352 (2011)
13. Sanz-Serna, J.M., Calvo, M.P.: Numerical Hamiltonian Problems. Applied Mathematics and Mathematical Computation, vol. 7. Chapman & Hall, London (1994)
14. Suris, Y.B.: Hamiltonian methods of Runge-Kutta type and their variational interpretation. Math. Model. **2**(4), 78–87 (1990)
15. Veselov. A.P.: Integrable systems with discrete time, and difference operators. Funct. Anal. Appl. **22**(2), 83–93 (1988)

A Block Compression Algorithm for Computing Preconditioners

Juana Cerdán, José Marín, and José Mas

Abstract To implement efficiently algorithms for the solution of large systems of linear equations in modern computer architectures, it is convenient to unravel the block structure of the coefficient matrix that is present in many applications of the physics and the engineering. This is specially important when a preconditioned iterative method is used to compute an approximate solution. Identifying such a block structure is a graph compression problem and several techniques have been studied in the literature. In this work we consider the *cosine* algorithm introduced by Y. Saad. This algorithm groups two rows of the matrix if the corresponding angle between them in the adjacency matrix is small enough. The modification that we propose considers also the magnitude of the nonzero entries of the rows with the aim of computing a better block partition.

1 Introduction

Very often Numerical Linear Algebra applications give raise to systems of linear equations

$$Ax = b\,, \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$ is a large and nonsingular sparse matrix. These problems are usually solved by iterative Krylov methods since they require less computational time and storage than counterpart direct methods based on Gaussian elimination. It is well known than the convergence of iterative Krylov methods is improved if a good preconditioner is used, see [10]. The aim of the preconditioning technique is to improve the condition number or the eigenvalue distribution of the coefficient matrix. On the other hand the cost of computing the preconditioner and its application inside the iterative method should be negligible and performed as efficiently as possible. To this end several block versions of the most popular preconditioners have

J. Cerdán (✉) • J. Marín • J. Mas
Instituto de Matemática Multidisciplinar, Universitat Politècnica de València, Camino de Vera s/n, E-46022 Valencia, Spain
e-mail: jcerdan@imm.upv.es; jmarinma@imm.upv.es; jmasm@imm.upv.es

F. Casas, V. Martínez (eds.), *Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 4, DOI 10.1007/978-3-319-06953-1_25

been proposed for a variety of structured problems. For example, matrices arising from the discretization of partial differential equations have often a natural block structure [5]. This structure consists of small and dense submatrices that can be treated as individual entries of the matrix. With appropriate sparse storage formats and using basic linear algebra subroutines (BLAS) the computational performance can be improved [8]. Examples of this block preconditioning approach can be found in the literature [2–4, 6, 7]. In every case an improvement in the efficiency compared to its point version is reported.

Finding the block structure of a matrix is a graph compression problem, and has been studied by different authors [1, 9]. In [11] the author propose the cosine algorithm that finds an approximate block structure since it allows some zero entries in the dense blocks. The goal of this paper is to modify the algorithm such that besides the nonzero pattern, the magnitude of the entries is considered.

The paper is organized as follows. The cosine algorithm is revised in Sect. 2. Section 3 is devoted to the motivation and description of the proposed modification. Numerical results of experiments with different matrices are shown in Sect. 4. Finally, section "Conclusions" summarizes the work.

2 The Cosine Algorithm

In this section we describe the cosine algorithm proposed by Y. Saad in [11]. Let us start with an example that illustrates the matrix graph compression problem. Consider the next symmetric nonzero pattern:

$$\left[\begin{array}{ccc|cc|c|c} * & * & * & 0 & 0 & 0 & * \\ * & * & * & 0 & 0 & 0 & * \\ * & * & * & 0 & 0 & 0 & * \\ \hline 0 & 0 & 0 & * & * & 0 & 0 \\ 0 & 0 & 0 & * & * & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & * & * \\ \hline * & * & * & 0 & 0 & * & * \end{array}\right],$$

where each $*$ represents a nonzero entry. This matrix has a block structure of sizes 3, 2, 1 and 1, respectively. The adjacency graph $G = (V, E)$ of a matrix consists of the node set V that corresponds to the rows or columns (unknowns) of the matrix, and the set E such that there is an edge from node i to node j if $a_{ij} \neq 0$. The nodes can be grouped into four subgroups: $Y_1 = \{1, 2, 3\}$, $Y_2 = \{4, 5\}$, $Y_3 = \{6\}$ and $Y_4 = \{7\}$. The corresponding restricted graph to each one of these subgroups

is complete and it can be represented by one entry on the adjacency matrix of the associated quotient graph

$$\begin{bmatrix} * & 0 & 0 & * \\ 0 & * & 0 & 0 \\ 0 & 0 & * & * \\ * & 0 & * & * \end{bmatrix}$$

where each $*$ in the position (i, j) corresponds to a dense block of size $|Y_i| \times |Y_j|$, where $|X|$ is the cardinality of the set X.

To detect the partition of the nodes and therefore the block structure of a matrix different algorithms have been proposed. Algorithms based on *hash* functions assign a different value to each nonzero row pattern. For instance in [1] it is used the hash function

$$\text{hash}(u) = \sum_{(u,w)\in E} w\,.$$

These values allow to group two nodes with the same hash value, though further refinements may be necessary. Hash-based algorithms are not useful to detect almost complete subgraphs as it is illustrated by the next example. It has been obtained from the previous one by introducing some zeros on the dense subblocks,

$$\left[\begin{array}{ccc|ccc|c} * & 0 & * & 0 & 0 & 0 & * \\ * & * & 0 & 0 & 0 & 0 & * \\ * & * & * & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & * & * & 0 & 0 \\ 0 & 0 & 0 & * & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & * & * \\ \hline * & * & * & 0 & 0 & * & * \end{array}\right].$$

Using a hash function we would not be able to find the same block structure because all the row patterns and their corresponding hash values are different. That is, no block structure will be found. To detect the same partition of the nodes the algorithm should allow some few zero entries in the dense subblocks. One technique that can be used to compute such a kind of approximate block structures is the cosine algorithm. It is based on the idea of computing the angle between the rows, or columns, of the adjacency matrix. For example, the adjacency matrix C of the previous matrix is

$$\left[\begin{array}{ccc|ccc|c} 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ \hline 1 & 1 & 1 & 0 & 0 & 1 & 1 \end{array}\right],$$

where each nonzero entry has been replaced by 1. We note that the (i, j) entry of the matrix CC^T is the inner product of the rows i and j. Thus, the cosine of the angle between two given rows of the adjacency matrix can be easily obtained by computing the upper triangular part of this matrix. Large cosine values correspond to rows with similar nonzero patterns and therefore, an approximate block structure can be detected.

Algorithm 3 shows the cosine algorithm. The vector *Group* stores the index of the group where each row belongs to. If *Group(i)* $= -1$ it means that row i is either not grouped or it is the leader row of the i-th group. Lines 5–10 implement the computation of the inner product between rows which is stored in vector $Count$. The cosine evaluation is done in lines 11–14. The parameter τ is the minimum value for the cosine between two rows such that they can be grouped together. Tipically $1 \geq \tau \geq 0.7$ even for some matrices smaller values may be needed to find blocks of moderate size. If $\tau = 1$ only rows with exactly the same nonzero pattern are grouped performing in that case as a hash-based algorithm. The biggest block sizes are obtained for small τ values but probably at the cost of introducing a large amount of nonzero entries on the block partition. Finally, $nz_C(i)$ indicates the number of nonzero elements of the i-th row of the matrix C and corresponds to its norm. Further details can be found in [11].

Algorithm 3 *Cosine Algorithm*

Input: Adjacency matrix C and tolerance τ; Output: block partition.
Compute the pattern C^T.
Set *Group(i)*$= -1$ and *Count(j)*$= 0$ for $i, j = 1, \ldots, n$
For $i = 1, \ldots, n$ if *Group(i)*$= -1$ Do:
  For $\{ j \mid c_{ij} \neq 0 \}$ Do:
    Let *row* the j-th row of C^T
    For $k = nz_{C^T(j)}$ to 1 Do:
      Let $col = row(k)$
      If ($col \leq i$) *break*
      If (*Group(col)*$== -1$) *Count(col)* ++
  For $\{col \mid Count(col) \neq 0\}$ Do:
    If ($Count(col)^2 > \tau * nz_C(i) * nz_C(col)$) Then:
      Group(col)$== i$; Update size of *Group(i)*
    Count(col)=0

3 Modified Cosine Algorithm

In this section we modify the cosine algorithm to take into account the magnitude of the nonzero entries of a matrix. In some scenarios, for instance when there is a big difference between the magnitude of the entries, considering only the nonzero pattern for grouping rows can lead to an approximate block partition that may

not be a good representation of the significant part structure of the matrix. With "significant" we mean the block structure induced by the largest entries of the matrix. For instance, by applying the cosine algorithm to the next two matrices one may find as result the same block structure since both share the same adjacency matrix (which is in fact the matrix on the right).

$$\left[\begin{array}{cc|cc|cc} 1 & 1 & 0 & \epsilon & 0 & \epsilon & 0 \\ 1 & -1 & 1 & \epsilon & 0 & \epsilon & 0 \\ \epsilon & \epsilon & 1 & 1 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & -1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & -1 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 \end{array}\right], \quad \left[\begin{array}{ccc|cc|cc} 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 \end{array}\right].$$

But, for relative small values of ϵ, it is reasonably to think that the rows 3 and 4 have closer patterns than those corresponding to rows 2 and 3. In that case it could be preferred to identify the slightly different block structure

$$\left[\begin{array}{cc|ccc|cc} 1 & 1 & 0 & \epsilon & 0 & \epsilon & 0 \\ 1 & -1 & 1 & \epsilon & 0 & \epsilon & 0 \\ \hline \epsilon & \epsilon & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & -1 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 \end{array}\right].$$

To allow the cosine algorithm to find this structure some changes must be done. Figure 1 illustrates the inner product of row i with all the columns of C^T, i.e., $c_i C^T$.

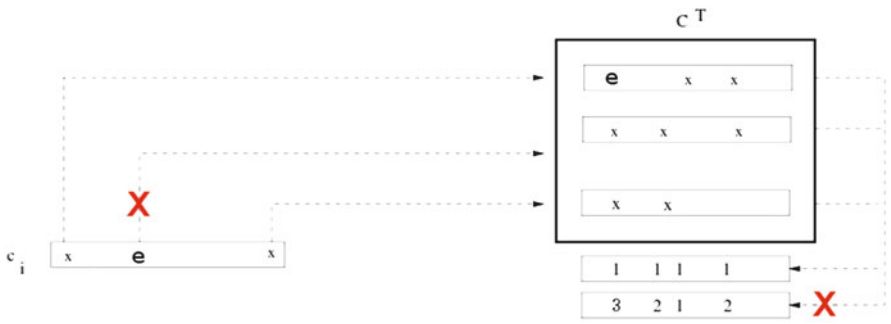


Fig. 1 Rowwise scheme for the inner product of row i with the columns of C^T. "X" indicates the differences in the application of the modified cosine algorithm with respect to the standard one

It is done rowwise and the computational cost is basically the sum of the number of nonzero entries of each row involved in this inner product. This example shows that taking into account the magnitude of the elements the result can be slightly different with respect to the sum obtained with the standard cosine. The differences come first, from the evaluation of which rows of C^T are involved. Then, the sum by columns may be also affected for the magnitude of the entries. Therefore, the corresponding inner product may be different and so the block partition for the matrix.

Algorithm 4 incorporates these modifications. In line 5 for each entry c_{ij} of the adjacency matrix C the magnitude of a_{ij} is checked. Row j of C^T is not considered for computing rowwise the product $c_i C^T$ if $|a_{ij}| < \epsilon$. Another modification is introduced in line 10 where the magnitude of the entries of a row of A^T is checked before adding its contribution to the sum by columns. Finally, we point out that to compute correctly the row norms used in line 12 to evaluate the cosine, the values of $nz_C(i)$ and $nz_C(col)$ must be decremented in the same quantity than the number of discarded entries in the corresponding rows.

Algorithm 4 *Modified Cosine Algorithm*
Input: matrix A, tolerance τ, tolerance ϵ; Output: Block partition.
Compute matrices A^T, C and C^T.
Set *Group(i)*$= -1$ and *Count(j)*$= 0$ for $i, j = 1, \ldots, n$
For $i = 1, \ldots, n$ if *Group(i)*$= -1$ Do:
For $\{j \mid |a_{ij}| > \epsilon\}$ Do:
Let *row* the j-th row of C^T
For $k = nz_{C^T(j)}$ to 1 Do:
Let *col* = *row(k)*
If (*col*$\leq i$) *break*
If ((*Group(col)*$== -1$) && ($|a^T_{j,\text{col}}| > \epsilon$)) *Count(col)*++
For $\{col \mid Count(col) \neq 0\}$ Do:
If ($Count(col)^2 > \tau * nz_C(i) * nz_C(col)$) Then:
Group(col)$== i$; Update the size of *Group(i)*
Count(col)=0

As additional notes we mention that, since the magnitude of the entries is needed, not only the adjacency graph of the matrix A but also its entries are needed as an input. Moreover, the matrix A^T must be computed internally. Using appropriate sparse storage schemes, as CSR (Compress Sparse Row), the adjacency matrices C and C^T are directly available without additional computational or memory cost. With respect to the value of ϵ we note that different choices for this parameter in lines 5 and 10 can be made. This observation makes a difference with respect to the possibility of sparsifying the matrix before the application of the standard cosine algorithm. Thus, the modified cosine is more flexible and opens more possibilities to the block partition computation.

4 Numerical Experiments

In this section the results of some numerical experiments conducted to determine the performance of the modified algorithm are presented. Table 1 shows the matrices used, its size (n), the number of nonzero entries (nnz) and the application field from they arise. The experiments have been done using MATLAB. The block version of the approximate inverse preconditioner AISM [6] has been used to precondition the BiCGSTAB iterative method. Table 2 shows, for different values of τ (minimum cosine value) and ϵ (tolerance to discard entries in the modified algorithm), the average block size obtained (γ), the preconditioner density (ρ) and the iteration count needed to reduce the initial residual by 10^{-8}. The parameters τ and ϵ have been chosen to get similar preconditioner densities for both algorithms.

From the results we observe an improvement on the convergence rate of the iterative method with a reduction in the number of iterations from 10 to 20 %. We also note that, in general, this improvement is obtained with a reduction on the preconditioner density and, therefore, a bigger reduction on the computational

Table 1 Used matrices (Available at http://math.nist.gov/MatrixMarket/)

Matrix	n	nnz	Application
HOR 131	434	4,710	Network flow
FS 541 4	541	4,285	Quemical kinetics
ORSIRR 1	1,030	6,858	Oil reservoir
UTM1700B	1,700	21,509	Plasma physics
UTM3060	3,060	42,211	Plasma physics
BCSSTK16	4,884	147,631	Structural engineering
ADD20	2,395	13,151	Computer component design
MEMPLUS	17,758	126,150	Computer component design
SAYLR4	3,564	22,316	Harwell-Boeing collection

Table 2 Results of the experiments

Matrix	*Cosine Alg.*				*Modified cosine Alg.*				
	γ	ρ	τ	*Its.*	γ	ρ	τ	ϵ	*Its.*
HOR 131	1.12	0.57	0.7	86	1.1	0.78	0.7	10^{-5}	79
FS 541 4	1.24	0.4	0.7	46	3.05	0.6	0.7	10^{-5}	42
ORSIRR 1	1.38	1.73	0.4	95	1.28	1.49	0.6	10^{-4}	81
UTM1700B	5.05	5.75	0.4	691	4.05	3.85	0.4	10^{-6}	626
UTM3060	6.02	4.89	0.4	1,337	4.53	4.44	0.4	10^{-4}	1,107
BCSSTK16	6.74	1.81	0.6	57	5.8	2.21	0.5	10^{-4}	52
ADD20	2.03	1.26	0.5	35	2.27	1.33	0.5	10^{-4}	31
MEMPLUS	11.8	7.79	0.2	86	10.25	4.92	0.2	10^{-6}	81
SAYLR4	4.74	3.15	0.2	160	2.59	2.29	0.2	10^{-3}	131

cost of the iterative solution process. We can conclude that it is possible to find a combination of the parameters τ and ϵ such that the performance of the iterative method can be significantly improved.

Conclusions
In this work we introduce a modification of the *cosine* algorithm to compress the graph of a sparse matrix that takes into account the magnitude of the nonzero entries of the rows. From the results of the numerical experiments one can deduce that with a good choice of the parameters ϵ and τ it is possible to reduce the number of iterations needed by the iterative method to get convergence.

Additional experiments to evaluate different choices for the parameter ϵ in different points inside the modified cosine algorithm will be done in the future.

Acknowledgements This work has been supported by The Spanish DGI grant MTM2010-18674.

References

1. Ashcraft, C.: Compressed graphs and the minimum degree algorithm. SIAM J. Sci. Comput. **16**, 1404–1411 (1995)
2. Barnard, S.T., Grote, M.J.: A block version of the SPAI preconditioner. In: Proceedings of the 9th SIAM Conference on Parallel Processing for Scientific Computing, San Antonio (1999)
3. Benzi, M., Kouhia, R., Tůma, M.: Stabilized and block approximate inverse preconditioners for problems in solid and structural mechanics. Comput. Methods Appl. Mech. Eng. **190**, 6533–6554 (2001)
4. Bridson, R., Tang, W.-P.: Refining an approximate inverse. J. Comput. Appl. Math. **123**, 293–306 (2000)
5. Chapman, A., Saad, Y., Wigton, L.: High-order ILU preconditioners for CFD problems. Int. J. Numer. Methods Fluids. **33**, 767–788 (2000)
6. Cerdán, J., Faraj, T., Malla, N., Marín, J., Mas, J.: Block approximate inverse preconditioners for sparse nonsymmetric linear systems. ETNA **37**, 23–40 (2010)
7. Chow, E., Saad, Y.: Approximate inverse techniques for block-partitioned matrices. SIAM J. Sci. Comput. **18**, 1657–1675 (1997)
8. Dongarra, J.J., Duff, I.S., Sorensen, D.C., van der Vorst, H.A.: Numerical Linear Algebra for High-Performance Computer. SIAM, Philadelphia (1998)
9. O'Neil, J., Szyld, B.D.: A block ordering method for sparse matrices. SIAM J. Sci. Comput. **11**, 811–823 (1990)
10. Saad, Y.: Iterative Methods for Sparse Linear Systems. PWS, Boston (1996)
11. Saad, Y.: Finding exact and approximate block structures for ILU preconditioning. SIAM J. Sci. Comput. **24**, 1107–1123 (2003)

Partially Implicit Runge-Kutta Methods for Wave-Like Equations

Isabel Cordero-Carrión and Pablo Cerdá-Durán

Abstract Runge-Kutta methods are used to integrate in time systems of differential equations. Implicit methods are designed to overcome numerical instabilities appearing during the evolution of a system of equations. We will present partially implicit Runge-Kutta methods for a particular structure of equations, generalization of a wave equation; the *partially implicit* term refers to this structure, where the implicit term appears only in a subset of the system of equations. These methods do not require any inversion of operators and the computational costs are similar to those of explicit Runge-Kutta methods. Partially implicit Runge-Kutta methods are derived up to third-order of convergence. We analyze their stability properties and show the practical applicability in several numerical examples.

1 Introduction

The evolution in time of many complex systems, governed by partial differential equations, implies, in a broad variety of cases, looking for the numerical solution of a system of ordinary differential equations. The most commonly used methods to integrate in time these systems are the well-known Runge-Kutta (RK) ones (see e.g. [4, 9] for a general review). Several classifications of the RK methods can be done, according to, e.g., their convergence order, the number of stages or their explicit/implicit structure.

Implicit methods are designed to overcome numerical instabilities appearing during the evolution of a system of equations. As an example, the so-called implicit-explicit RK (IMEX) methods have been used to evolve conservation laws with stiff terms or convection-diffusion-reaction equations (see, e.g., [1,2,12,13]). In our case,

I. Cordero-Carrión (✉)
Laboratoire Univers et Théories, CNRS/Laboratoire de Paris/Université Paris Diderot,
5 place Jules Janssen, F-92195 Meudon, France
e-mail: isabel.cordero@obspm.fr

P. Cerdá-Durán
Departamento de Astronomía y Astrofísica, Universidad de Valencia, C/Dr. Moliner, 50,
E-46100 Valencia, Spain
e-mail: pablo.cerda@uv.es

F. Casas, V. Martínez (eds.), *Advances in Differential Equations and Applications*,
SEMA SIMAI Springer Series 4, DOI 10.1007/978-3-319-06953-1_26

although we will not focus on equations with stiff source terms, a partially implicit treatment of the source terms will avoid the development of numerical instabilities in the numerical evolution of wave-like equations.

An implicit treatment offers a solution to get a stable evolution and involves, in general, an inversion of some operators. Depending on the complexity of the equations, the inversion can be even prohibitive in practice from a numerical point of view. We will focus on a particular structure of equations which does not require any analytical or numerical inversion. Therefore, these methods have a computational cost similar to the explicit Runge-Kutta methods (ERK).

2 Structure of the Equations

Let us consider the following system of PDEs,

$$\begin{cases} u_t = \mathscr{L}_1(u, v) \\ v_t = \mathscr{L}_2(u) + \mathscr{L}_3(u, v) \end{cases}, \tag{1}$$

being $\mathscr{L}_i, i = 1, 2, 3$, general non-linear differential operators. Let us denote by L_i their discrete operators. This particular structure is a generalization of a wave equation, written as a first-order system in time. $\mathscr{L}_1$ and $\mathscr{L}_3$ will be treated into an explicit way, whereas the $\mathscr{L}_2$ operator will be considered to contain the unstable terms and, therefore, treated implicitly. The *partially implicit* term refers to this structure, where the problematic term appears only in a subset of the system of equations.

Each stage of the derived partially implicit RK (PIRK) methods will proceed into two steps: (i) the variable u is evolved explicitly; (ii) the variable v is evolved taking into account the updated value of u for the evaluation of the $\mathscr{L}_2$ operator. The computational costs of the PIRK methods are comparable to those of the explicit ones. The resulting numerical schemes do not need any inversion of operators.

Numerical methods based on a nonlinear stability requirement are very desirable. Such methods are referred to as strong stability preserving (SSP) ones [8]. Given an evolution equation $\partial_t U = L(U)$, Gottlieb and Shu [7] proved that the classical second-order method,

$$U^{(0)} = U^n, \quad U^{(1)} = U^n + \Delta t\, L(U^n), \quad U^{n+1} = \frac{1}{2} U^n + \frac{1}{2} U^{(1)} + \frac{\Delta t}{2} L(U^{(1)}), \tag{2}$$

is the optimal second-order two-stage SSP ERK method, and that the third-order one due to Shu and Osher [14],

$$U^{(0)} = U^n,\ \ U^{(1)} = U^n + \Delta t\, L(U^n),\ \ U^{(2)} = \frac{3}{4}\, U^n + \frac{1}{4}\, U^{(1)} + \frac{\Delta t}{4}\, L(U^{(1)}),$$

$$U^{n+1} = \frac{1}{3}\, U^n + \frac{2}{3}\, U^{(2)} + \frac{2\Delta t}{3}\, L(U^{(2)}), \tag{3}$$

is the optimal third-order three-stage SSP ERK method. The optimal adjective refers, for a given number of stages, to a maximization of the corresponding Courant-Friedrichs-Lewy (CFL) value (1 in both cases). In the derivation of the PIRK methods, the previously described optimal SSP ERK methods are recovered when the $\mathscr{L}_2$ operator is neglected, i.e., when implicitly treated parts are not taken into account. The remaining coefficients associated to the $\mathscr{L}_2$ operator are chosen according to stability criteria. The PIRK methods will minimize the number of stages, two (three) for the second-order (third-order) method.

3 Numerical Methods and Stability Analysis

Let us denote by $(\bar{\alpha}_1 u, \bar{\alpha}_2 v)$, $\bar{\lambda} u$ and $(\bar{\gamma}_1 u, \bar{\gamma}_2 v)$ the associated linearized parts of the $\mathscr{L}_1$, $\mathscr{L}_2$ and $\mathscr{L}_3$ operators, respectively. The linearized system (1) is rewritten as

$$\begin{cases} u_t = \bar{\alpha}_1 u + \bar{\alpha}_2 v, \\ v_t = \bar{\gamma}_1 u + \bar{\gamma}_2 v + \bar{\lambda} u. \end{cases} \tag{4}$$

Let us denote $\alpha_i := \bar{\alpha}_i\, \Delta t$, $\lambda := \bar{\lambda}\, \Delta t$ and $\gamma_i := \bar{\gamma}_i\, \Delta t$. We assume that $|\omega_i| \leq 1$, where ω_i, $i = 1, 2$, denote the two eigenvalues of the following matrix

$$\begin{pmatrix} 1 + \alpha_1 & \alpha_2 \\ \gamma_1 & 1 + \gamma_2 \end{pmatrix}, \tag{5}$$

which represents the explicit terms of the system. We are going to focus here in the linear stability of the system; the analysis of the linear stability is the most simple case regarding the study of the stability of the system of equations, but if a method does not verify even this criteria it is obviously not stable in general. In most cases, the linear part of the operators is the dominant one and the results obtained in the analysis of the linear stability are reproduced in the numerical simulations. Previous matrix determinant, dex, and trace, trex, are bounded by $|\text{dex}| \leq 1$ and $|\text{trex}| \leq 2$. Let us denote M_i the matrix which updates values for a ith-order method,

$$\begin{pmatrix} u^{n+1} \\ v^{n+1} \end{pmatrix} = M_i \begin{pmatrix} u^n \\ v^n \end{pmatrix}. \tag{6}$$

Stability thus requires that the absolute value of the two eigenvalues associated to the matrix M_i are bounded by 1. However, in order to simplify the derivation of the PIRK methods, we are going to relax this condition on the eigenvalues of the

matrix M_i by a bound on its determinant, $|\det(M_i)| \leq 1$. The restriction onto the eigenvalues will be shown in the numerical experiments as the boundaries of the stability region. $\mathrm{Re}(\lambda\alpha_2) \leq 0$ is also assumed; this condition is satisfied for general wave-like equations written as a first-order system in time (see numerical example).

3.1 First-Order Method

The one-stage first-order method for the system (1) can be written in terms of one coefficient, c_1, as follows:

$$\begin{cases} u^{n+1} = u^n + \Delta t\, L_1(u^n, v^n), \\ v^{n+1} = v^n + \Delta t\, [(1-c_1)\, L_2(u^n) + c_1\, L_2(u^{n+1}) + L_3(u^n, v^n)]. \end{cases} \tag{7}$$

This method is a particular case for the system (1) of the IMEX-θ method (see, e.g., [10]). The matrix M_1 satisfies $\det(M_1) = \mathrm{dex} - \lambda\,\alpha_2\,(1-c_1)$. $c_1 = 1$ guarantees $|\det(M_1)| \leq 1, \forall(\lambda\,\alpha_2)$.

3.2 Second-Order Method

The two-stages second-order method for the system (1), imposing SSP optimal two-stages second-order method for the pure explicit parts, can be written in terms of two coefficients, c_1 and c_2, as follows:

$$\begin{cases} u^{(1)} = u^n + \Delta t\, L_1(u^n, v^n), \\ v^{(1)} = v^n + \Delta t\, [(1-c_1)\, L_2(u^n) + c_1\, L_2(u^{(1)}) + L_3(u^n, v^n)]. \end{cases} \tag{8}$$

$$\begin{cases} u^{n+1} = \dfrac{1}{2}[u^n + u^{(1)} + \Delta t\, L_1(u^{(1)}, v^{(1)})], \\ v^{n+1} = v^n + \dfrac{\Delta t}{2}\, [L_2(u^n) + 2c_2\, L_2(u^{(1)}) + (1-2c_2)\, L_2(u^{n+1}) \\ \qquad\qquad + L_3(u^n, v^n) + L_3(u^{(1)}, v^{(1)})]. \end{cases} \tag{9}$$

Matrix M_2 satisfies $\det(M_2) = \frac{1}{4}[(1-\mathrm{dex})^2 + \mathrm{trex}^2 + \lambda\,\alpha_2\,(1-\mathrm{dex})\,(1-2c_1 + 2c_2)]$. $|\det(M_2)| \leq 1$ cannot be guarantee $\forall(\lambda\,\alpha_2)$. We restrict to real numbers and consider the determinant of M_2 as a polynomial in $(\lambda\,\alpha_2)$; the extrema values of its coefficients can be analyzed. For $|\lambda\,\alpha_2| \ll 1$, the resulting optimal values for the coefficients are $c_1 = 1/2$ and $c_2 = 0$; we will denote this method by PIRK2a. For $|\lambda\,\alpha_2| \gg 1$, the optimal values for the coefficients are $c_1 = 1 - \sqrt{2}/2$ and $c_2 = (\sqrt{2}-1)/2$; we will denote this method by PIRK2b. If $|\lambda\,\alpha_2|$ is not too big, the choice $(c_1, c_2) = (1/2, 0)$ is convenient since it avoids to compute the term $L_2(u^{(1)})$ to obtain v^{n+1} in the final stage. Otherwise, the PIRK2b method is better.

3.3 Third-Order Method

The three-stages third-order method for the system (1), imposing SSP optimal three-stages third-order method for the pure explicit parts, can be written in terms of two coefficients, c_1 and c_2, as follows:

$$\begin{cases} u^{(1)} = u^n + \Delta t\, L_1(u^n, v^n), \\ v^{(1)} = v^n + \Delta t\, [(1-c_1)\, L_2(u^n) + c_1\, L_2(u^{(1)}) + L_3(u^n, v^n)]. \end{cases} \tag{10}$$

$$\begin{cases} u^{(2)} = \dfrac{1}{4}[3u^n + u^{(1)} + \Delta t\, L_1(u^{(1)}, v^{(1)})], \\ v^{(2)} = v^n + \dfrac{\Delta t}{4}\, [2(c_1 + 2c_2)\, L_2(u^n) + 4c_2\, L_2(u^{(1)}) + 2(1 - c_1 - 4c_2)\, L_2(u^{(2)}) \\ \qquad\qquad +L_3(u^n, v^n) + L_3(u^{(1)}, v^{(1)})]. \end{cases} \tag{11}$$

$$\begin{cases} u^{n+1} = \dfrac{1}{3}[u^n + 2u^{(2)} + 2\Delta t\, L_1(u^{(2)}, v^{(2)})], \\ v^{n+1} = v^n + \dfrac{\Delta t}{6}\, [L_2(u^n) + L_2(u^{(1)}) + 4L_2(u^{(2)}) \\ \qquad\qquad +L_3(u^n, v^n) + L_3(u^{(1)}, v^{(1)}) + 4L_3(u^{(2)}, v^{(2)})]. \end{cases} \tag{12}$$

Matrix M_3 satisfies

$$\begin{aligned} \det(M_3) = {} & \frac{1}{36}[14 + 2(\mathrm{trex} - 1)^3 + (\mathrm{dex} - 2)^3 + 6\,\mathrm{trex}^2 + 3\,\mathrm{dex}\,((\mathrm{trex} - 1)^2 - 2)] \\ & + \frac{1}{24}\lambda\,\alpha_2\,(-1 + c_1 - 4c_2)[(\mathrm{dex} - 2)^2 + (\mathrm{trex} - 1)^2 - 2] \\ & + \frac{1}{12}\lambda^2\,\alpha_2^2\,[c_1 - 4c_2 + (\mathrm{dex} - 1)(4c_2 - c_1^2 - 4c_1c_2)] \\ & - \frac{1}{72}\lambda^3\,\alpha_2^3\,[-1 + 3(1 - 2c_1)(c_1 + 4c_2)]. \end{aligned} \tag{13}$$

$|\det(M_3)| \leq 1$ cannot be guarantee $\forall(\lambda\,\alpha_2)$. We proceed as in the second-order method. For $|\lambda\,\alpha_2| \ll 1$, the resulting optimal values for the coefficients are $(c_1, c_2) = (1/4, 1/16)$; we will denote this method by PIRK3a. For $|\lambda\,\alpha_2| \gg 1$, the resulting optimal values for the coefficients are $(c_1, c_2) = ((3 - \sqrt{3})/6, (-1 + \sqrt{3})/8)$; we will denote this method by PIRK3b.

4 Numerical Experiments

In this section we show two examples of the application of PIRK methods to ODEs and PDEs, demostrating that the stability properties of the method hold in practice.

4.1 System of ODEs

Let us consider a system of ODEs of the following form:

$$u_t = c\,u + d\,v, \quad v_t = a\,u + b\,v, \tag{14}$$

where a, b, c and d are real constants. This system is interesting because it coincides with the linear part of the system of Eqs. (4) considered for our stability analysis, with $\bar{\alpha}_1 = c$, $\bar{\alpha}_2 = d$, $\bar{\gamma}_1 = 0$, $\bar{\gamma}_2 = b$ and $\bar{\lambda} = a$.

In the case $(b-c)^2 + 4\,a\,d < 0$ and $b + c \leq 0$, this system of equations has damped oscillatory solutions of the form,

$$u = \frac{\sqrt{-ad}}{a} v_0 \cos(\omega t + \phi) e^{\sigma t}, \quad v = v_0 \cos(\omega t) e^{\sigma t}, \tag{15}$$

being v_0, $\omega \equiv \frac{1}{2}\sqrt{-4\,a\,d - (b-c)^2}$, $\sigma \equiv \frac{b+c}{2}$ and $\tan\phi \equiv \frac{\omega}{\sigma - b}$ a constant set by the initial conditions, the frequency, decay rate and relative phase between u and v of the solution, respectively. This system corresponds to (1), with $\mathscr{L}_1(u,v) = u + v$, $\mathscr{L}_2(u) = a\,u$ and $\mathscr{L}_3(u,v) = b\,v$, and fulfills the applicability requirements of the PIRK methods, i.e. $\bar{\alpha}_2\,\bar{\lambda} < 0$, $|\mathrm{dex}| \leq 1$ and $|\mathrm{trex}| \leq 2$.

For our numerical experiment we will consider the case $\omega = 1$ and $a = -d$, without loss of generality, since it is equivalent to a rescaling of t and v. The remaining coefficients depend only on the values of σ and ϕ. We have performed numerical simulations for $\sigma = 0, -0.01, -0.1, -1$, and $\phi/\pi = 1/2, 1/3, 1/4, 1/10$, which are representative of all possible solutions of this set of equations.

Figure 1 shows the results for a representative test, comparing the first-order ERK with the PIRK. To estimate the relative error of the method we compute the time-averaged L_2-norm of the difference between the analytic and the numerical solution

$$L_2(u)(t) = \frac{1}{t}\sqrt{\sum_{t_n < t} [u_{\mathrm{num}}(t_n) - u_{\mathrm{ana}}(t_n)]^2 \Delta t^2 e^{-2\sigma t_n}}\ . \tag{16}$$

For this test the ERK is unconditionally unstable (see left panel) and decreasing the time step leads to an exponentially increasing amplitude, provided the integration time is sufficiently long. By comparison, the first-order PIRK is stable for $\Delta t < 2$, since $|u| \lesssim 1$. For longer time steps (e.g. $\Delta t = 0.1$) using the PIRK, the solution losses accuracy (in this case a phase shift) but it is still bounded (even at $t = 1{,}000$), and hence the numerical method is stable. We use the value of the time-averaged L_2-norm at time $t = 100$ as a measure of the stability of a numerical method, for a particular numerical test with a given time step. Values <1 (>1) usually indicate stability (instability). In Fig. 2 we compare the stability properties of ERK and PIRK methods observed in our numerical experiments. In all cases, the PIRK methods are superior to the ERK methods, as they can achieve

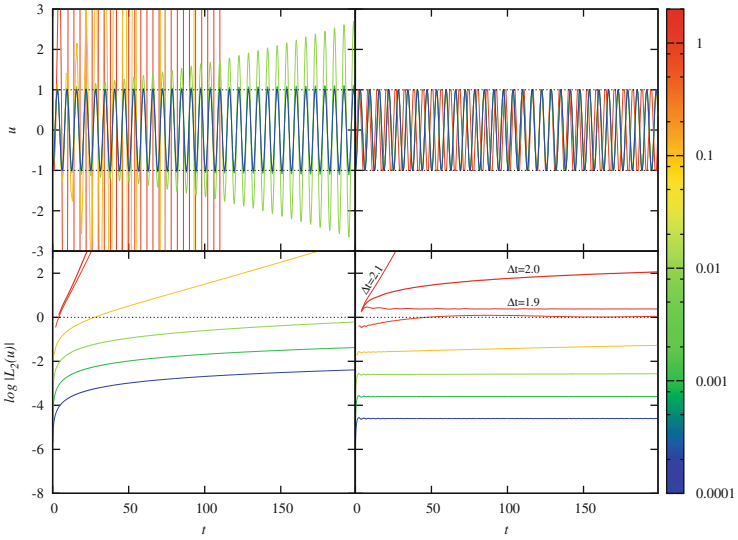


Fig. 1 Numerical integration of the previous ODEs with $\sigma = \phi = 0$, using a first-order ERK (*left panels*) and a first-order PIRK (*right panels*). *Upper panels* show the time evolution of u, $10^{-4} \leq \Delta t \leq 1$. *Dotted lines* are the amplitude of the oscillatory analytic solution. *Lower panels* show the time averaged L_2-norm of the difference between the numerical and analytical solutions, $10^{-4} \leq \Delta t \leq 2.1$

stable numerical evolutions with significantly longer time steps. For small time steps, all numerical methods follow the expected order of convergence. For first and second-order methods, ERK methods are unconditionally unstable; despite L_2-norm< 1 for small values of Δt, longer evolutions always lead to exponentially growing amplitudes in all studied cases. In contrast, first and second-order PIRKs are numerically stable in all simulations tested (up to $t = 1{,}000$), and only become unstable for Δt larger than a certain threshold. For the third-order methods, all the schemes are stable for small Δt, but the ERK becomes unstable at lower values of Δt than PIRK methods, which behave similar to the tested IMEX scheme.

A change of the value of σ, fixed $\phi = 0$, introduces a damping in the oscillatory solution, in a timescale of $1/\sigma$. As the parameters approach $|\sigma\omega| \sim 1$, the system becomes stiff, and the maximum time step providing stable evolutions decreases as expected. In the case of third-order methods (see upper panel of Fig. 3), and similarly for first and second-order ones, as we approach $\sigma = -1$, both ERK and PIRK methods behave almost identically. Despite of being *partially implicit*, the terms in Eq. (14) responsible for the stiffness cannot be included in the $\mathscr{L}_2$ operator, and both ERK and PIRK methods suffer from this stiffness.

In the case of varying ϕ, fixed $\sigma = 0$, all ERK schemes behave in an identical way (see lower-right panel of Fig. 3 for third-order schemes; first and second-order ones behave similarly). However, PIRK methods suffer from a significant reduction of the maximum time-step as $\phi \approx 0$ (see lower-middle and right panels of Fig. 3 for

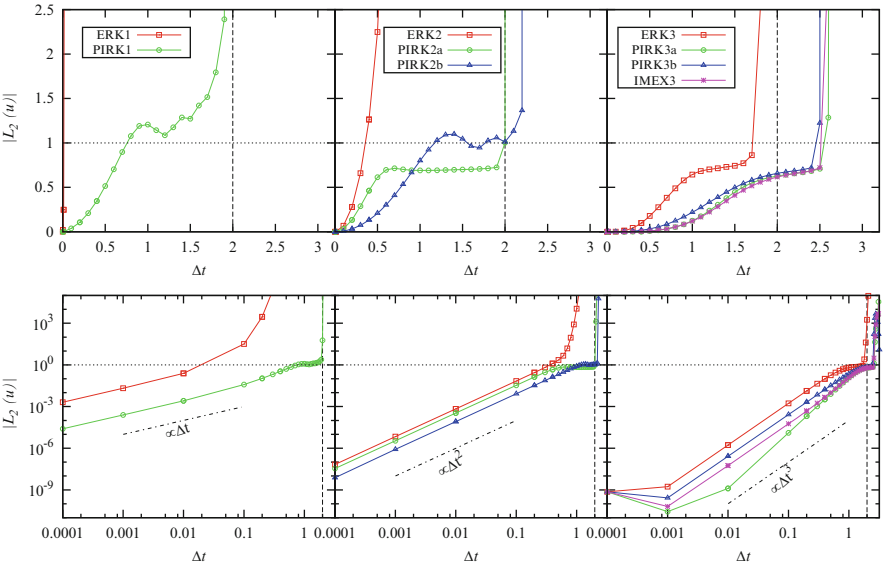


Fig. 2 Numerical error integrating a system of ODEs with $\sigma = 0$ and $\phi = 0$, using first (*left panels*), second (*middle panels*) and third (*right panels*) order methods. *Upper panels* show the transition between stable ($L_2 \ll 1$) and unstable ($L_2 \gg 1$) numerical evolutions. *Lower panels*, in logarithmic scale, show the behavior for small time steps, compared to the expected scaling for each method (*dashed-dotted lines*). As a reference, *vertical dashed line* at $\Delta t = 2$ corresponds to the maximum time step for the first-order PIRK to be stable

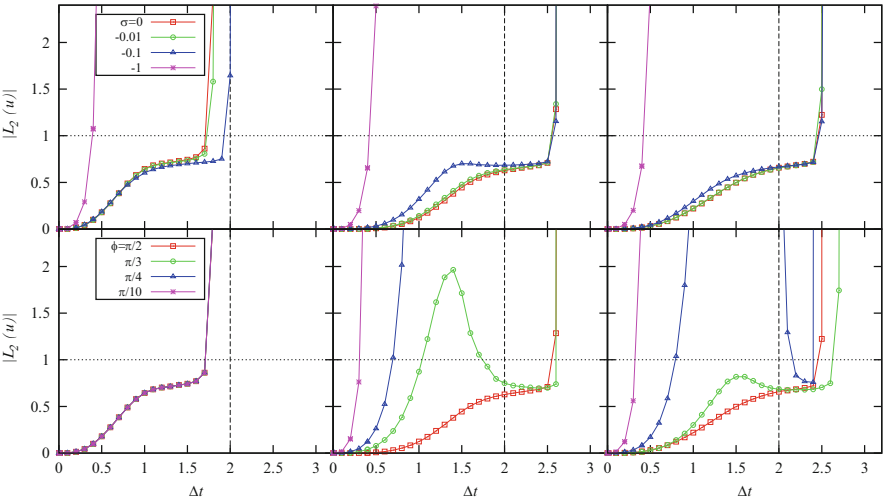


Fig. 3 Behavior of error for third-order schemes varying the value of σ (*upper panels*, $\phi = \pi/2$) and ϕ (*lower panels*, $\sigma = 0$). From *left* to *right*, third-order ERK, PIRK3a and PIRK3b

third-order schemes; first and second-order ones behave similarly). This is the only case in which ERK methods are superior to PIRK methods. Therefore, the class of systems for which PIRK methods are a good alternative to classical ERK methods are wave-like equations, in which the condition $\phi \approx \pi/2$ is fulfilled.

4.2 Wave Equation in Spherical Coordinates

In this section, the PIRK methods are applied to the case of the time evolution of a wave equation for a scalar, h, in spherical coordinates. The evolution equation for h can be written as $\partial_{tt} h = \triangle h$, where $\triangle$ denotes the Laplacian operator. This equation can be rewritten as a first-order system in time, with the addition of an extra auxiliary variable, A, as follows: $\partial_t h = A, \quad \partial_t A = \triangle h$. In this case, according to system (1), the variables can be identified as $(u, v) = (h, A)$, and the operators as $\mathscr{L}_1(h, A) = A$, $\mathscr{L}_2(h) = \triangle h$ and $\mathscr{L}_3(h, A) = 0$. Spherical coordinates are used. This equation has solutions of the form $h(r, \theta, \varphi, t) \sim j_l(kr)\, Y_{lm}(\theta, \varphi) \cos kt$, being j_l the spherical Bessel function of first kind of order l and Y_{lm} the spherical harmonics. The value of $k \in \mathbb{R}^+$ is determined by imposing boundary conditions. We search for solutions inside a sphere of radius unity imposing $h(r = 1, \theta, \varphi, t) = 0$. We have performed 1D, 2D and 3D simulations using as initial data solutions with $n = 1$ at $t = 0$. We use $(l, m) = (0, 0), (2, 0), (2, 2)$ for the 1D, 2D and 3D cases, respectively. We use a finite difference scheme and an equally-spaced grid with n_r, n_θ and n_φ grid points in the coordinate directions. At $r = 1$ the analytical solution is imposed as boundary condition. L_2-norm is used as a measure of the global absolute error,

$$L_2(h)(t) = \frac{1}{n_r n_\theta n_\varphi} \sqrt{\sum_{r,\theta,\varphi} [h_{\text{num}}(r, \theta, \varphi, t) - h_{\text{ana}}(r, \theta, \varphi, t)]^2 (k\, r)^2}. \tag{17}$$

We will analyze the numerical stability of the derived PIRK methods using $(n, l, m) = (1, 2, 0)$ for the initial data in 2D simulations with equatorial symmetry, $(n_r, n_\theta) = (100, 32)$ grid points and a fourth-order spatial discretization scheme (see more details in [6]). Let us denote CFL factor $= \frac{\Delta t}{\Delta l_{\min}} = \frac{\Delta t}{\Delta t_{\max}}$.

We study stability properties of the numerical solution depending on the coefficients of the methods and the time step Δt. The bound for the determinant is a necessary but not sufficient condition; the boundaries of the stability region correspond to the bounds for the eigenvalues. For the first-order PIRK method, the estimated optimal value of the coefficient, $c_1 = 1$, lays inside the stability region and is indeed the value such that the maximum CFL factor is achievable, as it can be checked in Fig. 4. The ERK method corresponds to $c_1 = 0$, and is always unstable.

We have studied the numerical stability of the second-order PIRK method. Figure 5 shows the stability region on the (c_1, c_2) plane, for $c_1, c_2 \in [-0.5, 1.5]$ and several CFL factors (0.5, 0.7, 0.8 and 0.9). The boundaries agree with the bounds

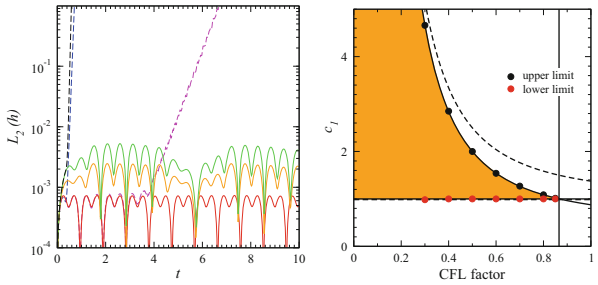


Fig. 4 Stability of the first-order PIRK method. *Left panel*: time evolution of the L_2-norm for simulations with CFL factor 0.5 and c_1 values of 0.9 (*blue*), 0.99 (*magenta*), 1 (*red*), 1.5 (*orange*), 2 (*green*) and 2.05 (*black*). *Solid* and *dashed lines* represent numerically stable and unstable simulations, respectively. *Right panel*: stability region depending on the values for c_1 and the CFL factor. *Solid lines* are the boundaries of the stability region (*orange area*). The boundary of the region $|\det(M_1)| \leq 1$ is also plotted (*dashed lines*)

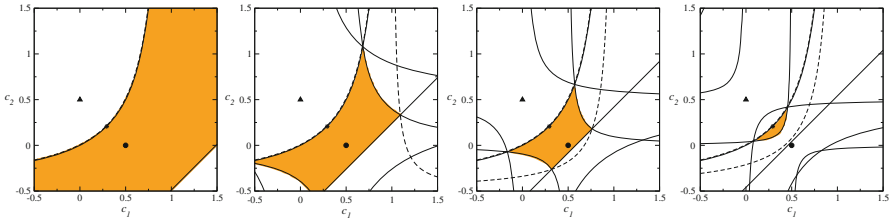


Fig. 5 Dependence of the numerically determined stability region (*orange area*) on the (c_1, c_2) coefficients using a second-order PIRK method for several CFL factors. Boundaries (*solid lines*) agree with the condition for the eigenvalues. The boundaries for the condition for the determinant (*dashed lines*), the optimal values for the coefficients, $(c_1, c_2) = (1/2, 0)$ (*black circle*) and $(c_1, c_2) = 1/2(2 - \sqrt{2}, \sqrt{2} - 1)$ (*star symbol*), and the ones corresponding to the second-order ERK method (*black triangle*) are also plotted

for the eigenvalues, and the condition for the determinant overestimates this region. The optimal values corresponding to the PIRK2b and PIRK2a methods lie in the stability region almost for all the cases and all the cases, respectively, as it can be checked in Fig. 5. The ERK method corresponds to $(c_1, c_2) = (0, 1/2)$ and is always unstable.

The same numerical stability analysis have been carried out for the third-order PIRK method, shown in Fig. 6 for several CFL factors. The boundaries of the stability region can be obtained in the same way as in the second-order method, the condition for the determinant being less restrictive. The optimal values of the coefficients lay inside the stability region for all CFL factors analyzed. For the coefficients corresponding to the third-order ERK method, stability is achieved if the CFL factor <0.751 (see Fig. 6).

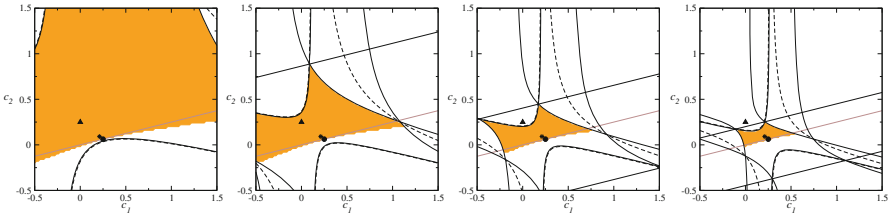


Fig. 6 Dependence of the numerically determined stability region (*orange area*) on the (c_1, c_2) coefficients using a third-order PIRK method for several CFL factors. Boundaries (*solid lines*) agree with the condition for the eigenvalues. The boundaries for the condition for the determinant (*dashed lines*), the optimal values for the coefficients, $(c_1, c_2) = (1/4, 1/16)$ (*black circle*) and $(c_1, c_2) = ((3-\sqrt{3})/6, (\sqrt{3}-1)/8)$ (*star symbol*), and the ones corresponding to the third-order ERK method (*black triangle*) are also plotted

We have studied the convergence of the PIRK methods by performing series of 1D, 2D and 3D simulations, with resolutions $n_r = 50$, $(n_\theta, n_\varphi) = (50, 16)$ and $(n_r, n_\theta, n_\varphi) = (50, 8, 32)$, respectively. We use CFL=0.8. The L_2-norm is used as an estimation of the error. Independently of the dimensionality of the simulation, the error falls with decreasing time step as expected from the convergence order of the PIRK method used.

Conclusions

PIRK methods, from first to third-order of convergence, have been derived to evolve in time wave-like systems of non-linear partial differential equations. Optimal SSP ERK methods are recovered when implicitly treated parts are neglected. No inversion is required and the computational costs of the PIRK methods are comparable to those of the ERK ones. The PIRK methods are stable for wave-like equations and larger time steps can be achieved. In contrast, first and second-order ERK methods result to be unconditionally unstable; third-order ERK method is stable, but the largest time step achievable is lower. PIRK methods are appropriate to evolve generalized complex wave equations in spherical coordinates, as it has been shown in [3, 5, 11] for the evolution of Einstein equations.

Acknowledgements This work has been funded by the SN2NS project ANR-10-BLAN-0503, the Spanish MICINN (AYA 2010-21097-C03-01), the Generalitat Valenciana (PROMETEO-2009-103 and PROMETEO-2011-083) and the ERC Starting Grant CAMAP-259276.

References

1. Asher, U.M., Ruuth, S.J., Wetton, B.T.R.: Implicit-explicit methods for time-dependent PDE's. SIAM J. Numer. Anal. **32**, 797–823 (1995)
2. Asher, U.M., Ruuth, S.J., Spiteri, R.J.: Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations. Appl. Numer. Math. **25**, 151–167 (1997)
3. Baumgarte, T.W., Montero, P., Cordero-Carrión, I., Müller, E.: Numerical relativity in spherical polar coordinates: evolution calculations with the BSSN formulation. Phys. Rev. D **87**, 044026 (2013)
4. Butcher, J.C.: Numerical Methods for Ordinary Differential Equations, 2nd edn. Wiley, Chichester (2008)
5. Cordero-Carrión, I., Cerdá-Durán, P., Ibáñez, J.M.: Gravitational waves in dynamical spacetimes with matter content in the fully constrained formulation. Phys. Rev. D **85**, 044023 (2012)
6. Cordero-Carrión, I., Cerdá-Durán, P.: Partially implicit Runge-Kutta methods for wave-like equations in spherical-type coordinates (2012, Preprint). arXiv:1211.5930
7. Gottlieb, S., Shu, C.-W.: Total variation diminishing Runge-Kutta schemes. Math. Comput. **67**, 73–85 (1998)
8. Gottlieb, S., Shu, C.-W., Tadmor, E.: Strong-stability-preserving high order time discretization methods. SIAM Rev. **43**, 89–112 (2001)
9. Hairer, E., Nørsett, S.P., Wanner, G.: Solving Ordinary Differential Equations I: Nonstiff Problems. Springer, Berlin (1987)
10. Hundsdorfer, W., Verwer, J.G.: Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations. Springer, Berlin (2003)
11. Montero, P., Cordero-Carrión, I.: BSSN equations in spherical coordinates without regularization: vacuum and nonvacuum spherically symmetric spacetimes. Phys. Rev. D **85**, 124037 (2012)
12. Pareshi, L.: Central differencing based numerical schemes for hyperbolic conservation laws with relaxation terms. SIAM J. Num. Anal. **39**, 1395–1417 (2001)
13. Pareshi, L., Russo, G.: Implicit-explicit Runge-Kutta methods and application to hyperbolic systems with relaxation. J. Sci. Comput. **25**, 129–155 (2005)
14. Shu, C.-W., Osher, S.: Efficient implementation of essentially non-oscillatory shock-capturing schemes. J. Comput. Phys. **77**, 439–471 (1988)

Operator-Splitting on Hyperbolic Balance Laws

Pedro González de Alaiza Martínez and María Elena Vázquez-Cendón

Abstract Operator-Splitting Methods or Fractional-Step Methods are based on the fact that hyperbolic balance laws can be split exactly into a homogeneous hyperbolic partial differential equation (PDE, advection) and an ordinary differential equation (EDO, evolution); which means that both advecting first and evolving next and evolving first and advecting next are equivalent to solve the whole problem directly [3, 4, 10, 11]. The key to this method is the physical flux which must be used in the advective part. If the problem is linear, it coincides with the physical flux of the problem and does not depend on whether the advection is solved before or after the evolution. However, this is no longer true for nonlinear problems: it is different from the flux of the problem and depends on the order [8]. In this work we will begin with the analysis of the splitting of multi-dimensional linear systems and we will end up explaining how exact nonlinear splitting can be obtained for one-dimensional scalar equations.

1 Introduction

For simplicity without loss of generality, we will introduce the idea of the splitting for scalar multidimensional hyperbolic equations. Let us consider the following initial value problem (IVP) in N dimensions:

$$\begin{cases} \dfrac{\partial w}{\partial t} + \displaystyle\sum_{i=1}^{N} \frac{\partial}{\partial x_i} f_i(w) = s(x_1, \ldots, x_N, t, w), \\ \\ w(x_1, \ldots, x_N, 0) = w_0(x_1, \ldots, x_N), \end{cases} \tag{1}$$

where $w = w(x_1, \ldots, x_N, t)$ is the conserved variable, $f_i(w)$ is the i-th component of the physical flux, $\lambda_i(w) := f_i'(w)$ is the i-th component of the wave-propagation

P. González de Alaiza Martínez (✉) • M.E. Vázquez-Cendón
Faculty of Mathematics, University of Santiago de Compostela, Lope Gómez de Marzoa s/n, Campus sur, 15782 Santiago de Compostela, Spain
e-mail: pgalaiza@hotmail.com; elena.vazquez.cendon@usc.es

F. Casas, V. Martínez (eds.), *Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 4, DOI 10.1007/978-3-319-06953-1_27

speed, $s(x_1, \ldots, x_N, t, w)$ is the source term and $w_0(x_1, \ldots, x_N)$ is the initial condition. This IVP has also been studied at references [2, 6, 9, 12, 15, 16].

The wave-progation speed defines the so-called characteristics curves $\boldsymbol{x}(t) = (x_1(t), \ldots, x_N(t))$ given by:

$$\begin{cases} \dfrac{d}{dt} x_i(t) = \lambda_i(w(\boldsymbol{x}(t), t)), \\ \\ x_i(0) = \xi_i, \end{cases} \qquad i = 1, \ldots, N \tag{2}$$

where $\boldsymbol{\xi} := \boldsymbol{x}(0)$ represents the foot of the characteristic. Along these curves, the evolution of the conserved variable is given by the following ODE:

$$\frac{d}{dt} w(\boldsymbol{x}(t), t) = s\left(\boldsymbol{x}(t), t, w(\boldsymbol{x}(t), t)\right). \tag{3}$$

Coupling the Eqs. (2) and (3) together, we obtain the exact solution of the IVP given by (1). The operator-splitting technique applied to hyperbolic balance laws is based on this fact. It consists of splitting the IVP into an *evolutive* part (ODE):

$$\frac{dw}{dt} = s(x_1, \ldots, x_N, t, w), \tag{4}$$

and an *advective* part (homogeneous PDE):

$$\frac{\partial w}{\partial t} + \sum_{i=1}^{N} \frac{\partial}{\partial x_i} \hat{f}_i(w) = 0, \tag{5}$$

where $\hat{\boldsymbol{f}}$ is a certain physical flux.

Depending on whether the advective part is solved in the first or second place, we define respectively two splittings. The first one will be called *Advection-Evolution Splitting* (AES) and it has the following structure of two steps:

$$\begin{aligned} &\textit{(advective step)} \begin{cases} \dfrac{\partial \hat{w}}{\partial t} + \displaystyle\sum_{i=1}^{N} \frac{\partial}{\partial x_i} \hat{f}_i^{AES}(\hat{w}) = 0, \\ \\ \hat{w}(x_1, \ldots, x_N, 0) = w_0(x_1, \ldots, x_N), \end{cases} \\ &\textit{(evolutive step)} \begin{cases} \dfrac{dw}{dt} = s(x_1, \ldots, x_N, t, w), \\ \\ w(x_1, \ldots, x_N, 0) = \hat{w}(x_1, \ldots, x_N, t_{end}), \end{cases} \end{aligned} \tag{6}$$

where t_{end} is the final instant at which $w(x,t)$ is calculated and $\hat{f}^{AES}$ is the physical flux of AES. The second one, the *Evolution-Advection Splitting* (EAS), has this analogous structure:

$$\begin{aligned} &(evolutive\ step) \begin{cases} \dfrac{d\hat{w}}{dt} = s(x_1,\ldots,x_N,t,\hat{w}), \\ \hat{w}(x_1,\ldots,x_N,0) = w_0(x_1,\ldots,x_N), \end{cases} \\ &(advective\ step) \begin{cases} \dfrac{\partial w}{\partial t} + \displaystyle\sum_{i=1}^{N} \frac{\partial}{\partial x_i} \hat{f}_i^{EAS}(w) = 0, \\ w(x_1,\ldots,x_N,0) = \hat{w}(x_1,\ldots,x_N,t_{end}), \end{cases} \end{aligned} \tag{7}$$

where $\hat{f}^{EAS}$ is the physical flux of EAS.

The graphical interpretation of both splittings (6) and (7) is shown in Fig. 1. It depicts, along a characteristic curve and in the x–w plane, the idea of the splitting for a one-dimensional IVP when we want to calculate the solution at $t = t_{end}$ from the values at $t = 0$. We can calculate the solution of the IVP directly by solving (2) and (3); by doing this, the values of the position x and the conserved variable w will be coupled together in time in a way depending on the flux and the source. Instead of solving the IVP in this way, which can be very complicated, we can apply the operator splitting to it. By doing this, we calculate the position and conserved variable separately: during the advection only the value of the position changes and during the evolution only the value of the conserved variable.

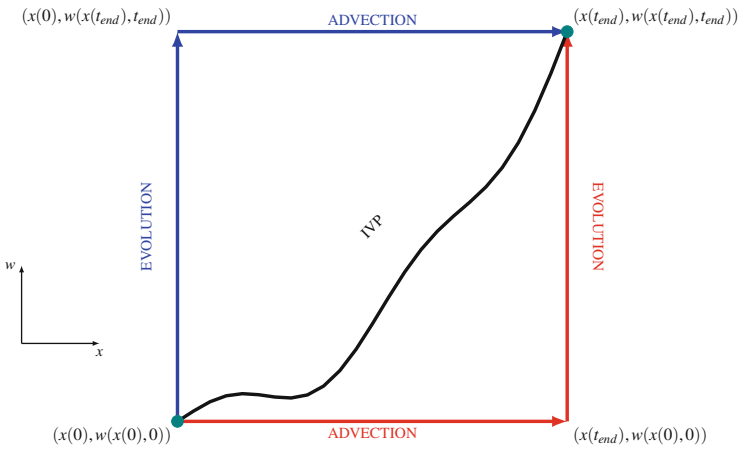


Fig. 1 Graphical interpretation of the equivalence between the two splittings (AES in *red*, EAS in *blue*) and the IVP (in *black*), along a characteristic curve and in one dimension

If the flux of the IVP given by (1) is linear it is well known that the problems (1), (6) and (7) are equivalent with the same fisical flux f in the advective parts of the splittings, that is:

Proposition 1 *Being f linear, the problems (1), (6) and (7) are equivalent to each other if $f = \hat{f}^{AES} = \hat{f}^{EAS}$.*

In Sect. 2 we will see that this equivalence is also true at discrete level.

However, if the flux of the IVP is nonlinear we have to modify the flux of the advective parts of AES and EAS in order to have equivalent splittings. In Sect. 3 we will deal with the exact splitting of scalar one-dimensional nonlinear equations.

2 Numerical Methods for Linear Systems

Proposition 1 establishes that, at continous level, any IVP defined by a linear hyperbolic system with source term is equivalent to the corresponding AES and EAS using its same flux. At discrete level, we will construct numerical schemes based on AES and EAS by means of a combination of a certain discrete advective operator (A) and a certain discrete evolutive operator (S): $\mathbf{W}^{n+1} = S^{(\Delta t)}A^{(\Delta t)}\mathbf{W}^n$ and $\mathbf{W}^{n+1} = A^{(\Delta t)}S^{(\Delta t)}\mathbf{W}^n$, respectively. If we add more steps and the operators are accurate enough, we could obtain high order schemes (Strang [13]); moreover, a dimensional splitting could be applied to the advection (Toro [14]).

A very interesting result is that, if we permute the order of the discrete operators, we get the same numerical scheme (maybe except for high-order terms); which means that AES and EAS are also equivalent between one to another at discrete level. We can illustrate this fact with the well-known scalar one-dimensional linear advection equation $w_t + \lambda w_x = s$. In a regular mesh (Δx, Δt) and for $\lambda > 0$, we will use Godunov's method for advecting and explicit Euler's method for evolving to obtain first-order accurate schemes. The first scheme is obtained from the IVP:

$$w_i^{n+1} = w_i^n - \mu\left(\phi_{i+\frac{1}{2}}^{n+\frac{1}{2}} - \phi_{i-\frac{1}{2}}^{n+\frac{1}{2}}\right) + \Delta t\left[\frac{\mu}{2}s_{i-1}^n + \left(1 - \frac{\mu}{2}\right)s_i^n\right], \tag{8}$$

where $\mu = \lambda\Delta t/\Delta x$ is the CFL number and $\phi_{i+1/2}^{n+1/2} = \lambda(w_i^n + \frac{\Delta t}{2}s_i^n)$ is the numerical flux at the boundary at instant $t_{n+1/2}$. Secondly, we get this scheme from AES:

$$w_i^{n+1} = w_{i-\mu}^n + s_{i-\mu}^n\Delta t, \tag{9}$$

where $w_{i-\mu}^n = (1-\mu)w_i^n + \mu w_{i-1}^n$ is the solution at the foot of the characteristic that passes through (x_i, t_{n+1}) and $s_{i-\mu}^n = (1-\mu)s_i^n + \mu s_{i-1}^n$ is the source at such foot approximated via an interpolation ($s_{i-\mu}^n = s(x_{i-\mu}, t_n, w_{i-\mu}^n)$ is also possible). Figure 2 shows its graphical interpretation as the solution along the characteristics

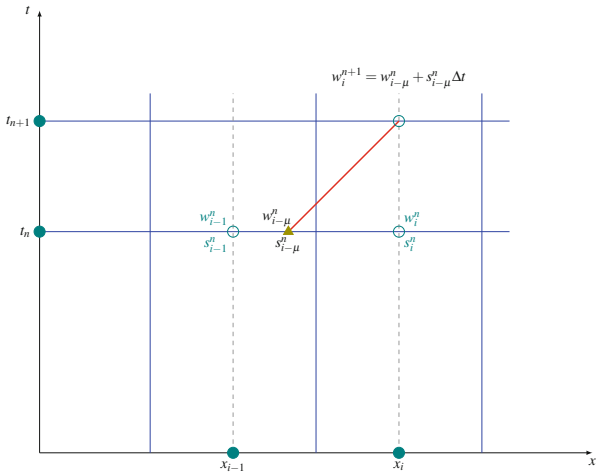


Fig. 2 Graphical interpretation of (9)

and it corresponds to an upwind scheme introduced by Bermúdez and Vázquez-Cendón [1]. Finally, the scheme from EAS is:

$$w_i^{n+1} = (1-\mu)\hat{w}_i^{n+1} + \mu\hat{w}_{i-1}^{n+1}, \tag{10}$$

where $\hat{w}_i^{n+1} = w_i^n + s_i^n \Delta t$ is the evolved solution in the cell. Figure 3 shows its graphical interpretation as the cell-averaged value of the advected evolved solution.

After easy manipulations, we can see that (8)–(10) are equivalent to each other. These resultings schemes are first-order accurate and stable if $\mu \leq 1$.

We shall show the results from a simulation:

Example 1 Let us consider the following two-dimensional IVP:

$$\begin{cases} \partial_t \mathbf{W} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \partial_x \mathbf{W} + \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix} \partial_y \mathbf{W} = \mathbf{S}, \\ \\ \mathbf{W}(x, y, 0) = \left(0,\, 0,\, \sin y - \sin x\right)^T, \end{cases} \tag{11}$$

where $\mathbf{S}(x, y, t, \mathbf{W}) = \left((\cos y - y \sin x)\cos t,\, (\cos x - x \sin y)\cos t,\, 0\right)^T$. Its solution is:

$$\mathbf{W}(x, y, t) = \left(y \sin x \sin t,\, x \sin y \sin t,\, (\sin y - \sin x)\cos t\right)^T. \tag{12}$$

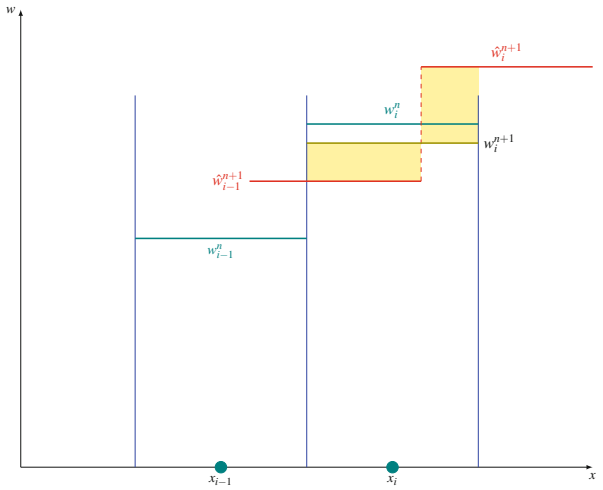


Fig. 3 Graphical interpretation of (10)

We solve (11) numerically by means of source (operator $S^{(\Delta t)}$) and dimensional (operators $X^{(\Delta t)}$ and $Y^{(\Delta t)}$) splittings in the domain $[0,1] \times [0,1]$ with a structured mesh, $\Delta x = \Delta y$, from $t = 0$ to $t = 1$:

- We construct this EAS-based first-order scheme, using Godunov to advect:

$$\mathbf{W}_{i,j}^{n+1} = Y^{(\Delta t)} X^{(\Delta t)} S^{(\Delta t)} \mathbf{W}_{i,j}^{n}, \tag{13}$$

$$\mathbf{W}_{i,j}^{n+1/3} = \mathbf{W}_{i,j}^{n} + \mathbf{S}_{i,j}^{n+1/2} \Delta t \quad (\textit{Leap-Frog} \text{ method}). \tag{14}$$

- We construct this EAES-based second-order scheme, using Lax-Wendroff to advect:

$$\mathbf{W}_{i,j}^{n+1} = S^{(\Delta t/2)} X^{(\Delta t/2)} Y^{(\Delta t)} X^{(\Delta t/2)} S^{(\Delta t/2)} \mathbf{W}_{i,j}^{n}, \tag{15}$$

$$\mathbf{W}_{i,j}^{n+1/5} = \mathbf{W}_{i,j}^{n} + \mathbf{S}_{i,j}^{n+1/2} \frac{\Delta t}{2}, \quad \mathbf{W}_{i,j}^{n+1} = \mathbf{W}_{i,j}^{n+4/5} + \mathbf{S}_{i,j}^{n+1/2} \frac{\Delta t}{2}. \tag{16}$$

As the eigenvalues are $\lambda_1 = 1$, $\lambda_2 = -1$ and $\lambda_3 = 0$, the schemes are stable if $\Delta t \leq \Delta x$. Table 1 shows the ∞-norm errors of the solution for different mesh sizes and $\mu = 1$. The ratios between errors verify the order of convergence [17].

Table 1 Errors in $||.||_\infty$ of the solution at $t = 1$

Δx	$\epsilon_{\mathrm{G}\,\Delta x}^{\infty}$	$\epsilon_{\mathrm{G}\,\Delta x}^{\infty}/\epsilon_{\mathrm{G}\,\frac{\Delta x}{2}}^{\infty}$	$\epsilon_{\mathrm{LW}\,\Delta x}^{\infty}$	$\epsilon_{\mathrm{LW}\,\Delta x}^{\infty}/\epsilon_{\mathrm{LW}\,\frac{\Delta x}{2}}^{\infty}$
0.025	4.62048×10^{-3}	–	5.50321×10^{-5}	–
0.0125	2.29361×10^{-3}	2.01	1.41283×10^{-5}	3.90
0.00625	1.14259×10^{-3}	2.01	3.61852×10^{-6}	3.90
0.003125	5.70237×10^{-4}	2.00	9.24710×10^{-7}	3.91

3 Splitting of Nonlinear Equations

In the nonlinear situation, we have to modify the flux of the balance law in order to achieve equivalent EAS and AES splittings. We shall justify this fact with the following illustrative example (Burger's equation):

$$\begin{cases} \dfrac{\partial w}{\partial t} + \dfrac{\partial}{\partial x}\left(\dfrac{w^2}{2}\right) = w, \qquad x \in \mathbb{R}, \\ \\ w(x,0) = w_0(x). \end{cases} \tag{17}$$

As a first approach, Langseth et al. [5] constructed non-equivalent splitting by using the same flux as the IVP; we will name them *AES and *EAS, respectively. Figure 4 compares the exact solution of the IVP with these two constructions, for (17) at $t = 0.5$ when $w_0(x) = e^{-x^2}$. It evinces the error made when extrapolating the linear constructions to the nonlinear case [7]: the exact solution of *AES is behind the exact solution (there is a delay) and the solution of *EAS is ahead (there is an advance). This means that we have to modify the flux in order to have an exact splitting; for (17) we introduce the fluxes:

$$\hat{f}^{AES}(w,t) = \frac{w^2}{2}e^{t}, \qquad \hat{f}^{EAS}(w,t) = \frac{w^2}{2}e^{-t}. \tag{18}$$

At discrete level, numerical schemes obtained from EAS and AES are also equivalent one to another like in the linear case (the evolution is solved exactly and the advection is calculated by Godunov's method):

$$\left(w_i^{n+1}\right)_{AES} = \left(w_i^{n+1}\right)_{EAS} = w_i^n e^{\Delta t} - \frac{e^{2\Delta t} - e^{\Delta t}}{\Delta x}\left(\phi_{i+\frac{1}{2}}^{n+\frac{1}{2}} - \phi_{i-\frac{1}{2}}^{n+\frac{1}{2}}\right), \tag{19}$$

where $\phi_{i\pm 1/2}^{n+1/2}$ is the numerical flux of Godunov's method for Burgers' equation (Toro [14]). We can contrast this first-order numerical scheme of AES and EAS

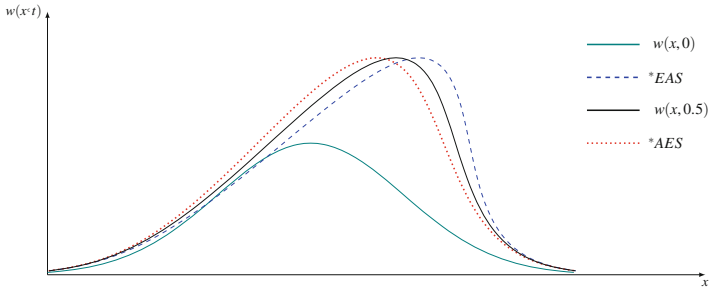


Fig. 4 Exact solution of (17) at $t = 0.5$ (in *black*) with $w_0(x) = e^{-x^2}$ (in *green*), compared with the non-equivalent *AES (in *red*) and *EAS (in *blue*) of Langseth et al.

with *AES and *EAS:

$$\left(w_i^{n+1}\right)^*_{AES} = w_i^n e^{\Delta t} - \frac{e^{\Delta t}\,\Delta t}{\Delta x}\left(\phi_{i+\frac{1}{2}}^{n+\frac{1}{2}} - \phi_{i-\frac{1}{2}}^{n+\frac{1}{2}}\right), \tag{20}$$

$$\left(w_i^{n+1}\right)^*_{EAS} = w_i^n e^{\Delta t} - \frac{e^{2\Delta t}\,\Delta t}{\Delta x}\left(\phi_{i+\frac{1}{2}}^{n+\frac{1}{2}} - \phi_{i-\frac{1}{2}}^{n+\frac{1}{2}}\right), \tag{21}$$

If we calculate their local truncation error, we see that the leading terms coincide in space but differ in time. This difference (of $\mathcal{O}(\Delta t)$, as Langseth et al. proved [5]) is the consequence of the lagging and leading phase errors depicted in Fig. 4 (these local truncation errors are calculated assuming $w(x,t) > 0$):

$$[\tau_i]^{AES}_{\Delta t} = [\tau_i]^{EAS}_{\Delta t} = \left(\frac{w_{tt}|_i^n - w|_i^n}{2} + \frac{3}{2} w|_i^n w_x|_i^n\right)\Delta t + \mathcal{O}(\Delta t^2), \tag{22}$$

$$[\tau_i]^{*AES}_{\Delta t} = \left(\frac{w_{tt}|_i^n - w|_i^n}{2} + w|_i^n w_x|_i^n\right)\Delta t + \mathcal{O}(\Delta t^2), \tag{23}$$

$$[\tau_i]^{*EAS}_{\Delta t} = \left(\frac{w_{tt}|_i^n - w|_i^n}{2} + 2w|_i^n w_x|_i^n\right)\Delta t + \mathcal{O}(\Delta t^2). \tag{24}$$

This result evidences that AES and EAS starts to be advantageous over *AES and *EAS when using at least second-order FV schemes.

For a general flux and source, it is highly complicated to obtain the formulae for AES and EAS. For example, if in (17) we consider the source term $s(x,t,w) = x$, the fluxes for the splittings are $\hat{f}_{AES}(w,t,x) = \frac{1}{2}w^2\cosh t + wx\sinh t$ and $\hat{f}_{EAS}(w,t,x) = \frac{1}{2}w^2\cosh^{-2} t - wx\cosh^{-2} t\tanh t$; and if we consider $s(x,t,w) = t$, then the fluxes are $\hat{f}_{AES}(w,t) = \frac{1}{2}w^2 + \frac{1}{2}wt^2$ and $\hat{f}_{EAS}(w,t) = \frac{1}{2}w^2 - wt^2$. Therefore further studies in this area will be needed to clarify this issue and, in this paper, we propose the following approximation of the flux of AES by a Taylor series:

Proposition 2 *The Taylor expansion of the flux of AES is:*

$$\begin{aligned} \hat{f}^{AES}(w,t,x) &= \int \hat{\lambda}(w,t,x)\,dw, \\ \hat{\lambda} &= \lambda + [\lambda' s]\,t + \tfrac{1}{2}\left[\lambda'' s^2 + \lambda'(s_x \lambda + s_t + s_w s)\right] t^2 + \ldots \end{aligned} \tag{25}$$

If we truncate the Taylor expansion at the degree N, we can obtain numerical schemes with splitting error up to order $N + 1$.

Acknowledgements The authors are indebted to Professor E. F. Toro for many valuable discussions. This work was financially supported by Spanish MICINN project CGL2011-28499-C03-01.

References

1. Bermúdez, A., Vázquez-Cendón, M.E.: Upwind methods for hyperbolic conservation laws with source terms. Comput. Fluids **23**(8), 1049–1071 (1994)
2. Chalabi, A.: Stable upwind schemes for hyperbolic conservation laws with source terms. IMA J. Numer. Anal. **12**(2), 217–241 (1992)
3. Hundsdorfer, W., Verwer, J.G.: A note on splitting errors for advection-reaction equations. Appl. Numer. Math. **18**(1), 191–199 (1995)
4. Hvistendahl Karlsen, K., Bursdal, K., Dahle, H.K., Evje, S., Lie, K.-A.: The corrected operator splitting approach applied to a nonlinear advection-diffusion problem. Comput. Method Appl. Mech. **167**(3–4), 239–260 (1998)
5. Langseth, J.O., Tveito, A., Winther, R.: On the convergence of operator-splitting applied to conservation laws with source terms. SIAM J. Numer. Anal. **33**(3), 843–863 (1996)
6. LeVeque, R.J.: Balancing source terms and flux gradients in high-resolution Godunov methods: the quasy-steady wave-propagation algorithm. J. Comput. Phys. **146**(1), 346–365 (1998)
7. LeVeque, R.J.: Finite-Volume Methods for Hyperbolic Problems. Cambridge Texts in Applied Mathematics, vol. 31. CUP, Cambridge (2002)
8. LeVeque, R.J., Yee, H.C.: A study of numerical methods for hyperbolic conservation laws with stiff source terms. J. Comput. Phys. **86**(1), 187–210 (1990)
9. Montecinos, G., Castro, C.E., Dumbser, M., Toro, E.F.: Comparison of solvers for the generalized Riemann problem for hyperbolic systems with source terms. J. Comput. Phys. **231**(19), 6472–6494 (2012)
10. Monthé, L.A.: A study of splitting scheme for hyperbolic conservation laws with source terms. J. Comput. Appl. Math. **137**(1), 1–12 (2001)
11. Peyroutet, F.: Splitting method applied to hyperbolic problem with source term. Appl. Math. Lett. **14**(1), 99–104 (2001)
12. Roe, P.L.: Upwind differencing schemes for hyperbolic conservation laws with source terms. Lect. Notes Math. **1270**, 41–51 (1987)
13. Strang, G.: On the construction and comparison of difference schemes. SIAM J. Numer. Anal. **5**(3), 506–517 (1968)
14. Toro, E.F.: Riemann Solvers and Numerical Methods for Fluid Dynamics: A Practical Introduction, 3rd edn. Springer, Heidelberg (2009)
15. Toro, E.F., Titarev, V.A.: Solution of the generalized Riemann problem for advection-reaction equations. Proc. R. Soc. Lond. A Math. **458**, 271–281 (2002)
16. Vázquez-Cendón, M.E.: Improved treatment of source terms in upwind schemes for the shallow water equations in channels with irregular geometry. J. Comput. Phys. **148**(2), 497–526 (1999)
17. Vázquez-Cendón, M.E., Cea, L.: Analysis of a new Kolgan-type scheme motivated by the shallow water equations. Appl. Numer. Math. **62**(4), 489–506 (2012)