SEMA SIMAI Springer Series 18

Juan Luis García Guirao
José Alberto Murillo Hernández
Francisco Periago Esparza *Editors*

# Recent Advances in Differential Equations and Applications

SeMA
SIMAI SOCIETÀ ITALIANA DI MATEMATICA APPLICATA E INDUSTRIALE
Springer

# SEMA SIMAI Springer Series

Series Editors: Luca Formaggia • Pablo Pedregal (Editors-in-Chief)
Mats G. Larson • Tere Martínez-Seara Alonso • Carlos Parés • Lorenzo Pareschi •
Andrea Tosin • Elena Vazquez • Jorge P. Zubelli • Paolo Zunino

Volume 18

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

Juan Luis García Guirao • José Alberto Murillo Hernández • Francisco Periago Esparza
Editors

# Recent Advances in Differential Equations and Applications

Springer

*Editors*

Juan Luis García Guirao
Department of Applied Mathematics
and Statistics
Technical University of Cartagena
Cartagena
Murcia, Spain

José Alberto Murillo Hernández
Department of Applied Mathematics
and Statistics
Technical University of Cartagena
Cartagena
Murcia, Spain

Francisco Periago Esparza
Department of Applied Mathematics
and Statistics
Technical University of Cartagena
Cartagena
Murcia, Spain

ISSN 2199-3041 ISSN 2199-305X (electronic)
SEMA SIMAI Springer Series
ISBN 978-3-030-00340-1 ISBN 978-3-030-00341-8 (eBook)
https://doi.org/10.1007/978-3-030-00341-8

Library of Congress Control Number: 2018965411

© Springer Nature Switzerland AG 2019
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

# Preface

This volume is one of the outcomes of the XXV Congress on Differential Equations and Applications (Spanish acronymized as CEDYA) and XV Congress on Applied Mathematics (CMA), which was held in Cartagena (Region of Murcia, Spain), June 26–30, 2017. This series of congress was originally an initiative of a group of researchers working in various applied mathematics areas, mainly ordinary and partial differential equations, numerical analysis and control and optimization, with the purpose of bring together people interested in these topics and diffuse their recent results. Nowadays this biennial international conference is the most important event organized by the Spanish Society of Applied Mathematics (SEMA).

Focussing on the XXV CEDAY/XV CMA, its main aim was to serve as a meeting place for all those who carry out their research work in the field of applied mathematics in the broad sense. This conference brought together an excellent group of international and national researchers (270 participants, with 8 plenary speakers) interested in the different branches of applied mathematics. The issues addressed include, among others: ordinary differential equations, partial differential equations, optimization and control, numerical analysis, scientific and computational calculus, models and industrial applications, approximation theory, discrete mathematics, numerical linear algebra, dynamical systems, ....

The collection of articles in this volume is based on a selection of the contributions presented at the conference. Every submitted paper has undergone a standard refereeing process. The volume does not strictly represent the acts of the congress, although it intends to reflect its structure and topics. It provides a good summary of the recent activity of the different Spanish research groups devoted to the applications of mathematics to certain branches of the experimental sciences and engineering.

This volume would not have been possible without the help of various people who contributed in different ways. First of all, we would like to thank the authors themselves for submitting their work to this issue. Special thanks go to the referees who agreed to take part in this process: their comments and suggestions have led to improvements in most of the contributions.

We would also like to express our gratitude to Francesca Bonadei from Springer for her patience, attention and constant support at every step in the editorial process.

Cartagena, Spain
September 2018

Juan Luis García Guirao
José Alberto Murillo Hernández
Francisco Periago Esparza

# Contents

# Contributors

**Euardo Alcaín** Departamento de Matemática Aplicada, Ciencia e Ingeniería de Materiales y Tecnología Electrónica, Universidad Rey Juan Carlos, ESCET, Móstoles, Madrid, Spain

**Jone Apraiz** Departamento de Matemáticas, Universidad del País Vasco, Leioa, Spain

**Iñigo Arregui** Departamento de Matemáticas, Universidade da Coruña, Coruña, Spain

**Jean Baccou** Institut de Radioprotection et de Sureté Nucléaire (IRSN), PSN-RES/SEMIA/LIMAR, CE Cadarache, Saint Paul Les Durance, France

**Patricia Barral** Departamento de Matemática Aplicada, Universidade de Santiago de Compostela, Santiago de Compostela, Spain

Technological Institute for Industrial Mathematics (ITMATI), Santiago de Compostela, Spain

**José Carlos Bellido** Departamento de Matemáticas, ETSII, Universidad de Castilla-La Mancha, Ciudad Real, Spain

**María Santos Bruzón** Departamento de Matemáticas, Universidad de Cádiz, Cádiz, Spain

**J. Jesús Cendán** Departamento de Matemáticas, Universidade da Coruña, Coruña, Spain

**Rafael de la Rosa** Departamento de Matemáticas, Universidad de Cádiz, Cádiz, Spain

**Alberto Donoso** Departamento de Matemáticas, ETSII, Universidad de Castilla-La Mancha, Ciudad Real, Spain

**Angel Durán** Departamento de Matemática Aplicada, Universidad de Valladolid, Valladolid, Spain

**Elisenda Feliu** Department of Mathematical Sciences, University of Copenhagen, København, Denmark

**María Isabel García-Planas** Universitat Politècnica de Catalunya, Barcelona, Spain

**Tamara M. Garrido** Departamento de Matemáticas, Universidad de Cádiz, Cádiz, Spain

**María González** Departamento de Matemáticas, Universidade da Coruña, Coruña, Spain

**Francisco Guillén-González** Departamento de Ecuaciones Diferenciales y Análisis Numérico, IMUS, Universidad de Sevilla, Sevilla, Spain

**Ángela Jiménez-Casas** Departamento de Matemática Aplicada, Grupo de Dinámica No lineal, Universidad Pontificia Comillas de Madrid, Madrid, Spain

**Zhiquing Kui** Aix Marseille Université, CNRS, Centrale Marseille, I2M, UMR 7353, Marseille, France

**Jacques Liandrat** Aix Marseille Université, CNRS, Centrale Marseille, I2M, UMR 7353, Marseille, France

**Ana Isabel Muñoz** Departamento de Matemática Aplicada, Ciencia e Ingeniería de Materiales y Tecnología Electrónica, Universidad Rey Juan Carlos, ESCET, Móstoles, Madrid, Spain

**Begoña Nicolás** Departamento de Matemática Aplicada, Universidade de Santiago de Compostela, Santiago de Compostela, Spain

**Luis Javier Pérez-Pérez** Departamento de Matemática Aplicada, Universidade de Santiago de Compostela, Santiago de Compostela, Spain

**Peregrina Quintela** Departamento de Matemática Aplicada, Universidade de Santiago de Compostela, Santiago de Compostela, Spain Technological Institute for Industrial Mathematics (ITMATI), Santiago de Compostela, Spain

**Iván Ramírez** Departamento de Matemática Aplicada, Ciencia e Ingeniería de Materiales y Tecnología Electrónica, Universidad Rey Juan Carlos, ESCET, Móstoles, Madrid, Spain

**Higinio Ramos** Grupo de Computación Científica, Universidad de Salamanca, Escuela Politécnica Superior de Zamora, Zamora, Spain

**María Victoria Redondo-Neble** Departamento de Matemáticas, Universidad de Cádiz, Cádiz, Spain

**José Rafael Rodríguez-Galván** Departamento de Matemáticas, Universidad de Cádiz, Cádiz, Spain

**David Ruiz** Departamento de Matemáticas, ETSII, Universidad de Castilla-La Mancha, Ciudad Real, Spain

**Meritxell Sáez** Department of Mathematical Sciences, University of Copenhagen, København, Denmark

**Emanuele Schiavi** Departamento de Matemática Aplicada, Ciencia e Ingeniería de Materiales y Tecnología Electrónica, Universidad Rey Juan Carlos, ESCET, Móstoles, Madrid, Spain

**Carsten Wiuf** Department of Mathematical Sciences, University of Copenhagen, København, Denmark

# About the Editors

**Juan Luis García Guirao** is a Full Professor of Applied Mathematics at the Technical University of Cartagena, Spain. He completed his Master's and PhD in Mathematics at the Universidad de Murcia, Spain in 2001 and 2004. Under the supervision of Professor Francisco Balibrea, Full Professor of Mathematical Analysis and founder of the Dynamical Systems Group of the Region of Murcia (http://www.um.es/sistdinamicos) he defended his Master's Thesis in 2001 and PhD Thesis in 2004. His doctoral studies also involved a research stay at the Instituto Superior Tecnico in Lisbon under the supervision of Professor José Sousa-Ramos, one of the main researchers in Dynamical Systems in Portugal.

He has held several positions at different universities, including the Universidad de Alicante, Universidad de Castilla-La Mancha, Universidad Autónoma de Barcelona, and Technical University of Cartagena.

In 2011, at the age of 33, he became Spain's youngest Full Professor of Mathematics.

The author of more than 100 research papers published in prominent journals, he also serves on the Editorial Board of several journals, including MATCH Commun. Math. Comput. Chem., Open Physics, and Discrete Dynamics in Nature and Society.

He has received a number of awards, e.g. the NSP 2017 for researchers younger than 40 years old, and the JDEA 2017 award for the best paper published in the field of Differential Equations. In addition, he has opened three research lines: discrete dynamical systems defined in low dimensional spaces, study of the periodic structure of smooth systems using the homological Lefschetz theory and analysis of Hamiltonian systems. Applications of the previous lines to problems from the fields of economics, chemistry and engineering are also considered. Further information is available at: http://www.jlguirao.es/.

**José Alberto Murillo Hernández** is an Associate Professor at the Department of Applied Mathematics and Statistics, Technical University of Cartagena, Spain. He received his Ph.D. in Mathematics from the University of Valencia, Spain and subsequently took part in a post-doctoral research stay at the Ecole Polytechnique in Paris, France supported by a grant from the European Union. His research interests include set-valued and nonsmooth analysis, viability theory and control. He is currently investigating the description and control of the evolution of sets driven by morphological equations.

**Francisco Periago Esparza** completed his PhD at the University of Valencia, Spain, in 1999. He is currently an Associate Professor at the Department of Applied Mathematics and Statistics and member of the Computational Mechanics and Scientific Computing group at the Technical University of Cartagena, Spain.

His main research interests include optimal control, optimal design and controllability, both at the theoretical level and as applied to engineering problems. In the past few years, he has chiefly focused on optimal control for random PDEs. Further information is available at: http://www.upct.es/mc3/en/dr-francisco-periago-esparza/.

# Applications of Observability Inequalities

**Jone Apraiz**

**Abstract** This article presents two observability inequalities for the heat equation over $\Omega \times (0, T)$. In the first one, the observation is from a subset of positive measure in $\Omega \times (0, T)$, while in the second, the observation is from a subset of positive surface measure on $\partial\Omega \times (0, T)$. We will provide some applications for the above-mentioned observability inequalities, the bang-bang property for the minimal time control problems and the bang-bang property for the minimal norm control problems, and also establish new open problems related to observability inequalities and the aforementioned applications.

**Keywords** Parabolic equations · Control theory · Controllability · Observability inequalities · Bang-Bang properties

**AMS 2010 Codes:** 49J20, 49J30, 58E25, 93B05, 93B07, 35K05

## 1 Introduction

This article serves as a review on observability inequalities from measurable sets for solutions to the heat equation. The purpose of trying to obtain the two observability inequalities that we will see and prove in this article, was that in control theory there is a very well known result, the Hilbert Uniqueness Method, that assures that the null controllability of an equation is equivalent to obtain an observability inequality for the adjoint equation. This result is attributed to J.L. Lion. In our previous research we were studying the null controllability of parabolic equations over measurable sets, so, for the Hilbert Uniqueness Method reason, we focused on proving the observability inequalities (Theorems 1 and 2) that we will see in this article.

J. Apraiz (✉)
Departamento de Matemáticas, Universidad del País Vasco, Leioa, Spain
e-mail: jone.apraiz@ehu.eus

© Springer Nature Switzerland AG 2019

J. L. García Guirao et al. (eds.), *Recent Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 18,
https://doi.org/10.1007/978-3-030-00341-8_1

In the next lines of the Introduction we will establish the type of problem we will work on, remember some a priori estimates for the parabolic equations and recall some previous results about this kind of work.

Then, in Sect. 2, we will establish and prove Theorems 1 and 2 which will give us two observability inequalities. We will continue, in Sect. 3, showing some applications of the observability inequalities we have proved, the bang-bang property for the minimal time control problems and the bang-bang property for the minimal norm control problems. Finally, with Sect. 4, we will finish the article establishing some open problems related to observability inequalities and their applications to control theory.

Let $\Omega$ be a bounded Lipschitz domain in $\mathbb{R}^n$ and $T$ be a fixed positive time. Consider the heat equation:

$$\begin{cases} \partial_t u - \Delta u = 0, \quad \text{in } \Omega \times (0,T), \\ \quad u = 0, \quad \text{on } \partial\Omega \times (0,T), \\ u(0) = u_0, \quad \text{in } \Omega, \end{cases} \tag{1}$$

with $u_0$ in $L^2(\Omega)$. The solution of (1) will be treated as either a function from $[0,T]$ to $L^2(\Omega)$ or a function of two variables $x$ and $t$. Two important a priori estimates for the above equation are as follows:

$$\|u(T)\|_{L^2(\Omega)} \leq N(\Omega, T, \mathcal{D}) \int_{\mathcal{D}} |u(x,t)|\, dxdt, \tag{2}$$

for all $u_0 \in L^2(\Omega)$, where $\mathcal{D}$ is a subset of $\Omega \times (0,T)$, and

$$\|u(T)\|_{L^2(\Omega)} \leq N(\Omega, T, \mathcal{J}) \int_{\mathcal{J}} |\frac{\partial}{\partial \nu} u(x,t)|\, d\sigma dt, \tag{3}$$

for all $u_0 \in L^2(\Omega)$, where $\mathcal{J}$ is a subset of $\partial\Omega \times (0,T)$. Such a priori estimates are called observability inequalities.

In the case that $\mathcal{D} = \omega \times (0,T)$ and $\mathcal{J} = \Gamma \times (0,T)$ with $\omega$ and $\Gamma$ accordingly open and nonempty subsets of $\Omega$ and $\partial\Omega$, both inequalities (2) and (3) (where $\partial\Omega$ is smooth) were essentially first established, via the Lebeau-Robbiano spectral inequalities in [6]. These two estimates were set up to the linear parabolic equations (where $\partial\Omega$ is of class $C^2$), based on the Carleman inequality provided in [5]. In the case when $\mathcal{D} = \omega \times (0,T)$ and $\mathcal{J} = \Gamma \times (0,T)$ with $\omega$ and $\Gamma$ accordingly subsets of positive measure and positive surface measure in $\Omega$ and $\partial\Omega$, both inequalities (2) and (3) were built up in [1] with the help of a propagation of smallness estimate from measurable sets for real-analytic functions first established in [10]. For $\mathcal{D} = \omega \times E$, with $\omega$ and $E$ accordingly an open subset of $\Omega$ and a subset of positive measure in $(0,T)$, the inequality (2) (when $\partial\Omega$ is smooth) was proved in [11] with the aid of the Lebeau-Robbiano spectral inequality, and it was then verified for heat equations (when $\Omega$ is convex) with lower terms depending on the time variable, through a frequency function method in [8]. When $\mathcal{D} = \omega \times E$, with $\omega$ and $E$ accordingly subsets of positive measure in $\Omega$ and $(0, T)$, the estimate (2) (when $\partial\Omega$ is real-analytic) was obtained in [12].

In [2], we established the inequalities (2) and (3) when $\mathcal{D}$ and $\mathcal{J}$ were arbitrary subsets of positive measure and of positive surface measure in $\Omega \times (0, T)$ and $\partial\Omega \times (0, T)$ respectively. Such inequalities not only are mathematically interesting but also have important applications in the control theory of the heat equation, such as the bang-bang control, the time optimal control, the null controllability over a measurable set and so on.

We will see how we proved the two above-mentioned inequalities. We start assuming that the Lebeau-Robbiano spectral inequality stands on $\Omega$. To introduce it, we write

$$0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_j \le \cdots$$

for the eigenvalues of $-\Delta$ with the zero Dirichlet boundary condition over $\partial\Omega$, and $\{e_j : j \ge 1\}$ for the set of $L^2(\Omega)$-normalized eigenfunctions, i.e.,

$$\begin{cases} \Delta e_j + \lambda_j e_j = 0, & \text{in } \Omega, \\ e_j = 0, & \text{on } \partial\Omega. \end{cases} \tag{4}$$

For $\lambda > 0$ we define

$$\mathcal{E}_\lambda f = \sum_{\lambda_j \le \lambda} (f, e_j)\, e_j \quad \text{and} \quad \mathcal{E}_\lambda^\perp f = \sum_{\lambda_j > \lambda} (f, e_j)\, e_j,$$

where

$$(f, e_j) = \int_\Omega f\, e_j\, dx, \text{ when } f \in L^2(\Omega),\ j \ge 1.$$

Throughout this paper the following notations are used:

$$(f, g) = \int_\Omega fg\, dx \quad \text{and } \|f\|_{L^2(\Omega)} = (f, f)^{\frac{1}{2}}.$$

$\nu$ is the unit exterior normal vector to $\Omega$. $d\sigma$ is surface measure on $\partial\Omega$. $B_R(x_0)$ stands for the ball centered at $x_0$ in $\mathbb{R}^n$ of radius $R$, $\triangle_R(x_0)$ denotes $B_R(x_0) \cap \partial\Omega$, $B_R = B_R(0)$ and $\triangle_R = \triangle_R(0)$. For measurable sets $\omega \subset \mathbb{R}^n$ and $\mathcal{D} \subset \mathbb{R}^n \times (0, T)$, $|\omega|$ and $|\mathcal{D}|$ stand for the Lebesgue measures of the sets. For each measurable set $\mathcal{J}$ in $\partial\Omega \times (0, T)$, $|\mathcal{J}|$ denotes its surface measure on the lateral boundary of $\Omega \times \mathbb{R}$. $\{e^{t\Delta} : t \ge 0\}$ is the semigroup generated by $\Delta$ with zero Dirichlet boundary condition over $\partial\Omega$. Consequently, $e^{t\Delta} f$ is the solution to the problem (1) with the initial state $f$ in $L^2(\Omega)$. The Lebeau-Robbiano spectral inequality is as follows:

*For each* $0 < R \leq 1$*, there is* $N = N(\Omega, R)$*, such that the inequality*

$$\|\mathcal{E}_\lambda f\|_{L^2(\Omega)} \leq N e^{N\sqrt{\lambda}} \|\mathcal{E}_\lambda f\|_{L^2(B_R(x_0))} \tag{5}$$

*holds, when* $B_{4R}(x_0) \subset \Omega$, $f \in L^2(\Omega)$ and $\lambda > 0$.

## 2 Observability Inequalities

Our main results related to the observability inequalities are stated as follows, but, first, we will define the real-analyticity of the set $\triangle_{4R}(q_0)$.

**Definition 1** Let $q_0 \in \partial\Omega$ and $0 < R \leq 1$. We say that $\triangle_{4R}(q_0)$ is real-analytic with constants $\varrho$ and $\delta$ if for each $q \in \triangle_{4R}(q_0)$, there are a new rectangular coordinate system where $q = 0$, and a real-analytic function $\phi : B'_\varrho \subset \mathbb{R}^{n-1} \to \mathbb{R}$ verifying

$$\begin{cases} \phi(0') = 0, \quad |\partial^\alpha \phi(x')| \leq |\alpha|!\delta^{-|\alpha|-1}, \\ \text{when} \quad x' \in B'_\varrho, \ \alpha \in \mathbb{N}^{n-1}, \\ B_\varrho \cap \Omega = B_\varrho \cap \{(x', x_n) : x' \in B'_\varrho, \quad x_n > \phi(x')\}, \\ B_\varrho \cap \partial\Omega = B_\varrho \cap \{(x', x_n) : x' \in B'_\varrho, \quad x_n = \phi(x')\}. \end{cases} \tag{6}$$

Here, $B'_\varrho$ denotes the open ball of radius $\varrho$ and with center at $0'$ in $\mathbb{R}^{n-1}$.

In the next two theorems, we establish two observability inequalities for the heat equation over $\Omega \times (0, T)$. In Theorem 1, the observation is from a subset of positive measure in $\Omega \times (0, T)$, while in Theorem 2, the observation is from a subset of positive surface measure on $\partial\Omega \times (0, T)$.

**Theorem 1** *Suppose that a bounded domain* $\Omega$ *verifies the condition (5) and* $T > 0$*. Let* $x_0 \in \Omega$ *and* $R \in (0, 1]$ *be such that* $B_{4R}(x_0) \subset \Omega$*. Then, for each measurable set* $\mathcal{D} \subset B_R(x_0) \times (0, T)$ *with* $|\mathcal{D}| > 0$*, there is a positive constant* $B = B(\Omega, T, R, \mathcal{D})$*, such that*

$$\|e^{T\Delta} f\|_{L^2(\Omega)} \leq e^B \int_{\mathcal{D}} |e^{t\Delta} f(x)| \, dxdt, \tag{7}$$

*when* $f \in L^2(\Omega)$.

**Theorem 2** *Suppose that a bounded Lipschitz domain* $\Omega$ *verifies the condition (5) and* $T > 0$*. Let* $q_0 \in \partial\Omega$ *and* $R \in (0, 1]$ *be such that* $\triangle_{4R}(q_0)$ *is real-analytic.*

*Then, for each measurable set $\mathfrak{J} \subset \triangle_R(q_0) \times (0, T)$ with $|\mathfrak{J}| > 0$, there is a positive constant $B = B(\Omega, T, R, \mathfrak{J})$, such that*

$$\|e^{T\Delta} f\|_{L^2(\Omega)} \leq e^B \int_{\mathfrak{J}} |\frac{\partial}{\partial \nu} e^{t\Delta} f(x)| \, d\sigma dt, \tag{8}$$

*when $f \in L^2(\Omega)$.*

Next, we will see some results that will be necessary in the proof of the previous Theorem 1.

**Lemma 1** *Let $B_R(x_0) \subset \Omega$ and $\mathcal{D} \subset B_R(x_0) \times (0, T)$ be a subset of positive measure. Set*

$$\mathcal{D}_t = \{x \in \Omega : (x, t) \in \mathcal{D}\}, \ E = \{t \in (0, T) : |\mathcal{D}_t| \geq |\mathcal{D}|/(2T)\}, \ t \in (0, T). \tag{9}$$

*Then, $\mathcal{D}_t \subset \Omega$ is measurable for a.e. $t \in (0, T)$, $E$ is measurable in $(0, T)$, $|E| \geq |\mathcal{D}|/2|B_R|$ and*

$$\chi_E(t)\chi_{\mathcal{D}_t}(x) \leq \chi_{\mathcal{D}}(x, t), \ \text{in } \Omega \times (0, T). \tag{10}$$

*Proof* From Fubini's theorem,

$$|\mathcal{D}| = \int_0^T |\mathcal{D}_t| \, dt = \int_E |\mathcal{D}_t| \, dt + \int_{[0,T]\setminus E} |\mathcal{D}_t| \, dt \leq |B_R||E| + |\mathcal{D}|/2.$$

□

**Theorem 3** *Let $x_0 \in \Omega$ and $R \in (0, 1]$ be such that $B_{4R}(x_0) \subset \Omega$. Let $\mathcal{D} \subset B_R(x_0) \times (0, T)$ be a measurable set with $|\mathcal{D}| > 0$. Write $E$ and $\mathcal{D}_t$ for the sets associated to $\mathcal{D}$ in Lemma 1. Then, for each $\eta \in (0, 1)$, there are $N = N(\Omega, R, |\mathcal{D}|/(T|B_R|), \eta)$ and $\theta = \theta(\Omega, R, |\mathcal{D}|/(T|B_R|), \eta)$ with $\theta \in (0, 1)$, such that*

$$\|e^{t_2\Delta} f\|_{L^2(\Omega)} \leq \left( N e^{N/(t_2-t_1)} \int_{t_1}^{t_2} \chi_E(s) \|e^{s\Delta} f\|_{L^1(\mathcal{D}_s)} \, ds \right)^{\theta} \|e^{t_1\Delta} f\|_{L^2(\Omega)}^{1-\theta}, \tag{11}$$

*when $0 \leq t_1 < t_2 \leq T$, $|E \cap (t_1, t_2)| \geq \eta(t_2 - t_1)$ and $f \in L^2(\Omega)$. Moreover,*

$$\begin{aligned} &e^{-\frac{N+1-\theta}{t_2-t_1}} \|e^{t_2\Delta} f\|_{L^2(\Omega)} - e^{-\frac{N+1-\theta}{q(t_2-t_1)}} \|e^{t_1\Delta} f\|_{L^2(\Omega)} \\ &\leq N \int_{t_1}^{t_2} \chi_E(s) \|e^{s\Delta} f\|_{L^1(\mathcal{D}_s)} \, ds, \quad \text{when } \ q \geq (N + 1 - \theta)/(N + 1). \end{aligned} \tag{12}$$

The reader can find the proof of the following Lemma 2 in either [7, pp. 256–257] or [8, Proposition 2.1].

**Lemma 2** *Let $E$ be a subset of positive measure in $(0, T)$. Let $l$ be a density point of $E$. Then, for each $z > 1$, there is $l_1 = l_1(z, E)$ in $(l, T)$ such that, the sequence $\{l_m\}$ defined as*

$$l_{m+1} = l + z^{-m}\,(l_1 - l)\,,\ m = 1, 2, \cdots,$$

*verifies*

$$|E \cap (l_{m+1}, l_m)| \geq \frac{1}{3}\,(l_m - l_{m+1})\,,\ \textit{when}\ m \geq 1. \tag{13}$$

*Proof (Theorem 1)* Let $E$ and $\mathcal{D}_t$ be the sets associated to $\mathcal{D}$ in Lemma 1 and $l$ be a density point in $E$. For $z > 1$ to be fixed later, $\{l_m\}$ denotes the sequence associated to $l$ and $z$ in Lemma 2. Because (13) holds, we may apply Theorem 3, with $\eta = 1/3$, $t_1 = l_{m+1}$ and $t_2 = l_m$, for each $m \geq 1$, to get that there are $N = N(\Omega, R, |\mathcal{D}|/(T|B_R|)) > 0$ and $\theta = \theta(\Omega, R, |\mathcal{D}|/(T|B_R|))$, with $\theta \in (0, 1)$, such that

$$\begin{aligned} & e^{-\frac{N+1-\theta}{l_m - l_{m+1}}}\|e^{l_m\Delta}f\|_{L^2(\Omega)} - e^{-\frac{N+1-\theta}{q\left(l_m - l_{m+1}\right)}}\|e^{l_{m+1}\Delta}f\|_{L^2(\Omega)} \\ & \leq N\int_{l_{m+1}}^{l_m}\chi_E(s)\|e^{s\Delta}f\|_{L^1(\mathcal{D}_s)}\,ds, \quad \text{when}\ \ q \geq \frac{N+1-\theta}{N+1}\ \ \text{and}\ \ m \geq 1. \end{aligned} \tag{14}$$

Setting $z = 1/q$ in (14) (which leads to $1 < z \leq \frac{N+1}{N+1-\theta}$) and

$$\gamma_z(t) = e^{-\frac{N+1-\theta}{(z-1)(l_1-l)t}},\ t > 0,$$

recalling that

$$l_m - l_{m+1} = z^{-m}\,(z-1)\,(l_1 - l)\,,\ \text{for}\ m \geq 1,$$

we have

$$\begin{aligned} & \gamma_z(z^{-m})\|e^{l_m\Delta}f\|_{L^2(\Omega)} - \gamma_z(z^{-m-1})\|e^{l_{m+1}\Delta}f\|_{L^2(\Omega)} \\ & \leq N\int_{l_{m+1}}^{l_m}\chi_E(s)\|e^{s\Delta}f\|_{L^1(\mathcal{D}_s)}\,ds, \quad \text{when}\ \ m \geq 1. \end{aligned} \tag{15}$$

Choose now

$$z = \frac{1}{2}\left(1 + \frac{N+1}{N+1-\theta}\right).$$

The choice of $z$ and Lemma 2 determines $l_1$ in $(l, T)$ and from (15),

$$\begin{aligned}&\gamma(z^{-m})\|e^{l_m\Delta}f\|_{L^2(\Omega)} - \gamma(z^{-m-1})\|e^{l_{m+1}\Delta}f\|_{L^2(\Omega)}\\ &\leq N\int_{l_{m+1}}^{l_m}\chi_E(s)\|e^{s\Delta}f\|_{L^1(\mathcal{D}_s)}\,ds, \quad \text{when} \quad m\geq 1.\end{aligned} \tag{16}$$

with

$$\gamma(t) = e^{-A/t} \text{ and } A = A(\Omega, R, E, |\mathcal{D}|/(T|B_R|)) = \frac{2\,(N+1-\theta)^2}{\theta\,(l_1-l)}.$$

Finally, because of

$$\|e^{T\Delta}f\|_{L^2(\Omega)} \leq \|e^{l_1\Delta}f\|_{L^2(\Omega)}, \ \sup_{t\geq 0}\|e^{t\Delta}f\|_{L^2(\Omega)} < +\infty, \ \lim_{t\to 0+}\gamma(t) = 0,$$

and (10), the addition of the telescoping series in (16) gives

$$\|e^{T\Delta}f\|_{L^2(\Omega)} \leq Ne^{zA}\int_{\mathcal{D}\cap(\Omega\times[l,l_1])}|e^{t\Delta}f(x)|\,dxdt, \quad \text{for} \quad f\in L^2(\Omega),$$

which proves (7) with $B = zA + \log N$. □

*Remark 1* The constant $B$ in Theorem 1 depends on $E$ because the choice of $l_1 = l_1(z, E)$ in Lemma 2 depends on the possible complex structure of the measurable set $E$ (See the proof of Lemma 2 in [8, Proposition 2.1]). When $\mathcal{D} = \omega\times(0, T)$, one may take $l = T/2$, $l_1 = T$, $z = 2$ and then,

$$B = A(\Omega, R, |\omega|/|B_R|)/T.$$

*Remark 2* The proof of Theorem 1 also implies the following observability estimate:

$$\sup_{m\geq 0}\sup_{l_{m+1}\leq t\leq l_m} e^{-z^{m+1}A}\|e^{t\Delta}f\|_{L^2(\Omega)} \leq N\int_{\mathcal{D}\cap(\Omega\times[l,l_1])}|e^{t\Delta}f(x)|\,dxdt,$$

for $f$ in $L^2(\Omega)$, and with $z$, $N$ and $A$ as defined along the proof of Theorem 1. Here, $l_0 = T$.

Next, we will see some results that will be necessary in the proof of the previous Theorem 2.

**Lemma 3** *Let* $q_0\in\partial\Omega$ *and* $\mathfrak{J}\subset\triangle_R(q_0)\times(0, T)$ *be a subset with* $|\mathfrak{J}| > 0$. *Set*

$$\mathfrak{J}_t = \{x\in\partial\Omega : (x, t)\in\mathfrak{J}\}\ t\in(0, T) \quad E = \{t\in(0, T) : |\mathfrak{J}_t|\geq|\mathfrak{J}|/(2T)\}.$$

*Then,* $\mathfrak{J}_t \subset \triangle_R(q_0)$ *is measurable for a.e.* $t \in (0, T)$*,* $E$ *is measurable in* $(0, T)$*,* $|E| \geq |\mathfrak{J}|/(2|\triangle_R(q_0)|)$ *and* $\chi_E(t)\chi_{\mathfrak{J}_t}(x) \leq \chi_{\mathfrak{J}}(x, t)$ *over* $\partial\Omega \times (0, T)$*.*

*Proof* From Fubini's theorem,

$$|\mathfrak{J}| = \int_0^T |\mathfrak{J}_t|\, dt = \int_E |\mathfrak{J}_t|\, dt + \int_{[0,T]\setminus E} |\mathfrak{J}_t|\, dt \leq |\triangle_R(x_0)||E| + |\mathfrak{J}|/2.$$

□

**Theorem 4** *Suppose that* $\Omega$ *verifies the condition* (5)*. Assume that* $q_0 \in \partial\Omega$ *and* $R \in (0, 1]$ *such that* $\triangle_{4R}(q_0)$ *is real-analytic. Let* $\mathfrak{J}$ *be a subset in* $\triangle_R(q_0) \times (0, T)$ *of positive surface measure on* $\partial\Omega \times (0, T)$*,* $E$ *and* $\mathfrak{J}_t$ *be the measurable sets associated to* $\mathfrak{J}$ *in Lemma* 3*. Then, for each* $\eta \in (0, 1)$*, there are* $N = N(\Omega, R, |\mathfrak{J}|/(T|\triangle_R(q_0)|), \eta)$ *and* $\theta = \theta(\Omega, R, |\mathfrak{J}|/(T|\triangle_R(q_0)|), \eta)$ *with* $\theta \in (0, 1)$*, such that the inequality*

$$\|e^{t_2\Delta} f\|_{L^2(\Omega)} \leq \left(Ne^{N/(t_2-t_1)} \int_{t_1}^{t_2} \chi_E(t) \|\tfrac{\partial}{\partial\nu} e^{t\Delta} f\|_{L^1(\mathfrak{J}_t)}\, dt\right)^{\theta} \|e^{t_1\Delta} f\|_{L^2(\Omega)}^{1-\theta}, \tag{17}$$

*holds, when* $0 \leq t_1 < t_2 \leq T$ *with* $t_2 - t_1 < 1$*,* $|E \cap (t_1, t_2)| \geq \eta(t_2 - t_1)$ *and* $f \in L^2(\Omega)$*. Moreover,*

$$\begin{aligned} &e^{-\frac{N+1-\theta}{t_2-t_1}} \|e^{t_2\Delta} f\|_{L^2(\Omega)} - e^{-\frac{N+1-\theta}{q(t_2-t_1)}} \|e^{t_1\Delta} f\|_{L^2(\Omega)} \\ &\leq N \int_{t_1}^{t_2} \chi_E(t) \|\tfrac{\partial}{\partial\nu}\, e^{t\Delta} f\|_{L^1(\mathfrak{J}_t)}\, dt, \quad \textit{when} \quad q \geq \tfrac{N+1-\theta}{N+1}. \end{aligned} \tag{18}$$

*Proof (Theorem* 2*)* Let $E$ and $\mathfrak{J}_t$ be the sets associated to $\mathfrak{J}$ in Lemma 3 and $l$ be a density point in $E$. For $z > 1$ to be fixed later, $\{l_m\}$ denotes the sequence associated to $l$ and $z$ in Lemma 2. Because of (13) and from Theorem 4 with $\eta = 1/3$, $t_1 = l_{m+1}$ and $t_2 = l_m$, with $m \geq 1$, there are $N = N(\Omega, R, |\mathfrak{J}|/\,(T|\triangle_R(q_0)|)) > 0$ and $\theta = \theta(\Omega, R, |\mathfrak{J}|/\,(T|\triangle_R(q_0)|))$, with $\theta \in (0, 1)$, such that

$$\begin{aligned} &e^{-\frac{N+1-\theta}{l_m-l_{m+1}}} \|e^{l_m\Delta} f\|_{L^2(\Omega)} - e^{-\frac{N+1-\theta}{q\left(l_m-l_{m+1}\right)}} \|e^{l_{m+1}\Delta} f\|_{L^2(\Omega)} \\ &\leq N \int_{l_{m+1}}^{l_m} \chi_E(s) \|\tfrac{\partial}{\partial\nu}\, e^{s\Delta} f\|_{L^1(\mathfrak{J}_s)}\, ds, \quad \text{when} \quad q \geq \frac{N+1-\theta}{N+1} \quad \text{and} \quad m \geq 1. \end{aligned}$$

Let

$$z = \frac{1}{2}\left(1 + \frac{N+1}{N+1-\theta}\right).$$

Then, we can use the same arguments as those in the proof of Theorem 1 to verify Theorem 2. □

*Remark 3* The proof of Theorem 2 also implies the following observability estimate:

$$\sup_{m\geq 0}\sup_{l_{m+1}\leq t\leq l_m} e^{-z^{m+1}A}\|e^{t\Delta}f\|_{L^2(\Omega)} \leq N\int_{\mathfrak{J}\cap(\partial\Omega\times[l,l_1])}\left|\tfrac{\partial}{\partial\nu}e^{t\Delta}f(x)\right|d\sigma dt,$$

for $f$ in $L^2(\Omega)$, with $A = 2(N+1-\theta)^2/[\theta(l_1-l)]$ and with $z$, $N$ and $\theta$ as given along the proof of Theorem 2. Here, $l_0 = T$.

## 3 Applications of Observability Inequalities

We will now show some applications of the Theorems 1 and 2 in the control theory of the heat equation. Specifically, we will focus on the uniqueness and bang-bang properties of the minimal time and minimal $L^\infty$-norm control problems.

In this section we assume that $T > 0$ and that $\Omega$ is a bounded Lipschitz domain verifying the condition (5).

First of all, we will show that Theorems 1 and 2 imply the null controllability with controls restricted over measurable subsets in $\Omega \times (0,T)$ and $\partial\Omega \times (0,T)$ respectively. Let $\mathfrak{D}$ be a measurable subset with positive measure in $B_R(x_0)\times(0,T)$ with $B_{4R}(x_0)\subset\Omega$. Let $\mathfrak{J}$ be a measurable subset with positive surface measure in $\triangle_R(q_0)\times(0,T)$, where $q_0\in\partial\Omega$, $R\in(0,1]$ and $\triangle_{4R}(q_0)$ is real-analytic. Consider the following controlled heat equations:

$$\begin{cases} \partial_t u - \Delta u = \chi_{\mathfrak{D}} v, \text{ in } \Omega\times(0,T], \\ u = 0, \text{ on } \partial\Omega\times[0,T], \\ u(0) = u_0, \text{ in } \Omega, \end{cases} \tag{19}$$

and

$$\begin{cases} \partial_t u - \Delta u = 0, \text{ in } \Omega\times(0,T], \\ u = g\,\chi_{\mathfrak{J}}, \text{ on } \partial\Omega\times[0,T], \\ u(0) = u_0, \text{ in } \Omega, \end{cases} \tag{20}$$

where $u_0 \in L^2(\Omega)$, $v \in L^\infty(\Omega \times (0,T))$ and $g \in L^\infty(\partial\Omega \times (0,T))$ are controls. *We say that u is the solution to* (20) *if* $v \equiv u - e^{t\Delta}u_0$ *is the unique solution defined in [4, Theorem 3.2] to*

$$\begin{cases} \partial_t v - \Delta v = 0, \text{ in } \Omega \times (0,T), \\ \\ v = g\chi_{\mathfrak{J}}, \text{ on } \partial\Omega \times (0,T), \\ \\ v(0) = 0, \text{ in } \Omega, \end{cases} \tag{21}$$

*with g in* $L^p(\partial\Omega \times (0,T))$ *for some* $2 \leq p \leq \infty$.

From now on, we always denote by $u(\cdot\,; u_0, v)$ and $u(\cdot\,; u_0, g)$ the solutions to problems (19) and (20) corresponding to $v$ and $g$ respectively.

**Corollary 1** *For each* $u_0 \in L^2(\Omega)$*, there are bounded control functions* $v$ *and* $g$ *with*

$$\|v\|_{L^\infty(\Omega\times(0,T))} \leq C_1\|u_0\|_{L^2(\Omega)},$$

$$\|g\|_{L^\infty(\partial\Omega\times(0,T))} \leq C_2\|u_0\|_{L^2(\Omega)},$$

*such that* $u(T; u_0, v) = 0$ *and* $u(T; u_0, g) = 0$*. Here* $C_1 = C(\Omega, T, R, \mathfrak{D})$ *and* $C_2 = C(\Omega, T, R, \mathfrak{J})$.

*Proof* We only prove the boundary controllability. Let $E$ be the measurable set associated to $\mathfrak{J}$ in Lemma 3. Write

$$\widetilde{\mathfrak{J}} = \{(x,t) : (x, T-t) \in \mathfrak{J}\} \quad \text{and} \quad \widetilde{E} = \{t : T-t \in E\}\,.$$

Let $l > 0$ be a density point of $\widetilde{E}$ (Hence, $T - l$ is a density point of $E$). We choose $z$, $l_1$ and the sequence $\{l_m\}$ as in the proof of Theorem 2 but with $\mathfrak{J}$ and $E$ accordingly replaced by $\widetilde{\mathfrak{J}}$ and $\widetilde{E}$. It is clear that

$$0 < l < \cdots < l_{m+1} < l_m \cdots < l_1 < l_0 = T, \ \lim_{m\to+\infty} l_m = l.$$

We set

$$\mathcal{M} = \mathfrak{J} \cap (\partial\Omega \times [T-l_1, T-l]) \subset \mathfrak{J}.$$

It is clear that $|\mathcal{M}| > 0$. The proof of Theorem 2, the change of variables $t = T - \tau$ and Remark 3 show that the observability inequality

$$\|\varphi(0)\|_{L^2(\Omega)} \leq e^B \int_{\mathcal{M}} |\tfrac{\partial\varphi}{\partial\nu}(p,t)|\, d\sigma dt, \tag{22}$$

holds, when $\varphi$ is the unique solution in $L^\infty([0,T],L^2(\Omega))\cap L^2([0,T],H_0^1(\Omega))$ to

$$\begin{cases}\partial_t\varphi+\Delta\varphi=0, & \text{in } \Omega\times[0,T),\\ \varphi=0, & \text{on } \partial\Omega\times[0,T),\\ \varphi(T)=\varphi_T, & \text{in } \partial\Omega,\end{cases}\tag{23}$$

for some $\varphi_T$ in $L^2(\Omega)$. Set

$$X=\{\tfrac{\partial\varphi}{\partial\nu}|_{\mathcal{M}}:\varphi(t)=e^{(T-t)\Delta}\varphi_T,\ \text{for } 0\le t\le T,\ \text{for some } \varphi_T\in L^2(\Omega)\}.$$

Since $\mathcal{M}\subset\partial\Omega\times[T-l_1,T-l]$, $X$ is a subspace of $L^1(\mathcal{M})$ and from (22), the linear mapping $\Lambda:X\longrightarrow\mathbb{R}$, defined by

$$\Lambda(\tfrac{\partial\varphi}{\partial\nu}|_{\mathcal{M}})=(u_0,\varphi(0)),$$

verifies

$$\left|\Lambda(\tfrac{\partial\varphi}{\partial\nu}|_{\mathcal{M}})\right|\le e^B\|u_0\|_{L^2(\Omega)}\int_{\mathcal{M}}|\tfrac{\partial\varphi}{\partial\nu}(p,t)|\,d\sigma dt,\ \text{when } \tfrac{\partial\varphi}{\partial\nu}|_{\mathcal{M}}\in X.$$

From the Hahn-Banach theorem, there is a linear extension $T:L^1(\mathcal{M})\longrightarrow\mathbb{R}$ of $\Lambda$, with

$$T(\tfrac{\partial\varphi}{\partial\nu}|_{\mathcal{M}})=(u_0,\varphi(0)),\ \text{when } \tfrac{\partial\varphi}{\partial\nu}|_{\mathcal{M}}\in X,$$
$$|T(f)|\le e^B\|u_0\|\|f\|_{L^1(\mathcal{M})},\ \text{for all } f\in L^1(\mathcal{M}).$$

Thus, $T$ is in $L^1(\mathcal{M})^*=L^\infty(\mathcal{M})$ and there is $g$ in $L^\infty(\mathcal{M})$ verifying

$$T(f)=\int_{\mathcal{M}}fg\,d\sigma dt,\ \text{for all } f\in L^1(\mathcal{M}) \text{ and } \|g\|_{L^\infty(\mathcal{M})}\le e^B\|u_0\|.$$

We extend $g$ over $\partial\Omega\times(0,T)$ by setting it to be zero outside $\mathcal{M}$ and denote the extended function by $g$ again. Then it holds that $u(T;u_0,g)=0$ provided that we know that

$$\int_\Omega u(T;u_0,g)\varphi_T\,dx=\int_\Omega u_0\varphi(0)\,dx-\int_{\mathcal{M}}g\,\tfrac{\partial\varphi}{\partial\nu}\,d\sigma dt,\ \text{for all } \varphi_T\in L^2(\Omega).\tag{24}$$

To prove (24), we first use the unique solvability for the problem

$$\begin{cases}\partial_t u-\Delta u=0, & \text{in } \Omega\times(0,T],\\ u=\gamma, & \text{on } \partial\Omega\times[0,T],\\ u(0)=0 & \text{in } \Omega,\end{cases}$$

with lateral Dirichlet data $\gamma$ in $L^p(\partial\Omega \times (0,T))$, $2 \leq p \leq \infty$, established in [4, Theorem 3.2] (See also [3, Theorems 8.1 and 8.3]). Then, because $g\chi_{\mathcal{M}}$ is bounded and supported in $\partial\Omega \times [T-l_1, T-l] \subset \partial\Omega \times (2\eta, T-2\eta)$ for some $\eta > 0$, the calculations leading to (24) can be justified via the regularization of $g\chi_{\mathcal{M}}$ and the approximation of $\Omega$ by smooth domains $\{\Omega_j;\ j \geq 1\}$ as in [3, Lemma 2.2]. □

## 3.1 Definition of the Minimal Time Control Problems and Main Results

In this section, we apply Theorems 1 and 2 to get the bang-bang property for the minimal time control problems usually called the first type of time optimal control problems; they are stated as follows. Let $\omega$ be a measurable subset with positive measure in $B_R(x_0)$ and $B_{4R}(x_0) \subset \Omega$. Suppose that $\triangle_{4R}(q_0)$ is real-analytic for some $q_0 \in \partial\Omega$ and $R \in (0,1]$ and let $\Gamma$ be a measurable subset with positive surface measure of $\triangle_R(x_0)$. For each $M > 0$, we define the following control constraint set:

$$\mathcal{U}^1_M = \{v \ \text{ measurable on } \ \Omega \times \mathbb{R}^+ : \ \ |v(x,t)| \leq M \ \text{ for a.e. } \ (x,t) \in \Omega \times \mathbb{R}^+\}.$$

$$\mathcal{U}^2_M = \{g \ \text{ measurable on } \ \partial\Omega \times \mathbb{R}^+ : \ \ |g(x,t)| \leq M \ \text{ for a.e. } \ (x,t) \in \partial\Omega \times \mathbb{R}^+\}.$$

Let $u_0 \in L^2(\Omega) \setminus \{0\}$. Consider the minimal time control problems:

$$(TP)^1_M : \ \ T^1_M \equiv \min_{v \in \mathcal{U}^1_M} \left\{ t > 0 : \ \ e^{t\Delta}u_0 + \int_0^t e^{(t-s)\Delta}(\chi_\omega v)\, ds = 0 \right\}$$

and

$$(TP)^2_M : \ \ T^2_M \equiv \min_{g \in \mathcal{U}^2_M} \{t > 0 : \ \ u(x,t;g) = 0 \ \text{ for a.e. } \ x \in \Omega\},$$

where $u(\cdot,\cdot;g)$ is the solution to

$$\begin{cases} \partial_t u - \Delta u = 0, & \text{in } \Omega \times \mathbb{R}^+, \\ u = g\chi_\Gamma, & \text{on } \partial\Omega \times \mathbb{R}^+, \\ u(0) = u_0, & \text{in } \Omega. \end{cases} \tag{25}$$

Any solution of $(TP)^i_M$, $i = 1, 2$, is called a minimal time control to this problem. According to Theorem 1 and Theorem 3.3 in [9], problem $(TP)^1_M$ has solutions. By Theorem 2, using the same arguments as those in the proof of Theorem 3.3 in [9], we can verify that there is $g \in \mathcal{U}^2_M$ such that for some $t > 0$, $u(x,t;g) = 0$ for a.e. $x \in \Omega$.

**Lemma 4** *Problem $(TP)^2_M$ has solutions.*

*Proof* Let $\{t_n\}_{n\geq 1}$, with $t_n \searrow T^2_M$, and $g_n \in \mathcal{U}^2_M$ be such that $u(x, t_n; g_n) = 0$ over $\Omega$. Hence, on a subsequence,

$$g_n \longrightarrow g^* \ \text{ weakly star in } \ L^\infty(\partial\Omega \times (0, t_1)). \tag{26}$$

It suffices to show that

$$u_n(x, t_n) \equiv u(x, t_n; g_n) \longrightarrow u^*(x, T^2_M) \equiv u(x, T^2_M; g^*), \ \text{ for all } x \in \Omega. \tag{27}$$

For this purpose, let $G(x, y, t)$ be the Green's function for $\triangle - \partial_t$ in $\Omega \times \mathbb{R}$ with zero lateral Dirichlet boundary condition. Reference [4, Theorems 1.3 and 1.4] and [4, p. 643] show that for $g \in \mathcal{U}^2_M$ and $(x, t) \in \Omega \times (0, T)$,

$$u(x, t; g) = e^{t\triangle} u_0 - \int_0^t \int_{\partial\Omega} \tfrac{\partial G}{\partial \nu_q}(x, q, t-s)\, \chi_\Gamma(q, s) g(q, s)\, d\sigma_q ds \tag{28}$$

and

$$\int_0^T \int_{\partial\Omega} |\tfrac{\partial G}{\partial \nu_q}(x, q, \tau)|^2\, d\sigma_q d\tau < +\infty, \ \text{ when } x \in \Omega, \ T > 0. \tag{29}$$

Also, by standard interior parabolic regularity there is $N = N(n, \epsilon)$ with

$$|u(x, t; g) - u(x, s; g)| \leq N|t-s| \left( \|g\|_{L^\infty(\partial\Omega\times(0,T))} + \|u_0\|_{L^2(\Omega)} \right) \tag{30}$$

when $d(x, \partial\Omega) > \sqrt{\epsilon}$ and $t > s \geq \epsilon$. Now, when $x \in \Omega$ with $d(x, \partial\Omega) > \sqrt{\epsilon}$, it holds that

$$|u_n(x, t_n) - u^*(x, T^2_M)| \leq |u_n(x, t_n) - u_n(x, T^2_M)| + |u_n(x, T^2_M) - u^*(x, T^2_M)|.$$

This, along with (26), (28), (29) and (30) indicates that (27) holds for all $x \in \Omega$ with $d(x, \partial\Omega) > \sqrt{\epsilon}$. Since $\epsilon > 0$ is arbitrary, (27) follows at once. □

Now, we can use the same methods as those in [11], as well as in Lemma 4, to get the following consequences of Theorems 1 and 2 respectively.

**Corollary 2** *Problem $(TP)^1_M$ has the bang-bang property: any minimal time control $v$ satisfies that $|v(x, t)| = M$ for a.e. $(x, t) \in \omega \times (0, T^1_M)$. Consequently, this problem has a unique minimal time control.*

**Corollary 3** *The problem $(TP)^2_M$ has the bang-bang property: any minimal time boundary control $g$ satisfies that $|g(x, t)| = M$ for a.e. $(x, t) \in \Gamma \times (0, T^2_M)$. Consequently, this problem has a unique minimal time control.*

### 3.2 Definition of the Minimal Norm Control Problems and Main Results

In this section, we apply Theorems 1 and 2 to get the bang-bang property for the minimal norm control problems; they are stated as follows. Let $\mathcal{D}$ and $\mathcal{J}$ be the subsets given at the beginning of this section. Let $u_0 \in L^2(\Omega)$, we define two control constraint sets as follows:

$$\mathcal{V}_{\mathcal{D}} = \left\{v \in L^\infty(\Omega \times (0,T)) : u(T; u_0, v) = 0\right\}$$

and

$$\mathcal{V}_{\mathcal{J}} = \left\{g \in L^\infty(\partial\Omega \times (0,T)) :\ u(T; u_0, g) = 0\right\}.$$

Consider the minimal norm control problems:

$$(NP)_{\mathcal{D}} : \quad M_{\mathcal{D}} \equiv \min\left\{\|v\|_{L^\infty(\Omega\times(0,T))} :\ \ v \in \mathcal{V}_{\mathcal{D}}\right\}$$

and

$$(NP)_{\mathcal{J}} : \quad M_{\mathcal{J}} \equiv \min\left\{\|g\|_{L^\infty(\partial\Omega\times(0,T))} :\ \ g \in \mathcal{V}_{\mathcal{J}}\right\}.$$

Any solution of $(NP)_{\mathcal{D}}$ (or $(NP)_{\mathcal{J}}$) is called a minimal norm control to this problem. According to Corollary 1, the sets $\mathcal{V}_{\mathcal{D}}$ and $\mathcal{V}_{\mathcal{J}}$ are not empty. Since $\mathcal{V}_{\mathcal{D}}$ is not empty, it follows from the standard arguments that Problem $(NP)_{\mathcal{D}}$ has solutions. Because $\mathcal{V}_{\mathcal{J}}$ is not empty, by using the similar arguments as those in the proof of Lemma 4, we can justify that Problem $(NP)_{\mathcal{J}}$ has solutions.

We can use the same methods as those in [8] to get the following consequences of Theorem 1 and Theorem 2 respectively:

**Corollary 4** *Problem $(NP)_{\mathcal{D}}$ has the bang-bang property: any minimal norm control $v$ satisfies that $|v(x,t)| = M_{\mathcal{D}}$ for a.e. $(x,t) \in \mathcal{D}$. Consequently, this problem has a unique minimal norm control.*

**Corollary 5** *The problem $(NP)_{\mathcal{J}}$ has the bang-bang property: any minimal norm boundary-control $g$ satisfies that $|g(x,t)| = M_{\mathcal{J}}$ for a.e. $(x,t) \in \mathcal{J}$. Consequently, this problem has a unique minimal norm control.*

## 4 Open Problems

In this section we will establish the heat equation with similar conditions to what we studied before, but in this case we will require it to verify other type of boundary conditions instead of Dirichlet boundary conditions.

Let $\Omega$ be a bounded Lipschitz domain in $\mathbb{R}^n$ and consider the following heat equation,

$$\begin{cases} \partial_t u - \Delta u = 0, & \text{in } \Omega \times (0,1), \\ \frac{\partial}{\partial \nu} u = 0, & \text{on } \partial\Omega \times (0,T), \\ u(0) = u_0, & \text{in } \Omega, \end{cases} \tag{31}$$

with Neumann boundary condition and

$$\begin{cases} \partial_t u - \Delta u = 0, & \text{in } \Omega \times (0,1), \\ \frac{\partial}{\partial \nu} u + \alpha u = 0, & \text{on } \partial\Omega \times (0,T), \\ u(0) = u_0, & \text{in } \Omega, \end{cases} \tag{32}$$

with Robin boundary condition, where $\alpha \in \mathbb{R}$ and $u_0$ in $L^2(\Omega)$.

We proved two observability inequalities (Theorems 1 and 2) for these kind of equations over measurable sets with Dirichlet boundary conditions, but if we change that condition to now use Neumann or Robin conditions, would we be able to prove some similar observability inequalities? And, if that's the case, could we apply them to prove some bang-bang properties?

The idea of facing these questions is to spread our mathematical knowledge about this kind of problems and also to discover new interesting ways or limitations in the techniques we are used to working with. It could also be physically interesting because of the physical meaning of these new boundary conditions, as we will see now.

The Dirichlet boundary condition states that we have a constant temperature at the boundary. This can be considered as a model of an ideal cooler in a good contact having infinitely large thermal conductivity.

With the Neumann boundary condition case for the heat flow, we can say that we have a constant heat flux at the boundary or that it corresponds to a perfectly insulated boundary. If the flux is equal to zero, the boundary condition describes the ideal heat insulator with the heat diffusion. For the Laplace equation and drum modes, we could think this corresponds to allowing the boundary to flap up and down but not move otherwise.

Finally, the Robin boundary condition is the mathematical formulation of Newton's law of cooling where the heat transfer coefficient $\alpha$ is utilized. The heat transfer coefficient is determined by details of the interface structure (sharpness, geometry) between two media. This law describes the boundary between metals and gas quite well and is good for the convective heat transfer.

## References

1. Apraiz, J., Escauriaza, L.: Null-control and measurable sets. ESAIM Control Optim. Calc. Var. **19**, 239–254 (2013)
2. Apraiz, J., Escauriaza, L., Wang, G., Zhang, C.: Observability inequalities and measurable sets. J. Eur. Math. Soc. **16**, 2433–2475 (2014)
3. Brown, R.M.: The method of layer potentials for the heat equation in Lipschitz cylinders. Am. J. Math. **111**, 339–379 (1989)
4. Fabes, E.B., Salsa, S.: Estimates of caloric measure and the initial-Dirichlet problem for the heat equation in Lipschitz cylinders. Trans. Am. Math. Soc. **279**, 635–650 (1983)
5. Fursikov, A.V., Yu Imanuvilov, O.: Controllability of Evolution Equations. Lecture Notes Series, vol. 34. Seoul National University, Seoul (1996)
6. Lebeau, G., Robbiano, L.: Contrôle exact de l'équation de la chaleur. Commun. Partial Differ. Equ. **20**, 335–356 (1995)
7. Lions, J.L.: Optimal Control for Systems Governed by Partial Differential Equations. Springer, Berlin (1971)
8. Phung, K.D., Wang, G.: An observability estimate for parabolic equations from a measurable set in time and its applications. J. Eur. Math. Soc. **15**, 681–703 (2013)
9. Phung, K.D., Wang, G., Zhang, X.: On the existence of time optimal controls for linear evolution equations. Discrete Contin. Dynam. Syst. Ser. B **8**, 925–941 (2007)
10. Vessella, S.: A continuous dependence result in the analytic continuation problem. Forum Math. **11**, 695–703 (1999)
11. Wang, G.: $L^\infty$-null controllability for the heat equation and its consequences for the time optimal control problem. SIAM J. Control Optim. **47**, 1701–1720 (2008)
12. Zhang, C.: An observability estimate for the heat equation from a product of two measurable sets. J. Math. Anal. Appl. **396**, 7–12 (2012)

# Optimal Design of Piezoelectric Microactuators: Linear vs Non-linear Modeling

David Ruiz, José Carlos Bellido, and Alberto Donoso

**Abstract** The main point of this work is the comparison between linear and geometrically non-linear elasticity modeling in the field of piezoelectric actuators fabricated at the micro-scale. Manufacturing limitations such as non-symmetrical lamination of the structure or minimum length scale are taken into account during the optimization process. The robust approach implemented in the problem also reduces the sensitivity of the designs to small manufacturing errors.

**Keywords** Piezoelectric actuators · Topology optimization · Electrode profile · Heterogeneous bimorph · Large displacements

## 1 Introduction

The conceptual tool of topology optimization was initially designed for structural design, nevertheless, nowadays its use is not restricted only to this purpose. Some fields where its contribution has been crucial are the design of compliant mechanisms [5, 25], dynamics [4, 19], band gaps [7, 29] and metamaterials [30], amongst others.

The topology optimization method has played an important role in the optimal design of MEMS (micro-electro-mechanical systems), where the size of the devices typically is smaller than 1 mm. In [25] is presented the optimal design of compliant mechanisms based on the topology optimization method, where these mechanisms were fabricated at macro- and micro-scale. Concerning piezoelectric effect, [33] suggested a procedure based on topology optimization and homogenization methods to optimize unit cells for piezocomposites. Regarding thermal and electrothermal actuators, [26] and [27] optimized microdevices composed of one and two materials,

D. Ruiz (✉) · J. C. Bellido · A. Donoso
Departamento de Matemáticas, ETSII, Universidad de Castilla-La Mancha, Ciudad Real, Spain
e-mail: David.Ruiz@uclm.es

© Springer Nature Switzerland AG 2019

J. L. García Guirao et al. (eds.), *Recent Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 18,
https://doi.org/10.1007/978-3-030-00341-8_2

respectively. A methodology of the design of MEMS under stochastic loads and boundary conditions is presented in [16].

In the last years many authors have also applied the topology optimization method to piezoelectric devices. A pioneering work is [31], where the authors designed the unit cell of 1–3 composites for hydrophone applications. In the field of piezoelectric actuators, [32] presented a method to design in-plane actuators by optimizing the host structure, but fixing the piezoelectric material layer. Kögl and Silva [12] considered the optimization of piezoelectric layer together with the polarization based on a three layer model. Carbonari et al. [3] and Luo et al. [15] optimized simultaneously the host structure and the piezoelectric distribution. The inclusion of a third variable, the spatial distribution of the control voltage (related to the polarization of the piezoelectric layers) in the optimization problem, was presented in [8] and improved in [9] by introducing an interpolation scheme in the tri-level actuation voltage term. New results are presented in [10] and [11] for in-plane and out-of-plane piezoelectric transducers, respectively. In [17] is optimized at the same time the structure and the piezoelectric profile in the context of actuators.

Recently, in prior works, [23] and [22] presented a systematic procedure to design static microtransducers and modal filters by optimizing simultaneously the structure layout and the polarization profile. The three layer model considers that both piezoelectric layers are perfectly bonded to the top and bottom of the host structure. Either in-phase or out-of-phase polarization of the two piezoelectric layers makes the structure move in-plane or out-of-plane, respectively. As shown in [13], at the micro-scale due to limitations in the fabrication techniques only one piezoelectric film can be deposited on the top of the surface. This is an important issue for actuators, since with only one piezoelectric layer the device moves in the in-plane and out-of-plane directions. This issue is overcome in [24], where unimorph piezoelectric microgrippers working in-plane are designed by optimizing simultaneously the host structure and the polarization profile of the piezoelectric layer.

Ruiz and Sigmund [21] improved the latter by using a geometrically non-linear model that is able to model large displacements. With regard to the topic of geometrical non-linearities in topology optimization, [2] was the pioneering work, using the total Lagrangian formulation. In [1] and [20] were presented the optimal design of compliant mechanisms taking into account this non-linearity. The robust design of compliant mechanisms that undergo large displacement was included in [14], adding random variations that model possible geometry errors. Wang et al. [36] suggests an interpolation scheme for fictitious domain and topology optimization approaches. The objective of the present paper is to present a comparison between the behavior of piezoelectric actuators when they are modeled by using linear elasticity and geometric non-linearities. The main novelty introduced in this work is the dependence of the external force (the piezoelectric one) on the design variables. Vertical displacements produced by a nonsymmetrical laminate are suppressed at some points of interest and a robust approach is used with the objective of controlling the minimum length scale and minimizing the effect of the small manufacturing errors.

The paper is organized as follows. In Sect. 2 the discrete formulation of the problem is described, including the finite element modeling and the robust approach implemented. Section 3 is focused on the numerical algorithm used to solve the problem. Examples with different boundary conditions are shown in Sect. 4. Section 5 is devoted to the comparison between the two different elasticity modeling, showing the importance of using the most suitable model. Finally, in Sect. 6 the conclusions of this work are presented.

## 2 Topology Optimization of Piezoelectric Microactuators

As design domain $\Omega$ we consider a rectangular plate clamped at its left side $\Gamma_u$, as represented in Fig. 1(top). A piezoelectric layer, that is sandwiched between two piezoelectric films, is perfectly bonded to the top surface of the host structure. This configuration is shown in Fig. 1(bottom).

When an input voltage $V_{in}$ is applied to the electrodes, the electric field generated polarizes the piezoelectric layer. Thanks to the direct piezoelectric effect, the resulting force deforms the host structure. Due to a non-symmetrical lamination of the device, an out-of-plane displacement distorts the genuine in-plane behavior of the gripper. In such a kind of lamination the piezoelectric effect is divided into

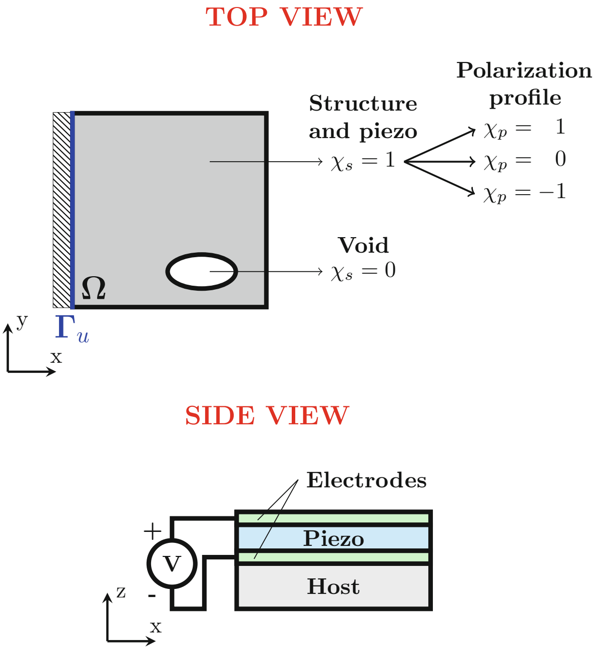


**Fig. 1** Top and side view of the piezoelectric device

two components: an initial strain that makes the structure expand or contract and a flexural moment that produces the aforementioned vertical displacement.

This manufacturing limitation leads us to try to suppress this vertical deformation (in the $z$-axis) at some points of interest. The output of the actuator is modeled by a spring of stiffness $k_{out}$ (that depends on the application). The points where the vertical displacement is suppressed and the output port will be defined for different examples in Sect. 4.

The goal of the problem is the maximization of the displacement at the output port, while the vertical displacement along the $z$-axis at some points of interest is suppressed as much as possible. This suppression is controlled by adding a constraint which relates the optimized and the canceled displacements. Finally, a volume constraint is used to control the amount of material used.

The optimization problem proposed involves two different design variables. The first one is a characteristic function $\chi_s \in \{0, 1\}$, that represents the structure layout and piezo as well ($\chi_s = 1$) and void ($\chi_s = 0$). The second variable is also a discrete function such that $\chi_p = \{-1, 0, 1\}$, meaning negative, null or positive polarity.

The role of the latter is crucial in this problem for two reasons. The first one is that only the part of the host structure covered by electrode is electrically affected and then subjected to the piezoelectric force. The second concerns the piezoelectric force: this variable is related with the sign of the force. In other words, $\chi_p$ controls which parts of the structure works under compression or traction, and this is key to suppress undesired displacements.

In this work two different elasticity models are used, for small and large in-plane displacements. Nevertheless, the out-of-plane displacements are expected to be small, and then it can be studied separately from the in-plane one (decoupled problems). It will be shown in the examples that the fact of suppressing the vertical displacement at the output port makes out-of-plane displacements small in general. This assumption leads us to model out-of-plane displacements in both cases with linear elasticity theory.

As usual, topology optimization problems lacks of classical solution and the characteristic functions $\chi_s$ and $\chi_p$ must be relaxed into density variables $\rho_s$ and $\rho_p$. The well-known SIMP (solid isotropic material with penalization) is used for this purpose. The Young's modulus $E_e$ of each element depends on the element density as follows:

$$E_e = (\bar{\rho}_{se})^p (E_0 - E_{min}) + E_{min}, \tag{1}$$

where $E_0$ is the Young's modulus of the base material, $E_{min} > 0$ is a small value used to avoid singularities of the stiffness matrix and $p$ is the penalization exponent (typically $p = 3$). Afterwards, the domain is discretized in $n_e$ finite elements with two variables per element.

The discrete formulation written as a topology optimization [21, 24] problem becomes:

$$\max_{\boldsymbol{\rho}_s, \boldsymbol{\rho}_p} : \quad u_1(\boldsymbol{\rho}_s, \boldsymbol{\rho}_p)$$

s.t.: In-plane and out-of-plane equilibrium equations:

$$\begin{cases} \mathbf{r}_{ip}(\boldsymbol{\rho}_s, \boldsymbol{\rho}_p, \mathbf{U}_{ip}) = \mathbf{0} \\ \mathbf{r}_{op}(\boldsymbol{\rho}_s, \boldsymbol{\rho}_p, \mathbf{U}_{op}) = \mathbf{0} \end{cases}$$

Displacements:

$$\begin{cases} u_1 = \mathbf{L}_1^T \mathbf{U}_{ip} \\ u_j = \mathbf{L}_j^T \mathbf{U}_{op}, \ \ j = 1, \ldots, n_c \end{cases}$$

Constraints:

$$\begin{aligned} &\left(\frac{u_j}{u_1}\right)^2 - \varepsilon_d^2 \leq 0, \ \ j = 1, \ldots, n_c \\ &\frac{\mathbf{1}^T \boldsymbol{\rho}_s}{V} - 1 \leq 0 \\ &\boldsymbol{\rho}_s \in [\mathbf{0}, \mathbf{1}] \\ &\boldsymbol{\rho}_p \in [-\mathbf{1}, \mathbf{1}], \end{aligned}$$

where $\boldsymbol{\rho}_s$ and $\boldsymbol{\rho}_p$ represent the structure layout and the polarization profile, respectively, $\mathbf{L}_1$ and $\mathbf{L}_j$ are vectors of zeros with 1 in the degree of freedom of interest, $u_1$ is the in-plane displacement to be optimized, $u_j$ is the out-of-plane displacements to be suppressed, $n_c$ is the number of points where the vertical displacement is suppressed, $\varepsilon_d$ is a small value that relates the suppressed and the optimized displacements, $V$ is the maximum volume allowed, $\mathbf{r}_{ip}$ and $\mathbf{r}_{op}$ are the residual vectors and finally, subscripts $ip$ and $op$ stand for the in-plane and out-of-plane case, respectively. The residual vectors are defined as the difference between the external forces (in our case the piezoelectric one) and the internal forces:

$$\begin{aligned} \mathbf{r}_{ip} &= \mathbf{f}_{ip}^{piezo} - \mathbf{f}_{ip}^{int} \\ \mathbf{r}_{op} &= \mathbf{f}_{op}^{piezo} - \mathbf{f}_{op}^{int}. \end{aligned} \tag{2}$$

**Linear Elasticity** For this particular case the dependence of the strain on the displacements is linear, and the elastic energy density can be defined as:

$$\phi = \frac{1}{2}\lambda\varepsilon_{kk}^2 + \mu\varepsilon_{ij}\varepsilon_{ij}, \tag{3}$$

where $\lambda$ and $\mu$ are the Lamé parameters and $\boldsymbol{\varepsilon}$ is the strain tensor. The stress tensor $\boldsymbol{\sigma}$ can be obtained by differentiating Eq. (3) with respect to the strain tensor. The discretization of the problem with finite elements is straightforward and the interested reader is referred to [18]. Finally, the residual vectors can be written as follows:

$$\begin{aligned}\mathbf{r}_{ip} &= \mathbf{f}_{ip}^{piezo}(\boldsymbol{\rho}_s, \boldsymbol{\rho}_p) - \big(\mathbf{K}_{ip}(\boldsymbol{\rho}_s) + \mathbf{1}_{out}k_{out}\big)\mathbf{U}_{ip} = \mathbf{0}\\ \mathbf{r}_{op} &= \mathbf{f}_{op}^{piezo}(\boldsymbol{\rho}_s, \boldsymbol{\rho}_p) - \mathbf{K}_{op}(\boldsymbol{\rho}_s)\mathbf{U}_{op} = \mathbf{0},\end{aligned}$$

where $\mathbf{1}_{out}$ is a zero matrix with 1 in the degree of freedom of the output port and $\mathbf{K}_{ip}$ and $\mathbf{K}_{op}$ are the stiffness matrices. The piezoelectric force depends on both design variables, $\boldsymbol{\rho}_p$ defines the sign of the force and $\boldsymbol{\rho}_s$ represents the fact that void areas are not electrically affected.

The piezoelectric force modeling can be found in [6] and it is not presented here. In addition, the same powerlaw dependence presented in Eq. (1) $\mathbf{R} = \boldsymbol{\rho}_s^p$ is used for the piezoelectric force generation. It is easy to see that the value of this interpolation function is 0 for void regions, and 1 for the solid ones. This interpolation scheme was presented in [22].

**Non-linear Elasticity** In this case the displacements of the in-plane case are supposed to be large and a linear elasticity model is not appropriate to represent the behavior of the device. Instead, we will use a geometrically non-linear model, taking the assumption of large displacements but small strains. The Saint-Venant-Kichhoff model is used to represent this behavior. The expression for the stored elastic energy density is:

$$\phi = \frac{1}{2}\lambda E_{kk}^2 + \mu E_{ij}E_{ij}, \tag{4}$$

where $\mathbf{E}$ is the Green-Lagrange strain tensor, that can be expressed as follows:

$$\mathbf{E} = \frac{1}{2}(\mathbf{F}\mathbf{F}^T - \mathbf{I}),$$

being $\mathbf{F}$ the deformation gradient tensor defined as $\mathbf{F} = \mathbf{I} + \partial\mathbf{u}/\partial\mathbf{y}$. The second Piola-Kirchhoff stress tensor $\mathbf{S}$ is defined as the derivative of the stored elastic energy density presented in Eq. (4) with respect to the Green-Lagrange strain tensor.

Finally, the expression for the internal forces presented in Eq. (2) is:

$$\mathbf{f}_{ip}^{int,e} = \frac{\partial \int_\Omega \phi_e \mathrm{d}\Omega}{\mathbf{u}_{ip}}.$$

where $\mathbf{u}_{ip}$ is the elemental displacement vector for element $e$.

Taking into account that the out-of-plane displacements are supposed to be small enough to be modeled with a linear elasticity model, the residual vector can be expressed as follows:

$$\begin{aligned}\mathbf{r}_{ip} &= \mathbf{f}_{ip}^{piezo} - \int_\Omega \mathbf{B}_{ip}(\mathbf{U}_{ip})\mathbf{S}\mathrm{d}\Omega\\ \mathbf{r}_{op} &= \mathbf{f}_{op}^{piezo} - \mathbf{K}_{op}\mathbf{U}_{op},\end{aligned}$$

where $\mathbf{B}_{ip}$ is the in-plane non-linear strain displacement matrix. The residual of the out-of-plane equation depends linearly on the displacements, then to obtain $\mathbf{U}_{op}$ we must solve a system of linear equations. For the in-plane case the system of equations is non-linear and an iterative method must be used. For this problem we use the Newton-Raphson method, where the non-linear system to solve is:

$$\mathbf{K}_t \boldsymbol{\Delta}\mathbf{U}_{ip} = \mathbf{r}_{ip},$$

where $\mathbf{K}_t$ is the tangent stiffness matrix defined as:

$$\mathbf{K}_t = -\frac{\partial \mathbf{r}_{ip}}{\partial \mathbf{U}_{ip}}$$

and the nodal displacement vector is updated by $\mathbf{U}_{ip} = \mathbf{U}_{ip} + \boldsymbol{\Delta}\mathbf{U}_{ip}$. The detailed computations of the nodal force vectors, the strain displacement matrix and the tangent stiffness matrix can be found in [37].

The interpolation presented in [36] for the elastic stored energy density is implemented in this work. The objective of this scheme is to alleviate the issue of distorted and ill-converged void region mesh. This paper suggests basing the analysis of the solid region on the non-linear analysis and on the linear analysis for the void one, thereby eliminating mesh distortion and ill-convergence issues in the low density domain. The interpolation scheme for element $e$ is:

$$\phi_e(\mathbf{u}_{ip}) = [\phi(\gamma_e \mathbf{u}_{ip}) - \phi_L(\gamma_e \mathbf{u}_{ip}) + \phi_L(\mathbf{u}_{ip})]E_e,$$

where $\phi(\cdot)$ is the stored elastic energy density, $\phi_L(\cdot)$ is the stored elastic energy density under small deformations, both with unit Young's modulus, and finally, $\gamma_e$ is the interpolation factor. In order to differ between solid and void regions in intermediate steps of the optimization process, a smoothed Heaviside projection is used:

$$\gamma_e = \frac{\tanh(\beta_1 \rho_0) + \tanh\big(\beta_1(\bar{\rho}_{se}^{p} - \rho_0)\big)}{\tanh(\beta_1 \rho_0) + \tanh\big(\beta_1(1 - \rho_0)\big)},$$

being $\beta_1$ the parameter that models the smoothness of the projection, $\rho_0$ the threshold and $\bar{\boldsymbol{\rho}}_s$ the physical density, which will be introduced in the next section.

## 2.1 Robust Formulation

Having in mind the manufacturability of the designs, a robust formulation has been used with two goals: the first is the minimization of the objective to small manufacturing errors, the second one is the control of the minimum length scale in both, void and solid, avoiding the appearance of hinges. This approach was presented in [28] and [35].

This formulation consists in the use of three different projections: eroded, intermediate and dilated. The expression for a projection is:

$$\bar{\boldsymbol{\rho}}_s = \frac{\tanh(\beta_0\eta) + \tanh\big(\beta_0(\tilde{\boldsymbol{\rho}}_s - \eta)\big)}{\tanh(\beta_0\eta) + \tanh\big(\beta_0(1-\eta)\big)},$$

where $\beta_0$ is a tunable parameter that represents the smoothness of the heaviside function, $\eta$ is the threshold which can take values between 0 and 1, and $\tilde{\boldsymbol{\rho}}_s$ is the filtered density that is expressed as follows:

$$\rho_{se} = \frac{\sum\limits_{i \in N_e} \omega(\mathbf{x}_i) v_i \rho_{si}}{\sum\limits_{i \in N_e} \omega(\mathbf{x}_i) v_i},$$

where $\mathbf{x}_i$ is the center of element $i$, $v_i$ the volume of element $i$, $N_e$ the neighborhood of element $e$ within a certain filter radius $r$ specified by $N_e = \{i|\ ||\mathbf{x}_i - \mathbf{x}_e|| \le r\}$, and $\omega(\mathbf{x}_i) = r - ||\mathbf{x}_i - \mathbf{x}_e||$.

From now on superscript $q$ stands for the projection, meaning $e$ erode, $d$ dilate and $i$ intermediate. The robust topology optimization problem is written as:

$$\max_{\boldsymbol{\rho}_s, \boldsymbol{\rho}_p} : \min_{q=\{e,i,d\}} \{u_1^q(\bar{\boldsymbol{\rho}}_s, \boldsymbol{\rho}_p)\}$$

s.t.: In-plane and out-of-plane equilibrium equations:

$$\begin{cases} \mathbf{r}_{ip}^q(\bar{\boldsymbol{\rho}}_s^q, \boldsymbol{\rho}_p, \mathbf{U}_{ip}^q) = \mathbf{0} \\ \mathbf{r}_{op}^q(\bar{\boldsymbol{\rho}}_s^q, \boldsymbol{\rho}_p, \mathbf{U}_{op}^q) = \mathbf{0} \end{cases}$$

Displacements:

$$\begin{cases} u_1^q = \mathbf{L}_1^T \mathbf{U}_{ip}^q \\ u_j^q = \mathbf{L}_j^T \mathbf{U}_{op}^q, \ j = 1, \ldots, n_c \end{cases}$$

Constraints:

$$\begin{aligned} &\left(\frac{u_j^q}{u_1^q}\right) - \varepsilon_d^2 \le 0, \ j = 1, \ldots, n_c \\ &\frac{\mathbf{1}^T \bar{\boldsymbol{\rho}}_s^d}{V_d^*} - 1 \le 0 \\ &\boldsymbol{\rho}_s \in [\mathbf{0}, \mathbf{1}] \\ &\boldsymbol{\rho}_p \in [-\mathbf{1}, \mathbf{1}] \\ &q \equiv \{e, i, d\}, \end{aligned}$$

where $V_d^*$ is the maximum volume for the dilated design. This value is computed at the beginning of the optimization process and then is update every 20 iterations following the next equation:

$$V_d^* = \frac{V^*}{V_i} V_d,$$

being $V^*$ the maximum volume prescribed for the intermediate design and $V_i$ and $V_d$ the volume of the intermediate and dilated designs, respectively.

Since the max-min function is not differentiable, the problem is rewritten by using the so-called bound formulation [19]:

$$\max_{\boldsymbol{\rho}_s, \boldsymbol{\rho}_p} : \quad \beta$$

s.t.: In-plane and out-of-plane equilibrium equations:

$$\begin{cases} \mathbf{r}_{ip}^q(\bar{\boldsymbol{\rho}}_s^q, \boldsymbol{\rho}_p, \mathbf{U}_{ip}^q) = \mathbf{0} \\ \mathbf{r}_{op}^q(\bar{\boldsymbol{\rho}}_s^q, \boldsymbol{\rho}_p, \mathbf{U}_{op}^q) = \mathbf{0} \end{cases}$$

Displacements:

$$\begin{cases} u_1^q = \mathbf{L}_1^T \mathbf{U}_{ip}^q \\ u_j^q = \mathbf{L}_j^T \mathbf{U}_{op}^q, \ \ j = 1, \ldots, n_c \end{cases}$$

Constraints:

$$\begin{aligned} & -u_1^q \leq \beta \\ & \left( \frac{u_j^q}{u_1^q} \right) - \varepsilon_d^2 \leq 0, \ \ j = 1, \ldots, n_c \\ & \frac{\mathbf{1}^T \bar{\boldsymbol{\rho}}_s^d}{V_d^*} - 1 \leq 0 \\ & \boldsymbol{\rho}_s \in [\mathbf{0}, \mathbf{1}] \\ & \boldsymbol{\rho}_p \in [\mathbf{-1}, \mathbf{1}] \\ & q \equiv \{e, i, d\}, \end{aligned}$$

being $\beta$ a new variable that does not depend on the design variables $\boldsymbol{\rho}_s$ and $\boldsymbol{\rho}_p$ and resolves the issue of the non-differentiability of the max-min function.

## 3 Numerical Implementation

The well-known MMA (Method of Moving Asymptotes, [34]) is used to solve the optimization problem. This algorithm is included inside the group of descent methods, which require information about the objective function, constraints and the sensitivities of both. The adjoint method is used to compute the sensitivities with respect to both design variables, $\boldsymbol{\rho}_s$ and $\boldsymbol{\rho}_p$, but these computations are not the objective of this work, and this is why they are not included here.

The complete process algorithm in a pseudo code looks like:

**Pre-process**

1. Define dimensions, boundary conditions and material properties of the plate. Input parameters like $V_{in}$, $\varepsilon_d$ and $V$ must be fixed.
2. Initialize both design variables, $\boldsymbol{\rho}_s$ with $\rho_{se} = V$ and $\boldsymbol{\rho}_p$ with $\rho_{pe} = 1$.

**Optimization Algorithm**

3. Compute the physical densities $\bar{\boldsymbol{\rho}}_s$ by filtering the structural density $\boldsymbol{\rho}_s$ and then projecting with three different thresholds.
4. Solve the finite element problems for the three different physical densities:
   - For the in-plane case.
   - For the out-of-plane case.
5. Extract the displacements $u_1^q$ and $u_j^q$ from the global displacements vectors and compute the value of the objective function and constraints.
6. Compute the sensitivities of the objective function and constraints.
7. Update design variables $\boldsymbol{\rho}_s$ and $\boldsymbol{\rho}_p$ by using MMA.
8. Until convergence, update parameters $\beta_0$ and $V_d^*$ and go back to step 3.

**Post-process**

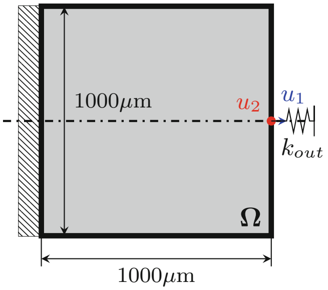
9. Plot the optimized designs.

## 4 Examples

In this section three actuators with different boundary conditions will be presented to show the validity of our method. For the sake of brevity, all the examples have been obtained using a geometrically non-linear modeling.

The materials used for the three different actuators remain constant. The host layer with a thickness of $t = 5\,\mu\text{m}$ is made of silicon with Young's modulus $E_0 = 130\,\text{GPa}$ and Poisson's coefficient $\nu = 0.3$. The piezoelectric film with thickness $t_p = 0.5\,\mu\text{m}$ is made of PZT with $E_0 = 67\,\text{GPa}$, $\nu = 0.3$ and $d_{31} = 190\,\text{pm}$. The stiffness of the void is chosen to be small compared to the stiffness of the solid one, $E_{min} = 10^{-9}E_0$. The minimum length scale is set to $22.5\,\mu\text{m}$ with $\eta = 0.3$ and

**Fig. 2** Boundary conditions for the first example



$\beta_0 = 1$ and doubling its value each 50 iterations. Concerning the interpolation scheme of the elastic stored energy density, the values of the parameters of the projection are fixed to $\beta_1 = 500$ and $\rho_0 = 0.01$. The maximum volume fraction of the designs is $V_0 = 40\%$. Finally, the values of the inputs are $V_{in} = 1000$ V, $k_{in} = 1000$ N/m and $\varepsilon_d = 5\%$.

The design domain $\Omega$ will change for the different examples and then its dimensions will be shown in each particular case.

## *4.1 Maximizing Displacement in Horizontal Direction*

A square plate-type structure clamped at its left side is considered as design domain $\Omega$. The objective is the maximization of the displacement $u_1$ while $u_2$ is suppressed. Boundary conditions and dimensions of the plate are shown if Fig. 2.

The result of the optimization process is shown in Fig. 3. For the sake of brevity, only the intermediate projection (blueprint design) is presented. The structural layout ($\bar{\boldsymbol{\rho}}_s$) is shown in Fig. 3(left), where black and white mean solid and void areas, respectively. There is no gray areas (microstructure) in the optimized design, meaning that the projection method is working properly. The minimum length scale imposed on both, solid and void, is represented with the black circle. The electrode profile ($\boldsymbol{\rho}_p$) is shown in Fig. 3(right), where cyan and orange represent parts of the structure with opposite polarization. The in-plane displacement is $u_1 = 122$ µm, while the out-of-plane $u_2$ is smaller than the 5% of $u_1$. As can be seen, the in-plane displacement is bigger than the 5% of the length of the plate. This value remarks the importance and is the motivation of using a geometrically non-linear modeling.

## *4.2 Maximizing Displacement in Horizontal Direction Including Void Passive Area*

There are two differences with respect to the previous example: the sense of the displacement to be optimized and the inclusion of a void passive area in the middle of the domain. Boundary conditions are presented in Fig. 4.

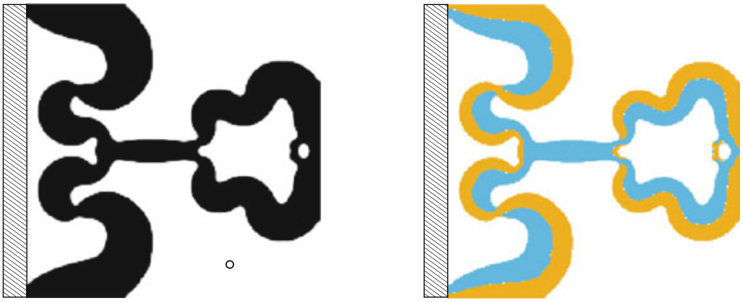


**Fig. 3** Structural layout (left) and polarization profile (right) for the first boundary conditions. The black circle indicates the minimum length scale

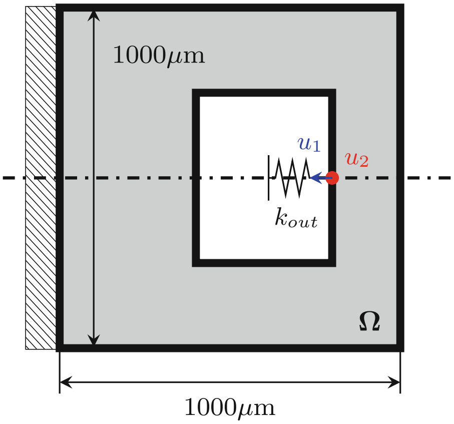


**Fig. 4** Boundary conditions for the second example

The optimized design are presented in Fig. 5. The structural layout is shown in Fig. 5(left) and the polarization profile in Fig. 5(right). In this case the in-plane displacement is $u_1 = 104\,\mu\text{m}$.

## *4.3 Maximizing Displacement in Vertical Direction Including Void Passive Area*

Boundary conditions and passive area are shown in Fig. 6. This configuration defines a particular case of actuators: grippers. Interested reader is referred to [24] and [21], where this problem is fully described. As can be seen one passive area has been added: a void passive area necessary to have enough space to grab objects.

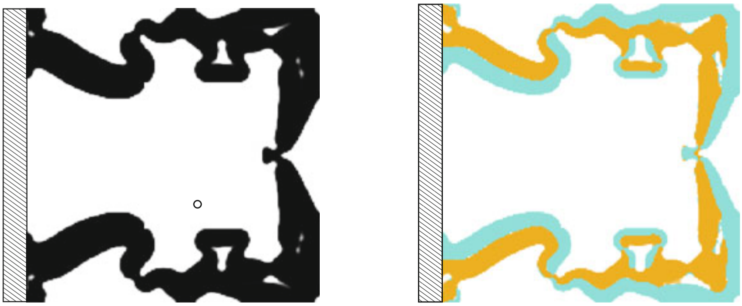


**Fig. 5** Structural layout (left) and polarization profile (right) for the second boundary conditions. The black circle indicates the minimum length scale

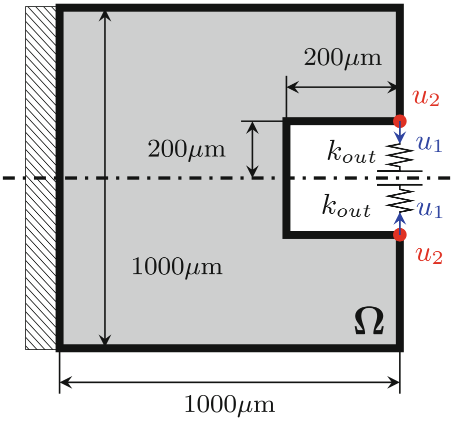


**Fig. 6** Boundary conditions for the third example

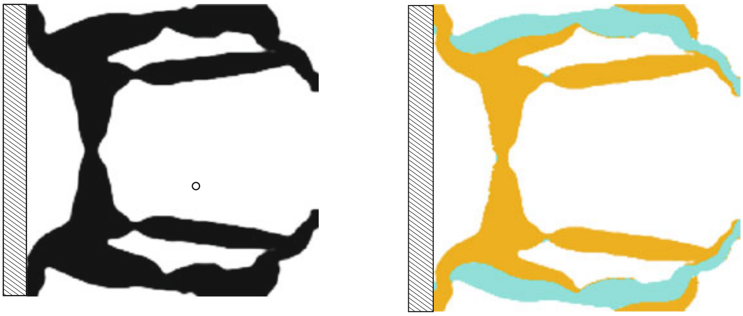


**Fig. 7** Structural layout (left) and polarization profile (right) for the third boundary conditions. The black circle indicates the minimum length scale

Optimized design for the new boundary conditions is shown in Fig. 7. Structural layout and polarization profile are presented in Fig. 7(left) and (right), respectively. The in-plane displacement of the gripper is $u_1 = 140\,\mu\text{m}$.

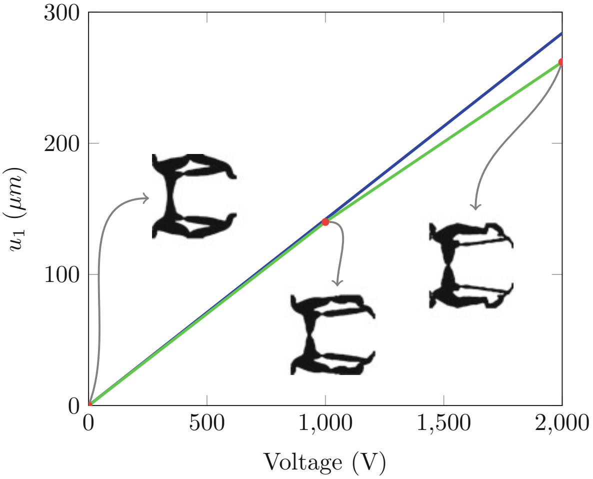


**Fig. 8** Evolution of the output displacement

## 5 Comparison Between Linear and Non-linear Modeling

In this section we show the importance of using the geometrically non-linear modeling when the displacements are big enough ($>$ 5% of the length of the domain). In Fig. 8 the evolution of the output displacement ($u_1$) is plotted as a function of the input voltage. Blue line represents the linear modeling of the displacements. In this case the optimized design does not change when the input voltage is increased and the evolution of the output is linear. The green line shows the geometrically non-linear modeling of the displacements. For this particular case, the optimized designs change when the input voltage in increased, actually, the higher the input, the greater the difference between both models. As expected, for low voltages, both designs are the same.

## 6 Conclusions

In this paper, different kind of piezoelectric actuators have been obtained by using the topology optimization method. The out-of-plane displacement caused by unsymmetrical lamination of the piezoelectric actuator (which is a real limitation when fabricating at micro-scale) is suppressed in order to get genuine in-plane actuators. This is obtained by adding an additional constraint to the optimization problem for each point where we want to cancel the out-of-plane bending.

Two different modelings have been used in this work, a linear model used for low inputs and a geometrically non-linear one for high voltages where the displacements are big compared to the size of the device. In both cases the out-of-plane displacement has been modeled with the linear approach, since these displacements are small in general and this allows us to save computational time. In addition, an elastic energy interpolation scheme has been used with the objective of alleviating convergence problems due to the excessive distortion of low density elements.

Finally, in order to ensure robustness (avoiding the appearance of hinges) and manufacturability (controlling the minimum length scale) of the optimized designs, a robust formulation approach has been used in this problem.

**Acknowledgements** This research has been founded through grant MTM2013-47053-P from the Spanish Ministerio de Economía y Competitividad. Special thanks to José Luis Sánchez-Rojas from Microsystems Actuators and Sensors Group (UCLM) and Ole Sigmund from the Department of Mechanical Engineering, Section of Solid Mechanics (DTU).

## References

1. Bruns, T.E., Tortorelli, D.A.: Topology optimization of non-linear elastic structures and compliant mechanisms. Comput. Methods Appl. Mech. Eng. **190**, 3443–3459 (2001)
2. Buhl, T., Pedersen, C., Sigmund, O.: Stiffness design of geometrically nonlinear structures using topology optimization. Struct. Multidiscip. Optim. **19**, 93–104 (2000)
3. Carbonari, R.C., Silva, E.C.N., Nishiwaki, S.: Optimum placement of piezoelectric material in piezoactuator design. Smart Mater. Struct. **16**, 207–220 (2007)
4. Díaz A.R., Kikuchi, N.: Solution to shape and topology eigenvalue optimization problems using a homogenization method. Int. J. Numer. Methods Eng. **35**, 1487–1502 (1992)
5. Frecker, M.I., Ananthasuresh, G.K., Nishiwaki, S., Kota, S.: Topological synthesis of compliant mechanisms using multi-criteria optimization. Trans. ASME **199**, 238–245 (1997)
6. Gibbs, G., Fuller, C.: Excitation of thin beams using asymmetric piezoelectric actuators. J. Acoust. Soc. Am. **92**, 3221–3227 (1992)
7. Jensen, J., Sigmund, O.: Topology optimization for nano-photonics. Laser Photonics Rev. **5**, 308–321 (2011)
8. Kang, Z., Tong, L.: Integrated optimization of material layout and control voltage for piezoelectric laminated plates. J. Intell. Mater. Syst. Struct. **19**, 889–904 (2008)
9. Kang, Z., Tong, L.: Topology optimization-based distribution design of actuation voltage in static shape control of plates. Comput. Struct. **86**, 1885–1893 (2008)
10. Kang, Z., Wang, R., Tong, L.: Combined optimization of bi-material structural layout and voltage distribution for in-plane piezoelectric actuation. Comput. Methods Appl. Mech. Eng. **200**, 1467–1478 (2011)
11. Kang, Z., Wang, X., Luo, Z.:. Topology optimization for static shape control of piezoelectric plates with penalization on intermediate actuation voltage. J. Mech. Des. **134**, 051006 (2012)
12. Kögl, M., Silva, E.C.N.: Topology optimization of smart structures: design of piezoelectric plate and shell actuators. Smart Mater. Struct. **14**, 387–399 (2005).
13. Kucera, M., Wistrela, E., Pfusterschmied, G., Ruiz-Díez, V., Manzaneque, T., Hernando-García, J., Sánchez-Rojas, J.L., Jachimowicz, A., Schalko, J., Bittner, A., Schmid, U.: Design-dependent performance of self-actuated and self-sensing piezoelectric-AlN cantilevers in liquid media oscillating in the fundamental in-plane bending mode. Sensors Actuators B Chem. **200**, 235–244 (2014)
14. Lazarov, B.S., Schevenels, M., Sigmund, O.: Robust design of large-displacement compliant mechanisms. Mech. Sci. **2**, 175–182 (2011)
15. Luo, Z., Gao, W., Song, C.: Design of multi-phase piezoelectric actuators. J. Intell. Mater. Syst. Struct. **21**, 1851–1865 (2010)

16. Maute, K., Frangopol, D.M.: Reliability-based design of MEMS mechanisms by topology optimization. Comput. Struct. **81**, 813–824 (2003)
17. Molter, A., Fonseca, J.S.O., dos Santos Fernandez, L.: Simultaneous topology optimization of structure and piezoelectric actuators distribution. Appl. Math. Model. **40**, 5576–5588 (2016)
18. Oñate, E.: Cálculo de Estructuras por el Método de Elementos Finitos, Análisis estático lineal, 2nd edn. CIMNE, Barcelona (1995)
19. Pedersen, N.L.: Maximization of eigenvalues using topology optimization. Struct. Multidiscip. Optim. **20**, 2–11 (2000)
20. Pedersen, C.B.W., Buhl, T., Sigmund, O.: Topology synthesis of large-displacement compliant mechanism. Int. J. Numer. Methods Eng. **50**, 2683–2705 (2001)
21. Ruiz, D., Sigmund, O.: Optimal design of robust piezoelectric microgrippers undergoing large displacements. Struct. Multidiscip. Optim. **57**, 71–82 (2018)
22. Ruiz, D., Bellido, J.C., Donoso, A.: Design of piezoelectric modal filters by simultaneously optimizing the structure layout and the electrode profile. Struct. Multidiscip. Optim. **53**, 715–730 (2016)
23. Ruiz, D., Donoso, A., Bellido, J.C., Kucera, M., Schmid, U., Sánchez-Rojas, J.L.: Design of piezoelectric microtransducers based on the topology optimization method. Microsyst. Technol. **22**, 1733–1740 (2016)
24. Ruiz, D., Díaz-Molina, A., Sigmund, O., Donoso, A., Bellido, J.C., Sánchez-Rojas, J.L.: Optimal design of robust piezoelectric unimorph microgrippers. Appl. Mech. Eng. **55**, 1–12 (2017)
25. Sigmund, O.: On the design of compliant mechanisms using topology optimization. Mech. Struct. Mach. **25**, 493–524 (1997)
26. Sigmund, O.: Design of multiphysics actuators using topology optimization. Part I: one-material structures. Comput. Methods Appl. Mech. Eng. **190**, 6577–6604 (2001)
27. Sigmund, O.: Design of multiphysics actuators using topology optimization. Part II: two-material structures. Comput. Methods Appl. Mech. Eng. **190**, 6605–6627 (2001)
28. Sigmund, O.: Manufacturing tolerant topology optimization. Acta Mech. Sin. **25**, 227–239 (2009)
29. Sigmund, O., Jensen, J.S.: Systematic design of phononic band gap materials and structures by topology optimization. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. **361**, 1001–1019 (2003)
30. Sigmund, O., Torquato, S.: Design of materials with extreme thermal expansion using a three-phase topology optimization method. J. Mech. Phys. Solids **45**, 1037–1067 (1997)
31. Sigmund, O., Torquato, S., Aksay, I.A.: On the design of 1–3 piezo-composites using topology optimization. J. Mater. Res. **13**, 1038–1048 (1998)
32. Silva, E.C.N., Kikuchi, N.: Design of piezoelectric transducers using topology optimization. Smart Mater. Struct. **8**, 350–364 (1999)
33. Silva, E.C.N., Fonseca, J.S.O., Kikuchi, N.: Optimal design of piezoelectric microstructures. Comput. Mech. **19**, 397–410 (1997)
34. Svanberg, K.: The method of moving asymptotes-a new method for structural optimization. Int. J. Numer. Meth. Eng. **24**, 359–373 (1987)
35. Wang, F., Lazarov, B.S., Sigmund, O.: On projection methods, convergence and robust formulations in topology optimization. Struct. Multidiscip. Optim. **43**, 767–784 (2011)
36. Wang, F., Lazarov, B., Sigmund, O., Jensen, J.S.: Interpolation scheme for fictitious domain techniques and topology optimization of finite strain elastic problems. Comput. Methods Appl. Mech. Eng. **276**, 453–472 (2014)
37. Zienkiewicz, O.C., Taylor, R.L., Fox, D.: The Finite Element Method for Solid and Structural Mechanics, 7th edn. Butterworth-Heinemann, Oxford (2014)

# Formulation and Analysis of a Class of Direct Implicit Integration Methods for Special Second-Order I.V.P.s in Predictor-Corrector Modes

**Higinio Ramos**

**Abstract** A detailed analysis of different formulations of a class of explicit direct integration methods in predictor-corrector modes for solving special second-order initial-value problems has been carried out (Comput. Phys. Comm. **181** (2010) 1833–1841), showing that the adequate combination of the involved formulas led to an increase in the order of the method. In this paper we consider different formulations of the implicit direct integration methods in predictor-corrector modes. An analysis of the accumulated truncation errors is made and the stability analysis is addressed, including the intervals of stability. Some numerical examples are presented to show the performance of the different formulations. These methods may constitute a reliable alternative to other methods in the literature for solving special second order problems.

**Keywords** Error analysis · Stability analysis · Falkner implicit methods · Direct integration methods · Second-order initial-value problems · Predictor-corrector modes

## 1 Introduction

Second-order differential equations appear frequently in the applied sciences. Examples of that are the mass movement under the action of a force, problems of orbital dynamics, or in general, any problem involving Newton's law.

H. Ramos (✉)
Grupo de Computación Científica, Universidad de Salamanca, Escuela Politécnica Superior de Zamora, Zamora, Spain
e-mail: higra@usal.es

© Springer Nature Switzerland AG 2019

J. L. García Guirao et al. (eds.), *Recent Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 18,
https://doi.org/10.1007/978-3-030-00341-8_3

Among the general procedures for direct integration of the so-called *special second-order* initial value problem (I.V.P.)

$$y''(x) = f(x, y(x)), \quad y(x_0) = y_0, \quad y'(x_0) = y'_0\,, \tag{1}$$

the Falkner methods [9] is a class of schemes that may be used for this purpose.

Although it is possible to integrate a second-order I.V.P. by reducing it to a first-order system and applying one of the methods available for such systems, it seems more natural to provide numerical methods in order to integrate the problem directly. The advantage of this procedure lies in the fact that they are able to exploit special information about ODEs, resulting in an increase in efficiency. For instance, it is well-known that Runge-Kutta-Nyström methods for (1) involve a real improvement as compared to standard Runge-Kutta methods for a given number of stages [13, p. 285], although the computational cost remains high because of the number of function evaluations. On the other hand, a linear $k$-step method for first-order ODEs becomes a $2k$-step method for (1), [13, p. 461], increasing the computational work. In the words of Krogh [17], "the direct integration of second order systems requires about half the number of function evaluations required for integrating the equivalent first order system".

The explicit Falkner method of $k$ steps consists in a couple of formulas that can be written in the form [7]

$$y_{n+1} = y_n + h\, y'_n + h^2 \sum_{j=0}^{k-1} \beta_j \, \nabla^j \, f_n\,, \tag{2}$$

$$y'_{n+1} = y'_n + h \sum_{j=0}^{k-1} \gamma_j \, \nabla^j \, f_n\,, \tag{3}$$

where $h$ is the stepsize, $y_n$ and $y'_n$ are approximations to the values of the solution and its derivative at $x_n = x_0 + n\,h$, $f_n = f(x_n, y_n)$, and $\nabla^j f_n$ is the standard notation for the backward differences. The coefficients $\beta_j$ and $\gamma_j$ can be obtained using the generating functions

$$G_\beta(t) = \sum_{j=0}^{\infty} \beta_j\, t^j = \frac{t + (1-t)\, Log(1-t)}{(1-t) Log^2(1-t)}\,,$$

$$G_\gamma(t) = \sum_{j=0}^{\infty} \gamma_j\, t^j = \frac{-t}{(1-t) Log(1-t)}\,,$$

which have been obtained similarly as that for the Störmer or Cowell methods (see [14, p. 291]).

The implicit Falkner method of $k$ steps consists in two formulas that may be written as [7]

$$y_{n+1} = y_n + h\, y_n' + h^2 \sum_{j=0}^{k} \beta_j^* \, \nabla^j \, f_{n+1} \,, \tag{4}$$

$$y_{n+1}' = y_n' + h \sum_{j=0}^{k} \gamma_j^* \, \nabla^j \, f_{n+1} \,, \tag{5}$$

with generating functions for the coefficients given respectively by

$$G_{\beta^*}(t) = \sum_{j=0}^{\infty} \beta_j^* \, t^j = \frac{t + (1-t)\, Log(1-t)}{Log^2(1-t)} \,,$$

$$G_{\gamma^*}(t) = \sum_{j=0}^{\infty} \gamma_j^* \, t^j = \frac{-t}{Log(1-t)} \,.$$

Note that the formulas in (3) and (5) are respectively the Adams-Bashforth and Adams-Moulton schemes for the problem $(y')' = f(x, y)$, which are used to follow the values of the derivative. All the above formulas are of multistep type, specifically $k$-step formulas, and so $k$ initial values must be provided in order to proceed with the methods (the Runge-Kutta methods are commonly used to obtain the starting values). In this paper, a rigorous analysis of the errors involved in the formulation of the implicit procedures in predictor-corrector modes (P-C) with fixed step size is made (for a variable step size version of these methods see [31], and for an adapted version of the methods in [23] in case of oscillatory systems see [21] and [20]. A new technique for stepsize changing in case of Adam's method has recently appeared in [2]. Other approaches concerning different formulations of Falkner-type methods have appeared in [25] or [24].

The implicit formulas in (4)–(5) may be used to generate methods for solving the I.V.P. in (1). A well-known procedure of this type is the Wilson method [32], which is one of the Newmark family, and is commonly used in molecular dynamics calculations. This method uses the two-step formula in (4), given by

$$y_{n+1} = y_n + h\, y_n' + \frac{h^2}{6} \left(2y_n'' + y_{n+1}''\right) \tag{6}$$

to obtain the positions, while the formula to update the velocities is the two-step method in (5) given by

$$y_{n+1}' = y_n' + \frac{h}{2}\, (y_n'' + y_{n+1}'') \,. \tag{7}$$

Another scheme that uses a Falkner formula is a method due to Gear [11], consisting of the two implicit formulas given by

$$y_{n+1} = y_n + h\, y'_n + \frac{h^2}{12}\left(y''_{n+1} + 6y''_n - y''_{n-1}\right) \tag{8}$$

$$y'_{n+1} = y'_n + \frac{h}{12}\,(5y''_{n+1} + 8y''_n - y''_{n-1})\,, \tag{9}$$

where the second formula is the three-step method in (5) but the first of these formulas is not the three-step method in (4).

The application of any of such procedures to (1) results in an implicit system that must be solved at each step, involving a great computational cost, but in practice, an explicit formulation in predictor-corrector mode is frequently used. In this way, the implicit methods in (4)–(5) are adequately combined with the explicit ones in (2)–(3) so as to avoid having to solve an algebraic system at each step. This P-C formulation may also be used as a starting method, as for example in [6], where the one-step implicit Falkner method in predictor-corrector mode is used to provide the starting values in the application of the De Vogelaere's method. Other examples of such procedures may be found in [1, 12] or [28].

In recent years it seems to be a competition to find the numerical method that performs better in solving different types of problems. This is not the goal of this article; in some cases the proposed methods will do better and in other cases they will perform worst than other specialized methods. But the question is how these general purposed methods can be formulated and why they do work or do not, developing an analysis of the errors involved and the stability characteristics. Anyway, we can say that the methods presented here may constitute a reliable alternative to other methods in the literature for solving special second order problems.

The paper is organized as follows. In the following section different formulations of the explicit and implicit Falkner methods are presented. The analysis of the explicit methods was done in [23] while the analysis of the implicit formulas formulated in P-C mode will be done here. Section 3 is devoted to the analysis of the local truncation errors and accumulated truncation errors. Section 4 deals with the stability analysis, which results of vital importance in the application of the methods. In Sect. 5 some examples are presented to show the performance of the different formulations. In the final section some conclusions put an end to the article.

## 2 Formulations of Falkner Methods in P-C Mode

In the application of P-C modes, P indicates the application of the explicit method, in our case the predictor given by the formula in (2), and C indicates the application of the implicit method, that is, the corrector given by the formula in (4). For the derivative, we will use P' to indicate the application of the explicit formula in (3), and C' to indicate the application of the implicit formula in (5). Finally, E refers to the evaluation of the function $f$ in order to use this value either for corrections, or in the next step.

Two different implementations for the explicit methods were considered in [23], and for the implicit methods three implementations will be considered here. They are summarized in what follows.

## 2.1 *Explicit Methods*

### 2.1.1 PP'E Mode

The fist formulation of the explicit Falkner method on each step for solving the problem in (1) consists in

1. Evaluate $y_{n+1}$ using the formula in (2)
2. Evaluate $y'_{n+1}$ using the formula in (3)
3. Evaluate $f_{n+1} = f(x_{n+1}, y_{n+1})$

### 2.1.2 PEC' Mode

The second formulation of the explicit Falkner method on each step for solving the problem in (1) reads

1. Evaluate $y_{n+1}$ using the formula in (2)
2. Evaluate $f_{n+1} = f(x_{n+1}, y_{n+1})$
3. Evaluate $y'_{n+1}$ using the formula in (5)

which can be accomplished due to the absence of the derivative on the function $f = f(x, y(x))$ . Thus, having obtained the value $y_{n+1}$ it is straightforward to obtain $f_{n+1}$ to be used in the formula in (5). Note that in this way the "implicit formula" in (5) is no longer implicit, resulting in an explicit formulation of the method.

For the PP'E mode the accumulated truncation error is of order $\mathcal{O}(h^k)$ while for the PEC' mode is of order $\mathcal{O}(h^{k+1})$ (for details see [23]). Thus, the second formulation of the explicit Falkner method (PEC' mode) provides a better performance.

## 2.2 *Implicit Methods*

### 2.2.1 P'PECE Mode

The first choice for the implementation of the implicit Falkner method is given by

1. Evaluate $y'_{n+1}$ using the formula in (3) to obtain $y'^{P}_{n+1}$
2. Evaluate $y_{n+1}$ using the formula in (2) to obtain $y^{P}_{n+1}$

3. Evaluate $f_{n+1} = f(x_{n+1}, y_{n+1}^P)$
4. Evaluate $y_{n+1}$ using the formula in (4) to obtain $y_{n+1}^C$
5. Evaluate $f_{n+1} = f(x_{n+1}, y_{n+1}^C)$

### 2.2.2 PEC'CE Mode

The second possibility is given by

1. Evaluate $y_{n+1}$ using the formula in (2) to obtain $y_{n+1}^P$
2. Evaluate $f_{n+1} = f(x_{n+1}, y_{n+1}^P)$
3. Evaluate $y'_{n+1}$ using the formula in (5) to obtain $y'^C_{n+1}$
4. Evaluate $y_{n+1}$ using the formula in (4) to obtain $y_{n+1}^C$
5. Evaluate $f_{n+1} = f(x_{n+1}, y_{n+1}^C)$

### 2.2.3 PECEC' Mode

The last implementation consists in

1. Evaluate $y_{n+1}$ using the formula in (2) to obtain $y_{n+1}^P$
2. Evaluate $f_{n+1} = f(x_{n+1}, y_{n+1}^P)$
3. Evaluate $y_{n+1}$ using the formula in (4) to obtain $y_{n+1}^C$
4. Evaluate $f_{n+1} = f(x_{n+1}, y_{n+1}^C)$
5. Evaluate $y'_{n+1}$ using the formula in (5) to obtain $y'^C_{n+1}$

Note that in the above formulations, once we have obtained the final value for $y_{n+1}$ given by $y_{n+1}^C$ we have to evaluate $f_{n+1} = f(x_{n+1}, y_{n+1}^C)$, which will be used in the next step. This evaluation is indicated by the last E in each of the different modes.

Obviously, many more choices could have been considered, taking formulas with different steps and therefore, a different order. Specifically, we have considered a predictor with $k + 1$ steps and a corrector with $k$ steps, but we have not obtained any improvements over the methods with $k$ steps in the predictor and the corrector. Or you could have considered further corrections, obtaining methods of the forms $P'P(EC)^n E$, $P(EC'C)^n E$ or $P(EC)^n EC'$. We have done different experiments, and in the latter case the interval of stability varies, resulting in an interval slightly higher, but with so little difference that the computational effort is not worth it. We have considered only P-C methods with the same number of steps in all the formulas (as it has been used by Shampine and Gordon [26]), where it is performed only a correction, as is usually done in the Adams methods in P-C mode (see [18]).

# 3 Error Analysis

## 3.1 Local Truncation Errors

Consider the formula resulting after approximating the function $f$ on the exact formula

$$y(x_{n+1}) = y(x_n) + hy'(x_n) + \int_{x_n}^{x_{n+1}} f(x, y(x))(x_{n+1} - x)\,dx \tag{10}$$

by the interpolating polynomial on the grid points $x_{n-(k-1)}, \ldots, x_n$, and also consider the formula in (2). Using the localization hypothesis that $y(x_j) = y_j,\ j = n-(k-1), \ldots, n$, and following a similar procedure as for the Adams formulas [18] we obtain that the local truncation error for the method in (2) is given by

$$\mathfrak{L}_P[y(x_n); h] = y(x_{n+1}) - y^P_{n+1} = h^{k+2} y^{(k+2)}(\xi)\beta_k, \tag{11}$$

where $\xi$ is an internal point of the smallest interval containing $x_{n-(k-1)}, \ldots, x_n$.

Similarly, for the formula in (3) the local truncation error reads

$$\mathfrak{L}_{P'}[y'(x_n); h] = y'(x_{n+1}) - y'^P_{n+1} = h^{k+1} y^{(k+2)}(\psi)\gamma_k, \tag{12}$$

where as before $\psi$ refers to an internal point.

For the formula in (4) the local truncation error is given by

$$\mathfrak{L}_C[y(x_n); h] = y(x_{n+1}) - y^C_{n+1} = h^{k+3} y^{(k+3)}(\xi)\beta^*_{k+1}, \tag{13}$$

and similarly, for the formula in (5) the local truncation error is

$$\mathfrak{L}_{C'}[y'(x_n); h] = y'(x_{n+1}) - y'^C_{n+1} = h^{k+2} y^{(k+3)}(\psi)\gamma^*_{k+1} \tag{14}$$

where we have denoted by $\xi$ or $\psi$ the internal points, or if they have to be different, we will use $\xi_j$ or $\psi_j$, as in the following section.

## 3.2 Accumulated Truncation Errors for the Implicit Falkner Method in P-C Modes

### 3.2.1 P'PECE Mode

Assuming that we have to integrate the problem in (1) on the interval $[x_0, x_N]$ and that we know in advance the starting values needed to apply the numerical scheme, we proceed by analyzing the accumulated errors on each successive application of the method along the grid points on the integration interval. We use a superscript P to indicate the application of the explicit method and a superscript C to indicate the application of the implicit one, regardless of whether the formula is for the solution or for the derivative.

Assuming the localization hypothesis and using the formulas for the local truncation errors in (11) and (12), for the first step ($n = 0$) we have that the differences between the true values, $y(x_1)$ , $y'(x_1)$, and the approximated ones, $y_1^P$ , $y_1'^P$, are given respectively by the local truncation errors, that is,

$$y'(x_1) - y_1'^P = h^{k+1} y^{(k+2)}(\psi_1)\gamma_k \,, \tag{15}$$

$$y(x_1) - y_1^P = h^{k+2} y^{(k+2)}(\xi_1)\beta_k \,, \tag{16}$$

where, for convenience, in the sequel the $\xi_j$ and $\psi_j$ will denote appropriate internal points.

After evaluating $f(x_1, y_1^P)$, using the Mean Value Theorem we can put that

$$f(x_1, y(x_1)) - f_1^P = \frac{\partial f}{\partial y}(x_1, \xi_1) \left( y(x_1) - y_1^P \right)$$

where $f_1^P = f(x_1, y_1^P)$. Assuming the localization hypothesis, we have

$$\begin{aligned} y(x_1) - y_1^C = {} & y(x_0) + h\, y'(x_0) + h^2 \sum_{j=0}^{k} \beta_j^* \, \nabla^j \, f(x_1, y(x_1)) \\ & + h^{k+3} y^{(k+3)}(\xi_1)\beta_{k+1}^* - \left( y_0 + h\, y_0' + h^2 \sum_{j=0}^{k} \beta_j^* \, \nabla^j \, f_1^P \right) \\ = {} & h^{k+3} y^{(k+3)}(\xi_1)\beta_{k+1}^* + \mathcal{O}(h^{k+4}) \,. \end{aligned} \tag{17}$$

After evaluating $f(x_1, y_1^C)$, for the next step ($n = 1$) we have that

$$\begin{aligned} y'(x_2) - y_2'^P = {} & y'(x_1) + h \sum_{j=0}^{k-1} \gamma_j \nabla^j f(x_1, y(x_1)) \\ & + h^{k+1} y^{(k+2)}(\psi_2)\gamma_k - \left( y_1'^P + h \sum_{j=0}^{k-1} \gamma_j \nabla^j f_1^C \right) \end{aligned} \tag{18}$$

where we have used the values of the step before, $y_1'^P$ and $y_1^C$, and have used the notation $f_1^C = f(x_1, y_1^C)$. After some calculations, using the formulas in (15) and (17), it results that

$$y'(x_2) - y_2'^P = h^{k+1}\gamma_k \left(y^{(k+2)}(\psi_1) + y^{(k+2)}(\psi_2)\right) + \mathcal{O}(h^{k+2})\,. \tag{19}$$

Similarly, for the predictor we have

$$y(x_2) - y_2^P = h^{k+2}\beta_k y^{(k+2)}(\xi_2) + h^{k+2} y^{(k+2)}(\xi_1)\gamma_k + \mathcal{O}(h^{k+3})\,. \tag{20}$$

And finally, for the corrector we get that

$$\begin{aligned} y(x_2) - y_2^C &= y(x_1) + h\, y'(x_1) + h^2 \sum_{j=0}^{k} \beta_j^* \, \nabla^j \, f(x_1, y(x_1)) \\ &\qquad + h^{k+3} y^{(k+3)}(\xi_2)\beta_{k+1}^* - \left( y_1^C + h\, y_1'^P + h^2 \sum_{j=0}^{k} \beta_j^* \, \nabla^j \, f_2^P \right) \\ &= h^{k+3}\beta_{k+1}^* \left(y^{(k+3)}(\xi_1) + y^{(k+3)}(\xi_2)\right) + h^{k+2} y^{(k+2)}(\psi_1)\gamma_k \\ &\qquad + \mathcal{O}(h^{k+4})\,. \end{aligned} \tag{21}$$

Repeating the procedure along the nodes on the integration interval we determine that the accumulated error at the final point $x_N$ is given by

$$\begin{aligned} y(x_N) - y_N^C = h^{k+2}\gamma_k \sum_{j=1}^{N-1} (N-j) y^{(k+2)}(\psi_j) + h^{k+3}\beta_{k+1}^* \sum_{j=1}^{N} y^{(k+3)}(\xi_j) \\ + \mathcal{O}(h^{k+4})\,, \end{aligned} \tag{22}$$

and for the derivative we have

$$y'(x_N) - y_N'^P = h^{k+1}\gamma_k \sum_{j=1}^{N} y^{(k+2)}(\psi_j) + \mathcal{O}(h^{k+2})\,. \tag{23}$$

Assuming that the derivatives $y^{(k+2)}(x)$ and $y^{(k+3)}(x)$ are continuous, after using the Mean Value Theorem, the above formulas may be rewritten as follows:

$$\begin{aligned} y(x_N) - y_N^C = \frac{1}{2} h^k \gamma_k y^{(k+2)}(\psi)(x_N - x_0)(x_N - x_1) \\ + h^{k+2}\beta_{k+1}^* y^{(k+3)}(\xi)(x_N - x_0) + \mathcal{O}(h^{k+4}) \end{aligned} \tag{24}$$

and

$$y'(x_N) - y_N'^P = h^k \gamma_k (x_N - x_0) y^{(k+2)}(\psi) + \mathcal{O}(h^{k+2}) . \tag{25}$$

### 3.2.2 PEC'CE Mode

In this mode, for the first step ($n = 0$), after the application of the predictor P and the corrector C we have the same results as before, that is,

$$y(x_1) - y_1^P = h^{k+2} y^{(k+2)}(\xi_1)\beta_k ,$$

and

$$y(x_1) - y_1^C = h^{k+3} y^{(k+3)}(\xi_1)\beta_{k+1}^* + \mathcal{O}(h^{k+4}) .$$

Now the difference with the mode before results in the application of the corrector C' to obtain the approximation for the derivative. We have

$$\begin{aligned} y'(x_1) - y_1'^C &= y'(x_0) + h \sum_{j=0}^{k} \gamma_j^* \nabla^j f(x_1, y(x_1)) \\ &\quad + h^{k+2} y^{(k+3)}(\psi_1)\gamma_{k+1}^* - \left( y_0'^C + h \sum_{j=0}^{k} \gamma_j^* \nabla^j f_1^P \right) \\ &= h^{k+2} y^{(k+3)}(\psi_1)\gamma_{k+1}^* + \mathcal{O}(h^{k+3}) . \end{aligned} \tag{26}$$

After the evaluation of $f(x_1, y_1^C)$ and the application of the predictor P to obtain an estimate $y_2^P$ for $y(x_2)$, the application of the corrector C results in

$$\begin{aligned} y(x_2) - y_2^C &= h^{k+3}\beta_{k+1}^* \left( y^{(k+3)}(\xi_1) + y^{(k+3)}(\xi_2) \right) + h^{k+3} y^{(k+3)}(\psi_1)\gamma_{k+1}^* \\ &\quad + \mathcal{O}(h^{k+4}) , \end{aligned} \tag{27}$$

while the application of the corrector C' produces

$$y'(x_2) - y_2'^C = h^{k+2}\gamma_{k+1}^* \left( y^{(k+3)}(\psi_1) + y^{(k+3)}(\psi_2) \right) + \mathcal{O}(h^{k+3}) . \tag{28}$$

Repeating the procedure along the nodes on the integration interval we obtain that the accumulated error at the final point $x_N$ is given by

$$y(x_N) - y_N^C = h^{k+3}\gamma_{k+1}^* \sum_{j=1}^{N-1} (N-j) y^{(k+3)}(\psi_j) + h^{k+3}\beta_{k+1}^* \sum_{j=1}^{N} y^{(k+3)}(\xi_j) + \mathcal{O}(h^{k+4}) \,, \tag{29}$$

while for the derivative we have the accumulated error given by

$$y'(x_N) - y_N^{\prime C} = h^{k+2}\gamma_{k+1}^* \sum_{j=1}^{N} y^{(k+3)}(\psi_j) + \mathcal{O}(h^{k+3}) \,. \tag{30}$$

Assuming that $y^{(k+3)}(x)$ is continuous, the above formulas for the accumulated errors at the final point $x_N$ may be written as

$$y(x_N) - y_N^C = \frac{1}{2} h^{k+1}\gamma_{k+1}^* y^{(k+3)}(\psi)(x_N - x_0)(x_N - x_1) + h^{k+2}\beta_{k+1}^* y^{(k+3)}(\xi)(x_N - x_0) + \mathcal{O}(h^{k+4}) \tag{31}$$

for the solution, and

$$y'(x_N) - y_N^{\prime C} = h^{k+1}\gamma_{k+1}^* (x_N - x_0) y^{(k+3)}(\psi) + \mathcal{O}(h^{k+3}) \tag{32}$$

for the derivative.

### 3.2.3 PECEC’ Mode

Now, the difference with the previous mode results from the application of the corrector C’ to obtain the values $y_n^{\prime C}$, using $f(x_n, y_n^C)$ instead of $f(x_n, y_n^P)$. The first difference with the previous mode is observed in the first step ($n = 0$) and results to be

$$\begin{aligned} y'(x_1) - y_1^{\prime C} &= y'(x_0) + h \sum_{j=0}^{k} \gamma_j^* \nabla^j f(x_1, y(x_1)) \\ &\quad + h^{k+2} y^{(k+3)}(\psi_1)\gamma_{k+1}^* - \left( y_0^{\prime C\,\prime} + h \sum_{j=0}^{k} \gamma_j^* \nabla^j f_1^C \right) \\ &= h^{k+2} y^{(k+3)}(\psi_1)\gamma_{k+1}^* + \mathcal{O}(h^{k+3}) \,. \end{aligned} \tag{33}$$

**Table 1** Principal terms of the accumulated errors for the different implementations of the explicit and implicit Falkner methods

| Method | $PT(y(x_n) - y_N)$ | $PT(y'(x_n) - y'_N)$ |
|---|---|---|
| $FEABk$ | $\frac{1}{2}\, h^k y^{(k+2)}(\psi)(x_N - x_0)(x_N - x_1)\gamma_k$ | $h^k y^{(k+2)}(\psi)(x_N - x_0)\gamma_k$ |
| $FEAMk$ | $\frac{1}{2}\, h^{k+1} y^{(k+3)}(\psi)(x_N - x_0)(x_N - x_1)\gamma^*_{k+1}$<br>$+h^{k+1} y^{(k+2)}(\xi)(x_N - x_0)\beta_k$ | $h^{k+1} y^{(k+3)}(\psi)(x_N - x_0)\gamma^*_{k+1}$ |
| $FI[1]k$ | $\frac{1}{2}\, h^k y^{(k+2)}(\psi)(x_N - x_0)(x_N - x_1)\gamma_k$ | $h^k y^{(k+2)}(\psi)(x_N - x_0)\gamma_k$ |
| $FI[2]k$<br>$FI[3]k$ | $\frac{1}{2}\, h^{k+1} y^{(k+3)}(\psi)(x_N - x_0)(x_N - x_1)\gamma^*_{k+1}$ | $h^{k+1} y^{(k+3)}(\psi)(x_N - x_0)\gamma^*_{k+1}$ |

Note that the only difference between the formulas in (26) and in (33) is that $f_1^P$ has been substituted by $f_1^C$, but this does not change the principal term of the errors. This means that the principal terms in the errors for the final formulas are the same as in the previous mode, and so the formulas in (31) and (32) remain valid in this mode. This, however, does not mean that the errors are equal, since the remaining terms are different. But since the principal terms are the same the errors will be similar.

For comparison purposes we have included in Table 1 the principal terms of the accumulated errors for the different implementations of the explicit and implicit Falkner methods. These terms will be denoted by $PT(y(x_n) - y_N)$ for the solution, and $PT(y'(x_n) - y'_N)$ for the derivative, where the approximated values of the solution and the derivative at the final point are indicated by $y_N$ and $y'_N$ respectively. The formulations of the explicit methods with $k$ steps are named $FEABk$ and $FEAMk$ as in [23], while the implicit method with $k$ steps formulated in P-C modes are named $FI[1]k$, $FI[2]k$ and $FI[3]k$ corresponding to the P'PECE, PEC'CE and PECEC' modes respectively in the above section.

*Remark 1* Note that the above methods may be applied to a system of second-order differential equations of the form

$$\begin{cases} \mathbf{y}''(x) = \mathbf{f}(x, \mathbf{y}(x)), & x \in [x_0, x_N] \\ \mathbf{y}(x_0) = \mathbf{y}_0 \end{cases}$$

where $\mathbf{y} : \mathbb{R}^m \to \mathbb{R}^m$, and $\mathbf{f} : [x_0, x_N] \times \mathbb{R}^m \to \mathbb{R}^m$, using a componentwise implementation.

## 4 Stability

In the context of ordinary differential equations, the concept of stability refers to what extent a numerical scheme is appropriate for solving an initial-value problem. Roughly speaking, a given method can be said to be stable if small changes in the data result in small changes in the solution obtained.

A procedure commonly used to study stability (zero-stability) consists in writing the difference equations of the method as a one-step recurrence in a space with high dimension and in an adequate norm to bound the finite powers of the resulting matrices. For the different combinations of the implicit method in (4) to get the above P-C modes, a similar procedure to that in [23] may be considered to obtain zero-stability. But, the zero stability is only a minimal condition for a numerical method; for second order equations the so-called P-stability is the type of concern together with A-stability.

In order to determine whether a numerical method will produce reasonable results with a given value of $h > 0$, we need a notion of stability that is different from zero-stability. The stability properties are analyzed by using the linear test equation introduced by Lambert and Watson [19]

$$y''(x) = -\mu^2\, y(x)\,, \quad \text{with } \mu > 0\,. \tag{34}$$

The application of the above predictor-corrector formulations of the implicit Falkner method to this problem yields the following recursion

$$Y_n = M[i]\, Y_{n-1}\,, \tag{35}$$

where $Y_n$ is the $k+1$-vector given by $Y_n = \left(y_{n+1}, y_n, \ldots, y_{n-(k-2)}, h\, y'_{n+1}\right)^T$, and $M[i]\,, i = 1, 2, 3,$ are $(k+1) \times (k+1)$ matrices whose elements depends on $H^2$ and the coefficients of the method, where $H = \mu\, h$. The matrices $M[i]$ are called the *stability matrices* and the index $i$ in $M[i]$ is used to denote each implementation according to its appearance in the above section, that is, $i = 1$ stands for the first P-C mode, $i = 2$ for the second P-C mode, and $i = 3$ for the last one.

The entries of each of the stability matrices are given by

$$M[1]_{1\,1} = 1 + H^2 \sum_{j=0}^{k} \left( j - 1 + H^2 \sum_{s=0}^{k-1} \beta_s \right) \beta_j^*$$

$$M[1]_{1\,l+1} = (-1)^l H^2 \sum_{j=0}^{k} \left( \binom{j}{l+1} + H^2 \sum_{s=l}^{k-1} \binom{s}{l} \beta_s \right) \beta_j^*\,,$$
$$l = 1, \ldots, k-1\,.$$

$$M[1]_{j\,l} = \delta_{j\,l+1}\,, \quad j = 2, \ldots, k \quad l = 1, \ldots, k+1\,.$$

$$M[1]_{k+1\,l+1} = (-1)^{l+1} H^2 \sum_{j=l}^{k-1} \binom{j}{l} \gamma_j\,, \quad l = 0, \ldots, k-1\,.$$

$$M[1]_{k+1\,k+1} = 1$$

where $\delta_{j\,l+1}$ is the Kronecker delta and the notations of type $\binom{j}{l}$ refer to the binomial coefficients.

For the matrices $M[2]$ and $M[3]$ all the entries are the same as for $M[1]$ except for the last row, which for $M[2]$ reads

$$M[2]_{k+1\,1} \quad = H^2 \sum_{j=0}^{k} \left( j-1+H^2 \sum_{s=0}^{k-1} \beta_s \right) \gamma_j^*$$

$$M[2]_{k+1\,l+1} = (-1)^l H^2 \sum_{j=0}^{k} \left( \binom{j}{l+1} + H^2 \sum_{s=l}^{k-1} \binom{s}{l} \beta_s \right) \gamma_j^* \,, \qquad l = 1, \ldots, k-1 \,.$$

$$M[2]_{k+1\,k+1} = 1 - H^2 \sum_{j=0}^{k} \gamma_j^* \,,$$

and, similarly, for $M[3]$ the last row results in

$$M[3]_{k+1\,1} \quad = H^2 \sum_{j=0}^{k} \left[ j-1+H^2 \sum_{s=0}^{k} \left( 1-s-H^2 \sum_{t=0}^{k-1} \beta_t \right) \beta_s^* \right] \gamma_j^*$$

$$M[3]_{k+1\,l+1} = (-1)^l H^2 \sum_{j=0}^{k} \left[ \binom{j}{l+1} - H^2 \sum_{t=0}^{k} \left( \binom{t}{l+1} + H^2 \sum_{s=l}^{k-1} \binom{s}{l} \beta_s \right) \beta_t^* \right] \gamma_j^* \,, \qquad l = 1, \ldots, k-1 \,.$$

$$M[3]_{k+1\,k+1} = 1 - H^2 \sum_{j=0}^{k} \left( 1 - H^2 \sum_{s=0}^{k} \beta_s^* \right) \gamma_j^* \,.$$

For example, when $k = 3$ the equation in (35) in case of $M[1]$ reads

$$Y_n = \begin{pmatrix} \frac{19(19H^2-132)H^2}{4320}+1 & \frac{H^2}{10}-\frac{19H^4}{432} & \frac{H^2(19H^2-28)}{1440} & 1-\frac{19H^2}{180} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -\frac{23H^2}{12} & \frac{4H^2}{3} & -\frac{5H^2}{12} & 1 \end{pmatrix} Y_{n-1} \,,$$

while the same equation in case of $M[2]$ results in

$$Y_n = \begin{pmatrix} \frac{19(19H^2-132)H^2}{4320}+1 & \frac{H^2}{10}-\frac{19H^4}{432} & \frac{H^2(19H^2-28)}{1440} & 1-\frac{19H^2}{180} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \frac{19H^4}{64}-\frac{7H^2}{6} & \frac{5\,H^2(4-3H^2)}{96} & \frac{H^2(9H^2-8)}{192} & 1-\frac{3H^2}{8} \end{pmatrix} Y_{n-1}\,,$$

and finally, for the third implementation the matrix $M[3]$ is given by

$$\begin{pmatrix} \frac{19(19H^2-132)H^2}{4320}+1 & \frac{H^2}{10}-\frac{19H^4}{432} & \frac{H^2(19H^2-28)}{1440} & 1-\frac{19H^2}{180} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \frac{2508H^4-361H^6-13440H^2}{11520} & \frac{95H^6-216H^4+1200H^2}{5760} & \frac{28H^4-19H^6-160H^2}{3840} & \frac{19H^4-180H^2}{480}+1 \end{pmatrix}$$

The behaviour of the numerical solution will depend on the eigenvalues of the matrices $M[i]$, and the stability properties of the methods will be characterized by the spectral radius, $\rho(M[i])$. According to the terminology introduced by Coleman and Ixaru [5] we have:

- $(0, H_s)$ is an interval of stability for a given method if for all $H \in (0, H_s)$ it is $|r_i| < 1$ where the $r_i$, $i = 1, \ldots, k+1$ are the eigenvalues of the stability matrix.
- $(0, H_p)$ is an interval of periodicity for a given method if for all $H \in (0, H_p)$ the eigenvalues of the stability matrix, $r_i$, $i = 1, \ldots, k+1$, satisfy

$$r_1 = e^{i\theta(H)}\,, \quad r_2 = e^{-i\theta(H)}\,, \quad |r_i| \leq 1\,,\; i > 2\,,$$

where $\theta(H)$ is a real function.

In particular, if the interval of stability is $(0, \infty)$ the method is *A-stable*, and if the interval of periodicity is $(0, \infty)$ the method is *P-stable*. Note that A-stability is more restrictive than P-stability, that is, A-stability implies P-stability.

The practical significance of the interval of stability is that, for a given $\mu$ in (34), there is no explosion of the error in the numerical solution when $0 < h < H_s/\mu$ [5]. The interval of periodicity defines the stepsize which can be used in order for the approximation of the solution of problems with high oscillatory or periodic solution to be of the same order as the algebraic order of the method. When $0 < h < H_p/\mu$ the numerical solution defined by (35) is also periodic, as is the exact solution of the test model (34) for all non-trivial initial-conditions on $y$ and $y'$.

A crucial difference between A-stability and P-stability is that for A-stable methods the stability matrix satisfies $(M[i])^n \to 0$ as $n \to \infty$ because $\rho(M[i]) < 1$, but for P-stable methods this fact is not possible because $\rho(M[i]) = 1$, (see [10]). Therefore, A-stable methods alleviate the initial errors whereas for P-stable methods the initial errors do not diminish when the integration progresses in time.

If we let the value $\mu$ in (34) to be complex, instead of intervals we obtain stability or periodicity regions in the H-complex plane.

The importance of determining the stability characteristics will be shown by considering the Gear's method in (8)–(9). This is an implicit method for which we need a predicted value to be formulated in P-C mode. Taking as predictor the two-step explicit Falkner method given by

$$y_{n+1} = y_n + h\, y'_n + \frac{h^2}{6}\left(4y''_n - y''_{n-1}\right)$$

to obtain $y^P_{n+1}$, after evaluating $f(x_{n+1}, y^P_{n+1})$ and applying the formula in (8) to obtain $y^C_{n+1}$ we can proceed in two ways:

1. we can evaluate $f(x_{n+1}, y^C_{n+1})$ with the best approximation we have for $y_{n+1}$ and then we use the formula in (9) to obtain the approximate value $y'_{n+1}$. We note this formulation as $P_2ECEC'$, where the subindex 2 corresponds to the order of the predictor.
2. or we can use the previous evaluation $f(x_{n+1}, y^P_{n+1})$ in the formula in (9) to obtain the approximate value $y'_{n+1}$. The notation for this formulation results in $P_2ECC'$.

For the first choice the stability matrix results in

$$\begin{pmatrix} \frac{H^4}{18} - \frac{7H^2}{12} + 1 & -\frac{1}{72}H^2\left(H^2-6\right) & 1 - \frac{H^2}{12} \\ 1 & 0 & 0 \\ \frac{-10H^6+105H^4-468H^2}{432} & \frac{H^2(5(H^2-6)H^2+72)}{864} & \frac{5(H^2-12)H^2+1}{144} \end{pmatrix}$$

for which there are no stability nor periodicity intervals.

In the second case the stability matrix is given by

$$\begin{pmatrix} \frac{H^4}{18} - \frac{7H^2}{12} + 1 & -\frac{1}{72}H^2\left(H^2-6\right) & 1 - \frac{H^2}{12} \\ 1 & 0 & 0 \\ \frac{1}{36}H^2\left(10H^2-39\right) & \frac{1}{72}\left(6H^2-5H^4\right) & 1 - \frac{5H^2}{12} \end{pmatrix}$$

and the stability interval is given by (0, 1.906123). Although at first glance one could intuitively suppose that the first procedure will be the best this is not true. Not only the first procedure requires more computational cost (two evaluations of the function $f$ per step versus one evaluation in the second case) but it also has worst stability properties.

Nevertheless, if we consider as predictor the three-step explicit Falkner method given by

$$y_{n+1} = y_n + h\, y'_n + \frac{h^2}{24}\left(19y''_n - 10y''_{n-1} + 3y''_{n-2}\right)$$

to obtain $y^P_{n+1}$, after evaluating $f(x_{n+1}, y^P_{n+1})$ and applying the formula in (8) to obtain $y^C_{n+1}$ we can also proceed in two ways:

1. we can evaluate $f(x_{n+1}, y^C_{n+1})$ with the best approximation we have for $y_{n+1}$ and then we use the formula in (9) to obtain the approximate value $y'_{n+1}$. Now the notation for this formulation is $P_3ECEC'$.
2. or we can use the previous value $f(x_{n+1}, y^P_{n+1})$ in the formula in (9) to obtain the approximate value $y'_{n+1}$. The notation for this formulation is $P_3ECC'$.

Now in the first case the stability matrix is given by

$$\begin{pmatrix} \frac{19H^4}{288} - \frac{7H^2}{12} + 1 & \frac{1}{144}H^2\left(12 - 5H^2\right) & \frac{H^4}{96} & 1 - \frac{H^2}{12} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \frac{-95H^6+840H^4-3744H^2}{3456} & \frac{H^2(25H^4-60H^2+144)}{1728} & -\frac{5H^6}{1152} & \frac{5(H^2-12)H^2+1}{144} \end{pmatrix}$$

and the stability interval is given by (0, 0.585417). In the second case the stability matrix is given by

$$\begin{pmatrix} \frac{19H^4}{288} - \frac{7H^2}{12} + 1 & \frac{1}{144}H^2\left(12 - 5H^2\right) & \frac{H^4}{96} & 1 - \frac{H^2}{12} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \frac{1}{288}H^2\left(95H^2 - 312\right) & \frac{1}{144}H^2\left(12 - 25H^2\right) & \frac{5H^4}{96} & 1 - \frac{5H^2}{12} \end{pmatrix}$$

for which there are no stability nor periodicity intervals. This time, the first procedure is the best one, concerning the stability characteristics.

In Fig. 1 the stability regions are depicted (left: taking as predictor the two-step explicit Falkner method and using $f(x_{n+1}, y^P_{n+1})$ in the formula in (9); right: taking as predictor the three-step explicit Falkner method and using $f(x_{n+1}, y^C_{n+1})$ in the formula in (9)).

These unexpected results show the importance of determining the stability regions in order to obtain appropriate results when applying a numerical method. The numerical results in the following section will confirm this claim (Tables 3 and 5 provide a confirmation of the theoretical expectations on the stability of the derived methods).

For the methods presented in the previous sections we have obtained that they are not A-stable nor P-stable up to $k = 14$ (it is also known that there are no A-stable Adams methods for $y' = f(x, y)$). There exists only one interval of periodicity corresponding to the third mode when $k = 1$ being this interval $[0, \sqrt{6}]$. In Table 2 the intervals of stability are presented from $k = 1$ up to $k = 14$, where $k$ refers to the number of steps, for the different formulations, named after $FI[1]k$, $FI[2]k$, and $FI[3]k$ respectively, according to the notation used in the above sections.

All these values have been obtained with the help of the Mathematica program. We have considered only the primary stability intervals although for different

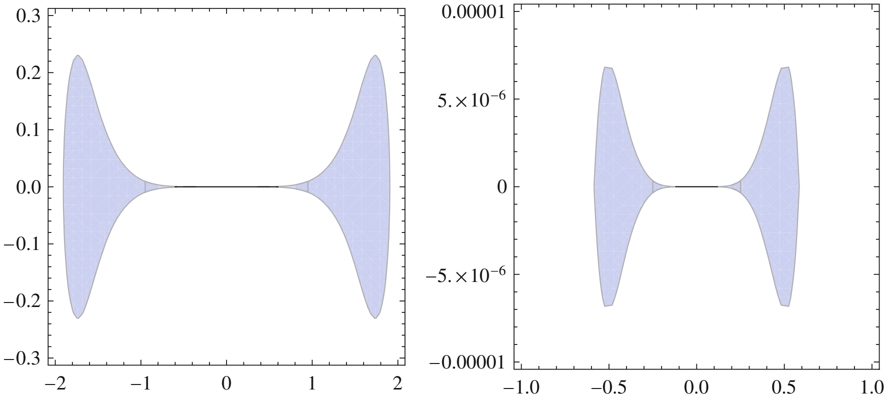


**Fig. 1** Absolute stability regions for different formulations of Gear's method in P-C mode (left: $P_2ECC'$ ; right: $P_3ECEC'$)

**Table 2** Intervals of stability for the different formulations of the Falkner implicit method in P-C modes

| $k$ | $FI[1]k$ | $FI[2]k$ | $FI[3]k$ |
|---|---|---|---|
| 1 | ∅ | (0, 2) | ∅ |
| 2 | ∅ | ∅ | ∅ |
| 3 | (0, 0.853006) | ∅ | ∅ |
| 4 | (0, 0.930949) | (0, 0.534947) | (0, 1.108998) |
| 5 | ∅ | (0, 0.925569) | (0, 1.409664) |
| 6 | ∅ | ∅ | ∅ |
| 7 | (0, 0.401364) | ∅ | ∅ |
| 8 | (0, 0.285341) | (0, 0.274630) | (0, 0.480033) |
| 9 | ∅ | (0, 0.521074) | (0, 0.821956) |
| 10 | ∅ | ∅ | ∅ |
| 11 | (0, 0.100705) | ∅ | ∅ |
| 12 | (0, 0.071103) | (0, 0.185094) | (0, 0.300133) |
| 13 | ∅ | (0, 0.362275) | (0, 0.390144) |
| 14 | ∅ | ∅ | ∅ |

methods may exist secondary stability intervals, as for the second P-C mode for $k = 2$, for which [1, 2] is an interval where $|r_i| < 1$ , $i = 1, 2, 3$.

We have included different plots showing the stability intervals. Figure 2 shows the absolute value of the eigenvalues of $M[2]$ and $M[3]$ for $k = 1$ versus $H$, and the resulting stability interval $(0, H_s)$. For the matrix $M[3]$ we have included the interval of periodicity $(0, H_p)$. Figure 3 shows the absolute value of the eigenvalues of $M[1]$ and $M[2]$ for $k = 3$ versus $H$. This time, for the $M[2]$ mode there is no primary stability interval, although a secondary interval of stability exists, (1.731413, 1.895953).

Figure 4 shows the absolute value of the eigenvalues of $M[1]$ and $M[3]$ for $k = 5$, showing in case of $M[1]$ a secondary stability interval given by (0.626085, 0.752697). Finally, Fig. 5 shows two eigenvalues limiting the stability

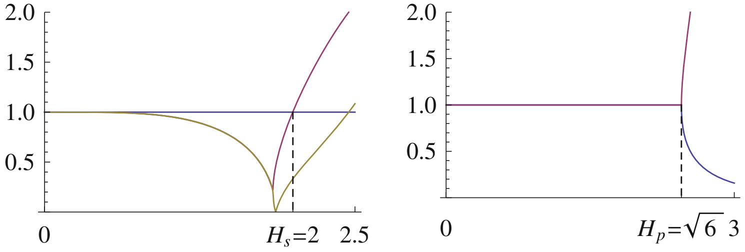


**Fig. 2** Absolute value of the eigenvalues $|r_i|$, $i = 1, 2$ of the stability matrices for the P-C modes: $M[2]$ (left) and $M[3]$ (right) when $k = 1$

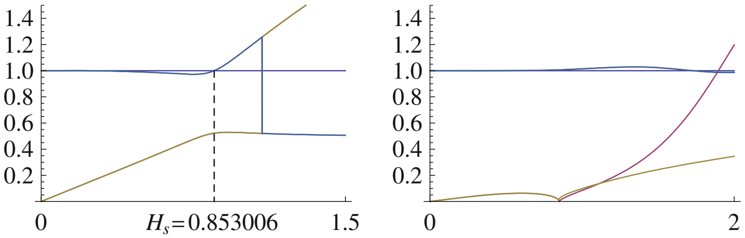


**Fig. 3** Absolute value of the eigenvalues $|r_i|$, $i = 1, 2, 3, 4$ of the stability matrices for the P-C modes: $M[1]$ (left) and $M[2]$ (right) when $k = 3$

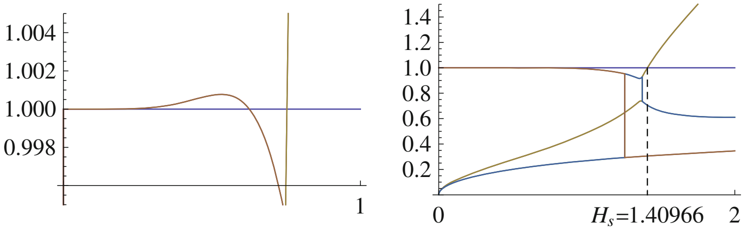


**Fig. 4** Absolute values of the eigenvalues $|r_i|$, $i = 1, \ldots, 6$ of the stability matrices for the P-C modes: $M[1]$ (left) and $M[3]$ (right) when $k = 5$

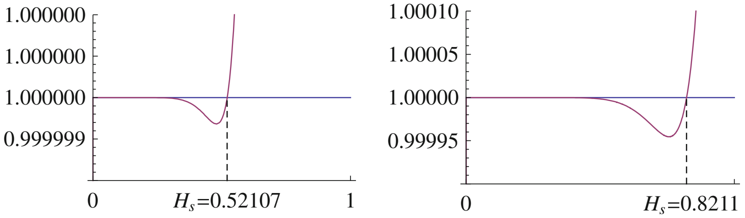


**Fig. 5** Absolute value of two eigenvalues limiting the stability region for each of the stability matrices for the P-C modes: $M[2]$ (left) and $M[3]$ (right) when $k = 9$

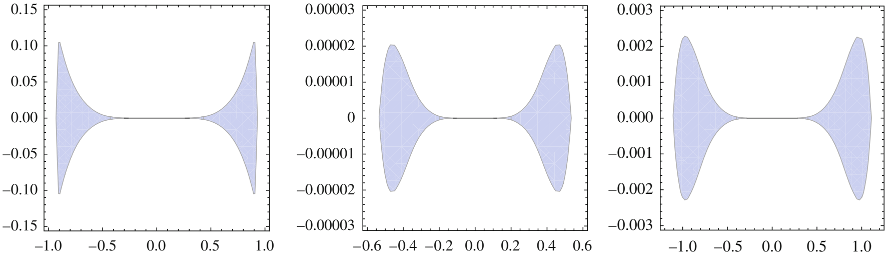


**Fig. 6** Stability regions for each of the methods $FI[i]4$ , $i = 1, 2, 3$ (from left to right)

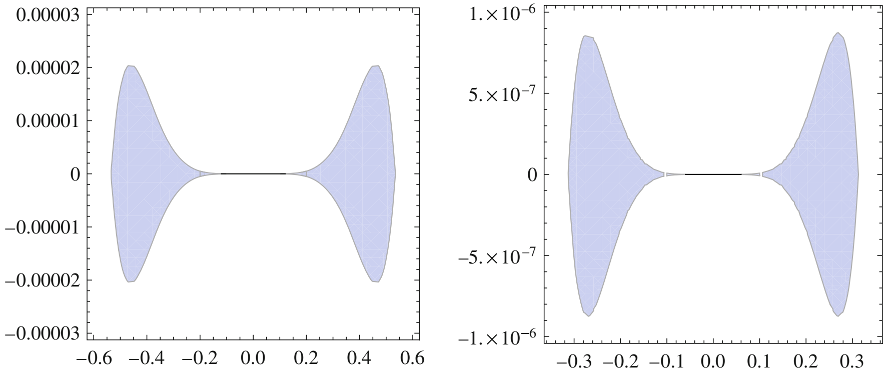


**Fig. 7** Stability regions for the methods $FI[2]4$ and $\widetilde{FI}[2]4$

regions in case of $M[2]$ and $M[3]$ when $k = 9$. We have included as a reference in all the plots the horizontal line at a distance of one unit above the horizontal axis.

In Fig. 6 appear the stability regions for the methods $FI[i]4$ , $i = 1, 2, 3$.

*Remark 2* In the application of the P-C modes we may choose to evaluate the final E in order to update the value of $f_{n+1}$ for the next step or not. If we do not make this further evaluation the mode $FI[1]k$ results to be P' PEC instead of P' PECE, but the modes $FI[2]k$ and $FI[3]k$ results to be the same, namely PEC' C, and will be denoted by $\widetilde{FI}[i]k$, $i = 1, 2$. We make the observation that the order of the accumulated truncation errors is the same with the evaluation or not of the final E, but the stability characteristics differ markedly in both cases. In fact, if we do not make the last evaluation, the stability regions are smaller. In Fig. 7 we include a plot with the stability regions for the modes $FI[2]4$ and $\widetilde{FI}[2]4$.

# 5 Numerical Examples

In this section we apply to some examples the above formulations in P-C mode to illustrate the performance of the implicit Falkner method. The numerical results obtained with these methods are compared with the results obtained with other numerical methods. The predictor-corrector formulations of the implicit Falkner methods in this paper will be denoted by $FI[j]k$, where $j$ refers to the implementations 1, 2 or 3 in the above sections, and $k$ refers to the number of steps of the method. That is, number 1 refers to the P'PECE mode in the above sections, 2 refers to the PEC'CE mode, and 3 refers to the PECEC' mode.

## *5.1 Example 1*

As a first example we take the classical problem used to test absolute stability, with particular initial values

$$y''(x) = -\mu^2 y(x)\,, \quad y(0) = 0\,, \quad y'(0) = 1 \tag{36}$$

whose exact solution is

$$y(x) = \frac{\sin(\mu x)}{\mu}\,.$$

The problem has been integrated on the interval $[0, 5\pi]$ with $\mu = 100$, considering the explicit and implicit implementations, using different values of $N$ and $k$. Table 3 shows the results obtained with $k = 7, 8$ and $N = 3000, 4000$. The values of $h\mu$, the CPU times (in seconds) and the maximum absolute error on the integration interval

$$MaxErr = \max_{j\in 0,\ldots,N} |y(x_j) - y_j|\,,$$

are included.

We observe that the time needed by implicit methods is bigger than for the explicit ones, as expected. Nevertheless, the accuracies obtained by the explicit eight order methods $FEAB8$ and $FEAM8$ are not acceptable. Concerning the $FEAB7$ and $FEAM7$ methods, for $N = 3000$ the errors are enormous, while for $N = 4000$ they are acceptable. These results are related with the stability intervals (see Table 2 in [23]) since when $N = 4000$ the value $h\mu = 0.3926$ is inside the absolute stability intervals of the methods $FEAB7$ and $FEAM7$, and it is not in other cases.

In the case of implicit methods there are only two situations in which the value of $h\mu$ is within the ranges of the primary interval of absolute stability: for $FI[1]7$ and for $FI[3]8$ when $N = 4000$ (see Table 2). Why in some of the other cases the

**Table 3** Maximum absolute error of the solution for different implementations of the explicit and implicit Falkner methods in P-C mode for solving the problem in (36)

| Method | N steps | $h\mu$ | $CPU(s.)$ | $MaxErr(y)$ |
|---|---|---|---|---|
| $FEAB7$ | 3000 | 0.5235 | 0.344 | $3.0150 \times 10^{50}$ |
| | 4000 | 0.3926 | 0.484 | $3.1289 \times 10^{-3}$ |
| $FEAB8$ | 3000 | 0.5235 | 0.391 | $6.0959 \times 10^{252}$ |
| | 4000 | 0.3926 | 0.547 | $2.7399 \times 10^{85}$ |
| $FEAM7$ | 3000 | 0.5235 | 0.359 | $3.2187 \times 10^{114}$ |
| | 4000 | 0.3926 | 0.500 | $2.7601 \times 10^{-4}$ |
| $FEAM8$ | 3000 | 0.5235 | 0.422 | $3.0925 \times 10^{367}$ |
| | 4000 | 0.3926 | 0.547 | $1.3965 \times 10^{160}$ |
| $FI[1]7$ | 3000 | 0.5235 | 0.594 | $1.9224 \times 10^{185}$ |
| | 4000 | 0.3926 | 0.813 | $3.2240 \times 10^{-3}$ |
| $FI[1]8$ | 3000 | 0.5235 | 0.703 | $2.3554 \times 10^{387}$ |
| | 4000 | 0.3926 | 0.906 | $1.8158 \times 10^{259}$ |
| $FI[2]7$ | 3000 | 0.5235 | 0.610 | $5.4653 \times 10^{-4}$ |
| | 4000 | 0.3926 | 0.813 | $4.9368 \times 10^{-5}$ |
| $FI[2]8$ | 3000 | 0.5235 | 0.688 | $2.4507 \times 10^{-4}$ |
| | 4000 | 0.3926 | 0.921 | $1.6627 \times 10^{-5}$ |
| $FI[3]7$ | 3000 | 0.5235 | 0.610 | $4.5078 \times 10^{-4}$ |
| | 4000 | 0.3926 | 0.812 | $4.3100 \times 10^{-5}$ |
| $FI[3]8$ | 3000 | 0.5235 | 0.672 | $1.9418 \times 10^{-4}$ |
| | 4000 | 0.3926 | 0.922 | $1.4224 \times 10^{-5}$ |

errors seem to behave well? There are two possible reasons, first, there may be a secondary stability interval of the form $[a, b]$ with $a > 0$, where the eigenvalues of the stability matrix are less than unity. Nevertheless, this situation does not occur with the data in Table 3 for this example.

On the other hand, it may occur that for some specific data, the eigenvalues of the stability matrix although more than one, are very close to 1. In this situation the errors grow very slowly in each step, indicating that for these values of N the instability does not affect the results yet. This is the case for the data in Table 3. For example, for the method FI[2]8 the largest eigenvalues are 1.0000079 when $N = 3000$, and 1.00000024 when $N = 4000$, which justifies the not so bad behavior of the method.

## 5.2 *Example 2*

The next example has appeared different times in the literature (see [4, 22]). The problem consists in

$$y''(t) = -w^2 y(t) + (w^2 - 1)\sin(t)\,, \quad y(0) = 1\,, \quad y'(0) = w + 1 \tag{37}$$

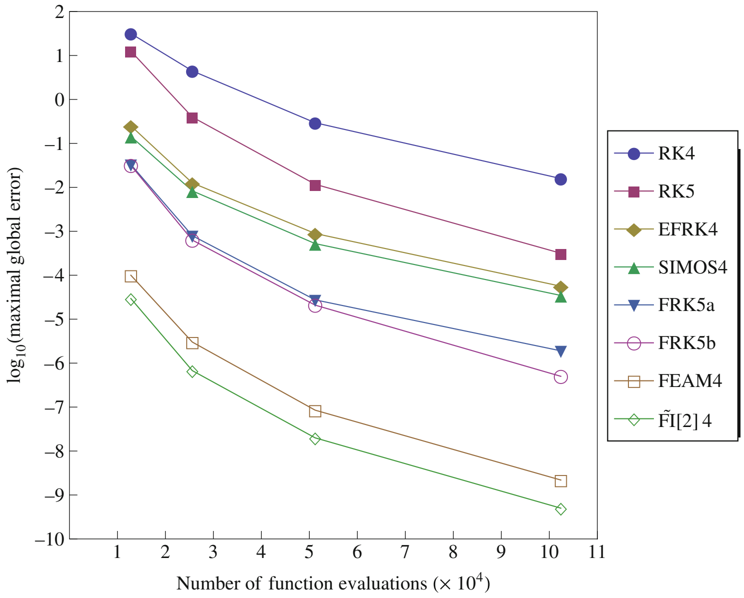


**Fig. 8** Numerical results for Problem (37) with different Runge-Kutta type methods and with the methods $FEAM4$ and $\widetilde{FI}[2]4$

whose exact solution is

$$y(t) = \cos(wt) + \sin(wt) + \sin(t)\,.$$

This problem has been solved in the interval [0, 100] with the methods $FEAM4$ and $\widetilde{FI}[2]4$ taking $w = 20$. Figure 8 shows the efficiency of different methods appeared in the literature, the horizontal axis stands for the number of function evaluations required and the vertical axis stands for the digital logarithm of the maximal global error for the solution. The methods used for comparison are

- RK4: the classical RK method of order four
- RK5: the classical RK method of order five presented in [13]
- EFRK4: the four-stage exponentially fitted RK method of order four given in [30]
- SIMOS4: the RK method of order four presented in [27]
- FRK5a: the seven-stage RK method of order five given by (4.17) and (4.20) in [4]
- FRK5b: the seven-stage RK method of order five given by (4.17) and (4.28) in [4]

It is obvious from the graphs that the Falkner methods perform better.

**Table 4** Data for the problem in (38) with different methods

| Method | F.Ev. | $MaxErr$ |
|---|---|---|
| Method 1 | 14,000 | $3.4 \times 10^{-10}$ |
| Method 2 | 16,000 | $3.2 \times 10^{-10}$ |
| Method 3 | 16,000 | $3.8 \times 10^{-10}$ |
| Method 4 | 12,000 | $1.8 \times 10^{-10}$ |
| Method 5 | 6000 | $4.1 \times 10^{-10}$ |

## 5.3 *Example 3*

We consider the following nonlinear problem studied by Chawla and Rao in [3]

$$y''(x) + 100y = \sin(y)\,, \quad y(0) = 0\,, \quad y'(0) = 1 \tag{38}$$

For this example, as there is not a known analytical solution, we perform the computation on the interval $[0, 20\pi]$ and measure the absolute error at the final point taking the value $y(20\pi) = 0.000392823991$ (see [29]). In Table 4 we present the absolute global error at $x = 20\pi$ with the number of function evaluations, F.Ev., for the following methods

- Method 1: the well known Runge-Kutta-Nyström method in [8]
- Method 2: the explicit eighth order method with maximal interval of periodicity in [29]
- Method 3: the sixth order hybrid method of Chawla and Rao in [3]
- Method 4: the P-C eight order method $FI[3]8$ in this paper
- Method 5: the P-C eight order method $\widetilde{FI}[3]8$ in this paper

We note that while the errors are similar with all the methods, the methods $FI[3]8$ and $\widetilde{FI}[3]8$ produce the better performances, being the formulation $\widetilde{FI}[3]8$ which needs the smallest number of function evaluations.

## 5.4 *Example 4*

The next example consists of the mildly stiff problem given by the linear system (see [15, 16])

$$y''(x) = 2498y + 4998z\,, \quad z''(x) = -2499y - 4999z \tag{39}$$
$$y(0) = 0\,, \quad y'(0) = 1\,, \quad z(0) = 0\,, \quad z'(0) = 1$$

with the exact solution $y = 2\cos x,\ z = -\cos x$.

The eigenvalues of the matrix of the system have modulus $|\lambda_1| = 2500$, and $|\lambda_2| = 1$. We have solved this problem in $[0, 2\pi]$ considering the methods $FEAMk$ and $FI[3]k$ for $k = 5, 6, 7, 8$ taking different number of steps $N = 200, 300, 400$.

**Table 5** Maximum absolute error of the solution for different implementations of the explicit and implicit Falkner methods in P-C mode for solving the problem in (39)

| Method | N steps | $CPU(s.)$ | $MaxErr(y)$ | $MaxErr(y')$ |
|---|---|---|---|---|
| $FEAM5$ | 300 | 0.125 | $1.6005 \times 10^{20}$ | $4.4213 \times 10^{21}$ |
| | 400 | 0.188 | $5.2427 \times 10^{-12}$ | $7.1552 \times 10^{-12}$ |
| $FI[3]5$ | 200 | 0.093 | $2.5509 \times 10^{4}$ | $1.4710 \times 10^{6}$ |
| | 300 | 0.172 | $5.8347 \times 10^{-12}$ | $7.5593 \times 10^{-12}$ |
| $FEAM6$ | 300 | 0.125 | $1.9647 \times 10^{52}$ | $4.8460 \times 10^{53}$ |
| | 400 | 0.187 | $9.3717 \times 10^{20}$ | $3.1892 \times 10^{22}$ |
| $FI[3]6$ | 200 | 0.109 | $5.0378 \times 10^{9}$ | $2.3892 \times 10^{11}$ |
| | 300 | 0.187 | $1.8252 \times 10^{-13}$ | $3.4094 \times 10^{-13}$ |
| $FEAM7$ | 300 | 0.125 | $3.2773 \times 10^{77}$ | $7.4364 \times 10^{78}$ |
| | 400 | 0.187 | $2.2564 \times 10^{58}$ | $7.1188 \times 10^{59}$ |
| $FI[3]7$ | 200 | 0.110 | $2.8934 \times 10^{17}$ | $1.1389 \times 10^{19}$ |
| | 300 | 0.203 | $3.6781 \times 10^{-13}$ | $5.7642 \times 10^{-13}$ |
| $FEAM8$ | 300 | 0.125 | $3.3732 \times 10^{97}$ | $7.1745 \times 10^{98}$ |
| | 400 | 0.188 | $3.8269 \times 10^{87}$ | $1.1378 \times 10^{89}$ |
| $FI[3]8$ | 200 | 0.109 | $1.0663 \times 10^{26}$ | $1.4867 \times 10^{28}$ |
| | 300 | 0.219 | $1.0280 \times 10^{-13}$ | $1.0943 \times 10^{-12}$ |

The results appear in Table 5 where we have included the errors measured in the norm $\|\cdot\|_\infty$, and the CPU time needed.

We have considered the methods and number of steps in Table 5 for this problem to highlight how important is to choose the method and the stepsize that suit the stability requirements in order to solve the problem properly.

## 5.5 *Example 5*

The last example consists in the system that models the motion of two bodies, given by (see [26])

$$\begin{cases} y_1''(t) = \dfrac{-y_1(t)}{r^3} \\ y_2''(t) = \dfrac{-y_2(t)}{r^3} \\ y_1(0) = 1\,, \quad y_1'(0) = 0\,, \quad y_2(0) = 0\,, \quad y_2'(0) = 1\,, \end{cases} \tag{40}$$

with $r = \sqrt{y_1^2 + y_2^2}$, and exact solution

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

**Table 6** Data for the problem (40) with the methods $FEAMk$ and $\widetilde{FI}[2]k$ for $k = 2, \ldots, 10$

| k | MaxErr $y_1$ | MaxErr $y_2$ | MaxErr $y_1'$ | MaxErr $y_2'$ |
|---|---|---|---|---|
| $FEAMk$ | | | | |
| 2 | $1.1651 \times 10^{-3}$ | $1.1781 \times 10^{-3}$ | $1.2740 \times 10^{-3}$ | $8.5695 \times 10^{-3}$ |
| 3 | $1.7458 \times 10^{-5}$ | $1.9902 \times 10^{-5}$ | $1.9762 \times 10^{-5}$ | $1.6453 \times 10^{-5}$ |
| 4 | $3.9115 \times 10^{-6}$ | $3.9177 \times 10^{-6}$ | $4.2581 \times 10^{-6}$ | $2.8593 \times 10^{-6}$ |
| 5 | $5.6869 \times 10^{-8}$ | $7.3229 \times 10^{-8}$ | $7.3142 \times 10^{-8}$ | $5.6222 \times 10^{-8}$ |
| 6 | $1.3264 \times 10^{-8}$ | $1.3101 \times 10^{-8}$ | $1.4338 \times 10^{-8}$ | $9.6002 \times 10^{-9}$ |
| 7 | $2.3774 \times 10^{-10}$ | $2.9070 \times 10^{-10}$ | $2.9288 \times 10^{-10}$ | $2.1161 \times 10^{-10}$ |
| 8 | $4.5591 \times 10^{-11}$ | $4.4313 \times 10^{-11}$ | $4.8903 \times 10^{-11}$ | $3.2613 \times 10^{-11}$ |
| 9 | $9.9675 \times 10^{-13}$ | $1.1674 \times 10^{-12}$ | $1.1874 \times 10^{-12}$ | $8.1706 \times 10^{-13}$ |
| 10 | $1.5953 \times 10^{-13}$ | $1.5451 \times 10^{-13}$ | $1.7053 \times 10^{-13}$ | $1.1368 \times 10^{-13}$ |
| $\widetilde{FI}[2]k$ | | | | |
| 2 | $5.3652 \times 10^{-4}$ | $5.4163 \times 10^{-4}$ | $5.8808 \times 10^{-4}$ | $3.9961 \times 10^{-4}$ |
| 3 | $3.1679 \times 10^{-6}$ | $1.8396 \times 10^{-6}$ | $2.6416 \times 10^{-6}$ | $1.4614 \times 10^{-6}$ |
| 4 | $8.5809 \times 10^{-7}$ | $8.3725 \times 10^{-7}$ | $9.2345 \times 10^{-7}$ | $6.2228 \times 10^{-7}$ |
| 5 | $1.4127 \times 10^{-8}$ | $1.0529 \times 10^{-8}$ | $1.3377 \times 10^{-8}$ | $8.3543 \times 10^{-8}$ |
| 6 | $1.7960 \times 10^{-9}$ | $1.6812 \times 10^{-9}$ | $1.8923 \times 10^{-9}$ | $1.2603 \times 10^{-9}$ |
| 7 | $5.3236 \times 10^{-11}$ | $4.2099 \times 10^{-11}$ | $5.2048 \times 10^{-11}$ | $3.2922 \times 10^{-11}$ |
| 8 | $4.1453 \times 10^{-12}$ | $3.6718 \times 10^{-12}$ | $4.2665 \times 10^{-12}$ | $2.7894 \times 10^{-12}$ |
| 9 | $1.9606 \times 10^{-13}$ | $1.5978 \times 10^{-13}$ | $1.9451 \times 10^{-13}$ | $1.2401 \times 10^{-13}$ |
| 10 | $2.1871 \times 10^{-14}$ | $2.0983 \times 10^{-14}$ | $2.3314 \times 10^{-14}$ | $1.5543 \times 10^{-14}$ |

**Table 7** Comparison of the Euclidean norms of the end-point global errors obtained by using Simos' method [27], the classical fourth-order RK method and the methods $FEAM4$ and $\widetilde{FI}[2]3$

| h | $FEAM4$ | $\widetilde{FI}[2]3$ | Simos' method | Classical RK method |
|---|---|---|---|---|
| 0.5 | 0.06 | 0.08 | 0.07 | 0.11 |
| 0.25 | $4.12 \times 10^{-3}$ | $5.54 \times 10^{-3}$ | $8.00 \times 10^{-3}$ | $1.10 \times 10^{-2}$ |
| 0.125 | $1.62 \times 10^{-4}$ | $1.80 \times 10^{-4}$ | $6.02 \times 10^{-4}$ | $8.29 \times 10^{-4}$ |
| 0.0625 | $5.53 \times 10^{-6}$ | $3.66 \times 10^{-6}$ | $4.05 \times 10^{-5}$ | $5.53 \times 10^{-5}$ |

This problem has been used as a test in a lot of articles presenting methods for solving initial value problems. We have solved the problem in the interval [0, 7], as was done in [30], using the methods $FEABk$ and $\widetilde{FI}[2]k$ for $k = 2, \ldots, 10$, taking a number of steps $N = 112$ in all cases, which results in a stepsize $h = 0.0625$. The numerical results appear in Table 6.

Finally, in Table 7 we show the data with the adapted Runge-Kutta method in [27], the classical Runge-Kutta method and methods $FEAM4$ and $\widetilde{FI}[2]3$. We have calculated at the endpoint of the integration interval the Euclidian norm of the error vector with components defined as the difference between the numerical and the exact solutions. We note that the methods $FEAM4$ and $\widetilde{FI}[2]3$ provide better results, even without considering that the RK methods need more evaluations of the function, and thus more computational effort.

## 6 Conclusions

Falkner methods are commonly used for solving the special second-order differential initial-value problem, particularly when the values of the derivatives are also needed (particularizations of these methods are the well-known Velocity-Verlet, Beeman's method or Wilson's method). On these approaches two formulas are needed for advancing the solution $y(x)$ and the derivative $y'(x)$. When the formulas are implicit, they are usually formulated in predictor-corrector modes.

We have considered three different modes to implement the implicit Falkner formulas and have made an analysis of the accumulated truncation errors. Considering the expressions of these errors for the solution and the derivative, we observe that the modes $FI[2]k$ and $FI[3]k$ show the best performance. The resulting errors in this case are both of order $\mathcal{O}(h^{k+1})$. In view of Table 1 we note that the errors of the explicit methods $FEAMk$ are also of order $\mathcal{O}(h^{k+1})$. On the other hand, if we do not perform the last evaluation of $f$ with the implicit Falkner method in P-C modes the accumulated errors are also of order $\mathcal{O}(h^{k+1})$, but the stability characteristics are compromised. This fact could suggest that the explicit method $FEAMk$ would be the best choice because explicit methods are in general less time consuming than implicit ones. Nevertheless, the stability is a factor that should be considered (compare different performances in Table 5). We can say that if the stability requirements are fulfilled than the $FEAMk$ should be used, but in other situation the $\widetilde{FI}[2]k$ or even the $FI[3]k$ should be used. Some numerical examples are included to make a comparison of the numerical performance of the different implementations, resulting that the proposed methods may be competitive with other methods in the literature. The numerical results agree with the theoretical analysis.

**Acknowledgement** The author is pleased to acknowledge a referee for carefully reading the manuscript.

## References

1. Beeman, D.: Some multistep methods for use in molecular dynamics calculations. J. Comput. Phys. **20**, 130–139 (1976)
2. Calvo, M., Montijano, J.I., Rández, L.: A new stepsize change technique for Adams methods. Appl. Math. Nonlinear Sci. **1**, 547–558 (2016)
3. Chawla, M.M., Rao, P.S.: Numerov-type method with minimal phase-lag for the integration of second order periodic initial-value problems II. Explicit method. J. Comput. Appl. Math. **15**, 329–337 (1986)

4. Chen, Z., You, X., Shu, X., Zhang, M.: A new family of phase-fitted and amplification-fitted Runge-Kutta type methods for oscillators. J. Appl. Math. (2012). https://doi.org/10.1155/2012/236281
5. Coleman, J.P., Ixaru, L.Gr.: P-stability and exponential-fitting methods for $y'' = f(x, y)$. IMA J. Numer. Anal. **16**, 179–199 (1996)
6. Coleman, J.P., Mohamed, J.: De Vogelaere's methods with automatic error control. Comput. Phys. Commun. **17**, 283–300 (1979)
7. Collatz, L.: The Numerical Treatment of Differential Equations. Springer, Berlin (1966)
8. Dormand, J.R., El-Mikkawy, M., Prince, P.J.: High order embedded Runge-Kutta-Nystrom formulae. IMA J. Numer. Anal. **7**, 423–430 (1987)
9. Falkner, V.M.: A method of numerical solution of differential equations. Phil. Mag. (7) **21**, 621–640 (1936)
10. Franco, J.M., Gómez, I.: Accuracy and linear stability of RKN methods for solving second-order stiff problems. Appl. Numer. Math. **59**, 959–975 (2009)
11. Gear, C.W.: Argonne National Laboratory, Report no. ANL-7126 (1966)
12. Gladwell, I., Thomas, R.: Stability properties of the Newmark, Houbolt and Wilson $\theta$ methods. I. J. Numer. Anal. Meth. Geomech. **4**, 143–158 (1980)
13. Hairer, E., Norsett, S.P., Wanner, G.: Solving Ordinary Differential Equations I. Springer, Berlin (1987)
14. Henrici, P.: Discrete Variable Methods in Ordinary Differential Equations. Wiley, New York (1962)
15. Kramarz, L.: Stability of collocation methods for the numerical solution of $y" = f(x, y)$. BIT **20**, 215–222 (1980). https://doi.org/10.1007/BF01933194
16. Krishnaiah, U.A.: Adaptive methods for periodic initial value problems of second order differential equations. J. Comput. Appl. Math. **8**, 101–104 (1982)
17. Krogh, F.T.: Issues in the Design of a Multistep Code. JPL Technical Report (1993). http://hdl.handle.net/2014/34958
18. Lambert, J.D.: Numerical Methods for Ordinary Differential Systems. Wiley, Chichester (1991)
19. Lambert, J.D., Watson, I.A.: Symmetric multistep methods for periodic initial value problems. J. Inst. Math. Appl. **18**, 189–202 (1976)
20. Li, J.: A family of improved Falkner-type methods for oscillatory systems. Appl. Math. Comput. **293**, 345–357 (2017)
21. Li, J., Wu, X.: Adapted Falkner-type methods solving oscillatory second-order differential equations. Numer. Algorithms **62**, 355–381 (2013)
22. Paternoster, B.: Runge-Kutta-Nyström methods for ODEs with periodic solutions based on trigonometric polynomials. Appl. Numer. Math. **28**, 401–412 (1998)
23. Ramos, H., Lorenzo, C.: Review of explicit Falkner methods and its modifications for solving special second-order IVPs. Comput. Phys. Commun. **181**, 1833–1841 (2010)
24. Ramos, H., Singh, G., Kanwar, V., Bhatia, S.: An efficient variable step-size rational Falkner-type method for solving the special second-order IVP. Appl. Math. Comput. **291**, 39–51 (2016)
25. Ramos, H., Mehta, S., Vigo-Aguiar, J.: A unified approach for the development of $k$-step block Falkner-type methods for solving general second-order initial-value problems in ODEs. J. Comput. Appl. Math. **318**, 550–564 (2017)
26. Shampine, L.F., Gordon, M.K.: Computer Solution of Ordinary Differential Equations. The Initial Value Problem. Freeman, San Francisco (1975)
27. Simos, T.E.: An exponentially-fitted Runge-Kutta method for the numerical integration of initial-value problems with periodic or oscillating solutions. Comput. Phys. Commun. **115**, 1–8 (1998)
28. Toxvaerd, S.: A new algorithm for molecular dynamics calculations. J. Comput. Phys. **47**, 444–451 (1982)
29. Tsitouras, Ch., Simos, T.E.: Explicit high order methods for the numerical integration of periodic initial-value problems. Appl. Math. Comput. **95**, 15–26 (1998)

30. Vanden Berghe, G., De Meyer, H., Van Daele, M., Van Hecke, T.: Exponentially fitted Runge-Kutta methods. J. Comput. Appl. Math. **125**, 107–115 (2000)
31. Vigo-Aguiar, J., Ramos, H.: Variable stepsize implementation of multistep methods for $y'' = f(x, y, y')$. J. Comput. Appl. Math. **192**, 114–131 (2006)
32. Wilson, E.L.: A computer program for the dynamic stress analysis of underground structures. SESM Report No. 68-1, Division Structural Engineering and Structural Mechanics, University of California, Berkeley (1968)

# Application of a Local Discontinuous Galerkin Method to the 1D Compressible Reynolds Equation

Iñigo Arregui, J. Jesús Cendán, and María González

**Abstract** In this work we present a numerical method to approximate the solution of the steady-state compressible Reynolds equation with additional first-order slip flow terms. This equation models the hydrodynamic features of read/write processes in magnetic storage devices such as hard disks. The numerical scheme is based on the local discontinuous Galerkin method proposed by Cockburn and Shu (SIAM J Numer Anal 35:2440–2463, 1998), which shows good properties in the presence of internal layers appearing in convection-diffusion problems. Several test examples illustrate the good performance of the method.

**Keywords** Compressible flows · Reynolds equation · Local discontinuous Galerkin method

## 1 Introduction

Contact between different surfaces is present in a large variety of common engines. As the contacting surfaces are generally moving with respect to each other, a lubricant is used to limit friction and wear; rolling element bearings, gears, cams and tappets are well known examples.

Although oil and grease are the most popular lubricants, other fluids can also play this role. In particular, the air acts as a lubricant in devices as a hard disk, for example. In this case, a reading head and a (rigid) disk are separated by a thin air layer, allowing the flow of information between both elements. A similar and more complex case arises when the data are stored in a flexible tape.

The configuration of the air gap is crucial for an optimal design of such devices. Manufacturers and engineers have realized that the introduction of some grooves or slots in the reading head leads to a more performing transfer of information.

I. Arregui (✉) · J. J. Cendán · M. González
Departamento de Matemáticas, Universidade da Coruña, Coruña, Spain
e-mail: arregui@udc.es; suceve@udc.es; mgtaboad@udc.es

© Springer Nature Switzerland AG 2019

J. L. García Guirao et al. (eds.), *Recent Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 18,
https://doi.org/10.1007/978-3-030-00341-8_4

Reynolds [12] published in 1886 the equation describing the fluid flow in narrow gaps, in order to explain the pressure generation in journal bearings. This equation became the basis of lubrication theory and the starting point of later studies (see [3, 6, 11], for example). Among them, the compressible Reynolds equation describes the pressure in a layer of gas confined between two solid bodies when the temperatures of the bodies' surfaces are equal and constant, and the gas is isothermal. When the thickness of the gaseous fluid layer is of the order of the molecular mean-free path of the gas, a modified compressible Reynolds lubrication equation for slip flow is used (see [6]). Both equations are time-dependent, nonlinear convection-diffusion equations.

In some applications, the thickness of the gaseous fluid layer is very small at one or more regions. This causes the appearance of internal layers near the points of minimum thickness [4]. Advanced techniques are needed to accurately simulate these internal layers.

Mathematical modelling and numerical simulation are powerful tools in lubrication, as well as in other fields of engineering. Among the most used techniques, discontinuous Galerkin methods have experienced an important development in the last 20 years. The possibility of treating discontinuities in a natural way is an advantage over other methods as finite differences or finite elements.

Local discontinuous Galerkin (LDG) methods are a generalization of the method introduced in [2] by Bassi and Rebay for the compressible Navier-Stokes equations. These methods are highly parallelizable and exhibit high-order accuracy for time-dependent, convection-dominated flows. Another interesting feature is that they are well-suited to use in combination with adaptive algorithms.

The basic idea of LDG methods consists in writing the problem equivalently as a first-order system, and then discretize this first-order system using discontinuous finite elements. In [8], Cockburn and Shu proved that it is possible to achieve nonlinear stability for some types of nonlinear, time-dependent convection-diffusion equations by properly rewriting the problem. Their method is $k$-th order accurate in the linear case, provided that polynomials of degree at most $k$ are used (they also showed that this rate of convergence is sharp for LDG methods).

We are interested in the adaptive solution of the (modified) compressible Reynolds lubrication equation using LDG methods. As a first step, we consider in this work the one-dimensional steady-state compressible Reynolds lubrication equation. We follow the ideas in [8] and propose a local discontinuous Galerkin method to solve the problem. Finally, we provide some numerical results that illustrate the performance of the method.

The structure of the paper is the following. In Sect. 2, we pose the mathematical model. Section 3 is devoted to the numerical method. In Sect. 4 we show some academic and realistic tests, and Sect. 5 provides the conclusions and future work.

## 2 The Model Problem

In magnetic storage devices, heads are designed so that a thin air film is generated between the head and the storage unit (hard disk or floppy disk). Thus, hydrodynamic and elastohydrodynamic lubrication theories govern these kinds of processes.

In this framework, we consider the one-dimensional steady-state modified compressible Reynolds lubrication equation for slip flow. After an adequate scaling of the variables [4], the equation can be written in the following form:

$$\begin{cases} \dfrac{d}{dx}(H\,P) - \dfrac{d}{dx}\left((\alpha + \beta\,H\,P)\,H^2\,\dfrac{dP}{dx}\right) = 0 & \text{in } \Omega = (\ell_1, \ell_2), \\ P = 1 & \text{on } \partial\Omega, \end{cases} \tag{1}$$

where the unknown $P$ represents the air pressure, while $H$ is the function that describes the gap between the two surfaces. As previously indicated we only study the rigid case, thus $H$ is a known function. Moreover, $x$ represents the spatial coordinate, while $\alpha$ and $\beta$ include some mechanical features of the device. More precisely,

$$\alpha := \frac{\lambda\,P_a}{\mu\,v} \quad \text{and} \quad \beta := (6\,\mu\,v)^{-1},$$

where $P_a$ is the ambient pressure, $v$ denotes the disk velocity, $\lambda$ is the particle mean-free length and $\mu$ is the air viscosity. In practical applications, $\alpha$ and $\beta$ are much smaller than unity and the differential equation in (1) becomes convection-dominated.

Different works in the literature are devoted to the mathematical analysis of the compressible Reynolds equation [7, 10]. It can be shown that this problem has a positive weak solution $P$ (cf. [5]). We remark that when $\alpha = 0$ ($\lambda = 0$), we recover the classical steady-state compressible Reynolds lubrication equation.

## 3 Numerical Method

We observe that problem (1) can be rewritten equivalently as follows:

$$\begin{cases} \dfrac{d}{dx}\big(f(u) - a(u)\,\dfrac{du}{dx}\big) = 0 & \text{in } \Omega, \\ u = H & \text{on } \partial\Omega, \end{cases} \tag{2}$$

where $u = PH$, $f(u) := u(1 + (\alpha + \beta u)H')$ and $a(u) := (\alpha + \beta u)H$. We remark that this problem has a positive weak solution $u^*$ (then, $a(u^*) \geq 0$).

Next, we follow [8] and introduce the new variable $q := a(u)^{\gamma} \frac{du}{dx}$ as a further unknown, where $\gamma \in (0, 1]$. For $\gamma = \frac{1}{2}$, we recover the method proposed in [8]. Then, problem (2) can be written as the following first-order system:

$$\begin{cases} \dfrac{d}{dx}\big(f(u) - a(u)^{1-\gamma}\, q\big) = 0 \quad \text{in } \Omega, \\ q - \dfrac{d}{dx} g(u) = 0 \quad \text{in } \Omega, \\ u = H \text{ on } \partial\Omega, \end{cases} \tag{3}$$

where $g(u) := \int^{u} a(s)^{\gamma}\, ds$. The LDG method for (1) is obtained by discretizing (3) with the discontinuous Galerkin method.

Let $\{x_{j+1/2}\}_{j=0}^{N}$ be a partition of $\Omega$. For $j = 1, \dots, N$, we denote $I_j = (x_{j-1/2}, x_{j+1/2})$ and $\Delta x_j = x_{j+1/2} - x_{j-1/2}$. We also denote $\Delta x = \max_{1 \le j \le N} \Delta x_j$. Finally, we let $\mathcal{P}_k(I)$ denote the space of polynomials in $I$ of degree at most $k$ and define

$$V_h := \{v \in L^1(\Omega) \,:\, v|_{I_j} \in \mathcal{P}_k(I_j), \quad j = 1, \dots, N\}\,.$$

We look for an approximation $\mathbf{w}_h = (u_h, q_h) \in V_h \times V_h$ to a solution $\mathbf{w} = (u, q)$ of (3).

Multiplying the two first equations of (3) by arbitrary smooth functions and integrating in $I_j$, we obtain the *flux formulation* of the problem:

$$\begin{cases} -\displaystyle\int_{I_j} h_u(\mathbf{w}) \frac{dv_u}{dx}\, dx \,+\, h_u(\mathbf{w}(x_{j+1/2}))\, v_u(x_{j+1/2}^{-}) - h_u(\mathbf{w}(x_{j-1/2}))\, v_u(x_{j-1/2}^{+}) = 0 \\ \displaystyle\int_{I_j} q\, v_q\, dx - \int_{I_j} h_q(\mathbf{w}) \frac{dv_q}{dx}\, dx \\ \qquad + h_q(\mathbf{w}(x_{j+1/2}))\, v_q(x_{j+1/2}^{-}) - h_q(\mathbf{w}(x_{j-1/2}))\, v_q(x_{j-1/2}^{+}) = 0 \end{cases}$$

where

$$\mathbf{h}(u, q) = \begin{pmatrix} h_u(u, q) \\ h_q(u, q) \end{pmatrix} = \begin{pmatrix} f(u) - a(u)^{1-\gamma}\, q \\ -g(u) \end{pmatrix}.$$

Next, we replace the smooth functions $v_u$ and $v_q$ by test functions, $v_{h,u}$ , $v_{h,q} \in V_h$, respectively, and the exact solution $\mathbf{w}$ by its approximation, $\mathbf{w}_h$. We recall that the components of $\mathbf{w}_h$ may be discontinuous and hence, we must replace the nonlinear flux $\mathbf{h}(\mathbf{w}(x_{j+1/2}))$ by a numerical flux, $\hat{\mathbf{h}}(\mathbf{w}_h)_{j+1/2}$. Then, the approximate solution $\mathbf{w}_h$ given by the LDG method can be obtained by solving the following set of equations ($j = 1, \dots, N$):

$$\left\{ \begin{aligned} & -\int_{I_j} h_u(\mathbf{w}_h) \frac{dv_{h,u}}{dx}\, dx \\ & \qquad + \hat{h}_u(\mathbf{w}_h)_{j+1/2}\, v_{h,u}(x^-_{j+1/2}) - \hat{h}_u(\mathbf{w}_h)_{j-1/2}\, v_{h,u}(x^+_{j-1/2}) = 0 \\ & \int_{I_j} q_h\, v_{h,q}\, dx - \int_{I_j} h_q(\mathbf{w}_h) \frac{dv_{h,q}}{dx}\, dx \\ & \qquad + \hat{h}_q(\mathbf{w}_h)_{j+1/2}\, v_{h,q}(x^-_{j+1/2}) - \hat{h}_q(\mathbf{w}_h)_{j-1/2}\, v_{h,q}(x^+_{j-1/2}) = 0 \end{aligned} \right. \tag{4}$$

for any $v_{h,u}, v_{h,q} \in \mathcal{P}_k(I_j)$, with $u_h = H$ on $\partial\Omega$.

The numerical fluxes are defined so that they reflect the convection-diffusion nature of the problem:

$$\hat{\mathbf{h}}(\mathbf{w}_h)_{j+1/2} = \begin{pmatrix} \hat{f}(u_h)_{j+1/2} \\ 0 \end{pmatrix} + \begin{pmatrix} -\left( \dfrac{[g(u_h)]_{j+1/2}}{[u_h]_{j+1/2}} \right)^{\frac{1-\gamma}{\gamma}} (\bar{q}_h)_{j+1/2} \\ -\overline{g(u_h)}_{j+1/2} \end{pmatrix}$$

$$+ \begin{pmatrix} -\, c(\mathbf{w}_h)_{j+1/2}\, [q_h]_{j+1/2} \\ c(\mathbf{w}_h)_{j+1/2}\, [u_h]_{j+1/2} \end{pmatrix} \tag{5}$$

where

$$\hat{f}(u_h)_{j+1/2} := \hat{f}(u_h(x^-_{j+1/2}), u_h(x^+_{j+1/2}))$$

and

$$c(\mathbf{w}_h)_{j+1/2} = c(\mathbf{w}_h(x^+_{j+1/2}), \mathbf{w}_h(x^-_{j+1/2}))$$

have to be chosen and, for a given scalar quantity $v$, we denote by $[v]_{j+1/2}$ and $\bar{v}_{j+1/2}$ (respectively) the jump and the average of $v$ in $x_{j+1/2}$:

$$[v]_{j+1/2} := v(x^+_{j+1/2}) - v(x^-_{j+1/2})\,,$$

$$\bar{v}_{j+1/2} := \frac{1}{2}\big(v(x^+_{j+1/2}) + v(x^-_{j+1/2})\big)\,.$$

In our numerical experiments, we have chosen

$$\hat{f}(u_h)_{j+1/2} := \overline{f(u_h)}_{j+1/2}$$

and

$$c(\mathbf{w}_h)_{j+1/2} := \frac{1}{2}\left(\frac{[g(u_h)]_{j+1/2}}{[u_h]_{j+1/2}}\right)^{\frac{1-\gamma}{\gamma}} .$$

Moreover, we solve the nonlinear system issued from (4)–(5) using Newton's method.

# 4 Numerical Examples

In this section, we show four numerical tests to assess the performance of the numerical methods described in the previous section. In all cases, Newton's method stopping criterium is the relative error between two consecutive approximations and the tolerance parameter is set to $10^{-4}$. Convergence is attained in a small number of iterations (3 in most cases).

## 4.1 Test 1: Numerical Convergence

In order to check the good behavior of the present method, we pose the following nonlinear problem: Given a gap function $H$, find $P$ verifying

$$\begin{cases} \dfrac{d}{dx}(H\,P) - \dfrac{d}{dx}\Big((\alpha + \beta\, H\, P)\, H^2\, \dfrac{dP}{dx}\Big) = \psi \quad \text{in } \Omega, \\ P = 1 \quad \text{on } \partial\Omega. \end{cases} \tag{6}$$

The only difference with problem (1) is the right hand side function $\psi$, which is computed for a known solution $P(x) = 1 + \sin(\pi x)$ in $\Omega = (0, 1)$.

Figures 1 and 2 show the $L^2$-errors in $u = PH$ and $q$ for a constant function $H(x) = 1$ and coefficients $\alpha = 10^{-6}$ and $\beta = 10^{-7}$. Moreover, we consider $\gamma = 0.5$. The dashed lines represent the theoretical 1st, 2nd, 3rd and 4th orders. As we can observe, a $k$-th order of convergence is attained for $P_1$ and $P_3$ approximations of $u$, while $(k+1)$-st order is attained for $P_2$ and $P_4$ polynomials. A similar behavior, with $k-1$ and $k$ orders, is observed in variable $q$ (see Fig. 2).

On the other hand, we can find in the literature (see [1, 9]) some references and numerical results obtained with a first degree gap function, $H(x) = 2 - x$, for which a right hand side function from an analytical solution is easily computed. Figure 3 shows the convergence for different polynomial approximations when the analytical solution is again $P(x) = 1 + \sin(\pi x)$ in $\Omega = (0, 1)$ and using the same parameters as before.

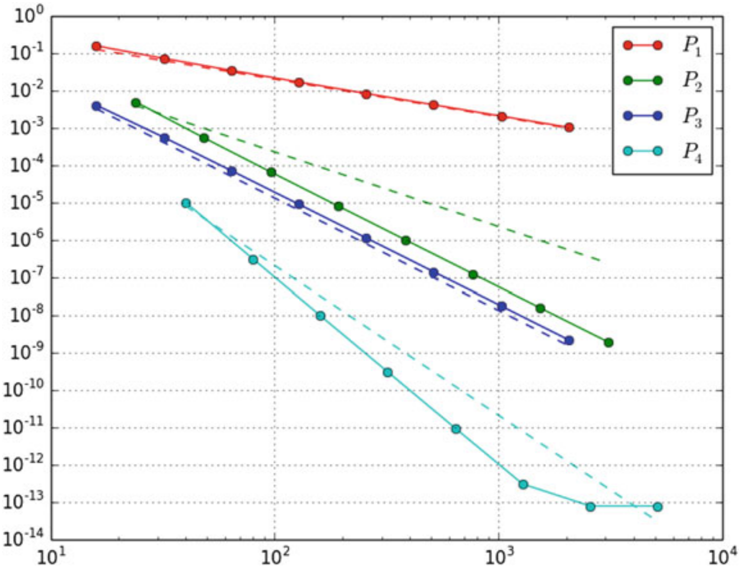


**Fig. 1** $L^2$-error in $u$ between exact and numerical solutions for different polynomial degrees with constant gap (Test 1)

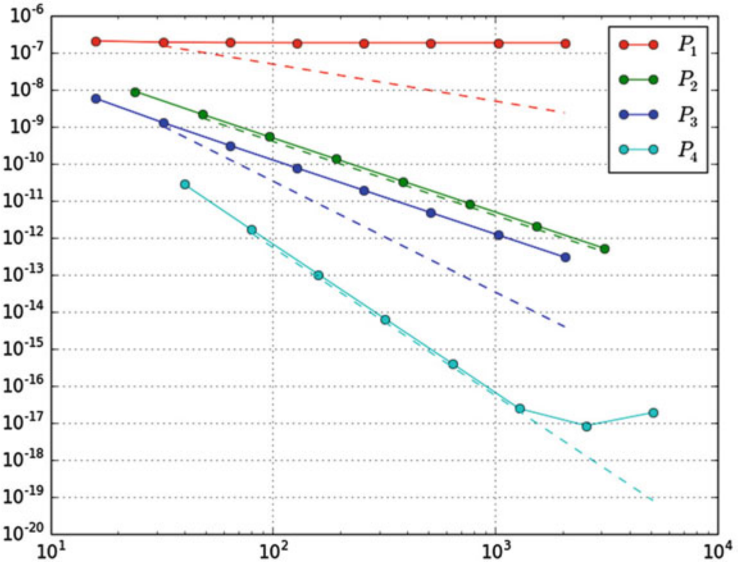


**Fig. 2** $L^2$-error in $q$ between exact and numerical solutions for different polynomial degrees with constant gap (Test 1)

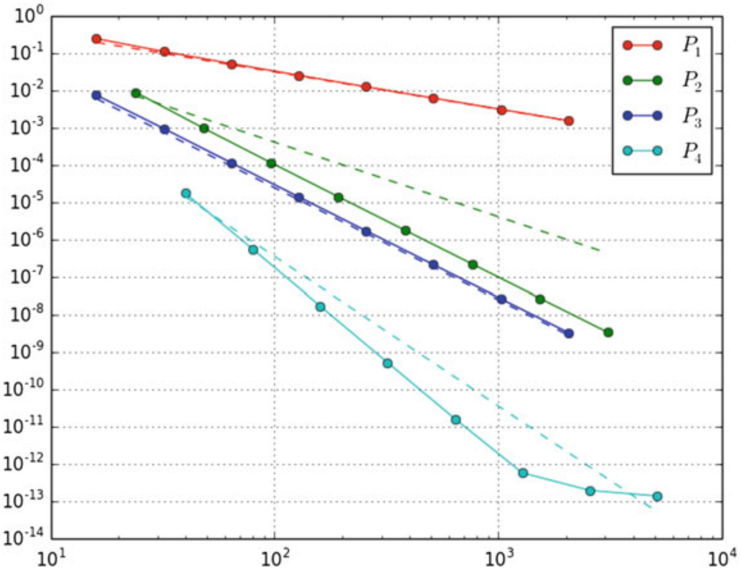


**Fig. 3** $L^2$-error in $u$ between exact and numerical solutions for different polynomial degrees with $H(x) = 2 - x$ (Test 1)

## 4.2 Test 2: p-Adaptivity

We now consider problem (1) with $H(x) = 2 - x$ in order to be closer to realistic applications. Figure 4 shows the numerical solution obtained with a $P_2$ approximation on a coarse mesh with only 15 elements. In this case, the coefficients of the PDE are $\alpha = 0.01$ and $\beta = 0.001$, while $\gamma = 0.5$.

This numerical solution is smooth in the left part of the domain, while it presents a discontinuity where the higher gradients occur. Thus, an adaptive strategy is justified. In Fig. 5 we show an improved approximation. We have initially used $P_1$ polynomials and have increased their degree for elements in which the norm of the gradient of the solution is larger; thus, after five refinement steps the solution in the last element is approximated by a sixth degree polynomial.

## 4.3 Test 3: A Rugous Surface Case

We now consider problem (1) with a different gap function:

$$H(x) = 2 - x + 0.6 \sin\left(\frac{2\pi x}{0.02}\right), \qquad x \in \Omega = (0, 1).$$

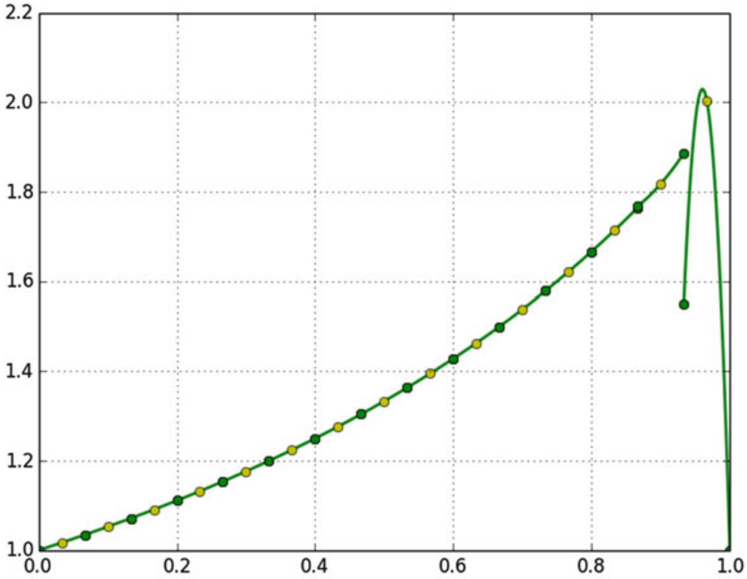


**Fig. 4** $P_2$ approximation of the pressure (Test 2)

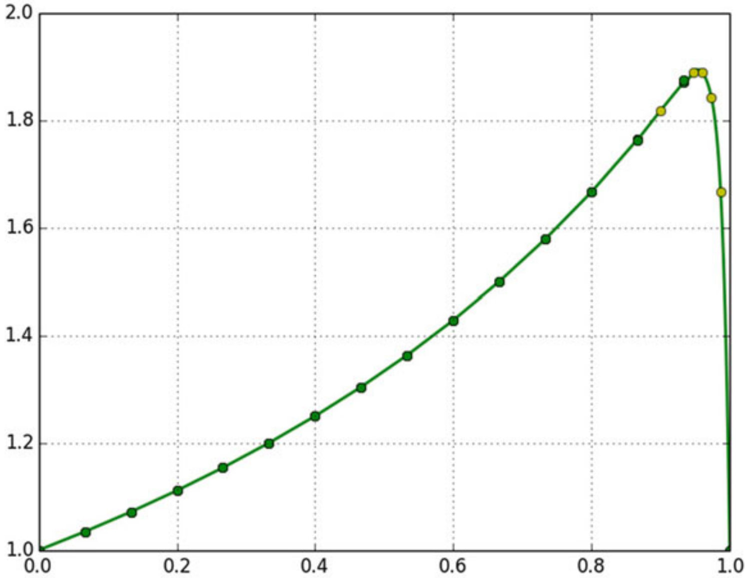


**Fig. 5** $p$-Adaptive approximation of the pressure (Test 2)

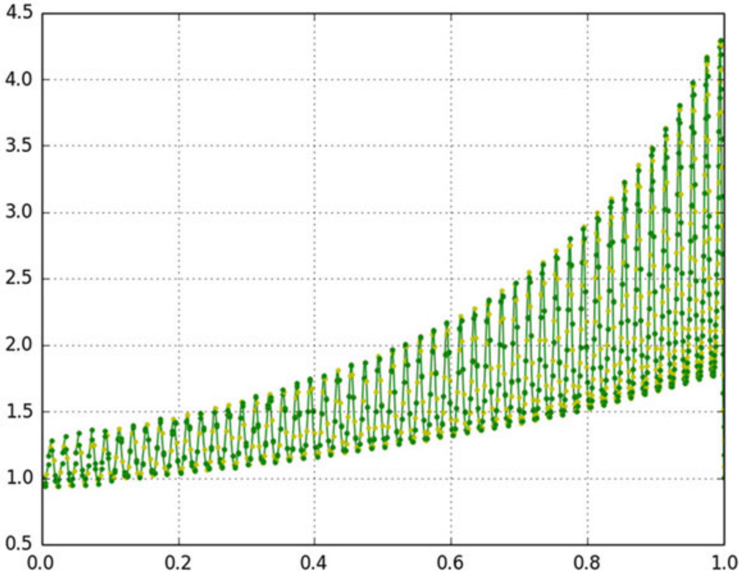


**Fig. 6** $P_2$ approximation of the pressure (Test 3)

The aim of this test is to simulate the flux in contact with a rugous surface, which can be present in real situations. The numerical solution (computed with $P_2$ polynomials on a 500 elements mesh) is shown in Fig. 6 and is in accordance with solutions obtained in [1] and [9].

In Sect. 2 we have remarked that the model problem has a positive solution $u^*$ and $a(u^*) \geq 0$. However, the approximations obtained in the different steps of the iterative algorithm do not always lead to a positive $a(u)$. That is the reason why we have used $\gamma = 0.4$ in this test, instead of the more standard value ($\gamma = 0.5$).

### *4.4 Test 4: A Head–Tape Recording System*

We now simulate a classical head–tape recording system. The domain in which Reynolds equation is considered is limited by two surfaces (the tape containing the magnetic support and the reading head) modelled by cylinders. Thus, the gap function is given by:

$$H(x) = \left[b + \sqrt{r^2 - (x - m)^2}\right] - \left[B + \sqrt{R^2 - (x - m)^2}\right],$$

with $x \in \Omega = (0.0347, 0.0497)$, $m$ being the midpoint of the interval, $R = 0.040$, $r = 0.0204$, $B = 0.0069 - R$ and $b = 0.0635 - r$. The PDE coefficients are $\alpha = 0.0332195$ and $\beta = 0.00070083$ (see [4]).

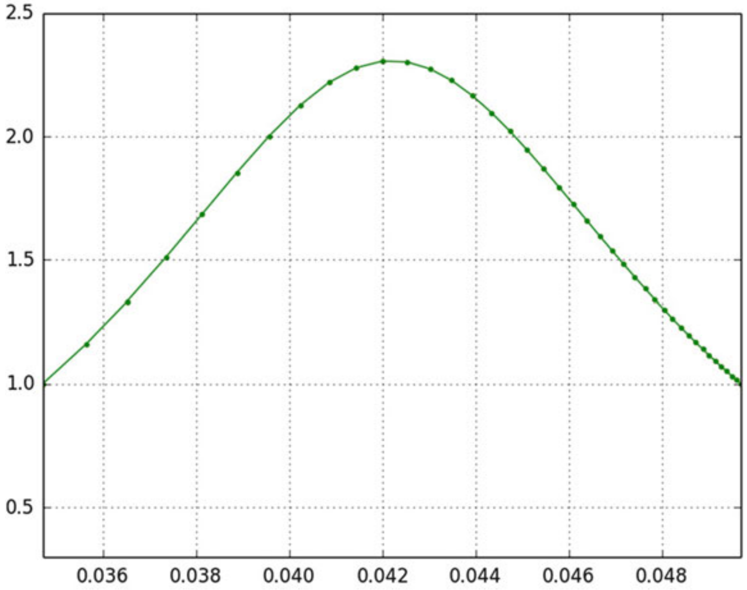


**Fig. 7** $P_1$ approximation of the pressure (Test 4)

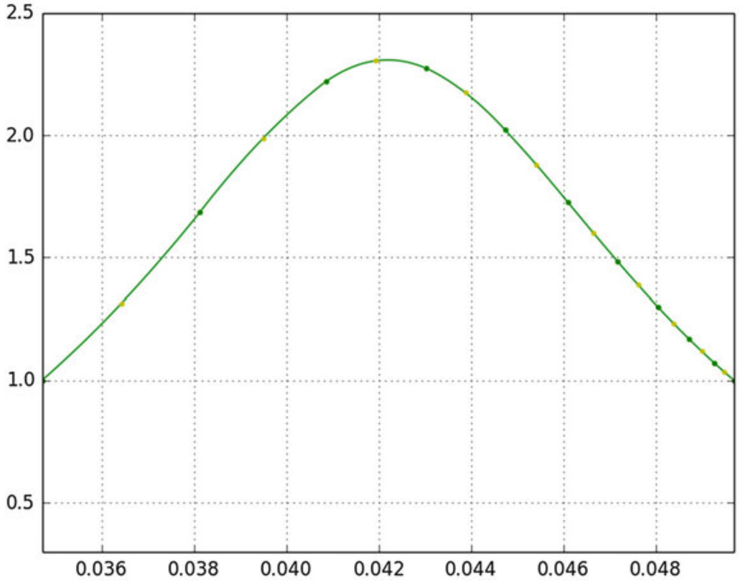


**Fig. 8** $P_2$ approximation of the pressure (Test 4)

Figure 7 shows the computed pressure using the $P_1$ local discontinuous Galerkin methods on a 40 elements mesh, while the solution obtained with $P_2$ approximation and 10 elements is plotted on Fig. 8. In both cases, $\gamma = 0.5$.

# 5 Conclusions

We propose in this work a slight modification of the method introduced in [8] to solve the modified steady-state compressible Reynolds equation for slip flow. The modification is motivated by the presence of non positive diffusion coefficient in some steps of the algorithm, due to the obtained intermediate numerical approximations.

We numerically study the convergence properties of the method. Moreover, we solve the problem of interest for different academic and realistic situations. In order to improve the efficiency of the proposed algorithm, we include an adaptive *p*-refinement. The obtained results are in accordance with those found in the literature, thus we can conclude that the local discontinuous Galerkin method is adequate to solve this kind of models.

In order to simulate industrial problems, we are considering the application of LDG methods to two-dimensional models as well as problems with discontinuous data.

**Acknowledgements** This research was partially supported by the Spanish Government (Ministerio de Economía y Competitividad) under projects MTM2013-47800-C2-1-P and MTM2016-76497-R.

# References

1. Arregui, I., Cendán, J.J., Vázquez, C.: A duality method for the compressible Reynolds equation. Application to simulation of read/write processes in magnetic storage devices. J. Comput. Appl. Math. **175**(1), 31–40 (2005)
2. Bassi, F., Rebay, S.: A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations. J. Comput. Phys. **131**, 267–279 (1997)
3. Bayada, G., Chambat, M.: The transition between the Stokes equations and the Reynolds equation: a mathematical proof. Appl. Math. Optim. **14**(1), 73–93 (1986)
4. Bhushan, B.: Tribology and Mechanics of Magnetic Storage Devices. Springer, New York (1990)
5. Buscaglia, G., Ciuperca, S., Jai, M.: Existence and uniqueness for several nonlinear elliptic problems arising in lubrication theory. J. Differ. Equ. **218**, 187–215 (2005)
6. Chipot, M., Luskin, M.: The compressible Reynolds lubrication equation. In: Metastability and Incompletely Posed Problems. IMA Volumes in Mathematics and Its Applications, vol. 3, pp. 61–75. Springer, New York (1987)
7. Ciuperca, S., Hafidi, I., Jai, M.: Analysis of a parabolic compressible first-order slip Reynolds equation with discontinuous coefficients. Nonlinear Anal. **69**, 1219–1234 (2008)
8. Cockburn, B., Shu, C.-W.: The local discontinuous Galerkin method for time-dependent convection-diffusion systems. SIAM J. Numer. Anal. **35**, 2440–2463 (1998)
9. Jai, M.: Homogenization and two-scale convergence of the compressible Reynolds lubrification equation modelling the flying characteristics of a rough magnetic head over a rough rigid-disk surface. Math. Model. Numer. Anal. **29**, 199–233 (1995)
10. Jai, M.: Existence and uniqueness of solutions of the parabolic nonlinear compressible Reynolds lubrication equation. Nonlinear Anal. **43**, 655–682 (2001)

11. Oden, J.T., Wu, S.R.: Existence of solutions to the Reynolds' equation of elastohydrodynamic lubrication. Int. J. Eng. Sci. **23**(2), 207–215 (1985)
12. Reynolds, O.: On the theory of lubrication and its application to Mr. Beauchamps tower's experiments, including an experimental determination of the viscosity of olive oil. Philos. Trans. **177**, 157–234 (1886)

# Classical Symmetries for Two Special Cases of Unsteady Flow in Nanoporous Rock

**Tamara M. Garrido, Rafael de la Rosa, and María Santos Bruzón**

**Abstract** In the present paper, we study two equations related to the theory of fluid and gas flow in nanoporous media. Both models are special cases of the basic equation of the unsteady flow in nanoporous rock, the first one is the case of weakly compressible fluid and the second one of the case of isothermic gas flow. Moreover, a generalization of the equations is presented to study thoroughly the physical phenomena. This new generalized equations involve arbitrary functions. Lie method is applied to the model of generalized unsteady flow and the model of generalized isothermic gas flow. Finally a classification in different cases depending on the arbitrary function is shown.

**Keywords** Lie point symmetries · Contact symmetries · Weakly compressible fluid · Isothermic gas flow · Nanoporous rock

## 1 Introduction

The field of multiphase phase flows in porous and nanoporous media is frequently studied because of their applications, such as, the migration of an organic contaminant in a porous media [1, 12], flow and heat transfer in biological tissues [13, 21], thermal transport in porous media [8, 20], fluid and gas flow in nanoporous media [4, 14], theory of counter flow capillary impregnation of a porous medium [10, 18, 22] and so on.

The mathematical models of these multiphase phase flows are reaching a high level in several fields and in the present paper we have focused our attention in one of them, specifically, in the study of two special cases from the basic model of the unsteady flow in nanoporous rock.

T. M. Garrido (✉) · R. de la Rosa · M. S. Bruzón
Departamento de Matemáticas, Universidad de Cádiz, Cádiz, Spain
e-mail: tamara.garrido@uca.es; rafael.delarosa@uca.es; m.bruzon@uca.es

© Springer Nature Switzerland AG 2019

J. L. García Guirao et al. (eds.), *Recent Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 18,
https://doi.org/10.1007/978-3-030-00341-8_5

To introduce both models, let us start from the equation of mass conservation in a permeable medium, from it the following relation was obtained in [15]

$$\phi\rho_t = \nabla\left(\frac{k\rho}{\mu}\nabla p\right), \tag{1}$$

where $\rho$ is the fluid density, $\phi$ is the porosity, $p$ is the pressure, $\mu$ is the dynamic fluid viscosity and $k$ the permeability of the rock.

Next, Monteiro et al. adapted the model (1) considering the following relation $k = A(p_x)^m$, where $A$ and $m$ are assumed to be constant, between pressure and density. Therefore, they obtained the basic equation of the unsteady flow in nanoporous rock

$$\phi\rho_t = \frac{A}{\mu}\left(\rho p_x^{m+1}\right)_x. \tag{2}$$

Now, Eq. (2) is adapted to two different cases. The first one with a weakly compressible fluid, so it fulfils $\rho = \rho_0(1 - \alpha(p - p_0))$ where $\alpha$ is the fluid compressibility coefficient and $\rho_0$, $p_0$ are the density and pressure values in the pristine estate. For this case Eq. (2) takes the form

$$p_t = \chi\left(p_x^{m+1}\right)_x. \tag{3}$$

On the other hand, considering an isothermic gas flow, Eq. (2) complies with the relation $\rho = Cp$ where $C$ is a given constant. Thus, the following equation arises

$$p_t = \kappa\left(p p_x^{m+1}\right)_x. \tag{4}$$

Equation (3) represents the case of weakly compressible fluid whereas Eq. (4) represents the case for isothermic gas flow. For these special cases, the constants are $\chi = \frac{A\rho_0}{\alpha\mu\phi}$ and $\kappa = \frac{A}{\mu\phi}$.

In [9], the analysis of the conservation laws admitted by Eqs. (3)–(4) was shown. It should be pointed out that as these conservative laws arise from low order multipliers, they can have a physical sense that helps to understand the physical meaning of the results of the model.

Moreover, with the aim of going deeper in the research and results of this field of unsteady flow in nanoporous rock, we have generalized the equation of weakly compressible fluid (3) and the equation of isothermic gas flow (4). So considering $f = f(p_x)$ an arbitrary function with $f \neq 0$ and $f' \neq 0$ the following equations have been studied

$$p_t = (f(p_x))_x \tag{5}$$

$$p_t = (pf(p_x))_x \tag{6}$$

Hence, (5) is the generalized equation of the case of weakly compressible fluid in nanoporous rock and (6) is the generalized equation of the case for isothermic gas flow in nanoporous rock.

On the other hand, due to its important applications in the context of differential equations [11], the Lie method is one of the best options to get results with applicability in physics of Eqs. (5)–(6).

Many renowned researchers applied this method to partial differential equations to understand and study in depth several physical phenomena. For example, de la Rosa et al. [7] considered an equation describing microwave heating and obtained its set of infinitesimal generators. Moreover, in [5, 6] solutions from a  Gardner and generalized variable-coefficient Gardner equation were obtained. Bruzón et al. [3] applied this method to the generalized KdV-Burgers-Kuramoto equation which models a dissipative, stroboscopic and unstable system in physics. Finally, Tracinà et al. [19] obtained symmetries of a system of dispersive evolution equations and so on.

In addition, in these cases of flow in nanoporous rock the Lie method is specially useful because there is an arbitrary function in both models (5)–(6) and while we are searching for symmetries it will provide a set of special forms for the unknown function $f$ where it is possible to choose.

The aim of this work is to apply the Lie classical method to the two models proposed in this introduction. In Sect. 2 the basis of the method is detailed and, we have emphasized the theory for both kind of symmetries, in Sect. 2.1 for Lie point symmetries and in Sect. 2.2 for Lie contact symmetries. Therefore, in Sect. 3 the Lie method have been applied to Eqs. (5)–(6) of generalized weakly compressible fluid in nanoporous rock , in which the case $f$ linear is omitted, and isothermic gas flow in nanoporous rock respectively. Then, in Sect. 3.1 results of Eq. (5) are presented and in Sect. 3.2 the theorems obtained for (6) are shown. Finally, conclusions are found in Sect. 4.

## 2 The Lie Method

In this section we have focused our attention on the Lie method. At the end of the nineteenth century, Lie introduced the properties of the parametric groups in the study of differential equations and that advance lead to the study of Lie groups.

The classical Lie method determine the symmetries of an ordinary differential equation or partial differential equation obtaining the uniparametric group of transformations that leave invariant the equation and transform the set of solutions in solutions. Therefore, symmetries allow us, for example, to reduce the number of variables of the equation or reduce the order of it.

We owe the first contributions in literature about the Lie method to Ovsiannikov [17] and in a more modern way to Olver [16]. Both books can be consulted for more details about the method.

## 2.1 Lie Point Symmetries

For a given partial differential equation (PDE) of order $n$

$$F(x, t, u, u_x, u_t \dots) = 0, \tag{7}$$

it is considered a one-parameter Lie group of infinitesimal point transformation in $(x, t, u)$ given by

$$\begin{aligned} x^* &= x + \epsilon\xi(x, t, u) + O(\epsilon^2),\\ t^* &= t + \epsilon\tau(x, t, u) + O(\epsilon^2),\\ u^* &= u + \epsilon\eta(x, t, u) + O(\epsilon^2), \end{aligned} \tag{8}$$

where $\epsilon$ is the group parameter and the associated vector field takes the following form

$$V = \xi(x, t, u)\partial_x + \tau(x, t, u)\partial_t + \eta(x, t, u)\partial_u. \tag{9}$$

Considering the condition that the transformation (8) with infinitesimal generator (9) leaves invariant the set of solutions of the equation (7), an overdetermined linear system of equations for the infinitesimals $\xi(x, t, u)$, $\tau(x, t, u)$ and $\eta(x, t, u)$ is obtained.

This condition is known as the criterion of invariance, it means that the set of solutions $u = u(x, t)$ of Eq. (7) is invariant under the transformations provided that

$$\mathrm{pr}^{(n)}V(F) = 0 \quad \text{when} \quad F = 0, \tag{10}$$

where $\mathrm{pr}^{(n)}V$ is the n-order prolongation of the vector field (9). This yields to the mentioned overdetermined linear system of equations for the infinitesimals whose solutions depend on the constants and functions involved in Eq. (7) if any.

## 2.2 Lie Contact Symmetries

Let us consider the given partial differential equation (7) of order $n$ and the one-parameter Lie group of contact transformation

$$\begin{aligned} \tilde{x} &= x + \epsilon\xi(x, t, u, u_x, u_t) + O(\epsilon^2),\\ \tilde{t} &= t + \epsilon\tau(x, t, u, u_x, u_t) + O(\epsilon^2),\\ \tilde{u} &= u + \epsilon\eta(x, t, u, u_x, u_t) + O(\epsilon^2), \end{aligned} \tag{11}$$

The contact transformation is characterized by the operator

$$X = \xi(x, t, u, u_x, u_t)\partial_x + \tau(x, t, u, u_x, u_t)\partial_t + \eta(x, t, u, u_x, u_t)\partial_u, \tag{12}$$

provided that the contact conditions

$$\begin{aligned}
\frac{\partial \eta}{\partial u_x} - \frac{\partial \xi}{\partial u_x}u_x - \frac{\partial \tau}{\partial u_x}u_t = 0,\\
\frac{\partial \eta}{\partial u_t} - \frac{\partial \xi}{\partial u_t}u_x - \frac{\partial \tau}{\partial u_t}u_t = 0,
\end{aligned}$$

are preserved.

To find the contact symmetries, a procedure similar to that shown for the point symmetries is followed, the criterion of invariance is applied. Invariance of the solution space of a given PDE under a contact symmetry is equivalent to invariance of the PDE under the corresponding infinitesimal contact symmetry generator (12)

$$\mathrm{pr}^{(n)}X(F)_{|\epsilon} = 0 \tag{13}$$

where $\mathrm{pr}^{(n)}X$ denotes the prolongation of the generator (12) in the solution space $\epsilon$ of Eq. (7). The prolongation formula has the form

$$\mathrm{pr}^{(n)}X = X + \eta^{(x)}\partial_{u_x} + \eta^{(t)}\partial_{u_t} + \eta^{(x,x)}\partial_{u_{xx}} + \eta^{(x,t)}\partial_{u_{xt}} + \eta^{(t,t)}\partial_{u_{tt}} + \dots$$

where

$$\begin{aligned}
\eta^{(x)} &= D_x\eta - u_x D_x\xi - u_t D_x\tau,\\
\eta^{(t)} &= D_t\eta - u_x D_t\xi - u_t D_t\tau,\\
\eta^{(x,x)} &= D_x\eta^{(x)} - u_{xx} D_x\xi - u_{tx} D_x\tau,\\
\eta^{(t,t)} &= D_t\eta^{(t)} - u_{tx} D_t\xi - u_{tt} D_t\tau,\\
&\vdots
\end{aligned}$$

So applying the contact transformation (11) and the criterion of invariance (13) yield to an overdetermined linear system of equations for the infinitesimals $(\xi,\ \tau,\ \eta)$.

It should be noted that the contact transformation (11) is equivalent to the point transformation (8) by introducing the first derivatives of the unknown function as new auxiliary unknowns. Therefore, the contact symmetries are extensions of the point symmetries under certain conditions. However, in the following, we will consider only proper contact symmetries, i.e. symmetries which are not to the point symmetries, the ones which depend essentially on derivatives.

# 3 Lie Group Classification

In this section, we have applied the Lie method to study the Lie point symmetries and contact symmetries admitted by two special cases of unsteady flow in nanoporous rock.

## 3.1 Case 1: Flow of Weakly Compressible Fluid in a Nanoporous Rock Equation

Once the Lie point transformation (8) is applied to the flow of weakly compressible fluid in a nanoporous rock equation (5) and considering that the set of solutions is invariant, the following overdetermined system for the infinitesimal is obtained

$$\begin{aligned}
-\tau_p p_x - \tau_x &= 0\\
-\tau_{pp} p_x^2 - 2\tau_{xp} p_x - \tau_{xx} &= 0\\
-\xi_{pp} f_{p_x} p_x^3 - 2\xi_{xp} f_{p_x} u_x^2 + \eta_{pp} f_{p_x} p_x^2 - \xi_{xx} f_{p_x} p_x + 2\eta_{xp} f_{p_x} p_x&\\
+\eta_{xx} f_{p_x} + \xi_t p_x - \eta_t &= 0\\
-\xi_p f_{p_x p_x} p_x^2 - \xi_x f_{p_x p_x} p_x + \eta_p f_{p_x p_x} p_x - 2\xi_p f_{p_x} p_x + \eta_x f_{p_x p_x} - 2\xi_x f_{p_x} + \tau_t f_{p_x} &= 0
\end{aligned}$$

Moreover, the solutions of the system are classified depending on the function $f = f(p_x)$ and its constants. Due to Eq. (5) reduces to the heat equation if $f$ is linear, this case is avoided. Hence, the classification of the Lie point symmetries is the following.

**Theorem 1** *The point symmetries admitted by Eq. (5) are generated by:*

- *Case 1. For $f = f(p_x)$ an arbitrary function, the infinitesimal generators are*

$$\begin{aligned}
v_1 &= \partial_x,\\
v_2 &= \partial_t,\\
v_3 &= \partial_p,\\
v_4 &= x\partial_x + 2t\partial_t + p\partial_p.
\end{aligned}$$

- *Case 2. For $f$ satisfying the differential equation*

$$\frac{f_{p_x}}{f} = \frac{a + 2bp_x}{c + dp_x - bp_x^2}$$

*being $a, b, c$ and $d$ arbitrary constants, we obtain besides $v_1$, $v_2$, $v_3$ and $v_4$ the following infinitesimal generator*

$$v_5 = bp\partial_x - at\partial_t + (cx + dp)\,\partial_p.$$

It should be remarked that Eq. (5) also describes the motion of a non-Newtonian, weakly compressible fluid in a porous medium and that its group classification was presented in [2].

In addition, applying the Lie contact transformation (11) to the flow of weakly compressible fluid equation (5) an overdetermined system is obtained, however, all the contact symmetries obtained reduce to point symmetries. Hence, the following result is obtained.

**Theorem 2** *The flow of weakly compressible fluid in a nanoporous rock equation (5) does not admit contact symmetries.*

## 3.2 Case 2: Isothermic Gas Flow in a Nanoporous Rock Equation

Applying the Lie group transformation (8) to the isothermic gas flow in a nanoporous rock equation (6) and by criterion of invariance (10) the following overdetermined system of equations for the infinitesimal is obtained

$$\begin{aligned}
\tau_p p_x - \tau_x &= 0\\
-f_{p_x}^2 p^2 \tau_{pp} p_x^2 - 2f_{p_x}^2 p^2 \tau_{xp} p_x - f_{p_x}^2 p\tau_p p_x^2 - f_{p_x}^2 p\tau_x p_x&\\
-f_{p_x} p\tau_x f - 2f_{p_x} p\xi_p p_x + p\eta_x f_{p_x p_x} + \tau_t f_{p_x} p - p^2\tau_{xx}&\\
+f_{p_x}\eta - 2f_{p_x} p\xi_x + p\tau_p f_{p_x p_x} f p_x^2 + p\tau_x f_{p_x p_x} f p_x&\\
-p\xi_p f_{p_x p_x} p_x^2 - p\xi_x f_{p_x p_x} p_x + p\eta_p f_{p_x p_x} p_x &= 0\\
2f_{p_x} p\xi_p f p_x^2 + f_{p_x} p\xi_x f p_x - f_{p_x}\eta f p_x + \xi_t p_x f_{p_x} p&\\
-2f_{p_x}^2 p^2 \xi_{xp} p_x^2 + f_{p_x}^2 p^2 \eta_{pp} p_x^2 - f_{p_x}^2 p\xi_p p_x^2 - f_{p_x}^2 p^2 \xi_{xx} p_x&\\
+2f_{p_x}^2 p^2 \eta_{xp} p_x - f_{p_x}^2 p\xi_x p_x^2 + f_{p_x}^2 p\eta_p p_x^2 + f_{p_x}^2 p\eta_x p_x&\\
+f_{p_x} p\eta_x f - f_{p_x}^2 p^2 \xi_{pp} p_x^3 + f_{p_x}^2 p^2 \eta_{xx} - \eta_t f_{p_x}^2 p&\\
+p\xi_p f_{p_x p_x} f p_x^3 + p\xi_x f_{p_x p_x} f p_x^2 - p\eta_p f_{p_x p_x} f p_x^2 - p\eta_x f_{p_x p_x} f p_x &= 0
\end{aligned}$$

The solutions of the system depend on the function $f = f(p_x)$ and six different cases have been obtained. Hence, the classification of the Lie point symmetries is the following.

**Theorem 3** *The point symmetries admitted by Eq. (6) are generated by:*

- *Case 1. For $f = f(p_x)$ an arbitrary function, the infinitesimal generators are*

$$\begin{aligned} v_1' &= \partial_x, \\ v_2' &= \partial_t, \\ v_3' &= x\partial_x + t\partial_t + p\partial_p. \end{aligned}$$

- *Case 2. For $f = \dfrac{(k_1k_2 - p_x)\, k_3 \left(\frac{p_x - k_1k_2}{p_x}\right)^{\frac{-1}{k_1}}}{(k_1 - 1)\, k_2 p_x} + k_4$, being $k_1,\ k_2,\ k_3,\ k_4$ arbitrary constants with $k_1 \neq 0, 1$ and $k_2 \neq 0$, the infinitesimal generators are $v_1',\ v_2',\ v_3'$ and*

$$v_4' = \left(-k_1k_4t + k_4t - k_1x + x + \frac{1}{k_2}p\right)\partial_x + p\partial_p.$$

- *Case 3. For $f = -\dfrac{1}{k_1}\exp\left(\dfrac{k_3 + p_x}{p_x}\right) + k_2$, being $k_1,\ k_2,\ k_3$ arbitrary constants with $k_1 \neq 0$ and $k_3 \neq 0$, the infinitesimal generators are $v_1',\ v_2',\ v_3'$ and*

$$\begin{aligned} v_5' &= \left(k_2t - \frac{1}{k_3}p\right)\partial_x + t\partial_t, \\ v_6' &= -\frac{1}{k_3}\ln(p)\partial_x + \partial_p. \end{aligned}$$

- *Case 4. For $f = -\dfrac{4k_1}{k_2 + p_x} + k_3$, being $k_1,\ k_2,\ k_3$ arbitrary constants with $k_2 \neq 0$, the infinitesimal generators are $v_1',\ v_2',\ v_3'$ and*

$$v_7' = \frac{1}{2k_2}\left(-k_2k_3t + 4k_1t + k_2x + p\right)\partial_x + t\partial_t.$$

- *Case 5. For $f = k_1p_x + k_2$, being $k_1,\ k_2$ arbitrary constants, the infinitesimal generators are $v_1',\ v_2',\ v_3'$ and*

$$v_8' = \left(\frac{1}{2}x - \frac{1}{2}k_2t\right)\partial_x + t\partial_t.$$

- *Case 6. For* $f = k_1 + \dfrac{k_2}{p_x}$, *being* $k_1,\ k_2$ *arbitrary constants with* $k_2 \neq 0$, *the infinitesimal generators are* $v'_1,\ v'_2,\ v'_3$ *and*

$$v'_9 = \ln(p)\partial_x,$$

$$v'_{10} = \left(\frac{1}{2}t - \frac{1}{2}\frac{p}{k_2}\right)\partial_x.$$

Moreover, the Lie contact transformation (11) has been applied to the isothermic gas flow in a nanoporous rock equation (6), applying the criterion of invariance (13) an overdetermined system for the infinitesimal was sought. But the contact symmetries obtained reduce to point symmetries, so the following result is obtained.

**Theorem 4** *The isothermic gas flow in a nanoporous rock equation (6) does not admit contact symmetries.*

## 4 Conclusions

In this paper we have considered two special cases derived from the basic equation of the unsteady flow in nanoporous rock, the model of a weakly compressible fluid and the model of flow of isothermal gas. Besides, the relationships from which the equation of a weakly compressible fluid in nanoporous rock and the equation of flow of isothermal gas in nanoporous rock arise are shown.

For each model, a generalization of the equation is proposed. Hence, in order to go deeper in the research and results of these models of unsteady flow in nanoporous rock, new generalized equations which involve arbitrary functions are presented.

Applying the Lie method to the generalize model of a weakly compressible fluid, Lie point symmetries and contact symmetries have been studied. With regard to Lie point symmetries, we have obtained four cases depending on the arbitrary function along with their infinitesimal generators. Furthermore, we have concluded that the flow of weakly compressible fluid in a nanoporous rock equation does not admit Lie contact symmetries other than prolongations of point symmetries.

On the other hand, the Lie method have been also applied to the generalize model of flow of isothermal gas in nanoporous rock. In relation to Lie point symmetries, a complete classification with six different cases depending on the arbitrary function of the model have been obtained. Finally, it has been asserted that the generalize model of flow of isothermal gas in nanoporous rock does not admit contact symmetries other than prolongations of point symmetries.

**Acknowledgements** The authors gratefully thank the Universidad Politécnica de Cartagena for their support. The authors are grateful to the Junta de Andalucía for the financial support of group FQM-201 and they express their sincere gratitude to the Plan Propio de Investigación de la Universidad de Cádiz. Authors warmly thank the referee for suggestions to improve the paper.

## References

1. Abriola, L.M., Pinder, G.F.: A multiphase approach to the modelling of porous media contamination by organic compounds, 1-equation development. Water Resour. Res. **21**, 11–18 (1985)
2. Akhatov, I.SH., Gazizov, R.K., Ibrahimov, N.KH.: Group classification of the equations of nonlinear filtration. Sov. Math. Dokl. **35**, 384–386 (1987)
3. Bruzón, M.S., Recio, E., Garrido, T.M., Márquez, A.P.: Conservation laws, classical symmetries and exact solutions of the generalized KdV-Burgers-Kuramoto equation. Open Phys. **15**, 433–439 (2017)
4. Cui, J., Sang, Q., Li, Y., Yin, C., Li, Y., Dong, M.: Liquid permeability of organic nanopores in shale: calculation and analysis. Fuel **202**, 426–434 (2017)
5. de la Rosa, R., Bruzón, M.S.: On the classical and nonclassical symmetries of a generalized Gardner equation. Appl. Math. Nonlinear Sci. **1**, 263–272 (2016)
6. de la Rosa, R., Bruzón, M.S.: Travelling wave solutions of a generalized variable-coefficient Gardner equation. In: Trends in Differential Equations and Applications. SEMA SIMAI Springer Series, pp. 405–417. Springer, Berlin (2016)
7. de la Rosa, R., Gandarias, M.L., Bruzón, M.S.: A study for the microwave heating of some chemical reactions through Lie symmetries and conservation laws. J. Math. Chem. **53**, 949–957 (2014)
8. Duval, F., Fichot, F., Quintard, M.: A local thermal non-equilibrium model for two-phase flows with phase-change in porous media. Int. J. Heat Mass Transf. **47**, 613–639 (2004)
9. Garrido, T.M., Bruzón, M.S.: Conservation laws for the Barenblatt-Gilman equation and for two special cases of unsteady flow in nanoporous rock. In: Libro de comunicaciones definitivas presentadas en CEDYA+CMA 2017, pp. 318–321 (2017). ISBN: 978-84-944402-1-2
10. Garrido, T.M., Kasatkin, A.A., Bruzón, M.S., Gazizov, R.K.: Lie symmetries and equivalence transformations for the Barenblatt–Gilman model. J. Comput. Appl. Math. **318**, 253–258 (2017)
11. Ibragimov, N.H.: CRC Handbook of Lie Group Analysis of Differential Equations, Vols. 1–3. CRC Press, Boca Raton (1994–1996)
12. Kaluarachchi, J.J., Parker, J.C.: Modeling multicomponent organic chemical transport in three-fluid-phase porous media. J. Contam. Hydrol. **5**, 349–374 (1990)
13. Khaled, A.R.A., Vafai, K.: The role of porous media in modeling flow and heat transfer in biological tissues. Int. J. Heat Mass Transf. **46**, 4989–5003 (2003)
14. Ma, J., Sanchez, J.P., Wu, K., Couples, G.D., Jiang, Z.: A pore network model for simulating non-ideal gas flow in micro- and nano-porous materials. Fuel **116**, 498–508 (2014)
15. Monteiro, P.J.M., Rycroft, C.H., Barenblatt, G.I.: A mathematical model of fluid and gas flow in nanoporous media. Proc. Natl. Acad. Sci. U. S. A. **109**, 20309–20313 (2012)
16. Olver, P.: Applications of Lie Groups to Differential Equations. Springer, New York (1993)
17. Ovsyannikov, L.V.: Group Analysis of Differential Equations. Academic, New York (1982)
18. Tomutsa, L., Silin, D.B., Radmilovic, V.: Analysis of chalk petrophysical properties by means of submicron-scale pore imaging and modeling. SPE Reserv. Eval. Eng. **10**, 285–293 (2007)
19. Tracinà, R., Bruzón, M.S., Gandarias, M.L., Torrisi, M.: Nonlinear self-adjointness, conservation laws, exact solutions of a system of dispersive evolution equations. Commun. Nonlinear Sci. Numer. Simul. **19**, 3036–3043 (2014)
20. Viskanta, R.: Thermal Transport in Highly Porous Cellular Materials. Handbook of Porous Media, pp. 535–558. CRC Press, Boca Ratón (2015)
21. Wessapan, T., Rattanadecho, P.: Flow and heat transfer in biological tissue due to electromagnetic near-field exposure effects. Int. J. Heat Mass Transf. **97**, 174–184 (2016)
22. Young, W.B.: Capillary impregnation into cylinder banks. J. Colloid Interface Sci. **273**, 576–580 (2004)

# Asymptotic Behaviour of Finite Length Solutions in a Thermosyphon Viscoelastic Model

**Ángela Jiménez-Casas**

**Abstract** A thermosyphon, in the engineering literature, is a device composed of a closed loop containing a fluid whose motion is driven by several actions, such as gravity and natural convection. In this work we prove some results about the asymptotic behaviour for solutions of a closed loop thermosyphon model with a viscoelastic fluid in the interior (Jiménez-Casas et al., Discrete Conti Dynam Syst (9th AIMS Conference Sool) 2013:375–384, 2013; Chaotic Model Simul 2:281–288, 2013). In this model a viscoelastic fluid described by the Maxwell constitutive equation is considered, this kind of fluids present elastic-like behavior and memory effects. Their dynamics are governed by a coupled differential nonlinear systems. In several previous works we have shown chaos in the fluid, even with this kind of viscoelastic fluid (Jiménez-Casas and Castro, Electron J Differ Equ (Conference 22), 53–61, 2015; Yasapan et al., Abstr Appl Anal 2013, Article ID: 748683, 2013; Discrete Conti Dynam Syst Ser B 20:3267–3299, 2015 among others). In this model, we consider a prescribed heat flux like Rodríguez-Bernal and Van Vleck (SIAM J Appl Math 58:1072–1093, 1998), Jiménez-Casas and Ovejero (Appl Math Comput 124:289–318, 2001) among others (all of them with Newtonian fluids). This work is, in some sense, a generalization of some previous results on standard (Newtonian) fluids obtained by Rodríguez-Bernal and Van Vleck (SIAM J Appl Math 58:1072–1093, 1998), when we consider a viscoelastic fluid.

**Keywords** Thermosyphon · Viscoelastic fluid · Asymptotic behaviour

A. Jiménez-Casas (✉)
Dpto. de Matemática Aplicada, Grupo de Dinámica No lineal, Universidad Pontificia Comillas de Madrid, Madrid, Spain
e-mail: ajimenez@comillas.edu

© Springer Nature Switzerland AG 2019

J. L. García Guirao et al. (eds.), *Recent Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 18,
https://doi.org/10.1007/978-3-030-00341-8_6

# 1 Introduction

In engineering literature, a thermosyphon is a device composed of a closed loop *pipe* containing a fluid whose motion is driven by the effect of several actions such as gravity and natural convection. The flow inside the loop is driven by an energetic balance between thermal energy and mechanical energy.

Here, we consider a thermosyphon model where the confined fluid is viscoelastic. This has some *a-priori* interesting peculiarities that could affect the dynamics with respect to the case of a Newtonian fluid. On one hand, the dynamics have memory, so its behavior depends on the its whole past history and, on the other hand, at small perturbations, the fluid behaves like an elastic solid and a characteristic resonance frequency could, eventually, be relevant (consider for instance the behavior of jelly or toothpaste).

The simplest approach to viscoelasticity comes from the so-called Maxwell model [11]. In this model, both Newton's law of viscosity and Hooke's law of elasticity are generalized and complemented through an evolution equation for the stress tensor, $\sigma$.

Viscoelastic behavior is common in polymeric and biological suspensions and, consequently, our results may provide useful information on the dynamics of this sort of systems inside a thermosyphon.

In a thermosyphon, the equations of motion can be greatly simplified because of the quasi-one-dimensional geometry of the loop. Thus, we assume that the section of the loop is constant and small compared with the dimensions of the physical device, so that the arc length co-ordinate along the loop $(x)$ gives the position in the circuit.

The velocity of the fluid is assumed to be independent of the position in the circuit, i.e. it is assumed to be a scalar quantity depending only on time.

This approximations come from the fact that the fluid is assumed to be incompressible. On the contrary, temperature is assumed to depend both on time and position along the loop.

The derivation of the thermosyphon equations of motion is similar to that in Ref. [9] and these equations are obtained in [7]. Finally, after adimensionalizing the variables (to reduce the number of free parameters) we get the following ODE/PDE system (see [7] and [1]):

$$\begin{cases} \varepsilon \dfrac{d^2 v}{dt^2} + \dfrac{dv}{dt} + G(v)v = \oint Tf, \\ \dfrac{\partial T}{\partial t} + v\dfrac{\partial T}{\partial x} \qquad = h(x, v, T) + \mu \frac{\partial^2 T}{\partial x^2}, \end{cases} \tag{1}$$

with $v(0) = v_0$, $\frac{dv}{dt}(0) = w_0$ and $T(0, x) = T_0(x)$.

Here, $v(t)$ is the velocity, $T(t, x)$ is the distribution of the temperature of the viscoelastic fluid into the loop, $G(v)$, is the friction law at the inner wall of the loop, and the function $f$ is the geometry of the loop and the distribution of gravitational forces. In this case $h(x, v, T) = h(x)$ as in [6, 7, 15, 16] is the general heat flux and $\mu$ is the temperature diffusion coefficient. Finally, $\varepsilon$ in Eq. (1) is the viscoelastic parameter, which is the dimensionless version of the viscoelastic time. Roughly speaking, it gives the time scale in which the transition from elastic to fluid-like occurs in the fluid.

We assume that $G(v)$ is positive and bounded away from zero. This function has been usually taken to be $G(v) = G_0$, a positive constant for the linear friction case [9], or $G(v) = |v|$ for the quadratic law [4, 10], or even a rather general function given by $G(v) = g(Re)|v|$,where $Re$ is a Reynolds-like number that is assumed to be large [14, 15] and proportional to $|v|$. The functions $G$, $f$, and $h$ incorporate relevant physical constants of the model, such as the cross sectional area, $D$, the length of the loop, $L$, the Prandtl, Rayleigh, or Reynolds numbers, etc. (see [15]). Usually $G$, $h$ are given continuous functions, such that $G(v) \geq G_0 > 0$, and $h(v) \geq h_0 > 0$, for $G_0$ and $h_0$ positive constants.

Finally, for physical consistency, it is important to note that all functions considered must be 1-periodic with respect to the spatial variable, and $\oint = \int_0^1 dx$ denotes integration along the closed path of the circuit. Note that $\oint f = 0$.

The main contribution of this paper (Sect. 3) is to prove that, under suitable conditions, any solution either converges to the rest state or the oscillations of velocity around $v = 0$ must be large enough. This result generalizes the one proposed in Rodríguez-Bernal and Van Vleck [14] for a thermosyphon model with a one-component **viscoelastic** fluid.

Besides, we note in some sense that Proposition 2 in the linear friction case, generalizes the results about the asymptotic behaviour for the dynamics in [7, 17] for functions $f$ and $h$ without the condition $K \cap J = \emptyset$. In this Proposition 2 we prove that, the finite length solutions ($\int_0^\infty |v(s)|ds < \infty$) converge to the rest equilibrium point, taking functions $f$ and $h$ with $K \cap J \neq \emptyset$, this is enough to consider the orthogonality condition (11), this is, $Re\left( \sum_{k\in(K\cap J)_+} \frac{b_k c_{-k}}{k^2} \right) = 0$ with $K \cap J$ a finite set.

## 2 Previous Results About Well Posedness and Global Attractor

First, we note that in this section we consider the case where all periodic functions in Eq. (1) have zero average, i.e. we work in $\mathcal{Y} = \mathbb{R}^2 \times \dot{H}^1_{per}(0, 1)$, where

$$\dot{L}^2_{per}(0, 1) = \{u \in L^2_{loc}(\mathbb{R}),\ u(x + 1) = u(x) a.e., \oint u = 0\},$$

$$\dot{H}^m_{per}(0, 1) = H^m_{loc}(\mathbb{R}) \cap \dot{L}^2_{per}(0, 1).$$

In effect, we observe that, for $\mu > 0$, if we integrate the equation for the temperature along the loop, we have that $\frac{d}{dt}(\oint T) = \oint h$ and then $\oint T(t) = \oint T_0 + t \oint h$. Therefore, the temperature is unbounded, as $t \to \infty$, unless $\oint h = 0$. However, taking $\theta = T - \oint T$ and $h^* = h - \oint h$ reduces to the case $\oint \theta = \oint \theta_0 = \oint h^* = 0$, since $\theta$ would satisfy:

$$\frac{\partial \theta}{\partial t} + v\frac{\partial \theta}{\partial x} = h(x) + \mu \frac{\partial^2 \theta}{\partial x^2}, \quad \theta(0) = \theta_0 = T_0 - \oint T_0 \tag{2}$$

Therefore, if we consider now $\theta = T - \oint T$ then from the second equation of system Eq. (1), we obtain that $\theta$ satisfies Eq. (2).

Finally, since $\oint f = 0$, we have that $\oint Tf = \oint \theta f$, and the equation for $v$ reads

$$\varepsilon \frac{d^2 v}{dt^2} + \frac{dv}{dt} + G(v)v = \oint \theta . f, \quad v(0) = v_0, \frac{dv}{dt}(0) = w_0. \tag{3}$$

Thus, from Eqs. (2) and (3) we have $(v, \theta)$ satisfies system Eq. (1) with $h^*, \theta_0$ replacing $h, T_0$ respectively and now $\oint \theta = \oint h = \oint \theta_0 = 0$.

Thus, we consider the system Eq. (1) with $\oint T_0 = \oint h = 0$ and $\oint T(t) = 0$ for every $t \geq 0$.

Also, if $\mu > 0$ the operator $-\mu \frac{\partial^2}{\partial x^2}$, together with periodic boundary conditions, is an unbounded, self-adjoint operator with compact resolvent in $L^2_{per}(0, 1)$ which is positive when restricted to the space of zero-average functions $\dot{L}^2_{per}(0, 1)$. Hence, the second equation in Eq. (1) is of parabolic type for $\mu > 0$.

Hereafter we denote by $w = \frac{dv}{dt}$ and we write the system (1) as the following evolution system for the acceleration, velocity and temperature:

$$\frac{d}{dt}\begin{pmatrix} w \\ v \\ T \end{pmatrix} + \begin{pmatrix} \frac{1}{\varepsilon} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -\mu \frac{\partial^2}{\partial x^2} \end{pmatrix}\begin{pmatrix} w \\ v \\ T \end{pmatrix} = \begin{pmatrix} F_1 \\ F_2 \\ F_3 \end{pmatrix} \tag{4}$$

with $F_1(w, v, T) = -\frac{1}{\varepsilon}G(v)v + \frac{1}{\varepsilon}\oint Tf$, $F_2(w, v, T) = w$ and $F_3(w, v, T) = -v\frac{\partial T}{\partial x} + h(v)$ and initial data $(w, v, T)^{\perp}(0) = (w_0, v_{0,0} T)^{\perp}$.

The operator $B = \begin{pmatrix} \frac{1}{\varepsilon} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -\mu \frac{\partial^2}{\partial x^2} \end{pmatrix}$ is a sectorial operator in $\mathcal{Y} = \mathbb{R}^2 \times \dot{H}^1_{per}(0, 1)$ with domain $D(B) = \mathbb{R}^2 \times \dot{H}^3_{per}(0, 1)$ and has compact resolvent, where

$$\dot{L}^2_{per}(0, 1) = \{u \in L^2_{loc}(\mathbb{R}), u(x + 1) = u(x) a.e., \oint u = 0\},$$

$$\dot{H}^m_{per}(0, 1) = H^m_{loc}(\mathbb{R}) \cap \dot{L}^2_{per}(0, 1).$$

Thus, we can apply the result about sectorial operator of [3] to prove the existence of solutions of system (1).

Moreover, if we consider some additionally hypothesis to add for the friction $G$, i.e. there exits a constant $h_0 \geq 0$ such that:

$$\limsup_{t\to\infty} \frac{|G'(t)|}{G(t)} = 0 \text{ and } \limsup_{t\to\infty} \frac{|tG'(t)|}{G(t)} \leq h_0. \tag{5}$$

Then, using the technique Lemma 3.1 in [7], we have the next result.

It is important to note that the conditions (5) are satisfied for all the friction functions $G$ considered in the previous works, i.e., the thermosyphon models where $G$ is constant or linear or quadratic law. Moreover, the conditions (5) are true for $G(s) \approx A|s|^n$, as $s \to \infty$.

**Proposition 1** *We suppose that $H(r) = rG(r)$ is locally Lipschitz, $f, h \in \dot{L}^2_{per}$ $h(v) \geq h_0 > 0$. Then, given $(w_0, v_0, T_0) \in \mathcal{Y} = \mathbb{R}^2 \times \dot{H}^1_{per}(0, 1)$, there exists a unique solution of (1) satisfying*

$$(w, v, T) \in C([0, \infty], \mathcal{Y}) \cap C(0, \infty, \mathbb{R}^2 \times \dot{H}^3_{per}(0, 1)),$$

$$(\dot{w}, w, \frac{\partial T}{\partial t}) \in C(0, \infty, \mathbb{R}^2 \times \dot{H}^{3-\delta}_{per}(0, 1)),$$

*where $w = \dot{v} = \frac{dv}{dt}$ and $\dot{w} = \frac{d^2v}{dt^2}$ for every $\delta > 0$. In particular, (1) defines a nonlinear semigroup, $S(t)$ in $\mathcal{Y}$, with $S(t)(w_0, v_0, T_0) = (w(t), v(t), T(t))$.*

*Moreover, from (5) (see [7]) Eq. (1) has a global compact and connected attractor, $\mathcal{A}$ , in $\mathcal{Y}$. Also if $h \in \dot{H}^m_{per}(0, 1)$ with $m \geq 1$, the global attractor $\mathcal{A} \subset \mathbb{R}^2 \times \dot{H}^{m+2}_{per}$ and is compact in this space.*

*Proof* This result has been proved in Theorem 2.1, Proposition 3.1 and Corollary 4.1 from Jiménez-Casas et al. [7]. □

## 3 Asymptotic Behaviour of Finite Length Solutions

In previous works, like Jiménez-Casas et al.[7], the asymptotic behaviour of the system Eq. (1) for large enough time is studied.

In this sense the existence of an inertial manifold associated to the functions $f$ (loop-geometry) and $h$ (prescribed heat flux) has been proved. In this case, we consider an **inertial manifold** $\mathcal{M}$ for the flow $S(t)(w_0, v_0, T_0) = (w(t), v(t), T(t))$ in the space $\mathcal{Y} = \mathbb{R}^2 \times \dot{H}^1_{per}(0, 1) \times \dot{L}^2_{per}(0, 1)$, as a smooth manifold $\mathcal{M} \subset \mathcal{Y}$ such that: (1) $\mathcal{M}$ is positively invariant, that is $S(t)\mathcal{M} \subset \mathcal{M}$ for every $t \geq 0$; (2) $\mathcal{M}$ contains the attractor, that is, $\mathcal{A} \subset \mathcal{M}$; (3) $\mathcal{M}$ is exponentially attracting in the sense that there exists a constant $\delta > 0$ such that for every bounded set $B \subset \mathcal{Y}$ there exists $C = C(B) \geq 0$ such that $dist(S(t)B, \mathcal{M}) \leq Ce^{-\delta t}$, for every $t \geq 0$.

The abstract operators theory (Henry [3], Foias et al. [2] and Rodríguez-Bernal [12, 13]) has been used for this purpose.

In Proposition 2 we consider the linear friction case [9] where $G(v) = G_0$, a positive constant and we prove some results under suitable conditions for the geometry of the loop $f$ and the heat flux $h$. These results rise an important consequence: in this condition, for large time the velocity reaches the equilibrium—null velocity—or takes a value to make its integral diverge.

Besides, Proposition 2 generalizes the result of thermosyphon model for Newtonian fluids of Rodríguez-Bernal and Van Vleck [14] in the case of a prescribed heat flux, i.e. $h = h(x)$, for a viscoelastic fluid.

## 3.1 Previous Results and Notations

In this section, we assume also that $G^*(r) = rG(r)$ is locally Lipschitz satisfying (5) (see [7]), and $f, h \in \dot{L}^2_{per}$ are given by following Fourier expansions

$$h(x) = \sum_{k\in\mathbb{Z}^*} b_k e^{2\pi kix}; \qquad f(x) = \sum_{k\in\mathbb{Z}^*} c_k e^{2\pi kix}; \tag{6}$$

where $\mathbb{Z}^* = \mathbb{Z} - \{0\}$, while $T_0 \in \dot{H}^2_{per}$ is given by $T_0(x) = \sum_{k\in\mathbb{Z}^*} a_{k0} e^{2\pi kix}$. Finally, assuming that $T(t,x) \in \dot{H}^2_{per}$ is given by

$$T(t,x) = \sum_{k\in\mathbb{Z}^*} a_k(t) e^{2\pi kix} \tag{7}$$

where $\mathbb{Z}^* = \mathbb{Z} - \{0\}$, we note that $\bar{a_k} = -a_k$ since all functions considered are real and also $a_0 = 0$ since they have zero average.

Now, we observe the dynamics of each Fourier mode and from Eq. (1), we get the following system for the new unknowns, $v$ and the coefficients $a_k(t)$.

$$\left\{ \begin{array}{l} \varepsilon \frac{d^2v}{dt^2} + \frac{dv}{dt} + G(v)v = \sum_{k\in\mathbb{Z}^*} a_k(t) c_{-k} \\ \dot{a}_k(t) + \left[2\pi kiv(t) + 4\mu\pi^2k^2\right] a_k(t) = b_k \end{array} \right. \tag{8}$$

- Assuming that the prescribed heat flux $h \in \dot{H}^m_{per}$, is given by

  $$h(x) = \sum_{k\in K} b_k e^{2\pi kix},$$

  and $b_k \neq 0$ for every $k \in K \subset \mathbb{Z}$ with $0 \notin K$, since $\oint h = 0$.

  We denote by $V_m$ the closure of the subspace of $\dot{H}^m_{per}$ generated by $\{e^{2\pi kix}, k \in K\}$.

Then we have from Theorem 4.2 in Jiménez-Casas et al.[7] the set $\mathcal{M} = \mathbb{R}^2 \times V_m$ is an **inertial manifold** for the flow of $S(t)(w_0, v_0, T_0) = (w(t), v(t), T(t))$ in the space $\mathcal{Y} = \mathbb{R}^2 \times \dot{H}^m_{per}$. By this, the dynamics of the flow is given by the flow in $\mathcal{M}$ associated to the prescribed heat flux $h$.

This is

$$\begin{cases} \varepsilon\frac{d^2v}{dt^2} + \frac{dv}{dt} + G(v)v = \sum_{k\in K} a_k(t)c_{-k} \\ \dot{a}_k(t) + \left[2\pi k i v(t) + 4\mu\pi^2k^2\right]a_k(t) = b_k,\ k \in K \end{cases} \tag{9}$$

- Moreover, we assume that the function associated to the geometry of the loop $f$, is given by

$$f(x) = \sum_{k\in J} c_k e^{2\pi k i x}$$

and $c_k \neq 0$ for every $k \in J \subset \mathbb{Z}$ with $0 \notin J$, since $\oint f = 0$. We note also that on the inertial manifold

$$\oint Tf = \sum_{k\in K\cap J} a_k(t)c_{-k}.$$

Thus, the dynamics of the system depend only on the coefficients in $K \cap J$.

First, from Eq. (8) we can observe the velocity of the fluid is independent of the coefficients for temperature $a_k(t)$ for every $k \in \mathbb{Z}\sqrt{}^* - (K \cap J)$. That is, the **relevant coefficients** for the velocity are only $a_k(t)$ with $k$ belonging to the set $K \cap J$. This important result about the asymptotic behaviour has been proved in Corollary 4.2 from Jiménez-Casas et al.[7].

- We also note that $0 \notin K \cap J$ and since $K = -K$ and $J = -J$ then the set $K \cap J$ has an even number of elements, which we denote by $2n_0$. Therefore, the number of the positive elements of $K \cap J$, $(K \cap J)_+$ is $n_0$.

Moreover, the equations for $a_{-k}$ are conjugates of the equations for $a_k$ and therefore we have:

$$\sum_{k\in K\cap J} a_{k(t)}c_{-k} = 2Re(\sum_{k\in(K\cap J)_+} a_k(t)c_{-k}).$$

Thus,

$$\oint Tf = 2Re(\sum_{k\in(K\cap J)_+} a_k(t)c_{-k}). \tag{10}$$

The aim is to prove Proposition 2, which generalizes, in some sense, the result of thermosyphon model for Newtonian fluids of Rodríguez-Bernal and Van Vleck [14] in the case of a prescribed heat flux, i.e. $h = h(x)$, for a viscoelastic fluid. To do so, we examine which are these steady-state solutions, also called *equilibrium points*.

We have to note the difference between equilibrium points (constants respect to the time) null velocity, called *rest equilibrium*, and equilibrium points with non-vanishing constant velocity.

*Equilibrium Conditions*

(a) The system Eq. (8) presents the *rest equilibrium* (rest stationary solutions) $w = v = 0$, $a_k = \frac{b_k}{4\mu\pi^2k^2}$ $\forall k \in K \cap J$ under the assumption of the following orthogonality condition:

$$I_0 = Re(\sum_{k\in(K\cap J)_+} \frac{b_k c_{-k}}{k^2}) = 0. \tag{11}$$

(b) Any other equilibrium position will have a non-vanishing velocity and the equilibrium is given by:

$$\begin{cases} G(v)v = 2Re(\displaystyle\sum_{k\in(K\cap J)_+} a_k c_{-k}) \\ a_k = \frac{b_k}{4\mu\pi^2k^2+2\pi kiv} \end{cases} \tag{12}$$

## 3.2 Asymptotic Behaviour

**Lemma 1** *If we assume that a solution of Eq. (8) satisfies $\int_0^\infty |v(s)|ds < \infty$, then for every $\eta > 0$ there exists $t_0$ such that*

$$\int_{t_0}^t e^{-4\mu\pi^2k^2(t-r)}(e^{-\int_r^t 2\pi ikv} - 1)dr \leq \eta, \tag{13}$$

*with $t \geq t_0$.*

*Proof* If $\int_0^\infty |v(s)|ds < \infty$, then for all $\delta$ there exists $t_0 > 0$ such that for every $t_0 \leq r \leq t$ we have $|\int_r^t v| \leq \delta$. Then, for any $\eta > 0$ we can take $t_0$ large enough such that

$$|e^{-\int_r^t 2\pi ikv} - 1| \leq \eta \text{ for all } t_0 \leq r \leq t. \tag{14}$$

Therefore, we get

$$\int_{t_0}^t e^{-4\mu\pi^2k^2(t-r)}(e^{-\int_r^t 2\pi ikv} - 1)dr \leq$$

$\leq \frac{\eta}{4\mu\pi^2k^2}(1 - e^{-4\mu\pi^2k^2(t-t_0)}) \leq \eta$ with $t \geq t_0$ and taking into account that $\eta \to 0$ for $t \to \infty$ and $\mu > 0$, we get Eq. (13). □

**Proposition 2** *We consider the linear friction case, i.e.* $G(v) = G_0$ *with* $G_0$ *a positive constant.*

*i) If we assume that* $I_0 = Re\Big(\sum_{k\in(K\cap J)_+} \frac{b_k c_{-k}}{k^2}\Big) = 0$, *with* $K \cap J$ *finite set, and that a solution of Eq. (8) satisfies* $\int_0^\infty |v(s)|ds < \infty$. *Then, the system reaches the rest stationary solution, that is:*

$$\begin{cases} v(t) \to 0, \ \textit{and} \ w(t) \to 0, \ \textit{as} \ t \to \infty \\ a_k(t) \to \frac{b_k}{4\mu\pi^2k^2}, \ \textit{as} \ t \to \infty \end{cases}$$

*Moreover,*

$$T(t,x) \to \theta_\infty \ \textit{in} \ \dot{H}^1_{per},$$

*where* $\theta_\infty$ *is the unique solution in* $\dot{H}^2_{per}(0,1)$ *of*

$$-\mu\frac{\partial^2\theta}{\partial x^2} = h(x).$$

*ii) Conversely, if* $I_0 = Re(\sum_{k\in(K\cap J)_+} \frac{b_k c_{-k}}{k^2}) \neq 0$ *then for every solution* $\int_0^\infty |v(s)|ds = \infty$, *and* $v(t)$ *does not converge to zero.*

*iii) If* $K \cap J = \emptyset$, *then the global attractor for system Eq. (1) in* $\mathbb{R}^2 \times \dot{H}^1_{per}(0,1)$ *is reduced to a point* $\{(0,0,\theta_\infty)\}$, *where* $\theta_\infty$ *is the unique solution in* $\dot{H}^2_{per}(0,1)$ *of*

$$-\mu\frac{\partial^2\theta}{\partial x^2} = h(x).$$

*Proof* First, we study the behaviour for large time of the coefficients $a_k(t)$.

The distance between the coefficients that represents the solution of the system, $a_k(t)$ to the values of those coefficients in the equilibrium, $\frac{b_k}{4\mu\pi^2k^2}$ is computed.

For $t_0$ enough large, we know that for every $t > t_0$ we have

$$a_k(t) = a_k(t_0)e^{-\int_{t_0}^t 2\pi ikv+4\mu\pi^2k^2} + b_k\int_{t_0}^t e^{-\int_r^t 2\pi ikv+4\mu\pi^2k^2}dr \tag{15}$$

and using

$$\int_{t_0}^t e^{-\int_r^t 4\mu\pi^2k^2} = \frac{1}{4\mu\pi^2k^2}(1 - e^{-4\mu\pi^2k^2(t-t_0)}), \tag{16}$$

we have that

$$a_k(t) - \frac{b_k(1 - e^{-4\mu\pi^2k^2(t-t_0)})}{4\mu\pi^2k^2} = a_k(t_0)e^{-\int_{t_0}^t 2\pi ikv+4\mu\pi^2k^2} +$$

$$+b_k \int_{t_0}^t e^{-\int_r^t 4\mu\pi^2k^2}(e^{-\int_r^t 2\pi ikv} - 1)dr.$$

Taking limits when $t \to \infty$, we get $a_k(t) - (1 - e^{-4\mu\pi^2k^2(t-t_0)})\frac{b_k}{4\mu\pi^2k^2} \to 0$, since $a_k(t_0)e^{-\int_0^t 2\pi ikv+4\mu\pi^2k^2} \to 0$ and from Eq. (13) (Lemma 1) we have that $b_k \int_{t_0}^t e^{-\int_r^t 4\mu\pi^2k^2}(e^{-\int_r^t 2\pi kiv} - 1) \to 0$.

Now taking into account that $(1 - e^{-4\mu\pi^2k^2(t-t_0)})\frac{b_k}{4\mu\pi^2k^2}$ converges to $\frac{b_k}{4\mu\pi^2k^2}$ for large time we conclude that:

$$\begin{cases} a_k(t) \to \frac{b_k}{4\mu\pi^2k^2} \\ I(t) = 2Re(\sum_{k\in(K\cap J)_+} a_k(t)c_{-k}) \to \dfrac{I_0}{2\mu\pi^2} = I_0^* \end{cases} \tag{17}$$

with

$$I_0^* = \frac{I_0}{2\mu\pi^2} = 2Re(\sum_{k\in(K\cap J)_+} \frac{b_kc_{-k}}{4\mu\pi^2k^2}).$$

We prove now

$$T(t,x) = \sum_k a_k(t)e^{2\pi kix} \to \theta_\infty = \sum_k \frac{b_k}{4\mu\pi^2k^2}e^{2\pi kix} \text{ in } \dot{H}^1_{per}(0,1)$$

and we also note

$$\frac{\partial^2\theta_\infty}{\partial x^2} = \frac{-1}{\mu}\sum_k b_ke^{2\pi kix} = \frac{-1}{\mu}h(x).$$

In effect, first from $T(t,x) = \sum_k a_k(t)e^{2\pi kix} \in \dot{H}^1_{per}(0,1)$ for every $t \geq t_0 \geq 0$ we have that $\sum_k k^2|a_k(t)|^2 < \infty$ for every $t \geq t_0 \geq 0$, and using $h(x) = \sum_k b_ke^{2\pi kix} \in \dot{L}^2_{per}(0,1)$ we also have that $\sum_k |b_k|^2 < \infty$, and then

$$\sum_{k=m+1}^{\infty} k^2|a_k(t)|^2 \to 0, \text{ for every } t \geq t_0 \geq 0 \text{ and } \sum_{k=m+1}^{\infty} |b_k|^2 \to 0 \text{ as } m \to \infty.$$

Next, we will prove that

$$\sum_{k=m+1}^{\infty} k^2|a_k(t)|^2 \to 0 \text{ as } m \to \infty, \text{ uniformly for } t \text{ large.} \tag{18}$$

From (15), taking into account that $|e^{-\int_r^t 2\pi ikv}| = 1$ together with (16), we get

$$|a_k(t)| \leq |a_k(t_0)|e^{-4\mu\pi^2k^2(t-t_0)} + \frac{|b_k|}{4\mu\pi^2k^2}(1 - e^{-4\mu\pi^2k^2(t-t_0)}).$$

Therefore, using now $e^{-4\mu\pi^2k^2(t-t_0)} \leq 1$ and $(1 - e^{-4\mu\pi^2k^2(t-t_0)}) \leq 1$, we get

$$|a_k(t)| \leq |a_k(t_0)| + \frac{|b_k|}{4\mu\pi^2k^2}.$$

Thus, we obtain that

$$\sum_{k=m+1}^{\infty} k^2|a_k(t)|^2 \leq C\Big(\sum_{k=m+1}^{\infty} k^2|a_k(t_0)|^2 + \frac{1}{16\mu^2\pi^4}\sum_{k=m+1}^{\infty} k^2|b_k|^2\Big)$$

since $\frac{1}{k^2} \leq 1$, with $C > 0$ independent of $k$, $m$ and $t$, and we conclude (18).

Finally, we note that

$$\|(T(t,x) - \theta_\infty)_x\|^2_{L^2_{per}(0,1)} \leq 4\pi^2\sum_k k^2|a_k(t) - \frac{b_k}{4\mu\pi^2k^2}|^2 \leq 4\pi^2\sum_{k=1}^{m} k^2|a_k(t) - \frac{b_k}{4\mu\pi^2k^2}|^2 +$$

$$+4\pi^2\sum_{k=m+1}^{\infty} k^2|a_k(t) - \frac{b_k}{4\mu\pi^2k^2}|^2 = 4\pi^2 S_m(t) + 4\pi^2 R_{m+1}(t),$$

where

$$R_{m+1}(t) = \sum_{k=m+1}^{\infty} k^2|a_k(t) - \frac{b_k}{4\mu\pi^2k^2}|^2 \leq C\sum_{k=m+1}^{\infty} k^2|a_k(t)|^2 + \frac{C}{16\mu^2\pi^4}\sum_{k=m+1}^{\infty} |b_k|^2 \to 0$$

as $m \to \infty$ uniformly for $t$ large thanks to (18).

Then, for every $\eta > 0$ there exists $m_0(\eta) > 0$ such that $4\pi^2R_{m_0+1}(t) < \frac{\eta}{2}$. Therefore, using again (17), we obtain $t_0(\eta) > 0$ enough large, such that we also have $4\pi^2S_{m_0}(t) < \frac{\eta}{2}$ for every $t \geq t_0$, where

$$S_{m_0}(t) = \sum_{k=1}^{m_0} k^2|a_k(t) - \frac{b_k}{4\mu\pi^2k^2}|^2.$$

This is, we get

$$\|(T(t,x)-\theta_\infty)_x\|^2_{\dot{L}^2_{per}(0,1)} \to 0 \text{ as } t \to \infty.$$

Analogously, we also prove that

$$\|T(t,x)-\theta_\infty\|^2_{\dot{L}^2_{per}(0,1)} \to 0 \text{ as } t \to \infty,$$

and we get that $T(t,x) \to \theta_\infty$ in $\dot{H}^1_{per}(0,1)$.

To conclude, we study now when the velocity $v(t)$ and the acceleration $w(t)$ go to zero. From (10) we can read the equation for $v$, the first equation of system Eq. (8), as

$$\varepsilon\frac{d^2v}{dt^2}+\frac{dv}{dt}+G_0v = I(t), \text{ with } I(t) = 2Re(\sum_{k\in(K\cap J)_+} a_k(t)c_{-k}).$$

(I) First, if we denote by $v_H(t)$ any solution of linear homogeneous equation given by:

$$\varepsilon\frac{d^2v}{dt^2}+\frac{dv}{dt}+G_0v = 0 \text{ with } \varepsilon>0, G_0>0$$

then,

$$v_H(t) = C_1e^{\lambda_1 t}+C_2e^{\lambda_2 t}$$

where $\lambda_1, \lambda_2$ are given by the solutions of $P(\lambda) = \varepsilon\lambda^2+\lambda+G_0 = 0$ and both are negative real numbers or complex numbers with negative real part. Therefore, we have that $v_H(t) \to 0$ as $t \to \infty$ for every constants $C_1, C_2$.

(II) Second, using (17) this is, $\delta(t) = I(t) - I_0^* \to 0$ as $t \to \infty$, we will prove now for every $v_p(t)$ solution of

$$\varepsilon\frac{d^2v_p}{dt^2}+\frac{dv_p}{dt}+G_0v_p = \delta(t), \tag{19}$$

satisfies that $v_p(t) \to 0$ as $t \to \infty$.

In effect, we can rewrite (19) as

$$\frac{dv*}{dt}+av* = \delta(t), \text{ with } v* = \varepsilon\frac{dv_p}{dt}+bv_p \tag{20}$$

where $a, b$ are given by

$$a = \frac{1 + \sqrt{1 - 4\varepsilon G_0}}{2\varepsilon}, \quad b = \frac{G_0}{a}. \tag{21}$$

Then:

$$v * (t) = e^{-a(t-t_0)} v * (t_0) + \int_{t_0}^{t} e^{-a(t-s)} \delta(s) ds \tag{22}$$

and now, taking into account that $a$, is positive real numbers or complex number with positive real part, together with $\delta(t) \to 0$; we get $v * (t) \to 0$ as $t \to \infty$.

Now, we note that:

$$\frac{dv_p}{dt} + \frac{b}{\varepsilon} v_p = v * (t), \text{ and } v_p(t) = e^{-\frac{b}{\varepsilon}(t-t_0)} v_p(t_0) + \int_{t_0}^{t} e^{-\frac{b}{\varepsilon}(t-s)} \frac{v * (s)}{\varepsilon} ds. \tag{23}$$

Working as above and taking into account that $b$, is positive real number or complex number with positive real part, together with $v{*}(t) \to 0$; we conclude that $v_p(t) \to 0$ as $t \to \infty$.

(III) Finally, we consider the equation for the velocity

$$\varepsilon \frac{d^2 v}{dt^2} + \frac{dv}{dt} + G_0 v = I(t) = (I(t) - I_0^*) + I_0^* \text{ with } I_0^* = \frac{I_0}{2\mu\pi^2} \tag{24}$$

and using the superposition principle of solutions, we have that:

$$v(t) = C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t} + v_p + \frac{I_0^*}{G_0}$$

for some constants $C_1, C_2$, where $v_H = C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t}$ is a solution for the homogeneous linear equation $\varepsilon \frac{d^2 v}{dt^2} + \frac{dv}{dt} + G_0 v = 0$, $\frac{I_0^*}{G_0}$ is a particular solution of linear equation $\varepsilon \frac{d^2 v}{dt^2} + \frac{dv}{dt} + G_0 v = I_0^*$ and $v_p$ is a particular solution of linear equation $\varepsilon \frac{d^2 v}{dt^2} + \frac{dv}{dt} + G_0 v = \delta(t) = (I(t) - I_0^*)$.

Then, from the above parts I and II, we conclude that

$$v(t) \to \frac{I_0^*}{G_0} \text{ as } t \to \infty.$$

(i) In particular, when $I_0 = 0$, i.e. $I_0^* = \frac{I_0}{2\mu\pi^2} = 0$, we get $v(t) \to 0$. Moreover, $\delta(t) = I(t) - G_0 v(t) \to 0$ as $t \to \infty$, since from (17) we have that $I(t) \to I_0^* = 0$, and using now (24) we obtain that,

$$\varepsilon \frac{dw}{dt} + w = \delta(t), \text{ and } w(t) = e^{-c(t-t_0)} w(t_0) + \int_{t_0}^{t} e^{-c(t-s)} c\delta(s) ds \qquad (25)$$

with $c = \frac{1}{\varepsilon} > 0$. Thus, we also prove that $w(t) \to 0$ as $t \to \infty$ and we conclude.

(ii) We also note that if $I_0 \neq 0$, then $I_0^* = \frac{1}{2\mu\pi^2} \neq 0$ and we get $liminf_{t\to\infty}|v(t)| > 0$, which implies that $\int_0^\infty |v(s)|ds = \infty$. This result is in contradiction with the initial condition $\int_0^\infty |v(s)|ds < \infty$, this implies it is not a valid hypothesis.

(iii) If $K \cap J = \emptyset$ from (10) we can read the equation for $v$, the first equation of system Eq. (8), as

$$\varepsilon \frac{d^2v}{dt^2} + \frac{dv}{dt} + G_0 v = 0$$

and then we get the velocity goes to zero. Moreover, if $v(t) \to 0$, working as above from $\varepsilon \frac{dw}{dt} + w = \delta(t)$ with $\delta(t) = -G_0 v(t) \to 0$ we get also that $w(t) \to 0$.

Finally, we can read the equation for $a_k(t)$, the second equation of system Eq. (8), as

$$\frac{da_k}{dt} + 4\mu\pi k^2 a_k(t) = b_k - \delta(t) \text{ with } \delta(t) = 2\pi k i v(t) \to 0 \text{ as } t \to \infty.$$

Moreover, using again the superposition principle of solutions, we have for some $C > 0$, positive constant depending on the initial condition $a_k(t_0)$, we have that

$$a_k(t) = Ce^{-4\mu\pi k^2 t} + \frac{b_k}{4\mu\pi i k^2} - a_k^p(t) \qquad (26)$$

where $a_k^p(t)$ is the particular solution of

$$\frac{da_k^p}{dt} + 4\mu\pi k^2 a_k^p(t) = \delta(t), \text{ with } \mu > 0 \text{ and } \delta(t) \to 0 \text{ as } t \to \infty,$$

hence $a_k^p(t) \to 0$ as $t \to \infty$. This is, from (26) we get

$$a_k(t) \to \frac{b_k}{4\mu\pi k^2} \text{ as } t \to \infty.$$

Working as above, we also get $T(t,x) \to \theta_\infty$ in $\dot{H}^1_{per}(0,1)$; therefore, in this case, for every solutions we get $(w(t), v(t), T(t,x)) \to (0,0,\theta_\infty)$. This is, the global attractor for system Eq. (8) is reduced to a point $(0,0,\theta_\infty)$ and we conclude. □

### 3.3 Concluding Remarks

Recalling that functions associated to circuit geometry, $f$, and to prescribed heat flux, $h$, are given by $f(x) = \sum_{k\in J} c_k e^{2\pi kix}$ and $h(x) = \sum_{k\in K} b_k e^{2\pi kix}$, respectively. In Jiménez-Casas et al. [7], using the operator abstract theory, it is proved that if $K \cap J = \emptyset$, then the global attractor for system Eq. (1) in $\mathbb{R}^2 \times \dot{H}^1_{per}$ is reduced to a point $\{(0,0,\theta_\infty)\}$, where $\theta_\infty$ is the unique solution in $\dot{H}^2_{per}(0,1)$ of $-\mu\frac{\partial^2\theta}{\partial x^2} = h(x)$.

In this sense, Proposition 2 offers the possibility to obtain the same asymptotic behaviour for the dynamics, i.e., the attractor is also reduced to a point taking functions $f$ and $h$ such that with $K \cap J = \emptyset$. Moreover, also without this condition, that is with $K \cap J \neq \emptyset$ a finite set, it is enough that $Re(\sum_{k\in(K\cap J)_+} \frac{b_k c_{-k}}{k^2}) = 0$, to get the solutions with finite length converge to the rest equilibrium, when we consider the linear friction case $G = G_0$. We note, the result about the inertial manifold (Jiménez-Casas et al. [7]) reduces the asymptotic behaviour of the initial system Eq. (1) to the dynamics of the reduced explicit system Eq. (9) with $k \in K \cap J$.

We also observe from the analysis above, that it is possible to design the geometry of circuit, $f$, and/or heat flux, $h$, so that the resulting system has an arbitrary number of equations of the form $N = 4n_0 + 1$ where $n_0$ is the number of elements of $(K \cap J)_+$ and we consider the real and imaginary parts of relevant coefficients for the temperature $a_k(t)$ with $k \in (K \cap J)_+$ .

Note that it may be the case that $K$ and $J$ are infinite sets, but their intersection is finite. Also, for a circular circuit we have $f(x) \sim asen(x) + bcos(x)$, i.e. $J = \{\pm 1\}$ and then $K \cap J$ is either $\{\pm 1\}$ or the empty set.

Recently, we have considered a thermosyphon model containing a viscoelastic fluid and we have shown chaos in some closed-loop thermosyphon model with one-component viscoelastic fluid not only in this model [7, 8], also in other kind of transfer law (Yasappan and Jiménez-Casas et al. [17]), and even in some cases with a viscoelastic binary fluid (Jiménez-Casas and Castro [5], Yasappan and Jiménez-Casas et al. [18]).

# 4 Numerical Experiments

## 4.1 Introduction

In this section, we will comment the results proved in [8] about the numerical experiments obtained for the resolution of the differential equations, using a fourth-order explicit Runge-Kutta method for stiffness equations, following the method used in previous works [14]. We solve a system of ordinary differential equations which are the projection of the partial differential equations on the inertial manifold derived in the preceding sections. All the variables and equations that we deal with are adimensional.

Specifically, we integrate the system of equations, where we consider only the coefficients of temperature $a_k(t)$ with $k \in K \cap J$ (relevant modes). Then,

$$\begin{cases} \frac{dw}{dt} + \frac{w}{\varepsilon} + \frac{G(v)v(t)}{\varepsilon} = \frac{2}{\varepsilon} Real \left( \sum_{k \in K \cap J} a_k(t) c_{-k} \right) \\ \frac{dv}{dt} = w \\ \dot{a}_k(t) + a_k(t)(2\pi k i v + \mu 4\pi^2 k^2) = b_k \end{cases}$$

where $a_{-k} = \bar{a}_k$, $b_{-k} = \bar{b}_k$ and $c_{-k} = \bar{c}_k$ since all the physical observable are real functions.

In particular, we will consider a thermosyphon with a circular geometry, so $J = \{\pm 1\}$ and $K \cap J = \{\pm 1\}$. Consequently, we can take $k = 1$ and omit the equation for $k = -1$. Hence,

$$\begin{cases} \frac{dw}{dt} = \frac{2a_1 c_{-1}}{\varepsilon} - \frac{w}{\varepsilon} - \frac{G(v)v(t)}{\varepsilon}, \\ \frac{dv}{dt} = w, \\ \dot{a}_1(t) + a_1(t)(2\pi i v + \mu 4\pi^2) = b_1 \end{cases}$$

where the unknowns are $w(t)$ (the acceleration of the fluid), $v(t)$ (velocity of the fluid) and $a_1(t)$ (the Fourier mode of the temperature). More complex geometries will result in higher dimensional dynamics on the inertial manifold.

In order to reduce the number of parameters we make the change of variables $a_1 c_{-1} \rightarrow a_1$ and we then define the real and imaginary parts of the equations in the following way:

$$a_1(t) = a^1(t) + i a^2(t), \tag{27}$$

$$b_1 = A + iB \tag{28}$$

with $A \in \mathbb{R}, B \in \mathbb{R}$. Hence, our central results correspond to the system of equations

$$\begin{cases} \frac{dw}{dt} = \frac{2a^1}{\varepsilon} - \frac{w}{\varepsilon} - \frac{G(v)v(t)}{\varepsilon}, \\ \dot{v} = w, \\ \dot{a}^1 = A - \mu 4\pi^2 a^1 + v 2\pi a^2, \\ \dot{a}^2 = B - \mu 4\pi^2 a^2 - v 2\pi a^1 \end{cases} \tag{29}$$

Note that it is a system of four equations with four unknowns where we need to make explicit choices for the constitutive laws for both the fluid-mechanical and thermal properties. For the friction law $G(v)$ and heat flux $h(x)$ we will take the one used in the reference [14]. For the numerical experiments which are of a similar model of thermosyphon for a fluid with one component, they use the function $G(v) = (|v| + 10^{-4})$.

The function $G(v)$ has a clear physical meaning; it interpolates between a low Reynolds number friction law (in which the overall friction $G(v)v$ is non-linear, Stokes friction law) and high Reynolds number (in which the friction is a quadratic law).

Besides, $A$ and $B$, which refer in this model to the position-dependant $(x)$ heat flux inside the loop will be used as tuning parameters. Without loss of generality, we will assume $A = 0$ in order to simplify, in analogy with the Lorenz's model, as it is shown in reference [14] (changing $A$ and $B$ simultaneously only results in a change in the *phase* of initial temperature profile).

We have carried out two different sets of numerical experiments with regard to heat diffusion.

The first set of numerical experiments are carried out keeping the heat diffusion to zero as it was done in [14]. And the second set of numerical experiments are performed with heat diffusion.

The initial conditions are fixed as $w(0) = 0, v(0) = 0, a_1(0) = 1, a_2(0) = 1$. This split would appear naive as diffusion tends to smooth the solution, however, as the order of the equations changes in the presence of diffusion (from first to second order, due to the Laplacian) it is worth studying both cases separately.

Numerical analysis has been carried out keeping $\varepsilon$ the viscoelastic coefficient as the tuning parameter ranging from 100 to 0.0001 and B the heat flux also as another tuning parameter ranging from 1 to 10,000. The impact of $\varepsilon$ on the system has been keenly observed for various intervals of time $t$, as short as 50 time units and as long as 5000 time units. We will show that in analogy with the classical Lorenz's system, as $\varepsilon$ varies, the dynamics of the model undergoes various transformations including steady asymptotic behavior, meta-stable chaos, i.e., transient irregular behavior followed by convergence to equilibria, periodic behaviors and chaotic progressions (see [8]).

## 4.2 Numerical Conclusions

The physical and mathematical implications of the resulting system of ODEs which describe the dynamics at the inertial manifold is analyzed numerically. The role of the parameter $\varepsilon$ which contains the viscoelastic information of the fluid was treated with special attention. We studied the asymptotic behavior of the system for different values of $\varepsilon$ the coefficient of viscoelasticity. We can conclude that for larger values of $\varepsilon$ the system behaves more chaotic.

**Acknowledgements** I want to thank M. Castro and J. Yasappan for the numerical experiments and the referee for their helpful comments. This research was partially supported by grants MTM2012-31298,MTM2016-75465-P and Project FIS2013-47949-C2-2 from Ministerio de Economia y Competitividad, Spain; by GR58/08 Grupo 920894 BSCH-UCM from Grupo de Investigación CADEDIF, Grupo de Dinámica No Lineal (U.P. Comillas)Spain and by the Project FIS2016-78883-C2-2-P (AEI/FEDER,UE).

## References

1. Bravo-Gutiérrez, M.E, Castro, M., Hernández-Machado, A., Poire, A.: Controlling viscoelastic flow in microchannels with slip. Langmuir (ACS Publ.) **27**, 2075–2079 (2011)
2. Foias, C., Sell, G.R., Temam, R.: Inertial manifolds for nonlinear evolution equations. J. Diff. Equ. **73**, 309–353 (1985)
3. Henry, D.: Geometric Theory of Semilinear Parabolic Equations. Lectures Notes in Mathematics, vol. 840. Springer, Berlin (1981)
4. Herrero, M.A., Velázquez, J.L.L.: Stability analysis of a closed thermosyphon. Eur. J. Appl. Math. **1**, 1–24 (1990)
5. Jiménez-Casas, A., Castro, M.: A thermosyphon model with a viscoelastic binary fluid. Electron. J. Differ. Equ. (Conference 22), 53–61 (2015). ISSN: 1072–6691
6. Jiménez-Casas, A., Ovejero, A.M.L.: Numerical analysis of a closed-loop thermosyphon including the Soret effect. Appl. Math. Comput. **124**, 289–318 (2001)
7. Jiménez-Casas, A., Castro, M., Yasappan, J.: Finite-dimensional behaviour in a thermosyphon with a viscoelastic fluid. Discrete Conti. Dynam. Syst. (9th AIMS Conference Sool.) **2013**, 375–384 (2013)
8. Jiménez-Casas, A., Castro, M., Yasappan, J.: Chaotic behavior of the closed loop thermosyphon model with memory effects. Chaotic Model. Simul. **2**, 281–288 (2013)
9. Keller, J.B.: Periodic oscillations in a model of thermal convection. J. Fluid Mech. **26**, 599–606 (1966)
10. Liñan, A.: Analytical description of chaotic oscillations in a toroidal thermosyphon. In: Velarde, M.G., Christov, C.I. (eds.) Fluid Physics, Lecture Notes of Summer Schools, pp. 507–523. World Scientific, River Edge (1994)
11. Morrison, F.: Understanding Rheology. Oxford University Press, Oxford (2001)
12. Rodríguez-Bernal, A.: Inertial manifolds for dissipative semiflows in Banach spaces. Appl. Anal. **37**, 95–141 (1990)
13. Rodríguez-Bernal, A.: Attractor and inertial manifolds for the dynamics of a closed thermosyphon. J. Math. Anal. Appl. **193**, 942–965 (1995)
14. Rodríguez-Bernal, A., Van Vleck, E.S.: Diffusion induced chaos in a closed loop thermosyphon. SIAM J. Appl. Math. **58**, 1072–1093 (1998)

15. Velázquez, J.J.L.: On the dynamics of a closed thermosyphon. SIAM J. Appl. Math. **54**, 1561–1593 (1994)
16. Welander, P.: On the oscillatory instability of a differentially heated fluid loop. J. Fluid Mech. **29**, 17–30 (1967)
17. Yasappan, J., Jiménez-Casas, A., Castro, M.: Asymptotic behavior of a viscoelastic fluid in a closed loop thermosyphon: physical derivation, asymptotic analysis and numerical experiments. Abstr. Appl. Anal. **2013**, Article ID: 748683 (2013)
18. Yasappan, J., Jiménez-Casas, A., Castro, M.: Stabilizing interplay between thermodiffusion and viscoelasticity in a closed-loop thermosyphon. Discrete Conti. Dynam. Syst. Ser. B **20**, 3267–3299 (2015)

# Conservation Laws and Potential Symmetries for a Generalized Gardner Equation

**Rafael de la Rosa, Tamara M. Garrido, and María Santos Bruzón**

**Abstract** In this paper, a generalized Gardner equation with nonlinear terms of any order has been analyzed from the point of view of group transformations and conservation laws. The generalized Gardner equation appears in many areas of physics and it is widely used to model a great variety of wave phenomena in plasma and solid state. By using the direct method of the multipliers, we have obtained an exhaustive classification of all low-order conservation laws which the generalized Gardner equation admits. Then, taking into account these conserved vectors we have determined the associated potential systems and we have searched for potential symmetries of the equation. Furthermore, we have determined and examined its first-level and second-level potential systems. From the first-level potential system we have found two new nonlocal conserved vectors.

**Keywords** Partial differential equations · Conservation laws · Symmetries · Potential symmetries

## 1 Introduction

The Korteweg-de Vries (KdV) equation

$$u_t + \lambda u u_x + \mu u_{xxx} = 0,$$

first emerged to look for the most efficient design for canal boats and it describes the physics and the dynamics of shallow water [9]. Through the years, many applications for the KdV equation have been found, for instance, it can be used as a mathematical model to analyse different phenomena in mathematical physics, nonlinear dynamics and plasma physics. The problem lies in the fact that when

R. de la Rosa (✉) · T. M. Garrido · M. S. Bruzón
Departamento de Matemáticas, Universidad de Cádiz, Cádiz, Spain
e-mail: rafael.delarosa@uca.es; tamara.garrido@uca.es; m.bruzon@uca.es

© Springer Nature Switzerland AG 2019

J. L. García Guirao et al. (eds.), *Recent Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 18,
https://doi.org/10.1007/978-3-030-00341-8_7

phenomena become more sophisticated, the KdV equation is considered too simple to model them.

For this reason, different generalizations of the KdV have been given. For example, the Gardner equation

$$u_t + uu_x + au^2u_x + u_{xxx} = 0,$$

also known as combined KdV-mKdV equation, is a generalization of the KdV equation which involves more than one nonlinear term. This equation is applied in several branches of physics, such as plasma physics, fluid dynamics and quantum field theory. Specifically, this equation is used to model a great variety of wave phenomena in plasma and solid state. Several papers have been devoted to analyse some families of generalized Gardner equations [11, 13, 18, 25].

In this paper we consider a generalized Gardner equation given by

$$u_t + a\,u^n\,u_x + b\,u^{2n}\,u_x + c\,u_{xxx} + d\,u_x + e\,u + f = 0, \tag{1}$$

where $n$ is a positive constant, $a$ and $b$ are not simultaneously equal to zero, $c \neq 0$, $d$, $e$ and $f$ are arbitrary constants.

Symmetry groups stand out because they can be used to determine exact solutions of partial differential equations (PDEs) [14, 15, 17, 23, 24]. The study of the generalized Gardner equation (1) started in a previous work [10] where classical and nonclassical symmetries of the considered equation were obtained. Moreover, taking into account the similarity variables and the similarity solutions, Eq. (1) was reduced into some ordinary differential equations and, finally, some travelling wave solutions were constructed.

On the other hand, for any evolution equation it is possible to obtain a complete classification of local low-order conservation laws [2, 4, 5, 16, 19–22]. In [12], an analysis of the conservation laws admitted by Eq. (1) was shown. Those conservation laws which could have a physical sense come from low-order multipliers. Moreover, conservation laws have many well-known applications in numerical methods and mathematical analysis, for instance, they can be used to determine the existence, uniqueness and stability of solutions of PDEs or to construct exact solutions.

Bluman et al. [6, 7] introduced a method to find a new class of symmetries for a PDE when it can be written in a conserved form. These symmetries are nonlocal symmetries called potential symmetries. Potential symmetries can be used to obtain nontrivial solutions of PDEs.

The aim of this work is to investigate the generalized Gardner equation (1) from conservation laws and potential symmetries. Making use of the direct method of the multipliers [1–5, 8], we have provided an exhaustive classification of all low-order conservation laws admitted by Eq. (1) depending on the different values of the parameters. We use some conservation laws of the Gardner equation (1) and we obtain the associated first-level potential systems. Next, we classify the local conservation laws of each of these potential systems, which yield second-level potential systems. Finally, we investigate nonlocal symmetries and nonlocal conservation laws of these potential systems.

## 2 Conservation Laws

A local conservation law for Eq. (1) is a space-time divergence expression

$$D_t T(t, x, u, u_t, u_x, \ldots) + D_x X(t, x, u, u_t, u_x, \ldots) = 0, \tag{2}$$

which is identically zero on all solutions of the PDE, where $T$ and $X$ represent the conserved density and the spatial flux respectively, whereas $D_t$ and $D_x$ denote the total derivative operators with respect to $t$ and $x$. Although this concept has its origin in physics, conservation laws have many applications in the study of differential equations or systems of differential equations. For instance, they can be used in mathematical analysis and numerical methods to enquire into the existence, uniqueness and stability of solutions of PDEs. Moreover, the integrability of a differential equation is strongly related to the existence of a large number of conservation laws.

Two conservation laws are considered to be locally equivalent if they differ by a locally trivial conservation law

$$\begin{aligned} \tilde{T} &= T + D_x \Theta(t, x, u, u_t, u_x, \ldots), \\ \tilde{X} &= X + D_t \Theta(t, x, u, u_t, u_x, \ldots), \end{aligned} \tag{3}$$

i.e. there is a function $\Theta(t, x, u, u_t, u_x, \ldots)$ so that $(D_x\Theta, -D_t\Theta)$ holds for every solution $u(t, x)$.

Anco and Bluman [2, 4, 5] showed a general method to determine conserved quantities of a given PDE. To apply this method is necessary the concept of multiplier. A multiplier is a non-singular function $Q(t, x, u, u_t, u_x, \ldots)$ satisfying that $(u_t + a\,u^n\,u_x + b\,u^{2n}\,u_x + c\,u_{xxx} + d\,u_x + e\,u + f)Q$ is a divergence expression for all functions $u(t, x)$, not only solutions of Eq. (1).

Each non-trivial conservation law can be stated in a general form

$$\frac{d}{dt}\int_{\Omega} T dx = -X\Big|_{\partial\Omega},$$

where $\Omega \subseteq \mathbb{R}$ is a fixed spatial domain. According to this method, if we move off of the set of solutions of Eq. (1), any conservation law can be expressed by using the characteristic form

$$D_t\tilde{T} + D_x\tilde{X} = \Big(u_t + a\,u^n\,u_x + b\,u^{2n}\,u_x + c\,u_{xxx} + d\,u_x + e\,u + f\Big)\,Q. \tag{4}$$

In particular, from the characteristic form (4) it follows that each conserved vector given by (2) derives from a multiplier $Q$ of Eq. (1). Multipliers $Q$ are obtained by requiring that the divergence condition must be verified identically

$$\frac{\delta}{\delta u}\left(\left(u_t + a\,u^n\,u_x + b\,u^{2n}\,u_x + c\,u_{xxx} + d\,u_x + e\,u + f\right)Q\right) = 0,$$

where $\frac{\delta}{\delta u} = \partial_u - D_x\partial_{u_x} - D_t\partial_{u_t} + D_x D_t\partial_{u_{xt}} + D_x^2\partial_{u_{xx}} + \ldots$, denotes the variational derivative. Divergence condition can be split in explicit form with respect to $u_t$, $u_{xxx}$ and their differential consequences which yields an overdetermined system in $Q$. Finally, the conserved vectors can be determined from the multipliers by integrating the characteristic equation (4) [2, 8].

The generalized Gardner equation (1) admits low-order multipliers $Q = Q(t, x, u, u_x, u_{xx})$ in the following cases

**Case 1:** $e \neq 0$ : In this case, the following multipliers are obtained

$$\begin{aligned} Q_1 &= exp(e\,t), \\ Q_2 &= exp(2\,e\,t)\left(u + \frac{f}{e}\right). \end{aligned}$$

The conserved vectors are given by

$$\begin{aligned} T_1 &= exp(e\,t)\left(u + \frac{f}{e}\right), \\ X_1 &= exp(e\,t)\left(c\,u_{xx} + \frac{a\,u^{n+1}}{n+1} + \frac{b\,u^{2n+1}}{2n+1} + d\,u\right). \end{aligned} \tag{5}$$

$$\begin{aligned} T_2 &= \frac{1}{2}exp(2\,e\,t)\left(u^2 + \frac{2\,f}{e}u + \frac{f^2}{e^2}\right), \\ X_2 &= exp(2\,e\,t)\left(c\,uu_{xx} + \frac{c\,f}{e}u_{xx} - \frac{c}{2}u_x^2 + \frac{d\,f}{e}u + \frac{d}{2}u^2 + \frac{b}{2n+2}u^{2n+2}\right. \\ &\quad \left. + \frac{b\,f}{e(2n+1)}u^{2n+1} + \frac{a}{n+2}u^{n+2} + \frac{a\,f}{e(n+1)}u^{n+1}\right). \end{aligned} \tag{6}$$

**Subcase 1.1:** Additionally, if $a = 0$ and $n = \frac{1}{2}$, besides $Q_1$ and $Q_2$ the following multiplier is obtained

$$Q_3 = \frac{exp(e\,t)}{b}\left(e\,(x - d\,t) + b\,(u + f\,t)\right).$$

The conserved vector is given by

$$T_3 = \frac{exp(e\,t)}{2\,b\,e^2}\Big(b\,e^2\,u^2 + 2\,e^2(b\,f\,t + e\,x - d\,e\,t)u$$

$$+2f(e^2\,x - d\,e^2\,t + e\,d + b\,e\,f\,t - b\,f)\Big),$$

$$X_3 = \frac{exp(e\,t)}{6\,b}\,(6\,b\,c\,u\,u_{xx} + 6\,c(b\,f\,t + e\,x - d\,e\,t)u_{xx} - 3\,b\,c\,u_x^2 - 6\,e\,c\,u_x + 2\,b^2\,u^3$$

$$+3\,b\,(b\,f\,t + e\,x \;\; -d\,e\,t + d)\,u^2 + 6\,d(b\,f\,t + e\,x - d\,e\,t)u\Big).$$

**Case 2:** $e = 0$: In this case, the following multipliers are obtained

$$Q_4 = 1,$$

$$Q_5 = u + f\,t,$$

$$Q_6 = c\,u_{xx} + f\,x - d\,f\,t + \frac{b}{2\,n+1}u^{2\,n+1} + \frac{a}{n+1}u^{n+1}.$$

Consequently, we obtain the following conserved vectors

$$T_4 = b\,t\,u^{2n}u_x + a\,t\,u^n u_x + d\,t\,u_x + u + f\,t,$$

$$X_4 = -b\,t\,u^{2n}u_t - a\,t\,u^n u_t - d\,t\,u_t + c\,u_{xx}.$$

$$T_5 = b\,t\,u^{2n+1}u_x + \frac{b\,f\,t^2}{2}\,u^{2n}u_x + a\,t\,u^{n+1}u_x + \frac{a\,f\,t^2}{2}\,u^n u_x + d\,t\,uu_x + \frac{d\,f\,t^2}{2}u_x$$

$$+\frac{1}{2}u^2 + f\,t\,u + \frac{f^2\,t^2}{2},$$

$$X_5 = -b\,t\,u^{2n+1}u_t - \frac{b\,f\,t^2}{2}\,u^{2n}u_t - a\,t\,u^{n+1}u_t - \frac{a\,f\,t^2}{2}\,u^n u_t - d\,t\,uu_t - \frac{d\,f\,t^2}{2}u_t$$

$$+c\,uu_{xx} + c\,f\,t\,u_{xx} - \frac{c}{2}u_x^2.$$

$$T_6 = -\frac{c}{2}u_x^2 + f(x-d\,t)u + \frac{b}{2(n+1)(2\,n+1)}u^{2\,n+2} + \frac{4\,a}{(n+1)(n+2)}u^{n+2}$$
$$+2\,f^2\,t(2\,x-d\,t),$$

$$X_6 = -c^2\,u_x u_{xxx} + \frac{c^2}{2}u_{xx}^2 + c\left(\frac{b}{2\,n+1}u^{2n+1} + \frac{a}{n+1}u^{n+1} + f(x-d\,t)\right)u_{xx}$$
$$-c\left(b\,u^{2\,n} + a\,u^n + \frac{d}{2}\right)u_x^2 - c\,f\,u_x + d\,f(x-d\,t)u + \frac{b^2}{2(2\,n+1)^2}u^{4\,n+2}$$
$$+\frac{b\,f(x-d\,t)}{2\,n+1}u^{2\,n+1} + \frac{a^2+b\,d}{2(n+1)^2}u^{2\,n+2} + \frac{a\,b}{(n+1)(2\,n+1)}u^{3\,n+2}$$
$$+\frac{a\,f(x-d\,t)}{n+1}u^{n+1} + \frac{a\,d}{(n+1)(n+2)}u^{n+2}.$$

**Subcase 2.1:** Moreover, if $a = 0$ and $n = \frac{1}{2}$, besides $Q_4$, $Q_5$ and $Q_6$ we obtain the following multiplier

$$Q_7 = x - \frac{b\,t}{2}\,(f\,t + 2\,u) - d\,t.$$

The conserved vector is given by

$$T_7 = -\frac{b\,t}{2}\,u^2 + \frac{1}{2}\left(2\,x - 2\,d\,t - b\,f\,t^2\right)u - \frac{b\,f^2\,t^3}{6} - \frac{d\,f\,t^2}{2} + f\,x\,t,$$

$$X_7 = -b\,c\,t\,uu_{xx} + \frac{c}{2}\left(2\,x - 2\,d\,t - b\,f\,t^2\right)u_{xx} - \frac{b^2\,t}{3}u^3 + \frac{b}{4}\left(2\,x - 4\,d\,t - b\,f\,t^2\right)u^2$$
$$+\frac{d}{2}\left(2\,x - 2\,d\,t - b\,f\,t^2\right)u + \frac{b\,c\,t}{2}u_x^2 - cu_x.$$

**Subcase 2.2:** Finally, if $f = 0$ and $n = 1$ besides $Q_4$, $Q_5$ and $Q_6$ we obtain the following multiplier

$$Q_8 = 6\,b\,c\,t\,u_{xx} + 2\,b^2\,t\,u^3 + 3\,a\,b\,t\,u^2 + \left(a^2\,t + 2\,b\,d\,t - 2\,b\,x\right)u + a\,d\,t - a\,x.$$

For $Q_8$, the corresponding conserved vector is given by

$$T_8 = 3\,b\,c\,t\,u_x^2 - \frac{b^2\,t}{2}u^4 - a\,b\,t\,u^3 - \left(\frac{a^2\,t}{2} + b\,d\,t - b\,x\right)u^2 - a\,(d\,t - x)\,u,$$

$$\begin{aligned} X_8 = & -3\,b\,c^2\,t\,u_{xx}^2 + a\,c\,\,(x - d\,t)\,u_{xx} - 2\,b^2\,c\,t\,u^3\,u_{xx} + c\left(2\,b\,(x - d\,t) - a^2\,t\right)u\,u_{xx} \\ & -3\,a\,b\,c\,t\,u^2\,u_{xx} + 6\,b^2\,c\,t\,u^2 u_x^2 + 6\,a\,b\,c\,t\,u\,u_x^2 - c\left(b\,(x - 4\,d\,t) - \tfrac{a^2\,t}{2}\right)u_x^2 \\ & +c\,(6\,b\,c\,t\,u_{xxx} - a)\,u_x - 2\,b\,c\,u\,u_x + a\,d\,(x - d\,t)\,u \\ & +\left(b\,d(x - d\,t) + \frac{a^2}{2}(x - 2\,d\,t)\right)u^2 + \left(a\,b\,(x - 2\,d\,t) - \frac{a^3\,t}{3}\right)u^3 \\ & +\left(b^2\left(\frac{x}{2} - d\,t\right) - a^2\,b\,t\right)u^4 - a\,b^2\,t\,u^5 - \frac{b^3\,t}{3}u^6. \end{aligned}$$

## 3 First Level Potential Systems and Symmetries

Bluman et al. introduced a method to find a new class of symmetries for a PDE [7] which can be written in conserved form. Given a conservation law of the form (2), we introduce an auxiliary potential variable $v$ and form an auxiliary potential system

$$\begin{aligned} v_x &= \quad T(t, x, u, u_x, u_t, \ldots), \\ v_t &= -X(t, x, u, u_x, u_t, \ldots). \end{aligned} \tag{7}$$

From conservation laws (5) and (6) admitted by Eq. (1) in the case $e \neq 0$ we determine the associated potential systems

$$\begin{aligned} v_x - u = 0, \\ v_t + \frac{a}{n+1}u^{n+1} + \frac{b}{2n+1}u^{2n+1} + cu_{xx} + du + ev + fx = 0, \end{aligned} \tag{8}$$

and

$$\begin{aligned} v_x - \frac{1}{2}u^2 + \frac{f}{e}u = 0, \\ v_t + \frac{af}{e(n+1)}u^{n+1} + \frac{a}{n+2}u^{n+2} + \frac{bf}{e(2n+1)}u^{2n+1} + \frac{b}{2n+2}u^{2n+2} + cuu_{xx} \\ +\frac{cf}{e}u_{xx} - \frac{c}{2}u_x^2 + \frac{d}{2}u^2 + \frac{df}{e}u + \frac{f^2}{e}x + 2ev = 0. \end{aligned} \tag{9}$$

Taking into account potential system (8) we can replace $u$ in terms of $v_x$ obtaining an auxiliary potential equation

$$v_t + \frac{a}{n+1} v_x^{n+1} + \frac{b}{2n+1} v_x^{2n+1} + c v_{xxx} + d v_x + e v + f x = 0. \tag{10}$$

In order to obtain a complete classification of all point symmetries of systems (8) and (9) we consider a one-parameter Lie group of transformations in the space $(t, x, u, v)$

$$\begin{aligned} t^* &= \; t + \epsilon\, \tau(t, x, u, v) + O(\epsilon^2), \\ x^* &= x + \epsilon\, \xi(t, x, u, v) + O(\epsilon^2), \\ u^* &= u + \epsilon\, \eta(t, x, u, v) + O(\epsilon^2), \\ v^* &= v + \epsilon\, \phi(t, x, u, v) + O(\epsilon^2), \end{aligned} \tag{11}$$

where $\epsilon$ is the group parameter. The vector field takes the following form

$$V = \tau(t, x, u, v)\partial_t + \xi(t, x, u, v)\partial_x + \eta(t, x, u, v)\partial_u + \phi(t, x, u, v)\partial_v. \tag{12}$$

The invariance of systems (8) and (9) under the one-parameter Lie group (11) with infinitesimal generator (12) leads to a system of determining equations for the infinitesimals $\tau(t, x, u, v)$, $\xi(t, x, u, v)$, $\eta(t, x, u, v)$ and $\phi(t, x, u, v)$. After having solved the determining system, omitting tedious calculations, we obtain the following results

**Theorem 1** *The point symmetries admitted by the potential system (8), for arbitrary a, $b \neq 0$, $c \neq 0$, d, $e \neq 0$, f and n, are generated by:*

$$\begin{aligned} V_1 &= \partial_t, \\ V_2 &= \partial_x - \tfrac{f}{e}\partial_v, \\ V_3 &= \exp(-et)\partial_v. \end{aligned} \tag{13}$$

*Furthermore, if $a = 0$ and $n = \frac{1}{2}$ additional point symmetries are generated by*

$$V_4 = \exp(-et)\left(\partial_x - \frac{e}{b}\partial_u + \left(\frac{e}{b}(dt - x) - ft\right)\partial_v\right).$$

Similarly, for system (9) we obtain

**Theorem 2** *The point symmetries admitted by the potential system (9), for arbitrary a, $b \neq 0$, $c \neq 0$, d, $e \neq 0$, f and n, are generated by:*

$$\begin{aligned} V_1' &= \partial_t, \\ V_2' &= \partial_x - \tfrac{f^2}{2e^2}\partial_v, \\ V_3' &= \exp(-2et)\partial_v. \end{aligned} \tag{14}$$

We recall that any Lie group of point transformations (11) yields a potential symmetry of Eq. (1) if it verifies

$$\tau_v^2 + \xi_v^2 + \eta_v^2 \neq 0. \tag{15}$$

If a point symmetry of systems (8) and (9) does not satisfy condition (15), the point symmetry projects to a point symmetry of Eq. (1). Furthermore, a point symmetry of system (8) can be projected to the corresponding potential equation (10) which leads us to a one-parameter Lie group of contact transformations in the space $(t, x, v, v_x)$ whose infinitesimal generator is given by $V = \tau(t, x, v, v_x)\partial_t + \xi(t, x, v, v_x)\partial_x + \eta(t, x, v, v_x)\partial_v + \phi(t, x, v, v_x)\partial_{v_x}$. This symmetry will be considered as a prolonged point symmetry solely when infinitesimals $\tau$, $\xi$ and $\eta$ do not depend on $v_x$. From Theorems 1 and 2 we obtain

**Proposition 1** *All the point symmetries admitted by systems (8) and (9) project to classical symmetries of the generalized Gardner equation (1).*

## 4 Second-Level Potential Systems and Symmetries

In this section, by using the multipliers method we determine all nontrivial low-order conservation laws of the potential equation (10). The potential equation (10) only admits multipliers of second order of the form $Q(t, x, v_{xx})$. We show the results on multipliers of Eq. (10) in the following theorem:

**Theorem 3** *Equation (10), for arbitrary a, $b \neq 0$, $c \neq 0$, d, $e \neq 0$, f and n, admits the following conservation law multiplier*

$$Q_1^* = \exp(2et)v_{xx}.$$

*The conserved vector associated to $Q_1^*$ is given by*

$$\begin{aligned} T_1^* &= -\frac{1}{2}\exp(2et)v_x^2, \\ X_1^* &= \frac{1}{2}\exp(2et)\left(-\frac{2a}{(n+2)}v_x^{n+2} - \frac{b}{(n+1)}v_x^{2n+2} - 2cv_x v_{xxx} + cv_{xx}^2 - dv_x^2 - 2fv\right). \end{aligned} \tag{16}$$

*Moreover, for $a = 0$ and $n = \frac{1}{2}$, besides $Q_1^*$ we obtain*

$$Q_2^* = \exp(et)\left(v_{xx} + \frac{e}{b}\right).$$

*Therefore, we obtain the following conserved vector*

$$
\begin{aligned}
T_2^* &= -\frac{1}{2}\exp(et)\left(v_x^2 - \frac{2e}{b}v\right),\\
X_2^* &= \frac{1}{2}\exp(et)\left(-\frac{2b}{3}v_x^3 - 2cv_xv_{xxx} + cv_{xx}^2 + \frac{2ec}{b}v_{xx}\right.\\
&\quad \left.-dv_x^2 + \frac{1}{b}\left(efx^2 - 2bfv + 2dev\right)\right).
\end{aligned}
\tag{17}
$$

*These conservation laws project to nonlocal conservation laws of the generalized Gardner equation (1), where $e \neq 0$ and $v = \int u\,dx$.*

Making use of Theorem 3, we associate the conserved vector (16) to the potential system given by

$$
\begin{aligned}
w_x + v_x^2 &= 0,\\
w_t - \frac{2a}{(n+2)}v_x^{n+2} - \frac{b}{(n+1)}v_x^{2n+2} - 2cv_xv_{xxx} + cv_{xx}^2 - dv_x^2 - 2fv + 2ew &= 0.
\end{aligned}
\tag{18}
$$

Similarly, for the conservation law (17), we have

$$
\begin{aligned}
w_x + v_x^2 - \frac{2e}{b}v &= 0,\\
w_t - \frac{2b}{3}v_x^3 - 2cv_xv_{xxx} + cv_{xx}^2 + \frac{2ec}{b}v_{xx} - dv_x^2 + \frac{1}{b}\left(efx^2 - 2bfv + 2dev\right) + ew &= 0.
\end{aligned}
\tag{19}
$$

An infinitesimal generator of systems (18) and (19) is given by

$$
W = \tau(t,x,v,w)\partial_t + \xi(t,x,v,w)\partial_x + \phi(t,x,v,w)\partial_v + \psi(t,x,v,w)\partial_w. \tag{20}
$$

By proceeding as in the previous section we obtain a linear system of determining equations involving the infinitesimals $\tau$, $\xi$, $\phi$ and $\psi$ for each potential system (18) and (19). Solving these systems, we obtain

**Theorem 4** *The potential system (18) admits point symmetries for arbitrary $a$, $b \neq 0$, $c \neq 0$, $d$, $e \neq 0$, $f$ and $n$, which are generated by:*

$$
\begin{aligned}
W_1 &= \partial_t,\\
W_2 &= \partial_x,\\
W_3 &= \exp(-2et)\partial_w,\\
W_4 &= F(t)\partial_v + 2f\exp(-2et)\int \exp(2et)F(t)\,dt\,\partial_w.
\end{aligned}
\tag{21}
$$

*where $F(t)$ is an arbitrary function.*

**Theorem 5** *The point symmetries admitted by the potential system (19), for arbitrary $b \neq 0$, $c \neq 0$, $d$, $e \neq 0$ and $f$, are spanned by:*

$$
\begin{aligned}
W_1' &= \partial_t, \\
W_2' &= \partial_x - \tfrac{f}{e}\partial_v - \tfrac{2f}{e^2 b}\left(be^2 + bf - de\right)\partial_w, \\
W_3' &= \exp(-et)\partial_w. \\
W_4' &= \exp(-et)\left(\partial_v + \tfrac{2}{b}\left(bft - edt + ex\right)\partial_w\right).
\end{aligned}
\tag{22}
$$

The vector field (20) projects to a symmetry of the generalized Gardner equation (1). This projected symmetry will be considered a point symmetry solely in the case that $\tau$ does not depend on $x$, $v$, $w$, and the other infinitesimals do not depend on $v$ and $w$. Thus:

**Proposition 2** *All the point symmetries admitted by systems (18) and (19) project to point symmetries of the generalized Gardner equation (1).*

## 5 Conclusions

In this paper, a generalized Gardner equation with nonlinear terms of any order involving six parameters $a$, $b$, $c$, $d$, $e$, $f$ and an arbitrary positive exponent $n$ has been considered. By using the multipliers method of Anco and Bluman we have provided a complete classification of low-order conservation laws admitted by Eq. (1) depending on the different values of the parameters. Making use of these conservation laws, Eq. (1) can be written in conserved form. From these conserved forms, related potential systems (7) and related integrated equations may be obtained. We focus our attention on the case $e \neq 0$ which yields conservation laws (5) and (6). We have determined and analyzed the associated first-level and second-level potential systems for this case. Moreover, we have shown that the point symmetries admitted by the potential systems (8)–(9) and (18)–(19) project to classical symmetries of Eq. (1). However, potential equation (10) yields two new nonlocal conserved vectors of the generalized Gardner equation (1).

**Acknowledgements** The authors gratefully acknowledge the financial support from the Universidad Politécnica de Cartagena. The authors also acknowledge the financial support from Junta de Andalucía group FQM-201 and they express their sincere gratitude to the Plan Propio de Investigación de la Universidad de Cádiz.

## References

1. Anco, S.C.: On the incompleteness of Ibragimov's conservation law theorem and its equivalence to a standard formula using symmetries and adjoint-symmetries. Symmetry **9**, 33 (28 pp.) (2017)
2. Anco, S.C.: Generalization of Noether's theorem in modern form to non-variational partial differential equations. In: Fields Institute Communications: Recent progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science, vol. 79. Springer, New York (2017)
3. Anco, S.C., Bluman, G.W.: Direct construction of conservation laws from field equations. Phys. Rev. Lett. **78**, 2869–2873 (1997)
4. Anco, S.C., Bluman, G.W.: Direct construction method for conservation laws of partial differential equations Part I: examples of conservation law classifications. Eur. J. Appl. Math. **13**, 545–566 (2002)
5. Anco, S.C., Bluman, G.W.: Direct construction method for conservation laws of partial differential equations Part II: general treatment. Eur. J. Appl. Math. **13**, 567–585 (2002)
6. Bluman, G.W., Kumei, S.: Symmetries and Differential Equations. Springer, New York (1989)
7. Bluman, G.W., Reid, G.J., Kumei, S.: New classes of symmetries for partial differential equations. J. Math. Phys. **29**, 806–811 (1988)
8. Bluman, G.W., Cheviakov, A., Anco, S.C.: Applications of Symmetry Methods to Partial Differential Equations. Springer, New York (2009)
9. Brauer, K.: The Korteweg-de Vries equation: history, exact solutions, and graphical representation. University of Osnabrück, Osnabrück (2000). http://math.arizona.edu/~gabitov/teaching/141/math_485/KDV.pdf
10. de la Rosa, R., Bruzón, M.S.: On the classical and nonclassical symmetries of a generalized Gardner equation. Appl. Math. Nonlinear Sci. **1**, 263–272 (2016)
11. de la Rosa, R., Bruzón, M.S.: Travelling wave solutions of a generalized variable-coefficient Gardner equation. In: Trends in Differential Equations and Applications. SEMA SIMAI Springer Series, pp. 405–417. Springer, Cham (2016)
12. de la Rosa, R., Bruzón, M.S.: On conservation laws of a generalized Gardner equation. In: Proceedings of XXV CEDYA/XV CMA, pp. 244–248 (2017)
13. de la Rosa, R., Gandarias, M.L., Bruzón, M.S.: Equivalence transformations and conservation laws for a generalized variable-coefficient Gardner equation. Commun. Nonlinear Sci. Numer. Simul. **40**, 71–79 (2016)
14. Dimas, S., Freire, I.L.: Study of a fifth order PDE using symmetries. Appl. Math. Lett. **69**, 121–125 (2017)
15. Gandarias, M.L., Khalique, C.M.: Symmetries, solutions and conservation laws of a class of nonlinear dispersive wave equations. Commun. Nonlinear Sci. Numer. Simul. **32**, 114–121 (2016)
16. Gandarias, M.L., de la Rosa, R., Rosa, M.: Conservation laws for a strongly damped wave equation. Open Phys. **15**, 300–305 (2017)
17. Garrido, T.M., Kasatkin, A.A., Bruzón, M.S., Gazizov, R.K.: Lie symmetries and equivalence transformations for the Barenblatt-Gilman model. J. Comput. Appl. Math. **318**, 253–258 (2017)
18. Molati, M., Ramollo, M.P.: Symmetry classification of the Gardner equation with time-dependent coefficients arising in stratified fluids. Commun. Nonlinear Sci. Numer. Simul. **17**, 1542–1548 (2012)
19. Olver, P.: Applications of Lie Groups to Differential Equations. Springer, New York (1993)
20. Ovsyannikov, L.V.: Group Analysis of Differential Equations. Academic, New York (1982)
21. Recio, E., Anco, S.C.: Conservation laws and symmetries of radial generalized nonlinear p-Laplacian evolution equations. J. Math. Anal. Appl. **452**, 1229–1261 (2017)
22. Rosa, M., Gandarias, M.L.: Multiplier method and exact solutions for a density dependent reaction-diffusion equation. Appl. Math. Nonlinear Sci. **1**(2), 311–320 (2016)

23. Torrisi, M., Tracinà, R.: Exact solutions of a reaction-diffusion system for Proteus mirabilis bacterial colonies. Nonlinear Anal. Real World Appl. **12**, 1865–1874 (2011)
24. Tracinà, R., Gandarias, M.L., Bruzón, M.S., Torrisi, M.: Nonlinear self-adjointness, conservation laws, exact solutions of a system of dispersive evolution equations. Commun. Nonlinear Sci. Numer. Simul. **19**, 3036–3043 (2014)
25. Vaneeva, O., Kuriksha, O., Sophocleous, C.: Enhanced group classification of Gardner equations with time-dependent coefficients. Commun. Nonlinear Sci. Numer. Simul. **22**, 1243–1251 (2015)

# On a Nonlocal Boussinesq System for Internal Wave Propagation

**Angel Durán**

**Abstract** In this paper we are concerned with a nonlocal system to model the propagation of internal waves in a two-layer interface problem with rigid lid assumption and under a Boussinesq regime for both fluids. The main goal is to investigate aspects of well-posedness of the Cauchy problem for the deviation of the interface and the velocity, as well as the existence of solitary wave solutions and some of their properties.

**Keywords** Internal wave propagation · Boussinesq systems · Solitary waves · Numerical approximation

## 1 Introduction

In this work a one-dimensional, nonlocal differential system for internal waves is considered. The system is derived in [5] and describes the propagation of internal waves in a two-layer interface problem with rigid lid assumption and under the Boussinesq regime for both fluids. The idealized model is sketched in Fig. 1.

This consists of two inviscid, homogeneous, incompressible fluids of depths $d_j, j = 1, 2$ and densities $\rho_j, j = 1, 2$ with $\rho_2 > \rho_1$. The system is bounded above and below by a rigid horizontal plane, with the origin of the vertical coordinate $z$ at the top. The deviation of the interface, denoted by $\zeta$, is assumed to be a graph over the bottom (described by the variable $x$) and surface tension effects are not considered. The approach in [5] is based on the reformulation of the Euler system with two nonlocal operators that link the velocity potentials associated to the layers at the interface. Then, by using suitable asymptotic expansions of these operators, several asymptotic models, consistent with the Euler system, are derived. They are associated to different physical regimes for the layers. The one considered

A. Durán (✉)
Departamento de Matemática Aplicada, Universidad de Valladolid, Valladolid, Spain
e-mail: angel@mac.uva.es

© Springer Nature Switzerland AG 2019

J. L. García Guirao et al. (eds.), *Recent Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 18,
https://doi.org/10.1007/978-3-030-00341-8_8

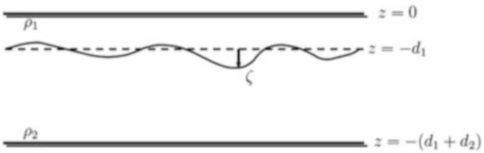


**Fig. 1** Idealized model of internal wave propagation in a two-layer interface

here is the so-called Boussinesq-Boussinesq (B/B) regime, for which the interfacial deformations are assumed to be of small amplitude for both the upper and lower fluid domains and, additionally, the flow has a Boussinesq structure with respect to the two layers, with the nonlinear and dispersive effects of the same size for both fluids [3–5].

One of the differential systems to model the B/B regime in 1D is given by

$$\begin{aligned}\zeta_t + \frac{1}{\gamma+\delta}\partial_x v_\beta + \left(\frac{\delta^2-\gamma}{(\delta+\gamma)^2}\right)\partial_x\left(\zeta v_\beta\right) &= 0,\\ (1-\beta\partial_{xx})(v_\beta)_t + (1-\gamma)\partial_x\zeta + \left(\frac{\delta^2-\gamma}{2(\delta+\gamma)^2}\right)\partial_x v_\beta^2 &= 0,\end{aligned} \tag{1}$$

where $\gamma = \rho_1/\rho_2 < 1, \delta = d_1/d_2$ are, respectively, the density and depth ratios, and $\beta = \dfrac{(1+\gamma\delta)}{3\delta(\gamma+\delta)}$. The variables $x$ and $t$ are proportional to distance along the channel and time respectively and $u(x,t)$ is the velocity variable with $v_\beta = (1-\beta\partial_{xx})^{-1}u$. When $\gamma = 0, \delta = 1$, (1) reduces to the classical Boussinesq system for surface water waves, see e.g. [3, 4, 6, 11, 17]. The purpose of this work is to deal with several properties of (1) as approximation of the Euler equations for internal waves:

1. The first point treated here is a review of the derivation of (1) without using the nonlocal operators defined and considered in [5] but with the original variables instead [9].
2. Then, well-posedness of the corresponding Cauchy problem is considered. After the analysis of the linear case (cf. [5]), the strategy used in [3, 4] for the surface wave problem is applied here to obtain a result of local existence and uniqueness of solution in suitable Sobolev spaces.
3. A final point of the paper is devoted to the solitary wave solutions of (1), that is, wave profiles traveling with permanent form and constant speed, decaying to zero at infinity. The study is divided into two parts. The first one is theoretical and proves the existence of such solutions for a range of speeds depending on the depth and density ratios of the two-fluid system, with the techniques considered in [11] for the case of surface waves. Since the theoretical study does not provide, in general, exact formulas for the waves, the second part is computational. A numerical technique, based on the Petviashvili iteration [14], along with extrapolation [15], is applied to analyze, by numerical means, some properties of the solitary wave profiles, mainly focused on the regularity of the waves, their asymptotic decay and the speed-amplitude relation.

All these results are displayed in the paper according to the following structure. In Sect. 2, an alternative to [5] for the derivation of (1) directly from the Euler equations, is provided. Section 3 is concerned with the analysis of well-posedness of the Cauchy problem. In Sect. 4 the existence result for solitary wave solutions is derived, while in Sect. 5 the computational study of the waves is carried out. We summarize the conclusions in Sect. 6.

The following notation will be used throughout the paper. From the $L^2$ based Sobolev space $H^s = H^s(\mathbb{R}), s \geq 0$ (with $H^0 = L^2(\mathbb{R})$, the space of squared integrable functions on $\mathbb{R}$) we consider the product $X_{s_1,s_2} = H^{s_1} \times H^{s_2}, s_1, s_2 \geq 0$, with associated norm

$$||u||_{s_1,s_2} = \left(||u_1||_{s_1}^2 + ||u_2||_{s_2}^2\right)^{1/2}, \quad u = (u_1, u_2) \in X_{s_1,s_2},$$

where $||\cdot||_s$ stands for the norm in $H^s$ and for simplicity we write $X_s := X_{s,s}, s \geq 0$. For $T > 0$ and $s \geq 0$, the space of continuous functions $v : (0, T] \to H^s$ is denoted by $C(0, T, H^s)$ and its norm by

$$||v||_{C(0,T,H^s)} = \max_{0<t\leq T} ||v(t)||_s.$$

For $s_1, s_2 \geq 0$, the corresponding product space, $X_T^{s_1,s_2} := C(0, T, H^{s_1}) \times C(0, T, H^{s_2})$, is provided by the norm

$$||(v, w)||_{X_T^{s_1,s_2}} := \left(||v||_{C(0,T,H^{s_1})}^2 + ||w||_{C(0,T,H^{s_2})}^2\right)^{1/2},$$

and where $X_T^s := X_T^{s,s}, s \geq 0$. We will additionally make use of the Fourier transform

$$\widehat{f}(k) = \int_{\mathbb{R}} f(x)e^{-ikx}dx, \quad k \in \mathbb{R}, \quad f \in H^0.$$

## 2 Derivation of the Model

In this section we shall derive the system (1) without using the approach based on nonlocal operators of [5].

### 2.1 Euler System for Internal Waves

Assuming that each flow is irrotational, let $\Phi_i, i = 1, 2$ be the velocity potential associated to the upper and lower fluid layer respectively. For the model idealized in Fig. 1, the Euler equations can be written in dimensional, unscaled variables as follows. For $t > 0$, the governing equations are

$$\Phi_{1xx} + \Phi_{1zz} = 0, \quad (x, z) \in \Omega_t^1, \tag{2}$$

$$\Phi_{2xx} + \Phi_{2zz} = 0, \quad (x, z) \in \Omega_t^2, \tag{3}$$

where

$$\Omega_t^1 = \{(x, z)/ - \infty < x < \infty, -d_1 + \zeta(x, t) < z < 0\},$$
$$\Omega_t^2 = \{(x, z)/ - \infty < x < \infty, -d_1 - d_2 < z < -d_1 + \zeta(x, t)\}.$$

The rigid lid assumptions mean that the normal component of both velocity potentials vanishes at the corresponding boundary; that is, for $t > 0$,

$$\Phi_{1z} = 0, \quad \text{at} \quad \Gamma_1 = \{(x, z)/ - \infty < x < \infty, -z = 0\}, \tag{4}$$

$$\Phi_{2z} = 0, \quad \text{at} \quad \Gamma_2 = \{(x, z)/ - \infty < x < \infty, z = -(d_1 + d_2)\}. \tag{5}$$

The kinematic conditions at the interface between the fluids are

$$\zeta_t + \Phi_{ix}\zeta_x = \Phi_{iz} \quad \text{at} \quad \Gamma_t = \{(x, z)/ - \infty < x < \infty, -z = -d_1 + \zeta(x, t)\}, \tag{6}$$

for $t > 0$ and $i = 1, 2$. Finally, the condition of continuity of the pressure can be written as

$$\rho_1\left(\Phi_{1t} + \frac{1}{2}(\Phi_{1x}^2 + \Phi_{1z}^2) + g\zeta\right) = \rho_2\left(\Phi_{2t} + \frac{1}{2}(\Phi_{2x}^2 + \Phi_{2z}^2) + g\zeta\right), \tag{7}$$

at $\Gamma_t$, where $g$ is the acceleration of gravity. From (6) and (7), two additional equations will be used throughout the section. The first one is obtained by subtracting the equations of (6) and eliminating $\zeta_t$:

$$\Phi_{1x}\zeta_x - \Phi_{1z} = \Phi_{2x}\zeta_x - \Phi_{2z} \quad \text{at} \quad \Gamma_t, \tag{8}$$

while (7) can also be written as

$$\Phi_{2t} - \gamma\Phi_{1t} + g(1 - \gamma)\zeta + \frac{1}{2}(\Phi_{2x}^2 - \gamma\Phi_{1x}^2 + \Phi_{2z}^2 - \gamma\Phi_{1z}^2) = 0 \quad \text{at} \quad \Gamma_t. \tag{9}$$

## 2.2 Non-dimensionalization

In [5], the system (2)–(7) is rewritten in terms of the potentials at the interface and two nonlocal operators which relate them. Instead, we directly nondimensionalize (2)–(7) by using the dimensionless parameters

$$\epsilon = \frac{a}{d_1}, \quad \mu = \left(\frac{d_1}{\lambda}\right)^2,$$

where $a$ is a typical amplitude and $\lambda$ a typical wavelength of the waves, as well as the dimensionless variables and unknowns

$$x = \lambda \widetilde{x}, \quad z = d_z \widetilde{z}, \quad t = \frac{\lambda}{\sqrt{g d_1}} \widetilde{t}, \quad \zeta = a \widetilde{\zeta}, \quad \Phi_i = a\lambda \sqrt{\frac{g}{d_1}} \widetilde{\Phi_i}.$$

Then the regions and boundaries are transformed into

$$\begin{aligned}
\Omega_t^1 &= \{(x,z)/-\infty < x < \infty, -1+\epsilon\zeta(x,t) < z < 0\},\\
\Omega_t^2 &= \{(x,z)/-\infty < x < \infty, -d_1-d_2 < z < -d_1+\zeta(x,t)\},\\
\Gamma_1 &= \{(x,z)/-\infty < x < \infty, z=0\}, \quad \Gamma_2 = \{(x,z)/-\infty < x < \infty, z = -1-\frac{1}{\delta}\},\\
\Gamma_t &= \{(x,z)/-\infty < x < \infty, z = -1+\epsilon\zeta(x,t)\},
\end{aligned}$$

while (2)–(7) are written as

$$mu\Phi_{1xx} + \Phi_{1zz} = 0, \quad (x,z) \in \Omega_t^1, \tag{10}$$

$$\mu\Phi_{2xx} + \Phi_{2zz} = 0, \quad (x,z) \in \Omega_t^2, \tag{11}$$

$$\Phi_{1z} = 0, \quad \text{at} \quad \Gamma_1, \tag{12}$$

$$\Phi_{2z} = 0, \quad \text{at} \quad \Gamma_2, \tag{13}$$

$$\zeta_t + \epsilon\Phi_{ix}\zeta_x = \frac{1}{\mu}\Phi_{iz} \quad \text{at} \quad \Gamma_t, \tag{14}$$

and (9) reads

$$\Phi_{2t} - \gamma\Phi_{1t} + (1-\gamma)\zeta + \frac{\epsilon}{2}(\Phi_{2x}^2 - \gamma\Phi_{1x}^2) + \frac{\epsilon}{2\mu}(\Phi_{2z}^2 - \gamma\Phi_{1z}^2) = 0, \tag{15}$$

at $\Gamma_t, t > 0$. (Tildes were dropped.)

## 2.3 *Boussinesq/Boussinesq Regime*

We assume $\delta \sim 1$ and that the deviation of the interface is long and of small amplitude for both fluids, which means that $\epsilon, \mu << 1$ as well as $\epsilon_2, \mu_2 << 1$, where

$$\epsilon_2 = \frac{a}{d_2} = \epsilon\delta, \quad \mu_2 = \left(\frac{d_2}{\lambda}\right)^2 = \frac{\mu}{\delta^2},$$

are the amplitude and wavelength parameters with respect to the lower fluid (and which are not independent of $\epsilon$, $\mu$). Furthermore, we are interested in the so-called Boussinesq/Boussinesq regime [5]. This means that the nonlinear and dispersive effects are of the same size for both fluids and thus

$$\epsilon \sim \mu \sim \epsilon_2 \sim \mu_2. \tag{16}$$

From this point, the derivation follows the strategy considered in [3] for surface waves, see also [2, 9, 13, 17]. The potentials $\Phi_i$, $i = 1, 2$, are formally expanded in the corresponding domains as

$$\Phi_1(x, z, t) = \sum_{m=0}^{\infty} f_{1m}(x, t) z^m, \quad (x, z) \in \Omega_t^1, \tag{17}$$

$$\Phi_2(x, z, t) = \sum_{m=0}^{\infty} f_{2m}(x, t)(z + h_\delta)^m, \quad (x, z) \in \Omega_t^2, \tag{18}$$

where $h_\delta = 1 + \dfrac{1}{\delta}$. Now, satisfaction of (10) and (11) implies

$$(m + 2)(m + 1) f_{i,m+2}(x, t) = -\mu(f_{im}(x, t))_{xx}, \quad i = 1, 2, \tag{19}$$

while, due to the boundary conditions (12) and (13), we have $f_{i,1}(x, t) = 0, i = 1, 2$, and therefore, according to (19), $f_{i,2k+1}(x, t) = 0, k \geq 0, i = 1, 2$. For the even terms, the application of (19) leads to

$$f_{i,2k}(x, t) = \frac{(-1)^k}{(2k)!} \mu^k \frac{\partial^{2k}}{\partial x^{2k}} F_i(x, t), \quad k \geq 0, i = 1, 2,$$

where $F_i(x, t) = f_{i,0}(x, t), i = 1, 2$. Therefore, (17), (18) can be written as

$$\Phi_1(x, z, t) = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} \mu^k \frac{\partial^{2k}}{\partial x^{2k}} F_1(x, t) z^m, \quad (x, z) \in \Omega_t^1, \tag{20}$$

$$\Phi_2(x, z, t) = \sum_{m=0}^{\infty} \frac{(-1)^k}{(2k)!} \mu^k \frac{\partial^{2k}}{\partial x^{2k}} F_2(x, t)(z + h_\delta)^m, \quad (x, z) \in \Omega_t^2, \tag{21}$$

In what follows, the expansions (20) and (21) will be evaluated at the interface $z = -1 + \epsilon\zeta$ and, according to (16), the linear terms in $\epsilon$, $\mu$ will be retained. Note first that (8), in nondimensional form, reads

$$\epsilon \Phi_{1x} \zeta_x - \frac{1}{\mu} \Phi_{1z} = \epsilon \Phi_{2x} \zeta_x - \frac{1}{\mu} \Phi_{2z} \tag{22}$$

at $\Gamma_t = \{(x,z)/-\infty < x < \infty, z = -1 + \epsilon\zeta(x,t)\}$, which, in terms of the expansions (20) and (21), leads to

$$\frac{1}{\delta}F_{2x} + F_{1x} = \epsilon\zeta(F_{1x} - F_{2x}) + \frac{\mu}{6}(F_{1xxx} + \frac{1}{\delta^3}F_{2xxx}) + O(\epsilon^2, \epsilon\mu, \mu^2). \quad (23)$$

On the other hand, we define $\Phi = \Phi_2 - \gamma\Phi_1$ and let $u = \Phi_x$. According to (20) and (21), we have

$$u = F_{2x} - \gamma F_{1x} - \frac{\mu}{2}\left(\frac{1}{\delta^2}F_{2xxx} - \gamma F_{1xxx}\right) + O(\epsilon^2, \epsilon\mu, \mu^2). \quad (24)$$

Formulas (23) and (24) are now used to determine iteratively $F_1$ and $F_2$ in terms of $\zeta$ and $u$ (cf. [9]). After some computations, we have

$$F_{1x} = \frac{1}{\delta+\gamma}\left(-u - \epsilon\frac{\delta(1+\delta)}{\delta+\gamma}\zeta u - \mu\frac{2+3\gamma\delta+\delta^2}{6\delta(\delta+\gamma)}u_{xx}\right) + O(\epsilon^2, \epsilon\mu, \mu^2). \quad (25)$$

$$F_{2x} = \frac{1}{\delta+\gamma}\left(\delta u - \epsilon\frac{\delta\gamma(1+\delta)}{\delta+\gamma}\zeta u + \mu\frac{1+3\delta+\delta^2(3\gamma-1)}{6\delta(\delta+\gamma)}u_{xx}\right) + O(\epsilon^2, \epsilon\mu, \mu^2). \quad (26)$$

Finally, the application of (25) and (26) to the first kinematic condition in (14) and, after differentiating in $x$, to (15), implies

$$\zeta_t + \frac{1}{\delta+\gamma}\partial_x(1+\beta\mu\partial_{xx})u + \epsilon\left(\frac{\delta^2-\gamma}{(\delta+\gamma)^2}\right)\partial_x(\zeta u) = O(\epsilon^2, \epsilon\mu, \mu^2), \quad (27)$$

$$u_t + (1-\gamma)\zeta_x + \frac{\epsilon}{2}\left(\frac{\delta^2-\gamma}{(\delta+\gamma)^2}\right)\partial_x(u^2) = O(\epsilon^2, \epsilon\mu, \mu^2), \quad (28)$$

where $\beta$ is defined in (1). Note that if $v_\beta = (1-\beta\mu\partial_{xx})^{-1}u$, then $v_\beta = (1 + \beta\mu\partial_{xx})u + O(\mu^2)$ and substitution into (27) and (28) leads to (1), assuming (16) and when the $O(\epsilon^2, \epsilon\mu, \mu^2)$ terms are dropped.

*Remark 1* Similarly, from system (27), (28), the three parameter family of Boussinesq systems for internal waves presented in [5] can alternatively be derived by adapting the procedure implemented in [3] for the case of surface waves.

# 3 Well-posedness

## *3.1 Linear well-posedness*

We now study the linear well-posedness of (1). The associated linear system, written in terms of $\zeta$ and $u$ is of the form

$$\begin{aligned}&\zeta_t + \frac{1}{\gamma+\delta}\partial_x(I-\beta\partial_{xx})^{-1}u_x = 0,\\ &u_t + (1-\gamma)\partial_x\zeta = 0.\end{aligned} \tag{29}$$

The application of the Fourier transform leads to the system

$$\frac{d}{dt}\begin{pmatrix}\widehat{\zeta}(k,t)\\ \widehat{u}(k,t)\end{pmatrix} + (ik)A(k)\begin{pmatrix}\widehat{\zeta}(k,t)\\ \widehat{u}(k,t)\end{pmatrix} = 0, \quad k\in\mathbb{R},$$

where

$$A(k) = \begin{pmatrix}0 & \omega(k)\\ 1-\gamma & 0\end{pmatrix}, \quad \omega(k) = \frac{1}{\delta+\gamma}\frac{1}{1+\beta k^2}, \tag{30}$$

Then well-posedness is determined by the behaviour of the matrix

$$e^{-ikA(k)t} = \begin{pmatrix}\cos(k\sigma(k)t) & -i\sqrt{\frac{\omega(k)}{1-\gamma}}\sin(k\sigma(k)t)\\ i\sqrt{\frac{1-\gamma}{\omega(k)}}\sin(k\sigma(k)t) & \cos(k\sigma(k)t)\end{pmatrix}, \tag{31}$$

where $\sigma(k) = \sqrt{(1-\gamma)\omega(k)}$. Specifically, the linearized problem (29) is well-posed when (31) is bounded in finite intervals of $t$. We define the order of $\sigma(k)$ as the integer $l$ such that

$$\sigma(k) \approx |k|^l, \quad |k|\to\infty,$$

and let $m_1 = \max\{0,-l\}$, $m_2 = \max\{0,l\}$. Since $\omega(k)$ has neither poles or zeros on the real axis then we have [3]:

**Theorem 1** *Let* $\beta > 0$. *System (1) is linearly well-posed for* $(\zeta, u) \in X_{s+1,s}$ *and therefore for* $(\zeta, v_\beta) \in X_{s+1,s+2}$, $s \geq 0$.

*Remark 2* Theorem 3.1 completes the linear well-posedness result established in [5] by specifying the spaces where that holds.

## 3.2 Local Well-posedness of the Full Nonlinear System

Taking the Fourier transform in $x$ to (1) for $\zeta$ and $u$ we have

$$\frac{d}{dt}\begin{pmatrix}\widehat{\zeta}\\ \widehat{u}\end{pmatrix}+ikA(k)\begin{pmatrix}\widehat{\zeta}\\ \widehat{u}\end{pmatrix}+ikF\begin{pmatrix}\widehat{\zeta}\\ \widehat{u}\end{pmatrix}(k)=0, \tag{32}$$

where $A$ is given by (30) and

$$F=K_{\gamma,\delta}\begin{pmatrix}\widehat{\left[\zeta(1-\beta\partial_{xx})^{-1}u\right]}(k)\\ \dfrac{\widehat{\left(\left[(1-\beta\partial_{xx})^{-1}u\right]^2\right)}}{2}(k)\end{pmatrix},\quad K_{\gamma,\delta}=\frac{\delta^2-\gamma}{(\delta+\gamma)^2}.$$

In order to study well-posedness of the system (1), we use a similar strategy to that of [4] by decoupling the linear part with the operator $\Sigma$ such that

$$\widehat{\Sigma f}(k)=\sqrt{\frac{\omega(k)}{1-\gamma}}\widehat{f}(k),\quad k\in\mathbb{R},$$

where $\omega$ is given in (30). Defining new variables $v$, $w$ such that

$$\zeta=\Sigma(v+w),\quad u=v-w,$$

then the system (32) is of the form

$$\frac{d}{dt}\begin{pmatrix}\widehat{v}\\ \widehat{w}\end{pmatrix}+ik\begin{pmatrix}\sigma(k) & 0\\ 0 & -\sigma(k)\end{pmatrix}\begin{pmatrix}\widehat{v}\\ \widehat{w}\end{pmatrix}+ikP^{-1}F=0, \tag{33}$$

with $\sigma(k)$ given in (31) and

$$P^{-1}=\frac{1}{2}\begin{pmatrix}\sqrt{\frac{1-\gamma}{\omega(k)}} & 1\\ \sqrt{\frac{1-\gamma}{\omega(k)}} & -1\end{pmatrix}. \tag{34}$$

Then the following local well-posedness result holds:

**Theorem 2** *Let* $(\zeta_0, v_{\beta,0})\in X_{s+1,s+2}$, $s\geq 0$. *Then there exists* $T>0$ *and a unique solution* $(\zeta,u)$ *in* $X_T^{s+1,s}$ *of (1) with initial condition* $(\zeta_0,u_0)$ *with* $u_0=(1-\beta\partial_{xx})v_{\beta,0}$.

*Proof* We consider the variables

$$v_0=\frac{\Sigma^{-1}(\zeta_0)+u_0}{2},\quad w_0=\frac{\Sigma^{-1}(\zeta_0)-u_0}{2}.$$

Then $(v_0, w_0) \in X_s$ and taking the inverse Fourier transform in (33) we have

$$\frac{d}{dt}\begin{pmatrix} v \\ w \end{pmatrix} + \mathcal{B}\begin{pmatrix} v \\ w \end{pmatrix} = \mathcal{F}\begin{pmatrix} v \\ w \end{pmatrix}, \tag{35}$$

where $\mathcal{B}$ is the operator with symbol

$$ik\begin{pmatrix} \sigma(k) & 0 \\ 0 & -\sigma(k) \end{pmatrix},$$

and

$$\mathcal{F}\begin{pmatrix} v \\ w \end{pmatrix} = -\mathcal{P}^{-1}K_{\gamma,\delta}\begin{pmatrix} [(1-\beta\partial_{xx})^{-1}(v-w)]\Sigma(v+w) \\ \frac{\left([(1-\beta\partial_{xx})^{-1}(v-w)]^2\right)}{2} \end{pmatrix},$$

with $\mathcal{P}$ the operator associated to $ikP(k)$, obtained from (34), as symbol. By Duhamel formula, the solution of (35) satisfies

$$\begin{pmatrix} v \\ w \end{pmatrix} = \mathcal{S}(t)\begin{pmatrix} v_0 \\ w_0 \end{pmatrix} + \int_0^t \mathcal{S}(t-\tau)\mathcal{F}\begin{pmatrix} v \\ w \end{pmatrix} d\tau,$$

where $\mathcal{S}(t)$ is the group generated by $\mathcal{B}$. Consider then the mapping $(\widetilde{v}, \widetilde{w}) \mapsto (v, w)$ where

$$\begin{pmatrix} v \\ w \end{pmatrix} = \mathcal{S}(t)\begin{pmatrix} v_0 \\ w_0 \end{pmatrix} + \int_0^t \mathcal{S}(t-\tau)\mathcal{F}\begin{pmatrix} \widetilde{v} \\ \widetilde{w} \end{pmatrix} d\tau. \tag{36}$$

Note first that $\mathcal{S}(t)$ is a unitary group on $X_s$. On the other hand, the operator $\mathcal{F}$ can be estimated by using Lemma 2.2 of [4], in such a way that if $T > 0$ and if $(\widetilde{v}_1, \widetilde{w}_1)$, $(\widetilde{v}_2, \widetilde{w}_2)$ are in a closed ball of radius $R$ centered at 0 in $X_T^s$ then there is $C(R) > 0$ for which

$$\|\mathcal{F}(\widetilde{v}_1, \widetilde{w}_1)(\tau) - \mathcal{F}(\widetilde{v}_2, \widetilde{w}_2)(\tau)\|_{X_s} \leq C(R)\left\|\begin{pmatrix} \widetilde{v}_1 \\ \widetilde{w}_1 \end{pmatrix}(\tau) - \begin{pmatrix} \widetilde{v}_2 \\ \widetilde{w}_2 \end{pmatrix}(\tau)\right\|_{X_s}, \tag{37}$$

for any $0 \leq \tau \leq T$. Now, (37) and the property $\mathcal{F}(0, 0) = (0, 0)$ imply that there is $T > 0$, sufficiently small, such that (36) is a contraction of the closed ball into itself. Then the result follows by using the Contraction Mapping Theorem. □

# 4 Solitary Wave Solutions

In this section, the existence of solitary wave solutions of (1) is studied. They are solutions of traveling-wave form $\zeta = \zeta(x - c_s t), u = u(x - c_s t)$ (or $v_\beta = v_\beta(x - c_s t)$) for some speed $c_s \neq 0$ and where the profiles $\zeta = \zeta(X), u = u(X), X = x - c_s t$ (or $v_\beta = v_\beta(X)$) with $\zeta, u \to 0$ as $|X| \to \infty$ must satisfy

$$\begin{pmatrix} c_s & \frac{-1}{\delta+\gamma} \\ \gamma - 1 & c_s(1 - \beta\partial_{xx}) \end{pmatrix} \begin{pmatrix} \zeta \\ v_\beta \end{pmatrix} = K_{\gamma,\delta} \begin{pmatrix} \zeta v_\beta \\ \frac{v_\beta^2}{2} \end{pmatrix}. \tag{38}$$

Solving the first equation of (38) for $\zeta$ and substituting into the second one lead to

$$v_\beta'' - \frac{1}{\beta} v_\beta + G'(v_\beta) = 0, \tag{39}$$

where $G(v) = \int_0^v \frac{1}{\beta c_s} g(z) dz$, $c_{\gamma,\delta} = \sqrt{\frac{1-\gamma}{\delta+\gamma}}$ and

$$g(v) = \frac{\delta^2 - \gamma}{2(\delta+\gamma)^2} v^2 + \frac{c_{\gamma,\delta}^2 v}{c_s - \frac{\delta^2-\gamma}{(\delta+\gamma)^2} v}.$$

## *4.1 Existence of Solitary Waves*

The analysis of (39) leads to the following result (cf. [11]).

**Theorem 3** *Let* $0 < \gamma < 1$, $\delta^2 - \gamma \neq 0$ *and* $c_s^2 - c_{\gamma,\delta}^2 > 0$. *Then (38) has a unique solution* $(\zeta_s(X), v_{\beta,s}(X))$ *which is even and goes to zero as* $|X| \to \infty$. *The profiles* $\zeta, v_\beta$ *are of elevation when* $\delta^2 - \gamma > 0$ *and of depression when* $\delta^2 - \gamma < 0$.

*Proof* We assume $c_s > 0$ (the arguments are similar in the case $c_s < 0$). Equation (39) is conservative with the energy given by

$$E = \frac{1}{2}(v_\beta')^2 + U(v_\beta),$$

where $U(x) = -\frac{x^2}{2\beta} + G(x)$ is the potential energy. Since $U(0) = U'(0) = 0$ and $U''(0) = -\frac{c_s^2 - c_{\gamma,\delta}^2}{\beta c_s^2} < 0$ then the phase plane analysis shows that the origin is a

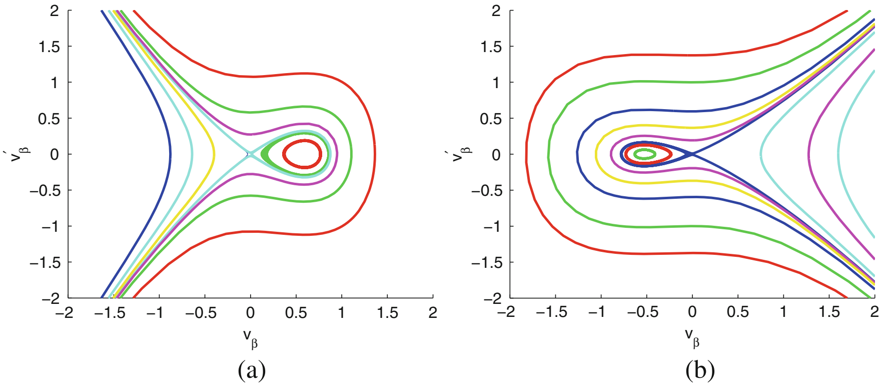


**Fig. 2** Phase portraits for (39) with (**a**) $K_{\gamma,\delta} > 0$ and (**b**) $K_{\gamma,\delta} < 0$

saddle point (see Fig. 2). Note that

$$G(v) = \frac{K_{\gamma,\delta} v^3}{6\beta c_s} + \left(\frac{c_s}{c_{\gamma,\delta}}\right)^2 \frac{c_s}{\beta K_{\gamma,\delta}} \left(-v + \frac{c_s}{K_{\gamma,\delta}} \ln\left(\frac{c_s/K_{\gamma,\delta}}{\frac{c_s}{K_{\gamma,\delta}} - v}\right)\right).$$

The sign of $K_{\gamma,\delta}$ determines the behaviour of $G$ in the following sense (cf. [11]):

(i) If $K_{\gamma,\delta} > 0$, then $G$ increases smoothly from 0 to $\infty$ as $0 < v < \dfrac{c_s}{K_{\gamma,\delta}}$.

(ii) If $K_{\gamma,\delta} < 0$, then $G$ decreases smoothly from $\infty$ to 0 as $\dfrac{c_s}{K_{\gamma,\delta}} < v < 0$.

In both cases, the function $G$ is always positive and therefore we can find $v^* = v^*(\gamma, \delta)$ such that

$$U(v^*) = G(v^*) - \frac{(v^*)^2}{2\beta} = 0,$$

with $0 < v^* < \dfrac{c_s}{K_{\gamma,\delta}}$ in the case (i) and $\dfrac{c_s}{K_{\gamma,\delta}} < v^* < 0$ in the case (ii). Thus, the solution of (39) with $v_\beta(0) = v^*,\ v_\beta'(0) = 0$ has zero energy, is even, positive when $K_{\gamma,\delta} > 0$, negative when $K_{\gamma,\delta} < 0$ and goes to zero as $|X| \to \infty$. Note that from (38) we have

$$\zeta_s = \frac{1}{K_{\gamma,\delta}(\gamma + \delta)} \frac{v_\beta}{\frac{c_s}{K_{\gamma,\delta}} - v_\beta},$$

and from (39)

$$u_s := (1 - \beta\partial_{xx})v_\beta = \beta G'(v_\beta).$$

Therefore, both $\zeta_s, u_s$ are also of elevation when $K_{\gamma,\delta} > 0$ and of depression when $K_{\gamma,\delta} < 0$. □

## 5 Numerical Approximation to Solitary Waves

Since explicit formulas for the solitary waves are, to our knowledge, in general unknown, then the study of the form and additional properties of the profiles will be done in this section by computational means. A numerical procedure used to this end will be first briefly described and then applied to analyze features of the waves, mainly concerned with their regularity, asymptotic decay and the relation between the amplitude and the speed of the profiles.

### 5.1 A Numerical Technique to Compute Solitary Wave Profiles

In order to compute approximate solitary wave profiles of (1), the system (38) is discretized on an interval $(-l, l)$, $l$ large, using Fourier pseudospectral approximation. Let $N \geq 1$ be integer and even, and let $(\zeta_h, v_h)$ be a $2N$-vector approximation to the profile $(\zeta, v_\beta)$ at the uniform grid of collocation points $x_j = -l + jh, h = 2l/N, j = 0, \ldots, N$. Then $(\zeta_h, v_h)$ satisfies the system

$$\mathcal{L}_h \begin{pmatrix} \zeta_h \\ v_h \end{pmatrix} = \mathcal{N}_h(\zeta_h, v_h) \tag{40}$$

$$\mathcal{L}_h = \begin{pmatrix} c_s I_N & -\frac{1}{\delta+\gamma} I_N \\ (\gamma-1) I_N & c_s(1-\beta D_N^2) \end{pmatrix}, \quad \mathcal{N}_h(\zeta_h, v_h) = K_{\gamma,\delta} \begin{pmatrix} \zeta_h . v_h \\ \frac{1}{2} v_h . v_h \end{pmatrix},$$

where $I_N$ stands for the $N \times N$ identity matrix, $D_N$ stands for the $N \times N$ pseudospectral differentiation matrix, [7], and the dot in $\mathcal{N}_h$ stands for the Hadamard product of vectors. System (40) is implemented in Fourier space, in such a way that if $\widehat{\zeta_h}(k), \widehat{v_h}(k), k = 0, \ldots, N-1$ denote the $k$-th discrete Fourier components of $\zeta_h$ and $v_h$ respectively, and $k' = \pi k/l$, then (40) leads to $2 \times 2$ algebraic systems of the form

$$\begin{pmatrix} c_s & -\frac{1}{\delta+\gamma} \\ (\gamma-1) & c_s(1+\beta(k')^2) \end{pmatrix} \begin{pmatrix} \widehat{\zeta_h}(k) \\ \widehat{v_h}(k) \end{pmatrix} = K_{\gamma,\delta} \begin{pmatrix} \widehat{\zeta_h . v_h}(k) \\ \frac{1}{2}\widehat{v_h . v_h}(k) \end{pmatrix}, \tag{41}$$

for $k = 0, \ldots, N-1$. Once (41) is solved for each $k = 0, \ldots, N-1$ the approximation $u_h$ to $u$ at the collocation points is defined as $u_h = (I_N - \beta D_N^2) v_h$ or, in Fourier components, as $\widehat{u_h}(k) = (1 + \beta(k')^2)\widehat{v_h}(k), k = 0, \ldots, N-1$.

The system (40), or its Fourier version (41), is numerically solved by iteration with the Petviashvili method [14]: given initial data $\zeta_h^{[0]}$, $v_h^{[0]}$, the $(\nu+1)$-th iteration $\zeta_h^{[\nu+1]}$, $v_h^{[\nu+1]}$, $\nu = 0, 1, \ldots,$ is the solution of

$$m_\nu = \frac{\langle \mathcal{L}_h \begin{pmatrix} \zeta_h^{[\nu]} \\ v_h^{[\nu]} \end{pmatrix}, \begin{pmatrix} \zeta_h^{[\nu]} \\ v_h^{[\nu]} \end{pmatrix} \rangle}{\langle \mathcal{N}_h(\zeta_h^{[\nu]}, v_h^{[\nu]}), \begin{pmatrix} \zeta_h^{[\nu]} \\ v_h^{[\nu]} \end{pmatrix} \rangle},$$

$$\mathcal{L}_h \begin{pmatrix} \zeta_h^{[\nu+1]} \\ v_h^{[\nu+1]} \end{pmatrix} = m_\nu^2 \mathcal{N}_h(\zeta_h^{[\nu]}, v_h^{[\nu]}), \tag{42}$$

where $\langle \cdot, \cdot \rangle$ stands for the Euclidean inner product in $\mathbb{R}^{2N}$. (See e.g. [12] for a justification of the formulas.) The iteration is supplemented with an extrapolation method [10, 15], a technique which has been revealed useful in the acceleration of the convergence when computing approximate solitary wave profiles [1]. The implementation of the iterative procedure is carried out in Fourier space and the accuracy of the method was checked in the standard way, see e.g. [8] for the details.

## 5.2 *Numerical Results*

The form of the profiles of depression and of elevation is illustrated in Fig. 3, which displays some computed waves for different speeds, close to the limiting value $c_{\gamma,\delta}$.

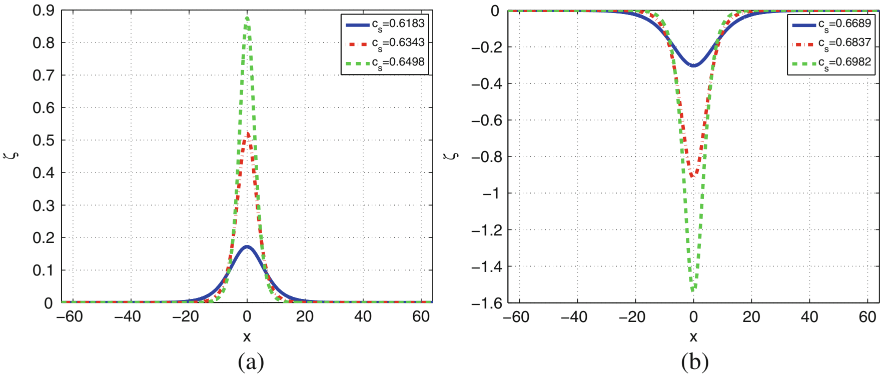
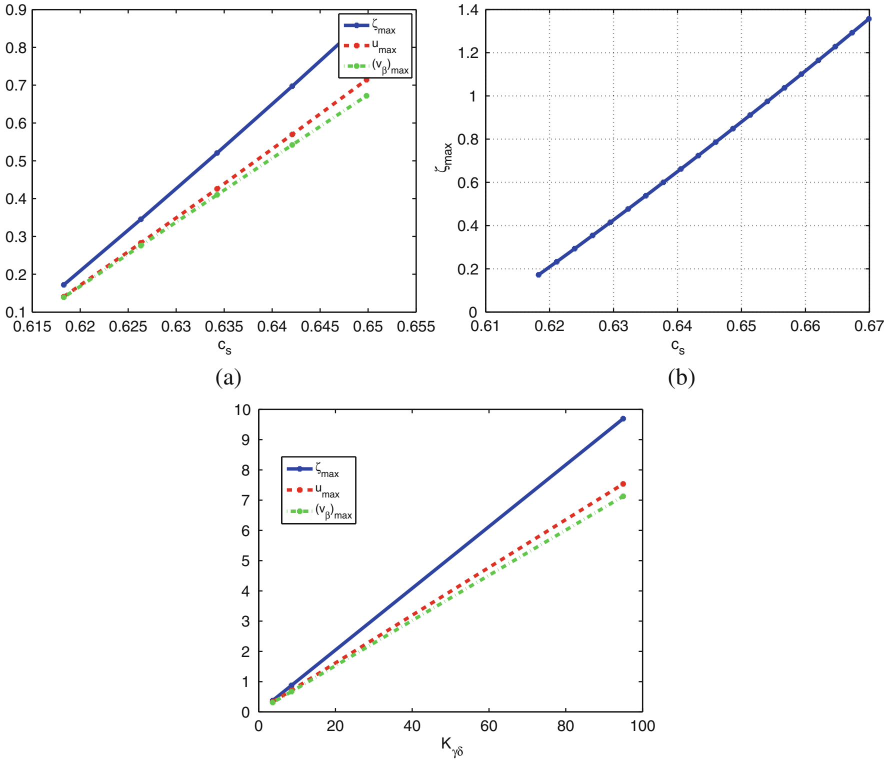


**Fig. 3** Approximate $\zeta$ solitary-wave profiles of (38): (**a**) corresponds to $\gamma = 0.5, \delta = 0.8$ ($K_{\gamma,\delta} > 0$, waves of elevation) while (**b**) corresponds to $\gamma = \delta = 0.5$ ($K_{\gamma,\delta} < 0$, waves of depression)



(a)

(b)

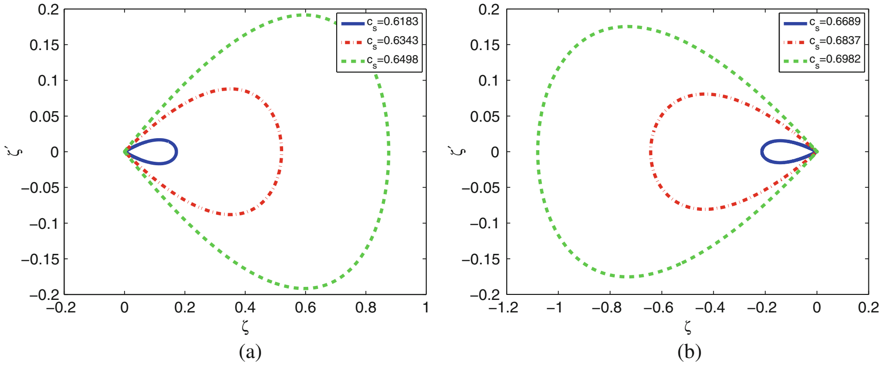
(c)

**Fig. 4** (**a**), (**b**) Amplitude vs. speed and (**c**) amplitude vs. $K_{\gamma,\delta}$ for the profiles of Fig. 3a

Note that the amplitude (the magnitude of the maximum negative excursion for the case of waves of depression) increases with the speed.

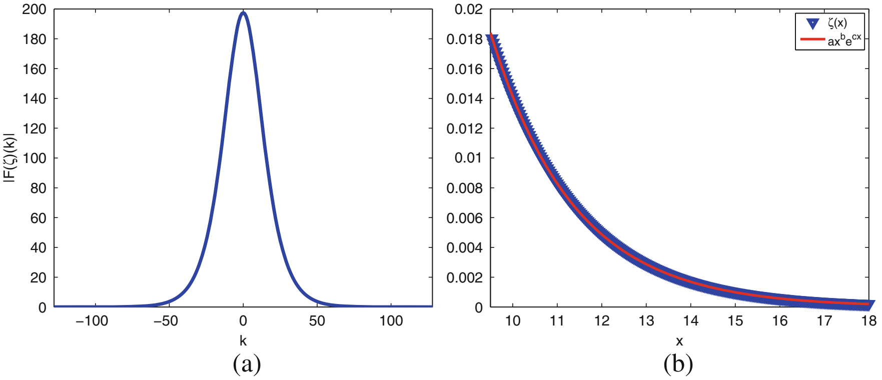
This speed-amplitude relation is confirmed in Fig. 4, which corresponds to the waves of elevation of Fig. 3a. In Fig. 4a we can observe the amplitude as increasing function of the speed in the deviation of the interface and both velocity variables. Figure 4b shows the speed-amplitude relation only for the interfacial deviation $\zeta$ and for a wider range of speeds. The goal is trying to specify a relation $\zeta_{max} \approx Ac_s^B + C$ by fitting the experimental data for the parameters $A$, $B$ and $C$. The resolution of the fitting problem gives $A = 18$, $B = 2.75$ and $C = -4.626$ with, as a measure of the goodness of fit, a sum of squares due to error (SSE) of about $4.146 \times 10^{-9}$, a R-squared value of 1 and a root mean squared error (RMSE) of approximately $2.434 \times 10^{-5}$.

On the other hand, in Fig. 4c, for $\gamma = 0.5$, various $\delta$ and $c_s = c_{\gamma,\delta} + 0.05$, the amplitude is displayed as function of the parameter $K_{\gamma,\delta}$ which determines the type (of elevation or of depression) of the profile. The amplitude also increases with $K_{\gamma,\delta}$



(a) (b)

**Fig. 5** Phase portraits of the approximate profiles shown in Fig. 3



(a) (b)

**Fig. 6** (**a**) Modulus of the Fourier transform $|\widehat{\zeta}(k)|$ of a profile vs. $k$; (**b**) Profile $\zeta(x)$ and fitting curve $ax^b e^{cx}$ vs. $x$

and Fig. 4c shows the dependence of the amplitude of the waves on the depth and density ratios of the two-fluid system (through the parameter $K_{\gamma,\delta}$).

Figure 5 shows the phase portraits of the approximate profiles of Fig. 3, computed by using pseudospectral differentiation [7]. We think that the smooth way how the orbits approach the origin suggests an exponential decay of the waves. This is also supported by the smooth form of the modulus of the Fourier transform of one of the profiles, displayed in Fig. 6a. If part of the wave is fitted to a curve of the form $ax^b e^{cx}$ (Fig. 6b), then the solution of the corresponding least squares problem confirms this behaviour (see Table 1). Actually, similar arguments to those of [11] for the case of surface water waves (by applying the Stable Manifold Theorem) might be used to show that the profiles decay exponentially to zero.

The decay of the Fourier transform can give more information on the regularity of the profiles. This is illustrated in Fig. 7. In linear-log and log-log scales, Fig. 7a shows the modulus of the Fourier transform versus the wavenumber, while in Fig. 7b, part of this modulus is fitted to a curve of the form $ak^b e^{ck}$, [16], and

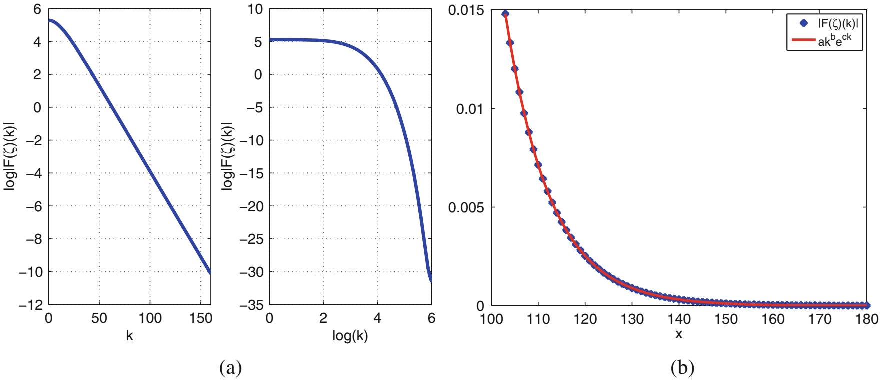


**Fig. 7** (**a**) Modulus of the Fourier transform vs. $k$ in linear-log and log-log scales [16]; (**b**) Modulus of the Fourier transform of a profile and fitting curve $ak^be^{ck}$ vs. $k$

**Table 1** Curves and goodness of fit for data from Figs. 6b (up) and 7b (down)

| *Fitting curve* | *g.o.f.* |
|---|---|
| $f(x) = ax^be^{cx}$ | $SSE = 1.967 \times 10^{-8}$ |
| $a = 2.8176, b = 0.0114$ | $R\text{-}squared = 1$ |
| $c = -0.5323$ | $RMSE = 4.233 \times 10^{-6}$ |
| $f(k) = ak^be^{ck}$ | $SSE = 3.548 \times 10^{-11}$ |
| $a = 1261.4, b = -0.1728$ | $R\text{-}squared = 1$ |
| $c = -0.1023$ | $RMSE = 4.896 \times 10^{-7}$ |

according to the goodness of fit (Table 1), the curve shows a reliable exponential decay, which suggests a smooth character of the solitary wave profile.

# 6 Conclusions

Considered in this paper is a nonlocal system for internal waves. The model was derived in [5] as a consistent approximation to the Euler equations for the wave propagation in a two-layer fluid system, with rigid lid assumptions and under a Boussinesq physical regime for both fluids. In this paper, having in mind the treatment of the case of surface wave propagation in the literature [3, 11], several mathematical aspects of the model are studied in more detail. After an alternative derivation of the differential system (which does not make use of the nonlocal operators considered in the original presentation in [5]), a result of local existence and uniqueness of solution in suitable Sobolev spaces is proved. The study is then focused on the solitary wave solutions. They are shown to exist for a range of the speeds which includes bidirectional propagation and depends on the depth and density ratios of the two-fluid problem. These ratios also determine the character of elevation or of depression of the wave. Finally, a numerical technique to generate approximate solitary wave profiles is introduced and applied to suggest some features: the waves look to be smooth, decay exponentially to zero at infinity and their amplitude looks to behave as an increasing power function of the speed.

The analysis, theoretical and numerical, of the solitary waves performed in the last part of the paper must be taken into account as starting point for a future research, mainly focused on the dynamics of the waves. In this sense, a computational work concerning experiments of stability under small and large perturbations, head-on and overtaking collisions, resolution of initial data into trains of solitary waves, etc., is essential for this purpose. This includes a detailed analysis of the codes designed and used to this end. From the theoretical point of view, the results, presented in the literature (see e.g. [11]) about stability are again a necessary tool to go more deeply into this topic.

**Acknowledgements** This research is supported by Spanish Ministerio de Economía y Competitividad under grant MTM2014-54710-P with the participation of FEDER and by Junta de Castilla y León under grant VA041P17. The author would like to thank Professors V. Dougalis, D. Dutykh and D. Mitsotakis for fruitful discussions and so important suggestions.

## References

1. Alvarez, J., Durán, A.: Petviashvili type methods for traveling wave computations: II. Acceleration with vector extrapolation methods. Math. Comput. Simul. **123**, 19–36 (2016)
2. Benjamin, T.B.: Lectures on nonlinear wave motion. In: Nonlinear Wave Motion (Proc. AMS-SIAM Summer Sem. Clarkson Coll. Tech., Potsdam, N. Y., 1972). Lectures in Applied Mathematics, vol 15, pp. 3–47. American Mathematical Society, Providence (1974)
3. Bona, J.L., Chen, M., Saut, J.-C.: Boussinesq equations and other systems for small-amplitude long waves in nonlinear dispersive media: I. Derivation and linear theory. J. Nonlinear Sci. **12**, 283–318 (2002)
4. Bona, J.L., Chen, M., Saut, J.-C.: Boussinesq equations and other systems for small-amplitude long waves in nonlinear dispersive media: II. The nonlinear theory. Nonlinearity **17**, 925–952 (2004)
5. Bona, J.L., Lannes, D., Saut, J.-C.: Asymptotic models for internal waves. J. Math. Pures Appl. **89**, 538–566 (2008)
6. Boussinesq, J.V.: Théorie générale des mouvements qui sont propagés dans un canal rectangulaire horizontal. C. R. Acad. Sci. Paris **73**, 256–260 (1871)
7. Canuto, C., Hussaini, M.Y., Quarteroni, A., Zang, T.A.: Spectral Methods in Fluid Dynamics. Springer, New York (1988)
8. Dougalis, V.A., Duran, A., Mitsotakis, D.E.: Numerical approximation of solitary waves of the Benjamin equation. Math. Comput. Simul. **127**, 56–79 (2016)
9. Grimshaw, R., Pudjaprasetya, S.R.: Hamiltonian formulation for the description of interfacial solitary waves. Nonlinear Process. Geophys. **5**, 3–12 (1998)
10. Jbilou, K., Sadok, H.: Vector extrapolation methods. Applications and numerical comparisons. J. Comput. Appl. Math. **122**, 149–165 (2000)
11. Pego, R.L., Weinstein, M.I.: Convective linear stability of solitary waves for Boussinesq equations. Stud. Appl. Math. **99**, 311–375 (1997)
12. Pelinovsky, D.E., Stepanyants, Y.A.: Convergence of Petviashvili's iteration method for numerical approximation of stationary solutions of nonlinear wave equations. SIAM J. Numer. Anal. **42**, 1110–1127 (2004)

13. Peregrine, D.H.: Equations for water waves and the approximation behind them. In: Meyer, R.E. (ed.) Waves on Beaches and Resulting Sediment Transport, pp. 95–121. Academic, New York (1972)
14. Petviashvili, V.I.: Equation of an extraordinary soliton. Sov. J. Plasma Phys. **2**, 257–258 (1976)
15. Smith, D.A., Ford, W.F., Sidi, A.: Extrapolation methods for vector sequences. SIAM Rev. **29**, 199–233 (1987)
16. Sulem, C., Sulem, P.-L., Frisch, H.: Tracing complex singularities with spectral methods. J. Comput. Phys. **50**, 138–161 (1983)
17. Whitham, G.B.: Linear and Nonlinear Waves. Wiley, New York (1974)

# Subdivision Schemes and Multiresolution Analyses: Focus on the Shifted Lagrange and Shifted PPH Schemes

**Zhiqing Kui, Jean Baccou, and Jacques Liandrat**

**Abstract** Subdivision schemes have been extensively developed since the eighties with very powerful applications for surface generation. To be implemented for compression, subdivision schemes have to be coupled with decimation operators sharing some consistency relation and with detail operators. The flexibility of subdivision schemes (they can be non-stationary, position or zone dependent, non-linear,. . . ) makes that the construction of consistent decimation operators is a difficult task. In this paper, following the first results introduced in Kui et al. (On the coupling of decimation operator with subdivision schemes for multi-scale analysis. In: Lecture notes in computer science, vol. 10521. Springer, Berlin, pp. 162–185, 2016), we present the construction of multiresolution analyses connected to general subdivision schemes with detailed application to a non-interpolatory linear scheme called shifted Lagrange (Dyn et al., A $C^2$ four-point subdivision scheme with fourth order accuracy and its extensions. In: Mathematical methods for curves and surfaces: Tromsø 2004. Citeseer, 2005) and its non-linear version called shifted PPH (Amat et al., Math. Comput. 80:959–959, 2011).

**Keywords** Subdivision schemes · Multiresolutions · Decimation · Non-linear

Z. Kui · J. Liandrat (✉)
Aix Marseille Université, CNRS, Centrale Marseille, I2M, UMR 7353, Marseille, France
e-mail: zhiqing.kui@centrale-marseille.fr; jliandrat@centrale-marseille.fr

J. Baccou
Institut de Radioprotection et de Sureté Nucléaire (IRSN), PSN-RES/SEMIA/LIMAR, CE Cadarache, Saint Paul Les Durance, France
e-mail: jean.baccou@irsn.fr

© Springer Nature Switzerland AG 2019

J. L. García Guirao et al. (eds.), *Recent Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 18,
https://doi.org/10.1007/978-3-030-00341-8_9

# 1 Introduction

Since 50 years, subdivision schemes have been developed, analyzed and used with very powerful applications such as curve generation, image processing or animation movies. One of their advantages stands in the flexibility of the construction of their masks. One can easily derive various types of subdivision schemes adapted to specific goals; one can cite for instance the development of position dependent schemes devoted to the treatment of local singularities in the data [3] or the development of non-linearly perturbed subdivision schemes for removing undesirable behaviours [1]. When used for compression, subdivision schemes are plugged into a multiresolution framework that involves a so-called decimation operator [9]. However, although the decimation operator is trivially defined as a sub-sampling when handling interpolatory subdivision schemes, it is a difficult task to derive it in other situations.

Following the first results presented in [10], the contribution of this paper is a generic approach to derive decimation operators that will be called consistent with a given general subdivision scheme. We will focus on two examples related to the shifted Lagrange subdivision scheme (that is linear and non-interpolatory) introduced in [8] and to the shifted PPH subdivision scheme (that is non-linear and non-interpolatory) introduced in [2]. After an overview on subdivision schemes and multiresolution analyses (Sect. 2), we focus on the construction of decimation operators (Sect. 3) and on the so-called prediction errors and details (Sect. 4). Section 5 is devoted to numerical tests obtained using the shifted Lagrange and the shifted PPH multiresolutions.

# 2 Subdivision Schemes and Associated Multiresolution Transform

## *2.1 Subdivision Schemes*

### 2.1.1 General Construction

We consider binary subdivision schemes [7] defined as operators $S : l^\infty(\mathbb{Z}) \to l^\infty(\mathbb{Z})$ constructed from real-valued sequences $(h_k)_{k\in\mathbb{Z}}$ having a finite number of non-zero values such that $(f_k)_{k\in\mathbb{Z}} \in l^\infty(\mathbb{Z}) \mapsto ((Sf)_k)_{k\in\mathbb{Z}} \in l^\infty(\mathbb{Z})$ with

$$(Sf)_k = \sum_{l\in\mathbb{Z}} h_{k-2l} f_l. \tag{1}$$

The set of non-zero values of $(h_k)_{k\in\mathbb{Z}}$ is called the mask of $S$.

Subdivision is generally iterated starting from an initial sequence $(f_k^{J_0})_{k\in\mathbb{Z}}$ to generate $(f_k^j)_{k\in\mathbb{Z}}$ for $j \geq J_0$ as

$$f^{j+1} = Sf^j, j \geq J_0 \ . \tag{2}$$

One of the advantages of subdivision schemes stands in the flexibility associated with the choice of the mask when changing position ($k$), scale ($j$) and data ($f^j$). The simplest strategy is to consider the same mask for every position, scale and data [7] and to set $h_{2k} = \delta_{k,0}$. It leads to linear (i.e. the mask is independent of $f^j$), uniform (i.e. the mask is independent of $k$), stationary (i.e. the mask is independent of $j$) and interpolatory (i.e. $f_{2k}^{j+1} = f_k^j$) schemes. The main limitation of this type of construction is the generation of Gibbs oscillations for discontinuous data. A better strategy is then to work with non-uniform or non-interpolatory schemes. The shifted Lagrange scheme [8] is non-interpolatory. Moreover, for each position and scale, the construction of the subdivision can be adapted according to the regularity of the data to specifically handle discontinuities. Among various schemes following this strategy, one can mention the ENO [4] or PPH[1] [1] approach: for each $j$ and $k$, the subdivision depends on the values of the sequence $(f_l^j)_{l\in\mathbb{Z}}$ for $l$ in the vicinity of $k$. It leads to non-linear schemes.

In this paper, we focus on these types of schemes and more specifically on the shifted Lagrange and shifted PPH schemes. The construction of PPH has been first performed in an interpolatory framework [1] by modifying the classical linear Lagrange 4-point interpolating scheme [6]. Then, a non-interpolatory version (called shifted PPH) has been introduced in [2]. The corresponding definitions are recalled in the remaining of this section.

### 2.1.2 Shifted Lagrange and Shifted PPH Schemes

The situation we consider here is a specific case where a non-linear subdivision scheme is constructed as a perturbation of a linear one.

We start from the 4-point Lagrange interpolation and consider its shifted version. More precisely, for any value of $k \in \mathbb{Z}$ we build

$$P_k(x) = L_{-1}(x)f_{k-1} + L_0(x)f_k + L_1(x)f_{k+1} + L_2(x)f_{k+2},$$

where $\{L_n(x)\}_{-1\leq n\leq 2}$ denotes the degree 3 Lagrange interpolatory polynomial associated with the stencil $\{-1, 0, 1, 2\}$.

[1] As it will become clear in the next section, this type of scheme is derived from a perturbation of a linear one. Therefore, a non-linear mask cannot be constructed but it is straightforward to extend the definition of a subdivision scheme provided by Expression (1).

The 4-point shifted Lagrange linear scheme is given by:

$$\begin{cases} (S_{LA}f)_{2k} &= P_k(\frac{1}{4}) \\ (S_{LA}f)_{2k+1} &= P_k(\frac{3}{4}) \end{cases} . \tag{3}$$

The shifted PPH scheme is constructed introducing $N_k(x)$, a degree 3 polynomial depending non-linearly on $(f_{k-1}, f_k, f_{k+1}, f_{k+2})$ and substituting in (3) $P_k$ by $P_k + N_k$.

More precisely, if

$$A(x, y) = \frac{x+y}{2}, \quad H(x, y) = \frac{xy}{x+y}\,(sgn(xy) + 1)$$

where $sgn$ denotes the sign function and $D_k = (f_{k+1} - f_k) - (f_k - f_{k-1})$ then, if $|D_k| \leq |D_{k+1}|$,

$$N_k(x) = 2L_2(x)\,(H(D_k, D_{k+1}) - A(D_k, D_{k+1}))\,,$$

if $|D_k| > |D_{k+1}|$,

$$N_k(x) = 2L_{-1}(x)\,(H(D_k, D_{k+1}) - A(D_k, D_{k+1}))\,.$$

Introducing the non-linear perturbation $S_N$ defined by

$$\begin{cases} (S_N f)_{2k} &= N_k(\frac{1}{4}) \\ (S_N f)_{2k+1} &= N_k(\frac{3}{4}) \end{cases} ,$$

the shifted PPH subdivision scheme can be written as

$$S_{PPHA}f = S_{LA}f + S_N f. \tag{4}$$

Expression (2) can be interpreted as a two-scale relation. It is therefore possible to establish a connection between subdivision schemes and local prediction operators in a multiresolution framework. This framework as well as the associated precisions required to construct subdivision-based multiresolution transforms are recalled in the next section.

## 2.2 Multiresolution Transform

A Harten multiresolution analysis [9] is characterized by the family of triplets $\left((V^j, D^j_{j+1}, P^{j+1}_j)\right)_{j\in\mathbb{Z}}$ where $V^j$ is a separable space ($j$ is a scale parameter) and $D^j_{j+1}$ (resp. $P^{j+1}_j$) is a decimation (resp. prediction) operator connecting

$V^{j+1}$ (resp. $V^j$) to $V^j$ (resp. $V^{j+1}$). If $f^j \in V^j$ is obtained after decimation of $f^{j+1} \in V^{j+1}$, $P_j^{j+1} f^j$ does not usually coincide with $f^{j+1}$. However the following consistency condition is required:

$$D_{j+1}^{j} P_j^{j+1} = I_{V^j} \tag{5}$$

where $I_{V^j}$ stands for the identity operator in $V^j$.

In order to recover $f^{j+1}$ after a decimation and a prediction, a sequence of prediction errors $e^{j+1} = \left(e_k^{j+1}\right)_{k \in \mathbb{Z}}$ is introduced and defined as:

$$e_k^{j+1} = f_k^{j+1} - \left(P_j^{j+1} D_{j+1}^{j} f^{j+1}\right)_k = \left(\left(I_{V^{j+1}} - P_j^{j+1} D_{j+1}^{j}\right) f^{j+1}\right)_k \tag{6}$$

The mapping $f^{j+1} \mapsto \{f^j, e^{j+1}\}$ is a key ingredient for the multiresolution transform.

Focussing on a given $j$ and denoting $\tilde{h} = D_{j+1}^{j}$ and $h = P_j^{j+1}$, the consistency relation (5), $\tilde{h}h = I_{V^j}$, leads in the linear case to the fact that $\tilde{h}e^{j+1} = 0$. Introducing $d^j$ and $o^j$ such that $\forall k \in \mathbb{Z}, d_k^j = e_{2k+1}^{j+1}$ and $\forall k \in \mathbb{Z}, o_k^j = e_{2k}^{j+1}$, we get a linear relation of type $L^j o^j = R^j d^j$ where $L^j$ and $R^j$ are two linear operators. As soon as $L^j$ is invertible, this relation allows to recover $o^j$ from $d^j$, and therefore $e^{j+1}$ from $d^j$. One can then substitute the previous mapping by the following bijective one $f^{j+1} \mapsto \{f^j, d^j\}$.

In the non-linear case, one can not, in general, deduce such a bijection from the consistency relation. We postpone to Sect. 4 for the construction of this bijection in the case of the shifted PPH scheme.

Finally, for a fixed level $J_0 \leq j$, the multiresolution decomposition of a sequence $f^{j+1}$ is the element $\{f^{J_0}, d^{J_0}, d^{J_0+1}, \ldots, d^j\}$. The advantage of this representation stands in the fact that generally the norm of the details $d^l$ decays exponentially with the level $l$. Similarly a reconstruction transform can be introduced to recover $f^{j+1}$ from $\{f^{J_0}, d^{J_0}, d^{J_0+1}, \ldots, d^j\}$.

Exploiting (2), the prediction from $V^j$ to $V^{j+1}$ can be performed using a subdivision scheme. The main advantage of the corresponding operator is that it inherits the interesting properties of the scheme such as the flexibility in the construction of the mask. One can therefore introduce interpolatory or non-interpolatory, linear or non-linear predictions according to the data, that generally improve the classical wavelet-based multi-analyses framework [5]. However, the full specification of the multiresolution transforms is more involved. The construction of a decimation still remains difficult to tackle for non-interpolatory schemes since, generally, a subsampling operator does not satisfy the consistency property (5). Moreover, as mentioned in the previous section, the storage of the prediction error is well described for linear operators but cannot be extended to the non-linear framework. Therefore, we propose in the next sections new contributions to handle these two open questions with a specific application to the shifted Lagrange and shifted PPH schemes recalled in Sect. 2.1.

# 3 Decimation Operators

## 3.1 The Linear Case

A linear and uniform decimation operator $\tilde{h}_L$ is defined through a sequence $(\tilde{h}_k)_{k\in\mathbb{Z}}$ having a finite number of non-zero values as $(f_k)_{k\in\mathbb{Z}} \in l^\infty(\mathbb{Z}) \mapsto ((Df)_k)_{k\in\mathbb{Z}} \in l^\infty(\mathbb{Z})$ with

$$(Df)_k = \sum_{l\in\mathbb{Z}} \tilde{h}_{l-2k} f_l \quad .$$

The set of non-zero values of $(\tilde{h}_k)_{k\in\mathbb{Z}}$ is called the mask of $D$ and denoted $M_{\tilde{h}}$.

A generic method to construct the mask of a decimation scheme consistent with a given linear uniform subdivision scheme has been proposed in [10]. The main results are recalled in the two following propositions. The first one is devoted to the construction of elementary operators (i.e. with masks of minimal number of non-zero values) while the second one describes how all consistent decimation operators can be recovered using linear combinations of translated versions of elementary operators.[2]

**Proposition 1** *Let $h$ be a prediction operator whose mask is constructed from the sequence*

$$\{h_{n-2\alpha}, h_{n-2\alpha+1}, \ldots, h_n, h_{n+1}\}$$

*with $h_{n-2\alpha}h_{n+1} \neq 0$ (i.e. mask of even length) or with $h_{n-2\alpha} = 0$ and $h_{n-2\alpha+1}h_{n+1} \neq 0$ (i.e. mask of odd length).*

*Introducing $H_{M_h} =$*

$$\begin{bmatrix} h_n & h_{n-2} & \cdots & h_{n-2\alpha} & 0 & \cdots & 0 \\ h_{n+1} & h_{n-1} & \cdots & h_{n-2\alpha+1} & 0 & \cdots & 0 \\ 0 & h_n & h_{n-2} & \cdots & h_{n-2\alpha} & \cdots & 0 \\ 0 & h_{n+1} & h_{n-1} & \cdots & h_{n-2\alpha+1} & \cdots & 0 \\ \vdots & & & \vdots & & & \\ 0 & 0 & \cdots & h_n & h_{n-2} & \cdots & h_{n-2\alpha} \\ 0 & 0 & \cdots & h_{n+1} & h_{n-1} & \cdots & h_{n-2\alpha+1} \end{bmatrix},$$

*if $det(H_{M_h}) \neq 0$, there exists at most $2\alpha$ consistent elementary decimation operators whose masks are of length not larger than $2\alpha$. These masks are given by each row of $H_{M_h}^{-1}$.*

[2] For any decimation operator $\tilde{h}$ constructed from $(\tilde{h}_k)_{k\in\mathbb{Z}}$ and any integer $t$, we define $T_t(\tilde{h})$ the translated decimation operator related to the sequence $(\tilde{h}_{k-t})_{k\in\mathbb{Z}}$.

**Proposition 2** *A subdivision operator h being fixed and satisfying the hypotheses of Proposition 1, we denote* $\{\tilde{h}^i\}_{1\leq i\leq 2\alpha}$ *the set of the elementary consistent decimation operators.*

*Then, all the consistent decimation operators can be constructed as*

$$\sum_{t\in\mathfrak{T}}\sum_{i\in\mathfrak{I}} c_{i,t}T_{2t}(\tilde{h}^i) \tag{7}$$

*with*

$$\forall t \in \mathfrak{T} \subset \mathbb{Z},\ \sum_{i\in\mathfrak{I}} c_{i,t} = \delta_{t,0},\ \ \textit{and}\ 0 \in \mathfrak{T}\ .$$

The large choice of decimation schemes offered by this approach is of prime importance in practice since it allows to optimize some specific characteristics of the scheme according to given objectives. In image compression for example, it is important to select decimation operators with minimal $L^\infty$-norm because this norm is involved in the stability of the multiresolution transforms.

We finally conclude this section by applying our approach to the shifted Lagrange linear subdivision scheme for which, up to now, there was no consistent decimation operator available in the literature.

*Decimation operators consistent with the shifted Lagrange scheme*

From (3), the shifted Lagrange scheme is given by:

$$\begin{cases} (S_{LA}f)_{2k} = -\frac{7}{128}f_{k-1} + \frac{105}{128}f_k + \frac{35}{128}f_{k+1} - \frac{5}{128}f_{k+2}, \\ (S_{LA}f)_{2k+1} = -\frac{5}{128}f_{k-1} + \frac{35}{128}f_k + \frac{105}{128}f_{k+1} - \frac{7}{128}f_{k+2}. \end{cases}$$

and its mask is therefore

$$M_h = \{h_{-4}, h_{-3}, h_{-2}, h_{-1}, h_0, h_1, h_2, h_3\}$$

$$= \{-\frac{5}{128}, -\frac{7}{128}, \frac{35}{128}, \frac{105}{128}, \frac{105}{128}, \frac{35}{128}, -\frac{7}{128}, -\frac{5}{128}\}.$$

A straightforward application of Proposition 1 leads to the set of consistent elementary decimation operators whose masks correspond to each row of the following matrix:

$$\tilde{H}_{LA} = \begin{bmatrix} \frac{24367}{1152} & -\frac{63605}{1152} & \frac{31115}{576} & -\frac{10325}{576} & -\frac{4165}{1152} & \frac{2975}{1152} \\ \frac{2975}{1152} & -\frac{4165}{1152} & \frac{1771}{576} & -\frac{565}{576} & -\frac{245}{1152} & \frac{175}{1152} \\ \frac{175}{1152} & -\frac{245}{1152} & \frac{875}{576} & -\frac{245}{576} & -\frac{133}{1152} & \frac{95}{1152} \\ \frac{95}{1152} & -\frac{133}{1152} & -\frac{245}{576} & \frac{875}{576} & -\frac{245}{1152} & \frac{175}{1152} \\ \frac{175}{1152} & -\frac{245}{1152} & -\frac{565}{576} & \frac{1771}{576} & -\frac{4165}{1152} & \frac{2975}{1152} \\ \frac{2975}{1152} & -\frac{4165}{1152} & -\frac{10325}{576} & \frac{31115}{576} & -\frac{63605}{1152} & \frac{24367}{1152} \end{bmatrix}.$$

Following Proposition 2, one can also increase the length of the decimation mask in order to reduce the $L^\infty$-norm of the operator. If for example the length is fixed to 8, the mask of minimal $L^\infty$-norm is given by:

$$\begin{aligned} M_{\tilde{h}} &= \{\tilde{h}_{-4}, \tilde{h}_{-3}, \tilde{h}_{-2}, \tilde{h}_{-1}, \tilde{h}_0, \tilde{h}_1, \tilde{h}_2, \tilde{h}_3\} \\ &= \{\frac{95}{2304}, -\frac{133}{2304}, -\frac{35}{256}, \frac{1505}{2304}, \frac{1505}{2304}, -\frac{35}{256}, -\frac{133}{2304}, \frac{95}{2304}\}. \end{aligned} \tag{8}$$

This decimation operator is often considered since it has the same length as the initial subdivision scheme and it is symmetrical.

## 3.2 Extension to the Non-linear Perturbation Case

Generalizing Expression (4) and using the notations of the subdivision-based multiresolution framework, the non-linear subdivision scheme considered in this paper satisfies:

$$hf^j = h_L f^j + h_N f^j, \tag{9}$$

where $h_L$ is the prediction associated to a linear subdivision scheme and $h_N$ stands for a non-linear perturbation operator.

Starting from a sequence $f^{j+1} \in V^{j+1}$, we can construct a consistent non-linear decimation operator by solving the following fixed-point equation:

$$f^j = \tilde{h}_L f^{j+1} - \tilde{h}_L h_N f^j, \tag{10}$$

where $\tilde{h}_L$ is a linear decimation operator consistent with $h_L$ and constructed by the method described in the previous section.

The contractivity of the operator $\tilde{h}_L h_N$ which means that there exists $c \in \mathbb{R}$, $|c| < 1$, such that $\forall (u, v) \in (l^\infty(\mathbb{Z}))^2$,

$$||\tilde{h}_L h_N u - \tilde{h}_L h_N v|| \leq c||u - v||,$$

is required to ensure the existence and uniqueness of $f^j$ according to the Banach fixed-point theorem.

Moreover, $f^j = \lim_{n \to +\infty} (f^j)_n$ where $((f^j)_n)_{n \in \mathbb{Z}}$ is constructed by induction:

$$\begin{cases} (f^j)_0 &= \tilde{h}_L f^{j+1} \\ (f^j)_{n+1} &= \tilde{h}_L f^{j+1} - \tilde{h}_L h_N (f^j)_n, \end{cases} . \tag{11}$$

Concerning the consistency of the decimation operator defined above, if $f^{j+1} = hg^j$, the unique solution of the fixed-point equation denoted $\hat{f}^j$ satisfies:

$$\hat{f}^j = \tilde{h}_L hg^j - \tilde{h}_L h_N \hat{f}^j = g^j + \tilde{h}_L h_N g^j - \tilde{h}_L h_N \hat{f}^j.$$

Since $g^j$ is solution of the previous equation, it is the unique solution. Therefore $g^j = \tilde{h} f^{j+1} = \tilde{h}hg^j$ that leads to the consistency of $\tilde{h}$.

When $h$ is the shifted PPH non-linear scheme, the following proposition provides that, for a suitable choice of $\tilde{h}_L$, the fixed-point is unique and can be reached using fixed-point iterations:

**Proposition 3** *If $h_L$ is the prediction operator associated with the shifted Lagrange linear scheme and $\tilde{h}_L$ is the consistent decimation whose mask is given by*

$$\begin{aligned}&\{\tilde{h}_{-6}, \tilde{h}_{-5}, \tilde{h}_{-4}, \tilde{h}_{-3}, \tilde{h}_{-2}, \tilde{h}_{-1}, \tilde{h}_0, \tilde{h}_1, \tilde{h}_2, \tilde{h}_3, \tilde{h}_4, \tilde{h}_5\}\\ =&\{\frac{19}{16128}, -\frac{19}{11520}, \frac{19}{576}, -\frac{19}{576}, -\frac{2623}{16128}, \frac{7639}{11520},\\ &\frac{7639}{11520}, -\frac{2623}{16128}, -\frac{19}{576}, \frac{19}{576}, -\frac{19}{11520}, \frac{19}{16128}\}\end{aligned}$$

*then the operator $\tilde{h}_L h_N$ is contractive with a Lipschitz constant bounded by $\frac{19657}{20160}$.*

# 4 Prediction Errors and Details

## 4.1 Construction of the Details

As mentioned in Sect. 2.2, the construction of the details is straightforward in the linear case and is based on a splitting of the prediction error that belongs to the kernel of the decimation operator. One can not apply such a construction in the non-linear case. However, for the shifted PPH scheme, the same idea can be exploited to address the problem of prediction error storage. It is stated by the following proposition.

**Proposition 4** *Let $h = h_L + h_N$ be a non-linear subdivision operator defined by (4) and $\tilde{h}$ be a consistent decimation operator constructed from Eq. (10).*

*For all $f^{j+1} \in l^\infty(\mathbb{Z})$, the associated prediction error $e^{j+1}$ satisfies*

$$\tilde{h}_L e^{j+1} = 0, \tag{12}$$

*and*

$$\tilde{h} e^{j+1} = 0. \tag{13}$$

*Proof* Under the contractivity condition, if $\hat{f}^j = \tilde{h} f^{j+1}$ denotes the unique solution to Eq. (10), the prediction error can be written as

$$\begin{aligned} e^{j+1} &= f^{j+1} - h\tilde{h} f^{j+1} \\ &= f^{j+1} - h_L \hat{f}^j - h_N \hat{f}^j \\ &= (I - h_L \tilde{h}_L) f^{j+1} - (I - h_L \tilde{h}_L) h_N \hat{f}^j . \end{aligned}$$

Then Eq. (12) is straightforward using the consistency condition linking $h_L$ and $\tilde{h}_L$.

Assuming $w^j = \tilde{h} e^{j+1}$,

$$w^j = \tilde{h}_L e^{j+1} - \tilde{h}_L h_N w^j = -\tilde{h}_L h_N w^j .$$

According to the Banach fixed-point theorem, $w^j = 0$ is the unique solution which leads to (13). □

Therefore, following what has been done in Sect. 2.2 and introducing the corresponding operators $L^j_L$ and $R^j_L$, a bijective mapping $f^{j+1} \mapsto (f^j, d^j)$ can be derived leading to the details $d^j$.

## 4.2 Prediction Error Decay

The following proposition, borrowed from [10], explains that for the linear case and a couple of consistent operators $(h, \tilde{h})$, the decay of the prediction error with the scale is fully controlled by the polynomial approximation property of the subdivision operator. Indeed, we have:

**Proposition 5** *Let $(V^j, h, \tilde{h})_{j\in\mathbb{Z}}$ define a linear multiresolution analysis with $h$ a subdivision operator associated to the real sequence $(h_k)_{k\in\mathbb{Z}}$ and $\tilde{h}$ a decimation operator constructed from the real sequence $(\tilde{h}_k)_{k\in\mathbb{Z}}$. We assume that the associated multi-scale transform is applied from a fine scale $J_{max}$ to a coarse one $J_0$.*

*If there exists $L \in \mathbb{N}$ such that $\forall n \in \{0, 1, \ldots, L\}$,*

$$\sum_{l\in\mathbb{Z}} h_{2l}(2l)^n = \sum_{l\in\mathbb{Z}} h_{2l+1}(2l+1)^n \quad , \tag{14}$$

*then for sufficiently large $j \in [J_0, J_{max} - 1]$,*

$$||e^j|| \le C 2^{-(L+1)j}, \tag{15}$$

*where $C$ does not depend on $j$.*

In the case of the 4-point shifted Lagrange scheme, it is easy to verify that $\forall n \in \{0, 1, 2, \ldots, 4\}$,

$$\sum_{i=-1}^{2} L_i(\frac{1}{4})(-2i)^n = \sum_{i=-1}^{2} L_i(\frac{3}{4})(-2i+1)^n,$$

and it follows from the previous proposition that the decay rate of the associated prediction error is 5.

When the subdivision is non-linear the consistency and the polynomial approximation property can not be combined to provide the decay of the prediction error, and therefore of the details. A direct analysis must be performed. In this paper, a numerical study is provided in the next section.

## 5 Numerical Results

This section provides a series of tests in order to evaluate the linear (shifted Lagrange) and non-linear (shifted PPH) multiresolutions. The results obtained using the Lagrange interpolatory scheme with mask

$$M_h = \{-1/16, 0, 9/16, 1, 9/16, 0, -1/16, 0\}$$

and consistent decimation $M_{\tilde{h}} = \{1\}$ are also provided as a comparison.

Initial data are sampled from related functions. Figure 1 deals with a piecewise regular function with a discontinuity at $x_0 = 0.5$ (Left). The multiresolution transform of this function is performed using the linear consistent decimation operator given by Proposition 3 while the non-linear one satisfies the fixed-point equation (10). Figure 1 (right) displays the convergence rate of the fixed-point in algorithm (11). The slope of the curve exhibits the convergence rate. It appears that very few iterations (less than 17) are required for convergence.

In Fig. 2, the prediction error associated to regular data (sampled from the $sin$ function) is numerically estimated and plotted versus the scale. For each approach, a multi-scale decomposition transform is applied from a fine level $J_{max} = 12$ to a coarse one $J_0 = 7$. It appears that the decay rate is larger for the linear approach (slope of 5.0379, to be compared with the theoretical value of 5) than for the non-linear one (slope of 4.21979).This can be explained by the presence of the non-linear perturbation term that reduces the degree of polynomial approximation. As a comparison, the interpolatory approach has a slope of 4.00717, to be compared with the theoretical value of 4.

For the functions of Fig. 3, we consider in Fig. 4 the prediction errors around the point $x_0 = 0, 5$. It appears that the stronger the discontinuity the better is the related performance of the non-linear scheme. This good behaviour of the non-linear scheme is confirmed by the plots in Fig. 5.

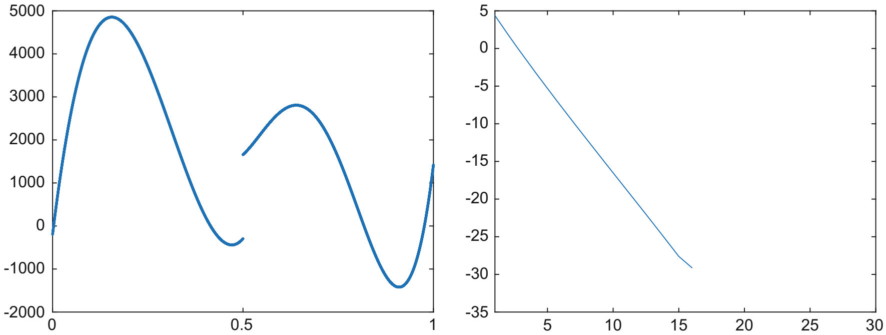


**Fig. 1** Left: Test function, Right: $\log ||f_n^j - f_{n-1}^j||_\infty$ versus $n$ for the fixed-point algorithm (11)

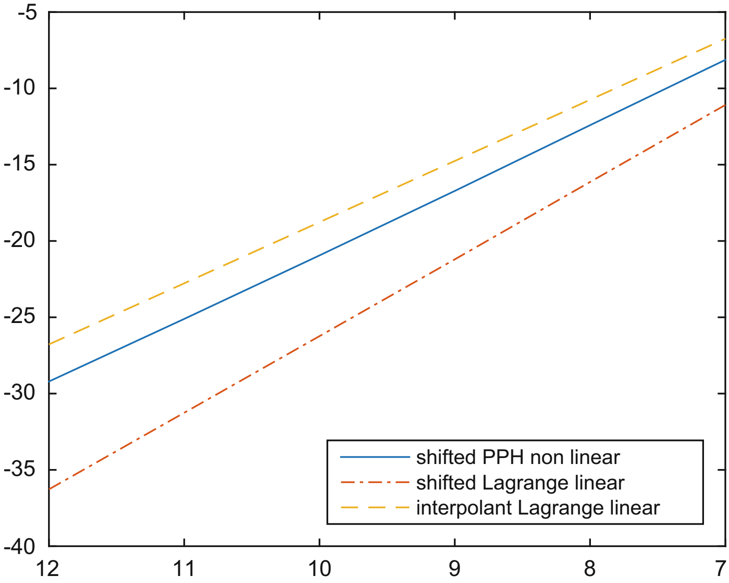


**Fig. 2** Log of the prediction error versus scale

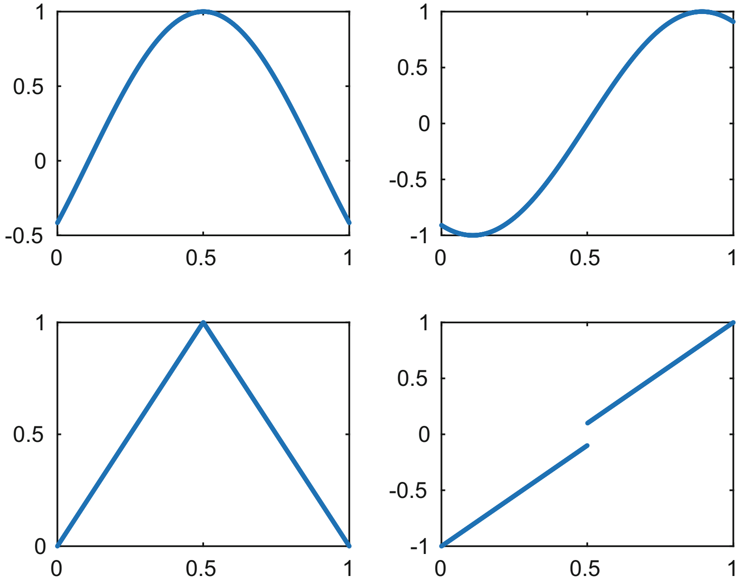


**Fig. 3** Test functions

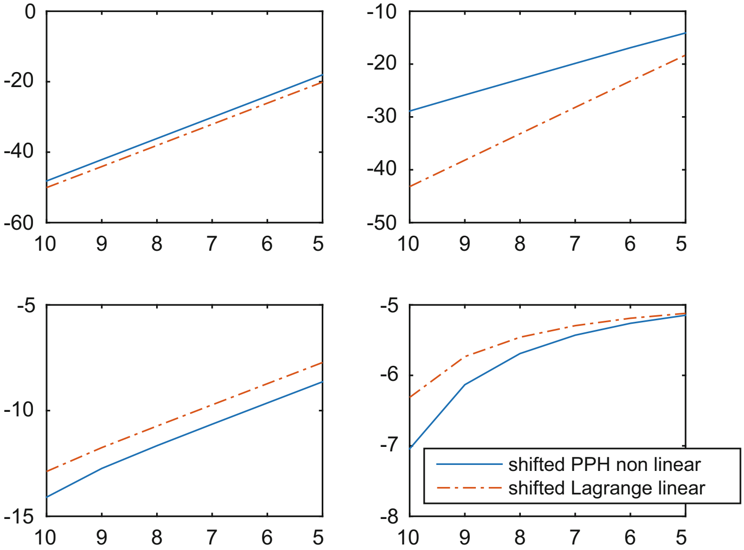


**Fig. 4** Log of the prediction error versus scale

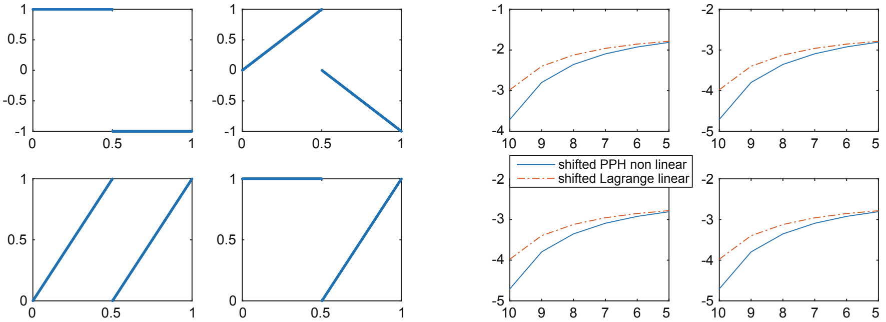


**Fig. 5** Left: Test functions, Right: Log of the prediction error versus scale

# 6 Conclusion

In this paper, we have defined the different elements required to construct a multiresolution analysis associated to linear or non-linear subdivision schemes. The key points are the generation of consistent decimation operators and the definition of the details starting from the prediction error. The construction of decimation is performed inverting a matrix in the linear case and next solving a fixed-point problem in the non-linear case. The detail operators are constructed exploiting a specific relation implying a linear decimation. We have described the construction of two schemes (shifted Lagrange and shifted PPH) for which no multiresolution was available before. The different numerical results show the interest of such frameworks for the compression of data.

# References

1. Amat, S., Donat, R., Liandrat, J., Trillo, J.C.: Analysis of a fully nonlinear multiresolution scheme for image processing. Found. Comput. Math. **6**, 193–225 (2006)
2. Amat, S., Dadourian, K., Liandrat, J.: On a nonlinear subdivision scheme avoiding Gibbs oscillations and converging towards $C^s$ functions with $s > 1$. Math. Comput. **80**, 959 (2011)
3. Baccou, J., Liandrat, J.: Position-dependent Lagrange interpolating multiresolutions. Int. J. Wavelet Multiresolution Inf. **5**, 513–539 (2005)
4. Cohen, A., Dyn, N., Matei, B.: Quasilinear subdivision schemes with applications to ENO interpolation. Appl. Comput. Harmon. Anal. **15**, 89–116 (2003)
5. Daubechies, I.: Ten Lectures on Wavelets. SIAM, Philadelphia (1992)
6. Deslaurier, G., Dubuc, S.: Interpolation dyadique. In: Fractals, dimensions non entières et applications, pp. 44–45. Masson, Paris (1987)
7. Dyn, N.: Subdivision schemes in computer-aided geometric design. In: Light, W.A. (ed.) Advances in Numerical Analysis II, Wavelets, Subdivision Algorithms and Radial Basis Functions, pp. 36–104. Clarendon Press, Oxford (1992)

8. Dyn, N., Floater, M.S., Hormann, K.: A $C^2$ four-point subdivision scheme with fourth order accuracy and its extensions. In: Mathematical Methods for Curves and Surfaces: Tromsø 2004. Citeseer (2005)
9. Harten, A.: Multiresolution representation of data: a general framework. SIAM J. Numer. Anal. **33**, 1205–1256 (1996)
10. Kui, Z., Baccou, J., Liandrat, J.: On the coupling of decimation operator with subdivision schemes for multi-scale analysis. Lecture Notes in Computer Science, vol. 10521, pp. 162–185. Springer, Berlin (2016)

# Modelling Sparse Saliency Maps on Manifolds: Numerical Results and Applications

**Euardo Alcaín, Ana Isabel Muñoz, Iván Ramírez, and Emanuele Schiavi**

**Abstract** Saliency detection is an image processing task which aims at automatically estimating visually salient object regions in a digital image mimicking human visual attention and eyes fixation. A number of different computational approaches for visual saliency estimation has recently appeared in Computer and Artificial Vision. Relevant and new applications can be found everywhere varying from automatic image segmentation and understanding, localization and quantification for biomedical and aerial images to fast video tracking and surveillance. In this contribution, we present a new variational model on finite dimensional manifolds generated by some characteristic features of the data. A Primal-Dual method is implemented for the numerical resolution showing promising preliminary results.

**Keywords** Saliency detection and segmentation · Superpixels · Non local total variation on graphs · Energy minimization · Primal-dual algorithm

## 1 Introduction

In the last years, there has been a growing interest in the production of computational methods for the detection of saliency objects in an image. A general overview of existing methods, drawbacks and virtues can be found in [5]. Salient (or foreground) objects are those objects which grasp the most interest when an image is considered. This estimation is normally used as a preprocessing step in a pipeline for a computer vision system. The concept of saliency has been applied to adaptive compression of images [7], image retrieval [2] and image cropping [14] to name just a few emerging applications. Saliency methods can be classified into three categories: biological based, purely computational and a mix from both methods.

E. Alcaín · A. I. Muñoz (✉) · I. Ramírez · E. Schiavi
Departamento de Matemática Aplicada, Ciencia e Ingeniería de Materiales y Tecnología Electrónica, Universidad Rey Juan Carlos, ESCET, Móstoles, Madrid, Spain
e-mail: e.alcain@alumnos.urjc.es; anaisabel.munoz@urjc.es; i.ramirez@alumnos.urjc.es; emanuele.schiavi@urjc.es

© Springer Nature Switzerland AG 2019

J. L. García Guirao et al. (eds.), *Recent Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 18,
https://doi.org/10.1007/978-3-030-00341-8_10

Although variational methods have achieved great success when applied in computer vision (denoising, deblurring, inpainting, etc.), there have been few saliency detection models which make use of the power of the variational setting. This introduces the main contribution of the paper which consists of a new variational model for automatic saliency detection in digital images based on Total Variation (TV) minimization on graphs.

We consider as a starting point of our modelling exercise the work in [15] where a non local $L_0$ minimization problem is proposed for bottom-up pure computationally saliency estimation. Based on the discrete $L_0$ norm on graphs, this model is appropriate to capture the sparse properties of saliency maps which are vectors, finite dimensional solutions which estimate the saliency of each superpixel in which the image has been previously partitioned. In fact, following the idea in [15] and with a view to fast (real time) video saliency detection, the data image is divided into superpixels for dimensional reduction and a non local graph is constructed in a feature space in which each vertex is a superpixel connected to its k-NN (k-nearest neighbour).

There has been a lot of effort in these years to develop the formalism of Non-local Calculus [9] and Non-local Operators so providing a new tool for the mathematical analysis of problems on graphs. This has generated a new kind of models where the influence of any pixel value is extended to all (or part) of the domain through the concept of Non-local derivatives. When a graph structure is generated, by considering some characteristics of a given data image, spatial proximity is lost or attenuated and Non-local operators are necessary to describe the manifold landscape.

In such a framework, our proposal is twofold. We first generalize the superpixels partition considered in [15] including edges as features in the graph by means of the use of the SCALP algorithm [12] where contour adherence is imposed using linear paths. Also, instead of considering a fixed threshold, the final saliency mask is obtained using the method proposed in [13]. Secondly, and more important, the $L_0$-norm of the non local gradients in the saliency model presented [15] is replaced by the NLTV (non local total variation) semi-norm on graphs defined by the non local total variation operator, which preserves edges and induces the sparsity of the gradients of saliency maps. This step amounts to a convexification of the problem which allows classical variational calculus to be applied. Finally, as in [15], a control map is computed in the graph to drive the solution towards salient regions. With all these ingredients a strictly convex energy functional is considered for minimization and the well-posedness of the model is guaranteed. The numerical resolution is based on a primal-dual algorithm that we designed to solve the associated minimization NLTV problem. The proposed algorithm proves to be faster in convergence to the solution than the one presented in [15] for the non local $L_0$ minimization problem, while obtaining results at least comparable to the ones in [15].

The paper is organized as follows: First, in Sect. 2 we introduce the concept of superpixel (SP) as a guided (informed) partition of the given image through the chosen characteristics. This is a critical step because the produced superpixels partitions (in number and characteristics) shall not be modified along the whole pipeline. In Sect. 3 the finite dimensional manifold (weighted graph) generated by the partition of the image into superpixels by using some characteristic features of the data is constructed. Sections 4 and 5 are devoted to the mathematical definitions of the control map and non-local operators used in the model. The main contribution is presented in Sect. 6 where the proposed Non-Local Total Variation Model (NLTVM) and its numerical resolution in terms of a primal-dual algorithm are described. The final saliency map segmentation step is considered in Sect. 7. Numerical results and discussion are presented in Sects. 8 and 9 in order to illustrate the performance of the proposed model.

## 2 Superpixel Segmentation of the Image

This section is based on the ideas and methods published in [12, 15]. In order to compute the graph structure, we first segment the given image into superpixels which are clusters of pixels partitioning the image following some relevant characteristics of the data. This will reduce the computation time of the saliency map providing a dimensional reduction of the problem and fast implementation. Notice also that the relevant information originally located into pixels is then encoded through the shape and values of the superpixel. The method used in [15] to generate superpixels is the SLIC method (Simple Linear Iterative Method [1]) which has been proven to be very efficient in the sense that it is fast, easy to use and it produces high quality partitions of the image. This method performs a local clustering of pixels taking into account the location and the values of the CIELAB colour space of the pixels of an image. The employed measure makes the superpixels shape to be compact and uniform. As an alternative to SLIC in [15] and in order to impose edges conservation we consider the SCALP method (superpixels with contour adherence using linear path [12]) which takes into consideration a contour prior with the aim of obtaining superpixels constrained by the existing contours, being this aspect not considered with SLIC (see Fig. 1 for comparison). In the SCALP algorithm, when trying to associate a pixel to a superpixel during clustering, the distance is enhanced by considering the linear path to the superpixel barycenter and a contour prior. The

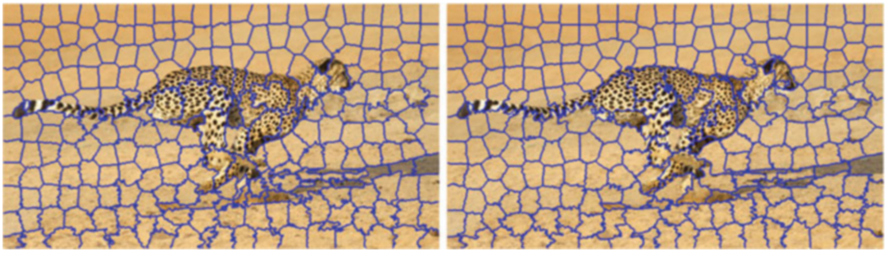


**Fig. 1** Comparison between SLIC (left) vs SCALP (right) algorithm with 300 superpixels. The edges and contours of the cheetah are better preserved in the partition provided by SCALP

prior contour can be calculated by some recent learning based approaches to detect edges like [8] or more classical variational approaches enhanced by implicit finite differences [3]. In this work we used the approach in [8] suggested in [12]. Different strategies shall be explored in future work.

# 3 Weighted Graph

Given an image and the superpixels partition generated by SLIC or SCALP method, we consider, as in [15], an undirected, symmetric and weighted graph $G$ in the space of the superpixels. The weighted graph $G = (V, E, w)$ consists of a finite set of superpixels $V$, the edges $E$ linking superpixels and their associated weights $w_{pq}$, $pq \in E$. The weights are given by:

$$w_{pq} = exp\left(-\frac{||\mathbf{f}_p - \mathbf{f}_q||^2}{2\sigma^2}\right),$$

where $\mathbf{f}_p$ is a feature vector at superpixel $p$ defined by $\mathbf{f}_p = (\alpha \mathbf{c}_p, \mathbf{l}_p)$, being $\mathbf{c}_p$ the mean of the superpixels in CIELAB colour space, and $\mathbf{l}_p$, the mean of the coordinates in the spatial space. In this case, $\alpha = 0.9$ is a parameter controlling the balance between the two features.

In order to reduce computational cost and to exploit local relationships in the feature space, once we have computed the weight for each superpixel $p$ with every other superpixel $q$, we consider its $k$-nearest neighbours the first $k$ vertices in a decreasing list with respect to weight) and keep these associated weights while setting to zero the remaining ones. Then, a superpixel $p$ is associated with a superpixel $q$ if the weight $w_{pq}$ is not zero. To go further in the suppression of the background, the boundaries are connected together as well those superpixels such that their initial saliency map has a value (score) below a user fixed threshold in the weight matrix. In our case the 25% of the less representative values in the control map are neglected (see Fig. 2).

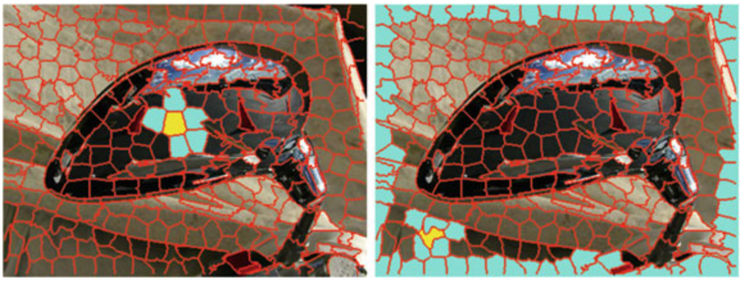


**Fig. 2** The image was divided by SLIC. On the left, we show a superpixel (yellow) and its k-neighbours. On the right, it is shown a superpixel within less than 25 % in the initial control map is associated to the boundaries as well to its k-neighbours

# 4 Control Map

We still follow the ideas in [15] to compute a saliency control map which we will include in the fidelity term of the variational model here presented. The control map $v^c$ is a vector of components $v_c = \{v_p^c,\ p \in G\}$, where $v_p^c$ is the value of the salient map in the superpixel $p$. This value, $v_p^c$, is computed as the product of a contrast prior $v_p^{con}$, which takes into consideration similarity of colours weighted by the distance to each superpixel and an object prior $v_p^{obj}$, which introduces the assumption of the most likely location of the salient object. Hence, the saliency control map is computed:

$$v_p^c = v_p^{con} \cdot v_p^{obj},$$

where

$$v_p^{con} = \sum_{q,pq \in E} w_{pq}^{(l)} ||c_p - c_q||^2, \quad w_{pq}^{(l)} = exp\left(-\frac{1}{2\sigma^2}||l_p - l_q||^2\right)$$

and

$$v_p^{obj} = exp\left(-\frac{||l_p - \bar{l}||^2}{2\sigma^2}\right).$$

where $\bar{l}$ are the coordinates of the center of the image (where the region of interest is usually located). The parameter $\sigma$ is empirically set as in [15] where $\sigma^2 = 0.05$.

# 5 Non Local Operators

In order to deal with variational models in a weighted graph, we need to introduce the notion of non local gradient and divergence operators in a weighted graph $G$. Details can be found in [9].

**Definition 1** Let $p \in G$, $v = (v_p)_{p \in G}$ be a real function defined on $G$ and taking values in $I\!R$, and $w_{p,q}$, for $p$ and $q \in G$, a nonnegative symmetric weight function. Then, the nonlocal gradient at a superpixel $p$, $\nabla_w v_p$, is defined as the vector of all partial differences $\nabla_w v_{p,q}$ at $p$:

$$\nabla_w v_p = \{\nabla_w v_{p,q},\ pq \in E\},$$

where

$$\nabla_w v_{p,q} = \sqrt{w_{p,q}}(v_q - v_p).$$

In fact, we are only interested in the components such that $w_{p,q} \neq 0$. Hence, the nonlocal gradient $\nabla_w v$ is defined as $\nabla_w v = (\nabla_w v_p)_{p \in G}$. Analogously, the divergence $div_w$ of a vector $d_p$, given a function $d : G \times G \to I\!R$ can be defined as:

$$div_w d_p = \sum_{q,\, qp \in E} (d_{p,q} - d_{q,p})\sqrt{w_{p,q}}.$$

Notice that, in our particular case, $d_p$ will be taken as $\nabla_w v_p$.

## 5.1 Non Local $L_0$ for Saliency Detection

In this section, we shall briefly present the model and the method of numerical resolution presented in [15]. This will allow the reader to notice where our contribution resides.

In [15] it is proposed the following nonlocal discrete $L_0$ minimization model to describe the sparse structure of the saliency maps:

$$\min_v \left( \sum_p ||\nabla_w v_p||_0 + \frac{\lambda}{2}||v - v^c||^2 \right), \tag{1}$$

where $v_p$ stands for the value of the saliency map $v$ at the superpixel $p$, $\nabla_w v_p = \{\nabla_w v_{p,q} : pq \in E\}$, as it was defined in the previous section, and the $L_0$ norm is given by $||\nabla_w v_p||_0 = \#\{q : \nabla_w v_{p,q} \neq 0,\ pq \in E\}$ (# is the cardinality of the set). The positive constant $\lambda$ is a parameter controlling the relative importance given to the fidelity term with respect to the sparsity inducing prior.

In order to solve the minimization problem, the authors in [15] use an Alternated Directions Method (ADM). Setting $d_p = \nabla_w v_p$ the equivalent minimization problem

$$\min_{v,d} \left( \sum_p ||d_p||_0 + \frac{\lambda}{2}(v_p - v_p^c)^2 \right).$$

is solved through the iterative scheme

$$v^{k+1} = \min_{v} \left( \sum_{p} \frac{\lambda}{2}(v_p - v^c_p)^2 + \frac{\rho}{2}||d^k_p - \nabla_w v_p + \frac{1}{\rho} y^k_p||^2 \right)$$

$$d^{k+1} = \min_{d} \left( \sum_{p} ||d_p||_0 + \frac{\rho}{2}||d_p - \nabla_w v^{k+1}_p + \frac{1}{\rho} y^k_p||^2 \right)$$

where the relaxation variable is given by $y^{k+1}_p = y^k_p + \rho(d^k_p - \nabla_w v^{k+1}_p)$. The solution of the problem for $v^{k+1}_p$ can be written as

$$(\lambda + 2\rho \sum_q w_{pq}) v^{k+1}_p = \lambda v^c_p$$

$$-\rho \sum_q \sqrt{w_{pq}}(d^k_{pq} - d^k_{qp} + \frac{1}{\rho}(y^k_{pq} - y^k_{qp})) + 2\rho \sum_q w_{pq} v^{k+1}_q .$$

Finally, the auxiliary problem for $d^{k+1}$ has a component-wise close solution is given by:

$$d^{k+1}_{pq} = \begin{cases} 0, & |\nabla_w v^{k+1}_{p,q} - \frac{1}{\rho} y^k_{pq}| \leq \sqrt{2}\rho, \\ \nabla_w v^{k+1}_{p,q} - \frac{1}{\rho} y^k_{pq}, & \text{otherwise.} \end{cases}$$

In [15], the iterative procedure is stopped when a fixed number of iterations is reached. This is not a convergence criteria and obscures the behaviour of the algorithm. In order to analyze and compare the results with the ones obtained with our NLTV model, we consider in both cases a typical convergence criteria: the computation is stopped when the difference between two consecutive iterations is less than a small value $\epsilon$. In our computations we have considered $\epsilon = 10^{-5}$. We refer the reader to [15] for more details. Notice that this convergence criteria do not reveal if the limit solution is a minimum of the original $L_0$ model problem (1). It just shows that the algorithm stabilizes and converges to a saliency vector which may not be a minimum of functional (1). This is due to the fact that, contrary to our model, the energy in (1) is not convex. In fact, a careful analysis of the behaviour of the energy functional reveals the existence of two different cases which we show in Figs. 3 and 4 where the total energy, the $L_0$ energy and the fidelity $L_2$ discrepancy norm are represented. We shall refer to case 1 when the total energy stabilizes to a

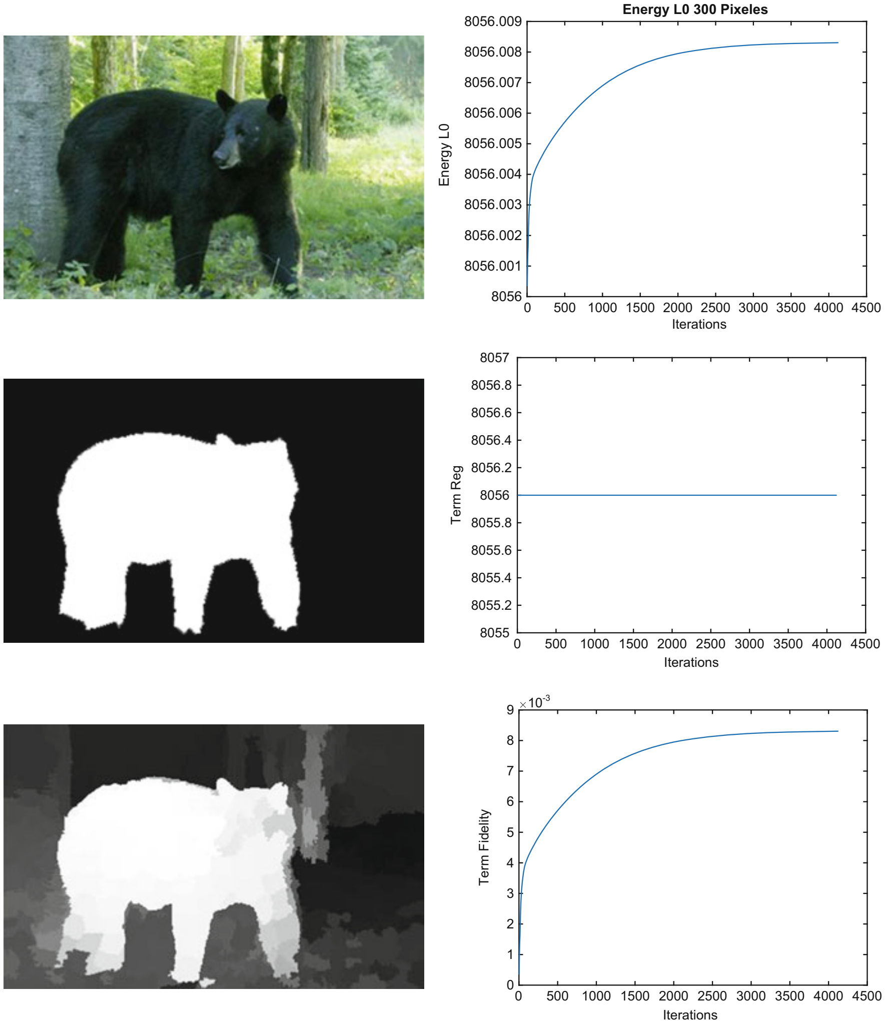


**Fig. 3** Case 1: From top to bottom and left to right: original image, ground truth, saliency map, energy curve, $L_0$ term of the energy functional and the fidelity term

value but do not converge to a minimum. The $L_0$ norm is constant and the energy is dominated by the fidelity term. Case 2 refers to convergence to a minimum. An oscillatory form is presented (see Fig. 4) and the energy is dominated by the $L_0$ norm.

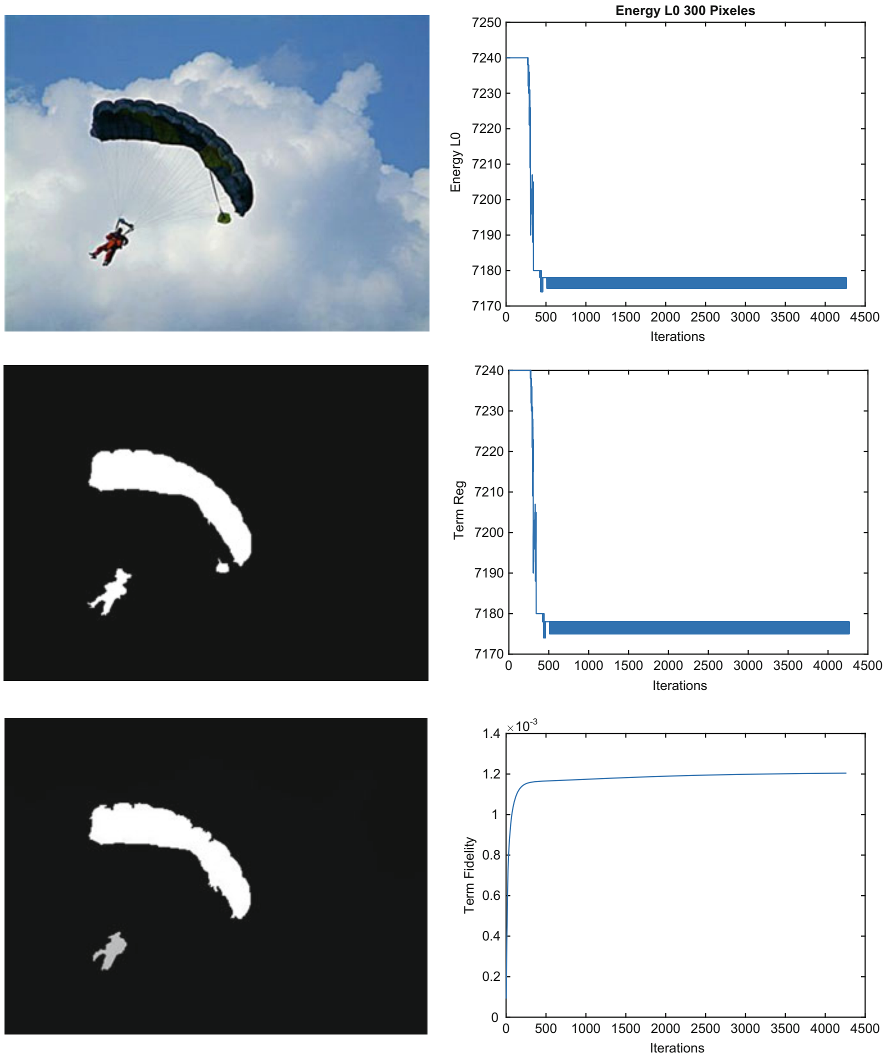


**Fig. 4** Case 2: From top to bottom and left to right: original image, ground truth, saliency map, energy curve, $L_0$ term of the energy functional and the fidelity term

## 6 Non Local Total Variation Model

As an alternative to the nonlocal $L_0$ norm, we propose the following saliency detection model based on the consideration of a non local version of the total variation suitable for weighted graphs, taken into consideration the well known properties of the total variation operator.

In order to do that, we shall first present the notion of NLTV norm of the weighted gradient $\nabla_w v$ introduced in Definition 1 (see [10]):

**Definition 2** The non local total variation norm in its discrete version can be defined as the following isotropic $L_1$ norm of the weighted graph gradient $\nabla_w v$:

$$J_{NLTV,w}(v) = \sum_{p \in G} \left( \sum_{q,\ pq \in E} w_{p,q}(v_q - v_p)^2 \right)^{1/2} .$$

The minimization problem based on the NLTV for saliency detection we shall consider is:

$$\min_v \left( \alpha \sum_{p \in G} \|\nabla_w v_p\|_2 + \frac{\lambda}{2} \sum_{p \in G} |v_p - v_p^c|^2 \right), \tag{2}$$

where

$$\|\nabla_w v_p\|_2 = \left( \sum_{q,\ pq \in E} w_{p,q}(v_q - v_p)^2 \right)^{1/2} .$$

In order to illustrate the mathematical foundations of our numerical method we rewrite our minimization problem in the general setting of the primal problem:

$$\min_v \ F(\nabla_w v) + G(v), \tag{3}$$

with

$$F(\nabla_w v) = \alpha \sum_{p \in G} \|\nabla_w v_p\|_2 \quad \text{and} \quad G(v) = \frac{\lambda}{2} \sum_{p \in G} |v_p - v_p^c|^2.$$

## *6.1 Numerical Resolution: Primal-Dual Algorithm*

In order to solve numerically the minimization problem (3), we generalize the Primal-Dual algorithm [4, 11]. We briefly review the mathematical setting. The primal-dual formulation of the nonlinear primal problem in (3) is the saddle-point problem

$$\min_v \max_d \langle \nabla_w v, d \rangle + G(v) - F^*(d) \tag{4}$$

for the primal variable $v$, the dual variable $d$ and $F^*$, the convex conjugate of $F$.

Assuming that these problems have a solution $(v, d)$, it satisfies

$$\nabla_w v \in \partial F^*(d), \qquad div_w(d) \in \partial G(v)$$

where $\partial F^*$ and $\partial G$ are the subdifferentials of $F^*$ and $G$. Introducing the resolvent operator defined through

$$\tilde{v} = (I + \tau \partial G)^{-1}(v) = \min_v \left( \frac{||v - \tilde{v}||}{2\tau} + G(v) \right)$$

the final algorithm is:

$$\begin{cases} d^{n+1} = (I + \sigma \partial F^*)^{-1}(d^n + \sigma \nabla_w v^n) \\ v^{n+1} = (I + \tau \partial G)^{-1}(v^n + \tau \, div_w d^{n+1}) \end{cases}$$

The algorithm therefore consists of an alternate minimization and maximization step, where the dual variable $d$ is updated in the maximization step and the solution $v$ is set in the minimization one. These two steps are repeated iteratively until the convergence is reached. Making explicit the above equations we have: Given the $k$-step solution $(v^k, d^k)$:

- Maximization step: For every superpixel q compute

$$d_q^{k+1} = \frac{d_q^k + \tau_d \nabla_w v_q^k}{max(1, |d_q^k + \tau_d \nabla_w v_q^k|_\infty)}.$$

Notice that $d_q^k$ is a vector of components $d_{q,p}^k$ for $p$ superpixel such that $qp \in E$.

- Minimization step: Fixed $d^{k+1}$, we compute $v^{k+1}$ for every pixel $q$:

$$v_q^{k+1} = (1 - \tau_p) v_q^k + \tau_p \left( \frac{1}{\lambda} div(d_q^{k+1}) + v_q^c \right),$$

where

$$div(d_q^{k+1}) = \sum_{p,qp \in E} (d_{q,p}^{k+1} - d_{p,q}^{k+1}) \sqrt{w_{p,q}} \tag{5}$$

The iterations are stopped when the difference between the values in two consecutive iterations is less than a fixed value $\epsilon$.

The parameters $\tau_d$ and $\tau_p$ refers to the gradient descent steps employed to solve the maximization and minimization problems associated to the dual $d$ and the primal variable $v$ respectively. A pseudo code of our primal dual algorithm for saliency is shown in Algorithm 1.

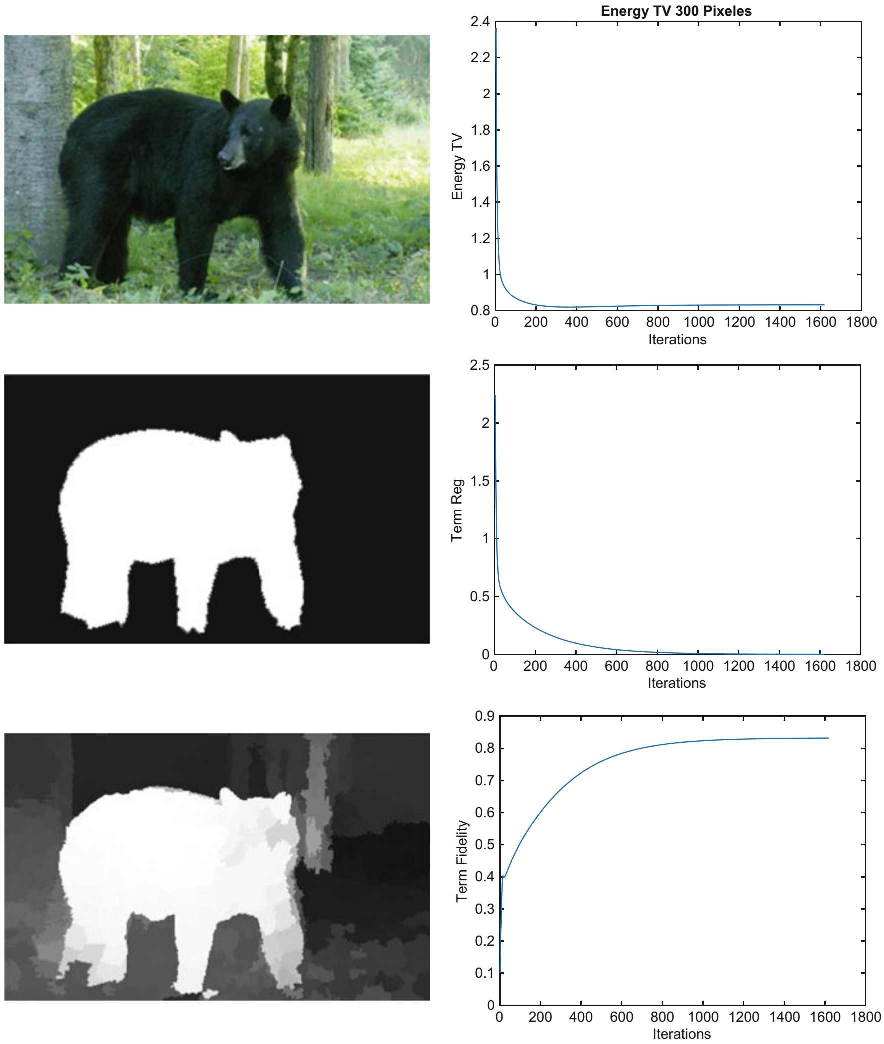


**Fig. 5** Results obtained with the NLTV model for the image consider in case 1 (Fig. 3): from top to bottom and left to right: original image, ground truth, saliency map, energy curve, the NLTV term of the energy functional and the fidelity term

In Figs. 5 and 6, we present the results obtained with our NLTV model for the examples shown in Figs. 3 and 4 regarding the $L_0$-mode. It can be seen that in this case, convergence to a minimum of the functional is achieved, and also that the convergence is faster than in the $L_0$ case.

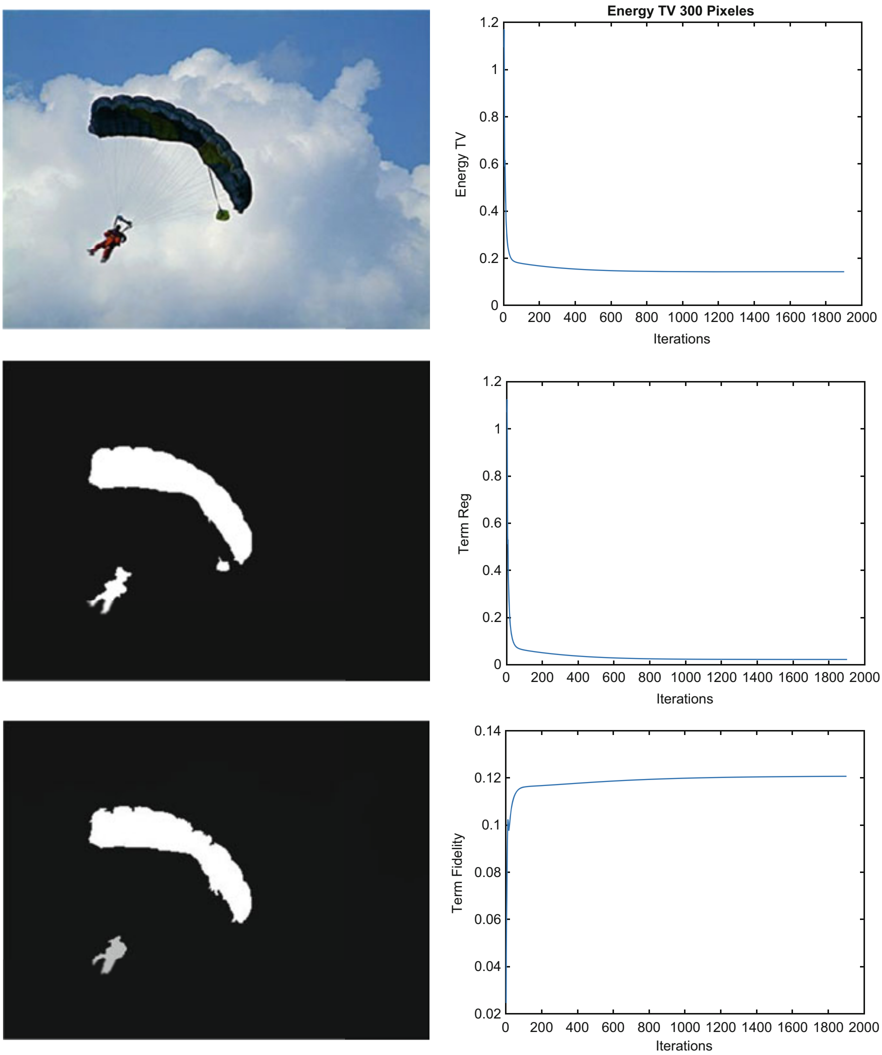


**Fig. 6** Results obtained with the NLTV model for the image consider in case 2 (Fig. 4): from top to bottom and left to right: original image, ground truth, saliency map, energy curve, the NLTV term of the energy functional and the fidelity term

**Algorithm 1** Saliency estimation on non local TV

```
procedure SALIENCYNLTV(inputImage,parameters)
    Calculate Superpixels
    wknn $\leftarrow$ Create knn graph
    controlMap $\leftarrow$ Calculate controlMap
    $v^k$ = controlMap;
    d = 0;
    repeat
        d = d + ($\tau_d$ · gradNLTV (wknn,$v^k$,NoSuperpixels) / max(1, | d + $\tau_d$ · gradNLTV (wknn,$v^k$,NoSuperpixels) $|_\infty$)) ;
        div d= (1/$\lambda$) · divNLTV(wknn,d,NoSuperpixels);
        $Prevv^k$=$v^k$;
        $v^k$ = (1 -$\tau_p$) · $v^k$ + $\tau_p$ ·( div d + controlMap);
        energy$v^k$= energyNLTV(wknn,$v^k$,controlMap ,$\lambda$);
        stopCriteria = | Prev$v^k$-$v^k$ |
        energyPrev$v^k$=energy$v^k$ ;
        $iter = iter + 1$;
    until stopCriteria $\leq$ tol
    return $v^k$;
end procedure
```

**end**

In Sect. 8, we shall show a comparison about the times of computation and also the results obtained for several measures with the $L_0$ and NLTV models.

## 7 Saliency Map Segmentation

The segmentation by fixed threshold $T_f \in [0, 255]$ is the simplest method to obtain the final saliency map of the input image. Varying $T_f$ provides also a fair methodology to compare other algorithms with the precision vs recall curves and confirm the efficiency. However, there exist methods which provide a final saliency segmentation given a saliency map such as the Saliency Binarization with Mean Shift or the Saliency Cut algorithm [6] to name a few. The Saliency Cut Algorithm is based on the GrabCut algorithm [13] and it can be initialized with the saliency map calculated with the algorithm. For this it was considered here to obtain the final saliency map (Fig. 7).

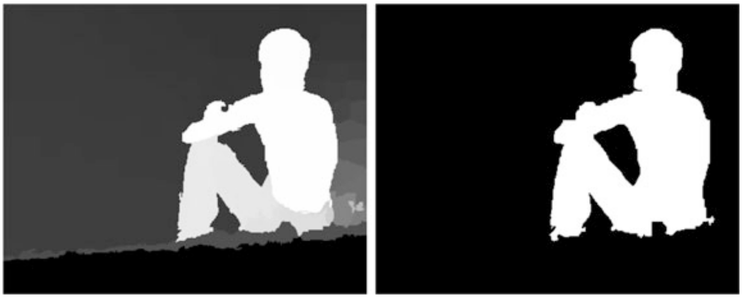


**Fig. 7** Saliency cut algorithm in an image from MSRA10K benchmark. On the left hand side our NLTV + SCALP + Bon and right hand side the result of applying SaliencyCut to this saliency map

## 8 Numerical Results

The results have been carried out on MSRA10K benchmark which has 10,000 images and each image has an unambiguous salient object. This benchmark provides the ground truth masks (salient objects) with pixel level accurately in comparison with MSRA where only the bounding box is provided. For $L_0$ algorithm, we take the parameters proposed by the authors in [15] $N = 300$ (number of superpixels), $k = 5$, $\lambda = 0.001$, $\rho = 0.0001$, $\alpha = 0.9$ and $\sigma^2 = 0.05$ for $NLTV$ algorithm $\tau_p = 0.3$, $\tau_d = 0.03$, $\lambda = 0.1$ and $N = 300$, $k = 5$, $\alpha = 0.9$ for fair comparison. There are three more parameters that have been modified (enable/disable) to see the influence in the variational methods:

- Normal: No boundaries, no location prior.
- Prior: The control map gives more importance to the center position of the superpixels in the image.
- Boundaries: Suppression of the background associating the external borders in the image and the superpixels with less value.

The naming conventions used in the experimental results are as follows:

- Variational method: NLTV and $L_0$
- Superpixel method: SLIC and SCALP
- Parameters: Loc (location prior is enabled) for the control map and Bon (Location prior is enabled as well as Boundaries) for the weights otherwise these parameters are disabled.

The results of applying the SaliencyCut algorithm to the whole benchmark MSRA10K for both saliency algorithms $L_0$ and NLTV is shown in Fig. 8. We used the precision, recall and $F_\beta$ measurements to compare our results with the ones obtained with the model proposed in [15], and illustrate the fact of including the location prior and the boundaries identification. On one hand, precision (positive predictive value) is the fraction of relevant instances among the retrieves instances.

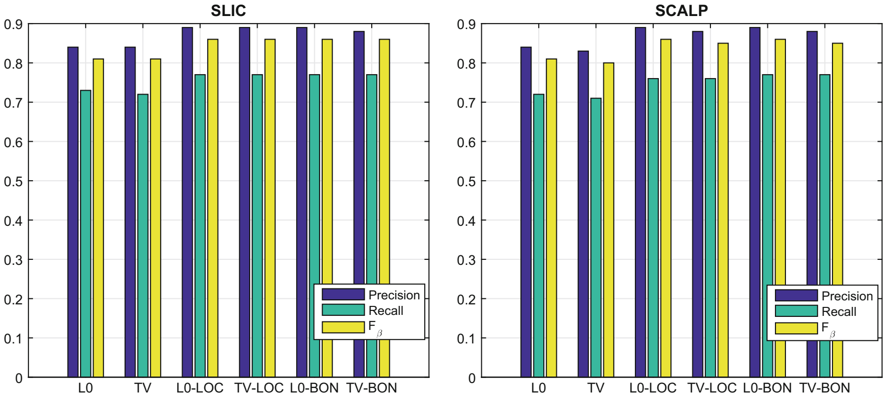


**Fig. 8** Precision, recall and F-measure for the complete data set MSRA10K using SLIC and SCALP as superpixels method and $L_0$ and NLTV as the regularization terms and Loc and Bon mean with Location prior and Boundaries background suppression enabled

On the other hand, recall (sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. The $F_\beta$ measure is given by

$$F_\beta = \frac{(1+\beta)^2 Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}, \qquad \beta = 0.3.$$

It can be seen that the results we have obtained are similar confirming the efficiency of our proposal.

The results in Fig. 9 demonstrate that the quality of the solution in both cases (case 1 and case 2, see Figs. 3, 4, 5 and 6) is good. We compare both methods with the same initial control map enabling a fair comparison. Although the case 2 presents oscillations in the energy functional, it achieves the best performance for these two images. We understand that the oscillations and the way to reach the convergence are not determinant for the quality in the solution. However, the method we proposed achieves a much better performance in computational time while keeping the same grade of quality in terms of precision and recall metrics.

Experimentation was performed on an Intel Xeon E5-1650v3, 3.5 GHz hexa-core processor, from the 2014 Intel Haswell architecture (Haswell-EP), 1.5 MB L2 cache and 15 MB L3 cache with 64 GB DDR3 RAM as a CPU platform using Microsoft Windows Server 2012R2 as operating system. this procesor can run up to 12 threads among its 6-cores simultaneously due to the Intel Hyper Threading technology (HTT). The Intel Xeon family of microprocessors belongs to the professional line instead of the more consumer oriented Intel Core family.

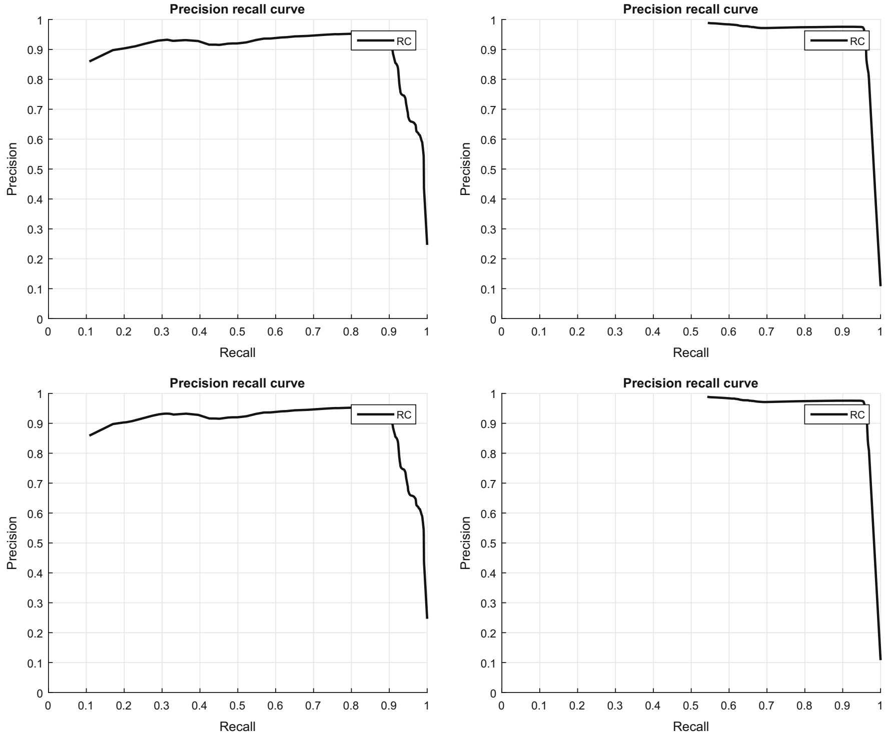


**Fig. 9** On the top of the image we present from left to right the results obtained in the $L_0$ model for the precision vs recall curves in the case 1 and case 2. On the bottom, the ones obtained with our model for the same cases

**Table 1** Computational time (in seconds) using the two variational methods for each case 1 (Bear) and case 2 (Parachute man)

| # Case | Method | Iterations | Time | Size |
|---|---|---|---|---|
| 1 | TV | 1620 | 1.38 s | [400×253] |
| | L0 | 4128 | 9.09 s | [400×253] |
| 2 | TV | 1903 | 1.45 s | [400×300] |
| | L0 | 4266 | 8.43 s | [400×300] |

As it can be seen NLTV method improves considerably both iterations and time performance

The algorithm has been implemented in C++ Visual Studio 2015 with Intel tools for compilation. No threads implementation has been used in our code. The elapsed time until convergence and the iterations in the cases 1 and 2 for $L_0$ and $TV$ are shown in Table 1.

## 9 Conclusions

In this work, we have presented a new saliency model as an alternative to the sparse gradient saliency detection model based on $L_0$ minimization proposed in [15]. An efficient primal dual variational method to obtain the saliency of an input image has also be described and implemented. A numerical comparison is presented based on the MSRA10K benchmark dataset. The results are qualitative and quantitatively comparable to [15], but the numerical resolution is faster opening the way to automatic real time saliency detection in video and multichannel images. We also included the edges and contours of the images to generate high quality superpixels using the SCALP algorithm. Some preliminary results indicate a clear improvement but the results are not conclusive in this benchmark.

Considering the model parameter we observe that the location prior parameter has more influence than the boundaries in the generation of accurate saliency binary partitions similar to the ground truth. From the results we can see that there is a clear improvement when applying the location prior (0.89 vs 0.83 when no location prior is imposed). The initial control map is also a key ingredient to obtain high quality results and machine learning techniques shall be used in future work to learn prior information about the object to be detected. This will allow to develop models tailored to specific saliency detection tasks.

The automatic segmentation has proven to be a robust method to efficiently segment the final saliency against the fixed threshold. Regarding the numerical performance the computation time is less than 0.5 s per image when the converge criteria is not that strict making this technique promising for real time systems.

In future research, we shall extend this computation to GPGPU[1] to be able to detect saliency in video in real time.

**Acknowledgements** This research was partially supported by projects TIN2015-69542-C2-1-R and MTM2014-57158-R.

## References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Suesstrunk, S.: SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. **34**, 2274–2282 (2012). http://dx.doi.org/10.1109/TPAMI.2012.120
2. Belongie, S., Carson, C., Greenspan, H., Malik, J.: Color- and texture-based image segmentation using EM and its application to content-based image retrieval. In: IEEE International Conference on Computer Vision (ICCV) (1998). http://dl.acm.org/citation.cfm?id=938978.939161
3. Belyaev, A.: On Implicit Image Derivatives and Their Applications (2012). http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.413.634

[1] General-purpose computing on graphics processing.

4. Chambolle, A., Pock, T.: A first order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**, 120–145 (2011)
5. Cheng, M., Mitra, N., Huang, X., Hu, S.: SalientShape: group saliency in image collections. Vis. Comput. **30**, 443–453 (2014). http://dx.doi.org/10.1007/s00371-013-0867-4
6. Cheng, M., Mitra, N., Huang, X., Torr, P., Hu, S.: Global contrast based salient region detection. In: IEEE TPAMI (2015). http://mmcheng.net/salobj/
7. Christopoulos, C., Skodras, A., Ebrahimi, T.: The JPEG2000 still image coding system: an overview. IEEE Trans. Consum. Electron. **46**, 1103–1127 (2000) http://ieeexplore.ieee.org/document/920468/
8. Dollár, P., Lawrence Zitnick, C.: Fast edge detection using structured forests. IEEE Trans. Pattern Anal. Mach. Intell. **37**, 1558–1570 (2015). http://dblp.uni-trier.de/rec/bib/journals/pami/DollarZ15
9. Elmoataz, A., Lezoray, O., Bougleux, S.: Nonlocal discrete regularization on weighted graphs: a framework for image and manifold processing. IEEE Trans. Image Process. **17**, 1047–1060 (2008)
10. Gilboa, G., Osher, S.: Nonlocal operators with applications to image processing. SIAM Multiscale Mod. Simul. (MMS) **7**, 1005–1028 (2008)
11. Martín, A., Garamendi, J.F., Schiavi, E.: Two efficient primal-dual algorithms for MRI rician denoising. In: Computational Modelling of Objects Represented in Images III, pp. 291–296 (2013). http://10.1201/b12753-54
12. Rémi, G., Vinh-Thong, T., Papadakis, N.: SCALP: superpixels with contour adherence using linear path. In: 23rd International Conference on Pattern Recognition (ICPR 2016), Cancún, México (2016). https://hal.archives-ouvertes.fr/hal-01349569
13. Rother, C., Kolmogorov, V., Blake, A.: GrabCut: interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. (2004). http://doi.acm.org/10.1145/1015706.1015720
14. Santella, A., Agrawala, M., DeCarlo, D., Salesin, D., Cohen, M.: Gaze-based interaction for semi-automatic photo cropping. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 771–780 (2006). http://doi.acm.org/10.1145/1124772.1124886
15. Wang, Y., Liu, R., Song, X., Zhixun, S.: Saliency detection via nonlocal $L_0$ minimization. In: Computer Vision ACCV 2014. Lecture Notes in Computer Vision, vol. 9004, pp. 521–535. Springer, New York (2014)

# Linear Elimination in Chemical Reaction Networks

**Meritxell Sáez, Elisenda Feliu, and Carsten Wiuf**

**Abstract** We consider dynamical systems arising in biochemistry and systems biology that model the evolution of the concentrations of biochemical species described by chemical reactions. These systems are typically confined to an invariant linear subspace of $\mathbb{R}^n$. The steady states of the system are solutions to a system of polynomial equations for which only non-negative solutions are of interest. Here we study the set of non-negative solutions and provide a method for simplification of this polynomial system by means of linear elimination of variables. We take a graphical approach. The interactions among the species are represented by an edge labelled graph. Subgraphs with only certain labels correspond to sets of species concentrations that can be eliminated from the steady state equations using linear algebra. To assess positivity of the eliminated variables in terms of the non-eliminated variables, a multigraph is introduced that encodes the connections between the eliminated species in the reactions. We give graphical conditions on the multigraph that ensure the eliminated variables are expressed as positive functions of the non-eliminated variables. We interpret these conditions in terms of the reaction network. The results are illustrated by examples.

**Keywords** Steady states · Elimination · Linear system · Reaction network · Noninteracting · Conservation law

## 1 Introduction

A reaction network describes the interactions among molecular species in a biochemical system. The evolution of the species concentrations is typically modelled by a system of (parametrized) polynomial ordinary differential equations (ODEs) $\dot{x} = f(x)$ in $\mathbb{R}^n_{\geq 0}$. The trajectories of this system are generally contained in invariant

M. Sáez · E. Feliu · C. Wiuf (✉)
Department of Mathematical Sciences, University of Copenhagen, København, Denmark
e-mail: meritxell@math.ku.dk; efeliu@math.ku.dk; wiuf@math.ku.dk

© Springer Nature Switzerland AG 2019

J. L. García Guirao et al. (eds.), *Recent Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 18,
https://doi.org/10.1007/978-3-030-00341-8_11

linear subspaces of dimension smaller than $n$. These subspaces are determined by the initial conditions of the system.

We are particularly interested in the steady states of the system, that is, the solutions to the steady state equations $0 = f(x)$. Since $f(x)$ is a polynomial, the steady states form an algebraic variety constrained to the non-negative orthant $\mathbb{R}^n_{\geq 0}$. In this context, two types of questions arise. On one hand, we would like to describe the set of steady states, for instance, by means of a parameterization [8, 15]. On the other hand, we would like to study the properties of the steady states in each invariant linear subspace of the system, which typically is a finite number of points.

The number of steady states in an invariant linear subspace is of biological relevance and relates to the function of the system. In particular, multistationarity, that is, the capacity of the system to have multiple steady states in an invariant linear subspace is linked to cellular decision making and memory-related on/off responses to graded input [7, 11, 16]. It is therefore of relevance to determine which invariant linear subspaces have multiple steady states and which have not.

Even for small systems, addressing the two problems above is a difficult task and case-by-case studies often appear to be the only way forward. It is therefore of interest to simplify the steady state equations (perhaps constrained to the invariant linear subspaces), for example, by eliminating variables. That is, we would like to express some variables (the eliminated) in terms of the remaining variables (the non-eliminated), by using the steady state equations and possibly also the equations describing the invariant linear subspaces. Since the system evolves in $\mathbb{R}^n_{\geq 0}$, positive parameterizations are of particular interest, that is, parameterizations such that the eliminated variables are positive provided the non-eliminated variables are positive. In this case we say that the eliminated variables are positively eliminated.

Standard elimination techniques from applied algebraic geometry such as Gröbner bases and the Shape Lemma are often not very fruitful. In some situations the expressions obtained are very long and difficult to simplify, or the positivity of the eliminated variables is hard to establish. In either case, the result is of little use. Further, the methods are computationally expensive. In this paper we identify subsets of variables that can be eliminated using only linear algebra, by exploiting the fact that these variables have degree at most one in the monomials of the system.

Generally, not much is known about the positivity of solutions to polynomial systems, let alone for linear systems [2, 13]. In a way this is surprising as many problems in applied mathematics require solutions to be non-negative (or positive). In the natural sciences, for instance, chemical concentrations, species abundances and gene frequencies are all non-negative quantities that typically are found by solving non-linear equations.

In [3], a method is presented, in which so-called noninteracting species are positively eliminated. In [14] we provide general graphical criteria for positivity of solutions to linear systems. The setting of this paper and that of [3] are special cases of [14]. We consider elimination of so-called reactant-noninteracting species and give conditions that ensure the positivity of the eliminated variables as functions of the non-eliminated variables. See also [4, 5] for related work.

Our approach starts with an undirected graph that shows the interactions of the species as reactants or products in the different reactions. This graph encodes the monomials that appear in the polynomial ODE system and it is used to identify sets of species that might be eliminated based on certain criteria. Such a set leads to a linear system in the concentrations of the species in the set, whereas the concentrations of the non-eliminated species are added to the coefficient field. We give easy-to-check conditions on the reaction network that ensure the positivity of the solution to this linear system.

If sufficiently many variables can be eliminated, then we obtain a parameterization of the positive part of the steady state variety. Such a parameterization can be used, for instance, to study the number of steady states in each invariant linear subspace [1].

In what follows we make the mathematical problem under consideration precise and present our results. We then show how our approach can be applied to a reaction network modelling an allosteric kinase [6], and to a particular type of phosphorylation networks that are abundant in cellular systems [12].

## 2 Chemical Reaction Networks

In this section we introduce the basic concepts for the study of chemical reaction networks.

A **reaction network** on a finite set $\mathcal{S} = \{S_1, \ldots, S_n\}$ is a digraph $(\mathcal{C}, \mathcal{R})$ where $\mathcal{C} \subseteq \mathbb{Z}^n_{\geq 0}$ and

1. $\mathcal{S}$ is an ordered set of elements, called *species*.
2. The nodes, elements of $\mathcal{C}$, are called *complexes*.
3. The edges, elements of $\mathcal{R}$, are called *reactions*.

We consider a complex as a linear combination of the species. The source of a reaction $r$ (that is, an edge) is called the reactant and denoted $y_r$, and the target is called the product and denoted $y'_r$. We identify the species with the unit vectors in $\mathbb{R}^n$ and the complexes with vectors in $\mathbb{R}^n$.

*Example 1* Consider the reaction network

$$S + E \rightleftharpoons ES \rightleftharpoons S_p + E$$

which models the reversible transformation of a substrate ($S$) into a modified substrate ($S_p$) catalysed by an enzyme ($E$). The conversion proceeds through the formation of an intermediate species ($ES$). This type of reaction network is found abundantly in molecular biology. There are four different species: $S$, $S_p$, $E$ and $ES$. Since the species do not have any particular order, we choose one: $\{S, S_p, E, ES\}$. The complexes are $S + E$ (or equivalently $(1, 0, 1, 0)$), $ES$ (or equivalently $(0, 0, 0, 1)$) and $S_p + E$ (or equivalently $(0, 1, 1, 0)$). There are four reactions: the reactant of the first reaction is $S + E$ and its product is $ES$. The second reaction is the reversed one, and so on.

We denote by $x_i$ the concentration of species $S_i$ and let $x = (x_1, \dots, x_n)$ be the vector of concentrations. We assume mass-action kinetics, that is, the rate of a reaction $r \in \mathcal{R}$ is

$$\kappa_r x^{y_r} = \kappa_r \prod_{i=1}^{n} x_i^{(y_r)_i},$$

where $\kappa_r > 0$ is called the *reaction rate constant*. By convention, $0^0 = 1$. The evolution of the species concentrations over time is then modelled by the following system of ODEs:

$$\dot{x} = \sum_{r \in \mathcal{R}} \kappa_r x^{y_r} (y_r' - y_r), \qquad x \in \mathbb{R}^n_{\geq 0}, \tag{1}$$

where $\dot{x} = (\dot{x}_1, \dots, \dot{x}_n)$ denotes the derivative of $x$ with respect to time. The non-negative orthant is invariant under the solutions to (1). In general, the reaction rate constants are not known, so we consider them as parameters of the system.

*Example 2* Consider the reaction network in Example 1. We let $x_1 = [S]$, $x_2 = [S_p]$, $x_3 = [E]$ and $x_4 = [ES]$. Let $\kappa_1, \dots, \kappa_4 \in \mathbb{R}_{>0}$ be the reaction rate constants. The ODEs for the species concentrations are:

$$\begin{aligned}
\dot{x}_1 &= -\kappa_1 x_1 x_3 + \kappa_2 x_4,\\
\dot{x}_2 &= \quad \kappa_3 x_4 - \kappa_4 x_2 x_3,\\
\dot{x}_3 &= -\kappa_1 x_1 x_3 + \kappa_2 x_4 + \kappa_3 x_4 - \kappa_4 x_2 x_3,\\
\dot{x}_4 &= \quad \kappa_1 x_1 x_3 - \kappa_2 x_4 - \kappa_3 x_4 + \kappa_4 x_2 x_3.
\end{aligned}$$

We note that $\dot{x}_1 + \dot{x}_2 + \dot{x}_4 = 0$ and $\dot{x}_3 + \dot{x}_4 = 0$, so the solutions to the ODE system are constrained by the linear equations: $x_1 + x_2 + x_4 = T_1$ and $x_3 + x_4 = T_2$, where $T_1, T_2 \in \mathbb{R}_{\geq 0}$ are determined by the initial condition of the system.

The **stoichiometric subspace** of a reaction network is the vector subspace of $\mathbb{R}^n$ given by

$$S = \langle y_r' - y_r \mid r \in \mathcal{R} \rangle \subseteq \mathbb{R}^n.$$

If $\omega \in S^\perp$, then it follows from (1) that $\omega \cdot \dot{x} = 0$. If $\{\omega^1, \dots, \omega^d\}$ is a basis of $S^\perp$, then the trajectory with initial concentration $x_0$ is confined to the linear subspace with equations

$$T_i = \omega^i \cdot x, \quad \text{where} \quad T_i = \omega^i \cdot x_0, \quad \text{for } i = 1, \dots, d.$$

Each of the previous equations is called a **conservation law** with total amount $T_i \in \mathbb{R}$, corresponding to $\omega^i \in S^{\perp}$. We are interested in the steady states in an arbitrary invariant linear subspace, so the total amounts are considered parameters of the system.

In Example 2, $S = \langle(-1, 0, -1, 1), (0, 1, 1, -1)\rangle$ and $S^{\perp} = \langle(1, 1, 0, 1), (0, 0, 1, 1)\rangle$. Thus the two linear constraints given in Example 2 are conservation laws and any other conservation law can be expressed as a linear combination of these two.

The **steady states** of system (1) are the solutions to the polynomial system

$$\sum_{r\in\mathcal{R}} \kappa_r x^{y_r}(y_r' - y_r) = 0 \ \text{ with } \ x \in \mathbb{R}^n_{\geq 0},$$

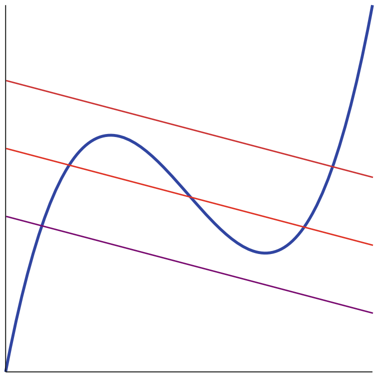
referred to as the *steady state equations*. Note that we consider only non-negative solutions. The set of solutions to the steady state equations is described by a real algebraic variety (called the steady state variety) intersected with the non-negative orthant.

In Fig. 1 we show an example of a steady state variety in $\mathbb{R}^2_{\geq 0}$ given by a cubic polynomial and such that the invariant linear subspaces are given by lines, that is, $\dim S^{\perp} = 1$. We show three different invariant linear subspaces. They are translates of each other, since they are obtained by varying the total amounts. One of the linear subspaces contains three steady states while the other two contain one steady state. By varying the total amounts there is, in this case, at least one steady state in each invariant linear subspace.

## 2.1 Elimination System

The aim of the paper is to simplify a system of equations by eliminating a subset of variables using linear algebra. That is, the eliminated variables will be given as functions of the remaining variables.

**Fig. 1** Steady state variety in $\mathbb{R}^2$ with three invariant linear subspaces



We will focus on either one of two situations: The steady state variety (defined by the steady state equations alone), or the intersection of the steady state variety with some of the conservation laws. In the first case the system of equations to be considered is given by the steady state equations alone, while in the second case it is given by the steady state equations and some of the conservation laws.

Let $\mathcal{U}$ be a subset of the species corresponding to the concentrations we wish to eliminate. Consider the system of equations given by the steady state equations for the species in $\mathcal{U}$, that is, without any conservation laws. If there is a vector $\omega \in S^{\perp}$ with support contained in $\mathcal{U}$, then the steady state equations for the species in the support of $\omega$ are linearly dependent. Therefore, removing one equation from the system does not change the set of solutions. Furthermore, this generally implies that the steady state equations for the species in $\mathcal{U}$ do not have a unique solution for the concentrations of the species in $\mathcal{U}$. This is in particular the case if the steady state equations are linear in the concentrations of the species in $\mathcal{U}$. If this is so, there might be a unique solution if some of the conservation laws are included, or if some of the species are removed from $\mathcal{U}$, thereby aiming to solve a smaller system.

These considerations lead to the following definition. The *elimination system* for $\mathcal{U}$ is the system consisting of the steady state equations for the species in $\mathcal{U}$ (or equivalently, a maximal set of independent steady state equations for the species in $\mathcal{U}$), enlarged with a maximal set of independent conservation laws that involve only species in $\mathcal{U}$ (if any). The concentrations of the species in $\mathcal{U}$ can be eliminated if the elimination system has a unique solution for fixed concentrations of the non-eliminated species (those that are not in $\mathcal{U}$).

We will say that an expression depending on parameters and/or variables is **positive** (resp. negative) if it takes positive (resp. negative) or zero values for all positive values of the parameters and/or variables.

*Example 3* The steady state equations for the reaction network in Example 1 are:

$$\begin{aligned}
0 &= -\kappa_1 x_1 x_3 + \kappa_2 x_4, \\
0 &= \quad \kappa_3 x_4 - \kappa_4 x_2 x_3, \\
0 &= -\kappa_1 x_1 x_3 + \kappa_2 x_4 + \kappa_3 x_4 - \kappa_4 x_2 x_3, \\
0 &= \quad \kappa_1 x_1 x_3 - \kappa_2 x_4 - \kappa_3 x_4 + \kappa_4 x_2 x_3.
\end{aligned} \tag{2}$$

Take $\mathcal{U} = \{S, ES\}$. Then, the elimination system for $\mathcal{U}$ is given by the first and fourth equation in (2), and its solution for $x_1$ and $x_4$ is

$$x_1 = \frac{\kappa_2 \kappa_4 x_2}{\kappa_1 \kappa_3}, \qquad x_4 = \frac{\kappa_4 x_2 x_3}{\kappa_3}. \tag{3}$$

The solution is clearly positive for positive values of the parameters and $x_2, x_3$. In this case we obtained a parameterization of the steady state variety. This parameterization further shows that the steady state variety has dimension 2, and hence, using only the steady state equations, it is not possible to express more than two variables in terms of the other variables.

The elimination system for $\mathcal{U} = \{E, ES\}$ consists of the third equation in (2) together with $x_3 + x_4 = T_1$. In this case we obtain the solution

$$x_3 = \frac{T_1(\kappa_2 + \kappa_3)}{\kappa_1 x_1 + \kappa_4 x_2 + \kappa_2 + \kappa_3},$$
$$x_4 = T_1 - x_3 = \frac{T_1(\kappa_1 x_1 + \kappa_4 x_2)}{\kappa_1 x_1 + \kappa_4 x_2 + \kappa_2 + \kappa_3}.$$

The elimination system for $\mathcal{U} = \{S, E, ES\}$ is given by the first and fourth equation in (2) together with $x_3 + x_4 = T_1$. It has solution

$$x_1 = \frac{\kappa_2\kappa_4 x_2}{\kappa_1\kappa_3}, \quad x_3 = \frac{T_1\kappa_3}{\kappa_4 x_2 + \kappa_3}, \quad x_4 = \frac{T_1\kappa_4 x_2}{\kappa_4 x_2 + \kappa_3}. \tag{4}$$

Also in this case, the solution is unique and positive for positive values of the parameters and $x_2$.

To find the full solution of the steady state equations together with the conservation laws, we make use of the only remaining non-trivial equation, $x_1+x_2+x_4 = T_2$. After substitution by (4) and simplification, we obtain the following polynomial equation in $x_2$:

$$\kappa_4\,(\kappa_1\kappa_3 + \kappa_2\kappa_4)\,x_2^2 + \kappa_3\,(\kappa_1\kappa_3 + \kappa_2\kappa_4)\,x_2 - T_2\kappa_1\kappa_3^2 = 0.$$

Since for any choice of positive parameter values there is exactly one change of sign in the list of the coefficients of the polynomial, we deduce, by Descartes' rule of signs, that there is exactly one positive real root for any choice of positive values of the parameters. Hence, using elimination we can deduce that this system has exactly one steady state in each invariant linear subspace.

Similarly, a positive parameterization could be obtained for the set $\mathcal{U} = \{S, S_p, ES\}$ using (3) and the conservation law $x_1 + x_2 + x_4 = T_2$, expressing $x_1, x_2, x_4$ in terms of $x_3$.

## 2.2 Some Further Concepts

We conclude this section with some concepts we will use in the next section.

Given a complex $\eta \in \mathcal{C} \subseteq \mathbb{Z}^n_{\geq 0}$, we call $\eta_i$ the **stoichiometric coefficient** of $S_i$ in $\eta$. We say that a complex **involves** a species $S$ if the stoichiometric coefficient of $S$ in the complex is non-zero. A reaction $r$ **involves** a species $S$ if it is involved in the reactant or product of $r$. We say that a pair of species **interact** as reactants (resp. as products) if there is a reactant (resp. product) that involves both of them. In the above example, the complex $S + E$ involves only $S$ and $E$ (both with stoichiometric coefficient equal to 1). Therefore $S$ and $E$ interact as reactants in the reaction $S + E \longrightarrow ES$ and as products in the reaction $ES \longrightarrow S + E$.

Let $S, S' \in \mathcal{S}$ be species (not necessarily different). We say that $S$ **produces** $S'$, and denote it by $S \rightsquigarrow S'$, if there exists a reaction that involves $S$ in the reactant and $S'$ in the product. Let $\mathcal{S}' \subseteq \mathcal{S}$ be a subset of species and $S, S' \in \mathcal{S}'$. We say that $S$ **ultimately produces** $S'$ via $\mathcal{S}'$ if there exist $S_{i_1}, \ldots, S_{i_k} \in \mathcal{S}'$ such that

$$S \rightsquigarrow S_{i_1} \rightsquigarrow \cdots \rightsquigarrow S_{i_k} \rightsquigarrow S'.$$

## 3 Reactant-Noninteracting Species

In this section we introduce certain subsets of species $\mathcal{U}$ whose elimination system is linear in the variables corresponding to the species in $\mathcal{U}$, as in Example 3. Subsequently, we give conditions that ensure the solution to this linear system is positive whenever the non-eliminated variables are positive.

**Definition 1** A subset $\mathcal{U} \subseteq \mathcal{S}$ is **reactant-noninteracting** if it does not contain a pair of species interacting as *reactants*, and the stoichiometric coefficient of every species in $\mathcal{U}$ in the *reactant* of any reaction is either 0 or 1.

Returning to Example 1, the sets $\mathcal{U} = \{S, ES\}, \mathcal{U} = \{E, ES\}$ and $\mathcal{U} = \{S, S_p, ES\}$ are reactant-non-interacting, but $\mathcal{U} = \{S, E, ES\}$ is not, as $S$ and $E$ are both in the reactant complex $S + E$.

Let $\mathcal{U} \subseteq \mathcal{S}$ be a reactant-noninteracting subset of species of cardinality $m$. Let $\rho \colon \mathbb{R}^n \to \mathbb{R}^m$ (resp. $\sigma \colon \mathbb{R}^n \to \mathbb{R}^{n-m}$) be the projection on the components corresponding to $\mathcal{U}$ (resp. to $\mathcal{S} \setminus \mathcal{U}$). We denote by $\mathcal{R}_{\mathcal{U}} \subseteq \mathcal{R}$ the set of reactions that involve species in $\mathcal{U}$. Let $S_{\mathcal{U}}^{\perp}$ be the subspace of $S^{\perp}$ defined by the vectors $\omega$ with $\sigma(\omega) = 0$, that is, the vectors that define conservation laws that relate the concentrations of the species in $\mathcal{U}$ only. Let $\{\omega^1, \ldots, \omega^d\}$ be a basis of $S_{\mathcal{U}}^{\perp}$ (which can be empty) giving rise to the conservation laws $\omega^i \cdot x = T_i, i = 1, \ldots, d$.

The ODE for the species $S_i \in \mathcal{U}$ is

$$\dot{x}_i = \sum_{j \mid S_j \in \mathcal{U}} \left( \sum_{\substack{r \in \mathcal{R}_{\mathcal{U}} \\ (y_r)_j \neq 0}} \kappa_r \sigma(x)^{\sigma(y_r)} (y_r' - y_r)_i \right) x_j \quad + \sum_{\substack{r \in \mathcal{R}_{\mathcal{U}} \\ \rho(y_r) = 0}} \sigma(x)^{\sigma(y_r)} (y_r')_i. \tag{5}$$

The right-hand side of this equation is linear in the concentrations of the species in $\mathcal{U}$. For simplicity, let $u = \rho(x)$ denote the vector of the concentrations of the species in $\mathcal{U}$. Then the elimination system is linear in $u$. We state it as

$$Au + b = 0, \tag{6}$$

and note that $A$ and $b$ depend on the parameters and the concentrations of the species that are not in $\mathcal{U}$. Thus their entries are in the coefficient field of the rational functions on the reaction rate constants and concentrations of the species that are not in $\mathcal{U}$. The matrix $A$ has size $(m+d) \times m$ and $b$ is a vector of size $m+d$. The first $m$ rows of $A$ correspond to the steady state equations and are given by the first term of (5). The last $d$ rows of $A$ are $\rho(\omega^i)$, $i = 1, \dots, d$. The first $d$ components of $b$ are given by the second term of (5) and the last $d$ components of $b$ are $-T_1, \dots, -T_d$.

The rank of $A$ is at most $m$, since the $d$ conservation laws imply that at least $d$ steady state equations can be written as linear combinations of the others. If $A$ has maximal rank $m$, then this system has a unique solution, which may be positive or not. To check whether $A$ has maximal rank, we remove $d$ of the rows of $A$ corresponding to redundant steady state equations. The resulting matrix is a square matrix of size $m$, and hence $A$ has maximal rank if and only if the determinant of this matrix is a nonzero polynomial in the parameters and the concentrations of the species that are not in $\mathcal{U}$.

Theorem 1 below gives a condition that guarantees positivity of the solution, in case it is unique. Before stating it, we need to introduce a few concepts.

Assume there exists a (possibly empty) basis of $S_{\mathcal{U}}^{\perp}$, $\{\omega^1, \dots, \omega^d\}$, given by vectors with *non-negative components*. We let

$$\mathcal{U}_0 \subseteq \mathcal{U}$$

denote the subset of species in $\mathcal{U}$ that are not in the support of any $\omega^i$. Let

$$\mathcal{T} \subseteq \mathcal{R}_{\mathcal{U}}$$

be the set of reactions defined as follows: $r \in \mathcal{T}$ if and only if the reactant of $r$ involves one species in $\mathcal{U}_0$ and the sum of the stoichiometric coefficients of the species in $\mathcal{U}_0$ in the product of $r$ is at least 2.

**Theorem 1 ([14])** *Let $\mathcal{U}$ be a reactant-noninteracting set of species. Assume $\{\omega^1, \dots, \omega^d\}$ is a basis of $S_{\mathcal{U}}^{\perp}$ given by vectors with non-negative components and that the matrix $A$ in the elimination system (6) has maximal rank $m$.*

*Assume that for each reaction $r \in \mathcal{T}$, there exists **at most one species** $S_i \in \mathcal{U}_0$ involved in the product of $r$ fulfilling:*

*if $S_j \in \mathcal{U}_0$ is involved in the reactant of $r$,*
*then $S_i$ ultimately produces $S_j$ via $\mathcal{U}_0$.*

*Further, assume that, if such a species $S_i$ exists, then its stoichiometric coefficient in the product of $r$ is* 1. *Then the solution to the elimination system is positive.*

As a consequence of the theorem, if the species in $\mathcal{U}$ do not interact (as reactants or products) and all stoichiometric coefficients are 0 or 1, then the elimination is positive, as already shown in [3]. Indeed, in this case the set $\mathcal{T}$ is empty. Such a set is called *noninteracting*.

*Example 4* Consider the following reaction network

$$\begin{array}{cc} X_1 + X_5 \xrightarrow{\kappa_1} X_2 + X_3 & X_3 \xrightarrow{\kappa_2} X_1 \\ X_4 + X_5 \xrightarrow{\kappa_3} X_2 + X_3 & X_2 \xrightarrow{\kappa_4} X_5 \\ 0 \underset{\kappa_6}{\overset{\kappa_5}{\rightleftharpoons}} X_1 & 0 \underset{\kappa_8}{\overset{\kappa_7}{\rightleftharpoons}} X_4 \end{array}$$

and the reactant-noninteracting set $\mathcal{U} = \{X_1, X_2, X_3, X_4\}$. We have $S^\perp = \langle(0, 1, 0, 0, 1)\rangle$ and hence $S^\perp_{\mathcal{U}} = \{0\}$. Thus we have $\mathcal{U}_0 = \mathcal{U}$. The set $\mathcal{T}$ consists of the reactions

$$X_1 + X_5 \xrightarrow{\kappa_1} X_2 + X_3 \quad \text{and} \quad X_4 + X_5 \xrightarrow{\kappa_3} X_2 + X_3.$$

For the first reaction, the species in $\mathcal{U}_0$ that is in the reactant is $X_1$ and the two species in $\mathcal{U}_0$ that are in the product are $X_2$ and $X_3$. Both have stoichiometric coefficient 1. The species $X_2$ does not ultimately produce $X_1$ via $\mathcal{U}_0$. The species $X_3$ does so due to the reaction $X_3 \xrightarrow{\kappa_2} X_1$. The condition of Theorem 1 holds for this reaction.

For the second reaction, the species in $\mathcal{U}_0$ that is in the reactant is $X_4$, and the two species in $\mathcal{U}_0$ that are in the product are $X_2$ and $X_3$. Neither $X_2$ nor $X_3$ ultimately produce $X_1$ via $\mathcal{U}_0$.

The coefficient matrix of the elimination system in this case has maximal rank. Hence, by Theorem 1, we deduce that the solution to the elimination system for the species in $\mathcal{U}$ is positive.

## 3.1 The Interaction Graph

To find variables to be eliminated, Theorem 1 tells us that a good place to start is to look for reactant-noninteracting sets of species. In order to find this type of sets, we introduce an edge labelled graph to represent the interactions among the species.

**Definition 2** The **interaction graph** for a reaction network with set of species $\mathcal{S}$ is the undirected graph with node set $\mathcal{S}$, and edge set as follows. There is

- a dotted edge connecting $S_i$ and $S_j$ for $i \neq j$ if they interact only as products,
- a dotted self-edge for $S_i$ if the stoichiometric coefficient of $S_i$ in a product is larger than 1, but in any reactant it is 0 or 1,
- a solid edge connecting $S_i$ and $S_j$ for $i \neq j$ if they interact as reactants, and
- a solid self-edge for $S_i$ if the stoichiometric coefficient of $S_i$ in a reactant is larger than 1.

A subset of species $\mathcal{U}$ is reactant-noninteracting if and only if the subgraph of the interaction graph induced by the subset of nodes $\mathcal{U}$ has no solid edges. Similarly, a set is noninteracting if and only if the induced subgraph has no edges.

If we are interested in finding maximal sets of concentrations to eliminate, then we look for node induced subgraphs with no solid edges and with the maximal number of nodes. Note, however, that if $\mathcal{U}_1$ and $\mathcal{U}_2$ are reactant-noninteracting sets, then $\mathcal{U}_1 \cup \mathcal{U}_2$ is not necessarily reactant-noninteracting. Hence, there might be several reactant-noninteracting sets that are maximal in the sense that they cannot be extended further. This is for example the case in Example 1. Here $\mathcal{U}_1 = \{E, ES\}$ and $\mathcal{U}_2 = \{S, S_p, ES\}$ are reactant-noninteracting, but their union is not.

Once we have chosen a set of species to eliminate, we assess whether the coefficient matrix $A$ has maximal rank and we assess positivity of the solution using Theorem 1. It might happen that only for some, or even for none, of the maximal sets the elimination is positive.

## 3.2 Ideas Underlying the Proof of Theorem 1

The proof of Theorem 1 is based on a multidigraph that encodes the transformations among the species in $\mathcal{U}$. The Laplacian of this multidigraph relates to the steady state equations for the concentrations of the species in $\mathcal{U}$. By means of the All Minors Matrix-Tree Theorem [10] and Cramer's rule, the solution to the elimination system can be written as a rational function whose numerator and denominator are linear combinations of the labels of certain spanning forests of the multidigraph. With that, we find conditions on the multidigraph that ensure that all negative labels cancel in the expressions, thereby providing positivity of the solutions to the elimination system. Finally, we translate these conditions on the multidigraph back into conditions on the reaction network, (as given in Theorem 1).

The main tool for the study of the positivity of the solution to the elimination system is the following multidigraph.

**Definition 3** Let $\mathcal{U}$ be a reactant-noninteracting set of species. We define the labeled **multidigraph** $\mathcal{G}_{\mathcal{U}}$ with node set $\mathcal{U} \cup \{*\}$ and edge set given, for every $r \in \mathcal{R}_{\mathcal{U}}$, by the edges

$$
\begin{aligned}
&S_i \xrightarrow{(y'_r)_j \kappa_r \sigma(x)^{\sigma(y_r)}} S_j && \text{if } (y_r)_i (y'_r)_j \neq 0,\\
&S_i \xrightarrow{\kappa_r \sigma(x)^{\sigma(y_r)}} * && \text{if } (y_r)_i = 1 \text{ and } \rho(y'_r) = 0,\\
&* \xrightarrow{(y'_r)_j \kappa_r \sigma(x)^{\sigma(y_r)}} S_j && \text{if } \rho(y_r) = 0 \text{ and } (y'_r)_j \neq 0,\\
&S_i \xrightarrow{-\lambda_r \kappa_r \sigma(x)^{\sigma(y_r)}} * && \text{if } (y_r)_i = 1 \text{ and } \lambda_r = \textstyle\sum_{j | S_j \in \mathcal{U}} (y'_r)_j - 1 > 0.
\end{aligned}
$$

By construction, each edge in $\mathcal{G}_{\mathcal{U}}$ corresponds to a reaction in $\mathcal{R}_{\mathcal{U}}$, and each reaction in $\mathcal{R}_{\mathcal{U}}$ corresponds to at least one edge in $\mathcal{G}_{\mathcal{U}}$, but there can be more. Only edges with target $*$ might have negative label. For every edge with negative label, any other edge corresponding to the same reaction is of the type $S_i \longrightarrow S_j$. Note that if a species is involved in the reactant as well as the product of a reaction, it gives a self-edge in the graph.

If we write the steady state equations for the concentrations of the species in $\mathcal{U}$ as $0 = \widehat{A}u + \widehat{b}$, then the Laplacian of $\mathcal{G}_{\mathcal{U}}$ is in block form,

$$L = \left(\begin{array}{c|c} \widehat{A} & \widehat{b} \\ \hline \cdots & \cdot \end{array}\right).$$

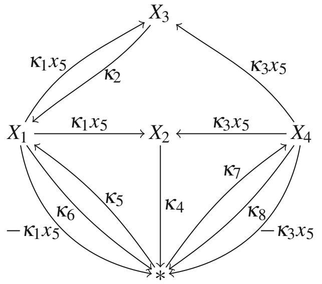
Using Cramer's rule and the Matrix-Tree theorem, we can write the solution to the elimination system as a rational function in the labels of certain spanning forests of $\mathcal{G}_{\mathcal{U}}$, the nonzero entries of the vectors $\omega^i$ and the total amounts $T_i$ for $i = 1, \dots, d$. The only negative terms in these expressions come from the edges in $\mathcal{G}_{\mathcal{U}}$ with negative labels. Theorem 1 is a consequence of the following result.

**Theorem 2 ([14])** *Consider the situation above. Assume that for every edge in $\mathcal{G}_{\mathcal{U}}$ with negative label, at most one other edge corresponding to the same reaction is contained in a directed cycle that does not contain $*$. Further, assume that, if such an edge exists, then the target of the edge has stoichiometric coefficient* 1 *in the product of the corresponding reaction. Then the solution to the elimination system is positive.*

As shown in [14], the condition of Theorem 2 is automatically fulfilled for edges with negative label and source node not in $\mathcal{U}_0$. This is due to the particular structure of the multidigraph imposed by $S_{\mathcal{U}}^{\perp}$ and its basis with vectors having non-negative components. Thus, it is sufficient to check the condition for the edges with negative label and source node in $\mathcal{U}_0$.

An edge with source (resp. target) node $*$ and positive label corresponds to a reaction that does not involve any species in $\mathcal{U}$ in the reactant (resp. the product). From this it follows that a species $S_i \in \mathcal{U}$ ultimately produces $S_j \in \mathcal{U}$ via $\mathcal{U}$ if and only if there is a path in $\mathcal{G}_{\mathcal{U}}$ from the node $S_i$ to the node $S_j$ that does not contain the node $*$. A self-edge is trivially a cycle that does not contain $*$. Additionally, an edge with source node in $\mathcal{U}_0$ and negative label corresponds to a reaction in the set $\mathcal{T}$. These two observations show that the condition on the edges in Theorem 2 is equivalent to the condition on the reactions of the network in Theorem 1.

*Example 5* To illustrate how we check the conditions in Theorem 2, we consider Example 4 again. For the reactant-noninteracting set $\mathcal{U} = \{X_1, X_2, X_3, X_4\}$, the graph $\mathcal{G}_{\mathcal{U}}$ is:



We show that the condition on the edges in Theorem 2 for the graph $\mathcal{G}_{\mathcal{U}}$ holds. There are two edges with negative label:

$$X_1 \xrightarrow{-\kappa_1 x_5} * \quad \text{and} \quad X_4 \xrightarrow{-\kappa_3 x_5} *.$$

The first edge corresponds to the reaction with reaction rate constant $\kappa_1$. There are two other edges corresponding to this reaction (those with label $\kappa_1 x_5$). Any cycle involving $X_1 \xrightarrow{\kappa_1 x_5} X_2$ involves $*$ as well. The cycle

$$X_1 \xrightarrow{\kappa_1 x_5} X_3 \xrightarrow{\kappa_2} X_1$$

shows that the other edge with label $\kappa_1 x_5$ is contained in a cycle that does not involve $*$. Since the stoichiometric coefficient of $X_3$ in the product of the reaction $X_1 + X_5 \xrightarrow{\kappa_1} X_2 + X_3$ is equal to one, the condition in Theorem 2 is satisfied for $X_1 \xrightarrow{-\kappa_1 x_5} *$.

The second edge with negative label corresponds to the reaction with reaction rate constant $\kappa_3$. The two edges with label $\kappa_3 x_5$ correspond to this reaction. Any cycle involving one of these edges contains also $*$. Therefore, the condition in Theorem 2 is satisfied for $X_4 \xrightarrow{-\kappa_3 x_5} *$.

# 4 Examples

## *4.1 Allosteric Kinase*

We consider a reaction network consisting of an allosteric kinase for one substrate [6]. This type of reaction network has been made famous by Monod, Wyman and Changeux in 1965. They proposed a model of this type to explain a number of puzzling experimental observations in molecular biology [9].

The kinase presents two possible conformations: $K_r$ (relaxed) and $K_t$ (tense). Each of these conformations catalyses the transformation of a substrate $S$ into its phosphorylated form $S_p$. The network has the following reactions:
Phosphorylation of $S$:

$$K_r + S \underset{\kappa_2}{\overset{\kappa_1}{\rightleftharpoons}} K_rS \xrightarrow{\kappa_3} K_r + S_p$$

$$K_t + S \underset{\kappa_5}{\overset{\kappa_4}{\rightleftharpoons}} K_tS \xrightarrow{\kappa_6} K_t + S_p$$

Dephosphorylation of $S_p$: $\quad S_p \xrightarrow{\kappa_7} S.$
Conformational change: $\quad K_r \underset{\kappa_9}{\overset{\kappa_8}{\rightleftharpoons}} K_t \quad K_rS \underset{\kappa_{11}}{\overset{\kappa_{10}}{\rightleftharpoons}} K_tS.$

We denote the concentrations of the different species as follows:

$$x_1 = [K_r], \qquad x_2 = [K_t], \qquad x_3 = [K_rS], \qquad x_4 = [K_tS],$$
$$x_5 = [S], \qquad x_6 = [S_p].$$

Under the law of mass-action, the dynamics of the concentrations over time is modelled by the following system of ODEs:

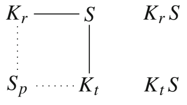
$$\begin{aligned}
\dot{x}_1 &= -\kappa_1 x_1 x_5 + (\kappa_2 + \kappa_3)x_3 - \kappa_8 x_1 + \kappa_9 x_2,\\
\dot{x}_2 &= -\kappa_4 x_2 x_5 + (\kappa_5 + \kappa_6)x_4 + \kappa_8 x_1 - \kappa_9 x_2,\\
\dot{x}_3 &= \quad \kappa_1 x_1 x_5 - (\kappa_2 + \kappa_3)x_3 - \kappa_{10} x_3 + \kappa_{11} x_4,\\
\dot{x}_4 &= \quad \kappa_4 x_2 x_5 - (\kappa_5 + \kappa_6)x_4 + \kappa_{10} x_3 - \kappa_{11} x_4,\\
\dot{x}_5 &= -\kappa_1 x_1 x_5 + \kappa_2 x_3 - \kappa_4 x_2 x_5 + \kappa_5 x_4 + \kappa_7 x_6,\\
\dot{x}_6 &= \quad \kappa_3 x_3 + \kappa_6 x_4 - \kappa_7 x_6.
\end{aligned}$$

The dynamics of the ODE system is confined to the linear subspace defined by the following two conservation laws:

$$x_1 + x_2 + x_3 + x_4 = T_1 \quad \text{and} \quad x_3 + x_4 + x_5 + x_6 = T_2,$$

where $T_1, T_2 > 0$ are positive total amounts of the kinase and the substrate, respectively.

The *interaction graph* for this network is:



A maximal reactant-noninteracting set is

$$\mathcal{U} = \{K_r, K_t, K_rS, K_tS, S_p\}$$

as there are no solid edges in the subgraph induced by $\mathcal{U}$. There is one conservation law among the species in $\mathcal{U}$, namely $x_1 + x_2 + x_3 + x_4 = T_1$. The elimination system becomes

$$\begin{aligned}
0 &= -(\kappa_1 x_5 + \kappa_8)\boldsymbol{x_1} + \kappa_9 \boldsymbol{x_2} + \kappa_2 \boldsymbol{x_3},\\
0 &= \kappa_8 \boldsymbol{x_1} - (\kappa_4 x_5 + \kappa_9)\boldsymbol{x_2} + (\kappa_5 + \kappa_6)\boldsymbol{x_4},\\
0 &= \kappa_1 x_5 \boldsymbol{x_1} - (\kappa_2 + \kappa_3 + \kappa_{10})\boldsymbol{x_3} + \kappa_{11}\boldsymbol{x_4},\\
0 &= \kappa_4 x_5 \boldsymbol{x_2} + \kappa_{10}\boldsymbol{x_3} - (\kappa_5 + \kappa_6 + \kappa_{11})\boldsymbol{x_4},\\
0 &= \kappa_3 \boldsymbol{x_3} + \kappa_6 \boldsymbol{x_4} - \kappa_7 \boldsymbol{x_6},\\
0 &= \boldsymbol{x_1} + \boldsymbol{x_2} + \boldsymbol{x_3} + \boldsymbol{x_4} - T_1.
\end{aligned}$$

The rank of the matrix $A$ is 5. This is checked by considering the coefficient matrix of the system where the first equation of the elimination system is removed. The determinant of this matrix is a nonzero polynomial in $\kappa_1, \dots, \kappa_{10}$ and $x_5$. Thus the elimination system has a unique solution.

In this case, $\mathcal{U}_0 = \{S_p\}$ and the set $\mathcal{T}$ is empty since there is no reaction where two species in $\mathcal{U}_0$ interact as products and the stoichiometric coefficient of $S_p$ in all complexes is 0 or 1. Therefore, the unique solution of the elimination system is positive by Theorem 1.

As in [6] one can then substitute the solution to the elimination system into the conservation law $x_3 + x_4 + x_5 + x_6 = T_2$. After simplifying the obtained expressions and removing the denominator, we obtain that the non-negative solutions to the steady state equations together with the two conservation laws are in one to one correspondence with the non-negative solutions of a degree three polynomial in $x_5$. The coefficients of the polynomial depend on the reaction rate constants and the total amounts.

By means of this polynomial, we can deduce that for appropriate values of the parameters there are linear invariant subspaces with three different non-negative steady states (see [6] for the details). There is thus multistationarity.

### 4.2 Phosphorylation and Dephosphorylation

The second example is an example of a cellular signalling system that is found ubiquitously in living organisms [12]:

$$\begin{gathered}
E + S \underset{\kappa_2}{\overset{\kappa_1}{\rightleftharpoons}} ES \xrightarrow{\kappa_3} E + S_p \\
E + S_p \underset{\kappa_5}{\overset{\kappa_4}{\rightleftharpoons}} ES_p \xrightarrow{\kappa_6} E + S \\
0 \underset{\kappa_8}{\overset{\kappa_7}{\rightleftharpoons}} S \qquad 0 \underset{\kappa_{10}}{\overset{\kappa_9}{\rightleftharpoons}} E.
\end{gathered}$$

A substrate $S$ is modified by phosphorylation into $S_p$ in a similar way to the network of Example 1. The reverse process of transforming $S_p$ into $S$ makes use of the same catalyst ($E$) as the forward process, but by means of a different intermediate molecule. Additionally, both substrate and catalyst might be regulated by the environment, which is modelled by the bottom two reactions.

There are no conservation laws. We denote $x_1 = [S]$, $x_2 = [S_p]$, $x_3 = [E]$, $x_4 = [ES]$ and $x_5 = [ES_p]$. The interaction graph is

$$S \text{———} E \text{———} S_p \qquad ES \qquad ES_p.$$

We deduce that a maximal reactant-noninteracting set is $\mathcal{U} = \{S, S_p, ES, ES_p\}$. In fact, it is noninteracting as the subgraph induced by $\mathcal{U}$ has no edges. The coefficient matrix of the elimination system has maximal rank 4. Hence, we directly deduce that the elimination is positive. Indeed, the solution is

$$x_1 = \frac{\kappa_7}{\kappa_8}, \qquad x_2 = \frac{\kappa_1\kappa_3\left(\kappa_5 + \kappa_6\right)\kappa_7}{\left(\kappa_2 + \kappa_3\right)\kappa_4\kappa_6\kappa_8},$$

$$x_4 = \frac{\kappa_1\kappa_7 x_3}{\left(\kappa_2 + \kappa_3\right)\kappa_8}, \qquad x_5 = \frac{\kappa_1\kappa_3\kappa_7 x_3}{\left(\kappa_2 + \kappa_3\right)\kappa_6\kappa_8}.$$

If we substitute these expressions in the equation $\dot{x}_3 = 0$ we obtain

$$-\kappa_{10}x_3 + \kappa_9 = 0.$$

Solving this equation leads to the conclusion that there is a single steady state for the reaction network:

$$x_1 = \frac{\kappa_7}{\kappa_8}, \quad x_2 = \frac{\kappa_1\kappa_3\left(\kappa_5 + \kappa_6\right)\kappa_7}{\left(\kappa_2 + \kappa_3\right)\kappa_4\kappa_6\kappa_8}, \quad x_3 = \frac{\kappa_9}{\kappa_{10}},$$

$$x_4 = \frac{\kappa_1\kappa_7\kappa_9}{\left(\kappa_2 + \kappa_3\right)\kappa_8\kappa_{10}}, \quad x_5 = \frac{\kappa_1\kappa_3\kappa_7\kappa_9}{\left(\kappa_2 + \kappa_3\right)\kappa_6\kappa_8\kappa_{10}}.$$

In this case, we have found the unique steady state of the system by linearly eliminating the concentrations of all species in two steps.

**Acknowledgements** This research was supported by the Lundbeck Foundation (Denmark) and the Danish Research Council.

## References

1. Conradi, C., Feliu, E., Mincheva, M., Wiuf, C.: Identifying parameter regions for multistationarity. PLoS Comput. Biol. **13** (2017). Available at arXiv:1608.03993
2. Farina, L., Rinaldi, S.: Positive Linear Systems: Theory and Applications. Series on Pure and Applied Mathematics. Wiley-Interscience, New York (2000)
3. Feliu, E., Wiuf, C.: Variable elimination in chemical reaction networks with mass-action kinetics. SIAM J. Appl. Math. **72**, 959–981 (2012)
4. Feliu, E., Wiuf, C.: Variable elimination in post-translational modification reaction networks with mass-action kinetics. J. Math. Biol. **66**, 281–310 (2013)
5. Feliu, E., Wiuf, C.: Simplifying biochemical models with intermediate species. J. R. Soc. Interface **10** (2013). https://doi.org/10.1098/rsif.2013.0484
6. Feng, S., Sáez, M., Wiuf, C., Feliu, E., Soyer, O.S.: Core signalling motif displaying multistability through multi-state enzymes. J. R. Soc. Interface **13** (2013). https://doi.org/10.1098/rsif.2016.0524
7. Laurent, M., Kellershohn, N.: Multistability: a major means of differentiation and evolution in biological systems. Trends Biochem. Sci. **24**, 418–422 (1999)
8. Millan, M.P., Dickenstein, A., Shiu, A., Conradi, C.: Chemical reaction systems with toric steady states. Bull. Math. Biol. **74**, 1027–1065 (2012)
9. Monod, J., Wyman, J., Changeux, J.P.: On the nature of allosteric transitions: a plausible model. J. Mol. Biol. **12**, 88–118 (1965)
10. Moon, J.: Some determinant expansions and the matrix-tree theorem. Discrete Math. **124**, 163–171 (1994)
11. Ozbudak, E.M., Thattai, M., Lim, H.N, Shraiman, B.I., Van Oudenaarden, A.: Multistability in the lactose utilization network of Escherichia coli. Nature **427**, 737–740 (2004)
12. Qian, H., Beard, D.A.: Metabolic futile cycles and their functions: a systems analysis of energy and control. IEEE Proc. Syst. Biol. **153**, 192–200 (2006)
13. Roman, S.: Positive solutions to linear systems: convexity and separation. In: Advanced Linear Algebra. Graduate Texts in Mathematica, vol. 135. Springer, New York (2005)
14. Sáez, M., Feliu, E., Wiuf, C.: Graphical criteria for positive solutions to linear systems. Linear Algebra Appl. **552**, 166–193 (2018). Available at arXiv:1709.01700
15. Thomson, M., Gunawardena, J.: The rational parameterization theorem for multisite post-translational modification systems. J. Theor. Biol. **261**, 626–636 (2009)
16. Xiong, W., Ferrell Jr, J.E.: A positive-feedback-based bistable ‘memory module’ that governs a cell fate decision. Nature **426**, 460–465 (2003)

# Minimal Set of Generators of Controllability Space for Singular Linear Systems

María Isabel García-Planas

**Abstract** In recent years, there has been increasing the interest in the descriptive analysis of singular (also called generalized) systems in the form $E\dot{x}(t) = Ax(t)$ because they play important roles in mathematical modelling problems permeating many aspects of daily life arising in a wide range of applications. Considerable advances have been obtained in the description of their structural and dynamical properties. However, much less effort has been devoted to studying the exact controllability measuring the minimum set of controls that are needed to steer the whole system $E\dot{x}(t) = Ax(t)$ toward any desired state. In this paper, we focus the study on the obtention of the set of all $B$ making the system $E\dot{x}(t) = Ax(t) + Bu(t)$ controllable.

**Keywords** Controllability · Exact controllability · Eigenvalues · Eigenvectors · Singular linear systems

## 1 Introduction

In these recent years, the study of the control of complex networks with linear dynamics has gained importance in both science and engineering. Controllability of a dynamical system has being largely studied by several authors and under many different points of view, (see [1, 3, 4, 9, 12, 14, 17] and [7], for example). Between different aspects in which we can study the controllability we have the notion of structural controllability that has been proposed by Lin [15] as a framework for studying the controllability properties of directed complex networks where the dynamics of the system is governed by a linear system: $\dot{x}(t) = Ax(t) + Bu(t)$ usually the matrix $A$ of the system is linked to the adjacency matrix of the network, $x(t)$ is a time dependent vector of the state variables of the nodes, $u(t)$ is the vector

M. I. García-Planas (✉)
Universitat Politècnica de Catalunya, Barcelona, Spain
e-mail: maria.isabel.garcia@upc.edu

© Springer Nature Switzerland AG 2019

J. L. García Guirao et al. (eds.), *Recent Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 18,
https://doi.org/10.1007/978-3-030-00341-8_12

of input signals, and $B$ which defines how the input signals are connected to the nodes of the network and it is the called input matrix. Structurally controllable means that there exists a matrix $\bar{A}$ in which is not allowed to contain a non-zero entry when the corresponding entry in A is zero such that the network can be driven from any initial state to any final state by appropriately choosing the input signals $u(t)$. Recent studies over the structural controllability can be found on [16].

Another important aspect of control is the notion of output controllability that describes the ability of an external data to move the output from any initial condition to any final in a finite time. Some results about can be found in [9].

In this article, we analyze the exact controllability concept that following definition given in [19], it is based on the maximum multiplicity to identify the minimum set of driver nodes required to achieve full control of networks with arbitrary structures and link-weight distributions. We were focusing the study on the obtention of the set of all matrices $B$ making the system $E\dot{x}(t) = Ax(t) + Bu(t)$ exact controllable. We have included several examples in order to make the work easier readable, and it is complete with an example in the case of an undirected network.

## 2 Notations

1. $M_{n\times m}(\mathbb{C})$: set of $n$-rows and $m$-columns matrices with entries in the set of complex numbers
2. $M_n(\mathbb{C})$: set of $n$-order square matrices with entries in the set of complex numbers
3. $Gl(n; \mathbb{C})$: set of invertible matrices with entries in the set of complex numbers
4. $\mathbb{C}^i = M_{i\times 1}(\mathbb{C})$: the set of column vectors of dimension $i$
5. The product of the matrices $X \in M_{n\times m}(\mathbb{C})$ and $Y \in M_{m\times p}(\mathbb{C})$ is denoted by $XY \in M_{n\times p}(\mathbb{C})$
6. The block column matrix of $X \in M_{n\times m}(\mathbb{C})$ and $Y \in M_{n\times p}(\mathbb{C})$ is denoted by $\begin{pmatrix} X & Y \end{pmatrix} \in M_{n\times(m+p)}(\mathbb{C})$
7. The diagonal block matrix of $X \in M_{n\times m}(\mathbb{C})$ and $Y \in M_{p\times r}(\mathbb{C})$ is denoted by $\operatorname{diag}(X, Y) = \begin{pmatrix} X & 0 \\ 0 & Y \end{pmatrix} = \begin{pmatrix} X & \\ & Y \end{pmatrix} \in M_{(n+p)\times(m+r)}(\mathbb{C})$

## 3 Preliminaries

It is well known that many complex networks have linear dynamics and they have a state space representation for its description:

$$E\dot{x}(t) = Ax(t) + Bu(t) \tag{1}$$

where $E, A \in M_n(\mathbb{C})$, $B \in M_{n\times m}(\mathbb{C})$ with $n, m \geq 1$, $x(t) \in \mathbb{C}^n$ is a time dependent vector column of the state variables and $u(t) \in \mathbb{C}^m$ is the vector column of input signals.

For simplicity, from now on we will write the system (1) as the triple of matrices $(E, A, B)$. If matrix $B$ does not appears in the system it is called homogeneous singular system and we will write simply as a pair of matrices $(E, A)$.

To obtain qualitative properties of the system we can make basis change into the basis space $x = Px_1$ as well premultiply the system by an invertible matrix $Q$. Operations that can mathematically be expressed as follows $(E_1, A_1) = (QEP, QAP)$ and permit us consider the homogeneous singular linear system $(E, A)$ as a matrix pencil $sE - A$ that will be called pencil associate to the system.

We will consider the case where the matrix pencil $sE-A$ is regular, as it is usual, in order to ensure that the system has a unique solution for any sinput function $u(t)$ having as many derivatives as needed [5]. Under this regularity assumption, there exist invertible matrices $Q, P \in Gl(n; \mathbb{C})$ such that $\bar{E} = QEP = \operatorname{diag}(I_r, N)$, $\bar{A} = QAP = \operatorname{diag}(J, I_{n-r})$, $J$ a Jordan matrix and $N$ a nilpotent matrix and we will say that the pencil is it is canonical reduced form.

So, considering $x(t) = P\bar{x}(t)$ and premultiplying the system (1) by $Q$ and calling $\bar{B} = QB$, the system can be written as $\bar{E}\dot{\bar{x}} = \bar{A}\bar{x}(t) + \bar{B}u(t)$, that is to say:

$$\begin{pmatrix} I_r & 0 \\ 0 & N \end{pmatrix} \begin{pmatrix} \dot{\bar{x}}_1(t) \\ \dot{\bar{x}}_2(t) \end{pmatrix} = \begin{pmatrix} J & 0 \\ 0 & I_{n-r} \end{pmatrix} \begin{pmatrix} \bar{x}_1(t) \\ \bar{x}_2(t) \end{pmatrix} + \begin{pmatrix} \bar{B}_1 \\ \bar{B}_2 \end{pmatrix} u(t) \tag{2}$$

In general we say that two systems $(E, A, B)$ and $(\bar{E}, \bar{A}, \bar{B})$ are equivalent if and only if, there exist invertible matrices $P$ and $Q$ such that $(\bar{E}, \bar{A}, \bar{B}) = (QEP, QAP, QB)$. This equivalence relation corresponds with strict equivalence of the pencil $s\begin{pmatrix} E & 0 \end{pmatrix} - \begin{pmatrix} A & B \end{pmatrix} = \begin{pmatrix} sE - A & B \end{pmatrix}$. So the collection of invariants of the pencil are the invariants for the system.

In particular, for the homogeneous systems $E\dot{x}(t) = Ax(t)$ the generalized eigenvalues of the system are the generalized eigenvalues of the pencil $sE - A$.

**Definition 1** $\lambda_0$ is a generalized eigenvalue of the system, if and only if $\operatorname{rank}(\lambda_0 E - A) < n$.

It is easy to observe that the generalized eigenvalues of $sE-A$ are the eigenvalues of the matrix $J$ in the reduced form (2):

$$\begin{aligned} \operatorname{rank}(\lambda_0 E - A) &= \operatorname{rank}(\lambda_0 Q^1 \bar{E} P^{-1} - Q^{-1}\bar{A}P^{-1}) \\ &= \operatorname{rank} Q^{-1}(\lambda_0 \bar{E} - \bar{A})P^{-1} = \operatorname{rank}(\lambda_0 \bar{E} - \bar{A}) \end{aligned}$$

and

$$\operatorname{rank}(\lambda_0 \bar{E} - \bar{A}) = \operatorname{rank} \begin{pmatrix} \lambda_0 I_r - J & \\ & \lambda_0 N - I_{n-r} \end{pmatrix} < n \quad \text{if and only if} \quad \operatorname{rank}(\lambda_0 I_r - J) < r.$$

In a more general form we have the following proposition.

**Proposition 1** *Let $(E, A)$ and $(\bar{E}, \bar{A})$ be two equivalent singular homogeneous systems, $\lambda_0$ is an eigenvalue of $(E, A)$ if and only if it is an eigenvalue of $(\bar{E}, \bar{A}) = (QEP, QAP)$.*

If $\lambda_0$ is a generalized eigenvalue of $(E, A)$, then there exists a vector $0 \neq w_0$ such that $(\lambda_0 E - A)w_0 = 0$.

**Definition 2** This vector is called the generalized eigenvector associated to $\lambda_0$.

**Proposition 2** *Let $(E, A)$ and $(\bar{E}, \bar{A})$ be two equivalent systems, $w_0$ is an eigenvector of $(E, A)$ if and only if $P^{-1}w_0$ is an eigenvector of $(\bar{E}, \bar{A}) = (QEP, QAP)$.*

*Proof* Let $(E, A)$ and $(\bar{E}, \bar{A}) = (QEP, QAP)$ two equivalent systems.

$(\lambda_0 E - A)w_0 = 0$ if and only if $(\lambda_0 Q^{-1}\bar{E}P^{-1} - Q^{-1}\bar{A}P^{-1})w_0 = 0$, equivalently if and only if $Q^{-1}(\lambda_0\bar{E} - \bar{A})P^{-1}w_0 = 0$, that is to say, if and only if $(\lambda_0\bar{E} - \bar{A})P^{-1}w_0 = 0$. □

In the particular case where the equivalent system is in the reduced form we have the following corollary.

**Corollary 1** *The corresponding eigenvector $v_0 = P^{-1}w_0$ to the eigenvalue $\lambda_0$, is in the form $v_0 = \begin{pmatrix} v_0^1 \\ 0 \end{pmatrix}$ with $v_0^1 \in \mathbb{C}^r$ and $0 \in \mathbb{C}^{n-r}$ and $v_0^1$ is an eigenvector of $J$.*

*Proof* Let $v_0 = \begin{pmatrix} v_0^1 \\ v_0^2 \end{pmatrix}$ with $v_0^1 \in \mathbb{C}^r$ and $v_0^2 \in \mathbb{C}^{n-r}$ be an eigenvector of a pair of matrices in its canonical reduced form corresponding to eigenvalue $\lambda_0$. Then,

$$\left( \begin{pmatrix} \lambda_0 I_r & \\ & \lambda_0 N \end{pmatrix} - \begin{pmatrix} J & \\ & I_{n-r} \end{pmatrix} \right) \begin{pmatrix} v_0^1 \\ v_0^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

So, clearly $v_0^2 = 0$ and $Jv_0^1 = \lambda_0 v_0^1$. □

The following result is important.

**Proposition 3** *Eigenvectors corresponding to different eigenvalues are independent.*

*Proof* Let $w_1, \ldots, w_\ell$ $\ell$ eigenvectors corresponding to $\lambda_1, \ldots, \lambda_\ell$ with $\lambda_i \neq \lambda_j$ for all $i \neq j$ and consider $\sum_{i=1}^{\ell} \alpha_i w_i$.

Then $\sum_{i=1}^{\ell} \alpha_i P^{-1} w_i = 0$ and $\bar{A}^k \sum_{i=1}^{\ell} \alpha_i P^{-1} w_i = \sum_{i=1}^{\ell} \alpha_i \lambda_i^k P^{-1} w_i = 0$

Solving the system

$$\left.\begin{array}{r} \sum_{i=1}^{\ell} \alpha_i P^{-1} w_i = 0 \\ \sum_{i=1}^{\ell} \alpha_i \lambda_i P^{-1} w_i = 0 \\ \vdots \\ \sum_{i=1}^{\ell} \alpha_i \lambda_i^{\ell-1} P^{-1} w_i = 0 \end{array}\right\},$$

we obtain $\alpha_i = 0$, for $i = 1, \dots \ell$. Consequently, the vectors are linearly independent. □

It is important to remark the following proposition corresponding to the eigenvectors of infinity.

**Proposition 4** *Let* $(E, A)$ *and* $(\bar{E}, \bar{A})$ *be two equivalent systems and* $0 \neq w \in \mathbb{C}^n$. $w \in \operatorname{Ker} E$ *if and only if* $P^{-1}w \in \operatorname{Ker} \bar{E}$ *with* $(\bar{E}, \bar{A}) = (QEP, QAP)$.

*Proof* Let $(E, A)$ and $(\bar{E}, \bar{A}) = (QEP, QAP)$ two equivalent systems.

$Ew = 0$ if and only if $Q^{-1}\bar{E}P^{-1}w = 0$, equivalently if and only if $\bar{E}P^{-1}w = 0$. □

## 3.1 Quasi-Weierstraß Form

The pair of matrices $(E, A)$ can be reduced to a weaker form called Quasi-Weierstraß form [2]

$$\begin{pmatrix} EV & AW \end{pmatrix}^{-1} (E, A) \begin{pmatrix} (V\ W) & \\ & (V\ W) \end{pmatrix} = \left( \begin{pmatrix} I_r & \\ & N \end{pmatrix}, \begin{pmatrix} A_r & \\ & I_{n-r} \end{pmatrix} \right) = (\tilde{E}, \tilde{A})$$

where $A_r$ is some matrix and $N$ is nilpotent. Matrices $V \in M_{n\times r}(\mathbb{C})$ and $W \in M_{n\times(n-r)}(\mathbb{C})$ are in such a way that the block column matrices $\begin{pmatrix} V & W \end{pmatrix}$ and $\begin{pmatrix} EV & AW \end{pmatrix}$ are invertible.

The vector spaces $\operatorname{Im} V$ and $\operatorname{Im} W$ are spanned by the generalized eigenvector at the finite and infinite eigenvalues respectively, and they are derived by the following recursive subspace iteration with a limited number of steps called Wong sequences [18].

$$\begin{array}{ll} V_0 = \mathbb{C}^n, & V_{i+1} = \{v \in \mathbb{C}^n \mid Av \in E(V_i)\} \\ W_0 = \{0\}, & W_{i+1} = \{v \in \mathbb{C}^n \mid Ev \in A(W_i)\} \end{array}$$

verifying

$$\begin{array}{l} V_0 \supseteq V_1 \supseteq \dots \supseteq V_\ell = V_{\ell+1} = \dots V_{\ell+q} = V^* \supseteq \operatorname{Ker} A \\ W_0 \subseteq W_1 \subseteq \dots \subseteq w_m = W_{m+1} = \dots W_{m+q} = W^* \end{array}$$

It is easy to prove that $\ell = m$ and satisfy $AV^* \subseteq EV^*$ and $EW^* \subseteq AW^*$.

Matrices $V$ and $W$ are defined in such away that $V^* = \operatorname{Im} V$ and $W^* = \operatorname{Im} W$.

*Example 1* Let $(E, A)$ a system with

$$E = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 1 & 1 & 2 \end{pmatrix} \qquad \text{and} \qquad A = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}$$

$$\begin{aligned} &W_0 = \{0\} \\ &W_1 = \operatorname{Ker} E = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} = W_2 = W \\ &V_0 = \mathbb{R}^3 \\ &V_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} = V_2 = V \end{aligned}$$

then,

$$\begin{pmatrix} 3 & 1 & 2 \\ 4 & 2 & 2 \\ 3 & 1 & -4 \end{pmatrix}^{-1} \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \tilde{E}$$

$$\begin{pmatrix} 3 & 1 & 2 \\ 4 & 2 & 2 \\ 3 & 1 & -4 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & -1 \end{pmatrix} = \begin{pmatrix} 2 & -2 & 0 \\ -5 & 5 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \tilde{A}$$

Similarly to Propositions 1, 2 and 4 we can prove the following results.

**Proposition 5** *Let $(E, A)$ be a system and $(\tilde{E}, \tilde{A})$ its quasi-Weierstraß form. $\lambda_0$ is an eigenvalue of $(E, A)$ if and only if it is an eigenvalue of $(\tilde{E}, \tilde{A})$.*

**Proposition 6** *Let $(E, A)$ be a system and $(\tilde{E}, \tilde{A})$ its quasi-Weierstraß form. $w_0$ is an eigenvector of $(E, A)$ if and only if $P^{-1}w_0$ is an eigenvector of $(\tilde{E}, \tilde{A})$ and $P^{-1}w_0 = \begin{pmatrix} v_0^1 \\ 0 \end{pmatrix}$ with $v_0^1 \in \mathbb{C}^r$ and $0 \in \mathbb{C}^{n-r}$ and $v_0^1$ is an eigenvector of $A_r$.*

It is important to remark the following proposition corresponding to the eigenvectors at the infinity.

**Proposition 7** *Let $(E, A)$ be a system and $(\tilde{E}, \tilde{A})$ its quasi-Weierstraß form. Let us consider a non-zero vector $w \in \mathbb{C}^n$. Then, $w \in \operatorname{Ker} E$ if and only if $P^{-1}w \in \operatorname{Ker} \tilde{E}$ with $(\tilde{E}, \tilde{A}) = (QEP, QAP)$.*

### 3.2 Controllability

An important concept concerning structural properties is the controllability that is defined as follows.

**Definition 3** The system (1) is called controllable if, for any $t_1 > 0$, $x(0) \in \mathbb{C}^n$ and $w \in \mathbb{C}^n$, there exists a control input $u(t)$ sufficiently smooth such that $x(t_1) = w$.

The controllability character can be computed by means of the generalized Hautus test for controllability of singular systems.

**Proposition 8 ([10])** *The system* (1) *is controllable if and only if:*

$$\left.\begin{array}{l} \operatorname{rank}\begin{pmatrix} E & B \end{pmatrix} = n \\ \operatorname{rank}\begin{pmatrix} sE - A & B \end{pmatrix} = n, \ \forall s \in \mathbb{C} \end{array}\right\}. \tag{3}$$

**Proposition 9 ([10])** *A system* $(E, A, B)$ *is controllable if and only if the matrix*

$$M = \begin{pmatrix} E & B & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\ A & 0 & B & E & B & 0 & 0 & 0 & 0 & \\ 0 & 0 & 0 & A & 0 & B & E & B & 0 & \\ 0 & 0 & 0 & 0 & 0 & 0 & A & 0 & B & \\ & & & & & & & & & \ddots \end{pmatrix} \in M_{n^2 \times ((n-1)n + 2(n-1)m)}(\mathrm{C}).$$

*has full rank.*

Observe that the matrix $M$ is constructed gluing matrix blocks $\begin{pmatrix} E & B & 0 \\ A & 0 & B \end{pmatrix}$ in the lower right corner.

## 4 Exact Controllability

There are many possible control matrices $B$ in the system (1) that satisfy the controllability condition. The goal is to find the set of all possible matrices $B$, having the minimum number of columns corresponding to the minimum number $n_B(E, A)$ of independent controllers required to control the whole network.

**Definition 4** Let $(E, A)$ be a pair of matrices. The exact controllability $n_B(E, A)$ is the minimum of the rank of all possible matrices $B$ making the system (1) controllable.

$$n_B(E, A) = \min \{\operatorname{rank} B, \forall B \in M_{n \times i} \ 1 \le i \le n, \ (E, A, B) \text{ controllable}\}. \tag{4}$$

If confusion is not possible we will write simply $n_B$.

Taking into account the generalized Hautus condition (3), it is straightforward the following proposition.

**Proposition 10** *The exact controllability $n_B$ is invariant under equivalence relation considered, that is to say: for any couple of invertible matrices $(Q, P)$,*

$$n_B(E, A) = n_B(QEP, QAP).$$

*Proof*

$$\begin{array}{l}\text{rank } \begin{pmatrix} QEP & QB \end{pmatrix} = \\ \text{rank } Q \begin{pmatrix} E & B \end{pmatrix} \begin{pmatrix} P & \\ & I \end{pmatrix} = \text{rank } \begin{pmatrix} E & B \end{pmatrix}\end{array}$$

$$\begin{array}{l}\text{rank } \begin{pmatrix} sQEP - QAP & QB \end{pmatrix} = \\ \text{rank } Q \begin{pmatrix} sE - A & B \end{pmatrix} \begin{pmatrix} P & \\ & I \end{pmatrix} = \text{rank } \begin{pmatrix} sE - A & B \end{pmatrix}\end{array}$$

□

As a consequence, if necessary we can consider $(E, A)$ in its canonical reduced form.

*Example 2*

1) If $E = A = 0$, $n_B = n$
2) If $E = I$ and $A = \text{diag}(\lambda_1, \dots, \lambda_n)$ with $\lambda_i \neq \lambda_j$ for all $i \neq j$, then $n_B = 1$, (it suffices to take $B = (1 \dots 1)^t$).
3) Not every matrix $B$ having $n_B$ columns is valid to make the system controllable. For example if $E = I$, $A = \text{diag}(1, 2, 3)$ and $B = (1, 0, 0)^t$, the system $(A, B)$ is not controllable, (rank $\begin{pmatrix} B & AB & A^2B \end{pmatrix} = 1 < 3$, or equivalently rank $\begin{pmatrix} A - \lambda I & B \end{pmatrix} = 2$ for $\lambda = 2, 3$.
4) If $E = \begin{pmatrix} 1 & \\ & 0 \end{pmatrix}$ and $A = \begin{pmatrix} 0 & \\ & 1 \end{pmatrix}$, $n_B = 1$. It suffices to consider $B = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

For standard systems we have the following result

**Proposition 11 ([19])**

$$n_B = \max_i \{\mu(\lambda_i)\}$$

*where* $\mu(\lambda_i) = \dim \text{Ker}\,(A - \lambda_i I)$ *is the geometric multiplicity of the eigenvalue* $\lambda_i$.

This proposition is a direct consequence of the Fattorini-Hautus test proved in [6] (in the infinite-dimensional version) and [11] (for finite-dimensional systems).

For singular systems, it is obvious that $n_B \geq n - \text{rank } E = n_E$, and we have the following theorem.

**Theorem 1** *Let $(E, A)$ be a singular system. The exact controllability $n_B$ is computed in the following manner.*

$$n_B = \max\{n_E, \mu(\lambda_i)\}$$

*where $\mu(\lambda_i) = \dim \operatorname{Ker}(\lambda_i E - A)$ and $\lambda_i$ (for each $i$) is the eigenvalue of pencil $sE - A$.*

*Proof* $\operatorname{rank}(\lambda_i E - A) = \operatorname{rank}\left(\lambda_i \begin{pmatrix} I & \\ & N \end{pmatrix} - \begin{pmatrix} J & \\ & I \end{pmatrix}\right) = n_2 + \operatorname{rank}(\lambda_i I - J)$. □

*Example 3* Let $(E, A)$ be a singular system with

$$E = \begin{pmatrix}
1&0&0&0&0&0&0&0&0&0&0&0&0\\
0&1&0&0&0&0&0&0&0&0&0&0&0\\
0&0&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&1&0&0&0&0&0&0&0&0&0\\
0&0&0&0&1&0&0&0&0&0&0&0&0\\
0&0&0&0&0&1&0&0&0&0&0&0&0\\
0&0&0&0&0&0&1&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&1&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&1&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&1&0
\end{pmatrix},$$

and

$$A = \begin{pmatrix}
3&0&0&0&0&0&0&0&0&0&0&0&0\\
1&3&0&0&0&0&0&0&0&0&0&0&0\\
0&1&3&0&0&0&0&0&0&0&0&0&0\\
0&0&0&3&0&0&0&0&0&0&0&0&0\\
0&0&0&1&3&0&0&0&0&0&0&0&0\\
0&0&0&0&0&2&0&0&0&0&0&0&0\\
0&0&0&0&0&1&2&0&0&0&0&0&0\\
0&0&0&0&0&0&0&1&0&0&0&0&0\\
0&0&0&0&0&0&0&0&1&0&0&0&0\\
0&0&0&0&0&0&0&0&0&1&0&0&0\\
0&0&0&0&0&0&0&0&0&0&1&0&0\\
0&0&0&0&0&0&0&0&0&0&0&1&0\\
0&0&0&0&0&0&0&0&0&0&0&0&1
\end{pmatrix}$$

We have:

$$\begin{aligned} &\operatorname{rank} E = 10 \\ &\operatorname{rank}(sE - A) = \begin{cases} 11 \text{ for } s = 3 \\ 12 \text{ for } s = 2 \\ 13 \text{ for all } s \neq 2, 3. \end{cases} \end{aligned}$$

So,

$n_E = 3$,
$\mu(3) = 2$,
$\mu(2) = 1$,

then $n_B = \max(3, 2, 1) = 3$.

In fact, taking

$$B = \begin{pmatrix} 1&0&0 \\ 0&0&0 \\ 0&0&0 \\ 0&0&1 \\ 0&0&0 \\ 1&0&0 \\ 0&0&0 \\ 1&0&0 \\ 0&0&0 \\ 0&1&0 \\ 0&0&0 \\ 0&0&1 \\ 0&0&0 \end{pmatrix}$$

$$\begin{aligned} &\operatorname{rank}\begin{pmatrix} E & B \end{pmatrix} = 13 \\ &\operatorname{rank}\begin{pmatrix} sE - A & B \end{pmatrix} = 13 \text{ for all } s \in \mathbb{C}. \end{aligned}$$

Obviously, for all matrix

$$B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \\ b_{41} & b_{42} \\ b_{51} & b_{52} \\ b_{61} & b_{62} \\ b_{71} & b_{72} \\ b_{81} & b_{82} \\ b_{91} & b_{92} \\ b_{101} & b_{102} \\ b_{111} & b_{112} \\ b_{121} & b_{122} \\ b_{131} & b_{132} \end{pmatrix}$$

$$\operatorname{rank}\begin{pmatrix} E & B \end{pmatrix} < 13.$$

## 5 Generators of Control Space

As we have discussed in the previous section, not every matrix B serves to make the system controllable, of all possible, we want to find those with the least number of columns.

**Proposition 12** *Let $(E, A)$ be a system with $E$ invertible. Then, the matrices $B$ making the system $(E, A, B)$ controllable are those that make the standard system $\dot{z}(t) = AE^{-1}z(t) + Bu(t)$, controllable.*

*Proof*

$$\begin{aligned} &\operatorname{rank}\begin{pmatrix} E & B \end{pmatrix} = n \text{ for all matrix } B \\ &\operatorname{rank}\begin{pmatrix} sE - A & B \end{pmatrix} = \operatorname{rank}\begin{pmatrix} sI - AE^{-1} & B \end{pmatrix}\begin{pmatrix} E & \\ & I \end{pmatrix} = \\ &\operatorname{rank}\begin{pmatrix} sI - AE^{-1} & B \end{pmatrix} \end{aligned}$$

Then, $\operatorname{rank}\begin{pmatrix} sE - A & B \end{pmatrix} = n$ if and only if $\operatorname{rank}\begin{pmatrix} sI - AE^{-1} & B \end{pmatrix} = n$. Now, it suffices to apply Hautus test for linear systems [11]. □

The minimal sets of matrices $B$ making a standard systems controllable are described in [8].

The solution of the problem for standard systems is linked to the eigenstructure of the matrix $AE^{-1}$. In our particular setup, the eigenstructure of the $AE^{-1}$ corresponds to the eigenstructure of the pair $(E, A)$ because of $\det(sI - AE^{-1}) \det E = \det((sI - AE^{-1})E) = \det(sE - A)$.

Then, we have the following proposition.

**Proposition 13** *Let $(E, A)$ with $E$ regular and $\lambda_1, \dots, \lambda_n$ with $\lambda_i \neq \lambda_j$ for $i \neq j$, the generalized eigenvalues of $(E, A)$ and $w_1, \dots, w_n$ the corresponding generalized eigenvectors. then $(E, A, B)$ with $B = [\sum_{i=1}^{n} \alpha_i E w_i]$ with $\alpha_i \neq 0$ is controllable.*

*Proof* It suffices to observe that if $w_i$ is a generalized eigenvector of $(E, A)$ for the generalized eigenvalue $\lambda_i$, then $Ew_i$ is an eigenvector of $AE^{-1}$ corresponding to the eigenvalue $\lambda_i$, and now to apply [8], proposition 2, or equivalently we can apply Kalman rank [13]. □

**Proposition 14** *Suppose that the the pair of matrices $(\bar{E}, \bar{A})$ defining the singular system is in its canonical reduced form and that the associate pencil to the pair $(\bar{E}, \bar{A})$ is regular and* rank $\bar{E} = n - 1$. *Let $\lambda_1, \dots, \lambda_r$ with $\lambda_i \neq \lambda_j$ for all $i \neq j$ are the generalized eigenvalues of the the pair of matrices $(\bar{E}, \bar{A})$. Then, the system $(\bar{E}, \bar{A}, \bar{B})$ with $\bar{B} = (\alpha_1, \dots, \alpha_n)^t$ and $\alpha_i \neq 0$ for $i = 1, \dots, r$ and $i = n$ is controllable*

*Proof* By hypothesis $N$ verifies that $N^{n-r} = 0$ but $N^{n-r-1} \neq 0$ and $J$ is diagonal.

i)

$$\operatorname{rank}\begin{pmatrix}\bar{E} & \bar{B}\end{pmatrix} = \operatorname{rank}\left(\begin{array}{ccc|cccc|c}
1 & & & & & & & \alpha_1 \\
 & \ddots & & & & & & \vdots \\
 & & 1 & & & & & \alpha_r \\
\hline
 & & & 0 & 1 & & & \alpha_{r+1} \\
 & & & & \ddots & & & \vdots \\
 & & & & & 0 & 1 & \alpha_{n-1} \\
 & & & & & & 0 & \alpha_n
\end{array}\right) = n$$

ii)

$$\operatorname{rank}\begin{pmatrix}s\bar{E} - \bar{A} & \bar{B}\end{pmatrix} =$$

$$\operatorname{rank}\left(\begin{array}{ccc|cccc|c}
s-\lambda_1 & & & & & & & \alpha_1 \\
 & \ddots & & & & & & \vdots \\
 & & s-\lambda_r & & & & & \alpha_r \\
\hline
 & & & -1 & s & & & \alpha_{r+1} \\
 & & & & \ddots & & & \vdots \\
 & & & & & -1 & s & \alpha_{n-1} \\
 & & & & & & -1 & \alpha_n
\end{array}\right) = n, \ \forall s \in \mathbb{C}.$$

□

*Remark 1* Notice that $v_1 = (1, 0, \dots, 0), \dots, v_r = (0, \dots, 0, \overset{\overset{r}{\smile}}{1}, 0, \dots, 0)$ verify

$$(\lambda_1 \bar{E} - \bar{A})v_1 = 0, \ \dots \ (\lambda_r \bar{E} - \bar{A})v_r = 0.$$

Then, $v_1, \dots, v_r$ are the generalized eigenvectors of $(\bar{E}, \bar{A})$, and $v_n \in \operatorname{Ker} \bar{E}^{n-r}$.

**Proposition 15** *Suppose that the the pair of matrices $(\bar{E}, \bar{A})$ defining the singular system is in its canonical reduced form and that the associate pencil to the pair $(\bar{E}, \bar{A})$ is regular of index one. Let $\lambda_1, \dots, \lambda_r$ with $\lambda_i \neq \lambda_j$ for all $i \neq j$ are the generalized eigenvalues of the the pair of matrices $(\bar{E}, \bar{A})$. Then, the system $(\bar{E}, \bar{A}, \bar{B})$ with $\bar{B} = (\sum_{i=1}^{r+1} v_i, v_{r+2}, \dots v_n)$ where $v_i = (0, \dots, 0, \overset{\overset{r}{\smile}}{1}, 0, \dots, 0)$, is controllable*

*Proof* By hypothesis $N$ verifies that $N = 0$ and $J$ is diagonal.

i)

$$\operatorname{rank}\begin{pmatrix}\bar{E} & \bar{B}\end{pmatrix} = \operatorname{rank}\left(\begin{array}{ccc|ccc|cccc}
1 & & & & & & 1 & 0 & \dots & 0 \\
 & \ddots & & & & & \vdots & & \ddots & \vdots \\
 & & 1 & & & & 1 & 0 & \dots & 0 \\
\hline
 & & & 0 & & & 1 & 0 & \dots & 0 \\
 & & & & 0 & & 0 & 1 & \dots & 0 \\
 & & & & & \ddots & \vdots & & \ddots & \vdots \\
 & & & & & 0 & 0 & 0 & & 1
\end{array}\right) = n$$

ii)

$$\operatorname{rank}\begin{pmatrix}s\bar{E} - \bar{A} & \bar{B}\end{pmatrix} =$$

$$\operatorname{rank}\left(\begin{array}{ccc|ccc|cccc}
s-\lambda_1 & & & & & & 1 & 0 & \dots & 0 \\
 & \ddots & & & & & \vdots & & \ddots & \vdots \\
 & & s-\lambda_r & & & & 1 & 0 & \dots & 0 \\
\hline
 & & & -1 & & & 1 & 0 & \dots & 0 \\
 & & & & & & 0 & 1 & \dots & 0 \\
 & & & & \ddots & & \vdots & & \ddots & \vdots \\
 & & & & & -1 & 0 & 0 & & 1
\end{array}\right) = n, \ \forall s \in \mathbb{C}.$$

□

**Proposition 16** *Suppose that the pencil $(\tilde{E}, \tilde{A})$ is in its quasi-Weirstraß form and it has a unique eigenvalue of multiplicity $r$ with a unique eigenvector and let $0 \neq u \in Ker\,(\bar{A} - \lambda I)^r \setminus Ker\,(\bar{A} - \lambda I)^{r-1}$ and $0 \neq v \in Ker\,\bar{E}^{n-r} \setminus Ker^{n-r-1}$. We consider* $\operatorname{Im} B = [u + v]$*, then $(\bar{E}, \bar{A}, \bar{B})$ is controllable.*

Let $P = \begin{pmatrix} V & W \end{pmatrix}$, $Q = \begin{pmatrix} EV & EW \end{pmatrix}^{-1}$ be invertible matrices such that $(QEP, QAP) = (\tilde{E}, \tilde{A})$ is in its quasi-Weierstraß form.

**Proposition 17** *Suppose that the associate pencil to the pair $(E, A)$ is regular of index one and* $\operatorname{rank} E = n - 1$. *Let $\lambda_1, \dots, \lambda_{n-1}$ with $\lambda_i \neq \lambda_j$ for all $i \neq j$ are the generalized eigenvalues and $w_1, \dots, w_{n-1}$ the corresponding generalized eigenvectors of the the pair of matrices $(E, A)$ and $0 \neq w_n \in \operatorname{Ker} E$. Then, the system $(E, A, B)$ with $B = Q^{-1}P^{-1}\sum_{i=1}^{n} \alpha_i w_i$, $\alpha_i \neq 0$ for $i = 1, \dots, n$ is controllable.*

*Proof* It suffices to apply Propositions 5, 6, 14 and 10. □

Let $P$ and $Q$ invertible matrices such that $(QEP, QAP) = (\bar{E}, \bar{A})$ is in its canonical reduced form.

**Proposition 18** *Suppose that the associate pencil to the pair $(E, A)$ is regular of index one. Let $\lambda_1, \dots, \lambda_r$ with $\lambda_i \neq \lambda_j$ for all $i \neq j$ are the generalized eigenvalues and $w_1, \dots, w_r$ the corresponding generalized eigenvectors of the the pair of matrices $(E, A)$ and let $\{w_{r+1}, \dots, w_n\}$ be a basis of Ker $E$. Then, the system $(E, A, B)$ with*

$$B = Q^{-1}P^{-1}\sum_{i=1}^{r+1} \alpha_i w_i, Q^{-1}P^{-1}w_{r+2}, \dots, Q^{-1}P^{-1}w_n,$$

*$\alpha_i \neq 0$ for $i = 1, \dots, r+1$ is controllable.*

*Proof* Let $v_i$ be as in Proposition 15.

Taking into account that $w_i$, $i = 1, \dots, r$ are generalized eigenvectors and $(w_{r+1}, \dots, w_n)$ is a basis of Ker $E$, we have that $P^{-1}w_i = \alpha_j v_j$ with $\alpha_j \neq 0$, for some $j = 1, \dots, n$ (we consider $v_i$ ordered in such a way that $P^{-1}w_i = \alpha_i v_i$).

On the other hand, we have that $(P^{-1}w_{r+1}, \dots, P^{-1}w_n)$ is a basis of Ker $\bar{E}$. So, $P^1 w_i = \sum_{j=r+1}^{n} \alpha_{ij} v_j$, and

$$(P^{-1}w_{r+1}, \dots, P^{-1}w_n) = \begin{pmatrix} 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \\ \alpha_{r+11} & \dots & \alpha_{r+1n-r} \\ \vdots & & \vdots \\ \alpha_{n1} & \dots & \alpha_{nn-r} \end{pmatrix}$$

with

$$\det \begin{pmatrix} \alpha_{r+11} & \dots & \alpha_{r+1n-r} \\ \vdots & & \vdots \\ \alpha_{n1} & \dots & \alpha_{nn-r} \end{pmatrix} \neq 0$$

Now it suffices to apply Proposition 15 for

$$\bar{B} = \begin{pmatrix} \alpha_1 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ \alpha_r & 0 & \dots & 0 \\ \alpha_{r+11} & \alpha_{r+12} & \dots & \alpha_{r+1n-r} \\ \vdots & \vdots & & \vdots \\ \alpha_{n1} & \alpha_{n2} & \dots & \alpha_{nn-r} \end{pmatrix}$$

and Proposition 10. □

*Example 4* Let $(E, A)$ be a system with

$$E = \begin{pmatrix} 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ 1 & 3 & 3 & 3 & 3 & 3 & 3 \\ 1 & 2 & 4 & 4 & 4 & 4 & 4 \\ 1 & 2 & 3 & 5 & 5 & 5 & 5 \\ 2 & 3 & 5 & 6 & 6 & 6 & 6 \\ 3 & 7 & 12 & 13 & 13 & 13 & 13 \\ 0 & 1 & 1 & 2 & 2 & 2 & 2 \end{pmatrix}$$

$$A = \begin{pmatrix} 4 & 4 & 4 & 4 & 4 & 4 & 4 \\ 2 & 8 & 8 & 8 & 8 & 8 & 8 \\ 2 & 5 & 13 & 13 & 13 & 13 & 13 \\ 2 & 5 & 9 & 19 & 19 & 19 & 19 \\ 4 & 7 & 15 & 20 & 24 & 24 & 24 \\ 6 & 18 & 38 & 43 & 44 & 46 & 46 \\ 0 & 3 & 3 & 8 & 8 & 9 & 10 \end{pmatrix}$$

$\operatorname{Ker} E = [w_5 = (0, 0, 0, -1, 0, 0, 1),\ w_6 = (0, 0, 0, 0, -2, 0, 2),\ w_7 = (0, 0, 0, 0, 0, -3, 3)]$

Taking

$$B = \begin{pmatrix} 2 & 0 & 0 \\ 3 & 0 & 0 \\ 4 & 0 & 0 \\ 5 & 0 & 0 \\ 10 & 0 & 0 \\ 16 & 4 & 0 \\ 4 & 4 & 3 \end{pmatrix}$$

The system is controllable.

## 6 Conclusion

In this work, given two $n$-order square matrices $E$, $A$ defining regular generalized systems $E\dot{x} = Ax$. We ask for minimal number of columns that must have a matrix $B$ in order to make the system $(E, A, B)$ controllable. Examples have been included to make the work easier readable. Finally, a characterization of controllable states for regular generalized systems of index at most one are presented.

## References

1. Assan, J., Lafay, A., Perdon, A.: Computation of maximal pre-controllability submodules over a Noetherian ring. Syst. Control Lett. **37**, 153–161 (1999)
2. Berger, T., Ilchmann, A., Trenn, S.: The quasi-Weierstraß form for regular matrix pencils. Linear Algebra Appl. **436**, 4052–4069 (2012)
3. Cardetti, F., Gordina, M.: A note on local controllability on lie groups. Syst. Control Lett. **57**, 978–979 (2008)
4. Chen, C.T.: Introduction to Linear System Theory. Holt, Rinehart and Winston Inc, New York (1970)
5. Dai, L.L.: Singular Control Systems. Springer, New York (1989)
6. Fattorini, H.O.: Some remarks on complete controllability. SIAM J. Control **4**, 686–694 (1966)
7. García-Planas, M.I.: Generalized controllability subspaces for time-invariant singular linear systems. In: Proceedings of Physcon 2011, IPACS Electronic Library (2011)
8. García-Planas, M.I.: Exact controllability of linear dynamical systems: a geometrical approach. Appl. Math. **1**, 37–47 (2017) https://doi.org/10.21136/AM.2016.0427-15
9. García-Planas, M.I., Domínguez, J.L.: Alternative tests for functional and pointwise output-controllability of linear time-invariant systems. Syst. Control Lett. **62**, 382–387 (2013)
10. García-Planas, M.I., Tarragona, S., Diaz, A.: Controllability of time-invariant singular linear systems. In: From Physics to Control through an Emergent View, pp. 112–117. World Scientific, Singapore (2010)
11. Hautus, M.L.J.: Controllability and observability conditions of linear autonomous systems. Nederl. Akad. Wetensch. (Reprinted form Proc. Ser. A **72** and Indag. Math. **31**) 443–448 (1969)
12. Heniche, A., Kamwa, I.: Using measures of controllability and observability for input and output selection. In: Proceedings of the 2002 IEEE International Conference on Control Applications, vol. 2, pp. 1248–1251 (2002)
13. Kalman, R.E., Falb, P.L., Arbib, M.R.: Topics in Mathematical Control Theory. McGraw-Hill, New York/Toronto, Ontario/London (1969)
14. Kundur, P.: Power System Stability and Control. McGraw-Hill, New York (1994)
15. Lin, C.: Structural controllability. IEEE Trans. Autom. Control **19**, 201–208 (1974)
16. Liu, Y., Slotine, J., Barabási, A.: Controllability of complex networks. Nature **473**, 167–173 (2011)
17. Shields, R., Pearson, J.: Structural controllability of multi-input linear systems. IEEE Trans. Autom. Control **21**, 203–212 (1976)
18. Wong, K.-T.: The eigenvalue problem $\lambda Tx + Sx$. J. Differ. Equ. **16**, 270–280 (1974)
19. Yuan, Z.Z., Zhao, C., Wang, W.X., Di, Z.R., Lai, Y.C.: Exact controllability of multiplex networks. New J. Phys. **16**, 1–24 (2014)

# On Stability of Discontinuous Galerkin Approximations to Anisotropic Stokes Equations

**Francisco Guillén-González, María Victoria Redondo-Neble, and José Rafael Rodríguez-Galván**

**Abstract** This work delves into the numerical approximation of Anisotropic Stokes equations (with small vertical diffusion coefficient), which is a generalization of the Hydrostatic Stokes equations (with zero vertical diffusion). It is known that the Ladyzhenskaya-Babuška-Brezzi condition is not sufficient to stabilize usual finite elements approximations, because a new stability condition appears. Here we extend to the Anisotropic Stokes equations the new approach given in Guillén González et al. (On stability of discontinuous galerkin approximations to the hydrostatic Stokes equations, 2018, submitted) for the Hydrostatic case. This approach is a symmetric interior penalty discontinuous Galerkin method (SIP DG) with adequate stability terms, approximating both velocity and pressure in the same Finite Element (FE) space ($\mathcal{P}_k$-discontinuous). Stability and well-posedness of this method is proven. Finally, we show some numerical tests in agreement with our numerical analysis.

**Keywords** Stokes equations · Anisotropic viscosity · Discontinuous Galerkin · Finite elements · Fluid mechanics · Stability

## 1 Introduction

In this work we delve into the stability of a discrete discontinuous Galerkin (DG) formulation for the steady Anisotropic Stokes equations, which can be written as:

$$-\nu \Delta \mathbf{u} + \nabla_{\mathbf{x}} p = \mathbf{f}, \tag{1}$$

F. Guillén-González
Departamento de Ecuaciones Diferenciales y Análisis Numérico, IMUS, Universidad de Sevilla, Sevilla, Spain
e-mail: guillen@us.es

M. V. Redondo-Neble · J. R. Rodríguez-Galván (✉)
Departamento de Matemáticas, Universidad de Càdiz, Càdiz, Spain
e-mail: victoria.redondo@uca.es; rafael.rodriguez@uca.es

© Springer Nature Switzerland AG 2019

J. L. García Guirao et al. (eds.), *Recent Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 18,
https://doi.org/10.1007/978-3-030-00341-8_13

$$-\varepsilon \Delta v + \partial_z p = g, \tag{2}$$

$$\nabla_{\mathbf{x}} \cdot \mathbf{u} + \partial_z v = 0, \tag{3}$$

in a domain $\Omega \subset \mathbb{R}^d$ ($d = 2$ or $d = 3$). Although the numerical analysis will be done in the general case, $d = 2$ or $d = 3$, due to the theoretical orientation of this work and to the high computing requirements of the $3D$ case, only $2D$ numerical experiments will be performed. Of course, in the isotropic case ($\varepsilon = \nu$), Eqs. (1)–(3) correspond to the usual Stokes equations for viscous flow. Here we consider the anisotropic case, assuming diffusion coefficients $\nu > 0$ and $\varepsilon > 0$ of different orders. Specifically, $\nu$ is considered of order one while $\varepsilon$ is vanishing. The unknowns are the pressure, $p : \Omega \to \mathbb{R}$, and the $3D$ velocity field, $(\mathbf{u}, v) : \Omega \to \mathbb{R}^3$, where $\mathbf{u} = (u_1, u_2)$ and $v$ are respectively the horizontal and vertical components of velocity. We introduce the notation $\nabla_{\mathbf{x}} = (\partial_x, \partial_y)^T$, $\nabla_{\mathbf{x}} \cdot \mathbf{u} = \partial_x u_1 + \partial_y u_2$. The term $\mathbf{f} = (f_1, f_2)^T$ models a given horizontal force while $g$ involves the force due to gravity (which can be written in a potential form and incorporated to the pressure term, hence it can be assumed $g = 0$).

This kind of equations arise in more general (transient non-linear Anisotropic) models, for instance, Navier-Stokes equations for geophysical fluid dynamics governing large-scale motion of the ocean, which are derived from the conservation laws from physics [5]. Typically, in these models, the coefficient $\varepsilon$ is related to the "small layer" hypothesis:

$$\widehat{\varepsilon} = \frac{\text{vertical scale}}{\text{horizontal scale}} \quad \text{is very small,}$$

for example, a few km over some thousand km, that is $\widehat{\varepsilon} \simeq 10^{-3}, 10^{-4}$. After a vertical scaling, the anisotropic domain is transformed into the following isotropic or adimensional (independent of $\widehat{\varepsilon}$) domain

$$\Omega = \left\{(\mathbf{x}, z) \in \mathbb{R}^3 \;/\; \mathbf{x} = (x, y) \in S,\ -D(\mathbf{x}) < z < 0\right\},$$

where $S \subset \mathbb{R}^2$ is the surface domain and $D = D(\mathbf{x})$ is the bottom function. We decompose the boundary into three parts: the surface, $\Gamma_s = \overline{S} \times \{0\}$, the bottom, $\Gamma_b = \{(\mathbf{x}, -D(\mathbf{x})) \;/\; \mathbf{x} = (x, y) \in S\}$, and the talus (if it exists) or lateral walls, $\Gamma_l = \{(\mathbf{x}, z) \;/\; \mathbf{x} \in \partial S, -D(\mathbf{x}) < z < 0\}$. Finally, a $\widehat{\varepsilon}$-dependent scaling of vertical velocity is introduced (see [4]), leading (in the simplified linear steady case) to the Anisotropic (1)–(3) equations, where $\varepsilon = \nu \widehat{\varepsilon}^2$. Note that the limit case $\varepsilon = 0$, corresponds to the simplified model that is known as *Hydrostatic Navier-Stokes* or *Primitive Equations* [7].

Equations (1)–(3) are endowed with the boundary conditions:

$$\nu \partial_z \mathbf{u}|_{\Gamma_s} = \mathbf{g}_s, \quad v|_{\Gamma_s} = 0, \tag{4}$$

$$\mathbf{u}|_{\Gamma_b \cup \Gamma_l} = 0, \quad v|_{\Gamma_b} = 0, \tag{5}$$

$$\varepsilon \nabla_{\mathbf{x}} v \cdot \mathbf{n}_{\mathbf{x}}|_{\Gamma_l} = 0, \tag{6}$$

where $\mathbf{g}_s$ represents the wind stress and $\mathbf{n}_{\mathbf{x}}$ is the horizontal part of the normal vector. For the sake of simplicity we consider $\nu = 1$ in what follows.

According to [3, 10–12], the strong anisotropy of these equations introduces some difficulties in addition to those already known for usual (isotropic) Stokes equations. In particular, standard continuous finite element approximations (like Taylor-Hood $\mathcal{P}_2/\mathcal{P}_1$ or bubble $\mathcal{P}_{1,b}/\mathcal{P}_1$ elements) are not stable when $\varepsilon \to 0$. More specifically, we consider the following variational formulation: find $(\mathbf{u}, v, p) \in \mathbf{U} \times V \times P$ such that

$$(\nabla \mathbf{u}, \nabla \overline{\mathbf{u}}) - (p, \nabla_{\mathbf{x}} \cdot \overline{\mathbf{u}}) = (\mathbf{f}, \overline{\mathbf{u}}) + (\mathbf{g}_s, \overline{\mathbf{u}})_{\Gamma_s} \quad \forall \overline{\mathbf{u}} \in \mathbf{U}, \tag{7}$$

$$\varepsilon (\nabla v, \nabla \overline{v}) - (p, \partial_z \overline{v}) = 0 \quad \forall \overline{v} \in V, \tag{8}$$

$$(\nabla \cdot (\mathbf{u}, v), \overline{p}) = 0 \quad \forall \overline{p} \in P. \tag{9}$$

Here and in what follows $(\cdot, \cdot)$ is the $L^2(\Omega)$ scalar product, $(\cdot, \cdot)_{\Gamma_s}$ is the $L^2(\Gamma_s)$ scalar product and

$$\mathbf{U} = \mathbf{H}^1_{b,l}(\Omega) = \Big\{ \mathbf{u} \in H^1(\Omega)^2 \mid \mathbf{u}|_{\Gamma_b \cup \Gamma_l} = 0 \Big\},$$

$$V = H^1_{z,0}(\Omega) = \Big\{ v \in L^2(\Omega) \mid \partial_z v \in L^2(\Omega),\ v|_{\Gamma_s \cup \Gamma_b} = 0 \Big\},$$

$$P = L^2_0(\Omega) = \Big\{ p \in L^2(\Omega) \mid \int_\Omega p = 0 \Big\}.$$

We emphasize that formulation (7)–(9) is purely differential and, unlike most usual formulations for Primitive Equations, it avoids vertical integration and also avoids hypotheses of the type $D(\mathbf{x}) > D_0 > 0$. Consequently, it facilitates the use of standard (not necessarily structured) meshes and also tools and techniques like mesh adaptivity.

The space $\mathbf{U}$ is endowed with the norm $\|\nabla \mathbf{u}\|_0$ (hereafter $\|\cdot\|_0$ denotes the $L^2(\Omega)$-norm) while in $V$ we consider $\|\partial_z v\|_0$, which is a norm owing to the homogeneous Dirichlet condition $v|_{\Gamma_s \cup \Gamma_b} = 0$ and a vertical Poincaré inequality. We take in $P$ the usual $L^2(\Omega)$-norm.

As stated in [10], the (uniformly when $\varepsilon \to 0$) well-posedness of (7)–(9) hinges on the following inf-sup conditions:

$$\beta_p \|p\|_0 \le \sup_{0 \ne (\mathbf{u}, v) \in \mathbf{U} \times V} \frac{(\nabla \cdot (\mathbf{u}, v), p)}{\|(\nabla \mathbf{u},\ \partial_z v)\|_0} \qquad \forall\, p \in P, \tag*{$(IS)^P$}$$

$$\beta_v \|\partial_z v\|_0 \le \sup_{0 \ne p \in P} \frac{(\partial_z v, p)}{\|p\|_0} \qquad \forall\, v \in V, \tag*{$(IS)^V$}$$

where $\|(\nabla\mathbf{u}, \partial_z v)\|_0$ is the norm in $\mathbf{U} \times V$, $\beta_p$ is the Stokes inf-sup or LBB (Ladyzhenskaya–Babuška–Brezzi) constant and $\beta_v = 1$. In fact, one always has

$$\|\partial_z v\|_0 = \sup_{0 \neq p \in P} \frac{(\partial_z v, p)}{\|p\|_0} \quad \forall\, v \in V.$$

Note that $(IS)^P$ is basically the well-known LBB inequality while $(IS)^V$ is a new "hydrostatic" condition.

For the discrete setting, it was shown in [10] that the discrete counterpart of the inf-sup condition $(IS)^P$ is no longer sufficient for stability of standard conforming FE approximations of (7)–(9), because it is also necessary to choose FE spaces satisfying the discrete counterpart of $(IS)^V$. The problem is that the standard FE spaces used in the Stokes framework (for instance, Taylor-Hood $\mathcal{P}_2$-$\mathcal{P}_1$ or ($\mathcal{P}_1$bubble)-$\mathcal{P}_1$ continuous FE approximations) do not satisfy the discrete version of $(IS)^V$, and then different FE must be considered (for instance, via a different approximation for horizontal and vertical velocity, see [10, 12]).

A different idea was introduced in [11], where discrete $(IS)^V$ condition is avoided by adding a consistent stabilizing term to the vertical momentum equation (8). In this way, stability for Stokes-LBB FE combinations is shown and error estimates are provided for Taylor-Hood $\mathcal{P}_2$–$\mathcal{P}_1$ FE and mini-element ($\mathcal{P}_1$bubble)–$\mathcal{P}_1$ approximations, showing optimal convergence order in the $\mathcal{P}_2$–$\mathcal{P}_1$ case.

The goal of this paper is to extend to the general case ($\varepsilon \neq 0$) the ideas introduced in [13], which is focused on the hydrostatic ($\varepsilon = 0$) case. Like in [13], we explore a new approach for the finite element approximation of (7)–(9): by considering symmetric interior penalty discontinuous Galerkin (SIP DG) methods, we can define approximations that satisfy both $(IS)^V$ and $(IS)^P$ restrictions.

In Sect. 2 we introduce the well-known SIP DG discontinuous Galerkin approximations for classical (isotropic) Stokes equations, where the velocity field and the pressure are composed of same order discontinuous $\mathcal{P}_k$ polynomials. The novel results are in Sect. 3, where this approximation is extended to anisotropic equations (1)–(3), showing stability and well-posedness, and in Sect. 4, where we show some numerical tests in agreement with our numerical analysis.

## 2 SIP DG Approximation of the Stokes Problem

In this section, we introduce the mixed-type SIP DG approximation of classical Stokes equations, as can be found in literature (see e.g. [8, 15, 16] and references therein included). Before presenting the bilinear forms arising in Stokes-type SIP DG approximations, let us introduce some notation and introduce the SIP DG schemes for linear second order elliptic equations.

## 2.1 SIP DG Discretization of Second Order Elliptic Equations

In the usual sense of Ciarlet [6], we denote by $\mathcal{T}_h$ a family of meshes of the domain $\Omega \subset \mathbb{R}^d$ into non-degenerate disjoint simplicial elements $K$ satisfying usual regularity assumptions:

$$\exists\, \rho > 0 \,/\, \rho\, h_K \leq r_K \quad \forall K \in \mathcal{T}_h,$$

where $h_K$ is the diameter of $K$ and $r_K$ is the radius of the largest ball inscribed in $K$. The number of edges (faces) of the elements is denoted by $N_\partial$.

We associate to each triangulation $\mathcal{T}_h$ the set of interior faces (edges in $2D$) $\mathcal{E}_h^0$ and the set of boundary faces $\mathcal{E}_h^\partial$, defined as follows: $e \in \mathcal{E}_h^0$ if there are two polyhedra $K^+$ an $K^- \in \mathcal{T}_h$ such that $e = K^+ \cap K^-$ and $e \in \mathcal{E}_h^\partial$ if there is $K \in \mathcal{T}_h$ such that $e = \partial K \cap \partial\Omega$. We define $\mathcal{E}_h = \mathcal{E}_h^0 \cup \mathcal{E}_h^\partial$.

Let $u$ be a scalar-valued function on $\Omega$ and assume that $u$ is smooth enough to admit on all $e \in \mathcal{E}_h^0$ a (possibly two-valued) trace. We define respectively the jump and the average of $v$ on $e \in \mathcal{E}_h^0$ as follows: if $e = K^+ \cap K^-$ then $[\![u]\!]_e = u|_{K^+} - u|_{K^-}$ and $\{\!\{u\}\!\}_e = \frac{1}{2}(u|_{K^+} + u|_{K^-})$. If $e \in \mathcal{E}_h^\partial$, we define $[\![u]\!] = \{\!\{u\}\!\} = u|_e$.

Let us define the following *broken discrete* Sobolev space, for each $m \geq 0$,

$$H^m(\mathcal{T}_h) = \big\{u \in L^2(\Omega) \mid u \in H^m(K)\; \forall K \in \mathcal{T}_h\big\},$$

the broken gradient operator $\nabla_h$ for each $u \in H^1(\mathcal{T}_h)$,

$$(\nabla_h u)|_K = \nabla(u|_K) \quad \forall\, K \in \mathcal{T}_h$$

and the following finite-dimensional subspace of $H^m(\mathcal{T}_h)$, for each $k \in \mathbb{N}$:

$$\mathcal{P}_h^k := \mathcal{P}_k(\mathcal{T}_h) = \Big\{u \in L^2(\Omega) \mid u \in \mathbb{P}_k(K),\ \forall K \in \mathcal{T}_h\Big\}.$$

The following discrete trace inequality shall be useful: for all $u_h \in \mathcal{P}_h^k$ and $K \in \mathcal{T}_h$,

$$h_K^{1/2} \|u_h|_K\|_{L^2(\partial K)} \leq C_{tr}\, \|u_h\|_{L^2(K)}, \tag{10}$$

where $C_{tr}$ is a constant independent of $h$ and $K$ (and depending on $k$, $d$ and $\rho$). See e.g. [8], Lemma 1.46 and Remark 1.47, for details. From this inequality, one has the following technical result.

**Lemma 1** *For every $p_h \in \mathcal{P}_h^k$, the following inequalities are satisfied:*

$$\Big(\sum_{e \in \mathcal{E}_h} h_e \int_e \{\!\{p_h\}\!\}^2\Big)^{1/2} \leq C \|p_h\|_{L^2(\Omega)}, \tag{11}$$

$$\Big( \sum_{e\in\mathcal{E}_h} h_e \int_e [\![p_h]\!]^2 \Big)^{1/2} \le C \| p_h \|_{L^2(\Omega)}, \tag{12}$$

*where $C > 0$ independent of $h$ (and dependent on $k$, $d$ and $\rho$) and $h_e$ is the diameter of the edge or face $e$.*

Let us consider the well-known symmetric interior penalty (SIP) bilinear form, which was introduced by D.N. Arnold [1] for discontinuous FE approximation of second order elliptic equations:

$$\begin{aligned} a_h^{\text{sip},\eta}(u,\overline{u}) = &\int_\Omega \nabla_h u \cdot \nabla_h \overline{u} \\ &- \sum_{e\in\mathcal{E}_h} \int_e \big( \{\!\{\nabla_h u\}\!\} \cdot \mathbf{n}_e \, [\![\overline{u}]\!] + [\![u]\!] \, \{\!\{\nabla_h \overline{u}\}\!\} \cdot \mathbf{n}_e \big) \\ &+ \eta \sum_{e\in\mathcal{E}_h} \frac{1}{h_e} \int_e [\![u]\!] \, [\![\overline{u}]\!] \,, \end{aligned} \tag{13}$$

for each $u, \overline{u} \in \mathcal{P}_h^k$. Here $\mathbf{n}_e = (\mathbf{n_x}, n_z) \in \mathbb{R}^d$ denotes the normal vector (in a fixed chosen sense) across the edge or face $e$ and $\eta > 0$ is a constant. The second term at RHS of (13) arises for consistency and symmetry, while the last one introduces a penalization for the jumps on interior faces which is needed to gain coercivity for the $\|\cdot\|_{\text{sip}}$ defined bellow. Specifically, one has the following coercivity result in $\mathcal{P}_h^k$ (see e.g. [8, Lemma 4.12]):

**Lemma 2 (Coercivity for $a_h^{\text{sip}}(\cdot,\cdot)$)** *Let us denote $\eta_* = C_{tr}^2$, with $C_{tr}$ given in (10). For all $\eta > \eta_*$, one has*

$$a_h^{\text{sip},\eta}(u_h,u_h) \ge C(\eta) \|u_h\|_{\text{sip}}^2, \quad \forall\, u_h \in \mathcal{P}_h^k, \tag{14}$$

*where $C(\eta) = (\eta - \eta_*)/(1+\eta)$.*

Here

$$\|u\|_{\text{sip}} = \big( \, \|\nabla_h u\|_0^2 + |u|_{\text{sip}}^2 \, \big)^{1/2},$$

where the SIP jump seminorm

$$|u|_{\text{sip}} = \Big( \sum_{e\in\mathcal{E}_h} \frac{1}{h_e} \int_e [\![u]\!]^2 \Big)^{1/2} \tag{15}$$

introduces a penalization for jumps of $u$. Thus, one of the keys in SIP DG is to use coercivity inequality (14) to regain continuity of infinity dimensional solution, when $h \to 0$.

One also has boundedness of $a_h^{\mathrm{sip},\eta}(u_h, \overline{u}_h)$ in $\mathcal{P}_h^k$ with respect to $\|\cdot\|_{\mathrm{sip}}$-norm (see e.g. [8, Lemmas 4.16 and 4.20]): there exists $C_{\mathrm{bnd}} > 0$ independent of $h$ (and depending on $\eta$) such that, for all $u_h, \overline{u}_h \in \mathcal{P}_h^k$,

$$a_h^{\mathrm{sip},\eta}(u_h, \overline{u}_h) \le C_{\mathrm{bnd}} \|u_h\|_{\mathrm{sip}} \|\overline{u}_h\|_{\mathrm{sip}}. \tag{16}$$

## 2.2 SIP DG Formulation of Stokes Equations

Let us focus in this section in the well-known DG SIP approximation of the classical steady Stokes equations for viscous flows in $\Omega$ (for simplicity, we fix isotropic viscosity $\nu = 1$ and homogeneous Dirichlet boundary conditions on $\partial\Omega$): find $(\mathbf{w}, p) \in \left(H_0^1(\Omega)\right)^d \times L_0^2(\Omega)$ such that

$$-\nu \Delta \mathbf{w} + \nabla p = \mathbf{f}, \tag{17}$$

$$\nabla \cdot \mathbf{w} = 0, \tag{18}$$

in a weak sense. It is well-known that this problem can be expressed as a mixed formulation which is well-posed in the sense of Hadamard. In this section, we consider a well-known SIP-DG discretization of (17)–(18) based on equal-order velocities and pressures (see e.g. [8, 16] and references therein included).

Specifically, for any $k \in \mathbb{N}$ we define the following discrete velocity and pressure spaces:

$$W_h = (\mathcal{P}_h^k)^d, \quad P_{h,0} = \mathcal{P}_h^k \cap L_0^2(\Omega)$$

and the following Stokes SIP DG bilinear forms: for all $\mathbf{w}_h, \overline{\mathbf{w}}_h \in W_h$, $p_h \in P_{h,0}$,

$$a_h^{\mathrm{Sto}}(\mathbf{w}_h, \overline{\mathbf{w}}_h) = \sum_{i=1}^{d} a_h^{\mathrm{sip},\eta}(w_i, \overline{w}_i), \tag{19}$$

$$b_h^{\mathrm{Sto}}(\mathbf{w}_h, p_h) = -\int_\Omega p_h \, \nabla_h \cdot \mathbf{w}_h + \sum_{e \in \mathcal{E}_h} \int_e [\![\mathbf{w}_h]\!] \cdot \mathbf{n}_e \; \{\!\{p_h\}\!\}\,. \tag{20}$$

Here we denote $\mathbf{w}_h = (w_i)_{i=1}^d$, $\nabla_h \cdot$ is the "broken" divergence operator (defined on each $K \in \mathcal{T}_h$) and the last term in (20) appears when Green's theorem is applied on each element.

A partial coercivity result follows from Lemma 2: for all $\eta > \eta_*$, there exists $C(\eta) > 0$ such that

$$a_h^{\mathrm{Sto}}(\mathbf{w}_h, \mathbf{w}_h) \ge C(\eta) \|\mathbf{w}_h\|_{\mathrm{sip}}^2, \quad \forall \mathbf{w}_h \in \mathbf{W}_h,$$

with the natural energy norm

$$\|\mathbf{w}_h\|_{\text{sip}}^2 := \sum_{i=1}^{d} \|w_i\|_{\text{sip}}^2.$$

It is not difficult to show the continuity of $b_h^{\text{Sto}}(\mathbf{w}_h, p_h)$ with respect to the norms $\|\mathbf{w}_h\|_{\text{sip}}$ and $\|p_h\|_0$ in $\mathbf{W}_h \times P_{h,0}$. Also, the following inf-sup condition holds (see [8, Lemma 6.10]): there exists $\beta > 0$ such that

$$\beta \, \|p_h\|_0 \le \sup_{\mathbf{w}_h \in \mathbf{W}_h \setminus \{0\}} \frac{b_h^{\text{Sto}}(\mathbf{w}_h, p_h)}{\|\mathbf{w}_h\|_{\text{sip}}} + |p_h|_P, \quad \forall p_h \in P_{h,0}. \tag{21}$$

Note that the norm $\|p_h\|_0$ in $P_{h,0}$ cannot be controlled uniquely in term of $b_h^{\text{Sto}}(\cdot, \cdot)$. In fact, the following extra semi-norm term

$$|p|_P = \Big( \sum_{e \in \mathcal{E}_h^0} h_e \int_e [\![p]\!]^2 \Big)^{1/2}$$

must be added to the usual discrete Stokes inf-sup condition. Note that (12) implies that $|p|_P \le C \, \|p_h\|_0$. Therefore, in order to apply (21), a stabilization term is added to the usual Stokes mixed variational formulation.

Specifically, the following discretization of (17)–(18) is considered for classical Stokes equations:

$$a_h^{\text{Sto}}(\mathbf{w}_h, \overline{\mathbf{w}}_h) + b_h^{\text{Sto}}(\overline{\mathbf{w}}_h, p_h) = (\mathbf{f}, \overline{\mathbf{w}}_h) \quad \forall \overline{\mathbf{w}}_h \in \mathbf{W}_h, \tag{22}$$

$$-b_h^{\text{Sto}}(\mathbf{w}_h, \overline{p}_h) + s_h^p(p_h, \overline{p}_h) = 0 \quad \forall \overline{p}_h \in P_{h,0}, \tag{23}$$

where the stabilization bilinear form

$$s_h^p(p_h, \overline{p}_h) = \sum_{e \in \mathcal{E}_h^0} h_e \int_e [\![p_h]\!] \, [\![\overline{p}_h]\!] \tag{24}$$

is introduced to control $\|p_h\|_0$ by the DG inf-sup condition (21). Then the following global inf-sup stability holds (see e.g. [8, Lemma 6.13]), which imply stability (and in particular well-posedness) of the scheme (22)–(23): there exists $\gamma > 0$ such that for all $(\mathbf{w}_h, p_h) \in \mathbf{X}_h = \mathbf{W}_h \times P_{h,0}$,

$$\gamma \, \|(\mathbf{w}_h, p_h)\|_{\mathbf{X}_h} \le \sup_{(\overline{\mathbf{w}}_h, \overline{p}_h) \in \mathbf{X}_h \setminus (0,0)} \frac{c_h^{\text{Sto}}((\mathbf{w}_h, p_h), (\overline{\mathbf{w}}_h, \overline{p}_h))}{\|(\overline{\mathbf{w}}_h, \overline{p}_h)\|_{\mathbf{X}_h}}, \tag{25}$$

where we denote

$$\|(\mathbf{w}_h, p_h)\|^2_{\mathbf{X}_h} = \|\mathbf{w}_h\|^2_{\text{sip}} + \|p_h\|^2_0 + |p_h|^2_P, \tag{26}$$

$$\begin{aligned} c_h^{\text{Sto}}((\mathbf{w}_h, p_h), (\overline{\mathbf{w}}_h, \overline{p}_h)) = a_h^{\text{Sto}}(\mathbf{w}_h, \overline{\mathbf{w}}_h) + b_h^{\text{Sto}}(\overline{\mathbf{w}}_h, p_h) \\ - b_h^{\text{Sto}}(\mathbf{w}_h, \overline{p}_h) + s_h^p(p_h, \overline{p}). \end{aligned} \tag{27}$$

We remark that $|p_h|_P \le C \|p_h\|_0$ implies that $|p_h|_P$ can be removed from (26). Anyway we keep explicitly this term, as usual in literature (see e.g. [8]). On the other hand, continuity of $b_h^{\text{Sto}}(\cdot, \cdot)$, and then of $c_h^{\text{Sto}}((\mathbf{w}_h, p_h), (\overline{\mathbf{w}}_h, \overline{p}_h))$, with respect to $\|\cdot\|_{\mathbf{X}_h}$ is straightforward.

# 3 SIP DG Approximation of Anisotropic Stokes Equations

This section is devoted to the generalization to the anisotropic case of the results presented in the previous section. Firstly, we study a way to generalize to the anisotropic case the functional framework which was introduced in Sect. 2.2 for Stokes equations. Then, in Sect. 3.2, we apply this new framework for defining a SIP DG scheme for the approximation of (7)–(9) and show an inf-sup inequality, similar to (25) which implies stability.

## *3.1 SIP DG Framework for Anisotropic Stokes Equations*

As in the previous section, we consider the same discrete spaces for velocity and pressure (all of them with the same polynomial order), while the following notation emphasizing the anisotropy of the velocity $\mathbf{w}_h = (\mathbf{u}_h, v_h)$ is applied:

$$\mathbf{U}_h = (\mathcal{P}_h^k)^{d-1}, \quad V_h = \mathcal{P}_h^k, \quad \mathbf{W}_h = \mathbf{U}_h \times V_h = (\mathcal{P}_h^k)^d, \quad P_{h,0} = \mathcal{P}_h^k \cap L_0^2(\Omega).$$

For each $\mathbf{w}_h = (\mathbf{u}_h, v_h)$ and $\overline{\mathbf{w}}_h = (\overline{\mathbf{u}}_h, \overline{v}_h) \in \mathbf{W}_h$, with $\mathbf{u}_h = (u_i)_{i=1}^{d-1}$ and $\overline{\mathbf{u}}_h = (\overline{u}_i)_{i=1}^{d-1}$, we define the following bilinear form associated to (7)–(9):

$$a_h^{\text{anis}}(\mathbf{w}_h, \overline{\mathbf{w}}_h) = \sum_{i=1}^{d-1} a_h^{\text{sip},\eta}(u_i, \overline{u}_i) + \varepsilon\, a_h^{\text{sip},\eta}(v_h, \overline{v}_h) + (1-\varepsilon) s_h^v(v_h, \overline{v}_h) \tag{28}$$

where $s_h^v(v_h, \overline{v}_h)$ is the following bilinear form

$$s_h^v(v_h, \overline{v}_h) = \sum_{e \in \mathcal{E}_h} \frac{1}{h_e} \int_e \Big( [\![ v_h n_z ]\!]\, [\![ \overline{v}_h n_z ]\!] \Big).$$

We remark that, for horizontal velocity $\mathbf{u}_h$, it corresponds to the SIP DG bilinear form defined for Stokes in (22), but for $v_h$ we introduce a new bilinear form taking into account the jump of $v_h$ in the $z$-normal direction. This former bilinear form (28) generalizes the hydrostatic ($\varepsilon = 0$) bilinear form which was analyzed in [13]. The effects of the presence or absence of this term is illustrated in numerical tests at Sect. 4.

The next step consists in the introduction of a suitable norm on $\mathbf{w}_h = (\mathbf{u}_h, v_h) \in \mathbf{W}_h$ for which a coercivity result of $a_h^{\text{anis}}(\cdot,\cdot)$ can be obtained. Note that Lemma 2 can be applied to obtain the following partial coercivity,

$$\sum_{i=1}^{d-1} a_h^{\text{sip},\eta}(u_i, \overline{u}_i) \geq C(\eta)\|\mathbf{u}_h\|_{\text{sip}}^2, \tag{29}$$

for $\|\mathbf{u}_h\|_{\text{sip}}^2 := \sum_{i=1}^{d-1} \|u_i\|_{\text{sip}}^2$. Then, it is easy to show:

**Lemma 3 ($\varepsilon$-Dependent Coercivity)** *For all $\eta > \eta_*$, with $\eta_*$ given in Lemma 2, and for all $\mathbf{w}_h = (\mathbf{u}_h, v_h) \in \mathbf{W}_h$,*

$$a_h^{\text{anis}}(\mathbf{w}_h, \mathbf{w}_h) \geq C(\eta)\big(\|\mathbf{u}_h\|_{\text{sip}}^2 + \varepsilon\|v_h\|_{\text{sip}}^2\big) + (1-\varepsilon)|v_h|_V^2 \tag{30}$$

*where $C(\eta)$ is given in Lemma 2 and*

$$|v_h|_V^2 = \sum_{e\in\mathcal{E}_h} \frac{1}{h_e}\int_e [\![ v_h n_z ]\!]^2. \tag{31}$$

Note that, applying (16), the following boundedness of $a_h^{\text{anis}}(\mathbf{w}_h, \overline{\mathbf{w}}_h)$ can be shown:

$$\begin{aligned} a_h^{\text{anis}}(\mathbf{w}_h, \overline{\mathbf{w}}_h) \leq C_{\text{bnd}}\big(&\|u_h\|_{\text{sip}}\|\overline{u}_h\|_{\text{sip}} + \varepsilon\, \|v_h\|_{\text{sip}}\|\overline{v}_h\|_{\text{sip}}\big) \\ &+(1-\varepsilon)|v_h|_V|\overline{v}_h|_V. \end{aligned} \tag{32}$$

Lemma 3 presents only $\varepsilon$-dependent coercivity respect to the norm $\|v_h\|_{\text{sip}}^2$, but we will get a uniform coercivity respect to $\partial_z v_h$ in $L^2(\Omega)$, whose deviation from the mean will be bounded in terms of $p_h$:

**Lemma 4 (Stability for $\partial_{z,h} v_h$ in $L^2(\Omega)$)** *For every $v_h \in V_h$, it holds*

$$\|\partial_{z,h} v_h - \langle \partial_{z,h} v_h\rangle_\Omega\|_0 = \sup_{\overline{p}_h\in P_{h,0}} \frac{\int_\Omega \overline{p}_h\, \partial_{z,h}\, v_h}{\|\overline{p}_h\|_0}, \tag{33}$$

*where $\langle \partial_{z,h} v_h\rangle_\Omega$ is the mean of $\partial_{z,h} v_h$ in $\Omega$.*

*Proof* Given $v_h \in V_h$, and using that $\int_\Omega \overline{p}_h \, \langle \partial_{z,h} v_h \rangle_\Omega = 0$ for any $\overline{p}_h \in P_{h,0}$,

$$\sup_{\overline{p}_h \in P_{h,0}} \frac{\int_\Omega \overline{p}_h \, \partial_{z,h} \, v_h}{\|\overline{p}_h\|_0} = \sup_{\overline{p}_h \in P_{h,0}} \frac{\int_\Omega \overline{p}_h \left( \partial_{z,h} \, v_h - \langle \partial_{z,h} v_h \rangle_\Omega \right)}{\|\overline{p}_h\|_0}.$$

Therefore, it suffices to note that the supreme on the right hand side is reached for $\overline{p}_h = \partial_{z,h} v_h - \langle \partial_{z,h} v_h \rangle_\Omega \in P_{h,0}$.

We define the norm

$$\|v_h\|_{\text{anis},\varepsilon} = \left( \varepsilon \|v_h\|_{\text{sip}}^2 + (1-\varepsilon) \|\partial_{z,h} v_h\|_0^2 + (1-\varepsilon) |v_h|_V^2 \right)^{1/2}, \tag{34}$$

where $\partial_{z,h}$ is the broken vertical derivative (defined similarly to $\nabla_h$). Taking into account equality

$$[\![ v_h ]\!] \, [\![ \overline{v}_h ]\!] = [\![ v_h \mathbf{n}_\mathbf{x} ]\!] \cdot [\![ \overline{v}_h \mathbf{n}_\mathbf{x} ]\!] + [\![ v_h n_z ]\!] \, [\![ \overline{v}_h n_z ]\!], \tag{35}$$

it is verified

$$\|v_h\|_{\text{anis},\varepsilon}^2 = \varepsilon \|\nabla_{\mathbf{x},h} v_h\|_0^2 + \|\partial_{z,h} v_h\|_0^2 + \sum_{e \in \mathcal{E}_h} \frac{1}{h_e} \int_e \left( \varepsilon \, [\![ v_h \mathbf{n}_\mathbf{x} ]\!]^2 + [\![ v_h n_z ]\!]^2 \right).$$

In particular, since $\varepsilon \le 1$, one has

$$\|v_h\|_{\text{anis},\varepsilon}^2 \le \|v_h\|_{\text{sip}}^2.$$

Similarly, we consider the following anisotropic velocity norm:

$$\|\mathbf{w}_h\|_{\text{anis},\varepsilon} = \left( \|\mathbf{u}_h\|_{\text{sip}}^2 + \|v_h\|_{\text{anis},\varepsilon}^2 \right)^{1/2}. \tag{36}$$

At this point, stability of the discrete problem, uniformly on $\varepsilon \to 0$, hinges on a bound of $\|p_h\|_0$. We adapt to the anisotropic norm (36) the SIP-DG discrete inf-sup condition which was introduced in (21):

**Lemma 5 (Stability for $p_h$)** *There exists $\gamma_p > 0$ independent of h, such that*

$$\gamma_p \, \|p_h\|_0 \le \sup_{\mathbf{w}_h \in \mathbf{W}_h \setminus \{0\}} \frac{b_h^{\text{Sto}}(\mathbf{w}_h, p_h)}{\|\mathbf{w}_h\|_{\text{anis},\varepsilon}} + |p_h|_P, \quad \forall p_h \in P_{h,0}. \tag{37}$$

The proof lies in the discrete DG inf-sup condition for Stokes equations (21) and the fact that (since $\varepsilon \le 1$)

$$\|\mathbf{w}_h\|_{\text{anis},\varepsilon}^2 \le \|\mathbf{w}_h\|_{\text{sip}}^2 = \|\mathbf{u}_h\|_{\text{sip}}^2 + \|v_h\|_{\text{sip}}^2.$$

See [13] for details in the particular case $\varepsilon = 0$. For $b_h^{\mathrm{Sto}}(\mathbf{w}_h, p_h)$, it is not difficult to show continuity with respect to $\|\mathbf{w}_h\|_{\mathrm{anis},\varepsilon}$.

### 3.2 Stability and Well-Posedness of the Discrete Scheme

Similarly to the isotropic SIP DG Stokes discretization (22)–(23) in Sect. 2.2, let us consider the following discretization of the Anisotropic Stokes problem (7)–(9): find $(\mathbf{w}_h, p_h) \in \mathbf{W}_h \times P_{h,0}$ such that

$$a_h^{\mathrm{anis}}(\mathbf{w}_h, \overline{\mathbf{w}}_h) + b_h^{\mathrm{Sto}}(\overline{\mathbf{w}}_h, p_h) = \int_\Omega \mathbf{f} \cdot \overline{\mathbf{u}}_h + \int_{\Gamma_s} \mathbf{g}_s \cdot \overline{\mathbf{u}}_h, \qquad \forall\, \overline{\mathbf{w}}_h = (\overline{\mathbf{u}}_h, v_h) \in \mathbf{W}_h, \tag{38}$$

$$-b_h^{\mathrm{Sto}}(\mathbf{w}_h, \overline{p}_h) + s_h^p(p_h, \overline{p}_h) = 0, \quad \forall\, \overline{p}_h \in P_{h,0}, \tag{39}$$

where $s_h^p(p_h, \overline{p}_h)$ is defined in (24).

Also similarly to the Stokes case, the above formulation can be rewritten in a vector form as follows: find $(\mathbf{w}_h, p_h) \in \mathbf{W}_h \times P_{h,0}$ such that

$$c_h^{\mathrm{anis}}\big((\mathbf{w}_h, p_h), (\overline{\mathbf{w}}_h, \overline{p}_h)\big) = \int_\Omega \mathbf{f} \cdot \overline{\mathbf{u}}_h + \int_{\Gamma_s} \mathbf{g}_s \cdot \overline{\mathbf{u}}_h, \tag{40}$$

for all $(\overline{\mathbf{w}}_h, \overline{p}_h) \in \mathbf{W}_h \times P_{h,0}$, where

$$c_h^{\mathrm{anis}}\big((\mathbf{w}_h, p_h), (\overline{\mathbf{w}}_h, \overline{p}_h)\big) = a_h^{\mathrm{anis}}(\mathbf{w}_h, \overline{\mathbf{w}}_h) + b_h^{\mathrm{Sto}}(\overline{\mathbf{w}}_h, p_h) - b_h^{\mathrm{Sto}}(\mathbf{w}_h, \overline{p}_h) + s_h^p(p_h, \overline{p}_h). \tag{41}$$

We consider the following norm in $\mathbf{X}_h = \mathbf{W}_h \times P_{h,0}$:

$$\|(\mathbf{w}_h, p_h)\|_{\varepsilon,\mathbf{X}_h}^2 = \|\mathbf{w}_h\|_{\mathrm{anis},\varepsilon}^2 + \|p_h\|_0^2 + |p_h|_P^2 = \|\mathbf{u}_h\|_{\mathrm{sip}}^2 + \|v_h\|_{\mathrm{anis},\varepsilon}^2 + \|p_h\|_0^2 + |p_h|_P^2.$$

According to Banach-Necas-Babuška theorem (see e.g. [9]) stability of discrete problem (38)–(39) follows from the following inf-sup result for $c_h^{\mathrm{anis}}(\cdot,\cdot)$.

**Theorem 1 (Discrete inf-sup Stability)** *Assume that the penalty parameter $\eta$ in $a_h^{sip,\eta}(\cdot,\cdot)$ is such that $\eta > \eta_*$, with $\eta_*$ defined in Lemma 2. Then, there exists $\gamma > 0$ independent of $h$ such that, for all $(\mathbf{w}_h, p_h) \in \mathbf{X}_h = \mathbf{W}_h \times P_{h,0}$, one has*

$$\gamma\, \|(\mathbf{w}_h, p_h)\|_{\varepsilon,\mathbf{X}_h} \le \sup_{(\overline{\mathbf{w}}_h, \overline{p}_h) \in \mathbf{X}_h \setminus \{0\}} \frac{c_h^{\mathrm{anis}}((\mathbf{w}_h, p_h), (\overline{\mathbf{w}}_h, \overline{p}_h))}{\|(\overline{\mathbf{w}}_h, \overline{p}_h)\|_{\varepsilon,\mathbf{X}_h}}. \tag{42}$$

*Remark 1* Since for all $\varepsilon > 0$,

$$\begin{aligned}\|(\mathbf{w}_h, p_h)\|^2_{\varepsilon,\mathbf{X}_h} \geq& \|(\mathbf{w}_h, p_h)\|^2_{0,\mathbf{X}_h}\\ =& \|\mathbf{u}_h\|^2_{\text{sip}} + \|\partial_{z,h} v_h\|^2_0 + |v_h|^2_V + \|p_h\|^2_0 + |p_h|^2_P,\end{aligned}$$

then (42) implies

$$\gamma \, \|(\mathbf{w}_h, p_h)\|_{\varepsilon,\mathbf{X}_h} \leq \sup_{(\overline{\mathbf{w}}_h, \overline{p}_h) \in \mathbf{X}_h \setminus \{0\}} \frac{c_h^{\text{anis}}((\mathbf{w}_h, p_h), (\overline{\mathbf{w}}_h, \overline{p}_h))}{\|(\overline{\mathbf{w}}_h, \overline{p}_h)\|_{0,\mathbf{X}_h}}.$$

Consequently, scheme (38)–(39) is stable uniformly on $\varepsilon \to 0$.

*Proof (Proof of Theorem 1)*

A particular case of this theorem ($\varepsilon = 0$) was shown in [13]. Here we outline that proof, focusing only on the new difficulties related with the general case $\varepsilon \neq 0$. For any $(\mathbf{w}_h, p_h) \in \mathbf{X}_h$, we denote $S(\mathbf{w}_h, p_h)$ the supreme given in the right hand side of (42). Owing to (30),

$$\begin{aligned}c_h^{\text{anis}}((\mathbf{w}_h, p_h), (\mathbf{w}_h, p_h)) &= a_h^{\text{anis}}(\mathbf{w}_h, \mathbf{w}_h) + s_h^p(p_h, p_h)\\ &\geq C(\eta)(\|\mathbf{u}_h\|^2_{\text{sip}} + \varepsilon\|v_h\|^2_{\text{sip}}) + (1-\varepsilon)|v_h|^2_V + |p_h|^2_P\end{aligned}$$

therefore

$$\begin{aligned}C(\eta)(\|\mathbf{u}_h\|^2_{\text{sip}} + \varepsilon\|v_h\|^2_{\text{sip}}) + (1-\varepsilon)|v_h|^2_V + |p_h|^2_P&\\ \leq S(\mathbf{w}_h, p_h)\|(\mathbf{w}_h, p_h)\|_{\varepsilon,\mathbf{X}_h}.& \qquad (43)\end{aligned}$$

The rest of the proof is divided into the following steps.

*Step 1: Estimate of* $\|p_h\|_0$ Definition of $c_h^{\text{anis}}(\cdot,\cdot)$ means that, for all $\overline{\mathbf{w}}_h \in \mathbf{W}_h$,

$$b_h^{\text{Sto}}(\overline{\mathbf{w}}_h, p_h) = c_h^{\text{anis}}((\mathbf{w}_h, p_h), (\overline{\mathbf{w}}_h, 0)) - a_h^{\text{anis}}(\mathbf{w}_h, \overline{\mathbf{w}}_h),$$

therefore inf-sup condition (37) implies

$$\gamma_p \, \|p_h\|_0 \leq \sup_{\overline{\mathbf{w}}_h \in \mathbf{w}_h} \frac{-a_h^{\text{anis}}(\mathbf{w}_h, \overline{\mathbf{w}}_h)}{\|\overline{\mathbf{w}}_h\|_{\text{anis},\varepsilon}} + S(\mathbf{w}_h, p_h) + |p_h|_P.$$

Boundedness of $a_h^{\text{anis}}(\cdot,\cdot)$ for $\|\cdot\|_{\text{anis},\varepsilon}$ follows from (32),

$$a_h^{\text{anis}}(\mathbf{w}_h, \overline{\mathbf{w}}_h) \leq C_{\text{bnd}}\Big(\|\mathbf{u}_h\|_{\text{sip}}\|\overline{\mathbf{u}}_h\|_{\text{sip}} + \varepsilon\|v_h\|_{\text{sip}}\|\overline{v}_h\|_{\text{sip}}\Big) + (1-\varepsilon)|v_h|_V|\overline{v}_h|_v,$$

so that

$$\|p_h\|_0 \lesssim \|\mathbf{u}_h\|_{\rm sip} + \varepsilon\|v_h\|_{\rm sip} + (1-\varepsilon)|v_h|_V + S(\mathbf{w}_h, p_h) + |p_h|_P. \tag{44}$$

Here and below, we used the following notation: $\Phi \lesssim \Psi$ if $\Phi \le C\,\Psi$ for some constant $C > 0$ independent of $h$ (and $\varepsilon$).

*Step 2: Estimate of* $\|\partial_{z,h} v - \langle \partial_{z,h} v_h\rangle_\Omega\|_0$ Definition of $c_h^{\rm anis}(\cdot,\cdot)$ and $b_h^{\rm Sto}(\cdot,\cdot)$ yields

$$\int_\Omega \overline{p}_h\,\partial_{z,h} v_h = -\int_\Omega \overline{p}_h\,\nabla_{\mathbf{x},h}\cdot\mathbf{u}_h + \sum_{e\in\mathcal{E}_h}\int_e [\![\mathbf{w}_h]\!]\cdot\mathbf{n}_e\ \{\!\{\overline{p}_h\}\!\}$$
$$+\, c_h^{\rm anis}((\mathbf{w}_h, p_h), (0, \overline{p}_h)) - s_h^p(p_h, \overline{p}_h)$$

for all $\overline{p}_h \in P_{h,0}$.

Applying inf-sup condition (33) one can proceed like in [13] obtaining

$$\|\partial_{z,h} v_h - \langle\partial_{z,h} v_h\rangle_\Omega\|_0 \lesssim \|\mathbf{u}_h\|_{\rm sip} + \|v_h\|_{{\rm anis},\varepsilon} + S(\mathbf{w}_h, p_h) + \|p_h\|_0 \tag{45}$$

*Step 3: Estimate of* $\|\langle\partial_{z,h} v_h\rangle_\Omega\|_0$ *by vertical jumps* Using (11) and (31) like in [13]:

$$\left|\int_\Omega \partial_{z,h} v_h\right| = \left|\sum_{K\in\mathfrak{T}_h}\int_K \partial_z v_h\right| \le \sum_{e\in\mathcal{E}_h}\int_e |\,[\![v_h n_z]\!]\,|$$
$$\le \Bigg(\sum_{e\in\mathcal{E}_h}\frac{1}{h_e}\int_e [\![v_h n_z]\!]^2\Bigg)^{1/2}\Bigg(\sum_{e\in\mathcal{E}_h} h_e\Bigg)^{1/2} \le C\,|\Omega|\,|v_h|_V,$$

therefore

$$\|\langle\partial_z v_h\rangle_\Omega\|_0 \le C|v_h|_V. \tag{46}$$

*Step 4:* Gathering (43)–(46), one has

$$\|(\mathbf{w}_h, p_h)\|^2_{\varepsilon,\mathbf{X}_h} \lesssim S(\mathbf{w}_h, p_h)\,\|(\mathbf{w}_h, p_h)\|_{\varepsilon,\mathbf{X}_h} + S(\mathbf{w}_h, p_h)^2.$$

The conclusion follows from Young's inequality. □

# 4 Numerical Tests

We have developed some numerical tests which agree with the stability results shown in previous sections. They are just $2D$ experiments, developed with the intention of supporting those theoretical results and in particular we are only looking for an adequate qualitative behavior of solutions. A more detailed suite of numerical tests, including 3D experiments, is left for a future work.

We used the FreeFem++ language [2, 14] to program three cavity tests, for $\varepsilon = 0$, $\varepsilon = 10^{-8}$ and $\varepsilon = 1$. In all cases we used discontinuous $\mathcal{P}_1$ elements for velocity and pressure and introduced the following parameters: $\Omega = (0, 1)^2 \subset \mathbb{R}^2$, unstructured mesh with $h \approx 1/30$, horizontal viscosity $\nu = 1$, and SIP stabilization parameter $\eta = 5$.

The following Dirichlet boundary conditions were used in all the experiments included in this section: $u_h = x(1 - x)$ and $v_h = 0$ on the surface boundary ($\Gamma_s = \{(x, 1) \ / \ x \in (0, 1)\}$); $u_h = 0$ and $v_h = 0$ on bottom ($\Gamma_b = \{(x, 0) \ / \ x \in (0, 1)\}$); $u_h = 0$ on lateral walls ($\Gamma_l = \{(x, y) \in \mathbb{R}^2 \ / x \in \{0, 1\}, y \in (0, 1)\}$). We consider the homogeneous Neumann boundary condition $\varepsilon \nabla v_h \cdot \mathbf{n} = 0$ on $\Gamma_l$. As $\mathbf{n} = (1, 0)$ or $\mathbf{n} = (-1, 0)$ on $\Gamma_l$, this represents the slip boundary condition $\varepsilon \partial_x v_h = 0$. Note that it is not related to the Dirichlet boundary condition $v_h = 0$, that would not make sense in the limit case $\varepsilon = 0$ (in which $v_h$ loses regularity).

As usual in DG methods, former boundary conditions are fixed weakly. Specifically, let us consider the bilinear form $a_h^{\text{sip}}(\cdot, \cdot)$, which is utilized in (28) for the components of velocity, and also the bilinear form $b_h^{\text{Sto}}(\cdot, \cdot)$ which couples velocity and pressure. We introduce in the bilinear forms those terms which are related to means and jumps of the solution in interior and in Dirichlet boundary edges. See [13] for details on implementation.

Also, for each term regarding to jumps and averages on Dirichlet boundary edges of the velocity, a corresponding term which includes the corresponding Dirichlet boundary value is introduced in the right hand side of the discrete equation. As Dirichlet boundary conditions are zero except $u|_{\Gamma_s}$, the only additional terms correspond to:

$$\sum_{e \in \mathcal{E}_h \cap \Gamma_s} \int_e x(1 - x)(-\nabla \overline{u}_h \cdot \mathbf{n} + \frac{\eta}{h_e} \overline{u}_h - n_x \overline{p}_h) ds$$

(former expression can be simplified even more, considering that $n_x = 0$ on $\Gamma_s$).

On the other hand, boundary terms regarding to averages of normal derivatives of velocity which appear in $a_h^{\text{sip}}(\cdot, \cdot)$, can be passed to the RHS and used to apply Neumann boundary conditions in a standard way. Since we only have one homogeneous Neumann condition, $\varepsilon \nabla v_h \cdot \mathbf{n} = 0$ on $\Gamma_l$, we add nothing to the RHS.

For our first numerical test, we fix $\varepsilon = \nu = 1$. Then we have no anisotropy and resulting velocity field and pressure iso-values reproduce the expected behaviour in a standard Stokes cavity test (Fig. 1).

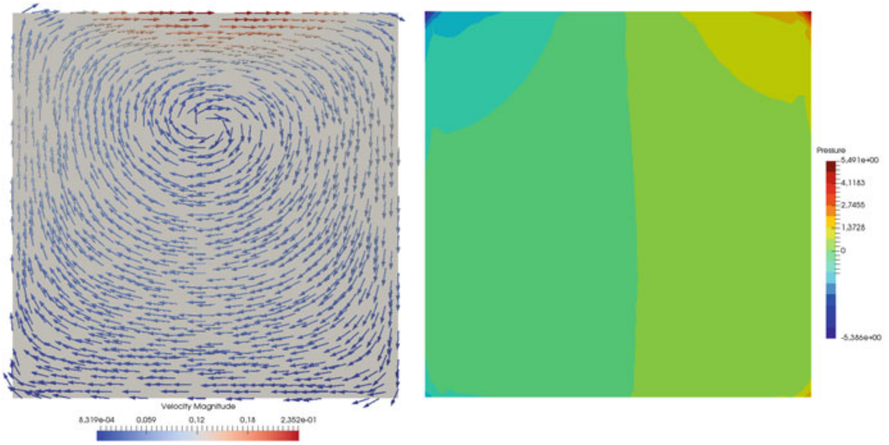


**Fig. 1** Cavity test, $\varepsilon = 1$

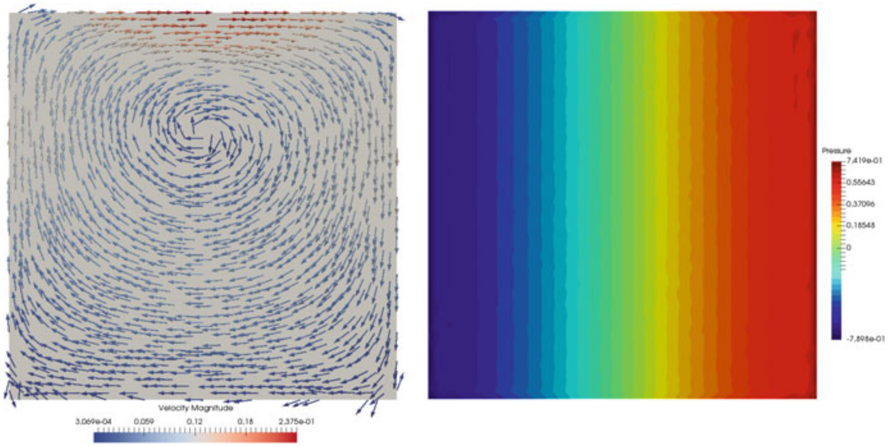


**Fig. 2** Cavity test, $\varepsilon = 10^{-8}$

In the second numerical test (Fig. 2) we fix the value $\varepsilon = 10^{-8}$. Note that, in the case of anisotropic equations derived from large scale geophysical fluid dynamics, this value would correspond to a realistic domain scale ratio, $\widehat{\varepsilon} =$ (vertical scale)/(horizontal scale) $= 10^{-4}$, see Sect. 1. Fluid is almost hydrostatic because diffusion in vertical momentum equation is negligible. Velocity field recirculation and hydrostatic (vertical) iso-values for pressure are obtained, according to expected qualitative behaviour.

We have also done an experiment with the value $\varepsilon = 0$, i.e. where fluid is purely hydrostatic. Graphics are not plotted because results are very similar to the case $\varepsilon = 10^{-8}$ (Fig. 2).

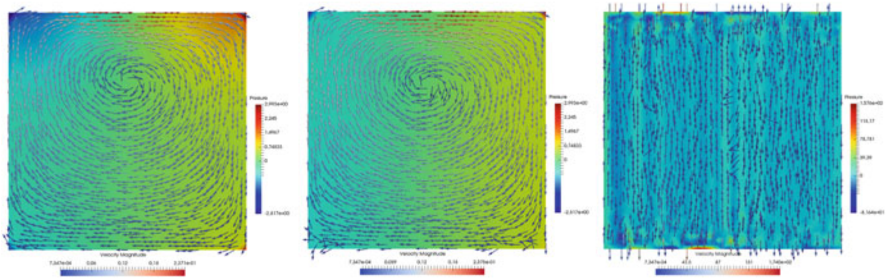


**Fig. 3** Cavity test, without the term $(1-\varepsilon)s_h^v(v_h, \overline{v}_h)$ for $\varepsilon = 1$, $\varepsilon = 10^{-2}$ and $\varepsilon = 10^{-4}$, respectively

Finally, we present a test where we illustrate numerically the necessity of addition of the anisotropic term

$$(1-\varepsilon)s_h^v(v_h, \overline{v}_h)$$

in the bilinear form $a_h^{\text{anis}}(\cdot,\cdot)$ (see (28)). Specifically, in Fig. 3 we present some simulations where this term is not included in the bilinear form. Results are good for the isotropic case $\varepsilon = 1$ but velocity is unstable in the anisotropic case $\varepsilon = 10^{-4}$.

**Acknowledgements** The first author has been partially financed by the MINECO grant MTM2015-69875-P (Spain) with the participation of FEDER. The second and third authors are also partially supported by the research group FQM-315 of Junta de Andalucía.

## References

1. Arnold, D.N.: An interior penalty finite element method with discontinuous elements. SIAM J. Numer. Anal. **19**, 742–760 (1982)
2. Auliac, S., Le Hyaric, A., Morice, J., Hecht, F., Ohtsuka, K., Pironneau, O.: FreeFem++. Third Edition, Version 3.31–2 (2014). http://www.freefem.org/ff++/ftp/freefem++doc.pdf
3. Azérad, P.: Analyse et approximation du problème de Stokes dans un bassin peu profond. C. R. Acad. Sci. Paris Sér. I Math. **318**, 53–58 (1994)
4. Azérad, P., Guillén, F.: Mathematical justification of the hydrostatic approximation in the primitive equations of geophysical fluid dynamics. SIAM J. Math. Anal. **33**, 847–859 (2001)
5. Besson, O., Laydi, M.R.: Some estimates for the anisotropic Navier-Stokes equations and for the hydrostatic approximation. Math. Mod. Num. Anal. **26**, 855–865 (1992)
6. Ciarlet, P.G.: The Finite Element Method for Elliptic Problems. North-Holland, Amsterdam (1978)
7. Cushman-Roisin, B., Beckers, J.M.: Introduction to Geophysical Fluid Dynamics - Physical and Numerical Aspects. Academic, London (2009)
8. Di Pietro, D.A., Ern, A.: Mathematical Aspects of Discontinuous Galerkin Methods. Springer, Berlin, New York (2012)
9. Ern, A., Guermond, J.L.: Theory and Practice of Finite Elements. Springer, New York (2004)

10. Guillén-González, F., Rodríguez-Galván, J.R.: Analysis of the hydrostatic Stokes problem and finite-element approximation in unstructured meshes. Numer. Math. **130**, 225–256 (2015)
11. Guillén González, F., Rodríguez Galván, J.R.: Stabilized schemes for the hydrostatic Stokes equations. SIAM J. Numer. Anal. **53**, 1876–1896 (2015)
12. Guillén-González, F., Rodríguez Galván, J.R.: On the stability of approximations for the Stokes problem using different finite element spaces for each component of the velocity. Appl. Numer. Math. **99**, 51–76 (2016)
13. Guillén González, F., Redondo Neble, M.V., Rodríguez Galván, J.R.: On stability of discontinuous galerkin approximations to the hydrostatic Stokes equations (2018, submitted)
14. Hecht, F.: New development in FreeFem++. J. Numer. Math. **20**, 251–265 (2012)
15. Kanschat, G.: Discontinuous Galerkin Methods for Viscous Incompressible Flow. Vieweg-Teubner, Wiesbaden (2008)
16. Rivière, B.: Discontinuous Galerkin methods for solving elliptic and parabolic equations: theory and implementation. Front. Appl. Math. SIAM, Philadelphia (2008)

# Numerical Simulation of Wear-Related Problems in a Blast Furnace Runner

**Patricia Barral, Begoña Nicolás, Luis Javier Pérez-Pérez, and Peregrina Quintela**

**Abstract** Two hydrodynamic problems related to the wear suffered by refractory linings at blast furnace runners during a stage of the steelmaking process are proposed. A thermo-hydrodynamic model is posed with the scope of finding the position of the critical isotherms inside the solid refractory layers. The computational domain is based on a runner at the ArcelorMittal Company, where the three-phase flow of slag, hot metal and air is solved using the SST $K-\omega$ turbulence model and the VOF method. Radiation heat transfer is accounted for using the S2S model. The impact of a jet of hot metal falling from the blast furnace on the runner is also analyzed using a similar hydrodynamic model. Shear stress, which is the main driving factor of the erosion rate, is computed at the impinging zone. Both models are solved using ANSYS Fluent.

**Keywords** Steelmaking · Simulation · Heat transfer · Radiation · Hydrodynamics · Multiphase · Free surface · Jet impact

## 1 Introduction

In the iron and steel industry, there are two major methods for steel production: *electric arc furnace steelmaking* and *basic oxygen steelmaking*. Our work is focused on a stage of the latter, which uses *hot metal* as feed material. Hot metal is produced

P. Barral · P. Quintela (✉)
Departamento de Matemática Aplicada, Universidade de Santiago de Compostela, Santiago de Compostela, Spain

Technological Institute for Industrial Mathematics (ITMATI), Santiago de Compostela, Spain
e-mail: patricia.barral@usc.es; peregrina.quintela@usc.es; peregrina.quintela@itmati.com

B. Nicolás · L. J. Pérez-Pérez
Departamento de Matemática Aplicada, Universidade de Santiago de Compostela, Santiago de Compostela, Spain
e-mail: bego.nicolas@usc.es; luisjavier.perez@usc.es

© Springer Nature Switzerland AG 2019

J. L. García Guirao et al. (eds.), *Recent Advances in Differential Equations and Applications*, SEMA SIMAI Springer Series 18,
https://doi.org/10.1007/978-3-030-00341-8_14

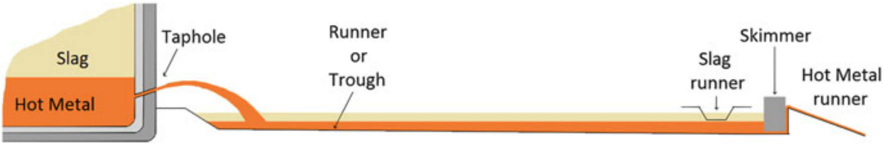


**Fig. 1** Schematic diagram of a longitudinal cut of the blast furnace runner

in the *blast furnace*, a large metallurgical furnace where some reductant sources like coke and coal are used to transform the iron-bearing components. In addition, the by-product known as slag is formed from ore gangue and coal ashes.

Both these materials, hot metal and slag, behave as liquids at the high operation temperatures, around 1500 °C, inside the hearth of the blast furnace. Their removal from the blast furnace takes place through the *taphole*, which is a drilled opening on its shell. After leaving the furnace, both fluids fall on a trench-like structure built with several layers of concrete refractories, known as the *blast furnace runner* or *main trough*. This structure is designed to enable separation of slag and hot metal as they flow downstream due to their density difference (see Fig. 1).

It is a major concern for the steel industry to minimize the wear suffered by refractory linings at blast furnace runners. Usually, the first layer of concrete, called the working layer, is completely replaced after 2 months of usage in order to prevent damage from extending to the remaining layers. This wear is due to a combination of factors: first, there is corrosion of the refractories due to the chemical attack of slag. Erosion also plays a role, both due to the falling jet impact on the trough and also to the flow of the fluids on it. Lastly, the high temperatures induce large thermal stresses. In fact, the wear of the runner is believed to be strongly related to the position of the *critical isotherms*, which are linked to the onset of chemical composition changes in the refractories. The nature of the problem, with extreme heat posing a big challenge to experimentation, motivates the use of numerical simulation in order to try to assess how these wear mechanisms affect the concrete refractories.

To the best knowledge of the authors, although there is a significant body of research about numerical simulation of processes inside the blast furnace, there is not much work available regarding the runner. Similar—albeit simpler—approaches to the critical isotherm computation were described in [13] and [15]. In the latter, a steady-state two-dimensional thermal problem was solved while in the former, a thermo-hydrodynamic model was solved for a symmetric three-dimensional domain. In [12], the hot metal jet behaviour travelling through the air and its mixing with the pool of hot metal was simulated. Additionally, in [11] the mechanical response of alumina refractory concrete used in a main trough was investigated under severe high temperature cyclic loading. Moreover, in [8], the characteristics of hot metal and slag flow in the runner were studied with scale models.

The scope of this work is twofold: first, to find the position of the critical isotherms inside the solid materials composing the runner by solving a thermo-hydrodynamic problem; and second, to study the shear stress exerted by the jet of hot metal on the runner, by solving a different hydrodynamic model that describes the dynamical behaviour of the jet. Shear stress is the main driving factor of mechanical erosion (see [2]). Thus, the two problems are treated separately in the subsequent discussion, although both of them are related to the described wear mechanisms.

## 2 Physical Problem

The blast furnace runner is composed of several layers of concrete refractories and a steel frame, as depicted in Fig. 2. The runner is covered with a refractory cover, which totally stops radiation emitted by the slag free surface from escaping to the surroundings, enhancing heat transfer towards the refractory vessel. The trough is refrigerated with air that circulates freely around both the steel frame and the cover. Additionally, air is aspirated next to the taphole in order to capture dust particles that can be ejected out of the furnace. The function of the cover is not only to avoid excessive cast iron cooling, but also to ensure the effective aspiration of these dust particles. Therefore, there is a forced flow of air below the cover that goes countercurrent in comparison to slag and hot metal.

To model the solid refractories composing the different layers, the constant values shown in Table 1 are used, according to the data supplied by ArcelorMittal. We denote the thermal conductivity, the specific heat, and the density as $k$, $c_p$ and $\rho$ respectively. For more details about refractories and their composition, we refer to [9]. Fluids are assumed to be incompressible and immiscible, with constant properties provided by ArcelorMittal, shown in Table 2. We denote the dynamic viscosity as $\mu$. These different fluids are regarded as distinct phases of the flow, separated by free surfaces whose position has to be found.

The tapping of the blast furnace is not a continuous process, as it is done in cycles. Typically, each cast lasts for about 90 min with its end being indicated by blast furnace gas bursting out through the taphole, the point at which it is plugged.

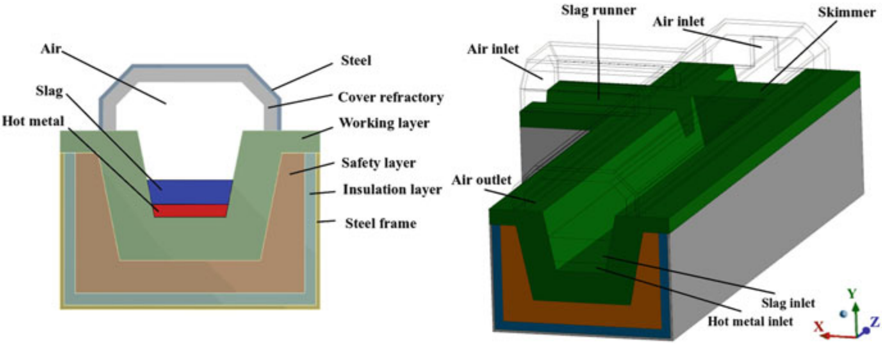


**Fig. 2** Transversal cut of the blast furnace runner (left). View of the computational domain for the thermo-hydrodynamic problem, outlining the cover and fluid subdomain (right)

**Table 1** Properties of the solid materials

| | $k$ (W/(m K)) | $c_p$ (J/(kg K)) | $\rho$ (kg/m$^3$) |
|---|---|---|---|
| Working layer | 2.60 | 1212 | 2500 |
| Safety layer | 1.80 | 1172 | 2900 |
| Insulation layer | 0.14 | 1050 | 480 |
| Cover refractory | 1.00 | – | 1950 |
| Steel | 54.00 | 465 | 7863 |

**Table 2** Properties of the fluid materials

| | $k$ (W/(m K)) | $c_p$ (J/(kg K)) | $\mu$ (kg/m$^3$) | $\rho$ (Pa s) |
|---|---|---|---|---|
| Hot metal | 16.50 | 850 | 7015.00 | $7.15e{-3}$ |
| Slag | 9.70 | 807 | 2600.00 | 0.50 |
| Air | $2.57e{-2}$ | 1005 | 1.21 | $1.82e{-5}$ |

Afterwards, there is a 15–30-min stop to allow the liquid level inside the hearth to rise back again. In spite of the stops, both experimental measures and numerical simulations show that the temperature in the bulk of the solids composing the runner reaches a steady state. Given that we are only interested in this final state of the temperatures for the computation of the critical isotherms, we consider a steady state thermal problem in Sect. 3.

Inside the hearth of the blast furnace, slag and hot metal are stratified, with slag remaining on top. Thus, at the beginning of each cast cycle only hot metal comes out. Moreover, given that there is a pool of fluids from prior casts covering the runner, the first cast is critical, as the jet of hot metal falls on the dry concrete, whereas in subsequent ones the impact is mitigated by the pool of fluids. In fact, it is believed that this first impact of the jet may cause high damage to the refractories. The computation of the corresponding shear stress is the main objective of Sect. 4.

## 3 Thermo-Hydrodynamic Problem

The scope of the thermo-hydrodynamic problem is to find the temperature field inside the runner in order to compute the position of the critical isotherms. A three-dimensional computational domain is chosen, as depicted in Fig. 2. It includes the latter half of the runner, part of the slag runner and the skimmer. With the purpose of ensuring that slag and hot metal are fully separated and stratified, the first part of the runner is not considered. The total length of the domain is 9.5 m, while the complete runner is about 15 m long. Moreover, it is 3 m wide and approximately 2 m deep.

The computational domain $\Omega$ is split in a fluid subdomain $\Omega_f$ and a solid subdomain $\Omega_s$, such that $\overline{\Omega} = \overline{\Omega}_f \cup \overline{\Omega}_s$. Since constant properties and incompressible fluids are considered, the energy equation is decoupled from the hydrodynamic model and can be solved afterwards to obtain the temperature field.

### 3.1 Free Surface Flow

In order to model the different phases of the flow, the *Volume of fluid method* (VOF) is used (see [5]). It is a *front-capturing method*, which means that the fluid phases are considered as a single fluid in a fixed domain, with properties described using piecewise constant functions. In the specific case of VOF, the characteristic function $\varphi_i : \Omega_f \times (0, T) \to \mathbb{R}$ indicates if phase $i$ is present at point $\mathbf{x}$ and time $t$:

$$\varphi_i(\mathbf{x}, t) = \begin{cases} 1, \ \mathbf{x} \in \Omega_{f,i}(t), \\ \\ 0, \ \mathbf{x} \notin \Omega_{f,i}(t), \end{cases}$$

where $\Omega_{f,i}(t)$ denotes the portion of $\Omega_f$ occupied by phase $i$ at time $t$. The points where $\varphi_i$ is discontinuous indicate the position of the free surface. Numerically, an approximation of $\varphi_i$ is taken on each computational cell, the so-called *volume fraction* $\alpha_i$:

$$\alpha_i = \frac{1}{\mathrm{Vol}(V)} \int_V \varphi_i dV,$$

where $\mathrm{Vol}(V)$ is the volume of the cell $V$. In virtue of this approximation, $\alpha_i$ is 1 when phase $i$ fills the cell and 0 whenever phase $i$ is not in the cell. Hence, any value between 0 and 1 indicates that a free surface is located inside the cell.

The properties of the fluids are taken such that the value of a scalar spatial field associated with an arbitrary property $\phi$ is

$$\tilde{\phi} = \sum_{i=1}^{N} \phi_i \alpha_i, \tag{1}$$

where $\phi_i$ is the property associated with the fluid phase $i$ and $N$ is the total number of phases, which for this problem is $N = 3$ (air, slag and hot metal). For a more detailed description of the VOF model, we refer to [14].

### 3.2 Hydrodynamic Model

The estimated Reynolds number of the flow is clearly above the critical values. Thus, with the aim of accounting for the turbulence-related phenomena without having to solve the full range of flow scales, the use of an adequate turbulence model is needed. In this problem, the SST $K - \omega$ turbulence model, proposed by Menter (see [10]), is used. The model includes two new variables, which are used to compute the *turbulent* or *eddy viscosity* $\mu_T$: $\omega$, known as *specific dissipation rate* and $K$, called the *turbulent kinetic energy*. The eddy viscosity and the turbulent kinetic energy are used to compute the value of the Reynolds stress tensor through the Boussinesq hypothesis, which describes the influence of the fluctuations in the mean flow (see [17]). The choice of the SST $K - \omega$ model is motivated on its well-documented performance in a wide range of wall-bounded flows, as it retains the advantages of the Standard $K - \omega$ and $K - \varepsilon$ models without suffering from some of their specific drawbacks. This is achieved by using blending functions that change the model depending on the wall distance.

Consequently, the PDE system to be solved in $\Omega_f$ is the following:

$$\operatorname{div}(\mathbf{V}) = 0, \tag{2}$$

$$\frac{\partial \alpha_i}{\partial t} + \operatorname{div}(\alpha_i \mathbf{V}) = 0, \quad 1 \leq i \leq N, \tag{3}$$

$$\begin{aligned} \frac{\partial (\tilde{\rho}\mathbf{V})}{\partial t} + \operatorname{div}(\tilde{\rho}\mathbf{V} \otimes \mathbf{V}) = \operatorname{div}\left(2\tilde{\mu}_T D(\mathbf{V}) - \frac{2}{3}\tilde{\rho} K \mathbf{I}\right) \\ - \operatorname{grad}(\Pi) + \operatorname{div}(2\tilde{\mu} D(\mathbf{V})) + \mathbf{f}, \end{aligned} \tag{4}$$

$$\frac{\partial (\tilde{\rho} K)}{\partial t} + \operatorname{div}(\tilde{\rho} K \mathbf{V}) = \operatorname{div}\left[\left(\tilde{\mu} + \frac{\tilde{\mu}_T}{\sigma_K}\right) \operatorname{grad}\,(K)\right] + G_K - Y_K, \tag{5}$$

$$\frac{\partial (\tilde{\rho} \omega)}{\partial t} + \operatorname{div}(\tilde{\rho} \omega \mathbf{V}) = \operatorname{div}\left[\left(\tilde{\mu} + \frac{\tilde{\mu}_T}{\sigma_\omega}\right) \operatorname{grad}(\omega)\right] + G_\omega - Y_\omega + D_\omega. \tag{6}$$

We denote as $\Pi$ and $\mathbf{V}$ the mean fields of the pressure $\pi$ and velocity $\mathbf{v}$, respectively, and as $D(\mathbf{V})$ the symmetric part of the mean velocity gradient. Furthermore, $\mathbf{f} = -\tilde{\rho} g \mathbf{e}_2$ is the body force due to gravity, where the orientation of the Cartesian coordinate axis is shown in Fig. 2. We omit the details of the turbulent constants $\sigma_K$ and $\sigma_\omega$, the $K$ and $\omega$ production terms $G_K$ and $G_\omega$, the dissipation terms $Y_K$ and $Y_\omega$ and the *cross-diffusion* term $D_\omega$. For the precise details on their definition for the SST $K - \omega$ turbulence model, see [1].

### 3.2.1 Initial Conditions

Since the main interest of the hydrodynamic model is to supply information to solve the steady state thermal problem, the initial conditions are chosen adequately to obtain a faster convergence to a quasi-steady state. We assume that the free surfaces are located initially at planes with constant $y$ coordinate, giving the initial condition for each $\alpha_i$. For the remaining flow variables, we set $\mathbf{V} = \mathbf{0}$ m/s, $K = 1$ m$^2$/s$^2$ and $\omega = 1$ s$^{-1}$.

### 3.2.2 Boundary Conditions

The position of inlets and outlets present at the computational domain is indicated in Fig. 2. We briefly summarize the main boundary conditions:

**Inlets** There are four different inlets in the domain, one for slag, another for hot metal and two for air. At each of these inlets, we set constant velocities following the normal direction. These velocities are extrapolated from daily slag and hot metal production, as detailed in [15]. Therefore, we take $\mathbf{V} \cdot \mathbf{n} = 0.074$ m/s and $\mathbf{V} \cdot \mathbf{n} = 0.035$ m/s, respectively. The air velocity is taken as $\mathbf{V} \cdot \mathbf{n} = 2$ m/s in both inlets. For the turbulent variables, standard values are chosen.

**Outlets** At outlets, pressure values are fixed. In the one corresponding to air, we set atmospheric pressure, whereas for slag and hot metal hydrostatic pressure profiles are used.

**Walls** For the walls, as a consequence of *no-slip* ($\mathbf{V} = \mathbf{0}$ m/s) taking place, we use wall laws for $\mathbf{V}$, $K$ and $\omega$. For more information about the wall laws and their implementation, we refer to [16].

## 3.3 Thermal Model

Once the hydrodynamic model is solved, the flow variables found as its solution are supplied to the steady state thermal problem. Turbulence plays a significant role, enhancing diffusive heat transport in the fluids due to the action of the velocity fluctuations. In order to account for this, the eddy viscosity $\mu_T$ is used to define the *turbulent conductivity* $k_T$, as detailed in [17]. Therefore, the equations that we solve in $\Omega_f \cup \Omega_s$ to obtain the temperature $\theta$ are the following:

$$\begin{aligned} &\operatorname{div}(\tilde{\rho}\tilde{c_p}\mathbf{V}\theta) = \operatorname{div}((\tilde{k} + \tilde{k}_T)\operatorname{grad}(\theta)), && \text{in } \Omega_f, \\ &\operatorname{div}(k\operatorname{grad}(\theta)) = 0, && \text{in } \Omega_s. \end{aligned} \tag{7}$$

At the interfaces between fluids and solids, continuity of both temperature and heat flux $\mathbf{q} = \tilde{k}\operatorname{grad}(\theta)$ has to be satisfied. This is true except for the boundary of the enclosure where the radiation heat transfer takes place, that is, the boundary of the air subdomain, which we denote as $\partial\Omega_{f,a}$. There, a radiation contribution $Q_{rad}$ to the heat flux is applied. Consequently, we assume that the *jump* of the heat flux at $\partial\Omega_{f,a}$, denoted as $[\mathbf{q}]_{\partial\Omega_{f,a}}$, verifies that:

$$[\mathbf{q}]_{\partial\Omega_{f,a}}(\mathbf{x}) \cdot \mathbf{n}_{f,a} = Q_{rad}(\mathbf{x}), \tag{8}$$

where $\mathbf{n}_{f,a}$ is the unit normal vector on $\partial\Omega_{f,a}$, pointing towards the exterior of the air subdomain. The jump is defined as follows:

$$[\mathbf{q}]_{\partial\Omega_{f,a}}(\mathbf{x}) := \lim_{\delta\to 0}\mathbf{q}(\mathbf{x}+\delta\mathbf{n}_{f,a}) - \lim_{\delta\to 0}\mathbf{q}(\mathbf{x}-\delta\mathbf{n}_{f,a}). \tag{9}$$

### 3.3.1 Radiation

A model accounting for radiation heat transfer has to be provided to compute the radiation contribution term $Q_{rad}$. With this purpose, we assume that both slag and hot metal are opaque to radiation, as are all the solids. Thus, radiation heat exchange takes place only in the region corresponding to the air subdomain, denoted by $\Omega_{f,a}$. It can be regarded as a *cavity* or *enclosure*, where the boundaries that are open to air flow—inlets and outlets—need special treatment. The boundary of the enclosure, denoted as $\partial\Omega_{f,a}$, is divided in $M$ open sets[1] $S_i$ in $\partial\Omega_{f,a}$, such that $\cup_{i=1}^{M}\overline{S_i} = \partial\Omega_{f,a}$.

The *surface-to-surface model* (S2S) is used. It is based on the *net radiation method*, described in [6], and has been used in similar problems, as in [7], where it enabled the computation of radiation fluxes inside an electric arc furnace. For each $i$, $S_i$ is assumed to be a gray, diffuse, opaque surface, characterized by a constant temperature. The interior of the cavity, which is filled with air, is considered to be totally transparent to incident radiation. Hence, the net radiation exchange between surfaces $i$ and $j$ may be obtained evaluating:

$$J_i = E_i + (1-\varepsilon_i)\sum_{j=1}^{M} F_{ij}J_j, \tag{10}$$

where $E_i = \sigma\varepsilon_i\theta_i^4$ is the *emissive power* on the surface $S_i$ according to Stefan-Boltzmann law,[2] $\theta_i$ is the temperature on the surface and $J_i$ is the outgoing radiation from the surface, known as *radiosity*. Lastly, $F_{ij}$ is the *view factor* between $i$ and $j$, which is computed evaluating the integral

$$F_{ij} = \frac{1}{A_i}\int_{A_i}\int_{A_j}\frac{\cos\psi_i\cos\psi_j}{\pi R^2}\delta_{ij}\,dA_j dA_i, \tag{11}$$

with $\psi_i$ being the plane angle between surface $i$ normal vector and the vector pointing from $i$ to $j$. $R$ denotes the distance among both surfaces, whereas the term

[1] These $M$ open sets, called *surfaces*, can be taken as the cell faces on the boundary, which arise from the discretization of the computational domain. Therefore, more accurate computations are obtained, given the constraint that each surface is assumed to have a constant temperature.

[2] With $\sigma$, we denote the Stefan-Boltzmann constant, i.e., $\sigma = 5.67\cdot 10^{-8}$ W/(m$^2$ K$^4$).

$\delta_{ij}$ accounts for the presence of other surfaces blocking the view. The view factor $F_{ij}$ represents the fraction of radiation leaving $i$ that is intercepted by surface $j$.

If the temperature and emissivity $\varepsilon_i$ at each surface are known, then equations (10) for all $i \in \{1, \ldots, M\}$ can be regarded as a linear system of $M$ equations and unknowns. Its solution yields the value of the radiosity, which then can be used to calculate the net heat flux due to radiation:

$$q_{rad,i} = J_i - \sum_{j=1}^{M} F_{ij} J_j. \tag{12}$$

Then, the $Q_{rad}(\mathbf{x})$ term, needed to compute the jump of the heat flux in (8), is obtained as a piecewise constant function:

$$Q_{rad}(\mathbf{x}) = q_{rad,i}, \tag{13}$$

where $\mathbf{x} \in S_i$, for each $i \in \{1, \ldots, M\}$.

### 3.3.2 Boundary Conditions

For the purpose of clarity, we distinguish between external walls, flow openings and other boundaries.

**External Walls** On the external walls of the runner, we set *convective* boundary conditions:

$$-k\frac{\partial \theta}{\partial \mathbf{n}} = h(\theta - \theta_{ext}),$$

with $h$ being the convection coefficients and $\theta_{ext} = 20\,°\mathrm{C}$ the external temperature. For lateral and bottom walls, the convection with the surrounding air is considered to be natural, and the estimated values of $h$ are 2 W/(m$^2$ K) for the bottom wall and 5 W/(m$^2$ K) for the lateral walls. In addition, the cover external wall is assumed to be cooled down by convection and by radiation emission to the environment:

$$-k\frac{\partial \theta}{\partial \mathbf{n}} = h(\theta - \theta_{ext}) + \sigma\varepsilon(\theta^4 - \theta_{ext,r}^4),$$

where $h = 15$ W/(m$^2$ K), $\theta_{ext} = \theta_{ext,r} = 293$ K (20 °C) and $\varepsilon = 0.7$.

**Flow Openings** On the flow openings corresponding to inlets, we set that $\theta = 1500\,°\mathrm{C}$ for hot metal and $\theta = 20\,°\mathrm{C}$ in the case of air. For slag, we use a profile which varies from 1327 °C at the free surface with air to 1500 °C at the free surface with hot metal. At the outlets, the normal temperature gradient is set to zero. Concerning the radiation model, $\varepsilon = 1$ is taken, which makes these surfaces behave as black bodies[3] emitting at $\theta_0 = 20\,^{\circ}\mathrm{C}$. In addition, for the air outlet, we choose an emissivity $\varepsilon = 0$. This means that we model it as a symmetry boundary for radiation, since all incoming radiation will be reflected in virtue of *Kirchhoff's law*, with no emission from the boundary.

**Other Boundaries** On the sections where the main trough and the slag runner are cut, we set zero heat flux.

## *3.4 Results*

The solution procedure is as follows: First, the transient hydrodynamic model is solved until a quasi-steady state is reached, which means that the computed position of all the free surfaces is stationary. Then, those surfaces corresponding to slag-air and hot metal-air are reconstructed as walls, splitting the fluid subdomain $\Omega_f$ in two regions, one corresponding to the region occupied by slag and hot metal and another for the air ($\Omega_{f,a}$). This enables an accurate computation of the view factors $F_{ij}$ (see 11), which are needed to account for the radiation heat transfer. Afterwards, the velocity and turbulence fields, obtained via the hydrodynamic model, have to be interpolated to a new mesh to solve the thermal one, described in Sect. 3.3, which provides the desired temperature field.

The proposed mathematical model is solved using the package *ANSYS Fluent* version 15.0, which uses a cell-centered finite volume discretization of the conservation equations. The pressure-based solver is used, with pressure-velocity coupling performed with the SIMPLE iterative algorithm. Face pressure values are interpolated using the *PRESTO!* scheme whereas gradient terms are discretized with the *least squares cell-based* scheme. Convective fluxes are integrated with *second order upwind schemes*. Advection and reconstruction of the free surfaces are done with the *geo-reconstruct* scheme (see [1]), which is a version of the well-known *PLIC* scheme proposed by Youngs (see [14]). The final mesh is composed of 21 million cells and is conformal between the fluid and solid subdomains.

The model was solved using 32 cores at a computer cluster composed of 25 Dell PowerEdge machines, ranging from R815 ones, equipped with AMD Opteron 6174 processors, to C6220 II systems with Intel Xeon E5-2637 v4 processors. Regarding computation times, the transient hydrodynamic model required 5 days, whereas the steady-state thermal one was solved in 7 h. This long computation time for a steady state problem owes to the relaxation factors that were required to ensure stability through the iterative procedure, as the radiation submodel is highly nonlinear.

In Fig. 3, the velocity modulus is shown along a longitudinal plane going through the middle section of the runner, including the streamlines for the air flow.

[3] This is reasonable, given that all radiation going out through those boundaries is expected to disperse around the casthouse and most incoming radiation would then be emitted by bodies at ambient temperature.

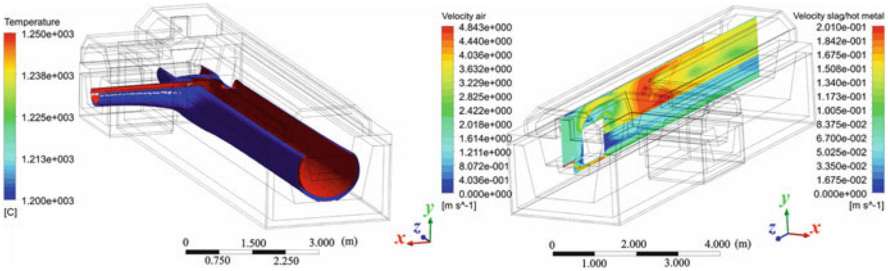


**Fig. 3** Critical isotherms computed position (left). Velocity modulus on a longitudinal cut of the runner and streamlines (right)

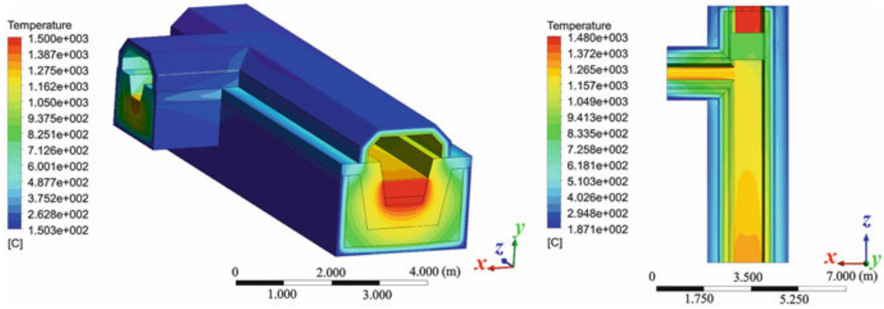


**Fig. 4** Temperature contours in the solid subdomain $\Omega_S$, hot metal and slag (left). Temperature contours in the slag-air free surface and the external part of the runner viewed from above (right)

The position of the critical isotherms (1200 and 1250 °C) is also depicted. Both isotherms are located inside the working lining, even though the 1200 °C one is close to the safety layer. In Fig. 4 (right), the temperature map in the computational domain is shown, seen from above and hiding both the air subdomain and the cover. Thus, the temperature at the slag-air and hot metal-air (after the skimmer) free surfaces is depicted, as well as the external part of the working layer. As slag flows downstream, its temperature decreases due to radiation emission and forced convection with air. However, hot metal remains at a high temperature during its voyage along the runner, as it is covered by the slag until it reaches the skimmer.

ArcelorMittal has provided temperature measurements obtained using six thermocouples embedded inside the insulation lining. These are shown in Table 3 alongside the computed values. Even though the agreement among the computed values and the measurements is good, it is not very significant because the thermocouples are placed far away from the working layer, where the highest temperatures are reached. This suboptimal placement is due to both economical

**Table 3** Thermocouple measurements, computed temperature, absolute error in (°C) and relative error

| | Measured | Numerical | Absolute error | Relative error |
|---|---|---|---|---|
| Thermocouple 1 | 576 | 596.2 | 20.2 | $3.5e{-}2$ |
| Thermocouple 2 | 669 | 617.9 | 51.1 | $7.6e{-}2$ |
| Thermocouple 3 | 755 | 767.4 | 12.4 | $1.6e{-}2$ |
| Thermocouple 4 | 770 | 759.1 | 10.9 | $1.4e{-}2$ |
| Thermocouple 5 | 644 | 645.8 | 1.8 | $2.8e{-}3$ |
| Thermocouple 6 | 626 | 625.8 | 0.2 | $3.2e{-}4$ |

concerns and access difficulties to better locations, as the thermocouples have to be placed during the main trough construction process.[4]

## 4 Jet Impact Problem

In order to assess how the jet impact affects the runner refractories, its dynamic behaviour has to be investigated. With the purpose of simulating and tracking the jet trajectory and its impact on the runner, a turbulent multiphase flow is considered. As described in Sect. 2, the first tapping of the blast furnace on a new working lining is believed to be critical, due to the jet of hot metal falling on dry concrete. Subsequent taps benefit from having a pool of fluids that dampens the impact.

The evolution of the free surface between the jet of hot metal and air is obtained using the same numerical techniques described in Sect. 3.1. In contrast with the previous problem, surface tension is computed using the *Continuum Surface Force* model (see [4]). The computational domain $\Omega_2$ is a longitudinal cut along the runner, as shown in Fig. 5 together with its boundaries.

### 4.1 Mathematical Model

The mathematical model of the jet is composed of the Navier-Stokes equations for an incompressible, turbulent, two-phase flow. Here, the Wilcox $K - \omega$ turbulence model is used (see [17]). Then, the same (2)–(6) equations are solved in $\Omega_2$, adjusting the turbulence constants and terms accordingly. The number of fluid

[4] It is also noteworthy that there is a degree of uncertainty on the actual position of the thermocouples. This is greatly worsened by the fact that the insulation layer has a low thermal conductivity (see Table 1), provoking that small deviations on the reported position of the thermocouples may imply high deviations on the computed temperature values due to the high temperature gradients.

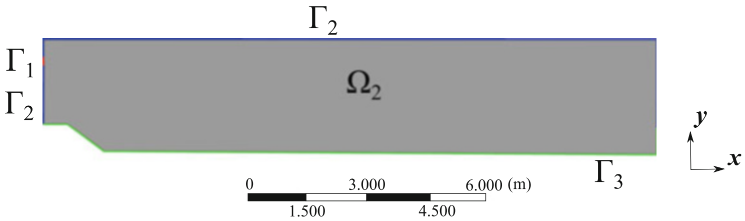


**Fig. 5** Computational domain $\Omega_2$. $\Gamma_1$ is the end of the taphole, $\Gamma_2$ corresponds to fictitious boundaries and $\Gamma_3$ is the runner surface

phases is $N = 2$, as we only consider hot metal and air. Surface tension is modeled adding a body force term $\mathbf{m}^{\sigma}$ to the momentum conservation equation.

### 4.1.1 Boundary Conditions

Concerning the boundary conditions (see Fig. 5), the following are taken:

- A velocity inlet condition is imposed on $\Gamma_1$, which corresponds to the taphole, using an estimation from production data, provided by ArcelorMittal. The modulus of the estimated velocity, $v_0 = 6.2$ m/s, is decomposed into its components in order to account for the 10° inclination of the taphole. The volume fraction for hot metal is equal to one, and standard values for turbulent variables are considered.
- On $\Gamma_2$ a fixed pressure condition is taken, with $\Pi = 0$ Pa.
- On the bottom surface of the runner, $\Gamma_3$, which is considered as a rigid wall, a *no-slip* wall condition is imposed. As the flow is turbulent, wall laws are used to compute both the velocity $\mathbf{V}$ and the turbulent variables $K$ and $\omega$.

### 4.1.2 Initial Conditions

To complete the mathematical model, it is necessary to provide initial conditions. Initial velocity is taken to be equal to zero all over the domain, whereas standard values are assumed for the turbulent variables: $K = 1$ m$^2$/s$^2$ and $\omega = 1$ s$^{-1}$.

## *4.2 Results*

For the numerical simulation of the jet behaviour, the model is also solved using the package *ANSYS Fluent*. Identical discretization schemes to those described in Sect. 3.4 are chosen. The computational mesh is composed of 120,000 cells and it is refined in the areas of interest, namely, the area where the jet is expected to travel

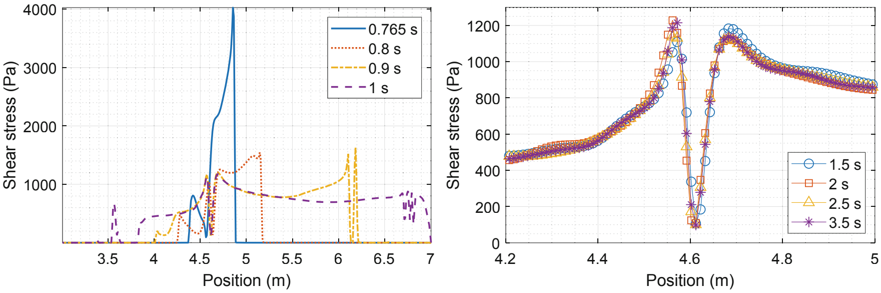


**Fig. 6** Shear stress profile on the runner (left), shear stress profile at the impact zone (right), for diverse time values

**Table 4** Maximum values of shear stress on both sides of the minimum, minimum values of shear stress and their position for different time values

| $t$<br>(s) | $\tau_{max_l}$<br>(Pa) | $x(\tau_{max_l})$<br>(m) | $\tau_{max_r}$<br>(Pa) | $x(\tau_{max_r})$<br>(m) | $\tau_{min}$<br>(Pa) | $x(\tau_{min})$<br>(m) |
|---|---|---|---|---|---|---|
| 1.5 | 1161.15 | 4.572 | 1132.57 | 4.683 | 100.75 | 4.612 |
| 2.0 | 1181.83 | 4.562 | 1121.66 | 4.683 | 102.07 | 4.612 |
| 2.5 | 1170.44 | 4.562 | 1125.61 | 4.683 | 100.20 | 4.612 |
| 3.0 | 1075.88 | 4.572 | 1115.30 | 4.683 | 96.42 | 4.612 |
| 3.5 | 1159.72 | 4.572 | 1129.19 | 4.683 | 96.28 | 4.612 |

through and the boundary layer on the runner, where an accurate computation of the velocity profile is needed to find the shear stress.

The computed time of impact is 0.765 s. At this instant, the point of the runner that suffers the maximum shear stress value is the one located 4.85 m away from the fluid inlet, where the shear stress is 4026 Pa. As time progresses, the shear stress becomes lower due to attenuation caused by hot metal being accumulated on the impact area.

In Fig. 6, shear stress profiles on the runner surface are depicted for several time values. Close to the impact zone, all curves show the same structure, with two maximums and a minimum in between. These profiles remain approximately constant and have a similar shape compared to those obtained experimentally by Beltaos and Rajaratnam (see [3]). They studied the shear stress produced by a normal submerged jet observing two symmetrical maximums on both sides of a minimum. This minimum corresponds to the impact point of the jet's centerline. In our case, the maximums are not completely symmetrical, which is due to the jet impact being oblique to the runner.

Minimum shear stress values, denoted as $\tau_{min}$, their location and evolution with time are collected in Table 4. The value and position of the maximum shear stress at the left ($\tau_{max_l}$) and right ($\tau_{max_r}$) of the minimum, as well as their positions are shown. All these positions remain essentially constant over time.

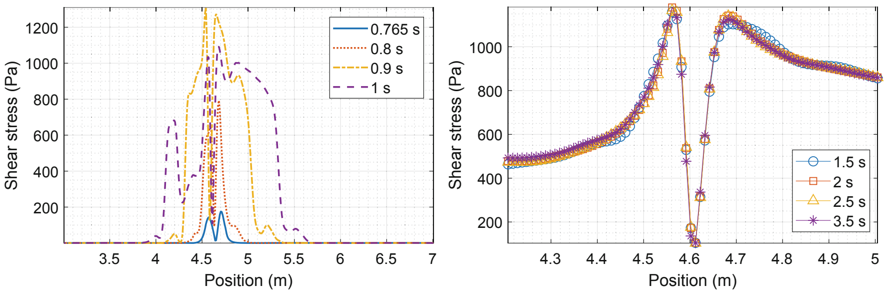


**Fig. 7** Shear stress profile on the runner (left), shear stress profile at the impact zone (right), for diverse time values considering a pool of hot metal covering the runner

Additional simulations were performed considering the presence of a 10-cm deep pool of hot metal. Results are shown in Fig. 7. It is significant that there are almost no differences between the shear stress values obtained in both cases over time.

## 5 Conclusion

The proposed decoupling methodology was used to successfully obtain the temperature field in the main trough of a blast furnace. A reasonable qualitative agreement with the measured temperatures was obtained despite the simplifications that were made. Nevertheless, the computed temperature field suggests that solidification may take place on the slag-air free surface, which would imply that the assumptions made about the slag flow being incompressible should be changed in future studies. Furthermore, the constant velocity inlet conditions are also likely to be unrealistic, since the flow of slag and hot metal could still be affected by the jet of slag and hot metal falling from the taphole, as evidenced by the results of the jet impact problem. Future work also aims to pose more realistic inlet conditions.

Regarding the jet impact on the runner, it was observed that the maximum values of shear stress were obtained for the impact instant, followed by a rapid decrease until a steady state is reached. The numerical shear stress profiles are in good agreement with those found in the bibliography for jets impacting normally on rigid surfaces.

**Acknowledgements** This work was partially supported by FEDER and Xunta de Galicia funds under the ED431C 2017/60 grant, by the Ministry of Economy, Industry and Competitiveness through the Plan Nacional de I+D+i (MTM2015-68275-R), the grant BES-2016-077228 and by the Vicerreitoría de Investigación e Innovación da Universidade de Santiago de Compostela via the Programa de Becas de Colaboración en Investigación 2016.

The authors would also like to acknowledge ArcelorMittal and personally thank Alejandro Lengomín and Sara Vázquez for their invaluable help to understand the phenomena in the complex industrial process that was modeled in this work.

## References

1. ANSYS, Inc.: ANSYS Fluent, Release 15.0, Theory Guide. ANSYS, Inc., Canonsburg (2013)
2. Ariathurai, R., Arulanandan, K.: Erosion rates of cohesive soils. J. Hydraul. Div. **104**, 279–283 (1978)
3. Beltaos, S., Rajaratnam, N.: Impinging circular turbulent jets. J. Hydraul. Div. ASCE, **100**, 1313–1328 (1974)
4. Brackbill, J., Kothe, D., Zemach, C.:A continuum method for modeling surface tension. J. Comput. Phys. **100**, 335–354 (1992)
5. Hirt, C., Nichols, B.: Volume of fluid (VOF) method for the dynamics of free boundaries. J. Comput. Phys. **39**, 201–225 (1981)
6. Howell, J., Menguc, M., Siegel, R.: Thermal Radiation Heat Transfer. Cambridge University Press, Cambridge (2010)
7. Khodabandeh, E., Ghaderi, M., Afzalabadi, A., Rouboa, A.: Parametric study of heat transfer in an electric arc furnace and cooling system. Appl. Therm. Eng. **123**, 1190–1200 (2017)
8. Kim, H., Ozturk, B.: Slag-metal separation in the blast furnace trough. ISIJ Int. **38**, 430–439 (1998)
9. Lee, W.E., Vieira, W., Zhang, S., Ahari, K.G., Sarpoolaky, H., Parr, C.: Castable refractory concretes. Int. Mater. Rev. **46**, 145–167 (2001)
10. Menter, F.R.: Two-equation eddy-viscosity turbulence models for engineering applications. AIAA J. **32**, 1598–1605 (1994)
11. Prompt, N., Ouedraogo, E.: High temperature mechanical characterisation of an alumina refractory concrete for Blast Furnace main trough: Part I. General context. J. Eur. Ceram. Soc. **28**, 2859–2865 (2008)
12. Rezende, P., Vicente, R., da Silva, A., Carvalho, F., Maliska, C.: The blast furnace trough two-phase flow and its influence in the refractory lining wear: mathematical modeling and numerical simulation. In: Proceedings of 19th International Congress of Mechanical Engineering, Brasilia, DF (2007)
13. Seoane, M.: Modelización de Fenómenos Térmicos que Afectan al Canal Principal del Horno Alto. Master's Thesis, Master in Industrial Mathematics, Universidade de Santiago de Compostela (2016)
14. Tryggvason, G., Scardovelli, R., Zaleski, S.: Direct Numerical Simulation of Gas–Liquid Multiphase Flows. Cambridge University Press, Cambridge (2011)
15. Vázquez-Fernández, S.: Modelización Matemática y Simulación Numérica de la Transferencia de Calor de una ruta de Horno Alto. Master's Thesis, Master in Industrial Mathematics, Universidade de Santiago de Compostela (2015)
16. Versteeg, K., Malalasekera, W.: An Introduction to Computational Fluid Dynamics: The Finite Volume Method. Pearson Education, London (2007)
17. Wilcox, D.: Turbulence Modeling for CFD. DCW Industries, La Canada (1998)