SEMA SIMAI Springer Series 17

Anna Doubova
Manuel González-Burgos
Francisco Guillén-González
Mercedes Marín Beltrán *Editors*

Recent Advances in PDEs: Analysis, Numerics and Control

In Honor of Prof. Fernández-Cara's 60th Birthday

SēMA SIMAI SOCIETÀ ITALIANA DI MATEMATICA APPLICATA E INDUSTRIALE Springer

SEMA SIMAI Springer Series

Volume 17

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

Anna Doubova • Manuel González-Burgos •
Francisco Guillén-González •
Mercedes Marín Beltrán
Editors

Recent Advances in PDEs: Analysis, Numerics and Control

In Honor of Prof. Fernández-Cara's 60th Birthday

Springer

Editors

Anna Doubova
Facultad de Matemáticas
Universidad de Sevilla
Sevilla, Spain

Manuel González-Burgos
Facultad de Matemáticas
Universidad de Sevilla
Sevilla, Spain

Francisco Guillén-González
Facultad de Matemáticas
Universidad de Sevilla
Sevilla, Spain

Mercedes Marín Beltrán
Departamento de Informática y
Análisis Numérico
Universidad de Córdoba
Córdoba, Spain

ISSN 2199-3041 ISSN 2199-305X (electronic)
SEMA SIMAI Springer Series
ISBN 978-3-319-97612-9 ISBN 978-3-319-97613-6 (eBook)
https://doi.org/10.1007/978-3-319-97613-6

Library of Congress Control Number: 2018957619

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

Introduction

This book contains the main results of the talks given at the workshop "Recent Advances in PDEs: Analysis, Numerics and Control", which took place in Sevilla (Spain) on January 25–27, 2017. The work comprises 13 contributions given by high-level researchers in the partial differential equation (PDE) area to celebrate the 60th anniversary of Enrique Fernández-Cara (University of Sevilla).

The aim of this book is to present a representative selection of the talks given at the workshop, aiming to disseminate the latest scientific results and to envisage new challenges in Control and inverse problems, Analysis of Fluid mechanics and Numerical Analysis.

The Editors warmly thank all the speakers and participants for their contributions to the Workshop, which ensured its success. In particular, we would like to acknowledge the efforts of all the speakers who have contributed to this volume. We are also grateful to the Scientific Committee, Tomás Chacón Rebollo (University of Seville), Tomás Caraballo Garrido (University of Seville), Oleg Imanuvilov (Colorado State University) and Nader Masmoudi (Courant Institute, New York) for their efforts during preparation of the Workshop. We extend our thanks and gratitude to all sponsors and supporting institutions for their valuable contributions: SEMA, SMAI, University of Seville, IMUS and the Spanish Ministry of Economy and Competitiveness, which awarded the grants MTM2015-69875-P, MTM2015-64577-C2-1-R, MTM2015-63723-P, MTM2014-53309-P and MTM2013-41286-P.

The Editors would also like to express their gratitude to Prof. Enrique Fernández-Cara for having agreed to receive this tribute in celebration of his 60th birthday and would like to present here the words of thanks that were expressed by Prof. Enrique Fernández-Cara to all participants of the Workshop, colleagues and collaborators.

Sevilla, Spain	Anna Doubova
Sevilla, Spain	Manuel González-Burgos
Sevilla, Spain	Francisco Guillén-González
Córdoba, Spain	Mercedes Marín Beltrán
May 2018	

Foreword

"Dear friends, dear colleagues,

With your permission, I would like to say some words.

First of all, I must express my deep gratitude to all of you for coming to Seville and participating in this meeting and, of course, for coming to this wonderful place to share these moments with me. It seems that the unique thing the organisers have not been able to arrange is the weather . . . .

Of course, very special thanks to the organisers of this workshop, my colleagues and former students Mercedes Marín, Francisco Guillén, Manolo G. Burgos and Anna Doubova. Also, many thanks to the members of the Scientific Committee and to all other colleagues and former and present students.

Definitively, I can see that the outcome of this meeting is much more than what I deserve. Indeed, the talks we are having these days confirm this to me. Very sincerely, I see that, this time, the tribute is of first class and I am not sure to be the same.

In my academic career, several crucial moments have determined where I am today. And I am very happy for this. Let me indicate them:

1. The first crucial moment was in 1978, when I decided to try to work in the differential equations and numerical analysis fields. I started a contact with our professor Antonio Valle, who unfortunately passed away in 2014.

   I will always be very grateful to Professor Valle. Thanks to him, I got a grant from the French Government to make a thesis in INRIA and Paris 6, under the direction of Roland Glowinski.
2. This was a second crucial moment. With me, Roland Glowinski solved a bi-objective problem as he is, as a master: he taught me a maximal amount of things using a minimal amount of time. Of course, this was for me the starting point of a long list of contacts with many people: first, Henri Berestycki, whom I began to work with; and then I met Americo Marrocco, Olivier Pironneau, Pierre-Louis Lions, Frédéric Hecht, María Jesús Esteban and also François Murat, Jean-Michel Coron, Jean-Pierre Puel, Lucio Boccardo, etc.

3. There was a third very relevant moment a few years later, in 1984. At that time, after defending several theses in our laboratory, we were scientifically a little bit disoriented and even lost. I remember that several of us made a 2-week visit to Paris, where we met François Murat. After a very pleasant and friendship conversation, François suggested to study in detail theoretical and numerical problems concerning the Navier-Stokes equations.

   And we did it.
4. A fourth important moment was my contact with Jacques Simon in 1989. Together, we gave a continuation to our analysis of the Navier-Stokes equations and we started to work on optimum design and related topics. I have always appreciated his ability to be at the same time deep and useful in analysis. I have had a lot of interesting conversations with him on Mathematics but also on other subjects (not always in agreement). Moreover, thanks to him, not only me but also many other colleagues had the chance to contact his former students Jerome Lemoine and Didier Bresch.
5. Then, towards 1994, I met again Enrique Zuazua and Jean-Pierre Puel. I knew them since the 1980s but it was only later that we began to work together. Thanks to them, I have learned a lot of things on control theory, in particular on the controllability of linear and nonlinear PDEs. With Enrique, we worked hard on the control of linear and semilinear heat equations. With Jean-Pierre (and then Oleg Imanuvilov and Sergio Guerrero), we found some results for the Navier-Stokes and related systems. In this area, my more recent contacts with Jean-Michel Coron and Arnaud Münch have also been decisive, making it possible to get new results.

   And I must also mention Assia Benabdallah and Chérif Ammar-Khodja and Otared Kavian ....
6. Finally, I cannot forget my contacts with a lot of people in Brazil, Mexico and Chile. This started in 2002, with Marko Antonio Rojas-Medar and Jose Luiz Boldrini. More or less at 2004, I met Luz De Teresa and then I met Fágner Araruna, Pablo Braz, Juan Límaco and others. All these collaborations have been fantastic to me from both the professional and personal viewpoints and I am also very grateful for this.

I have always believed that our activities must be guided by the following:

(a) First, we must teach. In fact, we are mainly paid for this. We must educate and train students and young people as much as we can. This will be our stimulus and maybe our legacy.

(b) Then, we must work together and collaborate. Today, it does not seem reasonable to isolate and work alone. Very probably, this is not the best way to be successful.

(c) Finally, we must progress. I understand this verb in the widest sense: progress in science and methods, progress in the choice of subjects and also progress in the way of life of our institutions. We have to be sensible to the evolution that our institutions need and we have to help them to achieve their goals.

I think that the situation of our laboratory is nowadays very satisfactory although, of course, many things remain still to be done. At a larger scale, I find that the situation is also very encouraging. Indeed, most of us are attached to an Institute of Mathematics with more than 100 members (about 40 of them are full professors) which is active in practice in all areas and is able to present a highly performant activity in the last years. In spite of a lot of difficulties, the work of several colleagues gave rise to these units, at present led by Manuel González Burgos and Tomás Chacón.

They seem excellent tools to grow in mathematics. We can dispose of structures sufficiently rich to receive students, young researchers and visitors and sufficiently powerful to support high-level programs with appropriate activities.

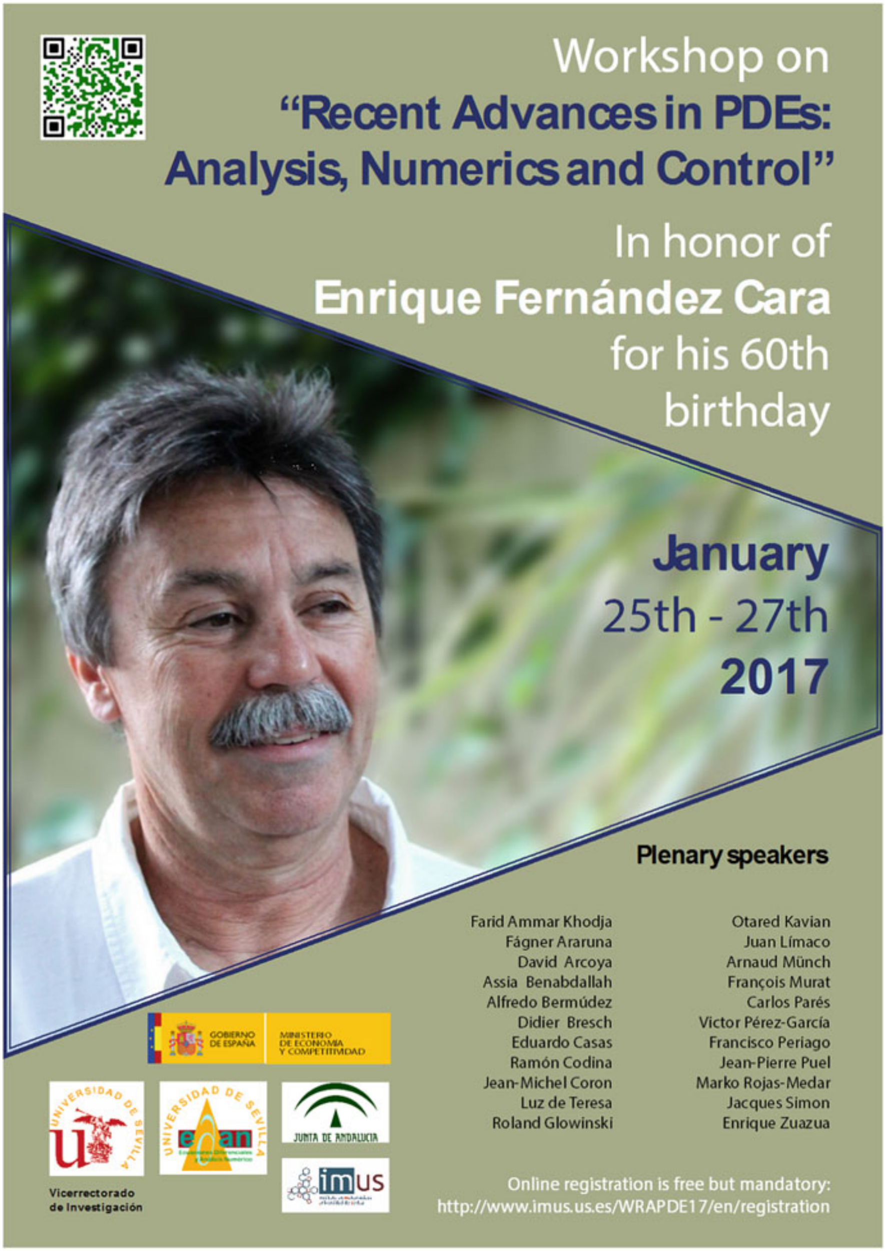
So, although very modestly, I feel proud to have been able to contribute to these tasks.

Thank you very much, Rosa.

And thank you very much to you all for your attention."

Sevilla, Spain
January 2017

Enrique Fernández-Cara



Workshop on
"Recent Advances in PDEs: Analysis, Numerics and Control"

In honor of
Enrique Fernández Cara
for his 60th birthday

January
25th - 27th
2017

Plenary speakers

Farid Ammar Khodja
Fágner Araruna
David Arcoya
Assia Benabdallah
Alfredo Bermúdez
Didier Bresch
Eduardo Casas
Ramón Codina
Jean-Michel Coron
Luz de Teresa
Roland Glowinski
Otared Kavian
Juan Límaco
Arnaud Münch
François Murat
Carlos Parés
Víctor Pérez-García
Francisco Periago
Jean-Pierre Puel
Marko Rojas-Medar
Jacques Simon
Enrique Zuazua

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMIA Y COMPETITIVIDAD

Vicerrectorado de Investigación

JUNTA DE ANDALUCIA

imus

Online registration is free but mandatory:
http://www.imus.us.es/WRAPDE17/en/registration

Contents

Short Biography

Anna Doubova is an Associate Professor in the Department of Differential Equations and Numerical Analysis (EDAN), Universidad de Sevilla, Spain. She holds a degree in Mechanics and Applied Mathematics from Lomonosov University (Moscow) and a PhD in Mathematics from the Universidad de Sevilla. Her research relates to analysis, control and inverse problems of PDEs with applications in physics, engineering, biology and other sciences. She publishes regularly in high-impact international journals.

Manuel González-Burgos, BS, PhD is Full Professor of Mathematical Analysis in the Department of Differential Equations and Numerical Analysis, Universidad de Sevilla, Spain. His fields of specialization include partial differential equations and control theory, with a particular focus on the controllability properties of scalar and non-scalar parabolic problems with controls exerted in a part of the domain or on a part of the boundary. He has published over 35 papers in peer-reviewed journals.

Francisco Guillén-González, BS, PhD is Full Professor of Mathematical Analysis in the Department of Differential Equations and Numerical Analysis, Universidad de Sevilla, Spain. His research interests include the mathematical and numerical analysis of PDE systems applied to other sciences. In particular, he has studied PDEs related to incompressible fluid mechanics, phase transitions and biological processes. He is the author of more than 70 papers in peer-reviewed journals.

Mercedes Marín Beltrán is an Associate Professor of Mathematical Analysis in the Department of Computing and Numerical Analysis at the Universidad de Córdoba, Spain. She received her BS and her PhD from the Universidad de Sevilla. She is the author of a number of publications in international journals on the development and analysis of advanced numerical methods for partial differential equations with applications to fluid dynamics and tumor growth.

Essential Spectrum and Null Controllability of Some Parabolic Equations

Farid Ammar Khodja and Cédric Dupaix

Dedicated to Prof. Enrique Fernández-Cara on the occasion of his 60th birthday.

Abstract We give some examples proving that if the underlying elliptic operator of a parabolic equation admits essential spectrum, then boundary or internal null controllability are not possible in general.

Keywords Controllability · Degenerate parabolic equations · Elliptic systems · Essential spectrum · Singular sequences

1 Introduction and Main Results

The aim of this paper is to show through two examples the effect of essential spectrum on controllability of parabolic systems.

Throughout this paper, $\Omega \subset \mathbb{R}^N$ will denote a bounded C^∞-domain and $\Gamma = \partial\Omega$ will be its boundary.

The first example we consider is the following control parabolic system:

$$\begin{cases} y' - \nabla \cdot (\sigma^\mu \nabla y) = u\chi_\omega, & Q_T := (0,T) \times \Omega, \\ \sigma^\mu \dfrac{\partial y}{\partial \nu} = 0, & \Sigma_T := (0,T) \times \Gamma, \\ y(0) = y_0, & \Omega, \end{cases} \tag{1}$$

where $\omega \subset \Omega$ is an open subset, $u \in L^2(Q_T)$ is the control function and $y_0 \in L^2(\Omega)$, $\mu \geq 0$ and ν is the exterior normal to Γ. We assume that the function

F. Ammar Khodja (✉) · C. Dupaix
Laboratoire de Mathématiques de Besançon, Besançon Cedex, France
e-mail: fammarkh@univ-fcomte.fr

A. Doubova et al. (eds.), *Recent Advances in PDEs: Analysis, Numerics and Control*, SEMA SIMAI Springer Series 17,
https://doi.org/10.1007/978-3-319-97613-6_1

$\sigma : \Omega \to (0, \infty)$ satisfies the following assumptions:

$$\sigma \in C^{\infty}(\Omega), \tag{2}$$

and there exist $\varepsilon_0, c > 0$ such that

$$\begin{aligned} &\tfrac{1}{c} d_\Gamma(x) \le \sigma(x) \le c d_\Gamma(x), \\ &\tfrac{1}{c} \le |\nabla\sigma(x)| \le c, \qquad \forall x \in V_{\varepsilon_0}(\Gamma). \\ &\lim_{d_\Gamma(x)\to 0} (\ln\sigma(x))(\sigma\Delta\sigma)(x) = 0, \end{aligned} \tag{3}$$

where $V_\varepsilon(\Gamma)$ denotes the ε-neighborhood of Γ :

$$V_\varepsilon(\Gamma) = \{x \in \Omega : d_\Gamma(x) < \varepsilon\},$$

and $d_\Gamma : \Omega \to \mathbb{R}_+$ is the distance function to Γ, defined by:

$$d_\Gamma(x) = \inf_{y\in\Gamma} |y - x|, \ x \in \Omega.$$

The null-controllability property (see the next sections for a definition of this property) of system (1) has been intensively studied these last years. The case $\mu < 2$ is by now well understood. Actually, it has been proved that solutions of the associated adjoint problem

$$\begin{cases} \varphi' + \nabla\cdot(\sigma^\mu\nabla\varphi) = 0, & Q_T := (0,T)\times\Omega, \\ \sigma^\mu \dfrac{\partial\varphi}{\partial\nu} = 0, & \Sigma_T := (0,T)\times\Gamma, \\ \varphi(T) = \varphi_0, & \Omega, \end{cases} \tag{4}$$

satisfy global Carleman estimates that imply the observability inequality

$$\forall T > 0, \exists C_T > 0 : \int_\Omega |\varphi(0,x)|^2 dx \le C_T \int_0^T \int_\omega |\varphi(t,x)|^2 dxdt, \ \forall\varphi_0 \in L^2(\Omega), \tag{5}$$

and it is classical that this observability inequality is equivalent to the null-controllability of system (1). For all these questions, we refer to Cannarsa et al. [6] and the references therein.

When $\mu \ge 2$, in [6, Proposition 16.5, p. 145], an example is provided which proves that, *in general*, the problem is not null-controllable. In this example, it should be noted that it is assumed that $\overline{\omega} \subset \Omega$. A natural question arises from this example: does this negative null-controllability result always hold when $\mu \ge 2$? We give an answer for system (1):

Theorem 1 *Assume that* $\mu > 2$ *and* σ *satisfies (2) and (3). If the open subset* $\omega \subset \mathbb{R}^n$ *is such that* $\overline{\omega} \subset \Omega$, *then (1) is not null controllable at any* $T > 0$.

The second control system we consider is the following:

$$\begin{cases} y' - \Delta y + (b \cdot \nabla + c)\, z = u\chi_{\omega}, & Q_T := (0,T) \times \Omega, \\ z' - az + (-\nabla \cdot b + c)\, y = 0 & Q_T \\ y = 0, & \Sigma_T := (0,T) \times \Gamma, \\ (y(0), z(0)) = (y_0, z_0), & \Omega, \end{cases} \tag{6}$$

with $a, b, c \in C^{\infty}(\Omega) \cap C^0\left(\overline{\Omega}\right)$ and $\omega \subset \Omega$ is an open subset. This system can be seen as a coupling between a parabolic equation and a first order differential equation. We have the following result:

Theorem 2 *Assume that $a(x) > 0$ for all $x \in \overline{\Omega}$. If $\Omega \backslash \overline{\omega} \neq \varnothing$, then (6) is not null-controllable.*

As we will see later, there is a common point between the two considered systems: in the two situations, the underlying elliptic operator is not uniformly elliptic and admits essential spectrum. This essential spectrum gives rise to singular sequences of functions with supports disjoint from the control domain ω.

The plan of this paper is the following. In Sect. 2, we first recall the definition of the essential spectrum of an operator and some of its elementary properties, and then give an abstract result where a sufficient condition of non null-controllability is provided when the underlying operator has essential spectrum (see Proposition 5). Sections 3 and 4 are devoted to the proof of Theorems 1 and 2 respectively. Each of these two last sections contains comments and open problems related to the studied systems.

2 Essential Spectrum, Singular Sequences and Controllability

Let H be a Hilbert space and $A : D(A) \subset H \to H$ be a closed unbounded operator. We denote by $\sigma(A)$ the spectrum of A. The discrete spectrum $\sigma_d(A)$ of A is the set of isolated eigenvalues of A with finite multiplicity. The essential spectrum $\sigma_{ess}(A)$ of A is

$$\sigma_{ess}(A) = \sigma(A) \backslash \sigma_d(A).$$

If A is a selfadjoint operator (so that, it is densely defined), we have Weyl's characterization of $\sigma_{ess}(A)$:

Proposition 3 *Assume that A is a selfadjoint operator on the Hilbert space H. Then the following assertions are equivalent:*

1. $\lambda \in \sigma_{ess}(A)$.

2. There exists a sequence $(\varphi_n)_{n\geq 1} \subset D(A)$ satisfying the following properties:

$$\|\varphi_n\|_H = 1, \ \forall n \geq 1, \tag{7}$$

$$\varphi_n \underset{n\to\infty}{\overset{w}{\rightharpoonup}} 0 \text{ in } H, \tag{8}$$

$$\lim_{n\to\infty} \|A\varphi_n - \lambda\varphi_n\|_H = 0 \tag{9}$$

where $\overset{w}{\rightharpoonup}$ denotes the weak limit.

The first item in this proposition ensures that the sequence (φ_n) has no (strongly) convergent subsequence. Those sequences satisfying the properties (7)–(9) are called *singular sequences* associated with $\lambda \in \sigma_{ess}(A)$.

If A is a densely defined closed operator (not necessarily selfadjoint), introduce the Weyl spectrum $W(A)$ of A:

$$W(A) = \left\{\lambda \in \mathbb{C} : \exists\, (\varphi_n)_{n\geq 1} \subset D(A) \text{ with } (7) - (8) - (9)\right\}$$

It is worthnoting that $W(A) \subset \sigma_{ess}(A)$ and the equality does not hold in general (see [12, Chapter 10] for more details).

Proposition 4 *Assume that $-A$ is the generator of a C^0-semigroup $\left(e^{-tA}\right)$ on H. If $\lambda \in W(A) \subset \sigma_{ess}(A)$ and (φ_n) is a singular sequence associated with λ, then*

$$\lim_{n\to\infty} \left\| e^{-\lambda t}\varphi_n - e^{-tA}\varphi_n \right\| = 0, \ t \geq 0.$$

Proof This readily follows from the formula:

$$\forall \varphi \in D(A), \ e^{-\lambda t}\varphi - e^{-tA}\varphi = \int_0^t e^{-\lambda(t-s)} e^{-sA}(A-\lambda)\varphi ds, \ t \geq 0. \tag{10}$$

■

We now turn to the connection between essential spectrum and observability (or controllability). Let $T > 0$, H and U be Hilbert spaces. Consider the control system:

$$\begin{cases} y' = -Ay + Bu, & (0,T) \\ y(0) = y_0 \in H. \end{cases} \tag{11}$$

Here, $-A : D(A) \subset H \to H$ is the generator of a C^0-semigroup denoted by e^{-tA} and $B : U \to H$ an admissible operator, i.e. an operator satisfying:

$$\exists C > 0, \ \int_0^T \left\| B^* e^{-tA^*}\varphi \right\|_U^2 dt \leq C \|\varphi\|^2, \ \forall \varphi \in D(A).$$

We write:

$$y\,(t;\,y_0,u)=e^{-tA}y_0+\int_0^t e^{-(t-s)A}Bu\,(s)\,ds,$$

the solution of (11) associated with $(y_0,u)\in H\times L^2\,(0,T;U)$.

The null-controllability issue is formulated as follows: given $T>0$ and $y_0\in H$, find $u\in U$ such that $y\,(T;\,y_0,u)=0$ in H, while the approximate controllability issue corresponds to the property: given $T,\varepsilon>0$ and $(y_0,y_1)\in H\times H$, find $u\in U$ such that$\|y\,(T;\,y_0,u)-y_1\|_H<\varepsilon$.

As is well-known, a dual formulation of these issues uses the adjoint problem

$$\begin{cases}\varphi'=-A^*\varphi, & (0,T)\\ \varphi\,(0)=\varphi_0\in H.\end{cases}\tag{12}$$

System (11) is null controllable if, and only if, the following observability inequality for the solutions of (12) holds true:

$$\exists C_T>0,\ \left\|e^{-TA^*}\varphi_0\right\|_H^2\le C_T\int_0^T\left\|B^*e^{-tA^*}\varphi_0\right\|_U^2\,dt,\forall\varphi_0\in H.\tag{13}$$

System (11) is approximately controllable if, and only if:

$$\begin{cases}\varphi'=-A^*\varphi, & (0,T)\,,\\ \varphi\,(0)=\varphi_0\in H, & \\ B^*\varphi=0, & (0,T)\,,\end{cases}\Rightarrow\varphi_0=0.$$

Proposition 5 *If for some* $\lambda\in W\,(A^*)$, $\lambda\ge 0$, *there exists an associated singular sequence* $\{\varphi_n\}\subset D\left(A^{*2}\right)$ *such that*

$$\lim_{n\to\infty}\|B^*\varphi_n\|_U^2=0,$$

then the observability inequality (13) fails to be true.

Proof If $\{\varphi_n\}\subset D\left(A^{*2}\right)$ is a singular sequence associated with $\lambda\in W\,(A^*)$, we can write:

$$\begin{aligned}\int_0^T\left\|B^*e^{-tA^*}\varphi_n\right\|_U^2\,dt&\le C_T\left(\int_0^T\left\|B^*\left(e^{-tA^*}-e^{-\lambda t}\right)\varphi_n\right\|_U^2\,dt+\int_0^T e^{-2\lambda t}\,dt\,\|B^*\varphi_n\|_U^2\right)\\&\le C_T\left(\int_0^T\left\|\int_0^t e^{-\lambda(t-s)}B^*e^{-sA^*}\left(A^*-\lambda\right)\varphi_n ds\right\|_U^2\,dt+\delta\,(\lambda,T)\,\|B^*\varphi_n\|_U^2\right)\\&\le C_T\delta\,(\lambda,T)\left(T\int_0^T\left\|B^*e^{-tA^*}\left(A^*-\lambda\right)\varphi_n\right\|_U^2\,dt+\|B^*\varphi_n\|_U^2\right)\\&\le C_T\delta\,(\lambda,T)\,(1+T)\left(\left\|\left(A^*-\lambda\right)\varphi_n\right\|^2\ +\|B^*\varphi_n\|_U^2\right),\end{aligned}\tag{14}$$

the last inequality following from the admissibility of B and

$$\delta(\lambda, T) = \begin{cases} \frac{1-e^{2\lambda T}}{2\lambda}, \text{ if } \lambda > 0, \\ \\ T, \quad \text{if } \lambda = 0. \end{cases}$$

It appears that

$$\lim_{n\to\infty} \left\| B^* \varphi_n \right\|_U^2 = 0 \Longrightarrow \lim_{n\to\infty} \int_0^T \left\| B^* e^{-tA^*} \varphi_n \right\|_U^2 \, dt = 0.$$

On the other hand, from Proposition 4,

$$\lim_{n\to\infty} \left\| e^{-\lambda T} \varphi_n - e^{-TA} \varphi_n \right\| = 0.$$

Thus

$$\lim \inf_{n\to\infty} \left\| e^{-TA} \varphi_n \right\| \geq \lim \inf_{n\to\infty} \left| e^{-\lambda T} \left\| \varphi_n \right\| - \left\| e^{-\lambda T} \varphi_n - e^{-TA} \varphi_n \right\| \right| = e^{\lambda T} > 0 \tag{15}$$

The proposition is then a consequence of (14), (15). ■

The proof of this last proposition is inspired from an inequality communicated to us by Morgan Morancey (Personal communication, 2016). Note that this result does not give any connection between approximate controllability and essential spectrum.

3 The Controllability Issue for Degenerate Parabolic Equations

3.1 Proof of Theorem 1

Let $A_\mu : L^2(\Omega) \to L^2(\Omega)$ be the operator defined by:

$$A_\mu y = -\nabla \cdot \left(\sigma^\mu \nabla y\right),$$
$$D\left(A_\mu\right) = H_\sigma^2(\Omega)$$

where

$$H_\sigma^1(\Omega) = \left\{ y \in L^2(\Omega) \cap H_{\text{loc}}^1(\Omega) : \int_\Omega \sigma^\mu \left|\nabla y\right|^2 < \infty \right\},$$

$$H_\sigma^2(\Omega) = \left\{ y \in H_\sigma^1(\Omega) \cap H_{\text{loc}}^2(\Omega) : \int_\Omega \left|\nabla \cdot \left(\sigma^\mu \nabla y\right)\right|^2 < \infty \right\}.$$

From Cannarsa et al. [6] (see also [16]), it appears that

$$y \in H^2_\sigma(\Omega) \Rightarrow \sigma^\mu \frac{\partial y}{\partial \nu} = 0,$$

and that $A_\mu \geq 0$ is a selfadjoint operator on $L^2(\Omega)$.

The proof of this theorem is based on a result due to Pang [15] which proves that if $\mu > 2$ then $0 \in \sigma_{ess}(A_\mu)$ (the essential spectrum of A_μ). Let us recall the main point of Pang's result. Let $f \in C^\infty(\mathbb{R}, [0, 1])$ be a function such that:

$$\exists k \in (0, 1), \; f(s) = \begin{cases} 1, & |s| \leq k \\ \\ 0, & |s| \geq 1 \end{cases}$$

For an integer $n \geq 1$ and a real number $m > 1$, define the function:

$$\phi_n(x) = f\left(\frac{\frac{1}{\sigma(x)} - n^{2m}}{n^2}\right), \; x \in \Omega. \tag{16}$$

Note that, in view of the definition of f, we have

$$\phi_n \in C_0^\infty(\Omega, \mathbb{R}), \; \text{supp}(\phi_n) = \left\{x \in \Omega : \frac{1}{n^{2m} + n^2} \leq \sigma(x) \;\; \leq \frac{1}{n^{2m} - n^2}\right\}. \tag{17}$$

For the sequence $\{\phi_n\}$, the author proves that

$$\frac{\phi_n}{\|\phi_n\|} \underset{n\to\infty}{\to} 0, \; \text{weakly } L^2(\Omega),$$

and

$$\lim_{n\to\infty} \frac{\int_\Omega \sigma^\mu |\nabla \phi_n|^2 \, dx}{\|\phi_n\|^2} = 0.$$

Thus $0 \in \sigma(A_\mu)$ (by minimax).

Proposition 6 *If* $\mu > 2$ *then* $\left\{\frac{\phi_n}{\|\phi_n\|_{L^2(\Omega)}}\right\}_{n>1}$ *is a singular sequence for* A_μ *associated with* $0 \in \sigma_{ess}(A_\mu)$.

Proof In [15, Proof of Theorem 5.1], it has been proved that there exist $C > 0$ such that

$$\|\phi_n\|^2_{L^2(\Omega)} \geq C n^{2-4m}, \; \forall n > 1. \tag{18}$$

Now, setting

$$g\left(n,\sigma\right)=\frac{\frac{1}{\sigma(x)}-n^{2m}}{n^{2}}$$

we have

$$\begin{aligned} A_{\mu}\phi_{n} &= -\nabla\cdot\left(\sigma^{\mu}\nabla\phi_{n}\right) \\ &= -\left(\frac{1}{n^{4}}f''\left(g\left(n,\sigma\right)\right)\sigma^{\mu-2}\left|\nabla\sigma\right|^{2}+\frac{\sigma^{\mu-3}}{n^{2}}f'\left(g\left(n,\sigma\right)\right)\Big(\left(2-\mu\right)\left|\nabla\sigma\right|^{2}-\sigma\Delta\sigma\Big)\right). \end{aligned}$$

From the definition of f and the assumptions on σ, we get:

$$\begin{aligned} \int_{\Omega}\left|A_{\mu}\phi_{n}\right|^{2}dx &\leq C\int_{\left\{\left(n^{2m}+n^{2}\right)^{-1}<\sigma(x)<\left(n^{2m}-n^{2}\right)^{-1}\right\}}\left(\frac{\sigma^{2(\mu-2)}}{n^{8}}+\frac{\sigma^{2(\mu-3)}}{n^{4}}\right)dx \\ &\leq C\int_{\left(n^{2m}+n^{2}\right)^{-1}}^{\left(n^{2m}-n^{2}\right)^{-1}}\left(\frac{r^{2(\mu-2)}}{n^{8}}+\frac{r^{2(\mu-3)}}{n^{4}}\right)dr \\ &\leq C\int_{\left(n^{2m}+n^{2}\right)^{-1}}^{\left(n^{2m}-n^{2}\right)^{-1}}\left(\frac{r^{2(\mu-2)}}{n^{8}}+\frac{r^{2(\mu-3)}}{n^{4}}\right)dr \\ &\leq C\left(\frac{\left(n^{2m}-n^{2}\right)^{-(2\mu-3)}-\left(n^{2m}+n^{2}\right)^{-(2\mu-3)}}{n^{8}\left(2\mu-3\right)}\right. \\ &\qquad\left.+\frac{\left(n^{2m}-n^{2}\right)^{-(2\mu-5)}-\left(n^{2m}+n^{2}\right)^{-(2\mu-5)}}{n^{4}\left(2\mu-5\right)}\right) \end{aligned}$$

But

$$\begin{aligned} \frac{\left(1-n^{2-2m}\right)^{-(2\mu-3)}-\left(1+n^{2-2m}\right)^{-(2\mu-3)}}{\left(2\mu-3\right)} &= 2n^{2-2m}+O\left(n^{3(2-2m)}\right), \\ \frac{\left(1-n^{2-2m}\right)^{-(2\mu-5)}-\left(1+n^{2-2m}\right)^{-(2\mu-5)}}{\left(2\mu-5\right)} &= 2n^{2-2m}+O\left(n^{3(2-2m)}\right) \end{aligned}$$

Thus:

$$\frac{\int_{\Omega}\left|A_{\mu}\phi_{n}\right|^{2}dx}{\left\|\phi_{n}\right\|_{L^{2}(\Omega)}^{2}}\leq C\left(\frac{1}{n^{4m\mu-8m+8}}+\frac{1}{n^{4m\mu-12m+4}}\right)\underset{n\to\infty}{\sim}\frac{C}{n^{4m\mu-12m+4}}$$

It follows that

$$\mu > 3 - \frac{1}{m} \Rightarrow \lim_{n\to\infty} \frac{\int_\Omega \left|A_\mu \phi_n\right|^2 dx}{\int_\Omega |\phi_n|^2 dx} = 0.$$

The conclusion follows from the fact that for any $\mu > 2$, we can find $m > 1$ such that $\mu > 3 - \frac{1}{m}$. ■

Proof (Proof of Theorem 1) Set

$$\psi_n = \frac{\phi_n}{\|\phi_n\|_{L^2(\Omega)}},\ n \geq 1,$$

where (ϕ_n) is the sequence defined in (16). From Proposition 6, (ψ_n) is a singular sequence associated with $0 \in \sigma_{ess}(A)$. According to Proposition 5, applied with $H = U = L^2(\Omega)$ and $B^* = 1_\omega \in \mathcal{L}\left(L^2(\Omega)\right)$, the observability inequality is not satisfied if

$$\lim_{n\to\infty} \int_\omega |\psi_n|^2 dx = 0.$$

But from the definition of ψ_n, this is immediate since it appears that if $\overline{\omega} \subset \Omega$, there exists $n_0 \geq 1$ such that

$$\operatorname{supp}(\psi_n) \cap \omega = \varnothing, \forall n \geq n_0.$$

Thus (5) is not satisfied for any $T > 0$. ■

3.2 Comments

1. **A variant of Theorem 1**. Let $\Gamma_0, \Gamma_1 \subset \Gamma$ be subsets of Γ such that $\Gamma = \Gamma_0 \cup \Gamma_1$ and $\Gamma_0 \cap \Gamma_1 = \varnothing$, and $\sigma : \Omega \cup \Gamma_1 \to (0, \infty)$ with $\sigma \in C^\infty(\Omega) \cap C^0\left(\overline{\Omega}\right)$ satisfying the following assumptions: there exist $\varepsilon_0, c > 0$ such that

$$\left.\begin{array}{l} \frac{1}{c} d_{\Gamma_0}(x) \leq \sigma(x) \leq c d_{\Gamma_0}(x), \\ \frac{1}{c} \leq |\nabla\sigma(x)| \leq c, \qquad \forall x \in V_{\varepsilon_0}(\Gamma_0). \\ \lim\limits_{d_{\Gamma_0}(x)\to 0} (\ln\sigma(x))(\sigma\Delta\sigma)(x) = 0, \end{array}\right. \tag{19}$$

Now, consider the system

$$\begin{cases} y' - \nabla \cdot (\sigma^{\mu} \nabla y) = u\chi_{\omega}, & Q_T, \\ \sigma^{\mu} \dfrac{\partial y}{\partial \nu}_{|\Gamma_0} = 0, \ y_{|\Gamma_1} = v & \Sigma_T, \\ y(0) = y_0, & \Omega, \end{cases} \tag{20}$$

The previous singular sequence of Proposition 6 is again a singular sequence for the operator underlying the adjoint system

$$\begin{cases} \varphi' + \nabla \cdot (\sigma^{\mu} \nabla \varphi) = 0, & Q_T, \\ \sigma^{\mu} \dfrac{\partial \varphi}{\partial \nu}_{|\Gamma_0} = 0, \ \varphi_{|\Gamma_1} = 0 & \Sigma_T, \\ \varphi(T) = \varphi_0, & \Omega. \end{cases}$$

With the same arguments, it follows that system (20) is not null-controllable if $\mu > 2$ and $\overline{\omega} \subset \Omega$.

2. **Approximate controllability**. In a forthcoming paper [1], the approximate controllability issue is considered for the system

$$\begin{cases} y' - \left(x^{\mu} y'\right)' = u\chi_{\omega}, & (0,T) \times (0,1), \\ \left(x^{\mu} y'\right)_{|x=0} = y(1) = 0, & (0,T), \\ y(0) = y_0, & (0,1), \end{cases}$$

where $\omega \subset (0,1)$ is an open subset. We prove that this system is approximately controllable for any $\mu \geq 0$ and any open subset ω. But, to our knowledge, this issue remains an open problem in higher dimension.

4 The Controllability Issue for Mixed Parabolic Systems

4.1 *Proof of Theorem 2*

The operator underlying (6) is given by:

$$\begin{array}{c} L_0 = \begin{pmatrix} -\Delta & P \\ P^* & a \end{pmatrix} : H := L^2(\Omega) \times L^2(\Omega) \to H \\ D(L_0) = H^2(\Omega) \cap H_0^1(\Omega) \times D(P), \\ P = b \cdot \nabla + c \end{array}, \tag{21}$$

where all the functions $a, b = (b_1, \ldots, b_n)$ and c are in $C^{\infty}(\Omega) \cap C\left(\overline{\Omega}\right)$. The operator L_0 is a very particular case among the operators studied in Grubb-Geymonat

[10]. It enters also in the class of matrix operators of the form $\begin{pmatrix} A & B \\ C & D \end{pmatrix}$ on $X = X_1 \times X_2$ where X_1 and X_2 are Banach spaces studied by Atkinson et al. [4]. Following these last paper, we can see that the closure of L_0 is given by

$$L = \begin{pmatrix} -\Delta & 0 \\ P^* & S_0 \end{pmatrix} \begin{pmatrix} I & (-\Delta)^{-1} P \\ 0 & I \end{pmatrix}$$
$$D(L) = \left\{ (y,z) \in H : y + (-\Delta)^{-1} Pz \in H^2(\Omega) \cap H_0^1(\Omega) \right\}.$$

Clearly, with our assumptions, the operator

$$S_0 = a - P^* \circ (-\Delta)^{-1} \circ P$$

defined on $D(P)$ can be extended to a bounded operator on $L^2(\Omega)$ and is selfadjoint. We can be more precise if we write:

$$\begin{aligned} S_0 &= a - (-b \cdot \nabla + (c - \nabla \cdot b)) \circ (-\Delta)^{-1} \circ (b \cdot \nabla + c) \\ &= a + b \cdot \nabla (-\Delta)^{-1} b \cdot \nabla + K \\ : &= \mathfrak{A} + K \end{aligned} \tag{22}$$

with

$$\mathfrak{A} = a + b \cdot \nabla (-\Delta)^{-1} b \cdot \nabla$$
$$K = b \cdot \nabla \circ (-\Delta)^{-1} \circ c - (c - \nabla \cdot b) \circ (-\Delta)^{-1} \circ b \cdot \nabla - (c - \nabla \cdot b)(-\Delta)^{-1} c \tag{23}$$

It is easy to see that since $(-\Delta)^{-1}$ is a compact operator on $L^2(\Omega)$, so is K. We first have:

Proposition 7 *Under the previous assumptions:*

$$\sigma_{ess}(L) = \sigma_{ess}(\mathfrak{A}) \neq \varnothing.$$

Proof This is a straightforward consequence of [4, Theorem 2.2, p. 9 and Corollary 2.3, p. 10.]. By this last result, we first have

$$\sigma_{ess}(L) = \sigma_{ess}(S_0).$$

From (22), since K is compact,

$$\sigma_{ess}(S_0) = \sigma_{ess}(\mathfrak{A})$$

using the invariance of the essential spectrum by compact perturbation (see for instance [12, 17]). ■

Going further in the computation of $\sigma_{ess}(L)$, we have:

Proposition 8 *Let $f \in C_0^\infty\left(\mathbb{R}^N\right)$ with $\int_{\mathbb{R}^N} f^2(x)\,dx = 1$. Then:*

1. $\sigma_{ess}(\mathfrak{A}) = \left\{\lambda(x,\xi) = a(x) + (b(x)\cdot\xi)^2,\ \ (x,\xi) \in \Omega \times S^1\right\}$.
2. If $\lambda = \lambda(x^*,\xi^*) \in \sigma_{ess}(\mathfrak{A})$, the sequence defined by

$$\varphi_n(x) = n^{N/2} f\left(n\left(x - x^*\right)\right) e^{in^2(x-x^*)\cdot\xi^*} \tag{24}$$

is a singular sequence of $\mathfrak{A}$ associated with λ.

Proof This result follows from [10, Proposition 6.4, p. 263] (see also [14, Theorem XVI 2, p. 243.]). ■

The last step is to construct a singular sequence for L using this singular sequence of $\mathfrak{A}$ associated with λ. We will need the following intermediate (classical) result:

Lemma 9 *There exists a properly supported operator $\Pi \in \mathcal{L}\left(L^2(\Omega)\right)$ and regularizing operators R, R' defined on $L^2(\Omega)$ such that*

$$\Delta \circ \Pi = I - R;\ \ \Pi \circ \Delta = I - R' \tag{25}$$

An operator Π satisfying (25) is a parametrix of Δ. A properly supported parametrix Π is an operator satisfying that for any compact set $K \subset \Omega$, there exists a compact set $K_1 \subset \Omega$ such that $\Pi\left(C_0^\infty(K)\right) \subset C_0^\infty(K_1)$. A proof of Lemma 9 can be found in [8, I.1.3, p. 14].

Corollary 10 *Let $(x^*,\xi^*) \in \Omega \times S^1$. The sequence (ψ_n) defined by*

$$\psi_n = c_n \begin{pmatrix} \Pi P \\ I \end{pmatrix} \varphi_n,\ \ c_n = \left(\|\Pi P\varphi_n\|^2_{L^2(\Omega)} + \|\varphi_n\|^2_{L^2(\Omega)}\right)^{1/2},$$

where Π is a parametrix of Δ given by (25) and (φ_n) is defined in (24), is a singular sequence for L associated with $\lambda = \lambda(x^*,\xi^*)$.

Proof Let $(x^*,\xi^*) \in \Omega \times S^1$ and $\lambda = \lambda(x^*,\xi^*)$. Then

$$\begin{aligned}(L-\lambda)\psi_n &= c_n \begin{pmatrix} (-\Delta\Pi + I)\,P\varphi_n - \lambda\Pi P\varphi_n \\ P^*(-\Delta)^{-1}(-\Delta\Pi + I)\,P\varphi_n + (S_0 - \lambda)\,\varphi_n \end{pmatrix} \\ &= c_n \begin{pmatrix} RP\varphi_n - \lambda\Pi P\varphi_n \\ P^*(-\Delta)^{-1} RP\varphi_n + (S_0 - \lambda)\,\varphi_n \end{pmatrix}.\end{aligned}$$

Now, by construction

$$\lim_{n\to\infty} \|(S_0 - \lambda)\,\varphi_n\|_{L^2(\Omega)} = 0.$$

Since R is a regularizing operator, we also have

$$\lim_{n\to\infty} \|RP\varphi_n - \lambda\Pi P\varphi_n\|_{L^2(\Omega)} = \lim_{n\to\infty} \left\| P^* \left(-\Delta\right)^{-1} RP\varphi_n \right\|_{L^2(\Omega)} = 0.$$

■

Remark 11 The sequence (ψ_n) can be chosen so that there exists a constant $c > 0$ such that

$$\operatorname{supp}(\psi_n) \subset B\left(x^*, \frac{c}{n}\right) := \left\{x \in \Omega : \left|x - x^*\right| < \frac{1}{n}\right\},\ n \geq 1.$$

This amounts to choose the parametrix Π in such a way that the support of a function is transported into a very close support.

Proof of Theorem 2 Let $\omega \subsetneqq \Omega$ and fix $(x^*, \xi^*) \in (\Omega \backslash \overline{\omega}) \times S^1$. Clearly, there exists $n_0 \geq 1$ such that

$$\operatorname{supp}(\psi_n) \cap \omega = \varnothing,\ \forall n \geq n_0.$$

Thus

$$\int_\Omega |1_\omega \psi_n|^2\, dx = 0,\ \forall n \geq n_0,$$

and again the conclusion follows from Proposition 5. ■

4.2 Comments

1. **Boundary null-controllability**. If the control acts on the boundary, the conclusion should be the same due to the concentration of the supports of the constructed singular sequence around points in Ω.
2. **The approximate controllability issue**. For system (6), the approximate controllability problem is still open. We can however mention the paper of Doubova and Fernandez-Cara [7] where the following system was considered:

$$\begin{cases} y' - \nu\Delta y + \nabla\cdot z = \nabla\pi + u\chi_\omega, & Q_T := (0,T)\times\Omega, \\ z' + az + b\left(\nabla y +^t \nabla y\right) = 0, & Q_T, \\ \nabla\cdot y = 0, & Q_T, \\ y = 0, & \Sigma_T := (0,T)\times\Gamma, \\ (y(0), z(0)) = (y_0, z_0), & \Omega, \end{cases}$$

where the coefficients of the system are real constants (ν, a, b), $y = (y_i)$ and $z = \left(z_{ij}\right)$ for some scalar functions y_i and z_{ij} with $z_{ij} = z_{ji}$ $(1 \leq i, j \leq N)$. The authors prove that the essential spectrum of the system is reduced to a unique point and it is the limit of a subsequence of eigenvalues. They were, however, able to prove that this system is approximately controllable. This last system can also be seen as a vectorial version of system (6). (See also Guerrero-Imanuvilov [11] where a proof of non null-controllability of a very close system is given using Fourier series.)

3. **The second order system**. In Geymonat-Valente [9], the second order (in time) system corresponding to (6) is proved to be non exactly controllable. In a particular setting allowing an explicit computation of the spectrum, the same second order system is considered in [3] and the set of exactly controllable initial data is characterized (see also [2] and [13]). In the general setting, here again the approximate controllability problem is open.
4. For other systems where controllability issues are considered for systems with continuous spectrum, see [5].

References

1. Ammar Khodja, F., Boussaïd, N., Dupaix, C.: (in preparation)
2. Ammar-Khodja, F., Geymonat, G., Münch, A.: On the exact controllability of a system of mixed order with essential spectrum. C. R. Acad. Sci. Paris Série I **346**, 629–634 (2008)
3. Ammar Khodja, F., Mauffrey, K., Münch, A.: Exact controllability of a system of mixed order with essential spectrum. SIAM J. Control **49**, 1857–1879 (2011)
4. Atkinson, F.V., Langer, H., Mennicken, R., Skalikov, A.A.: The essential spectrum of some matrix operator. Math. Nachr. **167**, 5–20 (1994)
5. Beauchard, K., Coron, J.-M., Rouchon, P.: Controllability issues for continuous-spectrum systems and ensemble controllability of Bloch equations. Commun. Math. Phys. **296**, 525–557 (2010)
6. Cannarsa, P., Martinez, P., Vancostenoble, J.: Global Carleman estimates for degenerate parabolic operators with applications. Mem. Am. Math. Soc. **239**(1133) (2016)
7. Doubova, A., Fernández-Cara, E.: On the control of viscoelastic Jeffreys fluids. Syst. Control Lett. **61**, 573–579 (2012)
8. Egorov, Yu.V., Shubin, M.A. (eds.) Partial Differential Equations II. Springer, Berlin/Heidelberg
9. Geymonat, G., Valente, V.: A noncontrollability result for systems of mixed order. SIAM J. Control Optim. **39**(3), 661–672 (2000)
10. Grubb, G., Geymonat, G.: The essential spectrum of elliptic systems of mixed order. Math. Ann. **227**, 247–276 (1977)
11. Guerrero, S., Imanuvilov, O.Yu.: Remarks on non controllability of the heat equation with memory. ESAIM: COCV **19**, 288–300 (2013)
12. Hislop, P.D., Sigal, I.M.: Introduction to Spectral Theory. Springer, Berlin (1996)
13. Münch, A.: A variational approach to approximate controls for system with essential spectrum: application to membranal arch. Evol. Equ. Control Theory **2**, 119–151 (2013)
14. Palais, R.S.: Seminar on the Atiyah-Singer Index Theorem. Princeton University Press, Princeton (1965)
15. Pang, M.M.H.: L^1 properties of two classes of singular second order elliptic operators. J. Lond. Math. Soc. (2) **38**, 525–543 (1988)

16. Višik, M.I.: Boundary value problems for elliptic equations degenerating on the boundary of a region. Mat. Sb. **35**, 513–568 (1954); English transl. in Am. Math. Soc. Transl. (2) 35 (1964)
17. Wolf, F.: On the essential spectrum of partial differential boundary problems. Commun. Pure Appl. Math. **XII**, 211–228 (1959)

Well-Posedness and Asymptotic Behavior for a Nonlinear Wave Equation

Fágner Dias Araruna, Frederico de Oliveira Matias,
Milton de Lacerda Oliveira, and Shirley Maria Santos e Souza

Dedicated to Prof. Enrique Fernández-Cara on the occasion of his 60th birthday.

Abstract We consider the initial boundary value problem for nonlinear damped wave equations of the form $u'' + M(\int_\Omega |(-\Delta)^s u|^2\,dx)\Delta u + (-\Delta)^\alpha u' = f$, with Neumann boundary conditions. We prove global existence of solutions, when $s \in [1/2, 1]$ and $\alpha \in (0, 1]$, and we show that the energy of these ones decays exponentially, as $t \to \infty$. The uniqueness of solutions is also obtained when $\alpha \in [1/2, 1]$.

Keywords Wave equation · Well-posedness · Asymptotic behavior

AMS Subject Classifications 35L70, 35B40, 74K10

1 Introduction

Problem on vibrations of the elastic bodies has been extensively studied in the last decades. We will look at the following nonlinear model for small deformations of an elastic membrane:

$$u'' + M\left(\int_\Omega \left|(-\Delta)^s u\right|^2 dx\right)\Delta u = f, \tag{1.1}$$

F. D. Araruna (✉) · F. O. Matias · M. L. Oliveira · S. M. S. e Souza
Departamento de Matemática, Universidade Federal da Paraíba, João Pessoa, PB, Brazil
e-mail: fagner@mat.ufpb.br; fred@mat.ufpb.br; milton@mat.ufpb.br; shirley@mat.ufpb.br

A. Doubova et al. (eds.), *Recent Advances in PDEs: Analysis, Numerics and Control*, SEMA SIMAI Springer Series 17,
https://doi.org/10.1007/978-3-319-97613-6_2

where $\Omega \subset \mathbb{R}^n$ is the region occupied by the membrane. In Eq. (1.1), the prime $'$ stands for temporal derivative, $M = M(\lambda)$ is a positive real function defined for all $\lambda \geq 0$ and connected with the initial tension and with the characteristic of the material of the membrane, and Δ is the Laplace operator. The unknown $u = u(x,t)$ represents the vertical displacement of a point x of the membrane at time t, and $f = f(x,t)$ is an external force. Equation (1.1) was derived by Kirchhoff [9] for the case $s = 1/2$ and by Carrier [4] for the case $s = 0$.

Equation (1.1) with different boundary conditions was studied by several authors. For the Kirchhoff equation ((1.1) with $s = 1/2$) we can mention the existence results of Bernstein [2] in the one dimensional case with some restrictions on Fourier series of the data, and the results by Lions [11] and Pohozhaev [24] which considered the data in a special class of analytic functions. Medeiros-Milla Miranda in [14] studied local well-posedness for (1.1) under very weak hypothesis on the data. The general case, when $s \in [0,1]$, was analyzed by Cousin et al. in [6], where the authors obtained existence of global solution in classes of Pohozhaev.

By adding a dissipative mechanism in Eq. (1.1), i.e.,

$$u'' + M\left(\int_\Omega \left|(-\Delta)^s u\right|^2 dx\right)\Delta u + (-\Delta)^\alpha u' = f, \tag{1.2}$$

we can cite several works which have obtained some decay rate of the solutions. For example, for the Kirchhoff equation ($s = 1/2$) with Dirichlet boundary conditions and $\alpha = 0$, we mention Brito [3], Nishihara-Yamada [20], Ono [22], and Yamada [26] which have obtained well-posedness and stability (as $t \to \infty$) results by considering data $(u_0, u_1, f) \in D(-\Delta) \times D((-\Delta)^{1/2}) \times L^2(0, T; D((-\Delta)^{1/2}))$ and satisfying a certain smallness conditions. Here and in what follows, $D(A)$ represents the domain of the operator A. Considering the case $\alpha = 1$ (strong dissipation) we cite the works of Matos-Pereira [13], Mimoni et al. [18], Nishihara [19], Ono [21], and Vasconcelos-Teixeira [25] which contain results of global solvability and exponential decay (as $t \to \infty$) of solutions. Still with Dirichlet boundary conditions and data $(u_0, u_1, f) \in H_0^1 \cap D((-\Delta)^\alpha) \times L^2 \times L^2(0, T; L^2)$, Medeiros and Milla Miranda in [15] obtained global existence and exponential decay (as $t \to \infty$) of solutions of (1.2) when $\alpha \in (0, 1]$. The uniqueness has been proved when $\alpha \in [1/2, 1]$. Considering Neumann boundary condition and $\alpha = 1$, Aassila in [1] studied the global existence and asymptotic behaviour (as $t \to \infty$) of solutions of the Kirchhoff equation. Relative to Carrier equation ($s = 0$) the literature is not so extensive, even so, we can mention Cousin et al. [7], Frota-Goldstein [8], Larkin [10], and Park et al. [23] which analyzed existence of global solutions and energy decay for this one with a nonlinear dissipative term. Besides all the previously mentioned works, we still indicated for the interested readers to consult the works by Medeiros et al. [16, 17] which contain an extensive list of results obtained for Kirchhoff-Carrier equation.

In this work, we consider a problem associated to (1.2) with Neumann boundary conditions, i.e.,

$$\begin{cases} u'' + M\left(\int_\Omega |(-\Delta)^s u|^2\, dx\right)\Delta u + (-\Delta)^\alpha u' = f \quad \text{in} \quad \Omega \times \mathbb{R}_+, \\ \dfrac{\partial u}{\partial \nu} = 0 \quad \text{on} \quad \Gamma \times \mathbb{R}_+, \\ u(\cdot, 0) = u_0, \quad u'(\cdot, 0) = u_1 \quad \text{in} \quad \Omega, \end{cases} \tag{1.3}$$

where Ω is a bounded open set of $\mathbb{R}^n$ with smooth boundary Γ, ν is the unit outward normal to Γ. The purpose of the present paper is to analyze the well-posedness and asymptotic behavior (as $t \to \infty$) of solutions for the problem (1.3) under the following conditions:

$$M(\lambda) \geq m_0 > 0, \quad \forall \lambda \geq 0, \tag{1.4}$$

$$0 < \alpha \leq 1 \quad \text{and} \quad 1/2 \leq s \leq 1, \tag{1.5}$$

and data $(u_0, u_1, f) \in D((-\Delta)^\alpha) \cap D((-\Delta)^{\alpha+s-\frac{1}{2}}) \times D((-\Delta)^{s-\frac{1}{2}}) \times L^2(0, T; D((-\Delta)^{s-\frac{1}{2}}))$. It is important to point out that, as in [15], the initial motivation was to obtain information as $\alpha \to 0$, but to our best knowledge, the existence of global solution of this system with $\alpha = 0$ and no restriction on the data is still an open (and seems to be difficult) problem. Returning to our results, to obtain the existence of solutions for (1.3), we need to construct a complete orthonormal system in a closed subspace of L^2 and to project the problem in this closed subspace. Thus, we can decompose the solutions of problem in two parts: one belonging to the kernel and another in the range of an operator, which corresponds to solutions of the projected problem. We also show that the projected solution decays in an exponential rate. The uniqueness of this solutions is obtained when $\alpha \in [1/2, 1]$. For $\alpha \in (0, 1/2)$ the uniqueness is still an open problem.

The paper is organized as follows. Section 2 contains some notations and essential results which we will apply in this work. In Sect. 3 we prove existence of global solution for (1.3) employing the Faedo-Galerkin method. The key point of the proof is to obtain the complete orthonormal system before mentioned. Concerning to uniqueness we will use energy method with a special regularization. Finally, in Sect. 4 we prove the exponential decay for the energy associated to projected solutions of the problem (1.3) making use of the perturbed energy method as in Zuazua [27].

2 Some Notations and Results

In this section we establish some important results that help us in the development of our work. Also we give some notations and we define the spaces and operators that we will use during the paper. We define the linear operator A_0 in $L^2(\Omega)$ as follows:

$$\left| \begin{array}{l} D(A_0) = \left\{ u \in H^2(\Omega);\ \dfrac{\partial u}{\partial \nu} = 0 \quad \text{on} \quad \Gamma = \partial\Omega \right\}, \\ A_0 u = -\Delta u, \quad \forall u \in D(A_0). \end{array} \right. \tag{2.1}$$

It is well know that the operator A_0 is nonnegative, selfadjoint and the resolvent $(I + \lambda A_0)^{-1}$ is compact for all $\lambda > 0$.

We recall a result from [1] that will be needed in the sequel.

Lemma 2.1 *Let $\mathcal{H}$ be a real Hilbert space with inner product $(\cdot, \cdot)$ and norm $|\cdot|$. Consider $\mathcal{A} : D(\mathcal{A}) \subsetneq \mathcal{H} \to \mathcal{H}$ a nonnegative selfadjoint operator with domain $D(\mathcal{A})$ and range $R(\mathcal{A})$ in H. Suppose that $(I + \mathcal{A})^{-1}$ is a compact operator. Then*

- (i) *$R(\mathcal{A})$ is closed and $\mathcal{H} = N(\mathcal{A}) \oplus R(\mathcal{A})$,*
- (ii) *The operator $\left[\mathcal{A}|_{D(\mathcal{A})\cap R(\mathcal{A})}\right]^{-1} : R(\mathcal{A}) \to R(\mathcal{A})$ is compact, where $\mathcal{A}|_{D(\mathcal{A})\cap R(\mathcal{A})}$ is the restriction of $\mathcal{A}$ to $D(\mathcal{A}) \cap R(\mathcal{A})$.*

According to Lemma 2.1, we can guarantee, for the operator A_0 defined in (2.1), that

$$R(A_0) = \left\{ v \in L^2(\Omega);\ \textstyle\int_\Omega v(x)dx = 0 \right\} \text{ is closed in } L^2(\Omega),$$

$$L^2(\Omega) = N(A_0) \oplus R(A_0), \text{ with } N(A_0) = \{v(x) = \text{constant a.e. in } \Omega\},$$

and

$$\left[A_0|_{D(A_0)\cap R(A_0)}\right]^{-1} : R(A_0) \to R(A_0) \text{ is compact.}$$

Let $P : L^2(\Omega) \to R(A_0)$ be the orthogonal projection of $L^2(\Omega)$ onto $R(A_0)$. Then

$$Pu(x) = u(x) - \overline{u}, \quad \forall u \in L^2(\Omega),$$

where $\overline{u} = \frac{1}{|\Omega|}\int_\Omega u(x)dx$ and $|\Omega|$ is the measure of Ω.

Let us denote by $(\cdot, \cdot)$ and $|\cdot|$ the inner product and norm in $L^2(\Omega)$, respectively.

We consider the system

$$\left\{ \begin{array}{l} u'' + M\left(\left|A_0^s u\right|^2\right) A_0 u + A_0^\alpha u' = f \quad \text{in} \quad L^2(\Omega), \\ u(0) = u_0, \quad u'(0) = u_1. \end{array} \right. \tag{2.2}$$

If $u(t)$ is a solution to (2.2), then by Lemma 2.1 we have $u(t) = u_1(t) + u_2(t)$, where $u_1(t) \in N(A_0)$ and $u_2(t) \in D(A_0) \cap R(A_0)$. Furthermore, we deduce that

$$\left\{ \begin{array}{l} u_1'' + u_2'' + M\left(\left|A_0^s u_2\right|^2\right) A_0 u_2 + A_0^\alpha u_2' = f \quad \text{in} \quad L^2(\Omega), \\ u(0) = u_1(0) + u_2(0) = u_{01} + u_{02}, \\ u'(0) = u_1'(0) + u_2'(0) = u_{11} + u_{12}, \end{array} \right.$$

where we have used the fact that $A_0^s u(t) = A_0^s u_2(t)$ and $A_0^\alpha u'(t) = A_0^\alpha u_2'(t)$. In this way, we can decompose the last system as follows:

$$\begin{cases} u_1''(t) = 0 \quad \text{in} \quad N(A_0), \\ u_1(0) = u_{01}, \quad u_1'(0) = u_{11}, \end{cases} \tag{2.3}$$

and

$$\begin{cases} u_2'' + M\left(|A^s u_2|^2\right) Au_2 + A^\alpha u_2' = f \quad \text{in} \quad R(A), \\ u_2\,(0) = u_{02}, \quad u_2'(0) = u_{12}, \end{cases} \tag{2.4}$$

where $A = A_0|_{D(A_0)\cap R(A_0)}$. If we can solve (2.3) and (2.4), we will get the solution $u(t) = u_1(t) + u_2(t)$ for (2.2).

For (2.3), we obtain the following explicit solution:

$$u_1(t) = u_{01} + u_{11}t.$$

The analysis of the well-posedness of global (weak) solutions of (2.4), when

$$u_0 \in V := D(A^s) \cap D(A^{\alpha+s-\frac{1}{2}}), \;\; u_1 \in H := D(A^{s-\frac{1}{2}}), \;\; \text{and} \;\; f \in L^2(0, T; H),$$

(α and s as in (1.5)) and the their asymptotic behavior, as $t \to \infty$, are our objectives in this paper. This will be done in the next two sections.

3 Well-Posedness

This section is devoted to show the well-posedness for the system (2.4). The following result holds.

Theorem 3.1 *Let us suppose $M \in C^0([0, \infty[, \mathbb{R})$, s, and α satisfying (1.4) and (1.5), and let us consider data $(u_0, u_1, f) \in V \times H \times L^2(0, T; H)$. Then there exists at least a function $u : \Omega \times [0, T] \to \mathbb{R}$ verifying the following conditions:*

$$u \in L^\infty(0, T; V) \cap L^2(0, T; D(A^{\frac{\alpha}{2}+s})), \tag{3.1}$$

$$u' \in L^\infty(0, T; H) \cap L^2(0, T; D(A^{\frac{\alpha}{2}+s-\frac{1}{2}})), \tag{3.2}$$

$$u'' + M\left(\left|A^s u\right|^2\right) Au + A^\alpha u' = f \quad \textit{in} \quad L^2(0, T; D(A^{\frac{\alpha}{2}+s-1}) \cap D(A^{-\frac{\alpha}{2}+s-\frac{1}{2}})), \tag{3.3}$$

$$u\,(0) = u_0, \quad u'\,(0) = u_1 \quad \textit{in} \quad \Omega. \tag{3.4}$$

Furthermore, if $M \in C^1([0,\infty[,\mathbb{R})$ and $\alpha \geq 1/2$, the function u satisfying (3.1)–(3.4) is unique.

Proof To prove the existence of solutions, we will use the Faedo-Galerkin method. For this, we consider $\{w_\nu\}_{\nu\in\mathbb{N}}$ a special basis in $R(A)$ formed by eigenvectors of A, whose eigenvalues $\{\lambda_\nu\}_{\nu\in\mathbb{N}}$ are such that $0 < \lambda_1 < \lambda_2 \leq \lambda_3 \leq \ldots \leq \lambda_\nu \leq \ldots$ with $\lim_{\nu\to\infty}(\lambda_\nu) = \infty$. We denote by $W_m = [w_1, w_2, \ldots, w_m]$ the subspace of V generated by the first m vectors of $\{w_\nu\}_{\nu\in\mathbb{N}}$. Let us find an approximate solution $u_m = u_m(t) \in W_m$ defined by $u_m(t) = \sum_{j=1}^m g_{jm}(t) w_j$, where $g_{jm}(t)$ are found as solutions of the following initial value problem for the system of ordinary differential equations:

$$\begin{cases} (u_m''(t), v) + M\left(|A^s u_m|^2\right)(Au_m(t), v) + (A^\alpha u_m'(t), v) = (f(t), v), \ \forall v \in W_m, \\ u_m(0) = u_{0m} \to u_0 \quad \text{in} \quad V, \\ u_m'(0) = u_{1m} \to u_1 \quad \text{in} \quad H. \end{cases} \tag{3.5}$$

System (3.5) has solutions u_m defined on a certain interval $[0, t_m]$, for $t_m < T$ (see, for example, [5, Th. 1.1, p. 43]). Moreover, the functions u_m and u_m' are absolutely continuous in this interval. Thus, we can guarantee the existence of u_m'' almost everywhere in $[0, t_m]$. This solution can be extended to whole interval $[0, T]$ by using the first estimate that we shall prove in the next step.

Estimate I Taking $v = 2A^{2s-1}u_m'(t)$ in $(3.5)_1$, we have

$$\begin{aligned} &(u_m''(t), 2A^{2s-1}u_m'(t)) + M\left(|A^s u_m(t)|^2\right)(Au_m(t), 2A^{2s-1}u_m'(t)) \\ &+(A^\alpha u_m'(t), 2A^{2s-1}u_m'(t)) = (f(t), 2A^{2s-1}u_m'(t)). \end{aligned}$$

So

$$\begin{aligned} &\frac{d}{dt}\left\{\left|A^{s-\frac{1}{2}}u_m'(t)\right|^2 + \widehat{M}\left(|A^s u_m(t)|^2\right)\right\} + 2\left|A^{\frac{\alpha}{2}+s-\frac{1}{2}}u_m'(t)\right|^2 \\ &\leq 2\left|A^{s-\frac{1}{2}}f(t)\right|\left|A^{s-\frac{1}{2}}u_m'(t)\right| \leq \left|A^{s-\frac{1}{2}}f(t)\right|^2 + \left|A^{s-\frac{1}{2}}u_m'(t)\right|^2, \end{aligned}$$

where $\widehat{M}(\lambda) = \int_0^\lambda M(t)dt$. Integrating from 0 to t, $0 \leq t \leq t_m$, we obtain

$$\begin{aligned} &\left|A^{s-\frac{1}{2}}u_m'(t)\right|^2 + \widehat{M}\left(|A^s u_m(t)|^2\right) + 2\int_0^t \left|A^{\frac{\alpha}{2}+s-\frac{1}{2}}u_m'(\xi)\right|^2 d\xi \\ &\leq \left|A^{s-\frac{1}{2}}u_{1m}\right|^2 + \widehat{M}\left(|A^s u_{0m}|^2\right) + 2\int_0^t \left|A^{s-\frac{1}{2}}f(\xi)\right|^2 d\xi \\ &+\int_0^t \left|A^{s-\frac{1}{2}}u_m'(\xi)\right|^2 d\xi. \end{aligned}$$

In this way, by (3.5)$_2$, (3.5)$_3$, and since $f \in L^2(0,T;H)$, the above inequality implies that

$$\begin{aligned}&\left|A^{s-\frac{1}{2}}u_m'(t)\right|^2 + m_0\,|A^s u_m(t)|^2 + 2\int_0^t \left|A^{\frac{\alpha}{2}+s-\frac{1}{2}}u_m'(\xi)\right|^2 d\xi\\ &\leq C + \int_0^t \left|A^{s-\frac{1}{2}}u_m'(\xi)\right|^2 d\xi,\end{aligned}$$

where $C > 0$ is a constant independent of m and t. Thus, by Gronwall's Lemma, we obtain

$$\left|A^{s-\frac{1}{2}}u_m'(t)\right|^2 + m_0\left|A^s u_m(t)\right|^2 + 2\int_0^t \left|A^{\frac{\alpha}{2}+s-\frac{1}{2}}u_m'(\xi)\right|^2 d\xi \leq C,$$

where $C > 0$ is a constant independent of m and t. Therefore

$$\left|\begin{array}{l}(u_m) \text{ is bounded in } L^\infty(0,T;D(A^s)),\\ (u_m') \text{ is bounded in } L^\infty(0,T;H)\cap L^2(0,T;D(A^{\frac{\alpha}{2}+s-\frac{1}{2}}))\end{array}\right. \tag{3.6}$$

and, consequently, we can prolong the approximate solution $u_m(t)$ for all t in $[0,T]$.

Estimate II Let us consider $v = A^{2s+\alpha-1}u_m(t)$ in (3.5)$_1$, then

$$\begin{aligned}&\frac{d}{dt}(u_m'(t), A^{2s+\alpha-1}u_m(t)) - \left|A^{\frac{\alpha}{2}+s-\frac{1}{2}}u_m'(t)\right|^2 + M\left(|A^s(u_m(t))|^2\right)\left|A^{s+\frac{\alpha}{2}}u_m\right|^2\\ &+\frac{1}{2}\frac{d}{dt}\left|A^{s+\alpha-\frac{1}{2}}u_m(t)\right|^2 = (f(t), A^{2s+\alpha-1}u_m(t)).\end{aligned}$$

Integrating this identity from 0 to t, $t \in [0,T]$, we get

$$\begin{aligned}&\frac{1}{2}\left|A^{s+\alpha-\frac{1}{2}}u_m(t)\right|^2 + \int_0^t M\left(|A^s u_m(\xi)|^2\right)\left|A^{s+\frac{\alpha}{2}}u_m(\xi)\right|^2 d\xi\\ &= -(A^{s-\frac{1}{2}}u_m'(t), A^{s+\alpha-\frac{1}{2}}u_m(t))\\ &+(A^{s-\frac{1}{2}}u_{1m}, A^{s+\alpha-\frac{1}{2}}u_{0m}) + \frac{1}{2}\left|A^{s+\alpha-\frac{1}{2}}u_{0m}\right|^2\\ &+\int_0^t \left|A^{\frac{\alpha}{2}+s-\frac{1}{2}}u_m'(\xi)\right|^2 d\xi + \int_0^t (A^{s-\frac{1}{2}}f(\xi), A^{s+\alpha-\frac{1}{2}}u_m(\xi))d\xi.\end{aligned}$$

Using the Young's inequality and (3.6)$_2$ we obtain

$$\begin{aligned}&\frac{1}{4}\left|A^{s+\alpha-\frac{1}{2}}u_m(t)\right|^2 + m_0\int_0^t \left|A^{s+\frac{\alpha}{2}}u_m(\xi)\right|^2 d\xi \leq C + 4\left|A^{s-\frac{1}{2}}u_m'(t)\right|^2\\ &+\frac{1}{2}\int_0^T \left|A^{s-\frac{1}{2}}f(t)\right|^2 dt + \frac{1}{2}\int_0^t \left|A^{s+\alpha-\frac{1}{2}}u_m(\xi)\right|^2 d\xi \leq C + \int_0^t \left|A^{s+\alpha-\frac{1}{2}}u_m(\xi)\right|^2 d\xi,\end{aligned}$$

where $C > 0$ is a constant independent of m and t, $t \in [0, T]$. So, applying again the Gronwall's Lemma, we can conclude that

$$\frac{1}{4}\left|A^{s+\alpha-\frac{1}{2}}u_m(t)\right|^2 + m_0\int_0^t \left|A^{s+\frac{\alpha}{2}}u_m(\xi)\right|^2 d\xi \leq C,$$

where $C > 0$ is constant independent of m and t, $t \in [0, T]$. Therefore

$$(u_m) \text{ is bounded in } L^\infty(0, T; D(A^{s+\alpha-\frac{1}{2}})) \cap L^2(0, T; D(A^{s+\frac{\alpha}{2}})). \tag{3.7}$$

Passage to the Limit From estimates (3.6) and (3.7), there exists a subsequence of (u_m), still denoted in the same form, such that

$$\left| \begin{array}{llll} u_m \to u & \text{weak} - * & \text{in} & L^\infty(0, T; D(A^s) \cap D(A^{s+\alpha-\frac{1}{2}})), \\ u_m \to u & \text{weakly} & \text{in} & L^2(0, T; D(A^{s+\frac{\alpha}{2}})), \\ u_m' \to u' & \text{weak} - * & \text{in} & L^\infty(0, T; H), \\ u_m' \to u' & \text{weakly} & \text{in} & L^2(0, T; D(A^{\frac{\alpha}{2}+s-\frac{1}{2}})). \end{array} \right. \tag{3.8}$$

To treat the convergence of the nonlinear term, we observe that, since the injections $D(A^{s+\frac{\alpha}{2}}) \subset D(A^s) \subset H$ are continuous and the embedding of $D(A^{s+\frac{\alpha}{2}})$ into $D(A^s)$ is compact, it follows by (3.6), (3.7), and Aubin-Lions' Compactness Theorem that exists a subsequence of (u_m), which we still denote by (u_m), and a function $u : [0, T] \to \mathbb{R}$, such that

$$u_m \to u \quad \text{strongly in} \quad L^2(0, T; D(A^s)).$$

Then there exists a subsequence of (u_m), which we still denote by (u_m), such that

$$\left|A^s u_m(t)\right|^2 \to \left|A^s u(t)\right|^2 \quad \text{a.e. in} \quad (0, T).$$

By the continuity of M, we have

$$M\left(\left|A^s u_m(t)\right|^2\right) \to M\left(\left|A^s u(t)\right|^2\right) \quad \text{a.e. in} \quad (0, T),$$

and

$$M\left(\left|A^s u_m(t)\right|^2\right) \leq C \quad \text{a.e. in} \quad (0, T).$$

Thus, by the Lebesgue's Dominated Convergence Theorem, we get

$$M\left(\left|A^s u_m(t)\right|^2\right) \to M\left(\left|A^s u(t)\right|^2\right) \quad \text{strongly in} \quad L^2(0, T). \tag{3.9}$$

The convergences (3.8) and (3.9) are sufficient to pass to the limit in $(3.5)_1$ and to obtain the function u satisfying (3.1)–(3.3). By standard arguments, we can verify the initial conditions (3.4).

Before proving the uniqueness, we consider the following lemma.

Lemma 3.1 *If* $\frac{1}{2} \leq \alpha \leq 1$, $u \in L^2(0, T; D(A^{s+\frac{\alpha}{2}})$, *and* $u' \in L^2(0, T; D(A^{s+\frac{\alpha}{2}-\frac{1}{2}}))$, *then*

$$\frac{d}{dt}\left|A^s u\right|^2 = 2\left(A^{s-\frac{\alpha}{2}+\frac{1}{2}}u,\, A^{s+\frac{\alpha}{2}-\frac{1}{2}}u'\right). \tag{3.10}$$

Proof *We consider the space* $W(0, T)$ *defined by*

$$W(0,T) = \left\{v;\ v \in L^2\left(0, T; D(A^{s+\frac{\alpha}{2}})\right),\ v' \in L^2(0, T; D(A^{s+\frac{\alpha}{2}-\frac{1}{2}}))\right\}$$

equipped with the norm

$$\|v\|^2_{W(0,T)} = \|v\|^2_{L^2(0,T;D(A^{s+\frac{\alpha}{2}}))} + \|v\|^2_{L^2(0,T;D(A^{s+\frac{\alpha}{2}-\frac{1}{2}}))}.$$

By Lions-Magenes [12, p. 13], we have that $\mathcal{D}([0, T]; D(A^{s+\frac{\alpha}{2}}))$ *is dense in* $W(0, T)$. *Taking* $\varphi \in \mathcal{D}([0, T]; D(A^{s+\frac{\alpha}{2}}))$, *it follows that* $\varphi' \in \mathcal{D}\left([0, T]; D(A^{s+\frac{\alpha}{2}})\right)$. *We also have* $D(A^{s+\frac{\alpha}{2}}) \subset D(A^{s-\frac{\alpha}{2}+\frac{1}{2}})$ *with continuous injections, because* $s + \frac{\alpha}{2} \geq s - \frac{\alpha}{2} + \frac{1}{2}$. *In this way, we can assert that* $A^{s+\frac{\alpha}{2}-\frac{1}{2}}\varphi, A^{s+\frac{\alpha}{2}-\frac{1}{2}}\varphi' \in D\left(A^{1-\alpha}\right)$ *and*

$$\begin{aligned}
\frac{d}{dt}\left|A^s\varphi\right|^2 &= \frac{d}{dt}\left(A^{s-\frac{\alpha}{2}+\frac{1}{2}}\varphi, A^{s+\frac{\alpha}{2}-\frac{1}{2}}\varphi\right)\\
&= \left(A^{s-\frac{\alpha}{2}+\frac{1}{2}}\varphi', A^{s+\frac{\alpha}{2}-\frac{1}{2}}\varphi\right) + \left(A^{s-\frac{\alpha}{2}+\frac{1}{2}}\varphi, A^{s+\frac{\alpha}{2}-\frac{1}{2}}\varphi'\right)\\
&= \left(A^{1-\alpha}\left(A^{s+\frac{\alpha}{2}-\frac{1}{2}}\varphi'\right), A^{s+\frac{\alpha}{2}-\frac{1}{2}}\varphi\right) + \left(A^{s-\frac{\alpha}{2}+\frac{1}{2}}\varphi, A^{s+\frac{\alpha}{2}-\frac{1}{2}}\varphi'\right)\\
&= \left(A^{s+\frac{\alpha}{2}-\frac{1}{2}}\varphi', A^{1-\alpha}\left(A^{s+\frac{\alpha}{2}-\frac{1}{2}}\varphi\right)\right) + \left(A^{s-\frac{\alpha}{2}+\frac{1}{2}}\varphi, A^{s+\frac{\alpha}{2}-\frac{1}{2}}\varphi'\right)\\
&= \left(A^{s+\frac{\alpha}{2}-\frac{1}{2}}\varphi', A^{s-\frac{\alpha}{2}+\frac{1}{2}}\varphi\right) + \left(A^{s-\frac{\alpha}{2}+\frac{1}{2}}\varphi, A^{s+\frac{\alpha}{2}-\frac{1}{2}}\varphi'\right)\\
&= 2\left(A^{s-\frac{\alpha}{2}+\frac{1}{2}}\varphi, A^{s+\frac{\alpha}{2}-\frac{1}{2}}\varphi'\right),
\end{aligned}$$

for all $\varphi \in \mathcal{D}\left([0, T]; D(A^{s+\frac{\alpha}{2}})\right)$. *In this way, using density arguments, we get (3.10) and this proves the lemma.* ■

Returning to uniqueness of solution, to prove it, we will make use of the energy method with a special regularization. In fact, firstly we observe that we can not multiply Eq. (3.3) by $A^{2s-1}u'$ directly because $A^{2s-1}u' \in L^{\infty}(0,T; D(A^{s-\frac{1}{2}})^{*}) \cap L^{2}(0,T, D(A^{s-\frac{\alpha}{2}-\frac{1}{2}})^{*})$ and $u'' \in L^{2}(0,T; D(A^{\frac{\alpha}{2}})^{*})$ and therefore the duality $\langle u'', A^{2s-1}u'\rangle$ does not make sense. To overcome this difficulty, let us consider the function u defined over $\mathbb{R}$ with the properties analogous with the properties of u over $[0,T]$ (which is possible by reflection). Let us consider a sequence of mollifiers $\{\rho_\varepsilon\}_{\varepsilon>0}$, that is, a sequence of functions $\rho_\varepsilon \geq 0$ on $\mathbb{R}$ such that

$$\rho_\varepsilon \in C_c^{\infty}(\mathbb{R}), \quad \operatorname{supp}\rho_\varepsilon \subset [-\varepsilon,\varepsilon], \quad \int_{-\infty}^{\infty}\rho_\varepsilon(s)\,ds = 1.$$

Taking

$$u_\varepsilon(x,t) = \int_{-\infty}^{\infty}\rho_\varepsilon(t-s)u(x,s)ds,$$

we can see that $u_\varepsilon'' \in L^{2}(0,T, D(A^{s-\frac{1}{2}}))$ and so the duality $\left\langle u_\varepsilon'', A^{2s-1}u_\varepsilon'\right\rangle$ makes sense.

Let us suppose that u and v are two solutions in the conditions of Theorem 3.1. Defining $w_\varepsilon = u_\varepsilon - v_\varepsilon$ and $w = u - v$, we have that

$$w_\varepsilon'' + \rho_\varepsilon * \left[M\left(\left|A^{s}u\right|^{2}\right)Au - M\left(\left|A^{s}v\right|^{2}\right)Av\right] + A^{\alpha}w_\varepsilon' = 0. \tag{3.11}$$

Making the duality between (3.11) and $A^{2s-1}w_\varepsilon'$, one has

$$\begin{aligned}&\left\langle w_\varepsilon'', A^{2s-1}w_\varepsilon'\right\rangle + \left\langle \rho_\varepsilon * M\left(|A^{s}u|^{2}\right)Aw, A^{2s-1}w_\varepsilon'\right\rangle \\ &+\left\langle \rho_\varepsilon * \left[M\left(|A^{s}u|^{2}\right) - M\left(|A^{s}v|^{2}\right)\right]Av, A^{2s-1}w_\varepsilon'\right\rangle + \left\langle A^{\alpha}w_\varepsilon', A^{2s-1}w_\varepsilon'\right\rangle = 0.\end{aligned} \tag{3.12}$$

Notice that

$$\begin{aligned}\left\langle \rho_\varepsilon * M\left(|A^{s}u|^{2}\right)Aw, A^{2s-1}w_\varepsilon'\right\rangle &= \left\langle \rho_\varepsilon * M\left(|A^{s}u|^{2}\right)A^{s-\frac{\alpha}{2}+\frac{1}{2}}w, A^{\frac{\alpha}{2}+s-\frac{1}{2}}w_\varepsilon'\right\rangle \\ &= \left\langle \rho_\varepsilon * M\left(|A^{s}u|^{2}\right)A^{s-\frac{\alpha}{2}+\frac{1}{2}}w, A^{\frac{\alpha}{2}+s-\frac{1}{2}}w_\varepsilon' - A^{\frac{\alpha}{2}+s-\frac{1}{2}}w'\right\rangle \\ &+ \left\langle \rho_\varepsilon * M\left(|A^{s}u|^{2}\right)A^{s-\frac{\alpha}{2}+\frac{1}{2}}w, A^{\frac{\alpha}{2}+s-\frac{1}{2}}w'\right\rangle.\end{aligned} \tag{3.13}$$

By Lemma 3.1, we have

$$\begin{aligned}\left\langle \rho_\varepsilon * M\left(|A^{s}u|^{2}\right)A^{s-\frac{\alpha}{2}+\frac{1}{2}}w, A^{\frac{\alpha}{2}+s-\frac{1}{2}}w'\right\rangle &= \frac{1}{2}\frac{d}{dt}\left\langle \rho_\varepsilon * M\left(|A^{s}u|^{2}\right)A^{s}w, A^{s}w\right\rangle \\ &- \left\langle \rho_\varepsilon * [M'(|A^{s}u|^{2})\left(A^{s-\frac{\alpha}{2}+\frac{1}{2}}u, A^{s+\frac{\alpha}{2}-\frac{1}{2}}u'\right)]A^{s}w, A^{s}w\right\rangle.\end{aligned} \tag{3.14}$$

Substituting (3.13) and (3.14) into (3.12) we get

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\left(\left|A^{s-\frac{1}{2}}w'_{\varepsilon}\right|^2+\left\langle\rho_{\varepsilon}*M(|A^s u|^2)A^s w,\,A^s w\right\rangle\right)+\left|A^{\frac{\alpha}{2}+s-\frac{1}{2}}w'_{\varepsilon}\right|^2\\
&=\left\langle\rho_{\varepsilon}*\left[M\left(|A^s v|^2\right)-M\left(|A^s u|^2\right)\right]A^{s-\frac{\alpha}{2}+\frac{1}{2}}v,\,A^{\frac{\alpha}{2}+s-\frac{1}{2}}w'_{\varepsilon}\right\rangle\\
&+\left\langle\rho_{\varepsilon}*\left[M'\left(|A^s u|^2\right)\left(A^{s-\frac{\alpha}{2}+\frac{1}{2}}u,\,A^{s+\frac{\alpha}{2}-\frac{1}{2}}u'\right)\right]A^s w,\,A^s w\right\rangle\\
&+\left\langle\rho_{\varepsilon}*M\left(|A^s u|^2\right)A^{s-\frac{\alpha}{2}+\frac{1}{2}}w,\,A^{\frac{\alpha}{2}+s-\frac{1}{2}}w'-A^{\frac{\alpha}{2}+s-\frac{1}{2}}w'_{\varepsilon}\right\rangle.
\end{aligned}\tag{3.15}$$

Integrating (3.15) from 0 to $t \le T$ we have

$$\begin{aligned}
&\frac{1}{2}\left(\left|A^{s-\frac{1}{2}}w'_{\varepsilon}\right|^2+\left\langle\rho_{\varepsilon}*M\left(|A^s u|^2\right)A^s w,\,A^s w\right\rangle\right)+\int_0^t\left|A^{\frac{\alpha}{2}+s-\frac{1}{2}}w'_{\varepsilon}\right|^2 ds\\
&=\int_0^t\left\langle\rho_{\varepsilon}*\left[M\left(|A^s v|^2\right)-M\left(|A^s u|^2\right)\right]A^{s-\frac{\alpha}{2}+\frac{1}{2}}v,\,A^{\frac{\alpha}{2}+s-\frac{1}{2}}w'_{\varepsilon}\right\rangle ds\\
&+\int_0^t\left\langle\rho_{\varepsilon}*\left[M'\left(|A^s u|^2\right)\left(A^{s-\frac{\alpha}{2}+\frac{1}{2}}u,\,A^{s+\frac{\alpha}{2}-\frac{1}{2}}u'\right)\right]A^s w,\,A^s w\right\rangle ds\\
&+\int_0^t\left\langle\rho_{\varepsilon}*M\left(|A^s u|^2\right)A^{s-\frac{\alpha}{2}+\frac{1}{2}}w,\,A^{\frac{\alpha}{2}+s-\frac{1}{2}}w'-A^{\frac{\alpha}{2}+s-\frac{1}{2}}w'_{\varepsilon}\right\rangle ds.
\end{aligned}$$

We can rewrite the above equality as follows

$$\begin{aligned}
&\frac{1}{2}\left(\left|A^{s-\frac{1}{2}}w'_{\varepsilon}\right|^2+\left\langle\rho_{\varepsilon}*M\left(|A^s u|^2\right)A^s w,\,A^s w\right\rangle\right)+\int_0^t\left|A^{\frac{\alpha}{2}+s-\frac{1}{2}}w'_{\varepsilon}\right|^2 ds\\
&=\int_0^t\left\langle\rho_{\varepsilon}*\left[M\left(|A^s v|^2\right)-M\left(|A^s u|^2\right)\right]A^{s-\frac{\alpha}{2}+\frac{1}{2}}v,\,A^{\frac{\alpha}{2}+s-\frac{1}{2}}w'\right\rangle ds\\
&+\int_0^t\left\langle\rho_{\varepsilon}*\left[M\left(|A^s v|^2\right)-M\,|A^s u|^2\right]A^{s-\frac{\alpha}{2}+\frac{1}{2}}v,\,A^{\frac{\alpha}{2}+s-\frac{1}{2}}w'_{\varepsilon}-A^{\frac{\alpha}{2}+s-\frac{1}{2}}w'\right\rangle ds\\
&+\int_0^t\left\langle\rho_{\varepsilon}*\left[M'\left(|A^s u|^2\right)\left(A^{s-\frac{\alpha}{2}+\frac{1}{2}}u,\,A^{s+\frac{\alpha}{2}-\frac{1}{2}}u'\right)\right]A^s w,\,A^s w\right\rangle ds\\
&+\int_0^t\left\langle\rho_{\varepsilon}*M\left(|A^s u|^2\right)A^{s-\frac{\alpha}{2}+\frac{1}{2}}w,\,A^{\frac{\alpha}{2}+s-\frac{1}{2}}w'-A^{\frac{\alpha}{2}+s-\frac{1}{2}}w'_{\varepsilon}\right\rangle ds.
\end{aligned}\tag{3.16}$$

Notice that, as $\varepsilon \to 0$, we have

$$\int_0^t\left\langle\rho_{\varepsilon}*\left[M\left(|A^s v|^2\right)-M\left(|A^s u|^2\right)\right]A^{s-\frac{\alpha}{2}+\frac{1}{2}}v,\,A^{\frac{\alpha}{2}+s-\frac{1}{2}}w'_{\varepsilon}-A^{\frac{\alpha}{2}+s-\frac{1}{2}}w'\right\rangle ds\to 0\tag{3.17}$$

and

$$\int_0^t\left\langle\rho_{\varepsilon}*M\left(\left|A^s u\right|^2\right)A^{s-\frac{\alpha}{2}+\frac{1}{2}}w,\,A^{\frac{\alpha}{2}+s-\frac{1}{2}}w'-A^{\frac{\alpha}{2}+s-\frac{1}{2}}w'_{\varepsilon}\right\rangle ds\to 0.\tag{3.18}$$

Thus, passing (3.16) to the limit, as $\varepsilon \to 0$, and taking into account the convergences in (3.17) and (3.18), we get

$$\begin{aligned}
&\frac{1}{2}\left|A^{s-\frac{1}{2}}w'\right|^2 + \left\langle M\left(|A^s u|^2\right)A^s w, A^s w\right\rangle + \int_0^t \left|A^{\frac{\alpha}{2}+s-\frac{1}{2}}w'\right|^2 ds \\
&= \int_0^t \left\langle \left[M\left(|A^s v|^2\right) - M\left(|A^s u|^2\right)\right] A^{s-\frac{\alpha}{2}+\frac{1}{2}} v, A^{\frac{\alpha}{2}+s-\frac{1}{2}}w'\right\rangle ds \\
&+2\int_0^t \left\langle M'\left(|A^s u|^2\right)\left(A^{s-\frac{\alpha}{2}+\frac{1}{2}}u, A^{s+\frac{\alpha}{2}-\frac{1}{2}}u'\right)A^s w, A^s w\right\rangle ds.
\end{aligned} \tag{3.19}$$

Using (1.4) and the fact that $s + \frac{\alpha}{2} \geq s + \frac{1}{2} - \frac{\alpha}{2}$, we have that there exists $\xi = \xi\,(t)$ between $|A^s u\,(t)|^2$ and $|A^s v\,(t)|^2$ such that

$$\begin{aligned}
&\left\langle \left[M\left(\left|A^s v\right|^2\right) - M\left(\left|A^s u\right|^2\right)\right] A^{s-\frac{\alpha}{2}+\frac{1}{2}} v, A^{\frac{\alpha}{2}+s-\frac{1}{2}}w'\right\rangle \\
&\leq \left|M'(\xi)\right| \left(\left|A^s u\right| + \left|A^s v\right|\right) \left|\left|A^s u\right| - \left|A^s v\right|\right| \left|\left\langle A^{s-\frac{\alpha}{2}+\frac{1}{2}} v, A^{\frac{\alpha}{2}+s-\frac{1}{2}}w'\right\rangle\right| \\
&\leq C\left|A^s w\right| \left|\left\langle A^{s+\frac{1}{2}-\frac{\alpha}{2}} v, A^{s-\frac{1}{2}+\frac{\alpha}{2}}w'\right\rangle\right| \\
&\leq C\left|A^s w\right| \left|A^{s+\frac{1}{2}-\frac{\alpha}{2}} v\right| \left|A^{s-\frac{1}{2}+\frac{\alpha}{2}}w'\right| \leq C\left|A^{s+\frac{1}{2}-\frac{\alpha}{2}} v\right|^2 \left|A^s w\,(t)\right|^2 + \frac{1}{2}\left|A^{s-\frac{1}{2}+\frac{\alpha}{2}}w'\right|^2 \\
&\leq C\left|A^{s+\frac{\alpha}{2}} v\right|^2 \left|A^s w\right|^2 + \frac{1}{2}\left|A^{s-\frac{1}{2}+\frac{\alpha}{2}}w'\right|^2.
\end{aligned} \tag{3.20}$$

We can also note by (1.4) that

$$\left\langle M'\left(\left|A^s u\right|^2\right)\left(A^{s-\frac{\alpha}{2}+\frac{1}{2}}u, A^{s+\frac{\alpha}{2}-\frac{1}{2}}u'\right)A^s w, A^s w\right\rangle \leq C\left(\left|A^{s+\frac{\alpha}{2}}u\right|^2 + \left|A^{s+\frac{\alpha}{2}-\frac{1}{2}}u'\right|^2\right)\left|A^s w\right|^2. \tag{3.21}$$

Combining (3.19)–(3.21), it follows that

$$m_0\left|A^s w\right|^2 + \frac{1}{2}\int_0^t \left|A^{\frac{\alpha}{2}+s-\frac{1}{2}}w'\right|^2 ds \leq C\int_0^t h(s)\left|A^s w(s)\right|^2 ds, \tag{3.22}$$

with $h\,(t) = \left|A^{s+\frac{\alpha}{2}}u\,(t)\right|^2 + \left|A^{s+\frac{\alpha}{2}}v\,(t)\right|^2 + \left|A^{s+\frac{\alpha}{2}-\frac{1}{2}}u'\,(t)\right|^2 \in L^1\,(0, T)$. Applying the Gronwall's Lemma in (3.22), we conclude that $w(t) = 0$, for all $t \in [0, T]$, and this gives the uniqueness. ■

Remark 3.1 As an immediate consequence of the estimates to obtain existence of solutions in the proof of Theorem 3.1, we have that if $f(x, \cdot)$ is defined in the interval $(0, \infty)$, then (3.1)–(3.3) hold when we consider $T = \infty$.

4 Asymptotic Behavior

The aim of this section is to study the asymptotic behavior, as $t \to \infty$, of the energy $E(t)$ associated to solution of the problem (2.4) (with $f = 0$). This energy is given by

$$E(t) = \frac{1}{2}\left|A^{s-\frac{1}{2}}u'(t)\right|^2 + \frac{1}{2}\widehat{M}\left(\left|A^s u(t)\right|^2\right), \quad \forall t \geq 0. \tag{4.1}$$

Recall that $\widehat{M}(\lambda) = \int_0^\lambda M(t)dt$. The main result of this section is the following.

Theorem 4.1 *Under the assumptions of Theorem 3.1 with $f = 0$, there exist positive constants C and γ such that the energy (4.1) satisfies*

$$E(t) \leq CE(0)e^{-\gamma t}, \quad \forall t \geq 0. \tag{4.2}$$

Proof A simple computation gives

$$E_m'(t) = -\left|A^{s+\frac{\alpha}{2}-\frac{1}{2}}u_m'\right|^2 \leq -\lambda_1^\alpha \left|A^{s-\frac{1}{2}}u_m'\right|^2, \tag{4.3}$$

where $E_m(t)$ is the energy similar to (4.1) associated to the approximated system (3.5) and λ_1 is the first eigenvalue of A. From (4.3), we see that $E_m(t)$ is non-increasing function.

For an arbitrary $\varepsilon > 0$, we define the perturbed energy

$$E_{m\varepsilon}(t) = (1 + \varepsilon c) E_m(t) + \varepsilon F(t), \tag{4.4}$$

with $c > 0$ being a constant to be determined later and

$$F(t) = \left(A^{s-\frac{1}{2}}u_m(t), A^{s-\frac{1}{2}}u_m'(t)\right).$$

Notice that

$$|F(t)| \leq C_1 E_m(t), \tag{4.5}$$

where $C_1 = \max\left\{C_0^2/m_0, c, 1\right\}$ and $C_0 > 0$ is the immersion constant of $D(A^s)$ into $D(A^{s-\frac{1}{2}})$. By (4.4) and (4.5)

$$|E_{m\varepsilon}(t) - (1 + \varepsilon c) E_m(t)| \leq \varepsilon C_1 E_m(t)$$

or

$$[1 + \varepsilon(c - C_1)] E_m(t) \leq E_{m\varepsilon}(t) \leq [1 + \varepsilon(c + C_1)] E_m(t).$$

Taking $0 < \varepsilon < \min\{1/2\,(C_1 - c)\,,\, 1/\,(C_1 + c)\}$, we get

$$\frac{1}{2}E_m\,(t) \leq E_{m\varepsilon}\,(t) \leq 2E_m\,(t)\,. \tag{4.6}$$

Considering the derivative of the function $F(t)$ and using $(3.5)_1$ (with $f = 0$), we obtain

$$\begin{aligned} F'\,(t) &= \left|A^{s-\frac{1}{2}}u_m'\right|^2 + \left(A^{s-\frac{1}{2}}u_m,\, A^{s-\frac{1}{2}}u_m''\right) = \left|A^{s-\frac{1}{2}}u_m'\right|^2 + \left(A^{2s-1}u_m,\, u_m''\right) \\ &= \left|A^{s-\frac{1}{2}}u_m'\right|^2 - \left(A^{2s-1}u_m,\, M\left(|A^s u_m|^2\right)Au_m\right) - \left(A^{s+\frac{\alpha}{2}-\frac{1}{2}}u_m,\, A^{s+\frac{\alpha}{2}-\frac{1}{2}}u_m'\right) \\ &= \left|A^{s-\frac{1}{2}}u_m'\right|^2 - M\left(|A^s u_m|^2\right)|A^s u_m|^2 - \left(A^{s+\frac{\alpha}{2}-\frac{1}{2}}u_m,\, A^{s+\frac{\alpha}{2}-\frac{1}{2}}u_m'\right). \end{aligned} \tag{4.7}$$

By (4.3) and (4.7) one has

$$\begin{aligned} E_\varepsilon'\,(t) + \varepsilon F'\,(t) \leq{}& -\lambda_1^\alpha\left|A^{s-\frac{1}{2}}u_m'\right|^2 + \varepsilon\left|A^{s-\frac{1}{2}}u_m'\right|^2 - \varepsilon M\left(\left|A^s u_m\right|^2\right)\left|A^s u_m\right|^2 \\ & - \varepsilon\left(A^{s+\frac{\alpha}{2}-\frac{1}{2}}u_m,\, A^{s+\frac{\alpha}{2}-\frac{1}{2}}u_m'\right). \end{aligned} \tag{4.8}$$

Notice that

$$\begin{aligned} \left|\left(A^{s+\frac{\alpha}{2}-\frac{1}{2}}u_m,\, A^{s+\frac{\alpha}{2}-\frac{1}{2}}u_m'\right)\right| &\leq \frac{\delta}{2}\left|A^{s+\frac{\alpha}{2}-\frac{1}{2}}u_m'\right|^2 + \frac{1}{2\delta}\left|A^{s+\frac{\alpha}{2}-\frac{1}{2}}u_m\right|^2 \\ &\leq -\frac{\delta}{2}E_m'\,(t) + \frac{1}{2\delta}\left|A^{s+\frac{\alpha}{2}-\frac{1}{2}}u_m\right|^2, \end{aligned} \tag{4.9}$$

with $\delta > 0$ being a constant to be chosen, and

$$\begin{aligned} \left|A^{s+\frac{\alpha}{2}-\frac{1}{2}}u_m\right|^2 &= \sum_{0<\lambda_\nu\leq 1}\lambda_\nu^{2s+\alpha-1}\left|(u_m, w_\nu)\right|^2 + \sum_{\lambda_\nu\geq 1}\lambda_\nu^{2s+\alpha-1}\left|(u_m, w_\nu)\right|^2 \\ &\leq |u_m|^2 + \left|A^s u_m\right|^2 \leq \left(\frac{1+\lambda_1}{m_0}\right)\widehat{M}\left(\left|A^s u_m\right|^2\right). \end{aligned} \tag{4.10}$$

We also have

$$- M\left(\left|A^s u_m\right|^2\right)\left|A^s u_m\right|^2 \leq -\frac{m_0}{\tau}\widehat{M}\left(\left|A^s u_m(t)\right|^2\right), \tag{4.11}$$

where $\tau = \max\left\{M(s);\ 0 \le s \le \frac{2E(0)}{m_0}\right\}$. Combining (4.8)–(4.11), it follows that

$$E_m'(t) + \varepsilon\frac{\delta}{2}E_m'(t) + \varepsilon F'(t) \le -\left(\lambda_1^\alpha - \varepsilon\right)\left|A^{s-\frac{1}{2}}u_m'\right|^2 + \left[\frac{\varepsilon}{2\delta}\left(\frac{1+\lambda_1}{m_0}\right) - \frac{\varepsilon m_0}{\tau}\right]\widehat{M}\left(\left|A^s u_m\right|^2\right). \quad (4.12)$$

Choosing $\delta = \frac{\tau(1+\lambda_1)}{m_0^2}$ and $c = \delta/2$, we obtain by (4.4) and (4.12) that

$$E_{m\varepsilon}'(t) \le -\left(\lambda_1^\alpha - \varepsilon\right)\left|A^{s-\frac{1}{2}}u_m'\right|^2 - \frac{\varepsilon m_0}{2\tau}\widehat{M}\left(\left|A^s u_m\right|^2\right). \quad (4.13)$$

Taking $\delta_0 = \min\left\{2(\lambda_1^\alpha - \varepsilon), \frac{\varepsilon m_0}{\tau}\right\}$, we can conclude by (4.6) and (4.13) that

$$E_{m\varepsilon}'(t) \le -\frac{\delta_0}{C_3}E_{m\varepsilon}(t), \quad \forall t \ge 0,$$

which implies

$$E_{m\varepsilon}(t) \le E_{m\varepsilon}(0)e^{-\frac{\delta_0}{C_3}t}, \quad \forall t \ge 0. \quad (4.14)$$

Combining (4.6) and (4.14) we get

$$E_m(t) \le \frac{C_3}{C_2}E_m(0)e^{-\frac{\delta_0}{C_3}t}, \quad \forall t \ge 0. \quad (4.15)$$

Taking the $\liminf\limits_{m\to\infty}$ in both sides of (4.15) and according the convergences $(3.5)_2$, $(3.5)_3$, (3.8), and (3.9), we deduce the inequality (4.2) and Theorem 4.1 is proved. ■

Acknowledgement The author “F. D. Araruna” is partially supported by INCTMat, CAPES, CNPq (Brazil).

References

1. Aassila, M.: On quasilinear wave equation with strong damping. Funkcial. Ekvac. **41**, 67–78 (1998)
2. Bernstein, S.: Sur une classe d’equations functionelles aux derivées partielles. Isv. Acad. Nauk SSSR, Serv. Math. **4**, 17–26 (1940)
3. Brito, E.H.: The damped elastic stretched string equation generalized: existence, uniqueness, regularity and stability. Appl. Anal. **13**, 219–233 (1982)

4. Carrier, G.F.: On the nonlinear vibration problem of the elastic string. Q. J. Appl. Math. **3**, 157–165 (1945)
5. Coddington, E.A., Levinson, N.: Theory of Ordinary Differential Equations. McGraw-Hill, New York (1987)
6. Cousin, A.T., Frota, C.L., Larkin, N.A., Medeiros, L.A.: On the abstract model of the Kirchhoff-Carrier equation. Commun. Appl. Anal. **1**(3), 389–404 (1997)
7. Cousin, A.T., Frota, C.L., Larkin, N.A.: Existence of global solutions and energy decay for the Carrier equation with dissipative term. Differential Integral Equation **12**(4), 453–469 (1999)
8. Frota, C.L., Goldstein, J.A.: Some nonlinear wave equations with acoustic boundary conditions. J. Differ. Equ. **164**, 92–109 (2000)
9. Kirchhoff, G.: Volersunger über Mechanik. Tauber Leipzig, Leipzig (1883)
10. Larkin, N.A.: Global regular solution for the nonhomogeneous Carrier equation. Math. Probl. Eng. **8**, 15–31 (2002)
11. Lions, J.-L.: On some questions in boundary value problems of mathematical physics. In: de la Penha, G.M., Medeiros, L.A. (eds.) Contemporary Development in Continuons Mechanics and Partial Differential Equations. North-Holland, London (1978)
12. Lions, J.-L., Magenes, E.: Problémes aux limites non homogénes et applications, vol. 1. Dunod, Paris (1968)
13. Matos, M.P., Pereira, D.: On a hyperbolic equation with strong dissipation. Funkcial. Ekvac. **34**, 303–331 (1991)
14. Medeiros, L.A., Milla Miranda, M.: Solutions for the equation of nonlinear vibrations in Sobolev spaces of fractionary order. Math. Appl. Comp. **6**, 257–276 (1987)
15. Medeiros, L.A., Milla Miranda, M.: On a nonlinear wave equation with damping. Rev. Math. Univ. Complu. Madrid **3**, 213–231 (1990)
16. Medeiros, L.A., Limaco, J., Menezes, S.B.: Vibrations of elastic string: mathematical aspects, part 1. J. Comput. Anal. Appl. **4**(2), 91–127 (2002)
17. Medeiros, L.A., Limaco, J., Menezes, S.B.: Vibrations of elastic string: mathematical aspects, part 2. J. Comput. Anal. Appl. **4**(3), 211–263 (2002)
18. Mimouni, S., Benaissa, A., Amroun, N.-E.: Global existence and optimal decay rate of solutions for the degenerate quasilinear wave equation with a strong dissipation. Appl. Anal. **89**(6), 815–831 (2010)
19. Nishihara, K.: Degenerate quasilinear hyperbolic equation with strong damping. Funkcial. Ekvac. **27**(1), 125–145 (1984)
20. Nishihara, K., Yamada, Y.: On global solutions of some degenerate quasi-linear hyperbolic equations with dissipative terms. Funkcial. Ekvac. **33**(1), 151–159 (1990)
21. Ono, K.: On decay properties of solutions for degenerate strongly damped wave equations of Kirchhoff type. J. Math. Anal. Appl. **381**(1), 229–239 (2011)
22. Ono, K.: On sharp decay estimates of solutions for mildly degenerate dissipative wave equations of Kirchhoff type. Math. Methods Appl. Sci. **34**(11), 1339–1352 (2011)
23. Park, J.Y., Bae, J.J., Jung, I.H.: On existence of global solutions for the carrier model with nonlinear damping and source terms. Appl. Anal. **77**(3–4), 305–318 (2001)
24. Pohozhaev, S.I.: On a class of quasilinear hyperbolic equations. Math. USSR Sbornik **25**, 145–158 (1975)
25. Vasconcelos, C.F., Teixeira, L.M.: Strong solution and exponential decay for a nonlinear hyperbolic equation. Appl. Anal. **55**, 155–173 (1993)
26. Yamada, Y.: On some quasilinear wave equations with dissipative terms. Nagoya Math. J. **87**, 17–39 (1982)
27. Zuazua, E.: Stability and decay for a class of nonlinear hyperbolic problems. Asymptot. Anal. **1**, 161–185 (1988)

A Second-Order Linear Newmark Method for Lagrangian Navier-Stokes Equations

Marta Benítez, Alfredo Bermúdez, and Pedro Fontán

Dedicated to Prof. Enrique Fernández-Cara on the occasion of his 60th birthday.

Abstract In this paper we propose a second-order pure Lagrange-Galerkin method for the numerical solution of free surface problems in fluid mechanics. We consider a viscous, incompressible Newtonian fluid in a time dependent domain which may present large deformations but no topological changes at interfaces. Pure-Lagrangian methods are useful for solving these problems because the convective term disappears, the computational domain is independent of time and modelling and tracking of the free surface is straightforward as far as there is no solid walls preventing the free motion of surface particles. Unfortunately, for moderate to high-Reynolds number flows and as a consequence of high distortion of the moved mesh, it can be necessary to re-mesh and re-initialize the motion each certain time. In this paper, a Newmark algorithm is considered for both, the time semi-discretization of equations in Lagrangian coordinates and the computation of initial conditions. The proposed scheme is pure-Lagrangian and can be written in terms of either material velocity and pressure or material acceleration and pressure or material displacement and pressure. The three formulations are stated. In order to assess the performance of the overall numerical method, we solve different problems in two space dimensions.

M. Benítez
Departamento de Matemáticas, Universidade da Coruña, Ferrol, Spain
e-mail: marta.benitez@udc.es

A. Bermúdez (✉)
Departamento de Matemática Aplicada, Universidade de Santiago de Compostela, Santiago de Compostela, Spain
e-mail: alfredo.bermudez@usc.es

P. Fontán
Instituto tecnológico de matemática industrial, Santiago de Compostela, Spain
e-mail: pedro.fontan@usc.es

A. Doubova et al. (eds.), *Recent Advances in PDEs: Analysis, Numerics and Control*, SEMA SIMAI Springer Series 17,
https://doi.org/10.1007/978-3-319-97613-6_3

In particular, numerical results of a dam break problem and a flow past a cylinder are presented.

Keywords free-surface problems · Lagrangian Navier-Stokes equations · second-order schemes · linear Newmark algorithm · Lagrange-Galerkin methods

1 Introduction

The main goal of the present paper is to introduce a new procedure for solving free surface problems based on the Newmark time integration algorithm. For this purpose, a Lagrangian framework is considered. Many problems in fluid mechanics involve free surfaces. In this paper we are interested in solving the Navier-Stokes equations in a time dependent domain which may involve large deformations. Notice that, some problems in engineering and applied sciences involve topological changes such as the breaking and/or merging of the interfaces. Although we are interested in dealing with these problems, for simplicity in this paper they will not be considered. Typically, the Navier-Stokes equations are written in Eulerian coordinates and in terms of the velocity. However, the Eulerian formulation of free surface problems presents two classical difficulties: the treatment of the convective term and the modelling and tracking of the free surface. The first one disappears if the problem is written in Lagrangian coordinates and also the second one at least if there is no walls preventing the free motion of surface particles. The more general case is beyond the scope of the present paper but at present is the subject of research in progress by the authors.

The methods of characteristics are extensively used for solving convection-diffusion problems with dominant convection (see the review paper [1]). These methods are based on time discretization of the time derivative along characteristic curves. When they are referred to a fixed domain (respectively, to a time dependent domain) they are called pure Lagrangian methods (respectively, semi-Lagrangian methods). These methods have been mathematically analyzed and applied to different problems with time independent domains by several authors (see [2–8]). For example, in [3, 4] the classical first-order characteristic method combined with finite elements applied to convection-diffusion equations is studied, and in [5, 6] and [7] second-order Lagrange-Galerkin methods are analyzed. Stability and optimal error estimates are proved.

The Eulerian framework of the classical characteristics methods is unduly cumbersome to solve problems with time dependent domains, such as free-surface flows or fluid-structure interaction problems. These problems have been solved with several Lagrangian approaches. For instance, the particle finite element method (PFEM) has been applied to the solution of fluid-dynamics problems including free surface flows and breaking waves [9], fluid-structure interactions [10, 11] or fluid-object interactions [12, 13]. The Eulerian classical formulation of the Navier-Stokes equations is considered and the classical technique to discretize the material derivative is adopted. However, a Lagrangian approach is used because the track of the locations of individual particles (which can be nodes) is kept and particles in current domain are viewed as moving points from previous domains. The particle positions are updated by using the values of velocity. At each time step, the problem to solve is non-linear because it is written in the current domain which is unknown (it depends on current velocity, unknown too). On the other hand, in [14] the Lagrangian form of the Navier-Stokes equations in terms of the motion is considered. A Newmark's algorithm for time discretization, combined with finite element for space discretization is proposed to solve the Lagrangian problem. At each time step, the obtained problem is non-linear and it is solved by Newton-Raphson iteration. In the present paper, we also consider the Lagrangian form of the Navier-Stokes equations but we propose a new strategy for time discretization so that at each time step the problem to solve is linear. Moreover, the obtained scheme can be written in terms of either material velocity and pressure or material acceleration and pressure or material displacement and pressure.

Recently, we have introduced new characteristics methods combined with finite elements applied to the solution of scalar linear convection-diffusion problems with time dependent domain [15–17], free surface flows [18] and fluid-structure interaction problems [19]. Numerical results showing the performance of these methods are shown. All of them are linear and are obtained by introducing a change of variable in the original problem. In particular, in [15, 16, 19] the Crank-Nicholson time discretization has been used to solve the considered problems in Lagrangian coordinates. Stability and optimal error estimates were proved for scalar linear convection-diffusion problems (see [15–17] for details). Moreover, in [17] and [18] we propose unified formulations to state pure-Lagrangian and semi-Lagrangian methods for solving scalar linear and vector non-linear convection-diffusion equations, respectively. More precisely, a quite general change of variable from the current configuration to an intermediate reference configuration, not necessarily the one of the initial time, is proposed obtaining another new strong formulation of the problem from which classical and new time discretization methods can be introduced in a natural way. In particular, in [18] we use the unified formulation to obtain two new second-order characteristics methods in terms of the displacement for solving the Navier-Stokes equation, one semi-Lagrangian and another one pure Lagrangian. The pure Lagrangian scheme has been used in [19] to solve fluid-structure interaction problems. In the present paper, a pure Lagrangian method as the one proposed in [18] is considered but new formulations in terms of material velocity and pressure, and material acceleration and pressure, are proposed. A more general new technique to obtain the initial conditions, the boundary conditions and the velocity is proposed. It is based on the Newmark algorithm and can be used in a natural way to introduce the initial conditions and the boundary conditions in the three formulations considered in this paper. Moreover, the procedure suggested in [18] for reinitialization is assessed by solving the problem of the flow past a cylinder.

The paper is organized as follows. In Sect. 2 an initial-boundary value problem is posed in a time dependent bounded domain and some hypotheses and notations concerning motion are recalled. In Sect. 3, the strong formulation of the problem is written in Lagrangian coordinates and in terms of the material velocity and pressure, and then the standard associated weak problem is obtained. In Sect. 4, we propose a second-order Newmark algorithm for the time semi-discretization of the Lagrangian problem and the initialization of variables. Three formulations are proposed, the first one is written in terms of material velocity and pressure, the second one in terms of the material acceleration and pressure and the last one in terms of the material displacement and pressure. Section 5 discusses the full discretized velocity/pressure pure Lagrange-Galerkin scheme using a finite element method. Moreover, a method to re-initialize this scheme is proposed. Finally, in Sect. 6 numerical examples are included showing the performance of the overall method.

2 Statement of Problem in Eulerian Coordinates and Notations

Let Ω be a bounded domain in $\mathbb{R}^d$ ($d = 2, 3$) with Lipschitz boundary Γ. Let us assume that Γ is divided into two parts: $\Gamma = \Gamma^D \cup \Gamma^N$, with $\Gamma^D \cap \Gamma^N = \emptyset$. Let $\mathsf{X} : \overline{\Omega} \times \mathbb{R} \longrightarrow \mathbb{R}^d$ be a *motion* in the sense of Gurtin [20]. For given $\mathscr{A} \subset \overline{\Omega}$, we denote $\mathscr{A}_t := \mathsf{X}(\mathscr{A}, t)$ (see Fig. 1). In practice, a bounded time interval is considered for the motion, namely, $[t_0, t_f]$, being t_0, t_f two non-negative numbers. For simplicity, in this paper we assume that $\mathsf{X}(\mathsf{p}, t_0) = \mathsf{p}\ \forall \mathsf{p} \in \overline{\Omega}$. Notice that in many cases the body is at rest until the initial time, i.e., $\mathsf{X}(\mathsf{p}, t) = \mathsf{p}\ \forall t \leq t_0\ \forall \mathsf{p} \in \overline{\Omega}$ and then the initial velocity is null. We will adopt the notation given in [18] for the trajectory of the motion ($\mathscr{T}$), the spatial velocity ($\mathbf{v}$), the material displacement ($\mathbf{u}$) and the deformation gradient ($\mathbf{F}$). We denote by P the reference map, by p the points in $\overline{\Omega}$ and by x the points in $\overline{\Omega}_t$. Moreover, fields defined in $\mathscr{T}$ (respectively, in $\overline{\Omega} \times [t_0, t_f]$) are called *spatial fields* (respectively, *material fields*).

Remark 2.1 For the sake of clarity, in expressions involving space and time derivatives we use the following notation (see, for instance, [20]).

- If Φ is a smooth material field, we denote by $\nabla\Phi$ (respectively, by $\operatorname{Div}\Phi$) the gradient (respectively, the divergence) with respect to the first argument (p), and by $\dot{\Phi}$ the partial derivative with respect to the second argument (t).

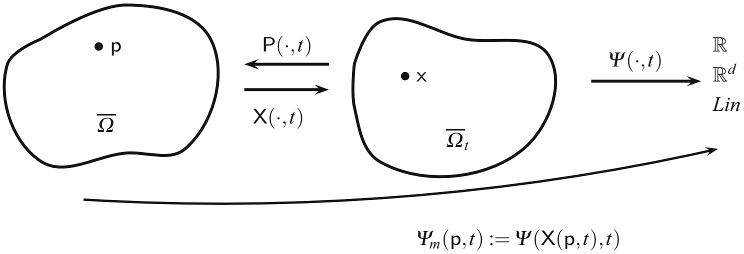


Fig. 1 Motion and material description of spatial fields

- If $\varPsi$ is a smooth spatial field, we denote by $\operatorname{grad}\varPsi$ (respectively, by $\operatorname{div}\varPsi$) the gradient (respectively, the divergence) with respect to the first argument (x), and by $\varPsi'$ the partial derivative with respect to the second argument (t).
- In some places where the above operators appear, we specify the differentiation variable as a subscript, e.g., $\nabla_{\mathsf{p}}\varPhi$, $\operatorname{grad}_{\mathsf{x}}\varPsi$ (respectively, $\operatorname{Div}_{\mathsf{p}}\varPhi$, $\operatorname{div}_{\mathsf{x}}\varPsi$) denote the gradient (respectively, the divergence) with respect to the first argument (p or x).
- If $\varPsi$ is a smooth spatial field, $\dot{\varPsi}$ denotes the *material time derivative* with respect to time, that is

$$\dot{\varPsi}(\mathsf{x},t) = \frac{\partial}{\partial t}\left(\varPsi(\mathsf{X}(\mathsf{p},t),t)\right)_{|\mathsf{p}=\mathsf{P}(\mathsf{x},t)}.$$

If $\varPsi$ is a spatial field, we define its *material description* $\varPsi_m$ by

$$\varPsi_m(\mathsf{p},t) := \varPsi(\mathsf{X}(\mathsf{p},t),t) \quad \forall(\mathsf{p},t) \in \overline{\varOmega} \times [t_0,t_f]. \tag{1}$$

This mapping is depicted in Fig. 1. Let us introduce the material description of the velocity and acceleration, namely,

$$\mathbf{v}_m(\mathsf{p},t) := \dot{\mathsf{X}}(\mathsf{p},t) \quad \forall(\mathsf{p},t) \in \overline{\varOmega} \times [t_0,t_f], \tag{2}$$

$$\mathbf{a}_m(\mathsf{p},t) := \ddot{\mathsf{X}}(\mathsf{p},t) \quad \forall(\mathsf{p},t) \in \overline{\varOmega} \times [t_0,t_f]. \tag{3}$$

Now, let us consider the following initial-boundary value problem (motion equation of a Newtonian fluid):

Eulerian Strong Problem (ESP) *Find two functions* $\mathbf{v} : \mathscr{T} \longrightarrow \mathbb{R}^d$ *and* $\pi : \mathscr{T} \longrightarrow \mathbb{R}$ *such that*

$$\rho\mathbf{v}' + \rho\operatorname{grad}\mathbf{v}\mathbf{v} - \operatorname{div}\left\{-\pi\mathbf{I} + \mu(\operatorname{grad}\mathbf{v} + \operatorname{grad}\mathbf{v}^t)\right\} = \mathbf{b} \quad \textit{in } \mathscr{T}, \tag{4}$$

$$\operatorname{div}\mathbf{v} = g \quad \textit{in } \mathscr{T}, \tag{5}$$

subject to the boundary conditions

$$\mathbf{v}(\cdot,t) = \mathbf{v}_D(\cdot,t) \quad \textit{on } \varGamma_t^D, \tag{6}$$

$$\left(-\pi(\cdot,t)\mathbf{I} + \mu(\cdot,t)(\operatorname{grad}\mathbf{v}(\cdot,t) + \operatorname{grad}\mathbf{v}^t(\cdot,t))\right)\mathbf{n}(\cdot,t) = \mathbf{h}(\cdot,t) \quad \text{on } \varGamma_t^N, \tag{7}$$

for $t \in [t_0,t_f]$, *and the initial condition*

$$\mathbf{v}(\cdot,t_0) = \mathbf{v}^0 \quad \textit{in } \overline{\varOmega}. \tag{8}$$

In the above equations, $\rho : \mathscr{T} \longrightarrow \mathbb{R}$, $\mu : \mathscr{T} \longrightarrow \mathbb{R}$, $\mathbf{b} : \mathscr{T} \longrightarrow \mathbb{R}^d$, $g : \mathscr{T} \longrightarrow \mathbb{R}$, $\mathbf{v}^0 : \overline{\varOmega} \longrightarrow \mathbb{R}^d$, $\mathbf{v}_D(\cdot,t) : \varGamma_t^D \longrightarrow \mathbb{R}^d$ and $\mathbf{h}(\cdot,t) : \varGamma_t^N \longrightarrow \mathbb{R}^d$, $t \in [t_0,t_f]$, are given spatial fields, $\mathbf{I}$ is the identity second order tensor and

$\mathbf{n}(\cdot, t)$ is the outward unit normal vector to Γ_t. Let us notice that for $g = 0$ the above equations are the incompressible Navier-Stokes equations. Otherwise, condition (5) with $g \neq 0$ appears when modelling low-Mach number flows as those arising in many combustion problems. In this case function g is obtained from the mass conservation equation and the state law of the gas mixture as a function of temperature which, in its turn, is computed by solving the energy conservation equation.

In the following $\mathscr{A}$ denotes a bounded domain in $\mathbb{R}^d$. Let us recall the definition of the Hilbert spaces $\mathrm{L}^2(\mathscr{A})$, $\mathrm{H}^1(\mathscr{A})$ and $\mathbf{H}(\mathrm{div}, \mathscr{A})$:

$$\mathrm{L}^2(\mathscr{A}) = \left\{ f : \mathscr{A} \longrightarrow \mathbb{R} \text{ measurable}, \int_{\Omega} f^2 dx < \infty \right\}, \tag{9}$$

$$\mathrm{H}^1(\mathscr{A}) = \left\{ f : \mathscr{A} \longrightarrow \mathbb{R} \text{ measurable}, f, \frac{\partial f}{\partial x_i} \in \mathrm{L}^2(\mathscr{A}),\ i = 1, \ldots, d \right\}, \tag{10}$$

$$\mathbf{H}(\mathrm{div}, \mathscr{A}) = \Big\{ \mathbf{w} \in (\mathrm{L}^2(\mathscr{A}))^d,\ \mathrm{div}\,\mathbf{w} \in \mathrm{L}^2(\mathscr{A}) \}. \tag{11}$$

We also introduce the notation $\mathbf{H}^1(\mathscr{A}) = \left(\mathrm{H}^1(\mathscr{A})\right)^d$ and denote by $\mathbf{H}^1_{\Gamma^P}(\mathscr{A})$ and $\mathbf{H}_{\Gamma^P}(\mathrm{div}, \mathscr{A})$ the subspaces of $\mathbf{H}^1(\mathscr{A})$ and $\mathbf{H}(\mathrm{div}, \mathscr{A})$, respectively, defined by

$$\mathbf{H}^1_{\Gamma^P}(\mathscr{A}) := \left\{ \mathbf{w} \in \mathbf{H}^1(\mathscr{A}),\ \mathbf{w}|_{\Gamma^P} \equiv 0 \right\}, \tag{12}$$

$$\mathbf{H}_{\Gamma^P}(\mathrm{div}, \mathscr{A}) := \left\{ \mathbf{w} \in \mathbf{H}(\mathrm{div}, \mathscr{A}),\ \mathbf{w} \cdot \mathbf{m}|_{\Gamma^P} \equiv 0 \right\}, \tag{13}$$

where Γ^P is a part of the boundary of $\mathscr{A}$ of non-null measure and $\mathbf{m}$ is the outward unit normal vector to Γ^P.

3 Strong Problem and Weak Formulation in Lagrangian Coordinates

We are going to develop some formal computations in order to write the above problem **(ESP)** in Lagrangian coordinates. Firstly, from the definition of the material time derivative and by using the chain rule, we get (see, for instance, [20])

$$\dot{\mathbf{v}}(\mathsf{x}, t) = \mathbf{v}'(\mathsf{x}, t) + \mathrm{grad}_{\mathsf{x}} \mathbf{v}(\mathsf{x}, t) \mathbf{v}(\mathsf{x}, t) = \dot{\mathbf{v}}_m(\mathsf{p}, t)|_{\mathsf{p}=\mathsf{P}(\mathsf{x},t)} \quad \forall (\mathsf{x}, t) \in \mathscr{T}. \tag{14}$$

Then, we use the divergence theorem, the change of variable $\mathsf{x} = \mathsf{X}(\mathsf{p}, t)$, the chain rule and the localization theorem, to obtain the equality

$$\begin{aligned} &- \operatorname{div}_{\mathsf{x}} \{-\pi(\mathsf{x}, t)\mathbf{I} + \mu(\mathsf{x}, t)(\operatorname{grad}_{\mathsf{x}}\mathbf{v}(\mathsf{x}, t) + \operatorname{grad}_{\mathsf{x}}\mathbf{v}^t(\mathsf{x}, t))\} \\ &= -\frac{1}{\det \mathbf{F}(\mathsf{p}, t)} \operatorname{Div}_{\mathsf{p}} \Big\{\Big(-\pi_m(\mathsf{p}, t)\mathbf{I} + \mu_m(\mathsf{p}, t) \Big(\nabla_{\mathsf{p}}\mathbf{v}_m(\mathsf{p}, t)\mathbf{F}^{-1}(\mathsf{p}, t) \\ &\quad +\mathbf{F}^{-t}(\mathsf{p}, t) \left(\nabla_{\mathsf{p}}\mathbf{v}_m\right)^t (\mathsf{p}, t)\Big)\Big) \det \mathbf{F}(\mathsf{p}, t)\mathbf{F}^{-t}(\mathsf{p}, t)\Big\} \Big|_{\mathsf{p}=\mathsf{P}(\mathsf{x},t)}, \end{aligned} \tag{15}$$

for $(\mathsf{x}, t) \in \mathscr{T}$. Next, by using the chain rule we obtain (see, for instance [20])

$$\operatorname{div}_{\mathsf{x}}\mathbf{v}(\mathsf{x}, t) = \nabla_{\mathsf{p}}\mathbf{v}_m(\mathsf{p}, t) \cdot \mathbf{F}^{-t}(\mathsf{p}, t)|_{\mathsf{p}=\mathsf{P}(\mathsf{x},t)} \quad (\mathsf{x}, t) \in \mathscr{T}. \tag{16}$$

Finally, by evaluating Eqs. (6) and (7) at point $\mathsf{x} = \mathsf{X}(\mathsf{p}, t)$ and using (8), we obtain the following material versions of the boundary and initial conditions:

$$\mathbf{v}_m = (\mathbf{v}_D)_m \ \text{ on } \Gamma^D \times [t_0, t_f], \tag{17}$$

$$\begin{aligned} &\left(-\pi_m\mathbf{I} + \mu_m \left(\nabla\mathbf{v}_m\mathbf{F}^{-1} + \mathbf{F}^{-t} \left(\nabla\mathbf{v}_m\right)^t\right)\right) \mathbf{F}^{-t}\mathbf{m} \\ &\qquad = |\mathbf{F}^{-t}\mathbf{m}|\mathbf{h}_m \ \text{ on } \Gamma^N \times [t_0, t_f], \end{aligned} \tag{18}$$

$$\mathbf{v}_m(\cdot, t_0) = \mathbf{v}^0 \ \text{ in } \overline{\Omega}, \tag{19}$$

where $\mathbf{m}$ is the outward unit normal vector to $\partial\Omega$. The second condition has been obtained by using the chain rule and noting that

$$\mathbf{n}(\mathsf{x}, t) = \left.\frac{\mathbf{F}^{-t}(\mathsf{p}, t)\mathbf{m}(\mathsf{p})}{\left|\mathbf{F}^{-t}(\mathsf{p}, t)\mathbf{m}(\mathsf{p})\right|}\right|_{\mathsf{p}=\mathsf{P}(\mathsf{x},t)} \qquad (\mathsf{x}, t) \in \Gamma_t \times [t_0, t_f].$$

As a consequence of these results and by evaluating equations of problem **(ESP)** at point $\mathsf{x} = \mathsf{X}(\mathsf{p}, t)$, we get the following formulation in $\overline{\Omega} \times [t_0, t_f]$:

Lagrangian Strong Problem (LSP) *Find two functions* $\mathbf{v}_m : \overline{\Omega} \times [t_0, t_f] \longrightarrow \mathbb{R}^d$ *and* $\pi_m : \overline{\Omega} \times [t_0, t_f] \longrightarrow \mathbb{R}$ *satisfying*

$$\rho_m\dot{\mathbf{v}}_m - \frac{1}{\det \mathbf{F}} \operatorname{Div} \left\{\left(-\pi_m\mathbf{I} + \mu_m \left(\nabla\mathbf{v}_m\mathbf{F}^{-1} + \mathbf{F}^{-t} \left(\nabla\mathbf{v}_m\right)^t\right)\right) \det \mathbf{F}\mathbf{F}^{-t}\right\} = \mathbf{b}_m, \tag{20}$$

$$\nabla\mathbf{v}_m \cdot \mathbf{F}^{-t} = g_m, \tag{21}$$

in $\Omega \times (t_0, t_f)$*, subjected to boundary conditions (17) and (18), and to initial condition (19).*

Now, we are going to obtain a weak formulation of (**LSP**). For that, we multiply (20) by $\det \mathbf{F}$ and by a test function $\mathbf{z} \in \mathbf{H}^1_{\Gamma^D}(\Omega)$, integrate in Ω, and apply the usual Green's formula and (18). Moreover, we multiply (21) by $\det \mathbf{F}$ and by a test function $q \in \mathrm{L}^2(\Omega)$, and integrate in Ω. The whole problem is the following:

$$\int_{\Omega} \rho_m \det \mathbf{F} \dot{\mathbf{v}}_m \cdot \mathbf{z}\, d\mathsf{p} - \int_{\Omega} \pi_m \det \mathbf{F}\mathbf{F}^{-t} \cdot \nabla \mathbf{z}\, d\mathsf{p}$$

$$+ \int_{\Omega} \mu_m \det \mathbf{F} \left(\nabla \mathbf{v}_m \mathbf{F}^{-1} + \mathbf{F}^{-t} \left(\nabla \mathbf{v}_m\right)^t \right) \mathbf{F}^{-t} \cdot \nabla \mathbf{z}\, d\mathsf{p}$$

$$= \int_{\Omega} \det \mathbf{F} \mathbf{b}_m \cdot \mathbf{z}\, d\mathsf{p} + \int_{\Gamma^N} |\mathbf{F}^{-t}\mathbf{m}| \det \mathbf{F} \mathbf{h}_m \cdot \mathbf{z}\, dA_{\mathsf{p}}, \tag{22}$$

$$\int_{\Omega} \det \mathbf{F} \nabla \mathbf{v}_m \cdot \mathbf{F}^{-t} q\, d\mathsf{p} = \int_{\Omega} \det \mathbf{F} g_m q\, d\mathsf{p}, \tag{23}$$

$\forall \mathbf{z} \in \mathbf{H}^1_{\Gamma^D}(\Omega)$ and $\forall q \in \mathrm{L}^2(\Omega)$. Numerical methods applied to formulations in material coordinates are called pure-Lagrangian methods. Thus, from (22)–(23), we can obtain different pure-Lagrangian numerical methods. These methods are useful, in particular, for solving free surface problems because the computational domain is known and time independent.

Remark 3.1 Notice that we can write Eqs. (22)–(23) in terms of the material displacement or acceleration instead of the material velocity, by replacing $\mathbf{v}_m$ with $\dot{\mathbf{u}}$ or $\dot{\mathbf{v}}_m$ with $\mathbf{a}_m$, respectively. Thus, from (22)–(23) we can obtain pure-Lagrangian methods whose unknowns are either the material velocity and pressure, or the material displacement and pressure, or the material acceleration and pressure. We will call *velocity methods*, *displacement methods* or *acceleration methods* to those written in terms of the velocity, the displacement or the acceleration, respectively. The classical characteristics methods for Navier-Stokes equations are semi-Lagrangian velocity schemes. In the next section, we are going to obtain, from (22)–(23), a second-order pure-Lagrangian method which can be written in terms of the velocity, the displacement or the acceleration.

4 Time Discretization: Linear Newmark Characteristic Method

In this section, we introduce a linear Newmark second-order scheme for time semi-discretization of (22)–(23).

The following notations will be used in the rest of the paper. Let us denote the number of time steps by N, the time step $\Delta t = (t_f - t_0)/N$, and the mesh-points $t_n = t_0 + n\Delta t$. We will use the notation $\varphi^l := \varphi(\cdot, t_l)$ for a material or spatial function φ. Similarly, for a given material or spatial field Φ we will denote by $\Phi^l_{\Delta t}$ an approximation of Φ^l obtained with a time-semidiscretized scheme.

In order to introduce time-semidiscretized schemes in terms of material acceleration or velocity, we use the following Newmark formulas which can be easily deduced by using Taylor expansions:

$$\mathbf{u}^{n+1} = \mathbf{u}^n + \Delta t \mathbf{v}_m^n + \Delta t^2\Big(\beta \mathbf{a}_m^{n+1} + \Big(\frac{1}{2} - \beta\Big)\mathbf{a}_m^n\Big) + O(\Delta t^3)\Big(\frac{1}{6} - \beta\Big) + O(\Delta t^4), \quad (24)$$

$$\mathbf{v}_m^{n+1} = \mathbf{v}_m^n + \Delta t\Big(\gamma \mathbf{a}_m^{n+1} + (1-\gamma)\mathbf{a}_m^n\Big) + O(\Delta t^2)\Big(\frac{1}{2} - \gamma\Big) + O(\Delta t^3). \quad (25)$$

By using the above formulas in (22)–(23), different velocity or acceleration pure-Lagrangian schemes can be obtained. They depend on the values of parameter (γ, β). An optimal accuracy is obtained with the choice $(\gamma, \beta) = (1/2, 1/6)$. However, in this case the criterion of unconditional stability: $2\beta - \gamma \geq 0$ and $\gamma \geq 1/2$, is not satisfied. Optimal accuracy compatible with unconditional stability is obtained with the choice $(\gamma, \beta) = (1/2, 1/4)$. Notice that, both choices lead to non-linear pure-Lagrangian schemes. In this paper we want to analyze the Newmark pure-Lagrangian scheme of high order and linear, which is achieved by taking $(\gamma, \beta) = (1/2, 0)$. More precisely, by evaluating (22)–(23) at time $t = t_{n+1}$ and then using (24) and (25) with $(\gamma, \beta) = (1/2, 0)$, and (17) and (19), we deduce the following time-semidiscretized scheme:

Velocity Pure Lagrangian Scheme (VPL) *Find two sequences of functions* $\widehat{\mathbf{v}}_{m,\Delta t} = \{\mathbf{v}_{m,\Delta t}^{n+1}\}_{n=0}^{N-1}$ *and* $\widehat{\pi}_{m,\Delta t} = \{\pi_{m,\Delta t}^{n+1}\}_{n=0}^{N-1}$ *such that*

$$\int_\Omega \rho^{n+1} \circ \mathsf{X}_{\Delta t}^{n+1} \det \mathbf{F}_{\Delta t}^{n+1} \left(\frac{2}{\Delta t}(\mathbf{v}_{m,\Delta t}^{n+1} - \mathbf{v}_{m,\Delta t}^n) - \mathbf{a}_{m,\Delta t}^n\right) \cdot \mathbf{z}\, d\mathsf{p} - \int_\Omega \pi_{m,\Delta t}^{n+1} \det \mathbf{F}_{\Delta t}^{n+1} (\mathbf{F}_{\Delta t}^{n+1})^{-t} \cdot \nabla \mathbf{z}\, d\mathsf{p}$$

$$+ \int_\Omega \mu^{n+1} \circ \mathsf{X}_{\Delta t}^{n+1} \det \mathbf{F}_{\Delta t}^{n+1} \nabla \mathbf{v}_{m,\Delta t}^{n+1} (\mathbf{F}_{\Delta t}^{n+1})^{-1} (\mathbf{F}_{\Delta t}^{n+1})^{-t} \cdot \nabla \mathbf{z}\, d\mathsf{p}$$

$$+ \int_\Omega \mu^{n+1} \circ \mathsf{X}_{\Delta t}^{n+1} \det \mathbf{F}_{\Delta t}^{n+1} (\mathbf{F}_{\Delta t}^{n+1})^{-t} (\nabla \mathbf{v}_{m,\Delta t}^{n+1})^t (\mathbf{F}_{\Delta t}^{n+1})^{-t} \cdot \nabla \mathbf{z}\, d\mathsf{p} = \int_\Omega \det \mathbf{F}_{\Delta t}^{n+1} \mathbf{b}^{n+1} \circ \mathsf{X}_{\Delta t}^{n+1} \cdot \mathbf{z}\, d\mathsf{p}$$

$$+ \int_{\Gamma^N} |(\mathbf{F}_{\Delta t}^{n+1})^{-t} \mathbf{m}| \det \mathbf{F}_{\Delta t}^{n+1} \mathbf{h} \circ \mathsf{X}_{\Delta t}^{n+1} \cdot \mathbf{z}\, dA_{\mathsf{p}} \quad \forall \mathbf{z} \in \mathbf{H}_{\Gamma^D}^1(\Omega), \quad (26)$$

$$\int_\Omega \det \mathbf{F}_{\Delta t}^{n+1} (\mathbf{F}_{\Delta t}^{n+1})^{-t} \cdot \nabla \mathbf{v}_{m,\Delta t}^{n+1}\, q\, d\mathsf{p} = \int_\Omega g^{n+1} \circ \mathsf{X}_{\Delta t}^{n+1} \det \mathbf{F}_{\Delta t}^{n+1}\, q\, d\mathsf{p} \quad \forall q \in \mathrm{L}^2(\Omega), \quad (27)$$

for $0 \le n \le N-1$, *subject to the initial and boundary conditions*

$$\mathbf{v}_{m,\Delta t}^{n+1} = \mathbf{v}_D^{n+1} \circ \mathsf{X}_{\Delta t}^{n+1} \quad \textit{on } \Gamma^D, \tag{28}$$

$$\mathbf{u}_{\Delta t}^{0} = \mathbf{0} \quad \textit{in } \overline{\Omega}, \tag{29}$$

$$\mathbf{v}_{m,\Delta t}^{0} = \mathbf{v}^{0} \quad \textit{in } \overline{\Omega}, \tag{30}$$

$$\mathbf{a}_{m,\Delta t}^{0} = \mathbf{a}^{0} \quad \textit{in } \overline{\Omega}, \tag{31}$$

and where

$$\mathsf{X}_{\Delta t}^{n+1}(\mathsf{p}) := \mathsf{p} + \mathbf{u}_{\Delta t}^{n+1}(\mathsf{p}), \tag{32}$$

$$\mathbf{F}_{\Delta t}^{n+1} := \mathbf{I} + \nabla \mathbf{u}_{\Delta t}^{n+1}, \tag{33}$$

$$\mathbf{u}_{\Delta t}^{n+1} := \mathbf{u}_{\Delta t}^{n} + \Delta t \mathbf{v}_{m,\Delta t}^{n} + \frac{\Delta t^2}{2} \mathbf{a}_{m,\Delta t}^{n}, \tag{34}$$

$$\mathbf{a}_{m,\Delta t}^{n+1} := \frac{2}{\Delta t}\left(\mathbf{v}_{m,\Delta t}^{n+1} - \mathbf{v}_{m,\Delta t}^{n}\right) - \mathbf{a}_{m,\Delta t}^{n}, \tag{35}$$

for $\mathsf{p} \in \Omega$ *and* $0 \le n \le N-1$.

Remark 4.1 Notice that in the above scheme an initial condition for the acceleration is required. In order to compute it, we evaluate (20) at time $t = t_0$ and differentiate (21) with respect to time variable, to obtain the equations

$$\rho^0 \mathbf{a}^0 + \nabla \pi^0 = \operatorname{Div}\left(\mu^0 \left(\nabla \mathbf{v}^0 + \left(\nabla \mathbf{v}^0\right)^t\right)\right) + \mathbf{b}^0, \tag{36}$$

$$\operatorname{Div} \mathbf{a}^0 = \nabla \mathbf{v}^0 \cdot (\nabla \mathbf{v}^0)^t + (\dot{g})^0, \tag{37}$$

where the unknowns are the initial acceleration and pressure. In the above equations we have used that $\mathbf{F}^0 = \mathbf{I}$ and the following equality

$$(\mathbf{F}^{-t})^{\cdot} = -(\operatorname{grad} \mathbf{v})_m^t \mathbf{F}^{-t} = -\left(\nabla \mathbf{v}_m \mathbf{F}^{-1}\right)^t \mathbf{F}^{-t} = -\mathbf{F}^{-t} \left(\nabla \mathbf{v}_m\right)^t \mathbf{F}^{-t}. \tag{38}$$

Notice that (36)–(37) is a Darcy-like problem (see, for instance, [21]). The *typical* functional setting for this problem is

$$\mathbf{a}^0 \in \mathbf{H}(\text{div}, \Omega), \quad \pi^0 \in \mathrm{L}^2(\Omega).$$

For this setting, we are going to obtain a weak formulation assuming enough regularity of the initial velocity $\mathbf{v}^0$. Let us multiply (36) by a test function $\mathbf{w} \in \mathbf{H}_{\Gamma^D}(\text{div}, \Omega)$, integrate in Ω, apply the usual Green's formula in the pressure term and use (18) at time t_0. Similarly, let us multiply (37) by a test function $q \in \mathrm{L}^2(\Omega)$ and integrate in Ω. The whole *mixed* problem is the following:

$$\int_\Omega \rho^0 \mathbf{a}^0 \cdot \mathbf{w}\, d\mathsf{p} - \int_\Omega \pi^0 \operatorname{Div} \mathbf{w}\, d\mathsf{p} = \int_\Omega \operatorname{Div}\left(\mu^0\left(\nabla \mathbf{v}^0 + \left(\nabla \mathbf{v}^0\right)^t\right)\right) \cdot \mathbf{w}\, d\mathsf{p}$$
$$+ \int_\Omega \mathbf{b}^0 \cdot \mathbf{w}\, d\mathsf{p} - \int_{\Gamma^N} \left(\left(\mu^0\left(\nabla \mathbf{v}^0 + (\nabla \mathbf{v}^0)^t\right)\mathbf{m} - \mathbf{h}^0\right) \cdot \mathbf{m}\right) \mathbf{w} \cdot \mathbf{m}\, dA_{\mathsf{p}}$$
$$\forall \mathbf{w} \in \mathbf{H}_{\Gamma^D}(\mathrm{div}, \Omega), \tag{39}$$

$$\int_\Omega \operatorname{Div} \mathbf{a}^0\, q\, d\mathsf{p} = \int_\Omega \nabla \mathbf{v}^0 \cdot (\nabla \mathbf{v}^0)^t\, q\, d\mathsf{p} + \int_\Omega (\dot{g})^0\, q\, d\mathsf{p} \quad \forall q \in \mathrm{L}^2(\Omega). \tag{40}$$

The normal component of the acceleration on Dirichlet boundary can be obtained by differentiating (17) with respect to time variable, namely

$$\mathbf{a}^0 \cdot \mathbf{m} = (\dot{\mathbf{v}}_D)^0 \cdot \mathbf{m} \quad \text{ on } \Gamma^D. \tag{41}$$

Notice that, velocity Newmark pure-Lagrangian methods can be rewritten in terms of acceleration by using (25). Moreover, for the choice $(\gamma, \beta) = (1/2, 0)$ a formulation in terms of the displacement can also be obtained. Indeed, by using (24) and (25) with $(\gamma, \beta) = (1/2, 0)$, we get

$$\mathbf{v}_{m,\Delta t}^{n+1} = \frac{\mathbf{u}_{\Delta t}^{n+2} - \mathbf{u}_{\Delta t}^{n}}{2\Delta t}, \tag{42}$$

and also

$$\mathbf{a}_{m,\Delta t}^{n+1} = \frac{\mathbf{u}_{\Delta t}^{n+2} - 2\mathbf{u}_{\Delta t}^{n+1} + \mathbf{u}_{\Delta t}^{n}}{\Delta t^2}. \tag{43}$$

Then, the above velocity pure Lagrangian scheme given by (26)–(35), can be rewritten in terms of the material acceleration and pressure or material displacement and pressure. More precisely,

Acceleration Pure Lagrangian Scheme (APL) *Find two sequences of functions* $\widehat{\mathbf{a}}_{m,\Delta t} = \{\mathbf{a}_{m,\Delta t}^{n+1}\}_{n=0}^{N-1}$ *and* $\widehat{\pi}_{m,\Delta t} = \{\pi_{m,\Delta t}^{n+1}\}_{n=0}^{N-1}$ *such that*

$$\int_\Omega \rho^{n+1} \circ \mathsf{X}_{\Delta t}^{n+1} \det \mathbf{F}_{\Delta t}^{n+1} \mathbf{a}_{m,\Delta t}^{n+1} \cdot \mathbf{z}\, d\mathsf{p} - \int_\Omega \pi_{m,\Delta t}^{n+1} \det \mathbf{F}_{\Delta t}^{n+1} (\mathbf{F}_{\Delta t}^{n+1})^{-t} \cdot \nabla \mathbf{z}\, d\mathsf{p}$$
$$+ \int_\Omega \mu^{n+1} \circ \mathsf{X}_{\Delta t}^{n+1} \det \mathbf{F}_{\Delta t}^{n+1} \Big(\frac{\Delta t}{2}(\nabla \mathbf{a}_{m,\Delta t}^{n+1} + \nabla \mathbf{a}_{m,\Delta t}^{n}) + \nabla \mathbf{v}_{m,\Delta t}^{n}\Big)(\mathbf{F}_{\Delta t}^{n+1})^{-1}$$
$$(\mathbf{F}_{\Delta t}^{n+1})^{-t} \cdot \nabla \mathbf{z}\, d\mathsf{p} + \int_\Omega \mu^{n+1} \circ \mathsf{X}_{\Delta t}^{n+1} \det \mathbf{F}_{\Delta t}^{n+1} (\mathbf{F}_{\Delta t}^{n+1})^{-t}$$

$$\begin{aligned}
&\left(\frac{\varDelta t}{2}(\nabla \mathbf{a}^{n+1}_{m,\varDelta t} + \nabla \mathbf{a}^{n}_{m,\varDelta t}) + \nabla \mathbf{v}^{n}_{m,\varDelta t}\right)^{t} (\mathbf{F}^{n+1}_{\varDelta t})^{-t} \cdot \nabla \mathbf{z}\, d\mathsf{p} \\
&\qquad = \int_{\varOmega} \det \mathbf{F}^{n+1}_{\varDelta t} \mathbf{b}^{n+1} \circ \mathsf{X}^{n+1}_{\varDelta t} \cdot \mathbf{z}\, d\mathsf{p} \\
&+ \int_{\varGamma^N} |(\mathbf{F}^{n+1}_{\varDelta t})^{-t}\mathbf{m}| \det \mathbf{F}^{n+1}_{\varDelta t} \mathbf{h} \circ \mathsf{X}^{n+1}_{\varDelta t} \cdot \mathbf{z}\, dA_{\mathsf{p}} \quad \forall \mathbf{z} \in \mathbf{H}^1_{\varGamma^D}(\varOmega),
\end{aligned} \tag{44}$$

$$\begin{aligned}
&\int_{\varOmega} \det \mathbf{F}^{n+1}_{\varDelta t} (\mathbf{F}^{n+1}_{\varDelta t})^{-t} \cdot \left(\frac{\varDelta t}{2}(\nabla \mathbf{a}^{n+1}_{m,\varDelta t} + \nabla \mathbf{a}^{n}_{m,\varDelta t}) + \nabla \mathbf{v}^{n}_{m,\varDelta t}\right) q\, d\mathsf{p} \\
&\qquad = \int_{\varOmega} g^{n+1} \circ \mathsf{X}^{n+1}_{\varDelta t} \det \mathbf{F}^{n+1}_{\varDelta t}\, q\, d\mathsf{p} \quad \forall q \in \mathrm{L}^2(\varOmega),
\end{aligned} \tag{45}$$

for $0 \le n \le N-1$, *subjected to the initial conditions* (29), (30) *and* (31), *and to the boundary condition*

$$\mathbf{a}^{n+1}_{m,\varDelta t} = \frac{2}{\varDelta t}\left(\mathbf{v}^{n+1}_{D} \circ \mathsf{X}^{n+1}_{\varDelta t} - \mathbf{v}^{n}_{D} \circ \mathsf{X}^{n}_{\varDelta t}\right) - \mathbf{a}^{n}_{m,\varDelta t} \quad \text{on } \varGamma^D, \tag{46}$$

and where $\mathsf{X}^{n+1}_{\varDelta t}$, $\mathbf{F}^{n+1}_{\varDelta t}$, $\mathbf{u}^{n+1}_{\varDelta t}$ *are updated by* (32), (33) *and* (34), *respectively and*

$$\mathbf{v}^{n+1}_{m,\varDelta t} := \mathbf{v}^{n}_{m,\varDelta t} + \frac{\varDelta t}{2}\left(\mathbf{a}^{n+1}_{m,\varDelta t} + \mathbf{a}^{n}_{m,\varDelta t}\right), \tag{47}$$

for $0 \le n \le N-1$.

Displacement Pure Lagrangian Scheme (DPL) *Find two sequences of functions* $\widehat{\mathbf{u}}_{\varDelta t} = \{\mathbf{u}^{n+2}_{\varDelta t}\}_{n=0}^{N-2}$ *and* $\widehat{\pi}_{m,\varDelta t} = \{\pi^{n+1}_{m,\varDelta t}\}_{n=0}^{N-1}$ *such that*

$$\begin{aligned}
&\int_{\varOmega} \rho^{n+1} \circ \mathsf{X}^{n+1}_{\varDelta t} \det \mathbf{F}^{n+1}_{\varDelta t} \frac{\mathbf{u}^{n+2}_{\varDelta t} - 2\mathbf{u}^{n+1}_{\varDelta t} + \mathbf{u}^{n}_{\varDelta t}}{\varDelta t^2} \cdot \mathbf{z}\, d\mathsf{p} - \int_{\varOmega} \pi^{n+1}_{m,\varDelta t} \det \mathbf{F}^{n+1}_{\varDelta t} \\
&(\mathbf{F}^{n+1}_{\varDelta t})^{-t} \cdot \nabla \mathbf{z}\, d\mathsf{p} + \int_{\varOmega} \mu^{n+1} \circ \mathsf{X}^{n+1}_{\varDelta t} \det \mathbf{F}^{n+1}_{\varDelta t} \frac{\nabla \mathbf{u}^{n+2}_{\varDelta t} - \nabla \mathbf{u}^{n}_{\varDelta t}}{2\varDelta t} (\mathbf{F}^{n+1}_{\varDelta t})^{-1} \\
&(\mathbf{F}^{n+1}_{\varDelta t})^{-t} \cdot \nabla \mathbf{z}\, d\mathsf{p} + \int_{\varOmega} \mu^{n+1} \circ \mathsf{X}^{n+1}_{\varDelta t} \det \mathbf{F}^{n+1}_{\varDelta t} (\mathbf{F}^{n+1}_{\varDelta t})^{-t} \left(\frac{\nabla \mathbf{u}^{n+2}_{\varDelta t} - \nabla \mathbf{u}^{n}_{\varDelta t}}{2\varDelta t}\right)^{t} \\
&\qquad (\mathbf{F}^{n+1}_{\varDelta t})^{-t} \cdot \nabla \mathbf{z}\, d\mathsf{p} = \int_{\varOmega} \det \mathbf{F}^{n+1}_{\varDelta t} \mathbf{b}^{n+1} \circ \mathsf{X}^{n+1}_{\varDelta t} \cdot \mathbf{z}\, d\mathsf{p} \\
&+ \int_{\varGamma^N} |(\mathbf{F}^{n+1}_{\varDelta t})^{-t}\mathbf{m}| \det \mathbf{F}^{n+1}_{\varDelta t} \mathbf{h} \circ \mathsf{X}^{n+1}_{\varDelta t} \cdot \mathbf{z}\, dA_{\mathsf{p}} \quad \forall \mathbf{z} \in \mathbf{H}^1_{\varGamma^D}(\varOmega),
\end{aligned} \tag{48}$$

$$\begin{aligned}
&\int_{\varOmega} \det \mathbf{F}^{n+1}_{\varDelta t} (\mathbf{F}^{n+1}_{\varDelta t})^{-t} \cdot \frac{\nabla \mathbf{u}^{n+2}_{\varDelta t} - \nabla \mathbf{u}^{n}_{\varDelta t}}{2\varDelta t}\, q\, d\mathsf{p} \\
&\qquad = \int_{\varOmega} g^{n+1} \circ \mathsf{X}^{n+1}_{\varDelta t} \det \mathbf{F}^{n+1}_{\varDelta t}\, q\, d\mathsf{p} \quad \forall q \in \mathrm{L}^2(\varOmega),
\end{aligned} \tag{49}$$

for $0 \le n \le N-2$, *subjected to the initial conditions* (29) *and*

$$\mathbf{u}^1_{\Delta t} := \Delta t \mathbf{v}^0 + \frac{\Delta t^2}{2}\mathbf{a}^0, \tag{50}$$

and to the boundary condition

$$\mathbf{u}^{n+2}_{\Delta t} = 2\Delta t \mathbf{v}^{n+1}_D \circ \mathsf{X}^{n+1}_{\Delta t} + \mathbf{u}^n_{\Delta t} \quad \text{on } \Gamma^D, \tag{51}$$

and where $\mathsf{X}^{n+1}_{\Delta t}$, $\mathbf{F}^{n+1}_{\Delta t}$, $\mathbf{v}^{n+1}_{\Delta t}$ *and* $\mathbf{a}^{n+1}_{m,\Delta t}$ *are updated by* (32), (33), (42) *and* (43) *respectively, for* $0 \le n \le N-2$.

Let us emphasize that while the continuous problem is non-linear, the methods **(VPL)**, **(APL)** and **(DPL)** are linear in their two unknowns, $\mathbf{v}^{n+1}_{m,\Delta t}$ and $\pi^{n+1}_{m,\Delta t}$, $\mathbf{a}^{n+1}_{m,\Delta t}$ and $\pi^{n+1}_{m,\Delta t}$, and $\mathbf{u}^{n+2}_{\Delta t}$ and $\pi^{n+1}_{m,\Delta t}$, respectively. Moreover, we notice that the three methods are exactly equivalent.

Remark 4.2 In [18] two second-order displacement methods are considered for solving Navier-Stokes equations, one semi-Lagrangian and the other one pure Lagrangian. More precisely, the Newmark algorithm with $(\gamma, \beta) = (1/2, 0)$ is considered for time semi-discretization. Then, a problem as (48)–(49) is solved at each time step for both, the semi-Lagrangian and the pure Lagrangian methods. However, other techniques are used to obtain the initial conditions, the boundary conditions, the velocity and the scheme. We notice that in this paper they are obtained in a natural way from Newmark algorithm. This approach is more general because could be applied for any choice of parameters (γ, β). However the procedure given in [18] is only available for the choice $(\gamma, \beta) = (1/2, 0)$. Let us notice that for the method (48)–(49) to be second-order in time not only for displacement but for the velocity as well, it is necessary to start with a third order approximation of displacement as (50). In [18] this is done by writing the problem in configuration $\Omega_{t_0-\Delta t/2}$ and using a third order centered formula to approximate $\mathbf{u}^{1/2}$.

5 Space Discretization: Finite Element Method

In this section, we propose a space discretization of the time semi-discretized problem (26)–(35) by using finite elements. In what follows, for a given material or spatial field Φ we will denote by $\Phi^l_{\Delta t,h}$ an approximation of Φ^l obtained with a fully discretized scheme.

Let us suppose Ω is a bounded domain in $\mathbb{R}^d$ with a Lipschitz polygonal boundary. Let us consider a suitable family of regular triangulations of $\overline{\Omega}$ to be denoted by $\mathfrak{T}_h$ consisting of elements T of diameter $\le h$. Moreover, we assume it is compatible with the partition of the boundary into Γ^D and Γ^N. We propose a space discretization of problem (26)–(35) by using finite element spaces $\mathbf{X}^k_h$ for the material velocity and V^k_h for the material pressure, where the positive integer k

is the "approximation degree" in the following sense. There exist two interpolation operators $\Upsilon_h : (C^0(\overline{\Omega}))^d \longrightarrow \mathbf{X}_h^k$ and $\zeta_h : C^0(\overline{\Omega}) \longrightarrow V_h^k$ satisfying

$$||\Upsilon_h \Phi - \Phi||_s \leq Q_1 h^{r-s} ||\Phi||_r \quad \forall \Phi \in (C^0(\overline{\Omega}))^d \cap \mathbf{H}^r(\Omega), \tag{52}$$

$$||\zeta_h \phi - \phi||_s \leq Q_2 h^{r-s} ||\phi||_r \quad \forall \phi \in C^0(\overline{\Omega}) \cap \mathrm{H}^r(\Omega), \tag{53}$$

where $0 \leq r \leq k+1$ and $s = 0, 1$, and being Q_1 and Q_2 two positive constants independent of h, $\mathbf{H}^0(\Omega) := (\mathrm{L}^2(\Omega))^d$, $\mathrm{H}^0(\Omega) := \mathrm{L}^2(\Omega)$ and $||\cdot||_m$ the usual norm in $\mathbf{H}^m(\Omega)$ and $H^m(\Omega)$. Let us define the following space:

$$\mathbf{X}_{0h}^k = \left\{ \mathbf{z}_h \in \mathbf{X}_h^k : \; \mathbf{z}_h = \mathbf{0} \text{ on } \Gamma^D \right\}. \tag{54}$$

In order to obtain a fully discrete scheme of the time semi-discretized problem (26)–(35) which will be denoted by (**VPLG**), we use spaces $\mathbf{X}_h^k$ and V_h^k to approximate function spaces for material velocity and pressure, respectively. In particular, we replace in (26) (respectively, (27)) the functional space $\mathbf{H}^1_{\Gamma^D}(\Omega)$ (respectively, $\mathrm{L}^2(\Omega)$) with $\mathbf{X}_{0h}^k$ (respectively, V_h^k), and consider the following initial and boundary conditions

$$\mathbf{v}_{m,\Delta t,h}^{n+1}(\mathsf{p}) = \mathbf{v}_D^{n+1} \circ \mathsf{X}_{\Delta t,h}^{n+1}(\mathsf{p}) \quad \text{for all node } \mathsf{p} \text{ on } \Gamma^D, \tag{55}$$

$$\mathbf{u}_{\Delta t,h}^0 = \mathbf{0} \quad \text{in } \overline{\Omega}, \tag{56}$$

$$\mathbf{v}_{m,\Delta t,h}^0 = \mathbf{v}^0 \quad \text{in } \overline{\Omega}, \tag{57}$$

$$\mathbf{a}_{m,\Delta t,h}^0 = \mathbf{a}_h^0 \quad \text{in } \overline{\Omega}, \tag{58}$$

where the initial acceleration $\mathbf{a}_h^0$ is obtained by solving the problem (39)–(41) which is discretized by using stable combinations of finite elements spaces for acceleration and pressure. As an example, the combination of first-order Raviart-Thomas finite element space for acceleration introduced in [22] with piecewise constant functions for pressure leads to stable approximations for Darcy's problem (see [23] for details).

In the fully discrete scheme, the approximations of the displacement ($\mathbf{u}_{\Delta t,h}^{n+1}$), motion ($\mathsf{X}_{\Delta t,h}^{n+1}$), and material acceleration ($\mathbf{a}_{m,\Delta t,h}^{n+1}$) at times $\{t_{n+1}\}_{n=0}^{N-1}$ are given, as in the time-semidiscretized scheme, by Eqs. (34), (32) and (35), respectively, but using the approximations of fully discrete scheme. By using these approximations, we can obtain approximations of the spatial description of the velocity, the pressure and acceleration at times $\{t_{n+1}\}_{n=0}^{N-1}$. These approximations will be considered as piecewise linear functions on the moved mesh and will be denoted by $\mathbf{v}_{\Delta t,h}^{n+1}$, $\pi_{\Delta t,h}^{n+1}$ and $\mathbf{a}_{\Delta t,h}^{n+1}$, respectively. More precisely, we will denote by $\{\mathsf{p}_i^h\}_{i=1}^{N^h}$ the vertices of mesh $\mathfrak{T}_h$ and by $\widetilde{\mathfrak{T}}_h^l$ the moved mesh at time t_l, being $\{\mathsf{X}_{\Delta t,h}^l(\mathsf{p}_i^h)\}_{i=1}^{N^h}$ the vertices of this mesh. Then, the values of $\mathbf{v}^{n+1}_{\Delta t,h}$, $\pi^{n+1}_{\Delta t,h}$ and $\mathbf{a}^{n+1}_{\Delta t,h}$ at vertices $\{\mathsf{X}^{n+1}_{\Delta t,h}(\mathsf{p}^h_i)\}_{i=1}^{N^h}$ are obtained as follows: firstly, we notice that

$$\psi^{n+1}(\mathsf{X}^{n+1}_{\Delta t,h}(\mathsf{p}^h_i)) \simeq \psi^{n+1}(\mathsf{X}^{n+1}(\mathsf{p}^h_i)) = \psi^{n+1}_m(\mathsf{p}^h_i),$$

for a given spatial field ψ, and then we use the approximations of the material descriptions of the velocity, pressure and acceleration: $\mathbf{v}^{n+1}_{m,\Delta t,h}$, $\pi^{n+1}_{m,\Delta t,h}$ and $\mathbf{a}^{n+1}_{m,\Delta t,h}$. More precisely,

- **Approximate velocity in spatial coordinates.** The values of $\mathbf{v}^{n+1}_{\Delta t,h}$ at vertices $\{\mathsf{X}^{n+1}_{\Delta t,h}(\mathsf{p}^h_i)\}_{i=1}^{N^h}$ are computed by

$$\mathbf{v}^{n+1}_{\Delta t,h}(\mathsf{X}^{n+1}_{\Delta t,h}(\mathsf{p}^h_i)) := \mathbf{v}^{n+1}_{m,\Delta t,h}(\mathsf{p}^h_i), \tag{59}$$

  for $0 \le n \le N-1$. Notice that $\bigcup_{T\in\tilde{\mathfrak{T}}^{n+1}_h} T \sim \overline{\Omega}_{t_{n+1}}$.
- **Approximate pressure in spatial coordinates.** The values of the approximate pressure at vertices $\{\mathsf{X}^{n+1}_{\Delta t,h}(\mathsf{p}^h_i)\}_{i=1}^{N^h}$ are computed by

$$\pi^{n+1}_{\Delta t,h}(\mathsf{X}^{n+1}_{\Delta t,h}(\mathsf{p}^h_i)) := \pi^{n+1}_{m,\Delta t,h}(\mathsf{p}^h_i), \tag{60}$$

  for $0 \le n \le N-1$.
- **Approximate acceleration in spatial coordinates.** The values of the approximate acceleration at vertices $\{\mathsf{X}^{n+1}_{\Delta t,h}(\mathsf{p}^h_i)\}_{i=1}^{N^h}$ are computed by

$$\mathbf{a}^{n+1}_{\Delta t,h}(\mathsf{X}^{n+1}_{\Delta t,h}(\mathsf{p}^h_i)) := \mathbf{a}^{n+1}_{m,\Delta t,h}(\mathsf{p}^h_i), \tag{61}$$

  for $0 \le n \le N-1$.

Remark 5.1 Notice that for pure-Lagrangian schemes, the computational domain is the same for all time steps. However, in order to calculate the velocity or the pressure in Eulerian coordinates the moved mesh has to be used. For real fluid mechanics problems, this mesh may have large deformations. When this happens it is necessary to remesh and reinitialize the motion. Next, we propose a method to do this for the pure Lagrange-Galerkin scheme that preserves the order of convergence. Let us assume that we have decided to reinitialize the problem at time t_r, $1 \le r \le N-1$, then the numerical solution at times $t_{n+1} > t_r$ is obtained by using an analogous scheme to (**VPLG**) where the initial time is t_r and the new reference domain is $\overline{\Omega}_{t_r}$. This new scheme will be denoted by (**VPLG**)$_r$. In general, domain and initial conditions at time t_r are unknown, but they are approximated by using the approximate solutions. More precisely, the proposed reinitialization algorithm consists of the following steps.

1. Compute the solution of problem (**VPLG**) for $n = r - 2$ and obtain $\mathbf{X}^r_{\varDelta t,h}$ by using $\mathbf{v}^{r-1}_{m,\varDelta t,h}$, namely

$$\mathsf{X}^r_{\varDelta t,h}(\mathsf{p}) = \mathsf{p} + \mathbf{u}^r_{\varDelta t,h}(\mathsf{p}), \tag{62}$$

   where

$$\mathbf{u}^r_{\varDelta t,h} = \mathbf{u}^{r-1}_{\varDelta t,h} + \varDelta t\mathbf{v}^{r-1}_{m,\varDelta t,h} + \frac{\varDelta t^2}{2}\mathbf{a}^{r-1}_{m,\varDelta t,h}. \tag{63}$$

2. Obtain an approximation of domain $\overline{\varOmega}_{t_r}$ by using $\mathsf{X}^r_{\varDelta t,h}$, namely

$$\overline{\varOmega}_{t_r} \sim \overline{\widetilde{\varOmega}}_{t_r} := \bigcup_{K \in \widetilde{\mathfrak{T}}^r_h} K,$$

   being $\{\mathsf{X}^r_{\varDelta t,h}(\mathsf{p}^h_i)\}_{i=1}^{N^h}$ the vertices of mesh $\widetilde{\mathfrak{T}}^r_h$.
3. Generate a new mesh of domain $\overline{\widetilde{\varOmega}}_{t_r}$. Notice that $\widetilde{\mathfrak{T}}^r_h$ is a mesh of this domain, but in general a new mesh must be considered in order not to have meshes with highly distorted elements.
4. Compute the solution of problem (**VPLG**) for $n = r - 1$ and obtain $\mathbf{v}^r_{\varDelta t,h}$ and $\mathbf{a}^r_{\varDelta t,h}$ from (59) and (61), respectively.
5. Obtain the initial conditions for the scheme $(\mathbf{VPLG})_r$ by using $\mathbf{v}^r_{\varDelta t,h}$ and $\mathbf{a}^r_{\varDelta t,h}$. More precisely,

$$\mathbf{u}^r_{r,\varDelta t,h} = \mathbf{0} \quad \text{in } \overline{\widetilde{\varOmega}}_{t_r}, \tag{64}$$

$$\mathbf{v}^r_{r,\varDelta t,h} = \mathbf{v}^r_{\varDelta t,h} \quad \text{in } \overline{\widetilde{\varOmega}}_{t_r}, \tag{65}$$

$$\mathbf{a}^r_{r,\varDelta t,h} = \mathbf{a}^r_{\varDelta t,h} \quad \text{in } \overline{\widetilde{\varOmega}}_{t_r}, \tag{66}$$

   where the fields with a subscript r are relative to the configuration $\overline{\widetilde{\varOmega}}_{t_r}$.
6. Solve the problem $(\mathbf{VPLG})_r$ and obtain approximations of the velocity, pressure, motion, displacement and acceleration relative to the configuration $\overline{\widetilde{\varOmega}}_{t_r}$ at time instants $t_{n+1} > t_r$.
7. By analogous procedures to the ones in this section, we obtain approximations of the spatial description of the velocity, the pressure and acceleration at time instants $t_{n+1} > t_r$ by using the solution of problem $(\mathbf{VPLG})_r$.

6 Numerical Results

In order to assess the performance of the numerical method introduced in this article, we solve three test problems in two space dimensions. The first one is an academic test example to check the order of the method. The second one is a free boundary problem: the so-called dam break problem. The third one is the standard flow past a cylinder. While the first two examples have been solved with only one mesh for the whole time interval, the latter needed remeshing and reinitializing each certain time in order to avoid the large distortion of the mesh that could lead to non-accurate approximations. We have chosen for space discretization of problems first-order finite element spaces, that is, $k = 1$.

Moreover, in Examples 1 and 2, we have also numerically checked the mass conservation of the scheme **(VPLG)**. In theses examples $\det \mathbf{F} = 1$ and therefore the area of the domain is conserved along the time ($area(\Omega_t) = area(\Omega)\ \forall t$). Then we calculate the l^∞ area error, i.e., we compute the error

$$\max_n \left| area(\widetilde{\Omega}_{t_{n+1}}) - area(\Omega) \right|,$$

where domains $\{\widetilde{\Omega}_{t_{n+1}}\}_n$ are calculated by using the approximate motion $\{\mathsf{X}^{n+1}_{\Delta t,h}\}_n$.

Example 1 This is an example aiming to check the rates of convergence of the scheme proposed in this paper. The spatial domain is $\Omega = (0, 1) \times (0, 1)$, $t_0 = 0$ and $t_f = 2$. The dynamic viscosity is $\mu = 0.1$ and $\rho = 1$. Functions $\mathbf{b}$ and g and Dirichlet boundary and initial conditions are taken such that the exact solution is

$$\begin{aligned} \pi(x, y, t) &= e^t \sin(x - 0.01e^t - 1), \\ v_1(x, y, t) &= 0.01e^t, \\ v_2(x, y, t) &= 0.01e^t \cos(x - 0.01e^t - 1). \end{aligned}$$

The problem has been solved by using the method **(VPLG)**. We denote by $\mathrm{L}^2_h(\Omega)$ the function space endowed with the approximation of the theoretical norm of $\mathrm{L}^2(\Omega)$ using quadrature formulas of degree 2. Furthermore, we introduce the notation $\mathbf{L}^2_h(\Omega) = (\mathrm{L}^2_h(\Omega))^2$ and denote by $l^\infty(\mathscr{A})$ the space of sequences in $\mathscr{A}$ equipped with the norm $||\widehat{\Psi}||_{\mathscr{A}} := \max_n ||\Psi^n||_{\mathscr{A}}$ being $\mathscr{A} = \mathrm{L}^2_h(\Omega), \mathbf{L}^2_h(\Omega)$. We calculate the $l^\infty(\mathbf{L}^2_h(\Omega))$ velocity error and $l^\infty(\mathrm{L}^2_h(\Omega))$ pressure error. In Fig. 2, we have fixed a uniform spatial mesh of 201 × 201 vertices and shown the velocity and pressure errors versus the number of time steps. These results show second-order accuracy in time for both velocity and pressure. In Fig. 3, we represent the velocity and pressure errors versus $1/h$ for a fixed small time step ($\Delta t = 1/2050$). Again, we can observe second-order accuracy in space for velocity and pressure. For this example we have also numerically observed that the scheme **(VPLG)** conserves the area: the area error at the final time $t_f = 4$ and the l^∞ area error, are $1.11 \cdot 10^{-16}$ and $2.22 \cdot 10^{-16}$, respectively, for a spatial mesh of 25 vertices and $\Delta t = 1$.

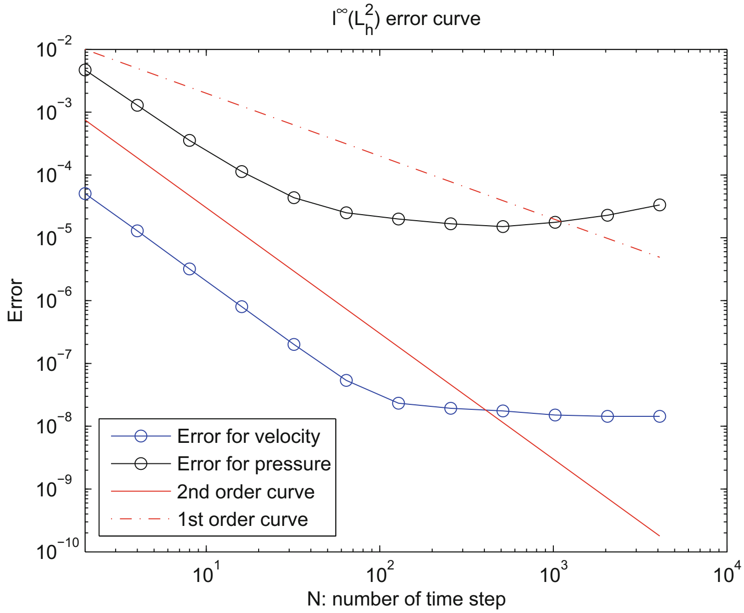


Fig. 2 Example 1: computed $l^\infty(L_h^2)$ errors, in log-log scale, versus the number of time steps for a fixed spatial mesh of 201 × 201 vertices

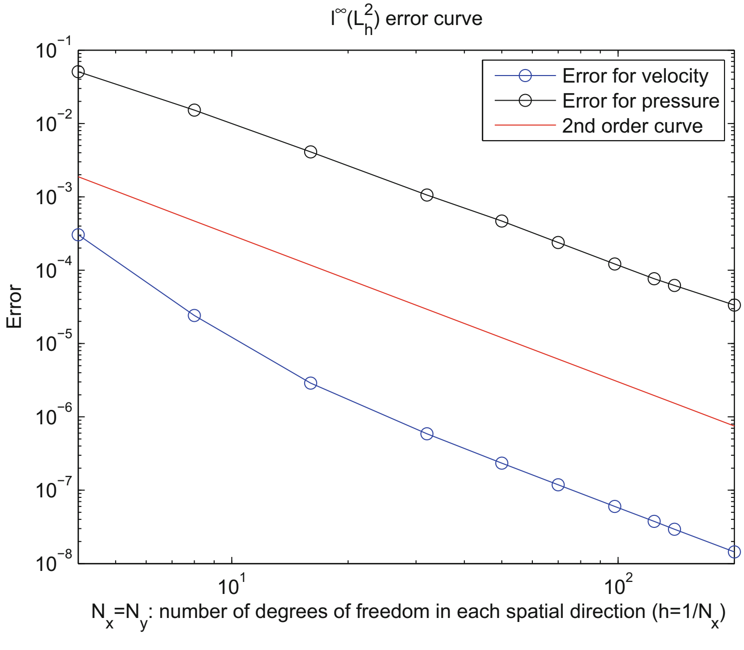


Fig. 3 Example 1: computed $l^\infty(L_h^2)$ errors, in log-log scale, versus $1/h$ for $\Delta t = 1/2050$

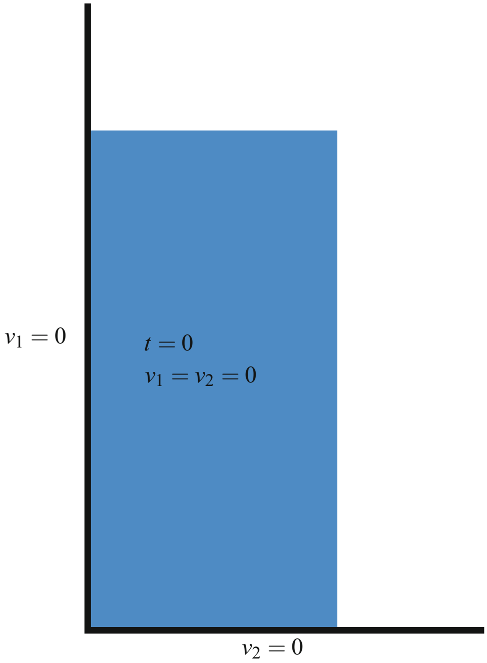


Fig. 4 Example 2: initial configuration and boundary conditions

Example 2 In this example, we consider the collapse of a water column also called the dam-break problem. This problem has long been used as a test case for free surface problems solvers because it has simple boundary conditions and initial configuration; they are shown in Fig. 4. More precisely, we impose a slip boundary condition at the lower horizontal and left vertical boundary and null Neumann condition (force-free) at the upper horizontal and right vertical boundary. The width of the water column is $L = 3.5$ cm and the height is $H = 7$ cm. Gravity is acting downwards with 980 cm/s^2. Notice that CGS units are considered and then $\rho = 1$ g/cm^3 and $\mu = 0.01$ cm^2/s. We have solved this problem by using the velocity pure Lagrange-Galerkin scheme **(VPLG)** without any reinitialization, and with a spatial mesh of 20 × 20 vertices and $\Delta t = 0.01$. In Fig. 5 we represent the time history of the horizontal position of the lower right corner. We compare our results with the numerical ones given in [24–27] and with the experimental data given in [28, 29]. Two groups of curves can be distinguished: those close to the Hirt and Nichols [28] experimental values and those close to the Martin and Moyce [29] ones.

Our results are in good agreement with those given in [28]. In Fig. 6 we represent three instantaneous configurations of the domain and the pressure.

Moreover, for this example we have also numerically checked that the scheme **(VPLG)** conserves the volume with second-order of accuracy. In Fig. 7 we represent the l^∞ area errors versus the number of time steps for a fixed uniform spatial mesh of 368 vertices. For this example, the time discretization error is dominant in the total error.

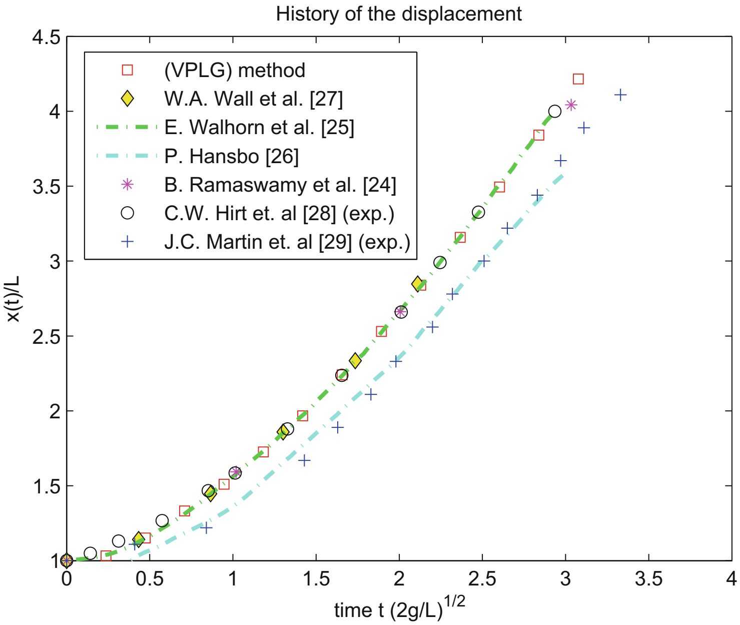


Fig. 5 Example 2: time history of the position of the lower right corner. Solution obtained using a spatial mesh of 20 × 20 vertices and $\Delta t = 0.01$

Example 3 In this example we consider the flow past an infinite cylinder. The cylinder is modeled as a circle and a rectangular domain is considered around the cylinder. Flow past a cylinder is a fundamental fluid mechanics test problem of practical importance. The flow field is symmetric at low values of Reynolds number. As the Reynolds number increases, typically above a value of about 90, flow begins to separate behind the cylinder causing vortex shedding which is an unsteady phenomenon. Incompressible fluid is considered. The width of the rectangular domain is 20 m, the depth is 0.5 m and the radius and the center of the cylinder are, respectively, 0.1 m and (0, 0). Moreover $\rho = 1\,\mathrm{kg/m^3}$. We consider velocity (1, 0) at the horizontal boundaries and periodic condition at the vertical ones. Then, we have the same fluid and the same domain for all time because the fluid going out of the rectangular domain through the right vertical boundary is entering through the left vertical boundary. We solve this problem for different viscosity coefficients ($\mu = 1$, $\mu = 0.01$, $\mu = 0.001$) by using the displacement pure Lagrange-Galerkin method but remeshing and reinitializing the transformation to the identity at certain time instants. The procedure proposed in [18] is used for initialization and reinitialization of the numerical scheme. This method will be denoted by **(LG)**. For $\mu = 1$ and $\mu = 0.01$ a symmetric steady solution is obtained, however for $\mu = 0.001$ eddies are shed continuously from each side of the cylinder, forming rows of vortices in its wake. In Fig. 8 we represent, the numerical solution streamlines for $\mu = 1$ and $\mu = 0.01$. In Fig. 9 we show the streamlines at four times for $\mu = 0.001$.

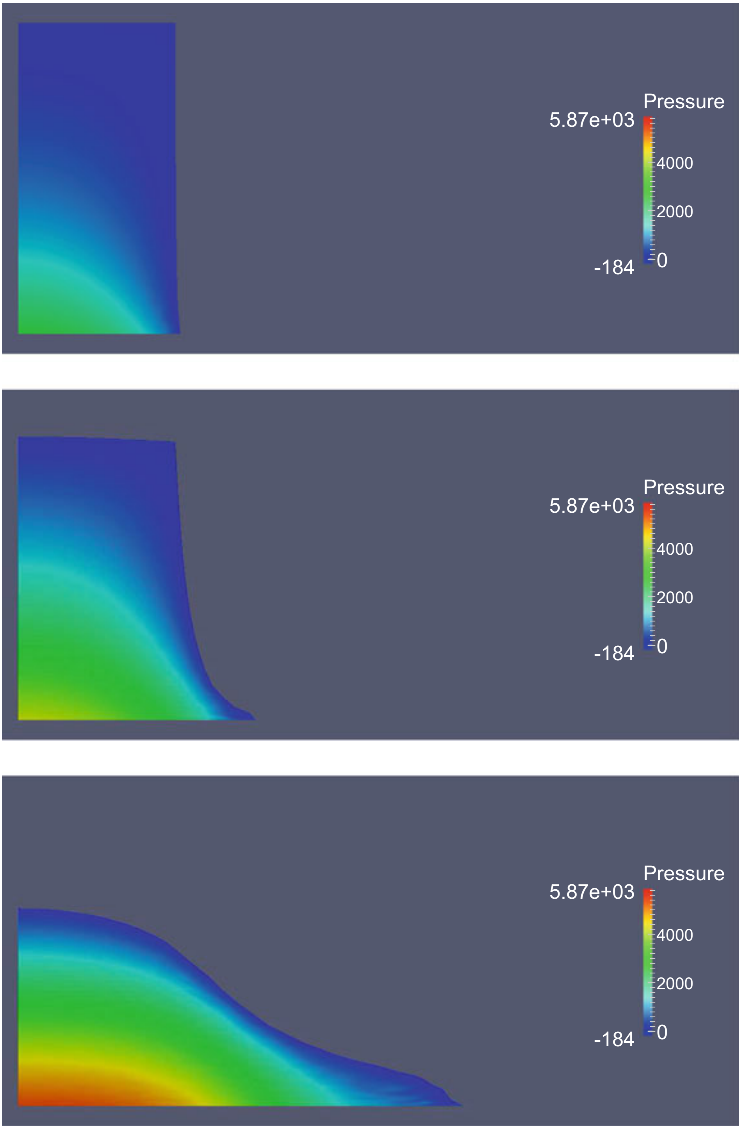


Fig. 6 Example 2: pressure and domain at $t = 0.01$, $t = 0.04$, $t = 0.09$

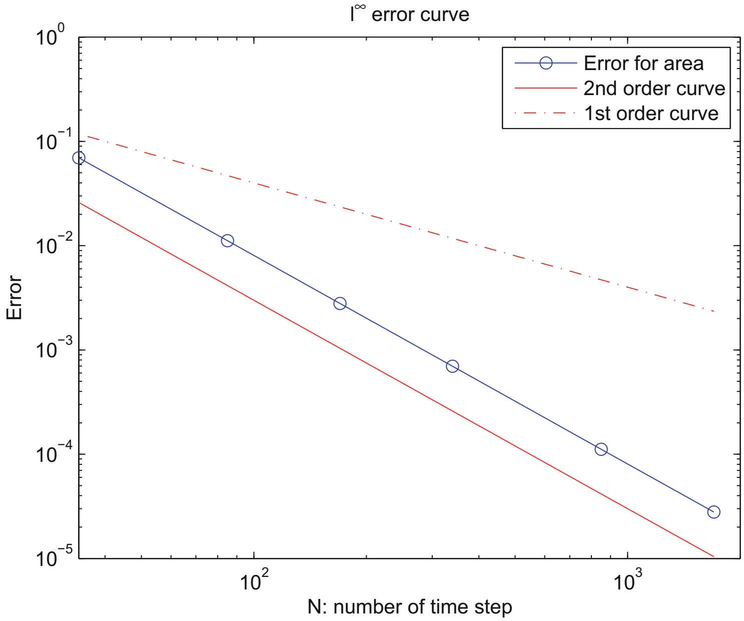


Fig. 7 Example 2: computed l^{∞} errors for area, in log-log scale, versus the number of time steps for a fixed spatial mesh of 368 vertices

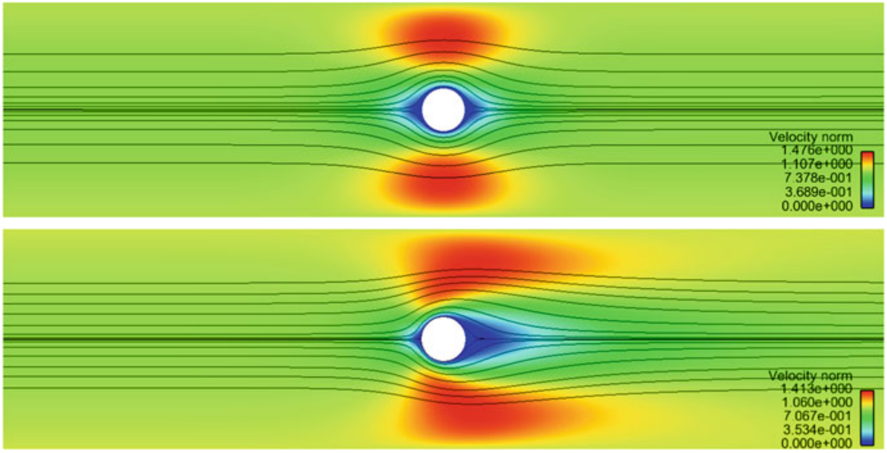


Fig. 8 Example 3: streamlines for $\mu = 1$ (top) and $\mu = 0.01$ (bottom), and for a spatial mesh of 38,910 vertices and $\Delta t = 0.001$

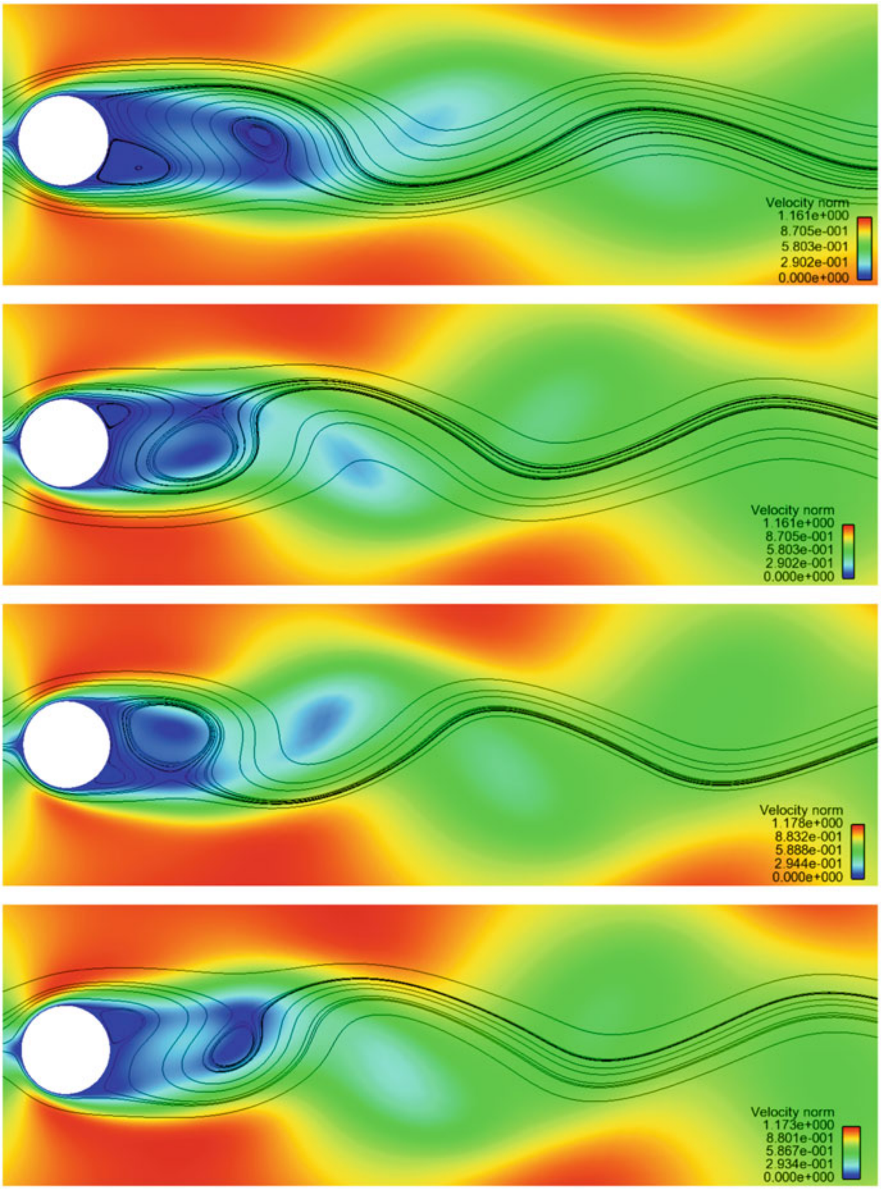


Fig. 9 Example 3: streamlines for $\mu = 0.001$ at four time instants, and for a spatial mesh of 38,910 vertices and $\Delta t = 0.0001$

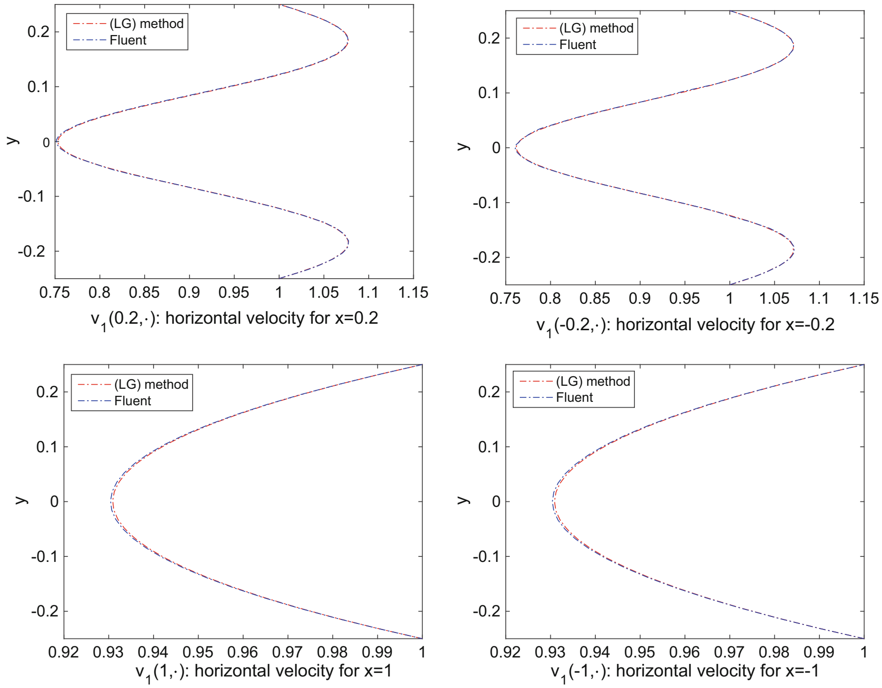


Fig. 10 Example 3: profiles of the horizontal velocity along the vertical lines: $x = 0.2$ (top left), $x = -0.2$ (top right), $x = 1$ (bottom left) and $x = -1$ (bottom right), computed using the pure Lagrange-Galerkin method with reinitializations **(LG)** and Fluent, for $\mu = 1$, and for a spatial mesh of $38{,}910$ vertices and $\Delta t = 0.001$

We also compare the results obtained with our code with those calculated using the commercial package Fluent. More precisely, in order to solve the problem with Fluent, a second-order stationary method was considered for $\mu = 1$ and $\mu = 0.01$. In Figs. 10 and 11 we represent, for $\mu = 1$ and $\mu = 0.01$, the horizontal velocity profiles along different vertical lines, calculated using our code and Fluent.

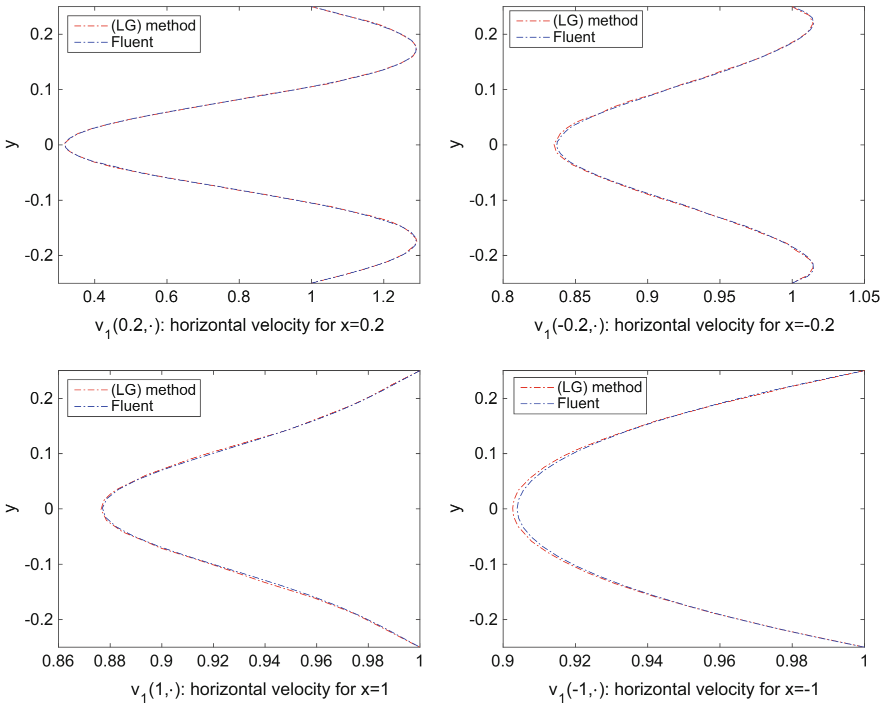


Fig. 11 Example 3: profiles of the horizontal velocity along the vertical lines: $x = 0.2$ (top left), $x = -0.2$ (top right), $x = 1$ (bottom left) and $x = -1$ (bottom right), computed using the pure Lagrange-Galerkin method with reinitializations **(LG)** and Fluent, for $\mu = 0.01$, and for a spatial mesh of 38,910 vertices and $\Delta t = 0.001$

Acknowledgements Authors wish to thank the referee for his/her valuable comments.

This work was partially funded by FEDER and the Spanish Ministry of Science and Innovation under research projects ENE2013-47867-C2-1-R and MTM2013-43745-R, and by FEDER and Xunta de Galicia under research project GRC2013/014.

References

1. Ewing, R.E., Wang, H.: A summary of numerical methods for time-dependent advection-dominated partial differential equations. J. Comput. Appl. Math. **128**, 423–445 (2001)
2. Douglas, J. Jr., Russell, T.F.: Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures. SIAM J. Numer. Anal. **19**, 871–885 (1982)
3. Pironneau, O.: On the transport-diffusion algorithm and its applications to the Navier-Stokes equations. Numer. Math. **38**, 309–332 (1982)
4. Süli, E.: Stability and convergence of the Lagrange-Galerkin method with non-exact integration. In: *The Mathematics of Finite Elements and Applications, VI*, pp. 435–442. Academic, London (1988)

5. Rui, H., Tabata, M.: A second order characteristic finite element scheme for convection-diffusion problems. Numer. Math. **92**, 161–177 (2002)
6. Bermúdez, A., Nogueiras, M.R., Vázquez, C.: Numerical analysis of convection-diffusion-reaction problems with higher order characteristics/finite elements. Part I: Time discretization. SIAM. J. Numer. Anal. **44**, 1829–1853 (2006)
7. Bermúdez, A., Nogueiras, M.R., Vázquez, C.: Numerical analysis of convection-diffusion-reaction problems with higher order characteristics/finite elements. Part II: fully discretized scheme and quadrature formulas. SIAM. J. Numer. Anal. **44**, 1854–1876 (2006)
8. Benítez, M., Bermúdez, A.: A second order characteristics finite element scheme for natural convection problems. J. Comput. Appl. Math. **235**, 3270–3284 (2011)
9. Idelsohn, S., Oñate, E., Del Pin, F.: The particle finite element method: a powerful tool to solve incompressible flows with free-surfaces and breaking waves. Int. J. Numer. Meth. Eng. **61**, 964–989 (2004)
10. Idelsohn, S., Oñate, E., Del Pin, F., Calvo, N.: Fluid-structure interaction using the particle finite element method. Comput. Methods Appl. Mech. Eng. **195**, 2100–2123 (2006)
11. Idelsohn, S., Marti, J., Limache, A., Oñate, E.: Unified Lagrangian formulation for elastic solids and incompressible fluids: application to fluid-structure interaction problems via the PFEM. Comput. Methods Appl. Mech. Eng. **197**, 1762–1776 (2008)
12. Del Pin, F., Idelsohn, S., Oñate, E., Aubry, R.: The Ale/Lagrangian particle finite element method: a new approach to computation of free-surface flows and fluid-object interactions. Comput. Fluids **36**, 27–38 (2007)
13. Oñate, E., Idelsohn, S., Celigueta, M.A., Rossi, R.: Advances in the particle finite element method for the analysis of fluid-multibody interaction and bed erosion in free surface flows. Comput. Methods Appl. Mech. Eng. **197**, 1777–1800 (2008)
14. Radovitzky, R., Ortiz, M.: Lagrangian finite element analysis of newtonian fluid flows. Int. J. Numer. Meth. Eng. **43**, 607–619 (1998)
15. Benítez, M., Bermúdez, A.: Numerical Analysis of a second-order pure Lagrange-Galerkin method for convection-diffusion problems. Part I: time discretization. SIAM. J. Numer. Anal. **50**, 858–882 (2012)
16. Benítez, M., Bermúdez, A.: Numerical Analysis of a second-order pure Lagrange-Galerkin method for convection-diffusion problems. Part II: fully discretized scheme and numerical results. SIAM. J. Numer. Anal. **50**, 2824–2844 (2012)
17. Benítez, M., Bermúdez, A.: Pure Lagrangian and semi-Lagrangian finite element methods for the numerical solution of convection-diffusion problems. Int. J. Numer. Anal. Mod. **11**, 271–287 (2014)
18. Benítez, M., Bermúdez, A.: Pure Lagrangian and semi-Lagrangian finite element methods for the numerical solution of Navier-Stokes equations. Appl. Numer. Math. **95**, 62–81 (2015)
19. Benítez, M., Bermúdez, A.: Second order pure Lagrange-Galerkin methods for fluid-structure interaction problems. SIAM J. Sci. Comput. **37**, B744–B777 (2015)
20. Gurtin, M.E.: An Introduction to Continuum Mechanics, vol. 158. Academic, San Diego (1981)
21. Badía, S., Codina, R.: Unified stabilized finite element formulations for the stokes and the darcy problems. SIAM J. Num. Anal. **47**, 1971–2000 (2009)
22. Raviart, P.A., Thomas, J.M.: A mixed-finite element method for second order elliptic problems. In: Mathematical Aspects of the Finite Element Method. Lecture Notes in Mathematics. Springer, New York (1977)
23. Brezzi, F., Fortin, M.: Mixed and Hybrid Finite Element Methods. Springer, Berlin (1991)
24. Ramaswamy, B., Kawahara, M.: Lagrangian finite element analysis applied to viscous free surface fluid flow. Int. J. Numer. Meth. Fluids **7**, 953–984 (1987)
25. Walhorn, E., Kölke, A., Hübner, B., Dinkler, D.: Fluid–structure coupling within a monolithic model involving free surface flows. Comput. Struct. **83**(25), 2100–2111 (2005)
26. Hansbo, P.: The characteristic streamline diffusion method for the time-dependent incompressible Navier-Stokes equations. Comput. Method. Appl. M. **99**(2–3), 171–186 (1992)
27. Wall, W.A., Genkinger, S., Ramm, E.: A strong coupling partitioned approach for fluid–structure interaction with free surfaces. Comput. Fluids **36**(1), 169–183 (2007)

28. Hirt, C.W., Nichols, B.D.: Volume of fluid (vof) method for the dynamics of free boundaries. J. Comput. Phys. **39**(1), 201–225 (1981)
29. Martin, J.C., Moyce, W.J.: Part IV. An experimental study of the collapse of liquid columns on a rigid horizontal plane. Philos. Trans. R. Soc. A **244**(882), 312–324 (1952)

Lubrication Theory and Viscous Shallow-Water Equations

Didier Bresch, Mathieu Colin, Xi Lin, and Pascal Noble

Dedicated to Prof. Enrique Fernández-Cara on the occasion of his 60th birthday.

Abstract In this proceeding dedicated to Enrique Fernández-Cara, we indicate how to derive rigorously lubrication equations studied in [A.L. Bertozzi, M.C. Pugh, *CPAM* (1998)] from the viscous shallow-water equations with drag terms and surface tension effect in the one-dimensional in space case. We consider strong and weak solutions and propose simple adaptation of recent relative entropy tools already developed by the first and third author. Note that the viscous shallow-water equation involves height dependent viscosity which vanishes if height vanishes. We choose such presentation because Enrique Fernández-Cara. worked with Francisco Guillén on non-homogeneous incompressible Navier-Stokes equations with density dependent viscosities and in some sense we want to show that in the compressible setting more general test functions are available so it helps to cover degenerate viscosity when height vanishes contrarily to the incompressible setting where such case is an open question.

D. Bresch (✉)
Université Grenoble Alpes, Université Savoie Mont Blanc, CNRS, LAMA, Chambéry, France
e-mail: Didier.Bresch@univ-smb.fr

M. Colin · X. Lin
Equipe INRIA, CARDAMOM, Institut Mathématiques de Bordeaux, Équipe EDP 351, Talence, France
e-mail: mcolin@math.u-bordeaux1.fr; xi.lin@inria.fr

P. Noble
Université de Lyon, Université Lyon 1, Institut Camille Jordan, UMR, CNRS, Villeurbanne Cedex, France
e-mail: noble@math.univ-lyon1.fr

A. Doubova et al. (eds.), *Recent Advances in PDEs: Analysis, Numerics and Control*, SEMA SIMAI Springer Series 17,
https://doi.org/10.1007/978-3-319-97613-6_4

Keywords Shallow-water equations · Navier-Stokes equations · Global weak solutions · Lubrication models

2000 MR Subject Classification 35B35

1 Introduction

It is a real pleasure for us to write a proceeding in honor of Enrique Fernandez-Cara's birthday: The first author is happy to have Enrique as friend and to have the opportunity in this special issue to precise that he started to work on compressible fluid system after discussion in Clermont-Ferrand with Enrique around 1997 in Jacques Simon' office and a first meet with Benoît Desjardins in 1998. The Navier-Stokes equations are considered as a basic mathematical model to describe the motion of a liquid. In his celebrated article "Sur le mouvement d'un liquide visqueux emplissant l'espace" published in *Acta Mathematica* in 1934, Jean Leray (1906–1998) introduced the concept of global in time weak solution with a precise definition of what could be an irregular solution of the system, and he has proved the existence of such weak solutions in the homogeneous (constant density) incompressible setting. We now talk about "solutions à la Leray" these solution of finite energy. Even if the global existence of weak solutions does not give too much information about the well posedness of the system, such analysis has a lot of practical interests. Beside the physical signification, because the assumed regularity on the data is minimal and strongly related to physical quantities well identified, the properties of stability of weak solutions on the continuous model help to better understand how to construct stable numerical schemes for which strong regularity estimates are not preserved.

The starting story correspond to the Leray solutions for the incompressible homogeneous Navier-Stokes equations in 1934, then results have been obtained for non-homogeneous incompressible Navier-Stokes equations by AM Kazhikhov (see [22] with initial density far from vacuum and bounded and with constant viscosity μ), J. Simon (see [28] with initial density with possible vacuum and with constant viscosity), Enrique Fernandez-Cara/Francisco Guillén-González (see [18] with initial density with possible vacuum and strictly positive density dependent viscosity $\mu(\rho)$ and the interesting review paper [16] by Enrique Fernandez-Cara. See also [19] in unbounded domains) and P.-L. Lions (see [25] for a full picture for non-homogeneous incompressible Navier-Stokes equations) and also the recent interesting result in [15] by R. Danchin and P. Mucha concerning global strong solution in the two dimensional setting for bound initial density and H^1 initial velocity. The starting story, in the multi-dimensional setting, for the compressible Navier-Stokes equations concerns **constant viscosities** μ and λ by P.-L. Lions (see [25]), E. Feireisl et al. (see [16]) with pressure laws as $P(\rho) = a\rho^\gamma$ or Van der Vaals type laws (which are increasing after a certain fixed density value). Recently it has been possible to obtain a more general result by D.B. and J.-E. Jabin covering thermodynamically unstable pressure laws (no monotonicity assumption) and some anisotropy in the viscosities, see for instance [10, 11].

Remark that **density dependent viscosities**, in the compressible setting, has been firstly studied by D.B., B. Desjardins based on an observation with C.K. Lin for the Korteweg-Navier-Stokes system in [8] that in the case where $\mu(\rho) = \rho$ and $\lambda(\rho) = 0$ we can control space derivative of the density if initial it is the case. It has been generalized in [7] to the case where $\mu(\rho)$ and $\lambda(\rho)$ are linked through the algebraic relation $\lambda(\rho) = 2(\mu'(\rho)\rho - \mu(\rho))$. It envolves a new mathematical entropy called now BD-entropy helping to control the gradient of a function of the density if initially it is the case. The method of proving global existence of weak solution is completely different and may be seen as a dual one compared to the constant viscosities case. The role and difficulties between the density and the velocity field are completely exchanged: In the constant viscosities case, the difficulties occur for compactness on the density to pass to the limit in the non-linear pressure law $p(\rho)$. In the density dependent viscosities case, the difficulties occur for compactness on the velocity field to pass to the limit in the non-linear quadratic term $\rho u \otimes u$. In the first case, control on $L \log L$ on the density through renormalization technic on the mass equation allow to get such compactness if the pressure law is an increasing function at least after a fixed value. In the second case, control on $L \log L$ quantity on the modulus of the velocity through renormalization technic on the momentum equations allow to get such compactness. The interested reader is referred to [5] and to the recent Bourbaki paper written by Rousset [27] for more information around density dependent viscosities and compressible Navier-Stokes equations and to the recent papers [24] and [29].

In this paper, we want to precise the limit between the viscous shallow-water equations with capillarity and drag terms envolved in [20] to the lubrication equation related to the height studied in [4] (see also works by Bertozzi et al. [1] and Bertozzi and Pugh [2]). Firstly, we get a weak convergence using the uniform bounds to a global weak solution of the lubrication system. Then we prove strong convergence to a strong solution of the lubrication system using a recent entropy inequality introduced in [13, 14] (and extended in [9] for Navier-Stokes-Korteweg system with compatibility condition between dispersive term and diffusive term). It is interesting to note here that in the compressible setting, density dependent viscosities vanishes if the density vanishes. This kind of dependency is not actually allowed in the non-homogeneous incompressible setting since a strictly positive properties is asked in the viscosity: see the very interesting review paper by Enrique Fernandez-Cara in [17].

2 Derivation from Shallow-Water Equations

Let us consider, in a periodic domain $\Omega = \mathbb{T}$, the shallow-water equations with linear drag term and surface tension:

$$\begin{aligned} &\partial_t h_\varepsilon + \partial_x(h_\varepsilon \overline{u}_\varepsilon) = 0, \\ &\partial_t(h_\varepsilon \overline{u}_\varepsilon) + \partial_x\left(h_\varepsilon \overline{u}_\varepsilon^2 + \frac{(h_\varepsilon)^2}{2\mathrm{Fr}^2}\right) = \frac{4}{\mathrm{Re}}\partial_x(h_\varepsilon \partial_x \overline{u}_\varepsilon) + \frac{1}{\mathrm{We}} h_\varepsilon \partial_x^3 h_\varepsilon - \alpha \overline{u}_\varepsilon, \end{aligned} \tag{1}$$

with $\alpha > 0$ where Re is the Reynolds number, We is the Weber number and Fr is the Froude number. Note that the terms in the right-hand side of the momentum equation represent respectively the viscous term, the capillarity term and the linear drag term. In the one-dimensional in space case, global existence of weak solutions of this system has been obtained by Bresch et al. [8] where the BD-entropy has been firstly introduced in the simplified setting. The more general BD entropy relation may be found in [7]. We consider the initial data

$$h_\varepsilon|_{t=0} = h_0^\varepsilon, \qquad (h_\varepsilon u_\varepsilon)|_{t=0} = m_0^\varepsilon.$$

In this paper, we consider the lubrication limit ($\varepsilon \ll 1$) with adimensionalized numbers under the form

$$\mathrm{We} := \varepsilon W_e, \quad \mathrm{Fr}^2 := \varepsilon F^2, \quad \alpha := \frac{\overline{\alpha}}{\varepsilon}$$

and the other dimensional numbers independent on ε. In the limit $\varepsilon \to 0$, on such system, assuming uniform bounds for all derivatives on the unknowns, we formally find

$$\overline{\alpha}\,\overline{u} = \frac{1}{W_e} h \partial_x^3 h - \frac{h \partial_x h}{F^2} \tag{2}$$

and

$$\partial_t h + \partial_x(h\overline{u}) = 0. \tag{3}$$

Combining Eq. (2) with (3), we obtain a lubrication equation

$$\partial_t h + \partial_x\left(\frac{1}{\overline{\alpha} W_e} h^2 \partial_x^3 h - \frac{1}{\overline{\alpha} F^2} h^2 \partial_x h\right) = 0. \tag{4}$$

The mathematical justification of such derivation is linked to the energy estimates and a mathematical entropy arising for the degenerate viscous shallow-water system that has been discovered in its first form in [8] and in its general form in [7]. We will discuss this derivation in the first section. Note that a similar asymptotic study has been performed recently in [23] focusing on singular van-der-Waals type pressure laws. Here we consider the standard shallow-water system occurring in geophysics, see for instance [20, 26] justified in [12].

Note that the energy estimate reads:

$$\frac{d}{dt}\left(\int_{\mathbb{T}} \varepsilon \frac{h_\varepsilon \overline{u}_\varepsilon^2}{2} + \frac{h_\varepsilon^2}{2F^2} + \frac{(\partial_x h_\varepsilon)^2}{2W_e}\right) + \int_{\mathbb{T}} \frac{4\varepsilon}{R_e} h_\varepsilon (\partial_x \overline{u}_\varepsilon)^2 + \overline{\alpha}\, \overline{u}_\varepsilon^2 \le 0. \tag{5}$$

This energy estimate is obtained multiplying the momentum equation by $\overline{u}_\varepsilon$ and adding the result to the following equation

$$\frac{1}{2}[\partial_t h_\varepsilon^2 + \partial_x (h_\varepsilon^2 \overline{u}_\varepsilon) + h_\varepsilon^2 \partial_x \overline{u}_\varepsilon] = 0$$

and then integrating in space. This last equation is obtained from the mass equation formally multiplying by h_ε and rewriting it. The BD entropy estimate is given by (see recall after proof):

$$\varepsilon \frac{d}{dt} \int_{\mathbb{T}} \frac{h_\varepsilon}{2} (\overline{u}_\varepsilon + 4(R_e)^{-1} \frac{\partial_x h_\varepsilon}{h_\varepsilon})^2 + \frac{d}{dt} \int_{\mathbb{T}} \Big(\frac{h_\varepsilon^2}{2F^2} + \frac{(\partial_x h_\varepsilon)^2}{2W_e} - \frac{4\overline{\alpha}}{R_e} \log_- h_\varepsilon\Big)$$
$$+ \frac{4}{R_e} \int_{\mathbb{T}} \frac{(\partial_x h_\varepsilon)^2}{F^2} + \frac{(\partial_x^2 h_\varepsilon)^2}{W_e} + \int_{\mathbb{T}} \overline{\alpha}\, \overline{u}_\varepsilon^2 \le 0 \tag{6}$$

Our result concerns weak solutions and is based on the following definition. The couple $(h_\varepsilon, u_\varepsilon)$ is called a global weak solutions of (1) if it satisfies (5)–(6) and

$$\int_0^\infty \int_{\mathbb{T}} h_\varepsilon \partial_t \psi + \int_{\mathbb{T}} h_0^\varepsilon \psi(\cdot, 0)\, dx = - \int_0^\infty \int_{\mathbb{T}} h_\varepsilon \overline{u}_\varepsilon \partial_x \psi \, dx dt \tag{7}$$

and

$$\varepsilon \Big(\int_0^\infty \int_{\mathbb{T}} h_\varepsilon \overline{u}_\varepsilon \partial_t \phi + \int_{\mathbb{T}} m_0^\varepsilon \phi(\cdot, 0)\, dx + \int_0^\infty \int_{\mathbb{T}} h_\varepsilon \overline{u}_\varepsilon^2 \partial_x \phi \, dx dt\Big) \tag{8}$$
$$- \frac{4\varepsilon}{R_e} \int_0^\infty \int_{\mathbb{T}} h_\varepsilon \partial_x \overline{u}_\varepsilon \partial_x \phi - \frac{1}{W_e} \int_0^\infty \int_{\mathbb{T}} \partial_x h_\varepsilon \partial_x^2 h_\varepsilon \phi \, dx dt$$
$$- \frac{1}{W_e} \int_0^\infty \int_{\mathbb{T}} h_\varepsilon \partial_x^2 h_\varepsilon \partial_x \phi \, dx dt + \frac{1}{F^2} \int_0^\infty \int_{\mathbb{T}} h_\varepsilon^2 \partial_x \phi \, dx dt - \overline{\alpha} \int_0^\infty \int_{\mathbb{T}} \overline{u}_\varepsilon \phi \, dx dt = 0$$

for all $\psi \in \mathcal{C}_0^\infty(\mathbb{T} \times [0, \infty))$ and $\phi \in \mathcal{C}_0^\infty(\mathbb{T} \times [0, \infty))$.

Let us first recall an existence result which may be found in [6].

Theorem *Let* $(h_0^\varepsilon, m_0^\varepsilon)$ *be such that* $h_0^\varepsilon \geq 0$ *and*

$$h_0^\varepsilon \in H^1(\Omega), \quad \varepsilon|m_0^\varepsilon|^2/h_0^\varepsilon \in L^1(\Omega), \quad \sqrt{\varepsilon}\partial_x\sqrt{h_0}^\varepsilon \in L^2(\Omega), \quad -\log_- h_0^\varepsilon \in L^1(\Omega)$$

where $\log_- \cdot = \log\min(\cdot, 1)$. *Then there exists a global weak solution of* (1) *in the sense of definition* (7)–(8).

Then we can give the following theorem which will be a straightforward application of bounds given by the energy and BD entropy. We will give the proof for reader's convenience.

Theorem *Let* $(h_\varepsilon, u_\varepsilon)$ *be a global weak solution of* (1) *as given in Theorem* 2 *with initial data satisfying the bounds uniformly. Then there exists a subsequence of* $(h_\varepsilon, \overline{u}_\varepsilon)$*, already denoted by* $(h_\varepsilon, \overline{u}_\varepsilon)$*, which converges to* $(h, \overline{u})$ *global weak solution of the lubrication system* (2)–(3) *satisfying the initial condition* $h|_{t=0} = h_0$ *with* h_0 *the weak limit in* $H^1(\Omega)$ *(up to a subsequence) of* h_0^ε.

Proof Due to the estimates, we have the following uniform bounds

$$\sqrt{\varepsilon}\|\sqrt{h}_\varepsilon\overline{u}_\varepsilon\|_{L^\infty(0,T;L^2(\Omega))} \leq C, \qquad \|h_\varepsilon\|_{L^\infty(0,T;H^1(\Omega))} \leq C,$$

$$\sqrt{\varepsilon}\|\sqrt{h}_\varepsilon\partial_x\overline{u}_\varepsilon\|_{L^2(0,T;L^2(\Omega))} \leq C, \qquad \|\overline{u}_\varepsilon\|_{L^2(0,T;L^2(\Omega))} \leq C.$$

Using this bounds, due to the BD entropy, the following extra uniform bounds

$$\sqrt{\varepsilon}\|\partial_x\sqrt{h}_\varepsilon\|_{L^\infty(0,T;L^2(\Omega))} \leq C, \quad \|h_\varepsilon\|_{L^2(0,T;H^2(\Omega))} \leq C.$$

Remark that, using the uniform $L^\infty(0, T; H^1(\Omega))$ bound of h_ε, we get

$$\|h_\varepsilon\|_{L^\infty} \leq C.$$

Using that $\partial_x(h_\varepsilon\overline{u}_\varepsilon) = h_\varepsilon\partial_x\overline{u}_\varepsilon + \overline{u}_\varepsilon\partial_x h_\varepsilon$ and the uniform bounds related to $\sqrt{h_\varepsilon}\partial_x\overline{u}_\varepsilon$ and h_ε and $\overline{u}_\varepsilon$, we get

$$\sqrt{\varepsilon}\|\partial_x(h_\varepsilon\overline{u}_\varepsilon)\|_{L^2(0,T;L^1(\Omega))} \leq C$$

Thus

$$\sqrt{\varepsilon}\|h_\varepsilon\overline{u}_\varepsilon\|_{L^2(0,T;W^{1,1}(\Omega))} \leq C.$$

Let us now pass to the limit in the weak formulation. In the mass equation, we use compactness on h_ε in $\mathcal{C}([0, T] \times \Omega)$ (due to bounds related to capillarity and estimates on $\partial_t h_\varepsilon$ looking at the mass equation in the distribution sense) and weak convergence in $L^2((0, T) \times \Omega)$ on $\overline{u}_\varepsilon$. Concerning the momentum equation, We easily pass to the limit in the third terms which will converge to 0 since they are multiplied by ε. The fourth one also converges to 0 since it concerns $\sqrt{h}_\varepsilon$ and

$\sqrt{\varepsilon}\sqrt{h_\varepsilon}\partial_x\overline{u}_\varepsilon$. Concerning the terms involving h_ε, we use the strong convergence of h_ε in $\mathcal{C}([0,T];H^s(\Omega))$ for all $s<1$ and in $L^2(0,T;H^s(\Omega))$ for $s<2$ and weak convergence of $\partial_x^2 h_\varepsilon$ in $L^2((0,T)\times\Omega)$. The last term is easy using the L^2 uniform bound on $\overline{u}_\varepsilon$.

Remark This is interesting to note that the BD entropy degenerates to similar entropy involved in lubrication theory for instance described in [1] and [2].

Recall Let us recall for reader's convenience how to derive the BD entropy. We differentiate the mass equation with respect to the space variable, it gives

$$\partial_t\partial_x h_\varepsilon + \partial_x(\overline{u}_\varepsilon\partial_x h_\varepsilon) + \partial_x(h_\varepsilon\partial_x\overline{u}_\varepsilon) = 0.$$

We remark that the last term is the same than the diffusive term in the momentum equation with an opposite sign if we multiply this equation by $4/\mathrm{Re}$. Thus multiplying by $4/\mathrm{Re}$ and adding the resulting equation with the momentum equation, we get

$$\partial_t\Big(h_\varepsilon(\overline{u}_\varepsilon+\frac{4}{\mathrm{Re}}\partial_x\log h_\varepsilon)\Big)+\partial_x\Big(h_\varepsilon(\overline{u}_\varepsilon+\frac{4}{\mathrm{Re}}\partial_x\log h_\varepsilon)\Big)+h_\varepsilon\Big(\frac{h_\varepsilon}{\mathrm{Fr}^2}\frac{\partial_x^3 h_\varepsilon}{\mathrm{We}}\Big)+\alpha\overline{u}_\varepsilon=0.$$

Multiplying this equation by $\overline{u}_\varepsilon+\frac{4}{\mathrm{Re}}\partial_x\log h_\varepsilon$ and using the mass equation $\partial_t h_\varepsilon + \partial_x(h_\varepsilon\overline{u}_\varepsilon)=0$, we get integrating by parts the BD entropy.

Comment Note that weak solutions to lubrication equations in the presence of strong slippage has been obtained in [23] from shallow-water equations. Strong slippage assumption with surface tension provides a height far from vanishing state (due to singular pressure laws and high derivative control of the height).

3 Relative Entropy and Strong Convergences

Let us now explain how to get better convergence result namely strong convergence from viscous shallow-water system to lubrication equation. More precisely let us consider again the following system

$$\begin{aligned}&\partial_t h_\varepsilon + \partial_x(h_\varepsilon\overline{u}_\varepsilon)=0,\\&\partial_t(h_\varepsilon\overline{u}_\varepsilon)+\partial_x\left(h_\varepsilon\overline{u}_\varepsilon^2+\frac{h_\varepsilon^2}{2\,\mathrm{Fr}^2}\right)=\frac{4}{\mathrm{Re}}\partial_x(h_\varepsilon\partial_x\overline{u}_\varepsilon)-\alpha\overline{u}_\varepsilon+\frac{1}{\mathrm{We}}h_\varepsilon\partial_x^3 h_\varepsilon,\end{aligned}\tag{9}$$

where Re is the Reynolds number, We is the Weber number and Fr is the Froude number. The coefficient α represents the friction due to the bottom and is assumed to be strictly positive. As explained before, such system has been studied initially in [8] and global existence of weak solutions has been proved in the one-dimensional setting using the linear drag term. Assuming the

$$\mathrm{We} := \varepsilon W_e, \qquad \mathrm{Fr}^2 := \varepsilon F^2, \qquad \alpha := \frac{\overline{\alpha}}{\varepsilon}$$

with W_e, F and $\overline{\alpha}$ fixed as explained in the introduction and looking at the limit $\varepsilon \to 0$, we can modulate the energy and BD-entropy in order to perform strong convergence between global weak solutions of the viscous shallow-water equation with damping and capillarity terms to strong solution $(\overline{h}, \overline{u})$ of the lubrication equations (4) using the fact that the height satisfies at the limit $h \geq c > 0$. Note that the limit quantity $(h, \overline{u})$ satisfies the following equations

$$\partial_t h + \partial_x (h\overline{u}) = 0. \tag{10}$$

Using (2), it also satisfies the momentum equation

$$\varepsilon\big(\partial_t(h\overline{u}) + \partial_x(h\overline{u}^2)\big) + \partial_x\Big(\frac{h^2}{2F^2}\Big) - \frac{4\varepsilon}{\mathrm{Re}}\partial_x(h\partial_x\overline{u}) - \frac{1}{W_e}h\partial_x^3 h + \overline{\alpha}\,\overline{u} = R_\varepsilon \tag{11}$$

where

$$R_\varepsilon = \varepsilon\big(\partial_t(h\overline{u}) + \partial_x(h\overline{u}^2)\big) - \frac{4\varepsilon}{\mathrm{Re}}\partial_x(h\partial_x\overline{u}).$$

Note that modulated technique for Navier-Stokes with density dependent viscosities has been recently developed in [13, 14] (extending an initial study by B. Haspot in [21] where the density dependent pressure law is assumed to be proportional to the density dependent viscosity). Namely such technic has been developed for the following system composed of mass equation

$$\partial_t \rho + \partial_x(\rho u) = 0$$

and momentum equation

$$\partial_t(\rho u) + \partial_x(\rho u^2) - \nu\partial_x(\rho\partial_x u) + \partial_x p(\rho) = 0$$

where $p(\rho) = a\rho^\gamma$. A weak-strong uniqueness result is also performed using this well defined modulated energy (relative entropy) control. Compared to what has been done in [13, 14], new terms here are therefore evolved namely the drag term $\overline{\alpha}\,\overline{u}$, the surface tension term $h\partial_x^3 h/We$ and the right-hand side R_ε.

Note that we also cannot use the recent work in [9] because in this paper capillarity coefficient and viscosity are assumed to be linked together in an appropriate way which is not satisfied by our model. To check if things work we need to look at new terms writing them in terms of the unknowns. Concerning R_ε, it is sufficient to assume that $R_\varepsilon \to 0$ when ε goes to zero in $L^1((0, T_\star); L^2(\mathbb{T}))$ where $T_\star$ is the existence time of strong solution of the lubrication equation where h

is strictly positive. To control this rest this ask for regularity properties on $(h, \overline{u})$ and this justified the fact that we consider strong limit solution of lubrication equation. Concerning the surface tension term, it suffices to write it as $h\partial_x^3 h = h\partial_x^2(hv)$ with $v = \partial_x \log h$. Concerning the drag term, since it is a linear one, it does not provide any difficulties. If we consider the following relative entropy

$$E(h_\varepsilon, \overline{u}_\varepsilon, \partial_x h_\varepsilon \;\big| h, \overline{u}, \partial_x h) = \tag{12}$$
$$\varepsilon \int_{\mathbb{T}} \frac{h_\varepsilon}{2}\Big(\big|\overline{u}_\varepsilon - \overline{u} + 4(R_e)^{-1}(\frac{\partial_x h_\varepsilon}{h_\varepsilon} - \frac{\partial_x h}{h})\big|^2 + |\overline{u}_\varepsilon - \overline{u}|^2\Big)$$
$$+ \int_{\mathbb{T}} \Big(\frac{|h_\varepsilon - h|^2}{F^2} + \frac{|\partial_x(h_\varepsilon - h)|^2}{2W_e} - \frac{4\overline{\alpha}}{R_e}\log_-(h_\varepsilon/h)\Big)$$

then by similar calculations than in [13] (we will not do such long but straightforward calculations again) we can get denoting the relative entropy $E_\varepsilon = E(h_\varepsilon, \overline{u}_\varepsilon, \partial_x h_\varepsilon \big| h, \overline{u}, \partial_x h)$ that

$$E_\varepsilon(t) \le \; E_\varepsilon(0)\exp[c(h)\int_0^t \|\partial_x\overline{u}\|_{L^\infty} + \|\partial_x \log h\|_{L^\infty}^2 + \|\partial_x^2 \log h\|_{L^\infty(\Omega)}]$$
$$+ \int_0^t \exp[c(h)\int_s^t (\|\partial_x\overline{u}\|_{L^\infty} + \|\partial_x \log h\|_{L^\infty}^2 + \|\partial_x^2 \log h\|_{L^\infty(\Omega)})d\tau]$$
$$\int_\Omega h_\varepsilon\big(R_\varepsilon(\overline{u}_\varepsilon - \overline{u} + 4(R_e)^{-1}(\frac{\partial_x h_\varepsilon}{h_\varepsilon} - \frac{\partial_x h}{h})) + R_\varepsilon(\overline{u}_\varepsilon - \overline{u})\big).$$

Thus assuming that $E_\varepsilon(0)$ goes to zero when ε go to zero, we get the convergence result using the convergence of R_ε to zero in $L^1(0, T_\star; L^2(\mathbb{T}))$.

4 Change of Time by Y. Brenier and X. Duan

In this last section, let us precise for reader's convenience a very interesting result by Y. Brenier and X. Duan concerning an appropriate quadratic change of variable that will provide a derivation of lubrication type model from the viscous-shallow system without any drag term initially present in the system. Let us consider the viscous shallow water without drag term namely the equation

$$\begin{aligned}
&\partial_t h_\varepsilon + \partial_x(h_\varepsilon \overline{u}_\varepsilon) = 0,\\
&\partial_t(h_\varepsilon \overline{u}_\varepsilon) + \partial_x\left(h_\varepsilon \overline{u}_\varepsilon^2 + \frac{h_\varepsilon^2}{2\,\mathrm{Fr}^2}\right) = \frac{4}{\mathrm{Re}}\partial_x(h_\varepsilon \partial_x \overline{u}_\varepsilon) + \frac{1}{\mathrm{We}} h_\varepsilon \partial_x^3 h_\varepsilon.
\end{aligned} \tag{13}$$

Let set

$$t \to \theta = t^2/2, \qquad \rho(t,x) \to \rho(\theta,x), \qquad v(t,v) \to v(\theta,x)\frac{d\theta}{dt}.$$

the system reads

$$\begin{aligned}
&\partial_\theta h_\varepsilon + \partial_x(h_\varepsilon \overline{u}_\varepsilon) = 0,\\
&h_\varepsilon \overline{u}_\varepsilon + 2\theta[\partial_\theta(h_\varepsilon \overline{u}_\varepsilon) + \partial_x \left(h_\varepsilon \overline{u}_\varepsilon^2\right)] + \partial_x(\frac{h_\varepsilon^2}{2\mathrm{Fr}^2}) = \frac{4\sqrt{2\theta}}{\mathrm{Re}}\partial_x(h_\varepsilon \partial_x \overline{u}_\varepsilon) + \frac{1}{\mathrm{We}} h_\varepsilon \partial_x^3 h_\varepsilon.
\end{aligned}$$

Thus letting formally θ goes to zero, we get the non-degenerate lubrication model

$$\partial_\theta h - \frac{1}{2\mathrm{Fr}^2}\partial_x(h\partial_x h) + \frac{1}{We}\partial_x(h\partial_x^3 h) = 0.$$

Obviously more general lubrication may be obtained starting from general Euler-Korteweg type systems. This would be interested to justify such asymptotic using the Relative entropy framework developed in [9] similarly than what has been done by Y. Brenier and X. Duan on curve-shortening flow in [3].

Acknowledgements Research of P. Noble was partially supported by French ANR project no. ANR-09-JCJC-0103-01. Research of M. Colin, X. Lin, D. Bresch has been partially supported by the ANR Project Viscap.

References

1. Bertozzi, A.L., Grün, G., Witelski, T.P.: Dewetting films: bifurcations and concentrations. Nonlinearity **14**, 1569–1592 (2001)
2. Bertozzi, A.L., Pugh, M.C.: Long-wave instabilities and saturation in thin film equations. Commun. Pure Appl. Math. **51**, 625–661 (1998)
3. Brenier, Y., Duan, X.: From conservative to dissipative systems through quadratic change of time, with application to the curve-shortening flow. ArXiv (2017)
4. Bresch, D., Colin, M., Msheik, K., Xi, L.: On a lubrication equation in one-dimension in space (2018, in preparation)
5. Bresch, D., Desjardins, B.: Weak solutions via the total energy formulation and their qualitative properties - density dependent viscosities. In: Y. Giga, Novotný, A. (eds.) Handbook of Mathematical Analysis in Mechanics of Viscous Fluids. Springer, Berlin (2017)
6. Bresch, D., Desjardins, B.: Existence of global weak solutions for a 2D viscous shallow water model and convergence to the quasigeostrophic model. Commun. Math. Phys. **238**(1–2), 211–223 (2003)
7. Bresch, D., Desjardins, B.: Quelques modèles diffusifs capillaires de type Korteweg. C. R. Acad. Sci. Paris Section Mécanique **332**(11), 881–886 (2004)
8. Bresch, D., Desjardins, B., Lin, C.K.: On some compressible fluid models: Korteweg, lubrication and shallow water systems. Commun. Part. Differ. Equ. **28**(3–4), 1009–1037 (2003)
9. Bresch, D., Gisclon, M., Lacroix-Violet, I.: On Navier-Stokes-Korteweg and Euler-Korteweg systems: application to quantum fluids models (2017, submitted)

10. Bresch, D., Jabin, P.-E.: Global existence of weak solutions for compressible Navier-Stokes equations: thermodynamical unstable pressure and anisotropy viscous stress tensor. Ann. Math. **188**, 577–684 (2018)
11. Bresch, D., Jabin, P.-E.: Global weak solutions of PDEs for compressible media: a compactness criterion to cover new physical situations. In: F. Colombini, D. Del Santo, D. Lannes (eds.) Shocks, Singularities and Oscillations in Nonlinear Optics and Fluid Mechanics. Springer INdAM-series, Spécial Issue Dedicated to G. Métivier, pp. 33–54. Springer, Cham (2017)
12. Bresch, D., Noble, P.: Mathematical derivation of viscous shallow-water equations with zero surface tension. Indiana Univ. J. **60**(4), 1137–1269 (2011)
13. Bresch, D., Noble, P., Vila, J.-P.: Relative entropy for compressible Navier-Stokes equations with density dependent viscosities and applications. C.R. Acad. Sci. Paris **354**(1), 45–49 (2016)
14. Bresch, D., Noble, P., Vila, J.-P.: Relative entropy for compressible Navier-Stokes equations with density dependent viscosities and various applications. In: ESAIM Proceedings (2017). https://doi.org/10.1051/proc/201758040
15. Danchin, R., Mucha, P.: The incompressible Navier-Stokes equations in vacuum. ArXiv:1705.06061 (2017)
16. Feireisl, E., Novotny, A., Petzeltova, H.: On the existence of globally defined weak solutions to the Navier-Stokes equations. J. Math. Fluid Mech. **3**(4), 358–392 (2001)
17. Fernández-Cara, E.: Motivation, analysis and control of the variable density Navier-Stokes equations. Discrete Contin. Dyn. Syst. Ser. S **5**, 1021–1090 (2012)
18. Fernández-Cara, E., Guillén-Gonzalez, F.: Some new existence results for the variable density Navier-Stokes Ann. Fac. Sci. Toulouse Math. Ser. 6 **2**(2), 185–204 (1993)
19. Fernández-Cara, E., Guillén-Gonzalez, F.: The existence of nonhomogeneous, viscous and incompressible flow in unbounded domains. Commun. Part. Differ. Equ. **17**(7 & 8), 1253–1265 (1992)
20. Gerbeau, J.F., Perthame, B.: Derivation of viscous Saint-Venant system for laminar shallow water: Numerical validation. Discrete Contin. Dyn. Syst. **1**, 89–102 (2001)
21. Haspot, B.: Weak-Strong uniqueness for compressible Navier-Stokes system with degenerate viscosity coefficient and vacuum in one dimension. Commun. Math. Sci. (2017, to appear)
22. Kazhikhov, A.: Resolution of boundary value problems for non homogeneous viscous fluids. Dokl. Akad. Nauk **216**, 1008–1010 (1974)
23. Kitavtsev, G., Laurencot, P., Niethammer, B.: Weak solutions to lubrication equations in the presence of strong slippage. Methods Appl. Anal. **18**, 183–202 (2011)
24. Li, J., Xin, Z.P.: Global existence of weak solutions to the barotropic compressible Navier-Stokes flows with degenerate viscosities. (2015, submitted) (see arXiv:1504.06826)
25. Lions, P.-L.: Mathematical Topics in Fluid Mechanics, vol. 1. Oxford Lecture Series in Mathematics and Its Applications. The Clarendon Press, Oxford University Press, New York (1996). Incompressible Models, Oxford Science Publications
26. Oron, A., Davis, S.H., Bankoff, S.G.: Long-scale evolution of thin liquid films. Rev. Mod. Phys. **69**(3), 931–980 (1997)
27. Rousset, F.: Solutions faibles de léquation de Navier-Stokes des fluides compressibles [d'après A. Vasseur et C. Yu]. Séminaire Bourbaki, no 135 (2016–2017)
28. Simon, J.: Nonhomogeneous viscous incompressible fluids: existence of velocity, density and pressure. SIAM J. Math. Anal. **21**(5), 1093–1117 (1990)
29. Vasseur, A., Yu, C.: Existence of global weak solutions for 3D degenerate compressible Navier-Stokes equations. Invent. Math. **206**, 935–974 (2016)

The Influence of the Tikhonov Term in Optimal Control of Partial Differential Equations

Eduardo Casas

Dedicated to Prof. Enrique Fernández-Cara on the occasion of his 60th birthday.

Abstract In this paper, we analyze the importance of the presence of the Tikhonov term in an optimal control problem. The influence of this term in several aspects of control theory is analyzed: existence and regularity of a solution, convergence of the numerical approximations, and second order optimality conditions.

Keywords Optimal control · Bang-bang controls · State constraints · Second order optimality conditions

1 Introduction

Throughout this paper, Ω denotes an open, bounded subset of $\mathbb{R}^n$, $1 \le n \le 3$, with a Lipschitz boundary Γ, and $0 < T < +\infty$ is fixed. We set $Q = \Omega \times (0, T)$ and $\Sigma = \Gamma \times (0, T)$. Let us consider the following control problem

$$\text{(P)} \quad \min\{J(u) : \alpha \le u(x,t) \le \beta \ \text{ for a.a. } (x,t) \in Q\},$$

where $-\infty \le \alpha < \beta \le +\infty$,

$$J(u) = \frac{1}{2}\int_Q (y_u - y_d)^2\,dx\,dt + \frac{\lambda}{2}\int_Q u^2\,dx\,dt$$

E. Casas (✉)
Departamento de Matemática Aplicada y Ciencias de la Computación, Universidad de Cantabria, Santander, Spain
e-mail: eduardo.casas@unican.es

A. Doubova et al. (eds.), *Recent Advances in PDEs: Analysis, Numerics and Control*, SEMA SIMAI Springer Series 17,
https://doi.org/10.1007/978-3-319-97613-6_5

with $y_d \in L^2(Q)$ and $\lambda \geq 0$. For every control u, we denote y_u the solution of

$$\begin{cases} \dfrac{\partial y}{\partial t} + Ay + a(x,t,y) = u \text{ in } Q, \\ y = 0 \text{ on } \Sigma, \\ y(0) = y_0 \text{ in } \Omega. \end{cases} \tag{1}$$

Here, A is the linear elliptic operator

$$Ay = -\sum_{i,j=1}^{n} \partial_{x_j}[a_{ij}(x)\,\partial_{x_i} y].$$

We make the following assumptions.

Assumption 1 The coefficients $a_{ij} \in L^\infty(\Omega)$ and satisfy

$$\exists \Lambda > 0 \text{ such that } \sum_{i,j=1}^{n} a_{ij}(x)\,\xi_i\,\xi_j \geq \Lambda\,|\xi|^2 \text{ for a.a. } x \in \Omega \text{ and } \forall \xi \in \mathbb{R}^n. \tag{2}$$

Assumption 2 The initial datum $y_0 \in L^\infty(\Omega)$, the target state $y_d \in L^{\hat{p}}(0,T;L^{\hat{q}}(\Omega))$, where $\hat{p},\hat{q} \in [2,+\infty]$ are such that $\frac{1}{\hat{p}} + \frac{n}{2\hat{q}} < 1$, and $a : Q \times \mathbb{R} \longrightarrow \mathbb{R}$ is a Carathéodory function of class C^2 with respect to the last variable, satisfying the following assumptions

$$\begin{cases} a(\cdot,\cdot,0) \in L^{\hat{p}}(0,T;L^{\hat{q}}(\Omega)) \text{ and } \exists C_a \leq 0 \text{ such that} \\ \dfrac{\partial a}{\partial y}(x,t,y) \geq C_a \text{ for a.a. } (x,t) \in Q \text{ and } \forall y \in \mathbb{R}, \end{cases} \tag{3}$$

$$\begin{cases} \forall M > 0\ \exists C_M > 0 \text{ such that} \\ \left|\dfrac{\partial^j a}{\partial y^j}(x,t,y)\right| \leq C_M \text{ for a.a. } (x,t) \in Q, \forall |y| \leq M, \text{ with } j = 1,2 \end{cases} \tag{4}$$

$$\begin{cases} \forall \rho > 0 \text{ and } \forall M > 0\ \exists \varepsilon_{M,\rho} > 0 \text{ such that for a.a. } (x,t) \in Q \\ \left|\dfrac{\partial^2 a}{\partial y^2}(x,t,y_2) - \dfrac{\partial^2 a}{\partial y^2}(x,t,y_1)\right| \leq \rho, \forall |y_i| \leq M, \text{ and } |y_2 - y_1| \leq \varepsilon_{M,\rho}. \end{cases} \tag{5}$$

Let us observe that the change of variable $\tilde{y} = \mathrm{e}^{C_a t} y$ transforms (1) in

$$\begin{cases} \dfrac{\partial \tilde{y}}{\partial t} + A\tilde{y} + \tilde{a}(x,t,\tilde{y}) = \mathrm{e}^{C_a t} u \text{ in } Q, \\ \tilde{y} = 0 \text{ on } \Sigma, \\ \tilde{y}(0) = y_0 \text{ in } \Omega. \end{cases} \tag{6}$$

where $\tilde{a}(x,t,y) = \mathrm{e}^{C_a t} a(x,t,\mathrm{e}^{-C_a t} y) - C_a y$. Now, we infer from (3) that

$$\frac{\partial \tilde{a}}{\partial y}(x,t,y) = \frac{\partial a}{\partial y}(x,t,\mathrm{e}^{C_a t} y) - C_a \geq 0.$$

Then, using the monotonicity of $\tilde{a}$ with respect to y, by classical arguments, we deduce the existence and uniqueness of a solution $\tilde{y}_u \in Y = W(0,T) \cap L^\infty(Q)$ for every $u \in L^{\hat{p}}(0,T;L^{\hat{q}}(\Omega))$; see, for instance, [4]. As usual we set

$$W(0,T) = \{y \in L^2(0,T;H_0^1(\Omega)) : \partial_t y \in L^2(0,T;H^{-1}(\Omega))\}.$$

From the equivalence between (1) and (6), we infer the existence and uniqueness of a solution $y_u \in Y$. In fact, we have the following result.

Theorem 1 *Under the above assumptions, for all $u \in L^{\hat{p}}(0,T;L^{\hat{q}}(\Omega))$ (1) has a unique solution $y_u \in Y$. The mapping $G : L^{\hat{p}}(0,T;L^{\hat{q}}(\Omega)) \longrightarrow Y$ defined by $G(u) = y_u$ is of class C^2. For all elements u, v, v_1 and v_2 of $L^{\hat{p}}(0,T;L^{\hat{q}}(\Omega))$, the functions $z_v = G'(u)v$ and $z_{v_1 v_2} = G''(u)(v_1,v_2)$ are the solutions of the problems*

$$\begin{cases} \dfrac{\partial z}{\partial t} + Az + \dfrac{\partial a}{\partial y}(x,t,y_u)z = v & \text{in } Q, \\ z = 0 & \text{on } \Sigma, \\ z(x,0) = 0 & \text{in } \Omega, \end{cases} \tag{7}$$

and

$$\begin{cases} \dfrac{\partial z}{\partial t} + Az + \dfrac{\partial a}{\partial y}(x,t,y_u)z + \dfrac{\partial^2 a}{\partial y^2}(x,t,y_u)z_{v_1} z_{v_2} = 0 & \text{in } Q, \\ z = 0 & \text{on } \Sigma, \\ z(x,0) = 0 & \text{in } \Omega, \end{cases} \tag{8}$$

respectively.

From this theorem we obtain easily the following result.

Theorem 2 *Under the above assumptions, the functional $J : L^{\hat{p}}(0,T;L^{\hat{q}}(\Omega)) \longrightarrow \mathbb{R}$ is of class C^2. For all u, v, v_1 and v_2 of $L^{\hat{p}}(0,T;L^{\hat{q}}(\Omega))$ we have*

$$J'(u)v = \int_Q (\varphi_u + \lambda u) v \, dx \, dt, \tag{9}$$

$$J''(u)(v_1,v_2) = \int_Q \left(1 - \varphi_u \frac{\partial^2 a}{\partial y^2}(x,t,y_u)\right) z_{v_1} z_{v_2} \, dx \, dt, \tag{10}$$

where $z_{v_i} = G'(u)v_i$, $i = 1, 2$, *and* $\varphi_u \in W(0, T) \cap C(\bar{Q})$ *is the solution of*

$$\begin{cases} -\dfrac{\partial \varphi}{\partial t} + A^*\varphi + \dfrac{\partial a}{\partial y}(x, t, y_u)\varphi = y_u - y_d & \text{in } Q,\\ \varphi = 0 & \text{on } \Sigma,\\ \varphi(x, T) = 0 & \text{in } \Omega, \end{cases} \tag{11}$$

For the proof of these theorems, the reader is referred to [8] and [20, Chapter 5]. In the sequel we denote

$$\mathbb{K}_{\alpha,\beta} = \{u \in L^{\hat{p}}(0, T; L^{\hat{q}}(\Omega)) : \alpha \leq u(x, t) \leq \beta \text{ for a.a. } (x, t) \in Q\}.$$

From (9) and the convexity of $\mathbb{K}_{\alpha,\beta}$ we infer the first order optimality conditions satisfied by a local minimum of (P); see, for instance, [20, Section §5.5].

Theorem 3 *Let* $\bar{u} \in \mathbb{K}_{\alpha,\beta}$ *be a local minimum of (P), then there exist elements* $\bar{y} \in Y$ *and* $\bar{\varphi} \in W(0, T) \cap C(\bar{Q})$ *such that*

$$\begin{cases} \dfrac{\partial \bar{y}}{\partial t} + A\bar{y} + a(x, t, \bar{y}) = \bar{u} & \text{in } Q,\\ \bar{y} = 0 & \text{on } \Sigma,\\ \bar{y}(0) = y_0 & \text{in } \Omega, \end{cases} \tag{12}$$

$$\begin{cases} -\dfrac{\partial \bar{\varphi}}{\partial t} + A^*\bar{\varphi} + \dfrac{\partial a}{\partial y}(x, t, \bar{y})\,\bar{\varphi} = \bar{y} - y_d & \text{in } Q,\\ \bar{\varphi} = 0 & \text{on } \Sigma,\\ \bar{\varphi}(T) = 0 & \text{in } \Omega, \end{cases} \tag{13}$$

$$\int_Q (\bar{\varphi} + \lambda\bar{u})(u - \bar{u})\, dx\, dt \geq 0 \quad \forall u \in \mathbb{K}_{\alpha,\beta}. \tag{14}$$

In this paper, we will analyze different issues of the control problem where λ plays a crucial role. The term $\frac{\lambda}{2}\|u\|^2_{L^2(Q)}$ in the cost functional J is called the Tikhonov term, and λ is the Tikhonov parameter. The presence of the Tikhonov term in the cost functional, i.e., $\lambda > 0$, changes very much the control problem: sometimes, it is essential to prove the existence of a solution; it produces a regularizing effect in the optimal control; the second order analysis produces the same results as for finite dimensional optimization problems, with a minimal gap between the necessary and sufficient second order conditions; it is possible to prove good properties of stability of the solutions of (P) with respect to perturbations in the data; we can prove error estimates for the numerical approximation of (P); and, finally, the numerical algorithms work much better when $\lambda > 0$.

When $\lambda = 0$, it is necessary to assume $-\infty < \alpha < \beta < +\infty$ to prove the existence of a solution; the optimal control is essentially discontinuous; the second order analysis is much more complicated and the research is still in progress; in general, it is neither possible to prove stability properties of the solutions of (P) with respect to perturbations of the data, nor can we get error estimates for the numerical approximation; and, finally, the numerical algorithms are unstable.

The first difference in the regularity of the optimal control is deduced from (14). Indeed, if $\lambda > 0$, it is easy to obtain from (14) the identity

$$\bar{u}(x,t) = \mathrm{Proj}_{[\alpha,\beta]}\Big(-\frac{1}{\lambda}\bar{\varphi}(x,t)\Big) \text{ for a.a. } (x,t) \in Q. \tag{15}$$

This implies that $\bar{u}$ inherits some regularity of $\bar{\varphi}$. Actually we have that $\bar{u} \in L^2(0,T;H^1(\Omega)) \cap C(\bar{Q})$. However, for $\lambda = 0$ and $-\infty < \alpha < \beta < +\infty$, (14) leads to

$$\bar{u}(x,t) = \begin{cases} \alpha \text{ if } \bar{\varphi}(x,t) > 0, \\ \beta \text{ if } \bar{\varphi}(x,t) < 0, \end{cases} \quad \text{for a.a. } (x,t) \in Q, \tag{16}$$

which shows that $\bar{u}$ is essentially discontinuous. If the set $\{(x,t) \in Q : \bar{\varphi}(x,t) = 0\}$ has zero Lebesgue measure, then $\bar{u}(x,t) \in \{\alpha,\beta\}$ a.e. in Q. These controls are known in the literature as bang-bang controls. It is frequent for an optimal control to be bang-bang when $\lambda = 0$. More differences will be shown in the rest of the paper.

The plan of this paper is as follows. In Sect. 2, we prove that, unlike it was believed, control constraints are not necessary to establish the existence of a solution of (P) if $\lambda > 0$. Of course, they are necessary if $\lambda = 0$. Moreover, the issue of uniqueness of an optimal control is analyzed. In Sect. 3, we add pointwise state constraints to the control problem. Assuming that $\lambda > 0$, we will prove an extra regularity for the optimal control, improving some existing results. Of course, the assumption $\lambda > 0$ is necessary to prove this additional regularity. In Sect. 4, we analyzed the convergence of the numerical approximation of the control problem. When $\lambda > 0$, the proof of a strong convergence of the controls is well known. However for $\lambda = 0$, only weak convergence is usually established; we prove that the convergence is strong if the continuous control is bang-bang. Finally, in Sect. 5, we show the existing very good results for the second order analysis when $\lambda > 0$, and the difficulties of this analysis when $\lambda = 0$.

2 About the Existence and Uniqueness of Optimal Controls

Theorem 1 establishes the existence of a unique solution y_u for every control $u \in L^{\hat{p}}(0,T;L^{\hat{q}}(\Omega))$. If we assume that the controls are only elements of $L^2(Q)$, then the analysis of Eq. (1) is much more involved. Though we could prove the existence of a solution $y_u \in W(0,T)$, this solution does not belong to $L^\infty(Q)$. Hence, the differentiability of the relation $u \in L^2(Q) \to y_u \in W(0,T)$ is not clear at all.

Therefore, the first and second order analysis of the control problem becomes too complicate or even impossible. To overcome this difficulty, it is usual to consider the controls in $L^\infty(Q)$. However, the cost functional J is not coercive in this space, and the existence of a solution to the control problem cannot be proved by standard arguments. This situation is solved by including control constraints of type $\alpha \leq u(x,t) \leq \beta$ with $-\infty < \alpha < \beta < +\infty$ in the formulation of the control problem. In particular, if $\lambda = 0$, the inclusion of control constraints is the only way to ensure the existence of a solution. Here, we prove that it is not necessary to include the control constraints to establish the existence of a solution if $\lambda > 0$. We will also address the issue of uniqueness of solution to (P) in the second part of the section.

2.1 Existence of Solution of (P)

Next, the goal is to prove the existence of a solution of (P). See [12] for a first proof of this result.

Theorem 4 *If* $\lambda > 0$*, then problem (P) has at least one solution* $\bar{u}$*.*

Proof For every real number $M > 0$ we consider the control problem

$$(\mathrm{P}_M) \quad \min\{J(u) : -M \leq u(x,t) \leq +M \text{ for a.a. } (x,t) \in Q\}.$$

This coincides with (P), where $\alpha = -M$ and $\beta = +M$. The existence of a solution $\bar{u}_M$ of this problem is proved by taking a minimizing sequence and following the classical arguments. We denote by $\bar{y}_M$ and $\bar{\varphi}_M$ the state and adjoint state associated with $\bar{u}_M$. Then, $(\bar{u}_M.\bar{y}_M, \bar{\varphi}_M)$ satisfies (12)–(14). Now, (15) is written as follows:

$$\bar{u}_M(x,t) = \mathrm{Proj}_{[-M,+M]}\Big(-\frac{1}{\lambda}\bar{\varphi}_M(x,t)\Big) \text{ for a.a. } (x,t) \in Q. \tag{17}$$

Since $\bar{u}_M$ is a solution of (P_M) and 0 is a feasible control of (P_M) we get that $J(\bar{u}_M) \leq J(0)$, hence

$$\|\bar{u}_M\|_{L^2(Q)} \leq \frac{1}{\sqrt{\lambda}}\|y^0 - y_d\|_{L^2(Q)} = C_0, \tag{18}$$

where y^0 denotes the state associated with 0. Now, by standard arguments, from (12) we infer with (18)

$$\begin{aligned}\|\bar{y}_M\|_{L^\infty(0,T;L^2(\Omega))} &\leq C_1\big(\|y_0\|_{L^2(\Omega)} + \|a_0(\cdot,\cdot,0)\|_{L^2(Q)} + \|\bar{u}_M\|_{L^2(Q)}\big)\\ &\leq C_1\big(\|y_0\|_{L^2(\Omega)} + \|a_0(\cdot,\cdot,0)\|_{L^2(Q)} + C_0\big) = C_2.\end{aligned} \tag{19}$$

Looking at the adjoint state equation (13), we get (see [15, Section §3.7]) with (19)

$$\|\bar\varphi_M\|_{L^\infty(Q)} \le C_3\big(\|\bar y_M\|_{L^\infty(0,T;L^2(\Omega))} + \|y_d\|_{L^{\hat p}(0,T;L^{\hat q}(\Omega))}\big)$$
$$\le C_3\big(C_2 + \|y_d\|_{L^{\hat p}(0,T;L^{\hat q}(\Omega))}\big) = C_4. \tag{20}$$

Finally, (17) and (20) imply

$$\|\bar u_M\|_{L^\infty(Q)} \le \frac{C_4}{\lambda} = C_\infty \;\; \forall M > 0. \tag{21}$$

Let us denote by $\bar u$ a solution of (P$_{M_\infty}$) for $M_\infty = C_\infty$. We conclude the proof by showing that $\bar u$ is a solution of (P). Given an arbitrary element $u \in L^\infty(Q)$ we set $M = \|u\|_{L^\infty(Q)}$. Any solution $\bar u_M$ of (P$_M$) satisfies (21), then it is a feasible control for (P$_{M_\infty}$) and, therefore, $J(\bar u) \le J(\bar u_M) \le J(u)$. Hence, $\bar u$ is a solution of (P). □

2.2 About the Uniqueness of Solution of (P)

Let us observe that the control problem (P) is not convex due to the nonlinearity of the state equation. The uniqueness of a solution is an open question up to now. There is a recent paper [1] where a uniqueness result is proved for a semilinear elliptic control problem under an structural assumption of the nonlinear function a in the state equation, and assuming that λ is large enough. A precise constant η only depending on λ and a is given such that the uniqueness holds whenever the $\|\bar\varphi\|_{L^q} \le \eta$ for a certain q depending on a.

If the cost functional is not convex with respect to y, then the uniqueness is false in general. Indeed, let us give an example. We consider the state equation

$$\begin{cases} \dfrac{\partial y}{\partial t} + Ay + y^3 = u & \text{in } Q,\\ y = 0 & \text{on } \Sigma,\\ y(0) = 0 & \text{in } \Omega. \end{cases} \tag{22}$$

We denote by $z \in Y$ the solution of the problem

$$\begin{cases} \dfrac{\partial z}{\partial t} + Az = 1 & \text{in } Q,\\ z = 0 & \text{on } \Sigma,\\ z(0) = 0 & \text{in } \Omega. \end{cases} \tag{23}$$

Take λ satisfying

$$0 < \lambda < \frac{1}{|Q|}\int_Q z^2(x,t)\,dx\,dt. \tag{24}$$

Finally, we define the cost functional

$$J(u) = \frac{1}{4}\int_Q (y_u^2 - 1)^2\,dx\,dt + \frac{\lambda}{2}\int_Q u^2\,dx\,dt,$$

and the associated control problem

$$\text{(P)} \quad \min_{u\in L^\infty(Q)} J(u),$$

Since the functional J is not quadratic with respect to y, the existence of a solution of (P) is not a consequence of Theorem 4. However, we can establish the existence of a solution by a small modification of the arguments of the proof of Theorem 4. First, we observe that given $u \in L^\infty(Q)$, (22) has a unique solution $y_u \in L^2(0,T; H_0^1(\Omega)) \cap L^\infty(Q)$. Hence, $\partial_t y_u + Ay_u \in L^2(Q)$ and consequently $y_u \in H^1(Q) \cap C([0,T]; H_0^1(\Omega))$; see [19, Section §III.2]. Then, multiplying (22) by y_u^3 we deduce that

$$\|y_u^3\|_{L^2(Q)} \le \|u\|_{L^2(Q)}.$$

Therefore, following again [19] we get

$$\|y_u\|_{L^\infty(0,T;H_0^1(\Omega))} \le C\|u - y_u^3\|_{L^2(Q)} \le 2C\|u\|_{L^2(Q)}.$$

On the other hand, looking at the right hand side of the adjoint state equation

$$\begin{cases} -\dfrac{\partial\varphi_u}{\partial t} + A^*\varphi_u + 3y_u^2\varphi_u = y_u(y_u^2 - 1) & \text{in } Q,\\ \varphi_u = 0 & \text{on } \Sigma,\\ \varphi_u(T) = 0 & \text{in } \Omega, \end{cases} \tag{25}$$

we get

$$\begin{aligned}\|\varphi_u\|_{L^\infty(Q)} &\le C'\|y_u(y_u^2-1)\|_{L^\infty(0,T;L^2(\Omega))} \le C'\big(\|y_u^3\|_{L^\infty(0,T;L^2(\Omega))} + \|y_u\|_{L^\infty(0,T;L^2(\Omega))}\big)\\ &\le C''\big(\|y_u\|^3_{L^\infty(0,T;H_0^1(\Omega))} + \|y_u\|_{L^\infty(0,T;L^2(\Omega))}\big) \le C''\|u\|_{L^2(Q)}\big(8C^3\|u\|^2_{L^2(Q)} + C'''\big).\end{aligned}$$

Using this estimate and arguing as in the proof of Theorem 4 we obtain the existence of a solution $\bar u$ of (P). Let us prove that $\bar u \not\equiv 0$. To this end we observe that (9) and (10) lead to

$$J'(u)v = \int_Q (\varphi_u \lambda u)v\,dx\,dt,$$

$$J''(u)v^2 = \int_Q \left[\left(3y_u^2 - 1 - 6\varphi_u y_u\right)z_v^2 + \lambda v^2\right] dx\,dt,$$

where z_v is the solution of the linearized equation

$$\begin{cases} \dfrac{\partial z}{\partial t} + Az + 3y_u^2 z = v & \text{in } Q, \\ z = 0 & \text{on } \Sigma, \\ z(0) = 0 & \text{in } \Omega, \end{cases} \tag{26}$$

For $\bar{u} \equiv 0$ we obtain the state $\bar{y} \equiv 0$. Then, the solution of (25) with $y_u = \bar{y} \equiv 0$ is $\bar{\varphi} \equiv 0$. Hence, the first order necessary optimality condition $J'(\bar{u})v = 0\ \forall v \in L^\infty(Q)$ holds. However, let us check that the second order necessary condition does not hold. We take $v \equiv 1$. Since $\bar{y} \equiv 0$, we have that the solution z_v of (26) coincides with z, solution of (23). Then, (24) implies

$$J''(\bar{u})v^2 = \int_Q [-z^2(x,t) + \lambda]\,dx\,dt < 0.$$

Hence, the second order necessary condition $J''(\bar{u})v^2 \geq 0\ \forall v \in L^\infty(Q)$ does not hold. Consequently, $\bar{u} \equiv 0$ is not a solution of (P). Finally, let us take $\tilde{u} = -\bar{u}$. We observe that the associated state $\tilde{y}$ satisfies $\tilde{y} = -\bar{y}$. Therefore, $J(\tilde{u}) = J(\bar{u})$ and $\bar{u} \neq \tilde{u}$, which proves that there exist at least two solutions of (P).

3 Regularity of Optimal Solutions of State-Constrained Control Problems

In this section, we assume that $\lambda > 0$ and include pointwise state constraints.

$$\text{(P)}\quad \min\{J(u) : u \in \mathbb{K}_{\alpha,\beta} \text{ and } a \leq y_u(x,t) \leq b\ \ \forall (x,t) \in \bar{Q}\},$$

where $-\infty < \alpha < \beta < +\infty$ and $-\infty < a < b < +\infty$. Here, we assume that the initial condition y_0 belongs to $C_0(\Omega)$ with

$$C_0(\Omega) = \{z \in C(\bar{\Omega}) : z = 0 \text{ on } \Gamma\}.$$

We also assume that $a < y_0(x) < b\ \forall x \in \bar{\Omega}$.

Due to the continuity of y_0, the fact that $a(\cdot,\cdot,0) \in L^{\hat{p}}(0,T;L^{\hat{q}}(\Omega))$ and $u \in L^\infty(Q)$, we have that $y_u \in W(0,T) \cap C(\bar{Q})$. This follows from [15, Sections §3.7 and §3.10]. With $M(Q)$ and $M(\Omega)$ we denote the spaces of real and regular Borel measures in Q and Ω, respectively. These space are identified with the dual spaces of $C_0(Q)$ and $C_0(\Omega)$, respectively; see, for instance, [18, Section §6.18]. Analogously to $C_0(\Omega)$, $C_0(Q)$ denotes the space of continuous functions in $\bar{Q}$ vanishing on ∂Q. Moreover, we have that

$$\|\mu\|_{M(Q)} = \sup\left\{\int_Q z\,d\mu : \|z\|_{C_0(Q)} \le 1\right\} = |\mu|(Q),$$

where $|\mu|(Q)$ is the total variation of μ. The analogous norm is defined in $M(\Omega)$.

Associated with the state constraints we define the set

$$\mathbb{C}_{a,b} = \{z \in C(\bar{Q}) : a \le z(x,t) \le b\ \forall (x,t) \in \bar{Q} \text{ and } z = 0 \text{ on } \Sigma\}.$$

Assuming that there exist at least one control $u \in \mathbb{K}_{\alpha,\beta}$ such that the associated state y_u belongs to $\mathbb{C}_{a,b}$, it is easy to prove the existence of a solution of (P). If $\bar{u}$ is solution of (P), we say that $\bar{u}$ satisfies the linearized Slater condition if

$$\exists u_0 \in \mathbb{K}_{\alpha,\beta} \text{ such that } a < \bar{y}(x,t) + z_{u_0-\bar{u}}(x,t) < b\ \ \forall (x,t) \in \bar{Q}, \tag{27}$$

where $\bar{y} = G(\bar{u})$ is the state associated with $\bar{u}$ and $z_{u_0-\bar{u}} = G'(\bar{u})(u_0-\bar{u})$ is the solution of (7) with $y_u = \bar{y}$ and $v = u_0 - \bar{u}$. The following optimality conditions are well known [4, 11, 17].

Theorem 5 *If $\bar{u}$ is a solution of (P) satisfying the linearized Slater condition (27), then there exist $\bar{y} \in W(0,T) \cap C(\bar{Q})$, $\bar{\varphi} \in L^p(0,T;W_0^{1,q}(\Omega))$ $\forall p,q \in [1,2)$ with $\frac{1}{p} + \frac{n}{2q} > \frac{n+1}{2}$, $\bar{\mu}_Q \in M(Q)$ and $\bar{\mu}_\Omega \in M(\Omega)$ such that*

$$\begin{cases} \dfrac{\partial \bar{y}}{\partial t} + A\bar{y} + a(x,t,\bar{y}) = \bar{u} & \text{in } Q,\\ \bar{y} = 0 & \text{on } \Sigma,\\ \bar{y}(0) = y_0 & \text{in } \Omega, \end{cases} \tag{28}$$

$$\begin{cases} -\dfrac{\partial \bar{\varphi}}{\partial t} + A^*\bar{\varphi} + \dfrac{\partial a}{\partial y}(x,t,\bar{y})\,\bar{\varphi} = \bar{y} - y_d + \bar{\mu}_Q & \text{in } Q,\\ \bar{\varphi} = 0 & \text{on } \Sigma,\\ \bar{\varphi}(T) = \bar{\mu}_\Omega & \text{in } \Omega, \end{cases} \tag{29}$$

$$\int_Q (z(x,t) - \bar{y}(x,t))\,d\bar{\mu}_Q + \int_\Omega (z(x,T) - \bar{y}(x,T))\,d\bar{\mu}_\Omega \le 0\ \ \forall z \in \mathbb{C}_{a,b}, \tag{30}$$

$$\int_Q (\bar{\varphi} + \lambda\bar{u})(u - \bar{u})\,dx\,dt \ge 0\ \forall u \in \mathbb{K}_{\alpha,\beta}. \tag{31}$$

We say that $\bar{\varphi} \in L^1(Q)$ is a solution of (29) if

$$\int_Q \bar{\varphi}\Big(\frac{\partial z}{\partial t} + Az + \frac{\partial a}{\partial y}(x,t,\bar{y})z\Big)\,dx\,dt = \int_Q z\,d\bar{\mu}_Q + \int_\Omega z(x,T)\,d\bar{\mu}_\Omega \ \ \forall z \in Z, \tag{32}$$

where

$$Z = \Big\{z \in L^2(0,T;H_0^1(\Omega)) : \frac{\partial z}{\partial t} + Az \in L^\infty(Q) \text{ and } z(x,0) = 0\Big\}.$$

We observe that $Z \subset C(\bar{Q})$ [15, Section §3.7 and §3.10], hence the right hand side of (32) is well defined. In [7], it is proved that there exists a unique solution $\bar{\varphi}$ of (29) in the sense above described, and $\bar{\varphi} \in L^p(0,T;W_0^{1,q}(\Omega))$ $\forall p,q \in [1,2)$ with $\frac{1}{p} + \frac{n}{2q} > \frac{n+1}{2}$. From (31) we deduce again the identity (15). From this identity we infer that $\bar{u} \in L^p(0,T;W^{1,q}(\Omega))$ $\forall p,q \in [1,2)$ with $\frac{1}{p} + \frac{n}{2q} > \frac{n+1}{2}$. For long time, this was the maximal regularity expected for the optimal solution of the control problem. However, by a simple argument that we show below, we obtain that $\bar{u} \in L^2(0,T;H^1(\Omega))$. This additional regularity is very important in the derivation of error estimates for the numerical approximation of the control problem. The proof is based on the following lemma.

Lemma 1 *Let $\bar{\varphi}$ be the solution of* (29)*. Given $M > 0$, we set*

$$\varphi_M(x,t) = \text{Proj}_{[-M,+M]}(\bar{\varphi}(x,t)).$$

Then, $\varphi_M \in L^2(0,T;H_0^1(\Omega))$ and there exists a constant $C = C(\Omega,\Lambda,C_a) > 0$ independent of M such that

$$\|\varphi_M\|_{L^2(0,T;H_0^1(\Omega))} \le C\Big[\|\bar{y} - y_d\|_{L^2(Q)} + \sqrt{M\big(\|\bar{\mu}_Q\|_{M(Q)} + \|\bar{\mu}_\Omega\|_{M(\Omega)}\big)}\Big]. \tag{33}$$

Proof Let us consider two sequences $\{f_k\}_{k=1}^\infty \subset L^2(Q)$ and $\{g_k\}_{k=1}^\infty \subset H_0^1(\Omega)$ satisfying

$$\|f_k\|_{L^1(Q)} \le \|\bar{\mu}_Q\|_{M(Q)} \text{ and } f_k \overset{*}{\rightharpoonup} \bar{\mu}_Q \text{ in } M(Q), \tag{34}$$

$$\|g_k\|_{L^1(\Omega)} \le \|\bar{\mu}_\Omega\|_{M(\Omega)} \text{ and } g_k \overset{*}{\rightharpoonup} \bar{\mu}_\Omega \text{ in } M(\Omega). \tag{35}$$

Now we consider the problem

$$\begin{cases} -\dfrac{\partial \varphi_k}{\partial t} + A^*\varphi_k + \dfrac{\partial a}{\partial y}(x,t,\bar{y})\,\varphi_k = \bar{y} - y_d + f_k & \text{in } Q, \\ \varphi_k = 0 & \text{on } \Sigma, \\ \varphi_k(T) = g_k & \text{in } \Omega. \end{cases} \tag{36}$$

The solution φ_k is unique and belongs to $H^1(Q) \cap C([0,T]; H_0^1(\Omega))$. From (34) and (35) we get the convergence

$$\varphi_k \rightharpoonup \bar{\varphi} \text{ in } L^p(0,T; W_0^{1,q}(\Omega))\ \forall p,q \in [1,2) \text{ with } \frac{1}{p} + \frac{n}{2q} > \frac{n+1}{2}, \tag{37}$$

$$\lim_{k\to\infty} \|\bar{\varphi} - \varphi_k\|_{L^r(Q)} = 0\ \ \forall 1 \le r < \frac{n+2}{n}; \tag{38}$$

see [7] for the proof.

Now, we define

$$\varphi_{M,k}(x,t) = \text{Proj}_{[-M,+M]}(\bar{\varphi}_k(x,t)).$$

Since $\varphi_k \in H^1(Q) \cap C([0,T]; H_0^1(\Omega))$, then $\varphi_{M,k}$ has the same regularity. From the $|\varphi_M(x,t) - \varphi_{M,k}(x,t)| \le |\bar{\varphi}(x,t) - \varphi_k(x,t)|$ and (38) we infer that $\varphi_{M,k} \to \varphi_M$ strongly in $L^r(Q)$ for every $1 \le r < \frac{n+2}{n}$. If we prove that $\{\varphi_{M,k}\}_{k=1}^\infty$ is bounded in $L^2(0,T; H_0^1(\Omega))$, then the convergence $\varphi_{M,k} \to \varphi_M$ in $L^r(Q)$ implies that $\varphi_M \in L^2(0,T; H_0^1(\Omega))$ as well. To prove this boundedness we multiply Eq. (36) by $\mathrm{e}^{-2C_a t}\varphi_{M.k}$, where C_a was introduced in (3). Then, we get

$$\begin{aligned}
&\int_Q -\mathrm{e}^{-2C_a t}\frac{\partial \varphi_k}{\partial t}\varphi_{M,k}\,dx\,dt + \sum_{i,j=1}^n \int_Q \mathrm{e}^{-2C_a t} a_{ij}\,\partial_{x_i}\varphi_k \partial_{x_j}\varphi_{M,k}\,dx\,dt \\
&+ \int_Q \frac{\partial a}{\partial y}(x,t,\bar{y})\mathrm{e}^{-2C_a t}\varphi_k\varphi_{M,k}\,dx\,dt \\
&= \int_Q \mathrm{e}^{-2C_a t}(\bar{y} - y_d)\varphi_{M,k}\,dx\,dt + \int_Q \mathrm{e}^{-2C_a t} f_k \varphi_{M,k}\,dx\,dt.
\end{aligned} \tag{39}$$

Now using that $\varphi_k\varphi_{M,k} \ge \varphi_{M,k}^2$ and $\varphi_k \partial_t \varphi_{M,k} = \varphi_{M,k}\partial_t\varphi_{M,k} = \frac{1}{2}\partial_t\varphi_{M,k}^2$, we obtain

$$\begin{aligned}
&\int_Q -\mathrm{e}^{-2C_a t}\frac{\partial \varphi_k}{\partial t}\varphi_{M,k}\,dx\,dt = -\int_0^T \frac{d}{dt}\int_\Omega \mathrm{e}^{-2C_a t}\varphi_k\varphi_{M,k}\,dx\,dt \\
&- 2C_a \int_Q \mathrm{e}^{-2C_a t}\varphi_k\varphi_{M,k}\,dx\,dt + \int_Q \mathrm{e}^{-2C_a t}\varphi_k \frac{\partial \varphi_{M,k}}{\partial t}\,dx\,dt \\
&= -\int_\Omega \mathrm{e}^{-2C_a T}\varphi_k(x,T)\varphi_{M,k}(x,T)\,dx + \int_\Omega \varphi_k(x,0)\varphi_{M,k}(x,0)\,dx \\
&- 2C_a \int_Q \mathrm{e}^{-2C_a t}\varphi_k\varphi_{M,k}\,dx\,dt + \frac{1}{2}\int_Q \mathrm{e}^{-2C_a t}\partial_t\varphi_{M,k}^2\,dx\,dt.
\end{aligned} \tag{40}$$

For the last term we have

$$\begin{aligned}
&\frac{1}{2}\int_Q \mathrm{e}^{-2C_a t}\partial_t\varphi_{M,k}^2\,dx\,dt\\
&=\frac{1}{2}\int_0^T\frac{d}{dt}\int_\Omega \mathrm{e}^{-2C_a t}\varphi_{M,k}^2\,dx\,dt + C_a\int_0^T\int_\Omega \mathrm{e}^{-2C_a t}\varphi_{M,k}^2\,dx\,dt\\
&\geq \frac{1}{2}\int_\Omega \mathrm{e}^{-2C_a T}\varphi_{M,k}^2(x,T)\,dx - \frac{1}{2}\int_\Omega \varphi_{M,k}^2(x,0)\,dx + C_a\int_Q \mathrm{e}^{-2C_a t}\varphi_k\varphi_{M,k}\,dx\,dt\\
&\geq -\frac{1}{2}\int_\Omega \varphi_{M,k}^2(x,0)\,dx + C_a\int_Q \mathrm{e}^{-2C_a t}\varphi_k\varphi_{M,k}\,dx\,dt. \qquad (41)
\end{aligned}$$

From (40) and (41) we deduce

$$\begin{aligned}
&\int_Q -\mathrm{e}^{-2C_a t}\frac{\partial\varphi_k}{\partial t}\varphi_{M,k}\,dx\,dt \geq -\mathrm{e}^{-2C_a T}\int_\Omega g_k(x)\varphi_{M,k}(x,T)\,dx\\
&+\frac{1}{2}\int_\Omega \varphi_{M,k}^2(x,0)\,dx - C_a\int_Q \mathrm{e}^{-2C_a t}\varphi_k\varphi_{M,k}\,dx\,dt\\
&\geq -\mathrm{e}^{-2C_a T}\int_\Omega g_k(x)\varphi_{M,k}(x,T)\,dx - C_a\int_Q \mathrm{e}^{-2C_a t}\varphi_k\varphi_{M,k}\,dx\,dt.
\end{aligned}$$

Inserting this inequality in (39) and using that $\partial_{x_i}\varphi_k\partial_{x_j}\varphi_{M,k} = \partial_{x_i}\varphi_{M,k}\partial_{x_j}\varphi_{M,k}$, we obtain with (2), Young's inequality, (34) and (35)

$$\begin{aligned}
&\Lambda\int_Q |\nabla\varphi_{M,k}|^2\,dx\,dt + \int_Q\Big[\frac{\partial a}{\partial y}(x,t,\bar{y}) - C_a\Big]\mathrm{e}^{-2C_a t}\varphi_k\varphi_{M,k}\,dx\,dt\\
&\leq \int_Q \mathrm{e}^{-2C_a t}(\bar{y}-y_d)\varphi_{M,k}\,dx\,dt + \int_Q \mathrm{e}^{-2C_a t}f_k\varphi_{M,k}\,dx\,dt + \mathrm{e}^{-2C_a T}\int_\Omega g_k\varphi_{M,k}(T)\,dx\\
&\leq \mathrm{e}^{-2C_a T}\Big[\|\bar{y}-y_d\|_{L^2(Q)}\|\varphi_{M,k}\|_{L^2(Q)} + M\big(\|f_k\|_{L^1(Q)} + \|g_k\|_{L^1(\Omega)}\big)\Big]\\
&\leq C\Big[\|\bar{y}-y_d\|_{L^2(Q)}^2 + M\big(\|\bar{\mu}_Q\|_{M(Q)} + \|\bar{\mu}_\Omega\|_{M(\Omega)}\big)\Big] + \frac{\Lambda}{2}\int_Q |\nabla\varphi_{M,k}|^2\,dx\,dt.
\end{aligned}$$

Finally, taking into account (3), we get from the above inequality that each $\varphi_{M,k}$ satisfies (33). Hence, φ_M also does it. □

Theorem 6 *Let $\bar{u} \in \mathbb{K}_{\alpha,\beta}$ satisfy (28)–(31). Then, $\bar{u} \in L^2(0,T;H^1(\Omega))$ and the inequality*

$$\|\bar{u}\|_{L^2(0,T;H^1(\Omega))} \leq C\big[\|\bar{y}-y_d\|_{L^2(Q)} + \sqrt{M_{\alpha,\beta}\big(\|\bar{\mu}_Q\|_{M(Q)} + \|\bar{\mu}_\Omega\|_{M(\Omega)}\big)}\big] \qquad (42)$$

holds, where $C = C(\Omega,\Lambda,C_a,\lambda) > 0$ and $M_{\alpha,\beta} = \max\{|\alpha|,|\beta|\}$.

Proof Let us take $M_{\alpha,\beta}$ as indicated in the statement of the theorem and set

$$\varphi_{M_{\alpha,\beta}}(x,t) = \text{Proj}_{[-M_{\alpha,\beta},+M_{\alpha,\beta}]}(\bar\varphi(x,t)).$$

Then, from Lemma 1 we know that $\varphi_{M_{\alpha,\beta}} \in L^2(0,T;H_0^1(\Omega))$ and

$$\|\varphi_{M_{\alpha,\beta}}\|_{L^2(0,T;H^1(\Omega))} \le C\big[\|\bar y - y_d\|_{L^2(Q)} + \sqrt{M_{\alpha,\beta}\big(\|\bar\mu_Q\|_{M(Q)} + \|\bar\mu_\Omega\|_{M(\Omega)}\big)}\big],$$

for a constant $C = C(\Omega,\Lambda,C_a,\lambda) > 0$. Now, from (31) we have

$$\bar u(x,t) = \text{Proj}_{[\alpha,\beta]}\big(-\frac{1}{\lambda}\bar\varphi(x,t)\big) = \text{Proj}_{[\alpha,\beta]}\big(-\frac{1}{\lambda}\varphi_{M_{\alpha,\beta}}(x,t)\big).$$

This implies that $\bar u \in L^2(0,T;H^1(\Omega))$ as well. Moreover, from the inequality

$$\|\bar u\|_{L^2(0,T;H^1(\Omega))} \le \|\varphi_{M_{\alpha,\beta}}\|_{L^2(0,T;H^1(\Omega))},$$

(42) follows. □

4 Convergence of the Numerical Approximations

In this section, we come back to the problem (P) formulated in Sect. 1 and assume that $-\infty < \alpha < \beta < +\infty$. This problem has at least a solution $\bar u$ for $\lambda \ge 0$. To compute an approximation of $\bar u$ we have to discretize the control problem. The goal in this section is to analyze the convergence of the approximations. The first difficulty of this analysis comes from the convergence of the discretization of the state equation. Here, the difficulty is due to the nonlinear term $a(x,t,y)$ and the low regularity of the solutions y. The main reference for that is [16]. Though the convergence analysis in [16] is carried out for two dimensional domains Ω, Boris Vexler has communicated me a modification of the proof to get a similar result in dimension 3. Using these results and assuming that $\lambda > 0$, then the strong convergence of the controls is proved in a standard way. The idea of the proof is the following. Let $\{u_k\}_{k=1}^\infty$ be a sequence of discrete optimal controls. Since every u_k satisfies the control constraints, the sequence is bound in $L^\infty(Q)$. Hence, we can take a subsequence, denoted in the same way, such that $u_k \overset{*}{\rightharpoonup} \bar u$ in $L^\infty(Q)$, for some control $\bar u$ satisfying the control constraints as well. Now, using [16], we get that the sequence of associated discrete states $\{y_k\}_{k=1}^\infty$ converges strongly to $\bar y$ in $L^2(Q)$, where $\bar y$ is the continuous state associated with $\bar u$. From these convergence properties and using the optimality of every u_k it is easy to prove that $\bar u$ is a solution of (P) and $J(u_k) \to J(\bar u)$. Since $\lambda > 0$, this convergence implies that $\|u_k\|_{L^2(Q)} \to \|\bar u\|_{L^2(Q)}$. This fact and the weak* convergence in $L^\infty(Q)$ imply the strong convergence $u_k \to \bar u$ in $L^2(Q)$. As a consequence of the boundedness of $\{u_k\}_{k=1}^{\infty}$ in $L^\infty(Q)$, we also deduce the strong convergence in $L^p(Q)$ for every $p < +\infty$.

The first part of the above argument can be repeated when $\lambda = 0$ and we obtain that $u_k \overset{*}{\rightharpoonup} \bar{u}$ in $L^\infty(Q)$ and $\bar{u}$ is a solution of (P). Obviously, the above argument to prove the strong convergence fails if $\lambda = 0$. As far as we know, the first result proving the strong convergence of the discrete controls when $\lambda = 0$ was given in [6]. We prove that the convergence of the optimal discrete controls to bang-bang optimal controls is strong. Though this is a very simple exercise, we have confirmed that most of the experts in the field had not realized about this property. The proof is an immediate consequence of the following proposition.

Proposition 1 *Let $\{u_k\}_{k=1}^{\infty}$ be a sequence satisfying $\alpha \le u_k(x,t) \le \beta$ for a.a. $(x,t) \in Q$, and $u_k \overset{*}{\rightharpoonup} \bar{u}$ in $L^\infty(Q)$. Assume that $\bar{u}$ is a bang-bang control. Then, the convergence $u_k \to \bar{u}$ strongly in $L^p(Q)$ holds for every $p < +\infty$.*

Proof Let us denote

$$Q_\alpha = \{(x,t) \in Q : \bar{u}(x,t) = \alpha\} \text{ and } Q_\beta = \{(x,t) \in Q : \bar{u}(x,t) = \beta\}.$$

Since, $\bar{u}$ is a bang-bang control, we have that $|Q| = |Q_\alpha| + |Q_\beta|$, where $|\cdot|$ denotes the Lebesgue measure. Hence, we deduce from the weak* convergence in $L^\infty(Q)$

$$\int_Q |\bar{u} - u_k|\,dx\,dt = \int_{Q_\alpha} (u_k - \bar{u})\,dx\,dt + \int_{Q_\beta} (\bar{u} - u_k)\,dx\,dt$$
$$= \int_Q \chi_{Q_\alpha}(u_k - \bar{u})\,dx\,dt + \int_Q \chi_{Q_\beta}(\bar{u} - u_k)\,dx\,dt \to 0,$$

where χ_{Q_α} and χ_{Q_β} denote the characteristic functions of Q_α and Q_β, respectively. This proves the strong convergence in $L^1(Q)$. Finally, it is enough to observe that

$$\int_Q |\bar{u} - u_k|^p\,dx\,dt \le (\beta - \alpha)^{p-1} \int_Q |\bar{u} - u_k|\,dx\,dt \to 0$$

to conclude the proof. □

5 Second Order Analysis

In this section, we give sufficient second order conditions for local optimality. The main goal is to show the difference between the cases $\lambda > 0$ and $\lambda = 0$. First, let us recall some issues concerning the second order analysis in infinite dimensional spaces. The material presented in this section is based on the papers [8] and [10]; see also [9].

It is well known that second order optimality conditions are an important tool in the numerical analysis of optimization problems. They are essential in proving superlinear or quadratic convergence of numerical algorithms, in deriving error estimates for the numerical discretization of infinite-dimensional optimization problems or just for the proof of local uniqueness of optimal solutions. Although there is an extensive literature on second order optimality conditions, there are still some open problems.

A study of the existing theory of first order optimality conditions reveals that the situation for finite-dimensional problems is very close to the infinite-dimensional one. However, there are big differences when we look at sufficient second order conditions. Let us mention some of these differences.

Consider a differentiable functional $J : U \longrightarrow \mathbb{R}$, where U is a Banach space. If $\bar{u}$ is a local minimum of J, then we know that $J'(\bar{u}) = 0$. This is a necessary condition. If J is not convex, we have to invoke a sufficient condition and should study the second derivative. In the finite-dimensional case, say $U = \mathbb{R}^n$, the first order optimality condition $J'(\bar{u}) = 0$ and the second order condition $J''(\bar{u})v^2 > 0$ for every $v \in U \setminus \{0\}$ imply that $\bar{u}$ is a strict local minimum of J. This second order condition says that the quadratic form $v \to J''(\bar{u})v^2$ is positive definite in $\mathbb{R}^n$, which is equivalent to the strict positivity of the smallest eigenvalue δ_m of the associated symmetric matrix. Moreover, the inequality $J''(\bar{u})v^2 \geq \delta_m \|v\|^2$ for every $v \in \mathbb{R}^n$ holds.

However, if U is an infinite-dimensional space, then the condition $J''(\bar{u})v^2 > 0$ is not equivalent to $J''(\bar{u})v^2 \geq \delta_m \|v\|^2$ for some $\delta_m > 0$. Is one of the two conditions sufficient for local optimality? The next example shows that the first condition is not sufficient for local optimality.

Example 1 Consider the optimization problem

$$(\mathrm{Ex}_1) \quad \min_{u \in L^\infty(0,1)} J(u) = \int_0^1 [tu^2(t) - u^3(t)]\, dt.$$

The function $\bar{u}(t) \equiv 0$ satisfies the first-order necessary condition $J'(\bar{u}) = 0$ and

$$J''(\bar{u})v^2 = \int_0^1 2tv^2(t)\, dt > 0 \ \ \forall v \in L^\infty(0,1) \setminus \{0\}.$$

However, $\bar{u}$ is not a local minimum of (Ex_1). Indeed, if we define

$$u_k(t) = \begin{cases} 2t \text{ if } t \in (0, \dfrac{1}{k}), \\ 0 \ \text{ otherwise,} \end{cases}$$

then it holds $J(u_k) = -\frac{1}{k^4} < J(\bar{u})$, and $\|u_k - \bar{u}\|_{L^\infty(0,1)} = \frac{2}{k}$.

However, it is well known that if J is of class C^2 in a neighborhood of $\bar{u}$, then the condition $J''(\bar{u})v^2 \geq \delta\|v\|^2$ $\forall v \in U$ with $\delta > 0$ is a sufficient condition for local optimality. This seems to solve completely the issue. Nevertheless, this conditions is not so simple in infinite dimensional optimization problems. Let us consider the following example.

Example 2 We discuss the optimization problem

$$(\mathrm{Ex}_2) \quad \min_{u \in L^2(0,1)} J(u) = \int_0^1 \sin(u(t))\, dt.$$

Obviously, $\bar{u}(t) \equiv -\pi/2$ is a global solution. Some fast but formal computations lead to

$$J'(\bar{u})v = \int_0^1 \cos(\bar{u}(t))v(t)\, dt = 0 \ \text{ and}$$

$$J''(\bar{u})v^2 = -\int_0^1 \sin(\bar{u}(t))v^2(t)\, dt = \int_0^1 v^2(t)\, dt = \|v\|^2_{L^2(0,1)} \ \forall v \in L^2(0,1).$$

If the second, stronger condition were sufficient for local optimality, $\bar{u}$ would be strict local minimum of (Ex$_2$). However, this is not true. Indeed, for every $0 < \varepsilon < 1$, the functions

$$u_\varepsilon(t) = \begin{cases} -\dfrac{\pi}{2} & \text{if } t \in [0, 1-\varepsilon], \\[2ex] +\dfrac{3\pi}{2} & \text{if } t \in (1-\varepsilon, 1], \end{cases}$$

are also global solutions of (Ex$_2$), with $J(\bar{u}) = J(u_\varepsilon)$ and $\|\bar{u} - u_\varepsilon\|_{L^2(0,1)} = 2\pi\sqrt{\varepsilon}$. Therefore, infinitely many different global solutions of (Ex$_2$) are contained in any L^2-neighborhood of $\bar{u}$ and $\bar{u}$ is not a strict solution.

What is wrong? The reason is that J is not of class C^2 in $L^2(0,1)$, our fast computations was too careless. Therefore we cannot apply the abstract theorem on sufficient conditions for local optimality in $L^2(0,1)$. On the other hand, J is of class C^2 in $L^\infty(0,1)$ and the derivatives computed above are correct in $L^\infty(0,1)$. However, the inequality $J''(\bar{u})v^2 \geq \delta\|v\|^2_{L^\infty(0,1)}$ does not hold for any $\delta > 0$.

This phenomenon is called the *two-norm discrepancy*: the functional J is twice differentiable with respect to one norm, but the inequality $J''(\bar{u})v^2 \geq \delta\|v\|^2$ holds in a weaker norm in which J is not twice differentiable; see, for instance, [14]. This situation arises frequently in infinite-dimensional problems but it does not happen for finite-dimensions because all the norms are equivalent in this case. The classical theorem on second order optimality conditions can easily be modified to deal with the two norm-discrepancy.

Theorem 7 *Let U be a vector space endowed with two norms, $\| \|_\infty$ and $\| \|_2$, such that $J : (U, \| \|_\infty) \mapsto \mathbb{R}$ is of class C^2 in a neighborhood of $\bar u$ and the following properties hold:*

$$J'(\bar u) = 0 \quad \textit{and} \ \exists \delta > 0 \ \textit{such that} \ J''(\bar u)v^2 \geq \delta \|v\|_2^2 \ \forall v \in U, \tag{43}$$

and there exists some $\varepsilon > 0$ such that

$$|J''(\bar u)v^2 - J''(u)v^2| \leq \frac{\delta}{2}\|v\|_2^2 \ \forall v \in U \ \ \textit{if} \ \ \|u - \bar u\|_\infty \leq \varepsilon. \tag{44}$$

Then, there holds

$$\frac{\delta}{4}\|u - \bar u\|_2^2 + J(\bar u) \leq J(u) \ \ \textit{if} \ \ \|u - \bar u\|_\infty \leq \varepsilon \tag{45}$$

so that $\bar u$ is a strictly locally optimal with respect to the norm $\| \cdot \|_\infty$.

The proof of this theorem is quite elementary.

Coming back to the control problem (P), we observe that the cost functional, in general, is not of class C^2 in $L^2(Q)$. Hence, the two-norm discrepancy appears in this case. As a consequence, for long time, it was believed that a second order condition of type $J''(\bar u)v^2 \geq \delta\|v\|^2_{L^2(Q)}$ implied a strict local optimality of $\bar u$ in $L^\infty(Q)$. This is a serious drawback in the numerical analysis of (P). More recently, it was proved in [8] that, under the assumption $\lambda > 0$, this condition implies the strict local optimality in $L^2(Q)$ as well; see [2] for a previous, but weaker result, in the case of elliptic control problems. However, the situation is completely different if $\lambda = 0$. Here, we show the difference in the second order analysis between the cases $\lambda > 0$ and $\lambda = 0$. We advance that the second order analysis of (P) with $\lambda > 0$ behaves essentially as the analysis for finite-dimensional optimization problems.

Before a correct formulation of the second order conditions for (P), we need to introduce the cone of critical directions. Given $\bar u$ a feasible control satisfying the first order optimality conditions (12)–(14), we define the cone of critical directions

$$C_{\bar u} = \left\{ v \in L^2(Q) : v(x,t) \begin{cases} \geq 0 \text{ if } \bar u(x,t) = \alpha, \\ \leq 0 \text{ if } \bar u(x,t) = \beta, \\ = 0 \text{ if } (\bar\varphi + \lambda \bar u)(x,t) \neq 0, \end{cases} \quad \text{a.e. in } Q \right\}.$$

It is not difficult to prove the necessary second order condition for local optimality $J''(\bar u)v^2 \geq 0 \ \forall v \in C_{\bar u}$; see, for instance, [3, Section §3.2] or [8]. This condition holds independently of the case $\lambda = 0$ or $\lambda > 0$. The differences appear in the statement for the second order sufficient conditions.

5.1 Case $\lambda > 0$

We start with the main theorem.

Theorem 8 *Let us assume that* $\bar u$ *is a feasible control for problem (P) satisfying the first order optimality conditions* (12)–(14)*. We also assume that* $\lambda > 0$*. If the condition* $J''(\bar u)v^2 > 0\ \forall v \in C_{\bar u} \setminus \{0\}$ *holds, then there exist* $\varepsilon > 0$ *and* $\kappa > 0$ *such that*

$$J(\bar u) + \frac{\kappa}{2}\|u - \bar u\|^2_{L^2(Q)} \le J(u) \quad \forall u \in \mathbb{K}_{\alpha,\beta} \cap \bar B_\varepsilon(\bar u), \tag{46}$$

where $\bar B_\varepsilon(\bar u)$ *denotes the* $L^2(Q)$*-ball centered at* $\bar u$ *and radius* ε*.*

Proof The reader is referred to, for instance, [8] for a detailed proof. Here, we present a sketch of the proof in order to show the role played by the Tikhonov parameter λ. We argue by contradiction, as in the finite dimensional case. If (46) does not hold, for every integer $k \ge 1$ we deduce the existence of $u_k \in \mathbb{K}_{\alpha,\beta}$ such that

$$\|\bar u - u_k\|_{L^2(Q)} < \frac{1}{k} \ \text{ and } \ J(\bar u) + \frac{1}{2k}\|\bar u - u_k\|^2_{L^2(Q)} > J(u_k) \quad \forall k \ge 1. \tag{47}$$

We set $\rho_k = \|\bar u - u_k\|_{L^2(Q)}$ and $v_k = \frac{1}{\rho_k}\|\bar u - u_k\|_{L^2(Q)}$. By taking a subsequence if necessary, we can assume that $v_k \rightharpoonup v$ in $L^2(Q)$. The proof is split into three steps.

Step 1: $v \in C_{\bar u}$ It is obvious that the set

$$S = \left\{v \in L^2(Q) : v(x,t) \begin{cases} \ge 0 \text{ if } \bar u(x,t) = \alpha, \\ \le 0 \text{ if } \bar u(x,t) = \beta, \end{cases} \text{a.e. in } Q\right\}$$

is convex and closed in $L^2(Q)$. Moreover, since $\alpha \le u_k(x,t) \le \beta$ a.e. in Q $\forall k \ge 1$, we deduce that $\{v_k\}_{k=1}^\infty \subset S$. Hence, $v \in S$ holds. It remains to prove that $v(x,t) = 0$ if $(\bar\varphi + \lambda\bar u)(x,t) \ne 0$ a.e. in Q. First, we observe that (14) implies that

$$\bar u(x,t) = \begin{cases} \alpha \text{ if } (\bar\varphi + \lambda\bar u)(x,t) > 0, \\ \beta \text{ if } (\bar\varphi + \lambda\bar u)(x,t) < 0, \end{cases} \text{for a.a.}(x,t) \in Q. \tag{48}$$

This implies that $(\bar\varphi + \lambda\bar u)(x,t)w(x,t) \ge 0$ a.e. in Q $\forall w \in S$. Therefore it also holds for $w = v$. Now, from (47) we infer

$$\frac{J(\bar u + \rho_k v_k) - J(\bar u)}{\rho_k} = \frac{J(u_k) - J(\bar u)}{\rho_k} < \frac{1}{2k}\|u_k - \bar u\|_{L^2(Q)}.$$

Passing to the limit when $k \to \infty$ we get

$$\int_Q |\bar\varphi + \lambda\bar u||v|\,dx\,dt = \int_Q (\bar\varphi + \lambda\bar u)v\,dx\,dt = \lim_{k\to\infty} \frac{J(\bar u + \rho_k v_k) - J(\bar u)}{\rho_k} \le 0.$$

This concludes the proof of $v \in C_{\bar u}$.

Step 2: $J''(\bar u)v^2 \le 0$ Using again (47), the fact that $J'(\bar u)w \ge 0\ \forall w \in C_{\bar u} \subset S$, and making a Taylor expansion of $J(u_k)$ around $\bar u$, we get

$$\begin{aligned}\frac{\rho_k^2}{2k} &= \frac{1}{2k}\|u_k - \bar u\|^2_{L^2(Q)} > J(u_k) - J(\bar u) = J(\bar u + \rho_k v_k) - J(\bar u)\\ &= \rho_k J'(\bar u)v_k + \frac{\rho_k^2}{2}J''(\bar u + \theta_k\rho_k v_k)v_k^2 \ge \frac{\rho_k^2}{2}J''(\bar u + \theta_k(u_k - \bar u))v_k^2.\end{aligned}$$

Dividing the above inequality by $\frac{\rho_k^2}{2}$ we obtain: $J''(\bar u + \theta_k(u_k - \bar u))v_k^2 < 1/k$. Hence, passing to the limit when $k \to \infty$, we conclude that $J''(\bar u)v^2 \le 0$.

Step 3: Contradiction The second order condition $J''(\bar u)v^2 > 0\ \forall v \in C_{\bar u} \setminus \{0\}$ along with the proved steps 1 and 2 implies that $v_k \rightharpoonup v = 0$. Let us set $\hat u_k = \bar u + \theta_k(u_k - \bar u)$, $\hat y_k$ its associated state, $\hat\varphi_k$ the corresponding adjoint state, and $\hat z_k$ the solution of (7), where u is replaced by $\hat u_k$ and v by v_k. The weak convergence $v_k \rightharpoonup v$ in $L^2(Q)$ implies the strong convergence $\hat z_k \to 0$ in $L^2(Q)$. Then, we deduce from (10) and the fact that $\|v_k\|_{L^2(Q)} = 1\ \forall k \ge 1$

$$0 < \lambda = \lim_{k\to\infty}\int_Q \Big[\big(1 - \hat\varphi_k\frac{\partial^2 a}{\partial y^2}(x,t,\hat y_k)\big)\hat z_k^2 + \lambda v_k^2\Big]\,dx\,dt = \lim_{k\to\infty} J''(\hat u_k)v_k^2 = J''(\bar u)v^2 \le 0,$$

which is a contradiction. □

Observe that it was not required a second order condition of type $J''(\bar u)v^2 \ge \delta\|v\|^2_{L^2(Q)}\ \forall v \in C_{\bar u}$ as one can expect for an infinite dimensional optimization problem. The issue is that this second order condition is equivalent to $J''(\bar u)v^2 > 0\ \forall v \in C_{\bar u}$. Once again, this equivalence is valid just because $\lambda > 0$. In fact the following result holds (see [8, 9]).

Theorem 9 *Let us assume that* $\lambda > 0$. *Then, the following statements are equivalent*

$$J''(\bar u)v^2 > 0\ \forall v \in C_{\bar u} \setminus \{0\}, \tag{49}$$

$$\exists\delta > 0 \text{ and } \tau > 0 \text{ such that } J''(\bar u)v^2 \ge \delta\,\|v\|^2_{L^2(Q)}\ \ \forall v \in C^\tau_{\bar u}, \tag{50}$$

$$\exists\delta > 0 \text{ and } \tau > 0 \text{ such that } J''(\bar u)v^2 \ge \delta\|z_v\|^2_{L^2(Q)}\ \ \forall v \in C^\tau_{\bar u}, \tag{51}$$

where z_v is the solution of (7) *corresponding to $y_u = \bar{y}$ and*

$$C_{\bar{u}}^{\tau} = \left\{ v \in L^2(Q) : v(x,t) \begin{cases} \geq 0 \textit{ if } \bar{u}(x,t) = \alpha, \\ \leq 0 \textit{ if } \bar{u}(x,t) = \beta, \\ = 0 \textit{ if } |(\bar{\varphi} + \lambda \bar{u})(x,t)| \geq \tau, \end{cases} \quad \textit{a.e. in } Q \right\}.$$

Observe that $C_{\bar{u}}$ is strictly contained in $C_{\bar{u}}^{\tau}$ $\forall \tau > 0$. Therefore (50) seems to be a stronger condition that the usual one: $J''(\bar{u}) \geq \delta \|v\|_{L^2(Q)}^2$ $\forall v \in C_{\bar{u}}$. But, actually they are equivalent as follows from the previous theorem.

5.2 Case $\lambda = 0$

When $\lambda = 0$, the proof of Theorem 8 fails precisely at the last step. There is no way to get the contradiction. In this case, Theorem 9 is also false. Since, the condition (49) is not enough to prove that $\bar{u}$ is a local minimum, one can try to check if the condition (50) is sufficient for a local optimality. However, looking at the expression of $J''(\bar{u})v^2$

$$J''(u)v^2 = \int_Q \Big[\Big(1 - \bar{\varphi} \frac{\partial^2 a}{\partial y^2}(x,t,\bar{y})\Big) z_v^2 \Big] \, dx \, dt,$$

this condition seems quite difficult to be fulfilled. In fact, it was proved in [5] that it does not hold, except maybe in a few extreme cases. Finally, the condition (51) makes sense if we compare with the second derivative $J''(\bar{u})v^2$. In [5], it was proved that (51) is a sufficient second order optimality condition. More precisely, assuming that $\bar{u} \in \mathbb{K}_{\alpha,\beta}$ satisfies (12)–(14) and (51), then there exist $\varepsilon > 0$ and $\kappa > 0$ such that

$$J(\bar{u}) + \frac{\kappa}{2} \|y_u - \bar{y}\|_{L^2(Q)}^2 \leq J(u) \quad \forall u \in \mathbb{K}_{\alpha,\beta} \cap B_\varepsilon(\bar{u}), \tag{52}$$

where $B_\varepsilon(\bar{u})$ denotes again the $L^2(Q)$-ball.

Unlike (46), the above inequality does not allow to prove, in general, neither stability of the optimal control with respect to perturbations of the data of the control problem, nor we can derive error estimates in the control for the numerical approximation. However, they are useful to prove stability of the states or to get error estimates for the states.

The main drawback of the condition (51) is that the gap with the necessary second order optimality condition is big. However, for $\lambda > 0$, this gap is minimal, in fact, the same as in finite dimension. Recently, some results have been obtained for bang-bang controls where the gap is smaller; see [13]. However, the problem is not completely solved and some research is in progress.

Acknowledgements The author was partially supported by Ministerio Español de Economía, Industria y Competitividad under the research projects MTM2014-57531-P and MTM2017-83185-P.

References

1. Ali, A.A., Deckelnick, K., Hinze, M.: Global minima for semilinear optimal control problems. Comput. Optim. Appl. **65**, 261–288 (2016)
2. Bonnans, J.F.: Second-order analysis for control constrained optimal control problems of semilinear elliptic systems. Appl. Math. Optim. **38**, 303–325 (1998)
3. Bonnans, F., Shapiro, A.: Perturbation Analysis of Optimization Problems. Springer, Berlin (2000)
4. Casas, E.: Pontryagin's principle for state-constrained boundary control problems of semilinear parabolic equations. SIAM J. Control Optim. **35**(4), 1297–1327 (1997)
5. Casas, E.: Second order analysis for bang-bang control problems of PDEs. SIAM J. Control Optim. **50**(4), 2355–2372 (2012)
6. Casas, E., Chrysafinos, K.: Error estimates for the approximation of the velocity tracking problem with bang-bang controls. ESAIM Control Optim. Calc. Var. **23**, 1267–1291 (2017)
7. Casas, E., Kunisch, K.: Parabolic control problems in space-time measure spaces. ESAIM Control Optim. Calc. Var. **22**(2), 355–370 (2016)
8. Casas, E., Tröltzsch, F.: Second order analysis for optimal control problems: improving results expected from abstract theory. SIAM J. Optim. **22**(1), 261–279 (2012)
9. Casas, E., Tröltzsch, F.: Second order optimality conditions and their role in PDE control. Jahresber. Dtsch. Math. Ver. **117**(1), 3–44 (2015)
10. Casas, E., Tröltzsch, F.: Second order optimality conditions for weak and strong local solutions of parabolic optimal control problems. Vietnam J. Math. **44**(1), 181–202 (2016)
11. Casas, E., Raymond, J.P., Zidani, H.: Pontryagin's principle for local solutions of control problems with mixed control-state constraints. SIAM J. Control Optim. **39**(4), 1182–1203 (2000)
12. Casas, E., Mateos, M., Rösch, A.: Finite element approximation of sparse parabolic control problems. Math. Control Relat. Fields **7**(3), 393–417 (2017)
13. Casas, E., Wachsmuth, D., Wachsmuth, G.: Sufficient second-order conditions for bang-bang control problems. SIAM J. Control Optim. **55**(5), 3066–3090 (2017)
14. Ioffe, A.D.: Necessary and sufficient conditions for a local minimum. III. Second order conditions and augmented duality. SIAM J. Control Optim. **17**(2), 266–288 (1979). MR 525027 (82j:49005c)
15. Ladyzhenskaya, O.A., Solonnikov, V.A., Ural'tseva, N.N.: Linear and Quasilinear Equations of Parabolic Type. American Mathematical Society, Providence (1988)
16. Neitzel, I., Vexler, B.: A priori error estimates for space-time finite element discretization of semilinear parabolic optimal control problems. Numer. Math. **120**, 345–386 (2012)
17. Raymond, J.P., Zidani, H.: Pontryagin's principle for state-constrained control problems governed by parabolic equations with unbounded controls. SIAM J. Control Optim. **36**(6), 1853–1879 (1998)
18. Rudin, W.: Real and Complex Analysis. McGraw-Hill, London (1970)
19. Showalter, R.E.: Monotone Operators in Banach Space and Nonlinear Partial Differential Equations. Mathematical Surveys and Monographs, vol. 49. American Mathematical Society, Providence (1997). MR 1422252 (98c:47076)
20. Tröltzsch, F.: Optimal Control of Partial Differential Equations: Theory, Methods and Applications. Graduate Studies in Mathematics, vol. 112. American Mathematical Society, Philadelphia (2010)

On the Design of Algebraic Fractional Step Methods for Viscoelastic Incompressible Flows

Ramon Codina

Dedicated to Prof. Enrique Fernández-Cara on the occasion of his 60th birthday.

Abstract Classical fractional step methods for viscous incompressible flows aim to uncouple the calculation of the velocity and the pressure. In the case of viscoelastic flows, a new variable appears, namely, a stress, which has an elastic and a viscous contribution. The purpose of this article is to present two families of fractional step methods for the time integration of this type of flows whose objective is to permit the uncoupled calculation of velocities, stresses and pressure, both families designed at the algebraic level. This means that the splitting of the equations is introduced once the spatial and the temporal discretizations have been performed. The first family is based on the extrapolation of the pressure and the stress in order to predict a velocity, then the calculation of a new stress, the pressure and then a correction to render the scheme stable. The second family has a discrete pressure Poisson equation as starting point; in this equation, velocities and stresses are extrapolated to compute a pressure, and from this pressure stresses and velocities can then be computed. This work presents an overview of methods previously proposed in our group, as well as some new schemes in the case of the second family.

Keywords Fractional step schemes · Viscoelastic flows · Pressure extrapolation · Velocity extrapolation · Inexact factorization

R. Codina (✉)
Universitat Politècnica de Catalunya, Barcelona, Spain
e-mail: ramon.codina@upc.edu

A. Doubova et al. (eds.), *Recent Advances in PDEs: Analysis, Numerics and Control*, SEMA SIMAI Springer Series 17,
https://doi.org/10.1007/978-3-319-97613-6_6

1 Introduction

From the computational point of view, the key aspect in the complexity of the approximation of the incompressible Navier-Stokes equations is the coupling between the velocity and the pressure degrees of freedom. Apart from the difficulties in choosing a spatial interpolation for both variables that renders the final scheme stable, once the discrete problem needs to be solved one has to face with unknowns with different behavior from the standpoint of algebraic solvers. In incompressible flows, it is usually the pressure the variable that drives the whole iterative behavior of linear solvers, and it is certainly a waste of effort that the velocity be dragged in this process as a coupled variable. Moreover, special solvers with special preconditioners could be used for the pressure if it could solved in an uncoupled manner.

The interest in fractional step methods in incompressible flows, also known as splitting methods, started with the works of Chorin [6] and Temam [16], who attempted the uncoupling of velocity and pressure at the continuous level, segregating the calculation of the pressure from the momentum equation and then understanding the final velocity correction as a projection onto the space of solenoidal fields. Since then, many works have been devoted to a proper understanding of the original schemes, their numerical analysis, their extension to order higher than one in time and to the design of adequate boundary conditions. The reader is referred to the survey [11] for the description of all these works.

There is also the possibility to look at the problem from the purely algebraic point of view, when the equations have already been discretized in space and in time. This way to approach the problem emerged after the identification in [13] of the classical pressure segregation method as an inexact factorization of the system arising after discretization. Several authors followed this path; for a review, see [1]. This point of view has clear advantages, as for example its generality or the fact that it avoids any issue related to boundary conditions, but also some inconveniences from the convergence point of view, since estimates depend on derivatives of discrete functions whose boundedness is not easy to prove.

In the case of viscoelastic flows, the main difficulty is the appearance of a new variable, a stress, that evolves in time. Thus, there are three variables (velocity, stress and pressure) that are in principle coupled and for which uncoupling algorithms need to be devised. Obviously, the uncoupling needs to satisfy two main conditions: it has to maintain the stability of the underlying time discretization (otherwise, a simple explicit treatment of adequate terms would suffice) and it has to maintain also its temporal order of accuracy. Surprisingly, even though several fractional schemes have been proposed for this problem (see for example [15], perhaps one of the first attempts), they either do not uncouple all the variables or are not natural extensions of the most popular schemes used for viscous Newtonian flows; see for example the bibliography cited in [5].

In [5] we proposed fractional step methods for viscoelastic flows based on the segregation of the pressure and the stress in the momentum equation. The approach proposed there is completely algebraical, working with the problem arising from spatial and temporal discretization of the original initial and boundary value problem. We designed schemes of first, second and third order in time, and all motivated from two perspectives: either the extrapolation in time of variables to allow their segregation or the inexact factorization of the linear system to be solved at each time step. All the schemes were tested in convergence tests, to check the predicted order of accuracy, and in more realistic examples to experiment their robustness.

The purpose of this article is to present some fractional step methods for viscoelastic flows designed from the pure algebraic point of view. Two families of approaches will be described. The first is the same as in [5], considering pressure (and stress) extrapolations to allow for the calculation of an intermediate velocity, whereas the second is based on the extrapolation of the velocity to allow for the calculation of the pressure. This second new approach is based in the design of fractional step schemes based on a discrete pressure Poisson equation that was proposed for viscous Newtonian flows in [2, 12].

The spatial approximation will not be discussed in detail. To fix ideas, we will describe how the approximation can be done using the finite element method using inf-sup stable approximations, although we favor the stabilized finite element approximation presented in [4]; minor modifications to the schemes to be described need to be introduced in case this stabilized formulation is used. Likewise, we will assume that the temporal discretization is performed using backward difference (BDF) schemes, although any other time integration could be employed.

The outline of the paper is as follows. In Sect. 2 we state the continuous problem, its finite element approximation in space and its numerical integration in time. In Sect. 3 we describe the schemes based on pressure extrapolation proposed in [5], whereas in Sect. 4 we present new schemes based on velocity extrapolation. Even though our objective is not the numerical analysis of the resulting methods, but only their design, some comments on their stability are also included in Sect. 5. In Sect. 6 we also explain how to view the schemes as inexact factorizations of the fully discrete system. Finally, some conclusions are drawn in Sect. 7.

2 Problem Statement and Numerical Approximation

Let Ω be a bounded domain of $\mathbb{R}^d$ $(d = 2, 3)$ where the flow takes place, and let $[0, t_{\mathrm{f}}[$ be the time interval of analysis. The viscoelastic (Oldroyd-B) flow problem we wish to consider consists of finding a velocity $u : \Omega \times]0, t_{\mathrm{f}}[\rightarrow \mathbb{R}^d$, a pressure $p : \Omega \times]0, t_{\mathrm{f}}[\rightarrow \mathbb{R}$ and a stress $\sigma : \Omega \times]0, t_{\mathrm{f}}[\rightarrow \mathbb{R}^d \otimes \mathbb{R}^d$ such that

$$\rho \frac{\partial u}{\partial t} + \rho u \cdot \nabla u - \nabla \cdot T + \nabla p = f \tag{1}$$

$$\nabla \cdot u = 0 \tag{2}$$

with $T = 2\beta\eta_0\nabla^s u + \sigma$ and

$$\frac{\lambda}{2\eta_0}\frac{\partial\sigma}{\partial t} + \frac{1}{2\eta_0}\sigma - (1-\beta)\,\nabla^s u + \frac{\lambda}{2\eta_0}\Big(u\cdot\nabla\sigma - \sigma\cdot\nabla u - (\nabla u)^T\cdot\sigma\Big) = 0 \tag{3}$$

In these equations, which hold in $\Omega\times]0, t_{\mathrm{f}}[$, f is the body force, ρ the fluid density, β, η_0 and λ are positive physical parameters ($0 \le \beta \le 1$) and ∇^s denotes the symmetric part of the gradient of a vector field. Appropriate initial and boundary conditions need to be added to close the problem (see [10], for example).

To write the weak form of the problem, let $\mathscr{V}$, $\mathscr{Q}$ and Υ be the spaces where velocities, pressures and stresses, respectively, have to belong for each $t \in]0, t_{\mathrm{f}}[$. Considering for example homogeneous velocity boundary conditions, $\mathscr{V} = H_0^1(\Omega)^d$, $\mathscr{Q} = L^2(\Omega)/\mathbb{R}$ and Υ is the space of tensor fields with components in $L^2(\Omega)$, such that the last term in parenthesis in (3) has components in $L^2(\Omega)$ and satisfying the appropriate boundary conditions. Let $(\cdot,\cdot)$ denote the inner product in $L^2(\Omega)$ (for scalars, vectors or tensors) and $\langle\cdot,\cdot\rangle$ the integral of the product of two functions. The weak form of the problem consists then of finding $[u, p, \sigma] :]0, t_{\mathrm{f}}[\to \mathscr{X} := \mathscr{V}\times\mathscr{Q}\times\Upsilon$ such that the initial conditions are satisfied and

$$\left(\rho\frac{\partial u}{\partial t}, v\right) + 2\left(\beta\eta_0\nabla^s u, \nabla^s v\right) + \langle\rho u\cdot\nabla u, v\rangle + \left(\sigma, \nabla^s v\right) - (p, \nabla\cdot v) = \langle f, v\rangle \tag{4}$$

$$(q, \nabla\cdot u) = 0 \tag{5}$$

$$\left(\frac{\lambda}{2\eta_0}\frac{\partial\sigma}{\partial t}, \tau\right) + \left(\frac{1}{2\eta_0}\sigma, \tau\right) - \left((1-\beta)\,\nabla^s u, \tau\right) + \frac{\lambda}{2\eta_0}\Big(u\cdot\nabla\sigma - \sigma\cdot\nabla u - (\nabla u)^T\cdot\sigma, \tau\Big) = 0 \tag{6}$$

for all $[v, q, \tau] \in \mathscr{X}$, where it is assumed that f is such that $\langle f, v\rangle$ is well defined.

The fractional step schemes to be presented can be used in conjunction with any space discretization. For the sake of conciseness, suppose that the finite element method is used. From a finite element partition of the computational domain Ω we may construct conforming finite element subspaces of $\mathscr{V}$, $\mathscr{Q}$ and Υ, that we respectively denote by $\mathscr{V}_h$, $\mathscr{Q}_h$ and Υ_h, the subscript h referring to the size of the partition. We assume that these spaces render a stable approximation in space, a point that turns out to be crucial and poses stringent requirements on the choice of the finite element spaces (in the form of two inf-sup conditions). This can be circumvented by using a stabilized finite element method, in which the discrete variational form of the problem is modified with respect to the continuous form, and therefore also the final algebraic system presented below is modified. Nevertheless, since the spatial approximation is not our focus, we assume hereafter that the so-called standard Galerkin method is used and refer to [4, 8] for further discussion.

Once $\mathscr{X}$ has been approximated by $\mathscr{X}_h := \mathscr{V}_h \times \mathscr{Q}_h \times \Upsilon_h$, the unknowns and test functions can be expressed as a combination of the basis functions of each space and the arrays of nodal values. We shall respectively denote the nodal values of u_h, p_h and σ_h as U, P and Σ; these arrays are time-dependent functions before the time discretization.

Considering the time discretization prior to the splitting, any alternative could be used. To fix ideas, and to simplify the notation, we will assume that backward difference schemes (BDF) of order $k \geq 1$ are used. Let us consider a uniform partition of the interval $[0, t_{\mathrm{f}}]$ of size δt, and let us denote with a superscript the time step level at which functions are approximated. A BDF scheme of order k is based on the k-th difference of a function, which when evaluated at $t^{n+1} = (n+1)\delta t$ reads

$$\delta_k g^{n+1} = \frac{1}{\gamma_k}\left(g^{n+1} - \sum_{i=0}^{k-1}\varphi_k^i g^{n-i}\right) =: \frac{1}{\gamma_k} g^{n+1} - g^{*,n}$$

for a generic function g, and where γ_k and φ_k^i are parameters that depend on k. In particular, we will be interested in the cases $k = 1, 2, 3$.

We will also use the extrapolation operators of order k, defined as $\widehat{g}_k^{n+1} = g^{n+1} + \mathscr{O}(\delta t^k)$, which for $k = 1, 2$ and 3 are given by

$$\begin{aligned} \widehat{g}_1^{n+1} &= g^n \\ \widehat{g}_2^{n+1} &= 2g^n - g^{n-1} \\ \widehat{g}_3^{n+1} &= 3g^n - 3g^{n-1} + g^{n-2} \end{aligned}$$

As for the k-th difference of a function, proper initializations are required in the first time steps.

Assuming space is discretized using the standard Galerkin method and time using a BDF scheme of order k, the resulting algebraic structure of the approximation to problem (4)–(6) is

$$M_u \frac{\delta_k}{\delta t} U^{n+1} + K_u\left(U^{n+1}\right) U^{n+1} + G_u P^{n+1} - D_\sigma \Sigma^{n+1} = F^{n+1} \tag{7}$$

$$D_u U^{n+1} = 0 \tag{8}$$

$$M_\sigma \frac{\delta_k}{\delta t} \Sigma^{n+1} + K_\sigma\left(U^{n+1}\right) \Sigma^{n+1} - G_\sigma U^{n+1} = 0 \tag{9}$$

The identification of the matrices and arrays appearing in these algebraic equations with the terms arising from the discretization of (4)–(6) is straightforward. Let us remark that matrices G_u and D_u, coming from the gradient of the pressure and the divergence of the velocity, respectively, are related by $G_u = -D_u^T$. Similarly, matrices G_σ and D_σ coming from the symmetric gradient of the velocity and the divergence of the stress, respectively, are related by $(1-\beta)G_\sigma = -D_\sigma^T$. We have explicitly displayed the dependence of matrices K_u and K_σ on U, in the first case due to the convective term in (1) and in the second to the convective and rotational terms in (3).

Equations (7)–(9) can be written in compact form as

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & 0 \\ A_{31} & 0 & 0 \end{bmatrix} \begin{bmatrix} U^{n+1} \\ \Sigma^{n+1} \\ P^{n+1} \end{bmatrix} = \begin{bmatrix} F_1^{n+1} \\ F_2^{n+1} \\ 0 \end{bmatrix} \tag{10}$$

where only the unknowns at time step $n+1$ have been left in the left-hand-side and the identification of the different matrices and arrays is obvious.

3 Schemes Based on Pressure Extrapolation

The first family of schemes to be presented can be introduced using pressure and stress extrapolation in the momentum equation. This implies that these terms are solved explicitly, and therefore this would lead to an at most conditionally stable time integration scheme. To keep the stability properties of the original BDF scheme employed, a velocity correction is required once pressure and stress have been obtained from their corresponding equations. We elaborate this idea in the next subsection, where we present the schemes already proposed in [5]. Then we write the problem posed in terms of the end-of-step unknowns, what we call equivalent monolithic formulation, which allows us to foresee the order in time of the splitting error.

3.1 Formulation of the Algorithms

To motivate the schemes based on pressure extrapolation, let us write the algebraic system (7)–(9) in the equivalent form

$$M_u \frac{\delta_k}{\delta t} \tilde{U}^{n+1} + K_u(\tilde{U}^{n+1})\tilde{U}^{n+1} + G_u \hat{P}_{k'-1}^{n+1} - D_\sigma \hat{\Sigma}_{k'-1}^{n+1} = F^{n+1} \tag{11}$$

$$M_u \frac{1}{\gamma_k \delta t}(U^{n+1} - \tilde{U}^{n+1}) + N_u^{n+1} + G_u(P^{n+1} - \hat{P}_{k'-1}^{n+1}) - D_\sigma(\Sigma^{n+1} - \hat{\Sigma}_{k'-1}^{n+1}) = 0 \tag{12}$$

$$M_\sigma \frac{\delta_k}{\delta t}\tilde{\Sigma}^{n+1} + K_\sigma(\tilde{U}^{n+1})\tilde{\Sigma}^{n+1} - G_\sigma \tilde{U}^{n+1} = 0 \tag{13}$$

$$M_\sigma \frac{1}{\gamma_k \delta t}(\Sigma^{n+1} - \tilde{\Sigma}^{n+1}) + N_\sigma^{n+1} - G_\sigma(U^{n+1} - \tilde{U}^{n+1}) = 0 \tag{14}$$

$$\begin{aligned} &- D_u \tilde{U}^{n+1} + \gamma_k \delta t D_u M_u^{-1} N_u^{n+1} + \gamma_k \delta t D_u M_u^{-1} G_u (P^{n+1} - \hat{P}_{k'-1}^{n+1}) \\ &\qquad - \gamma_k \delta t D_u M_u^{-1} D_\sigma (\Sigma^{n+1} - \hat{\Sigma}_{k'-1}^{n+1}) = 0 \end{aligned} \tag{15}$$

where

$$N_u^{n+1} := K_u(U^{n+1})U^{n+1} - K_u(\tilde{U}^{n+1})\tilde{U}^{n+1}$$
$$N_\sigma^{n+1} := K_\sigma(U^{n+1})\Sigma^{n+1} - K_\sigma(\tilde{U}^{n+1})\tilde{\Sigma}^{n+1}$$

and $\tilde{U}^{n+1}$ and $\tilde{\Sigma}^{n+1}$ are intermediate unknowns. That this system is equivalent to (7)–(9) can be checked as follows: adding (11) and (12) we exactly recover (7), adding (13) and (14) we exactly recover (9), and (15) is obtained multiplying (12) by $\gamma_k \delta t D_u M_u^{-1}$ and making use of (8). The order k' used in the extrapolated variables can in principle be different from k.

Equations (11)–(15) motivate the following algorithm, which is only an approximation to (7)–(9) but allows one to compute the different variables sequentially:

1. Compute $\tilde{U}^{n+1}$ from (11).
2. Compute $\tilde{\Sigma}^{n+1}$ from (13).
3. Compute *an approximation* to P^{n+1} by solving (15) *neglecting* N_u^{n+1} and *replacing* Σ^{n+1} by $\tilde{\Sigma}^{n+1}$.
4. Compute an approximation to U^{n+1} from (12) neglecting N_u^{n+1}.
5. Compute an approximation to Σ^{n+1} from (14) neglecting N_σ^{n+1}.

Several remarks are in order:

- Steps 1 to 5 above allow one to uncouple the calculation of the different variables.
- Matrix $D_u M_u^{-1} G_u$ appearing in the pressure Poisson equation can be approximated by the classical Laplacian matrix, L, with a reduced stencil. This introduces a further approximation, except if an iterative scheme is employed where L is simply used as a preconditioner (see [2, 7]).
- For $k' = k$, the resulting scheme is of order $\mathscr{O}(\delta t^k)$ for a given spatial discretization. We will come back to this point in the following subsection.
- The resulting scheme is only stable for $k' = 1, 2$. For $k' = 3$, the extrapolation $\hat{P}_2^{n+1} = 2P^n - P^{n-1}$ is known to yield an unstable scheme (see the discussion in [1, 12]).

- For $k = 1$ we have an extension to viscoelastic flows of the classical first order fractional step method, whereas for $k = 2$ we have an extension of the second order method.

In view of these comments, we consider $k' = k = 1, 2$, obtaining the following system of equations:

First and second order pressure extrapolation schemes:

$$M_u \frac{\delta_k}{\delta t} \tilde{U}^{n+1} + K_u(\tilde{U}^{n+1})\tilde{U}^{n+1} + G_u \hat{P}^{n+1}_{k-1} - D_\sigma \hat{\Sigma}^{n+1}_{k-1} = F^{n+1} \tag{16}$$

$$M_\sigma \frac{\delta_k}{\delta t} \tilde{\Sigma}^{n+1} + K_\sigma(\tilde{U}^{n+1})\tilde{\Sigma}^{n+1} - G_\sigma \tilde{U}^{n+1} = 0 \tag{17}$$

$$\begin{aligned} &- D_u \tilde{U}^{n+1} + \gamma_k \delta t D_u M_u^{-1} G_u (P^{n+1} - \hat{P}^{n+1}_{k-1}) \\ &\qquad - \gamma_k \delta t D_u M_u^{-1} D_\sigma (\tilde{\Sigma}^{n+1} - \hat{\Sigma}^{n+1}_{k-1}) = 0 \end{aligned} \tag{18}$$

$$\frac{1}{\gamma_k \delta t} M_u (U^{n+1} - \tilde{U}^{n+1}) + G_u (P^{n+1} - \hat{P}^{n+1}_{k-1}) - D_\sigma (\tilde{\Sigma}^{n+1} - \hat{\Sigma}^{n+1}_{k-1}) = 0 \tag{19}$$

$$\frac{1}{\gamma_k \delta t} M_\sigma (\Sigma^{n+1} - \tilde{\Sigma}^{n+1}) - G_\sigma (U^{n+1} - \tilde{U}^{n+1}) = 0 \tag{20}$$

These are the first and second order pressure extrapolation algorithms proposed in [5]. In fact, it is not only the pressure, but also the stress, the variable extrapolated in the first equation.

A third order scheme can be obtained with a different approximation to (11)–(15), which can be related to Yosida's factorization (see [5]). The steps are the following:

1. Compute $\tilde{U}^{n+1}$ from (11) with $k = 3$ and $k' = 2$.
2. Compute $\tilde{\Sigma}^{n+1}$ from (13) with $k = 3$ and $k' = 2$.
3. Compute *an approximation* to P^{n+1} by solving (15) *neglecting* N_u^{n+1}, *replacing* Σ^{n+1} by $\tilde{\Sigma}^{n+1}$ and taking $k = 3$ and $k' = 2$.
4. Compute an approximation to U^{n+1} from (12) *without neglecting* N_u^{n+1}.
5. Compute an approximation to Σ^{n+1} from (14) *neglecting* N_σ^{n+1}.

Even if only a first order extrapolation is used for the pressure and the elastic stresses in the momentum equation, including N_u^{n+1} in the fourth step allows one to obtain *third* order accuracy.

The system of equations to be solved is presented next.

Third order pressure extrapolation scheme:

$$M_u \frac{\delta_3}{\delta t}\tilde{U}^{n+1} + K_u(\tilde{U}^{n+1})\tilde{U}^{n+1} + G_u P^n - D_\sigma \Sigma^n = F^{n+1} \tag{21}$$

$$M_\sigma \frac{\delta_3}{\delta t}\tilde{\Sigma}^{n+1} + K_\sigma(\tilde{U}^{n+1})\tilde{\Sigma}^{n+1} - G_\sigma \tilde{U}^{n+1} = 0 \tag{22}$$

$$\begin{aligned} &- D_u\tilde{U}^{n+1} + \gamma_3\delta t D_u M_u^{-1} G_u(P^{n+1} - P^n) \\ &\qquad - \gamma_3\delta t D_u M_u^{-1} D_\sigma(\tilde{\Sigma}^{n+1} - \Sigma^n) = 0 \end{aligned} \tag{23}$$

$$\begin{aligned} &\frac{1}{\gamma_3\delta t} M_u(U^{n+1} - \tilde{U}^{n+1}) + K_u(U^{n+1})U^{n+1} - K_u(\tilde{U}^{n+1})\tilde{U}^{n+1} \\ &\qquad + G(P^{n+1} - P^n) - D_\sigma(\tilde{\Sigma}^{n+1} - \Sigma^n) = 0 \end{aligned} \tag{24}$$

$$\frac{1}{\gamma_3\delta t} M_\sigma(\Sigma^{n+1} - \tilde{\Sigma}^{n+1}) - G_\sigma(U^{n+1} - \tilde{U}^{n+1}) = 0 \tag{25}$$

3.2 Equivalent Monolithic Formulations

A way to predict formally the order of approximation of the splitting schemes introduced is to write the equations for the final unknowns, after the correction steps, and see which is the perturbation with respect to the original monolithic equations. Let us start with the first and second order schemes introduced earlier. Adding up (16) and (19) on the one hand, and (17) and (20) on the other, we obtain

$$\begin{aligned} M_u \frac{\delta_k}{\delta t} U^{n+1} &+ K_u(U^{n+1})U^{n+1} + G_u P^{n+1} - D_\sigma \Sigma^{n+1} \\ &- N_u^{n+1} - D_\sigma(\tilde{\Sigma}^{n+1} - \Sigma^{n+1}) = F^{n+1} \\ M_\sigma \frac{\delta_k}{\delta t}\Sigma^{n+1} &+ K_\sigma(U^{n+1})\Sigma^{n+1} - G_\sigma U^{n+1} - N_\sigma^{n+1} = 0 \end{aligned}$$

from where we observe that the perturbation of the momentum equation is $-N_u^{n+1} - D_\sigma(\tilde{\Sigma}^{n+1} - \Sigma^{n+1})$ and the perturbation of the stress equation is $-N_\sigma^{n+1}$, as it could be expected from the steps followed. These are the only perturbations, since from (18) and (19) it follows that

$$D_u U^{n+1} = 0$$

i.e., the continuity equation is not perturbed (it would be perturbed if the classical Laplacian matrix L is used, as mentioned earlier).

Let us analyze which is the expected order of accuracy. Combining (19) and (20) we get

$$\left[\frac{1}{\gamma_k \delta t} M_u + \gamma_k \delta t D_\sigma M_\sigma^{-1} G_\sigma\right] (U^{n+1} - \tilde{U}^{n+1}) \\ + G_u(P^{n+1} - \hat{P}_{k-1}) - D_\sigma(\Sigma^{n+1} - \hat{\Sigma}_{k-1}) = 0$$

from where we see that $U^{n+1} - \tilde{U}^{n+1}$ is of order $\mathscr{O}(\delta t^k)$ (in an adequate norm). Knowing this, it follows from (20) that $\Sigma^{n+1} - \tilde{\Sigma}^{n+1}$ is of order $\mathscr{O}(\delta t^{k+1})$. From this we conclude that *the perturbation terms* $-N_u^{n+1} - D_\sigma(\tilde{\Sigma}^{n+1} - \Sigma^{n+1})$ *and* $-N_\sigma^{n+1}$ *are of order* $\mathscr{O}(\delta t^k)$ *and, in fact, the correction step (20) is not needed to have a splitting error of order* $\mathscr{O}(\delta t^k)$. This last remark is relevant, since using the classical factorization point of view described in [5] this last step does not appear.

Let us move our attention to the third order pressure extrapolation scheme. Adding up (21) and (24) on the one hand, and (22) and (25) on the other, we obtain

$$M_u \frac{\delta_3}{\delta t} U^{n+1} + K_u(U^{n+1})U^{n+1} + G_u P^{n+1} - D_\sigma \Sigma^{n+1} - D_\sigma(\tilde{\Sigma}^{n+1} - \Sigma^{n+1}) = F^{n+1}$$

$$M_\sigma \frac{\delta_3}{\delta t} \Sigma^{n+1} + K_\sigma(U^{n+1})\Sigma^{n+1} - G_\sigma U^{n+1} - N_\sigma^{n+1} = 0$$

from where it follows that the perturbation of the momentum equation is only $-D_\sigma(\tilde{\Sigma}^{n+1} - \Sigma^{n+1})$ and the perturbation of the stress equation is $-N_\sigma^{n+1}$. Combining (23) and (24) one gets

$$D_u U^{n+1} + \gamma_3 \delta t D_u M_u^{-1} N_u^{n+1} = 0$$

Let us verify formally which should be the order of accuracy of the scheme. Combining (24) and (25) we get

$$\left[\frac{1}{\gamma_3 \delta t} M_u + K_u(U^{n+1}) - K_u(\tilde{U}^{n+1}) + \gamma_3 \delta t D_\sigma M_\sigma^{-1} G_\sigma\right] (U^{n+1} - \tilde{U}^{n+1}) \\ + G_u(P^{n+1} - P^n) - D_\sigma(\Sigma^{n+1} - \Sigma^n) = 0$$

Noting that $K_u(U)$ is linear in U, from this expression it follows that $U^{n+1} - \tilde{U}^{n+1}$ is of order $\mathscr{O}(\delta t^2)$ (in an adequate norm). Knowing this, from (25) it follows that $\Sigma^{n+1} - \tilde{\Sigma}^{n+1}$ is of order $\mathscr{O}(\delta t^3)$. Contrary to the first and second order schemes, the correction step (25) is now crucial, since it guarantees that the perturbation of the momentum equation is $\mathscr{O}(\delta t^3)$, which is of the same order as the perturbation of the stress equation and the perturbation of the continuity equation of the monolithic scheme. Therefore, we can expect (21)–(25) to be a third order fractional step scheme. This was numerically checked in [5].

4 Schemes Based on Velocity Extrapolation

In the schemes presented heretofore, pressure and stress have been extrapolated in the momentum equation. This permits to compute a first guess for the velocity that needs to be corrected. The idea now is to write an equation for the pressure and extrapolate the velocity and the stress. That should allow one to compute a first guess for the pressure, that may need to be corrected (or not). But such an equation for the pressure is not explicit in (1)–(2), and so we will start reformulating the continuous problem, although we shall see that it is not an appropriate option.

4.1 The Continuous Problem

We may replace the continuous equation (2) by the equation that is obtained taking the divergence of (1) and using the fact that u must be divergence free. This leads to:

$$\varDelta p = \nabla \cdot (f + 2\beta\eta_0 \nabla \cdot \nabla^s u - \rho u \cdot \nabla u + \nabla \cdot \sigma)$$

which has to hold in $\varOmega$ and in the time interval $]0, t_{\mathrm{f}}[$. The appropriate boundary condition for this equation turns out to be that the normal derivative of the pressure on $\partial\varOmega$ be equal to the normal component of the term within parenthesis. If q is a pressure test function, the weak form of this equation reads:

$$(\nabla q, \nabla p) = (\nabla q, f + 2\beta\eta_0 \nabla \cdot \nabla^s u - \rho u \cdot \nabla u + \nabla \cdot \sigma) \tag{26}$$

for all test functions q. The continuous variational problem determined by Eqs. (4), (5) and (6) can be replaced by the problem made by Eqs. (4), (26) and (6). However, two remarks are needed:

- The regularity of the problem has changed. This is obvious from (26). It is well posed for example for pressures in $H^1(\varOmega)$ in space, not only in $L^2(\varOmega)$, velocities in $H^2(\varOmega)$ and stresses in $H(\mathrm{div}; \varOmega)$. The regularity of these variables could be relaxed at the expenses of taking q in $H^2(\varOmega)$. This additional need of regularity is not only a theoretical problem, but also could complicate enormously the numerical approximation.
- For divergence free velocities, $\nabla \cdot \nabla \cdot \nabla^s u = 0$, and therefore the term $2\beta\eta_0 \nabla \cdot \nabla^s u$ could be removed from (26). However, this does not only change the natural boundary condition, but also yields an ill-posed problem (see the discussion and references in [1]).

In view of these comments, it seems clear that system (4), (26) and (6) is not a good alternative. However, we could mimic the obtention of (26) at the algebraic level, and design effective fractional step schemes from the resulting equations.

4.2 Formulation of the Algorithms

Let us consider problem (7)–(9). Multiplying the first equation by $\gamma_k \delta t D_u M_u^{-1}$ and using the fact that $D_u U^{n+1} = 0$ we obtain

$$\begin{aligned}\gamma_k \delta t D_u M_u^{-1} G_u P^{n+1} \\ &= \gamma_k \delta t D_u M_u^{-1}(F^{n+1} - K_u(U^{n+1})U^{n+1} + D_\sigma \Sigma^{n+1}) + D_u U^{*,n} \end{aligned} \tag{27}$$

$$M_u \frac{\delta_k}{\delta t} U^{n+1} + K_u \left(U^{n+1}\right) U^{n+1} + G_u P^{n+1} - D_\sigma \Sigma^{n+1} = F^{n+1} \tag{28}$$

$$M_\sigma \frac{\delta_k}{\delta t} \Sigma^{n+1} + K_\sigma \left(U^{n+1}\right) \Sigma^{n+1} - G_\sigma U^{n+1} = 0 \tag{29}$$

This system is equivalent to (7)–(9), with the difference that now we have an equation for the pressure in terms of the velocity and the stress that is invertible, of Poisson type, obtained from the original monolithic discretization of the problem. To this system we can apply the same ideas as for the algorithms based on pressure extrapolation:

- Compute an approximation to the pressure using a velocity and a stress extrapolation in (27).
- Compute an approximation to the velocity using the pressure obtained and a stress extrapolation in (28).
- Compute the stress using the velocity obtained in (29).
- Correct the velocity to cancel the effect of the extrapolated stress in (28).
- Correct the pressure to cancel the effect of the extrapolated velocity and stress in (27).

To have an overall scheme of order k, the extrapolations need to be of order $k-1$. The equations to be solved are thus the following:

First, second and third order velocity extrapolation schemes:

$$\begin{aligned}\gamma_k \delta t D_u M_u^{-1} G_u \tilde{P}^{n+1} \\ &= \gamma_k \delta t D_u M_u^{-1}(F^{n+1} - K_u(\hat{U}_{k-1}^{n+1})\hat{U}_{k-1}^{n+1} + D_\sigma \hat{\Sigma}_{k-1}^{n+1}) + D_u U^{*,n} \end{aligned} \tag{30}$$

(continued)

$$M_u \frac{\delta_k}{\delta t} \tilde{U}^{n+1} + K_u(\tilde{U}^{n+1})\tilde{U}^{n+1} + G_u \tilde{P}^{n+1} - D_\sigma \hat{\Sigma}_{k-1}^{n+1} = F^{n+1} \tag{31}$$

$$M_\sigma \frac{\delta_k}{\delta t} \Sigma^{n+1} + K_\sigma(\tilde{U}^{n+1})\Sigma^{n+1} - G_\sigma \tilde{U}^{n+1} = 0 \tag{32}$$

$$\frac{1}{\gamma_k \delta t} M_u(U^{n+1} - \tilde{U}^{n+1}) - D_\sigma(\Sigma^{n+1} - \hat{\Sigma}_{k-1}^{n+1}) = 0 \tag{33}$$

$$\begin{aligned} D_u M_u^{-1} G_u(P^{n+1} - \tilde{P}^{n+1}) &= D_u M_u^{-1}(K_u(\hat{U}_{k-1}^{n+1})\hat{U}_{k-1}^{n+1} - D_\sigma \hat{\Sigma}_{k-1}^{n+1}) \\ &\quad + D_u M_u^{-1}(-K_u(U^{n+1})U^{n+1} + D_\sigma \Sigma^{n+1}) \end{aligned} \tag{34}$$

This algorithms admits several modifications and requires some remarks:

- Matrix $D_u M_u^{-1} G_u$ has a wide stencil. In principle, one could use the approximation $D_u M_u^{-1} G_u \tilde{P}^{n+1} \approx L\tilde{P}^{n+1} + (D_u M_u^{-1} G_u - L)\hat{P}_{k-1}$. However, the resulting scheme turns out to be unstable for $k = 3$ because of the second order pressure extrapolation, and thus it cannot be used to design a third order formulation. The alternative could be to use L only as a preconditioner in an iterative scheme. See [12] for further discussion.
- If instead of $\hat{\Sigma}_{k-1}^{n+1}$ one uses $\hat{\Sigma}_k^{n+1}$ in (30), the fourth step (33) would be unnecessary from the accuracy point of view. However, stability would be affected, since the intermediate velocities obtained from (31) depend on extrapolated stresses, i.e., to an explicit treatment of the stress in the momentum equation.
- For the exact problem, $D_u U^{*,n} = 0$. However, this does not hold with the approximations done, and the term $D_u U^{*,n}$ has to be kept in (30) to obtain a stable scheme (see [2]).
- A very important point from the computational point of view is that the fifth step (34) is in fact *not needed*, since pressure is not an evolution variable for incompressible flows. However, it is formally convenient to maintain (34), since it shows how the pressure should be corrected in case it is needed.

4.3 Equivalent Monolithic Formulation

As for the schemes based on pressure extrapolation, let us obtain the equivalent monolithic system solved by (30)–(33). The resulting momentum equation is obtained adding up (31) and (33) and the resulting stress equation is directly (32). These equations can be written as

$$M_u \frac{\delta_k}{\delta t} U^{n+1} + K_u(U^{n+1})U^{n+1} + G_u \tilde{P}^{n+1} - D_\sigma \Sigma^{n+1} - N_u^{n+1} = F^{n+1} \tag{35}$$

$$
\begin{aligned}
&M_\sigma \frac{\delta_k}{\delta t} \Sigma^{n+1} + K_\sigma(U^{n+1})\Sigma^{n+1} - G_\sigma U^{n+1} \\
&\qquad + [K_\sigma(\tilde{U}^{n+1}) - K_\sigma(U^{n+1})]\Sigma^{n+1} - G_\sigma(\tilde{U}^{n+1} - U^{n+1}) = 0 \qquad (36)
\end{aligned}
$$

Multiplying (35) by $\gamma_k \delta t D_u M_u^{-1}$ and making use of (30) it is found that

$$
\begin{aligned}
&D_u U^{n+1} + \gamma_k \delta t D_u M_u^{-1}[K_u(U^{n+1})U^{n+1} - K_u(\hat{U}_{k-1}^{n+1})\hat{U}_{k-1}^{n+1}] \\
&\qquad + \gamma_k \delta t D_u M_u^{-1}[-D_\sigma(\Sigma^{n+1} - \hat{\Sigma}_{k-1}^{n+1}) - N_u^{n+1}] = 0 \qquad (37)
\end{aligned}
$$

From (33) it follows that $U^{n+1} - \tilde{U}^{n+1}$ is or order δt^k (in the appropriate norm). Identifying $\tilde{P}^{n+1}$ with the pressure to be computed, we observe that (35)–(37) is a perturbation of the original system (7)–(9) with all the perturbation terms of order δt^k.

5 Comments on Stability

The obvious way to undertake the numerical analysis of the algorithms presented is to evaluate the stability and convergence properties of the segregated schemes with respect to their monolithic counterpart, and then rely on the estimates of stability and convergence of the monolithic formulations with respect to the continuous problem. The difficulty of this approach relies on the fact that convergence estimates of the first step will depend on norms of discrete solutions. While in some cases it is possible to prove bounds for these norms (see [3] for an application of this technique to a first order scheme), in general this boundedness has to be assumed. The order of accuracy of the formulations has to be based solely on the formal derivation presented before, comparing the fractional step schemes with their monolithic versions.

However, stability can be proved rigorously and at the pure algebraic level. This was shown first in [7] for Newtonian fluids, and then the approach was followed in [2, 9] with other schemes (see [1] for a review and additional references).

It is outside the scope of this article to present the stability proofs of the different schemes presented. We will just describe the results that can be obtained in a descriptive manner. To this end, given arrays X and Y of m components and a positive definite $m \times m$ matrix A, we define

$$
(X, Y)_A := X^T A Y, \quad \|X\|_A := (X^T A X)^{1/2}, \quad \|X\|_{-A} := \sup_{Y \neq 0} \frac{X^T Y}{\|Y\|_A}
$$

Given a sequence of arrays $\{X^n\}$, $n = 1, 2, \ldots, N$, we define

$$\{X^n\} \in \ell^\infty(A) \iff \|X^n\|_A < \infty \text{ for all } n = 1, 2, \ldots, N$$

$$\{X^n\} \in \ell^p(A) \iff \sum_{n=1}^{N} \delta t \|X^n\|_A^p < \infty \quad 1 \leq p < \infty$$

where $\delta t = t_\mathrm{f}/N$. We will apply these definitions to the sequences $\{U^n\}$, $\{\tilde{U}^n\}$, $\{\Sigma^n\}$, $\{\tilde{\Sigma}^n\}$ and $\{P^n\}$ obtained using the *first and second order* schemes presented. The third order formulations proposed have been based on the third order BDF time integration scheme, which is only conditionally stable; therefore, unconditional stability for the split schemes cannot be expected.

Let us denote by $K_{u,0}$ the symmetric part of K_u. From the original term in (4) from which matrix comes, it is seen that it is zero when $\beta = 0$. The results one can prove for all the methods presented are the following:

$$\{U^n\} \in \ell^\infty(M_u), \quad \{\tilde{U}^n\} \in \ell^\infty(M_u) \cap \ell^2(K_{u,0})$$

$$\{\Sigma^n\} \in \ell^\infty(M_\sigma), \quad \{\tilde{\Sigma}^n\} \in \ell^\infty(M_\sigma)$$

provided $\sum_{n=1}^{N} \delta t \|F^n\|_{M_u}^2 < \infty$. If $\beta > 0$, the stability for $\{\tilde{U}^n\}$ is optimal, and in fact one only needs to have $\sum_{n=1}^{N} \delta t \|F^n\|_{-L_+}^2 < \infty$, where $L_+ = -L$ and L is the Laplacian matrix, as before, but now extended to vector fields (the sequence of arrays $\{F^n\}$ comes from the approximation of the forcing term f). However, if $\beta = 0$ (or β is very small), we do not have stability of $\{\tilde{U}^n\}$ in the discrete counterpart of $L^2(0, t_\mathrm{g}; H^1(\Omega)^d)$, which is precisely $\ell^2(K_{u,0})$ if $\beta > 0$ or $\ell^2(L_+)$ if $\beta = 0$ (again, L_+ is applied to vector fields).

To obtain the missing stability one has to make use of the inf-sup conditions that need to be satisfied between the approximation of pressures and velocities on the one hand and on the approximation of velocities and stresses on the other. Alternatively, one can use stabilized finite element formulations (see [8] and references therein for further discussion). Using the first option, the conditions that need to be satisfied can be written as follows. Let P be an array in the space coming from the discretization of the pressure and let M_p be the matrix coming from the $L^2(\Omega)$ inner product in the pressure space. Let also U, V be generic arrays in the space coming from the discretization of the velocity and Ψ an array in the space coming from the discretization of the stress. Then, we assume that there exist $\beta_1 > 0$ and $\beta_2 > 0$, constants, such that

$$\text{For all } P \text{ there exists } V \text{ such that } \beta_1 \|P\|_{M_p} \|V\|_{L_+} \leq P^T D_u V$$

$$\text{For all } U \text{ there exists } \Psi \text{ such that } \beta_2 \|U\|_{L_+} \|\Psi\|_{M_\sigma} \leq \Psi^T G_\sigma U$$

Under this assumption, one can prove that

$$\{\tilde{U}^n\} \in \ell^2(L_+), \quad \{P^n\} \in \ell^2(M_p)$$

With this, we have all the stability results that could be expected. In fact, for schemes based on velocity extrapolation one can prove some additional stability results that do not have a counterpart at the continuous level (see the review in [1]).

6 The Inexact Factorization Point of View

Let us apply the inexact factorization point of view to fractional step schemes for viscoelastic flows. This idea was proposed in [13]; see also [14] for an interesting elaboration.

Let A be the matrix of system (10), which we may factorize as $A = L_A U_A$, with L_A lower diagonal per blocks and U_A upper diagonal. Writing (10) as $AX^{n+1} = R^{n+1}$, we may solve the sequence $L_A \tilde{X}^{n+1} = R^{n+1}$ and $U_A X^{n+1} = \tilde{X}^{n+1}$, the advantage being that in each system we can solve sequentially for the different unknowns. The problem is that this process involves the inversion of A_{11} and A_{22}, which is computationally expensive. Therefore, the idea of inexact factorizations is to *approximate* A_{11}^{-1} and A_{22}^{-1}, this yielding approximations to L_A and U_A respectively denoted by L^* and U^*. Thus, the matrix of the approximate factorization is $A^* = L^* U^*$, and the error matrix is $E^* = A - A^*$. We will apply this idea to the first order schemes based on pressure extrapolation and on velocity extrapolation. For the application to second and third order schemes based on pressure extrapolation, see [5].

6.1 First Order Pressure Extrapolation Scheme as Inexact Factorization

To simplify the notation, let us introduce the abbreviations

$$B := D_u M_u^{-1} G_u, \quad C_u := \frac{1}{\delta t} M_u + K_u, \quad C_\sigma := \frac{1}{\delta t} M_\sigma + K_\sigma$$

It is understood in all what follows that matrices K_u and K_σ are evaluated with $\tilde{U}^{n+1}$.

If in algorithm (16)–(20) we take $k = 1$ and replace (20) by $\Sigma^{n+1} = \tilde{\Sigma}^{n+1}$ (that can be done for the reasons explained in Sect. 3.2), we may understand this algorithm as the sequence of solving first $L^*\tilde{X}^{n+1} = R^{n+1}$:

$$
\begin{aligned}
&C_u\tilde{U}^{n+1} = F_1^{n+1}\\
&C_\sigma\tilde{\Sigma}^{n+1} - G_\sigma\tilde{U}^{n+1} = F_2^{n+1}\\
&-D_u\tilde{U}^{n+1} + \delta t B\tilde{P}^{n+1} - \delta t D_u M_u^{-1} D_\sigma\tilde{\Sigma}^{n+1} = 0
\end{aligned}
$$

and then solving $U^* X^{n+1} = \tilde{X}^{n+1}$:

$$
\begin{aligned}
&P^{n+1} = \tilde{P}^{n+1}\\
&\Sigma^{n+1} = \tilde{\Sigma}^{n+1}\\
&U^{n+1} + \delta t M_u^{-1} G_u P^{n+1} - \delta t M_u^{-1} D_\sigma \Sigma^{n+1} = \tilde{U}^{n+1}
\end{aligned}
$$

Matrices L^* and U^* are now given by

$$
L^* = \begin{bmatrix} C_u & 0 & 0 \\ -G_\sigma & C_\sigma & 0 \\ -D_u & -\delta t D_u M_u^{-1} D_\sigma & \delta t B \end{bmatrix}, \quad U^* = \begin{bmatrix} I & -\delta t M_u^{-1} D_\sigma & \delta t M_u^{-1} G_u \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}
$$

Thus, matrix A has effectively been approximated by $A \approx A^* = L^* U^*$, where

$$
A^* = \begin{bmatrix} C_u & -D_\sigma - \delta t K_u M_u^{-1} D_\sigma & G_u + \delta t K_u M_u^{-1} G_u \\ -G_\sigma & C_\sigma + \delta t G_\sigma M_u^{-1} D_\sigma & -\delta t G_\sigma M_u^{-1} G_u \\ -D_u & 0 & 0 \end{bmatrix}
$$

The error matrix of the splitting scheme is

$$
E^* := A - A^* = \begin{bmatrix} 0 & \delta t K_u M_u^{-1} D_\sigma & -\delta t K_u M_u^{-1} G_u \\ 0 & -\delta t G_\sigma M_u^{-1} D_\sigma & \delta t G_\sigma M_u^{-1} G_u \\ 0 & 0 & 0 \end{bmatrix}
$$

This error matrix allows us to observe which are the terms approximated and that they are of first order in time.

6.2 First Order Velocity Extrapolation Scheme as Inexact Factorization

Schemes based on pressure extrapolation can be cast as a classical inexact *LU* factorization. However, velocity correction schemes fit better as inexact *general* factorizations of the system matrix into block triangular matrices. For Newtonian flows, it was shown in [1] that they can be written as a factorization of the system matrix A into two block triangular matrices, but not the canonical *LU* factorization. In the case of viscoelastic flows, it is convenient to organize the unknowns as $(\Sigma^{n+1}, U^{n+1}, P^{n+1})$ and split the matrix of the system to be solved as the product of *three triangular matrices*. If in algorithm (30)–(34) we take $k = 1$ and neglect (34) (for the reasons explained Sect. 4.2) this splitting is as follows:

$$A = \begin{bmatrix} C_\sigma & -G_\sigma & 0 \\ -D_\sigma & C_u & G_u \\ 0 & D_u & 0 \end{bmatrix}$$

$$\approx \begin{bmatrix} I_\sigma & 0 & 0 \\ 0 & I_u & 0 \\ 0 & \delta t D_u M_u^{-1} & -\delta t D_u M_u^{-1} G_u \end{bmatrix} \begin{bmatrix} C_\sigma & -G_\sigma & 0 \\ 0 & C_u & G_u \\ 0 & 0 & I_p \end{bmatrix} \begin{bmatrix} I_\sigma & 0 & 0 \\ -\delta t M_u^{-1} D_\sigma & I_u & 0 \\ 0 & 0 & I_p \end{bmatrix} \tag{38}$$

$$= \begin{bmatrix} C_\sigma & -G_\sigma & 0 \\ -D_\sigma - E_{u\sigma} & C_u & G_u \\ -E_{p\sigma} & D_u - E_{pu} & 0 \end{bmatrix} =: A^*$$

where I_σ, I_u and I_p are the identity matrices corresponding to stress, velocity and pressure, respectively, and the error terms are:

$$E_{u\sigma} = \delta t K_u M_u^{-1} D_\sigma = \mathscr{O}(\delta t)$$

$$E_{p\sigma} = \delta t^2 D_u M_u^{-1} C_u M_u^{-1} D_\sigma = \mathscr{O}(\delta t)$$

$$E_{pu} = -\delta t M_u^{-1} K_u = \mathscr{O}(\delta t)$$

which are all of order δt.

In order to check that this splitting corresponds to (30)–(33), let us write now the approximate factorization (38) as $A^* = T_{(1)} T_{(2)} T_{(3)}$, where matrices $T_{(i)}$, $i = 1, 2, 3$, are all block triangular. This is what allows us to solve for the different unknowns in an uncoupled way. Problem $T_{(1)} X_{(1)} = R^{n+1}$, with $X_{(1)} = (\Sigma_{(1)}^{n+1}, U_{(1)}^{n+1}, P_{(1)}^{n+1})$ and $R^{n+1} = (\frac{1}{\delta t} M_\sigma \Sigma^n, F^{n+1} + \frac{1}{\delta t} M_u U^n, 0)$ yields:

$$\Sigma_{(1)}^{n+1} = \frac{1}{\delta t} M_\sigma \Sigma^n$$

$$U_{(1)}^{n+1} = F^{n+1} + \frac{1}{\delta t} M_u U^n$$

$$\delta t D_u M_u^{-1} G_u P_{(1)}^{n+1} = \delta t D_u M_u^{-1} U_{(1)}^{n+1} = \delta t D_u M_u^{-1} F^{n+1} + D_u U^n$$

from where it follows that $P_{(1)}^{n+1} = \tilde{P}^{n+1}$ is the solution of (30) (with $k = 1$). Solving now $T_{(2)}X_{(2)} = X_{(1)}$ yields:

$$\begin{aligned}
&P_{(2)}^{n+1} = P_{(1)}^{n+1} = \tilde{P}^{n+1} \\
&C_u U_{(2)}^{n+1} + G_u P_{(2)}^{n+1} = U_{(1)}^{n+1} \iff C_u U_{(2)}^{n+1} = F^{n+1} + \frac{1}{\delta t} M_u U^n - G_u \tilde{P}^{n+1} \\
&C_\sigma \Sigma_{(2)}^{n+1} - G_\sigma U_{(2)}^{n+1} = \Sigma_{(1)}^{n+1} \iff C_\sigma \Sigma_{(2)}^{n+1} - G_\sigma U_{(2)}^{n+1} = \frac{1}{\delta t} M_\sigma \Sigma^n
\end{aligned}$$

from where it follows that $U_{(2)}^{n+1} = \tilde{U}^{n+1}$ is the solution of (31) and $\Sigma_{(2)}^{n+1} = \Sigma^{n+1}$ the solution of (32), with $k = 1$ in both cases. Finally, solving $T_{(3)}X_{(3)} = X_{(2)}$ yields:

$$\begin{aligned}
&\Sigma_{(3)}^{n+1} = \Sigma_{(2)}^{n+1} = \Sigma^{n+1} \\
&- \delta t M_u^{-1} D_\sigma \Sigma_{(3)}^{n+1} + U_{(3)}^{n+1} = U_{(2)}^{n+1} \iff U_{(3)}^{n+1} = \tilde{U}^{n+1} + \delta t M_u^{-1} D_\sigma \Sigma^{n+1} \\
&P_{(3)}^{n+1} = P_{(2)}^{n+1} = \tilde{P}^{n+1}
\end{aligned}$$

from where $U_{(3)}^{n+1} = U^{n+1}$ is the solution of (33) with $k = 1$. Therefore, $X_{(3)}$ is the solution of the first order version of (30)–(33), thus proving that this algorithm corresponds to the inexact factorization (38).

7 Conclusions

In this article we have explained the main aspects related to the design of fractional step schemes for viscoelastic flows at the purely algebraic level. The design of the algorithms has taken as starting point the fully discrete problem, discretized both in space and in time. The driving idea in all cases is to extrapolate one variable to allow the uncoupled calculation of the others and then to make a correction to maintain the implicitness of the original time integration. Two families of schemes have been presented, one based on pressure (and stress) extrapolation and the other based on velocity (and stress) extrapolation. In the former case, the modifications required to design a third order scheme have been explained, whereas the latter has been motivated from a discrete pressure Poisson equation that does not have the theoretical difficulties of the continuous one.

A first way to understand the properties of the schemes proposed, and in particular their order of accuracy, is to write the equivalent monolithic problem. This shows which equations of the original system are approximated a how. The interpretation of the schemes as inexact factorization serves the same target, and is also a source of inspiration to design other fractional steps schemes.

Comments about the stability of the schemes have been also provided. Summarizing, one can prove at the discrete level the same stability results as those that hold for the continuous counterpart, although using purely algebraic concepts.

Many of the points treated deserve further research. Related to the last point, for example, the stability of third order schemes has not been undertaken, and the analysis of either inf-sup stable or stabilized formulations has many gaps to be filled, although we have tried to explain the main lines. The same happens with the identification of inexact factorizations for all the schemes proposed, and even the analysis of modifications that these factorizations suggest. Needless to say that all what has been presented could be applied to time integration schemes other than BDF. The usefulness of algebraic fractional step schemes to design preconditioners for linear solvers has not even been touched. Nevertheless, our objective has been to provide a global picture of this way to approach fractional step methods in computational fluid mechanics, particularly applied to viscoelastic fluids.

Acknowledgements The author wishes to acknowledge the financial support received from project Elastic-Flow, ref. DPI2015-67857-R, from the *Retos Investigación* program of the Spanish Ministerio de Economía y Competitividad.

References

1. Badia, S., Codina, R.: Algebraic pressure segregation methods for the incompressible Navier-Stokes equations. Arch. Comput. Methods Eng. **15**, 1–52 (2007)
2. Badia, S., Codina, R.: Pressure segregation methods based on a discrete pressure Poisson equation. An algebraic approach. Int. J. Numer. Methods Fluids **56**, 351–382 (2008)
3. Blasco, J., Codina, R.: Error estimates for an operator splitting method for incompressible flows. Appl. Numer. Math. **51**, 1–17 (2004)
4. Castillo, E., Codina, R.: Variational multi-scale stabilized formulations for the stationary three-field incompressible viscoelastic flow problem. Comput. Methods Appl. Mech. Eng. **279**, 579–605 (2014)
5. Castillo, E., Codina, R.: First, second and third order fractional step methods for the three-field viscoelastic flow problem. J. Comput. Phys. **296**, 113–137 (2015)
6. Chorin, A.J.: A numerical method for solving incompressible viscous flow problems. J. Comput. Phys. **2**, 12–26 (1967)
7. Codina, R.: Pressure stability in fractional step finite element methods for incompressible flows. J. Comput. Phys. **170**, 112–140 (2001)
8. Codina, R.: Finite element approximation of the three-field formulation of the Stokes problem using arbitrary interpolations. SIAM J. Numer. Anal. **47**, 699–718 (2009)
9. Codina, R., Badia, S.: On some pressure segregation methods of fractional-step type for the finite element approximation of incompressible flow problems. Comput. Methods Appl. Mech. Eng. **195**, 2900–2918 (2006)
10. Fernández-Cara, E., Guillén, F., Ortega, R.R.: Mathematical modeling and analysis of viscoelastic fluids of the Oldroyd kind. In: Handbook of Numerical Analysis, VIII. North-Holland, Amsterdam (2002)
11. Guermond, J.L., Minev, P., Shen, J.: An overview of projection methods for incompressible flows. Comput. Methods Appl. Mech. Eng. **195**, 6011–6045 (2006)
12. Owen, H., Codina, R.: A third-order velocity correction scheme obtained at the discrete level. Int. J. Numer. Methods Fluids **69**, 57–72 (2012)

13. Perot, J.B.: An analysis of the fractional step method. J. Comput. Phys. **108**, 51–58 (1993)
14. Quarteroni, A., Saleri, F., Veneziani, A.: Factorization methods for the numerical approximation of Navier-Stokes equations. Comput. Methods Appl. Mech. Eng. **188**, 505–526 (2000)
15. Saramito, P.: A new θ-scheme algorithm and incompressible FEM for viscoelastic fluid flows. ESAIM Math. Model. Numer. Anal. Modél. Math. Anal. Numér. **28**, 1–35 (1994)
16. Temam, R.: Sur l'approximation de la solution des équations de Navier-Stokes par la méthode des pas fractionaires (I). Arch. Ration. Mech. Anal. **32**, 135–153 (1969)

Some Remarks on the Hierarchic Control for Coupled Parabolic PDEs

Víctor Hernández-Santamaría and Luz de Teresa

Dedicated to Prof. Enrique Fernández-Cara on the occasion of his 60th birthday.

Abstract In this paper, we study Stackelberg-Nash strategies to control a system of two coupled parabolic equations. We assume that we act in the system by means of a hierarchy of controls. First, a *leader* (vectorial) control achieve their objectives, and then other controls, named *followers*, react optimally to the leader action. We prove an observability inequality for an extended system, which yields the Stackelberg-Nash optimization. Then, we remove the action of one of the components of the leader control. In this way, we control a system of various equations by acting only on the first component.

Keywords Controllabilty · Stackelberg-Nash strategies · Carleman inequalities · parabolic systems

1 Introduction

In the last years, there has been an increasing interest in studying multi-objective control problems for PDEs. In game theory, Stackelberg [24] formulated non cooperative decision problems where one of the participants act as a *leader* and the others react according to the decisions of the leader (these participants are

V. Hernández-Santamaría
Depto. de Control Automático, CINVESTAV-IPN, Ciudad de México, Mexico

L. de Teresa (✉)
Instituto de Matemáticas, Universidad Nacional Autónoma de México, Mexico City, Mexico
e-mail: ldeteresa@im.unam.mx

A. Doubova et al. (eds.), *Recent Advances in PDEs: Analysis, Numerics and Control*, SEMA SIMAI Springer Series 17,
https://doi.org/10.1007/978-3-319-97613-6_7

named *followers*). If in addition, the followers have an optimality objective, it will be desirable to have a Nash equilibrium [23].

In the seminal work of Lions, see [20, 21], the method of hierarchic control was introduced as a tool to address multi-objective problems by combining the concepts of optimal control and controllability. This technique used the notion of Stackelberg optimization. Later, several other papers, see for instance, [10, 16, 19–21], applied the hierarchic control methodology to solve a wide variety of problems. In particular, in [5] the authors developed the first hierarchical results within the controllability to trajectories framework for parabolic equations.

Most of the previous works have one thing in common: they deal with hierarchical control of a single equation. The aim of this paper is to develop a hierarchic control strategy for a non-scalar system of parabolic equations. The systems analyzed here represent a linear version of more complex models arising in mathematical biology: chemotaxis (see for instance [9, 22]) or treatment of tumors [8]. As far as we know, there are two papers dealing with coupled systems: in [4] the authors study a Stackelberg-Nash strategy for two coupled equations of fluid mechanics, with controls acting on both equations and with an approximate controllability objective for the leaders. In [17], the authors deal with a Stackelberg-Nash strategy for a cascade system of parabolic equations acting only in the first equation, but with a suitable weight on the follower control. This requirement is not necessary when dealing with a single parabolic equation (see [5]). However when dealing with systems, hierarchic control becomes more intricate and other control strategies are required.

2 The Problem and Its Formulation

Let Ω be an open and bounded domain of $\mathbb{R}^N$ with boundary $\partial\Omega$ of class C^2 and ω be an open and nonempty subset of Ω. Given $T > 0$, we consider the following system of coupled parabolic PDEs with leader controls localized in ω and follower controls localized in $\omega_1, \omega_2 \subset \Omega$ with $\omega_i \cap \omega = \emptyset$. More precisely

$$\begin{cases} y_{1,t} - \Delta y_1 + a_{11}y_1 + a_{12}y_2 = h_1\chi_\omega + v^1\chi_{\omega_1} + v^2\chi_{\omega_2} \quad \text{in } Q = \Omega \times (0,T), \\ y_{2,t} - \Delta y_2 + a_{21}y_1 + a_{22}y_2 = h_2\chi_\omega \quad \text{in } Q = \Omega \times (0,T), \\ y_j = 0 \text{ on } \Sigma = \partial\Omega \times (0,T), \ \ j = 1,2, \\ y_j(x,0) = y_j^0(x) \text{ in } \Omega, \ \ j = 1,2, \end{cases} \tag{1}$$

where $a_{ij} = a_{ij}(x,t) \in L^\infty(Q)$ and $y_j^0 \in L^2(\Omega)$ are given.

In system (1), $y = (y_1, y_2)^t$ is the state, $v^j = v^j(x,t)$ and $h = (h_1(x,t), h_2(x,t))^t$ are the followers and leader control functions, respectively, while χ_ω and χ_{ω_j} denote the characteristic functions of ω and ω_j.

Observe that for each $h_j \in L^2(\omega \times (0,T))$, $v^j \in L^2(\omega_j \times (0,T))$, and $y_{j,0} \in L^2(\Omega)$, $j = 1,2$, system (1) admits a unique weak solution $y \in [C\left([0,T]; L^2(\Omega)\right) \cap L^2\left(0,T; H_0^1(\Omega)\right)]^2$, hereinafter denoted as

$$y = y(x,t; h, v^1, v^2).$$

In the case where only a (leader) control is exerted on ω, i.e. $v^1 \equiv v^2 \equiv 0$, there exist several papers devoted to the controllability of non-scalar parabolic systems, see for instance [1, 2], or [3] for a recent survey on the controllability of coupled parabolic problems. In particular, in [15] the authors proved that it is possible to get a null controllability result acting only on the first component of the system. That is, the following system

$$\begin{cases} y_{1,t} - \Delta y_1 + a_{11}y_1 + a_{12}y_2 = h\chi_\omega & \text{in } Q, \\ y_{2,t} - \Delta y_2 + a_{21}y_1 + a_{22}y_2 = 0 & \text{in } Q, \\ y_j = 0 \text{ on } \Sigma, \ j = 1,2, \\ y_j(x,0) = y_j^0(x) \text{ in } \Omega, \ j = 1,2, \end{cases}$$

is null controllable as long as a_{21} has a fixed sign on an open subset of ω.

In this paper we are interested in a Stackelberg-Nash multi-objective control strategy for system (1). In what follows, we give a precise description of the problem.

Given $h \in [L^2(Q)]^2$ and $\mathscr{O}_{1,d}, \mathscr{O}_{2,d} \subset \Omega$ two open subsets—representing the observation domains of the followers—localized arbitrarily in Ω, we define the followers functionals

$$\begin{aligned} J_i(v^1, v^2; h) = \frac{\alpha_i}{2} \iint_{\mathscr{O}_{i,d}\times(0,T)} \left(|y_1 - y_{1,d}^i|^2 + |y_2 - y_{2,d}^i|^2\right) dxdt \\ + \frac{\mu_i}{2} \iint_{\omega_i\times(0,T)} |v^i|^2 dxdt, \ i = 1,2, \end{aligned} \tag{2}$$

where $\alpha_i, \mu_i > 0$ are constants and $y_d^i = (y_{1,d}^i, y_{2,d}^i)^t$ is a given function in $L^2(\mathscr{O}_{1,d} \times (0,T)) \times L^2(\mathscr{O}_{2,d} \times (0,T))$.

We consider also the leader functional

$$J(h) = \frac{1}{2} \iint_{\omega\times(0,T)} |h_1|^2 dxdt + \frac{1}{2} \iint_{\omega\times(0,T)} |h_2|^2 dxdt.$$

The main objective is to choose h minimizing J subject to the null controllability constraint

$$y(\cdot, T; h, v^1, v^2) = 0 \quad \text{in } \Omega. \tag{3}$$

The second objective is the following. Given the functions h and y_d^i, we want to choose the control v^i minimizing J_i. Intuitively, this is that throughout the interval $t \in (0, T)$

$$\begin{gathered} y(x,t;h,v^1,v^2) \text{ “do not deviate much” from } y_d^i(x,t), \\ \text{in the observability domain } \mathscr{O}_{i,d}. \end{gathered} \tag{4}$$

To achieve simultaneously (3) and (4), the control process can be described as follows:

- For a fixed leader control h, find controls $(\overline{v}^1, \overline{v}^2)$ (depending on h) and the corresponding state solution $y = y(h, \overline{v}^1, \overline{v}^2)$ to (1) satisfying the Nash equilibrium related to the functionals (J_1, J_2). That is, given h, find $(\overline{v}^1, \overline{v}^2)$ such that

$$J_1(h, \overline{v}^1, \overline{v}^2) \leq J_1(h, v^1, \overline{v}^2), \quad \forall v^1 \in L^2(\omega_1 \times (0,T)),$$
$$J_2(h, \overline{v}^1, \overline{v}^2) \leq J_2(h, \overline{v}^1, v^2), \quad \forall v^2 \in L^2(\omega_2 \times (0,T)),$$

or equivalently

$$J_1\left(h, \overline{v}^1, \overline{v}^2\right) = \min_{v^1} J_1\left(h, v^1, \overline{v}^2\right), \tag{5}$$

$$J_2\left(h, \overline{v}^1, \overline{v}^2\right) = \min_{v^2} J_2\left(h, \overline{v}^1, v^2\right). \tag{6}$$

Any pair $(\overline{v}^1, \overline{v}^2)$ satisfying (5)–(6) is called a Nash equilibrium for (J_1, J_2). Thanks to the linearity of system (1), J_1 and J_2 are strictly convex functionals. Then $(\overline{v}^1, \overline{v}^2)$ is a Nash equilibrium with respect to (J_1, J_2) if and only if

$$\left(\frac{\partial J_1}{\partial v^1}(h, \overline{v}^1, \overline{v}^2), v^1\right) = 0 \quad \forall v^1 \in L^2(\omega_1 \times (0,T)), \tag{7}$$

$$\left(\frac{\partial J_2}{\partial v^2}(h, \overline{v}^1, \overline{v}^2), v^2\right) = 0 \quad \forall v^2 \in L^2(\omega_2 \times (0,T)). \tag{8}$$

- After identifying the Nash equilibrium and the associated state $y = y(h, \overline{v}^1(h), \overline{v}^2(h))$ for each h, we look for an optimal control $\widehat{h}$ such that

$$J(\widehat{h}) = \min_{h} J\left(h, \overline{v}^1(h), \overline{v}^2(h)\right) \tag{9}$$

subject to the restriction

$$y(\cdot, T; h, \overline{v}^1(h), \overline{v}^2(h)) = 0 \quad \text{in } \Omega. \tag{10}$$

In the previous paper [17], the authors studied a Stackelberg-Nash strategy for system (1) when the leader control is exerted only on the first equation of the system, i.e. when $h = (h_1, 0)^t$. However, the followers minimize a modified functional defined as

$$\begin{aligned} J_i^\star(h, v^1, v^2) = \frac{\alpha_i}{2} \iint_{\mathscr{O}_{i,d}\times(0,T)} \left(|y_1 - y^i_{1,d}|^2 + |y_2 - y^i_{2,d}|^2\right) dxdt \\ + \frac{\mu_i}{2} \iint_{\omega_i\times(0,T)} \rho_\star^2(t)|v^i|^2 dxdt,\ i = 1, 2, \end{aligned} \tag{11}$$

with an appropriate positive function $\rho_\star(t)$ blowing up at $t = 0$ and $t = T$. The results in [17] are valid when $\mathscr{O}_{i,d} = \mathscr{O}_d$, $\mathscr{O}_d \cap \omega \neq \emptyset$ and the sign condition

$$a_{21} \geq a_0 > 0 \quad \text{or} \quad -a_{21} \geq a_0 > 0 \quad \text{in } (\mathscr{O}_d \cap \omega) \times (0, T) \tag{12}$$

holds. The penalizing weight function in (11) forces the control v^i to vanish exponentially as $t \to 0$ and $t \to T$ and then the leader $h = (h_1, 0)^t$ finds no obstruction to control the system.

2.1 Main Results

The main contributions of this paper can be stated as follows. Assume that

$$\mathscr{O}_{1,d} = \mathscr{O}_{2,d}, \tag{13}$$

denoted in the following sections as $\mathscr{O}_d$. Our first result is the following:

Theorem 1 *Suppose that* (13) *holds,* $\mathscr{O}_d \cap \omega \neq \emptyset$ *and that* μ_i, $i = 1, 2$, *are large enough. Then, there exists a positive function* $\rho = \rho(t)$ *blowing up at* $t = T$ *such that for any* $y^i_d \in [L^2(\mathscr{O}_d \times (0, T))]^2$ *satisfying*

$$\iint_{\mathscr{O}_d\times(0,T)} \rho^2|y^i_{j,d}|^2 dxdt < +\infty, \quad i, j = 1, 2, \tag{14}$$

and any $y^0 \in L^2(\Omega)^2$, *there exists a control* $h = (h_1, h_2)^t \in [L^2(\omega \times (0, T))]^2$ *and its associated Nash equilibrium* $(\bar{v}^1, \bar{v}^2)$ *such that the solution of (1) satisfies (10).*

Observe that in the previous result, the leader control h has two components, one for each equation in the system. When dealing with the controllability of non-scalar parabolic systems, one of the main questions is if it is possible to control many equations with few controls. There are various positive answers in the classical context of controllability problems (see [3] for a survey on this topic). Therefore, in the case of hierarchic control, it is natural to ask if we can remove the action of one of the leader controls.

Here, following the spirit of [7], we consider the modified follower functionals

$$\widetilde{J}_i(v^1, v^2; h) = \frac{\alpha_i}{2} \iint_{\mathscr{O}_d \times (0,T)} |y_2 - y^i_{2,d}|^2 dxdt + \frac{\mu_i}{2} \iint_{\omega_i \times (0,T)} |v^i|^2 dxdt, \ i = 1, 2. \tag{15}$$

In this way, we only consider the second variable of system (1) for the optimization problem of the followers. We will prove that by introducing this new functional, we can also eliminate the action on the second component of the leader control.

We have the following:

Theorem 2 *Suppose that* (13) *holds,* $\mathscr{O}_d \cap \omega = 0$ *and* μ_i*,* $i = 1, 2$*, are large enough. If the sign condition* (12) *is verified, then there exists a positive function* $\rho = \rho(t)$ *blowing up at* $t = T$ *such that if* (14) *holds, then for any* $y^0 \in L^2(\Omega)^2$ *there exists a control* $h = (h_1, 0)^t \in [L^2(\omega \times (0, T))]^2$ *and its associated Nash equilibrium* $(\bar{v}^1, \bar{v}^2)$*—for the functionals given by* (15)*—such that the solution of (1) satisfies (10).*

Remark 1 Some remarks are in order.

- Just as in [5], the condition $\rho y^i_{j,d} \in L^2(Q)$ seems natural and it means that the follower objectives $y^i_{j,d}$ approach 0 as $t \to T$. This is because the leader control h should not find any obstruction to control the system. It remains an open problem to verify if this condition is necessary, even in the scalar case.
- Condition (12) is exactly the one employed on [15] to prove the null controllability of (1) $[v^1 = v^2 = 0]$ when the control is exerted on the first component of the system, i.e. $h = (h_1, 0)^t$. Moreover, such condition can be applied repeatedly to study the null controllability for non-scalar parabolic problems of m equations in cascade form, see [15].
- Recently in [6] the authors eliminate the condition $\mathscr{O}_{1,d} = \mathscr{O}_{2,d}$ in the scalar case. It is not clear that the same arguments hold in the case of coupled systems.
- Unlike other papers as [16] (in the scalar case) or [4] (in the coupled case), we are supposing that the follower controls are being applied in some sets ω_i disjoint of the leader set ω. This leads to a more realistic situation, because otherwise once the followers choose a policy, the leader modifies its behavior at the same points.

The rest of the paper is organized as follows. We devote Sect. 3 to prove Theorem 1, we briefly review the existence and uniqueness of Nash equilibrium, as well as its characterization. Then, we prove that the leader controls solve the problem of null controllability. In Sect. 4, we prove Theorem 2. Lastly, we present some concluding remarks in Sect. 5.

3 Proof of Theorem 1

3.1 Optimality Condition for the Followers

Here, we briefly recall some results about the follower controls. Hereinafter, we assume (13), i.e., $\mathscr{O}_{1,d} = \mathscr{O}_{2,d}$.

Following the arguments of [5] and [17], the existence and uniqueness of follower controls for system (1) is guaranteed if the parameters μ_i, $i = 1, 2$, in Eq. (2) are large enough.

Since the functionals (2) are continuous, coercive and strictly convex, we have that $(\overline{v}^1, \overline{v}^2)$ is a Nash equilibrium (in the sense of (7)–(8)) if and only if

$$\begin{aligned}&\alpha_i \iint_{\mathscr{O}_d \times (0,T)} \left(y_1 - y^i_{1,d}\right) \widehat{y}^i_1 + \left(y_2 - y^i_{2,d}\right) \widehat{y}^i_2 dxdt \\ &+\mu_i \iint_{\omega_i \times (0,T)} \overline{v}^i \widehat{v}^i dxdt = 0, \quad \forall \widehat{v}^i \in L^2(\omega_i \times (0,T)),\ i = 1, 2,\end{aligned} \tag{16}$$

where $\widehat{y}^i = \left(\widehat{y}^i_1, \widehat{y}^i_2\right)^t$ is the solution of system

$$\begin{cases} \widehat{y}^i_{1,t} - \Delta \widehat{y}^i_1 + a_{11}\widehat{y}^i_1 + a_{12}\widehat{y}^i_2 = \widehat{v}^i \chi_{\omega_i} & \text{in } Q, \\ \widehat{y}^i_{2,t} - \Delta \widehat{y}^i_2 + a_{21}\widehat{y}^i_1 + a_{22}\widehat{y}^i_2 = 0 & \text{in } Q, \\ \widehat{y}^i_j(0) = 0 \text{ in } \Omega, \quad \widehat{y}^i_j = 0 \text{ on } \Sigma,\ j = 1, 2. \end{cases} \tag{17}$$

Let us introduce the adjoint state to (17), that is, $p^i = \left(p^i_1, p^i_2\right)^t$ solution of

$$\begin{cases} -p^i_{1,t} - \Delta p^i_1 + a_{11} p^i_1 + a_{21} p^i_2 = \alpha_i \left(y_1 - y^i_{1,d}\right) \chi_{\mathscr{O}_d} & \text{in } Q, \\ -p^i_{2,t} - \Delta p^i_2 + a_{12} p^i_1 + a_{22} p^i_2 = \alpha_i \left(y_2 - y^i_{2,d}\right) \chi_{\mathscr{O}_d} & \text{in } Q, \\ p^i_j(T) = 0 \text{ in } \Omega, \quad p^i_j = 0 \text{ on } \Sigma,\ j = 1, 2. \end{cases} \tag{18}$$

If we multiply (18) by $\widehat{y}^i$ in $L^2(Q)^2$ and integrate by parts, we obtain

$$\iint_Q \alpha_i \left(y_1 - y^i_{1,d}\right) \chi_{\mathscr{O}_d} \widehat{y}^i_1 - a_{21} p^i_2 \widehat{y}^i_1 dxdt = \iint_Q p^i_1 \left(\widehat{v}^i \chi_{\omega_i} - a_{12}\widehat{y}^i_2\right) dxdt,$$

$$\iint_Q \alpha_i \left(y_2 - y^i_{2,d}\right) \chi_{\mathscr{O}_d} \widehat{y}^i_2 dxdt = \iint_Q (-a_{21} p^i_2 \widehat{y}^i_1 + a_{12} p^i_1 \widehat{y}^i_2) dxdt.$$

Adding up the above expressions and replacing on (16) we have

$$\iint_{\omega_i \times (0,T)} p^i_1 \widehat{v}^i dxdt + \mu_i \iint_{\omega_i \times (0,T)} \overline{v}^i \widehat{v}^i dxdt = 0,$$

which implies that

$$(p_1^i + \mu_i \overline{v}^i)|_{\omega_i} = 0.$$

Therefore, given $h \in [L^2(\omega \times (0,T))]^2$, the pair $(\overline{v}^1, \overline{v}^2)$ is a Nash equilibrium for problem (5)–(6) if and only if

$$\overline{v}^i = -\frac{1}{\mu_i} p_1^i|_{\omega_i}, \quad i = 1, 2,$$

where p_1^i can be found from (y, p^i) solution to the coupled system

$$\begin{cases} y_{1,t} - \Delta y_1 + a_{11} y_1 + a_{12} y_2 = h_1 \chi_\omega - \frac{1}{\mu_1} p_1^1 \chi_{\omega_1} - \frac{1}{\mu_2} p_1^2 \chi_{\omega_2} & \text{in } Q, \\ y_{2,t} - \Delta y_2 + a_{21} y_1 + a_{22} y_2 = h_2 \chi_\omega & \text{in } Q, \\ -p_{1,t}^i - \Delta p_1^i + a_{11} p_1^i + a_{21} p_2^i = \alpha_i \left(y_1 - y_{1,d}^i \right) \chi_{\mathcal{O}_d} & \text{in } Q, \\ -p_{2,t}^i - \Delta p_2^i + a_{12} p_1^i + a_{22} p_2^i = \alpha_i \left(y_2 - y_{2,d}^i \right) \chi_{\mathcal{O}_d} & \text{in } Q, \\ y_j(0) = y_j^0, \ p_j^i(T) = 0, \quad y_j = p_j^i = 0 \text{ on } \Sigma, \ i, j = 1, 2. \end{cases} \tag{19}$$

3.2 The Leader Controls

Recall that the main goal in the hierarchic methodology is to prove the null controllability of (y_1, y_2) at time T. However, the computation of the follower controls satisfying (5)–(6) added four additional equations coupled to the original system under study. Hence, we now look for $h = (h_1, h_2) \in [L^2(\omega \times (0,T))]^2$ such that the solution of (19) satisfies (9)–(10).

It is classical by now that null controllability is related to the observability of a proper adjoint system (see, for instance, [12, 25]). For our particular case, let us consider the adjoint system

$$\begin{cases} -\varphi_{1,t} - \Delta \varphi_1 + a_{11} \varphi_1 + a_{21} \varphi_2 = (\alpha_1 \theta_1^1 + \alpha_2 \theta_1^2) \chi_{\mathcal{O}_d} & \text{in } Q, \\ -\varphi_{2,t} - \Delta \varphi_2 + a_{12} \varphi_1 + a_{22} \varphi_2 = (\alpha_1 \theta_2^1 + \alpha_2 \theta_2^2) \chi_{\mathcal{O}_d} & \text{in } Q, \\ \theta_{1,t}^i - \Delta \theta_1^i + a_{11} \theta_1^i + a_{12} \theta_2^i = -\frac{1}{\mu_i} \varphi_1 \chi_{\omega_i} & \text{in } Q, \\ \theta_{2,t}^i - \Delta \theta_2^i + a_{21} \theta_1^i + a_{22} \theta_2^i = 0 & \text{in } Q, \\ \varphi_j(T) = f_j, \ \theta_j^i(0) = 0 \text{ in } \Omega, \quad \varphi_j = \theta_j^i = 0 \text{ on } \Sigma, \ j = 1, 2. \end{cases} \tag{20}$$

The main task is to prove an observability inequality for system (20).

We have the following result:

Proposition 1 *Under assumptions of Theorem 1, there exist a positive constant C and a positive weight function $\rho = \rho(t)$ blowing up at $t = T$ such that*

$$\|\varphi_1(0)\|^2_{L^2(\Omega)} + \|\varphi_2(0)\|^2_{L^2(\Omega)} + \sum_{i=1}^{2} \iint_Q \rho^{-2} \left(|\theta_1^i|^2 + |\theta_2^i|^2\right) dxdt \\ \leq C \left(\iint_{\omega\times(0,T)} \left(|\varphi_1|^2 + |\varphi_2|^2\right) dxdt \right), \tag{21}$$

for any $(f_1, f_2) \in [L^2(\Omega)]^2$, where (φ, θ^i) is the associated solution to (20).

The proof of Proposition 1 relies on various well-known arguments. For the moment, suppose that the proposition holds and let us end the proof of Theorem 1. There are several ways to prove that inequality (21) implies the existence of a pair (h_1, h_2) of minimum norm. We sketch one of them. It is clear that

$$\|(f_1, f_2)\|_W = \left(\iint_{\omega\times(0,T)} \left(|\varphi_1|^2 + |\varphi_2|^2\right) dxdt \right)^{1/2},$$

where (φ_1, φ_2) are the first two components of the solution to (20), is a semi-norm. From (21), which gives a unique continuation property, it is straightforward to see it defines a norm in $[L^2(\Omega)]^2$. We define W as the completion of $[L^2(\Omega)]^2$ with this norm and set

$$\mathscr{I}(f_1, f_2) = \frac{1}{2}\|(f_1, f_2)\|^2_W + \int_\Omega y_1^0 \varphi_1(0) dx + \int_\Omega y_2^0 \varphi_2(0) dx \\ - \sum_{i=1}^{2} \alpha_i \iint_{\mathscr{O}_d\times(0,T)} \left(\theta_1^i y_{1,d}^i + \theta_2^i y_{2,d}^i\right) dxdt,$$

where (φ, θ^i) is the solution to (24). It is clear that $\mathscr{I}$ is continuous and strictly convex. Moreover, the observability inequality (21) allows to prove that

$$\mathscr{I}(f_1, f_2) \geq \frac{1}{4}\|(f_1, f_2)\|^2_W - C\Bigg(\int_\Omega |y_1^0|^2 dx + \int_\Omega |y_2^0|^2 dx \\ + \sum_{i=1}^{2} \alpha_i^2 \iint_Q \rho^2 \left(|y_{1,d}^i|^2 + |y_{2,d}^i|^2\right) dxdt \Bigg),$$

where C and ρ are provided by Proposition 1. Therefore, $\mathscr{I}$ is coercive in W. Note that here, we have used the growth assumption (14). Consequently, from classical results (see, for instance, [12]), the existence of a minimizer $(\widehat{f_1}, \widehat{f_2})$ solution to

$$\mathscr{I}(\widehat{f_1}, \widehat{f_2}) = \min_{(f_1, f_2)\in W} \mathscr{I}(f_1, f_2)$$

is guaranteed. Thus, the pair $(h_1, h_2) = (\widehat{\varphi}_1 \chi_\omega, \widehat{\varphi}_2 \chi_\omega)$, where $(\widehat{\varphi}_1, \widehat{\varphi}_2)$ is the solution to (24) corresponding to this minimizer solves the leader problem (9)–(10). This concludes the proof of Theorem 1.

3.3 Proof of the Observability Inequality

This section is devoted to the proof of Proposition 1. Before stating the results of this section, let us introduce several weight functions that will be useful in the remainder of this paper. We introduce a special function whose existence is guaranteed by the following result [14, Lemma 1.1].

Lemma 1 *Let $\mathscr{B} \subset\subset \Omega$ be a nonempty open subset. Then there exists $\eta^0 \in C^2(\overline{\Omega})$ such that*

$$\begin{cases} \eta^0(x) > 0 & \textit{all } x \in \Omega, \qquad \eta^0|_{\partial\Omega} = 0, \\ |\nabla \eta^0| > 0 & \textit{for all } x \in \overline{\Omega\backslash\mathscr{B}}. \end{cases}$$

Then, for some positive number λ, we introduce the weight functions

$$\alpha(x,t) = \frac{e^{4\lambda\|\eta^0\|_\infty} - e^{\lambda(2\|\eta^0\|_\infty + \eta^0(x))}}{t(T-t)}, \quad \xi(x,t) = \frac{e^{\lambda(2\|\eta^0\|_\infty + \eta^0(x))}}{t(T-t)}. \tag{22}$$

The following notation will be used to abridge the estimates

$$I_m(s,\lambda;z) := \iint_Q e^{-2s\alpha}(s\xi)^{m-2}\lambda^{m-1}|\nabla z|^2 + \iint_Q e^{-2s\alpha}(s\xi)^m\lambda^{m+1}|z|^2,$$

$$I_{m,\mathscr{B}}(s,\lambda;z) := \iint_{\mathscr{B}\times(0,T)} e^{-2s\alpha}(s\xi)^m\lambda^{m+1}|z|^2,$$

for some parameter $s > 0$.

We state a Carleman estimate, due to [18], for solutions to the heat equation:

Lemma 2 *Let $\mathscr{B} \subset\subset \Omega$ be a nonempty open subset. For any $m \in \mathbb{R}$, there exist constants $s_m > 0$, λ_m, and $C_m > 0$ such that, for any $s \geq s_m$, $\lambda \geq \lambda_m$, $F \in L^2(Q)$ and every $z^0 \in L^2(\Omega)$, the solution z to*

$$\begin{cases} z_t - \Delta z = F & \text{in } Q, \\ z = 0 & \text{on } \partial\Omega \times (0,T), \\ z(x,0) = z^0(x) & \text{in } \Omega, \end{cases}$$

satisfies

$$I_m(s,\lambda;z) \leq C_m \left(I_{m,\mathscr{B}}(s,\lambda;z) + \iint_Q e^{-2s\beta}(s\lambda\xi)^{m-3}|F|^2 dxdt \right). \tag{23}$$

Furthermore, C_m only depends on ω, $\mathscr{B}$ and m and s_m can be taken of the form $s_m = \sigma_m(T+T^2)$ where σ_m only depends on ω, $\mathscr{B}$ and m.

Remark 2 Note that by changing t for $T-t$, Lemma 2 remains valid for linear backward in time systems. Therefore, we can apply it interchangeably in what follows.

The observability inequality (21) is consequence of a global Carleman inequality and some energy estimates. We begin by simplifying (20) as follows

$$\begin{cases} -\varphi_{1,t} - \Delta\varphi_1 + a_{11}\varphi_1 + a_{21}\varphi_2 = \psi_1\chi_{\mathscr{O}_d} & \text{in } Q, \\ -\varphi_{2,t} - \Delta\varphi_2 + a_{12}\varphi_1 + a_{22}\varphi_2 = \psi_2\chi_{\mathscr{O}_d} & \text{in } Q, \\ \psi_{1,t} - \Delta\psi_1 + a_{11}\psi_1 + a_{12}\psi_2 = -\left(\frac{\alpha_1}{\mu_1}\chi_{\omega_1} + \frac{\alpha_2}{\mu_2}\chi_{\omega_2}\right)\varphi_1 & \text{in } Q, \\ \psi_{2,t} - \Delta\psi_2 + a_{21}\psi_1 + a_{22}\psi_2 = 0 & \text{in } Q, \\ \varphi_j(T) = f_j,\ \psi_j(0) = 0 \text{ in } \Omega, \quad \varphi_j = \psi_j = 0 \text{ on } \Sigma,\ j = 1,2, \end{cases} \tag{24}$$

where $\psi_j = \alpha_1\theta_j^1 + \alpha_2\theta_j^2$ for $j = 1,2$. Using the notation introduced before, we present below a Carleman inequality for the solutions to system (24). This will be the main ingredient to prove the observability inequality (21).

Proposition 2 *Under assumptions of Theorem 1. There exist positive constants C and σ_1 such that (φ,ψ) solution to (24) satisfies*

$$\begin{aligned} &I_3(s,\lambda;\varphi_1) + I_3(s,\lambda;\varphi_2) + I_3(s,\lambda;\psi_1) + I_3(s,\lambda;\psi_2) \\ &\qquad \leq C\left(\iint_{\omega\times(0,T)} e^{-2s\alpha}s^7\lambda^8\xi^7\left(|\varphi_1|^2 + |\varphi_2|^2\right)dxdt \right), \end{aligned} \tag{25}$$

for any $s \geq s_1 = \sigma_1(T + T^2 + T^2[\max_{1\leq i,j\leq 2} \|a_{ij}\|_\infty^{2/3}])$, *any* $\lambda \geq C$ *and every* $(f_1, f_2) \in [L^2(\Omega)]^2$.

Proof Let us define $\omega_0 = \omega \cap \mathscr{O}_d$. Since $\omega_0 \neq \emptyset$, there exists some subset $\omega' \subset\subset \omega_0$. We start by applying Carleman inequality (23) to each equation in system (24) with $m = 3$ and $\mathscr{B} = \omega'$. By adding them up, we obtain

$$\begin{aligned}
&I_3(s,\lambda;\varphi_1) + I_3(s,\lambda;\varphi_2) + I_3(s,\lambda;\psi_1) + I_3(s,\lambda;\psi_2)\\
&\quad\leq C\left(\sum_{j=1}^{2}\left(I_{3,\omega'}(s,\lambda;\varphi_j) + I_{3,\omega'}(s,\lambda;\psi_j)\right) + \sum_{j=1}^{2}\iint_Q e^{-2s\alpha}|\psi_j\chi_{\mathscr{O}_d}|^2 dxdt\right.\\
&\qquad + \iint_Q e^{-2s\alpha}\left| -\tfrac{\alpha_1}{\mu_1}\varphi_1\chi_{\omega_1} - \tfrac{\alpha_2}{\mu_2}\varphi_1\chi_{\omega_2}\right|^2 dxdt\\
&\qquad \left.+\sum_{i=1}^{2}\sum_{j=1}^{2}\iint_Q e^{-2s\alpha}\left(\|a_{ji}\|_\infty^2|\varphi_j|^2 + \|a_{ij}\|_\infty^2|\psi_j|^2\right)dxdt\right).
\end{aligned}$$

Taking the parameters s and λ large enough we can absorb the lower order terms into the left-hand side in the previous inequality. More precisely, we have

$$\begin{aligned}
&I_3(s,\lambda;\varphi_1) + I_3(s,\lambda;\varphi_2) + I_3(s,\lambda;\psi_1) + I_3(s,\lambda;\psi_2)\\
&\qquad\qquad \leq C\left(\sum_{j=1}^{2} I_{3,\omega'}(s,\lambda;\varphi_j) + \sum_{j=1}^{2} I_{3,\omega'}(s,\lambda;\psi_j)\right),
\end{aligned} \tag{26}$$

valid for every $\lambda \geq C$ and every

$$s \geq s_1 = \sigma_1(T + T^2 + T^2[\max_{1\leq i,j\leq 2} \|a_{ij}\|_\infty^{2/3}]).$$

The next step is to eliminate the local terms corresponding to ψ_1 and ψ_2. We will reason out as in [15] and [11]. First, note that from the definition of the weight functions (22), we have that, for $s \geq C(T + T^2)$

$$(e^{-2s\alpha}\xi^3)_t \leq Cs^2 e^{-2s\alpha}\xi^5, \quad \Delta(e^{-2s\alpha}\xi^3) \leq Cs^2\lambda^2 e^{-2s\alpha}\xi^5. \tag{27}$$

We consider a function $\zeta \in C^\infty(\mathbb{R}^N)$ verifying:

$$0 \leq \zeta \leq 1 \text{ in } \Omega, \quad \zeta \equiv 1 \quad \text{in } \omega', \quad \operatorname{supp}\zeta \subset \omega_0, \tag{28}$$

$$\frac{\Delta\zeta}{\zeta^{1/2}} \in L^\infty(\Omega), \quad \frac{\nabla\zeta}{\zeta^{1/2}} \in L^\infty(\Omega)^N. \tag{29}$$

Such function exists. It is sufficient to take $\zeta = \tilde{\zeta}^4$ with $\tilde{\zeta} \in C_0^\infty(\Omega)$ verifying (28).

Define $u := e^{-2s\alpha}s^3\lambda^4\xi^3$. Then, we multiply the equations satisfied by φ_1 and φ_2 in system (24) by $u\zeta\psi_1$ and $u\zeta\psi_2$, respectively, and integrate over Q. We add those expressions to obtain

$$\iint_Q u\zeta\left(|\psi_1|^2+|\psi_2|^2\right)\chi_{\mathscr{O}_d} = \iint_Q u\zeta\psi_1(-\varphi_{1,t}-\Delta\varphi_1+a_{11}\varphi_1+a_{21}\varphi_2)$$
$$+\iint_Q u\zeta\psi_2(-\varphi_{2,t}-\Delta\varphi_1+a_{12}\varphi_1+a_{22}\varphi_2).$$

We can integrate several times with respect to the time and space variables in the right hand side of the above expression. Using Hölder and Young inequalities together with (27)–(29) we obtain the following

$$I_{3,\omega'}(s,\lambda;\psi_1)+I_{3,\omega'}(s,\lambda;\psi_2)\le \varepsilon C_A\left(I_3(s,\lambda;\psi_1)+I_3(s,\lambda;\psi_2)\right)$$
$$+C_{\varepsilon,A}\left(\iint_{\omega_0\times(0,T)} e^{-2s\alpha}s^7\lambda^8\xi^7|\varphi_1|^2+\iint_{\omega_0\times(0,T)} e^{-2s\alpha}s^7\lambda^8\xi^7|\varphi_2|^2\right), \tag{30}$$

where $\varepsilon > 0$ and C_A, $C_{\varepsilon,A}$ are new constants only depending on Ω, ω', ω and $\|a_{ij}\|_\infty$. Replacing (30) in (26) with ε small enough and noting that $\omega_0\subset\omega$, we obtain the desired inequality. This concludes the proof of Proposition 2.

Now, we are going to improve inequality (25) in the sense that the weight functions do not vanish at $t=0$. We consider the function

$$l(t)=\begin{cases} T^2/4 & \text{for} \quad 0\le t\le T/2,\\ t(T-t) & \text{for} \quad T/2\le t\le T,\end{cases}$$

and the functions

$$\beta(x,t)=\frac{e^{4\lambda\|\eta^0\|_\infty}-e^{\lambda(2\|\eta^0\|_\infty+\eta^0(x))}}{l(t)},\qquad \gamma(x,t)=\frac{e^{\lambda(2\|\eta^0\|_\infty+\eta^0(x))}}{l(t)},$$
$$\beta^*(t)=\max_{x\in\overline{\Omega}}\beta(x,t),\qquad \gamma^*(t)=\min_{x\in\overline{\Omega}}\gamma(x,t).$$

With these definitions, we have the following

Proposition 3 *Let s and λ as in Proposition 2 and μ_i be large enough. Then there exists a positive constant C depending on Ω, ω, $\mathscr{O}_d$, s, λ and T such that*

$$\|\varphi_1(0)\|^2_{L^2(\Omega)}+\|\varphi_2(0)\|^2_{L^2(\Omega)}+\iint_Q e^{-2s\beta^*}(\gamma^*)^3\left(|\varphi_1|^2+|\varphi_2|^2\right)dxdt$$
$$+\iint_Q e^{-2s\beta^*}(\gamma^*)^3\left(|\psi_1|^2+|\psi_2|^2\right)dxdt\le C\left(\iint_{\omega\times(0,T)} e^{-2s\beta}\gamma^7\left(|\varphi_1|^2+|\varphi_2|^2\right)dxdt\right), \tag{31}$$

for any $(f_1, f_2) \in [L^2(\Omega)]^2$*, where* (φ, ψ) *is the associated solution to* (24).

Proof We follow several well-known arguments, see for instance, [13]. First, by construction, $\alpha = \beta$ and $\xi = \gamma$ in $\Omega \times (T/2, T)$, hence

$$\begin{aligned}&\int_{T/2}^{T}\int_{\Omega} e^{-2s\alpha}\xi^3\left(|\varphi_1|^2+|\varphi_2|^2\right)dxdt + \int_{T/2}^{T}\int_{\Omega} e^{-2s\alpha}\xi^3\left(|\psi_1|^2+|\psi_2|^2\right)dxdt\\ &= \int_{T/2}^{T}\int_{\Omega} e^{-2s\beta}\gamma^3\left(|\varphi_1|^2+|\varphi_2|^2\right)dxdt + \int_{T/2}^{T}\int_{\Omega} e^{-2s\beta}\gamma^3\left(|\psi_1|^2+|\psi_2|^2\right)dxdt.\end{aligned}$$

Therefore, from (25) and the definition of β and γ we obtain

$$\begin{aligned}&\int_{T/2}^{T}\int_{\Omega} e^{-2s\beta}\gamma^3\left(|\varphi_1|^2+|\varphi_2|^2\right)dxdt + \int_{T/2}^{T}\int_{\Omega} e^{-2s\beta}\gamma^3\left(|\psi_1|^2+|\psi_2|^2\right)dxdt\\ &\qquad\leq C\left(\iint_{\omega_0\times(0,T)} e^{-2s\beta}\gamma^7\left(|\varphi_1|^2+|\varphi_2|^2\right)dxdt\right).\end{aligned} \tag{32}$$

On the other hand, for the domain $\Omega \times (0, T/2)$, we will use energy estimates for system (24). In fact, let us introduce a function $\eta \in C^1([0, T])$ such that

$$\eta = 1 \text{ in } [0, T/2], \quad \eta = 0 \text{ in } [3T/4, T], \quad |\eta'(t)| \leq C/T.$$

Using classical energy estimates for $\eta\varphi_1$ and $\eta\varphi_2$ solution to the first and second equation of system (24) we obtain

$$\begin{aligned}&\|\varphi_1(0)\|^2_{L^2(\Omega)} + \|\varphi_2(0)\|^2_{L^2(\Omega)} + \|\varphi_1\|^2_{L^2(0,T/2;H_0^1(\Omega))} + \|\varphi_2\|^2_{L^2(0,T/2;H_0^1(\Omega))}\\ &\leq C\left(\frac{1}{T^2}\|\varphi_1\|^2_{L^2(T/2,3T/4;L^2(\Omega))} + \frac{1}{T^2}\|\varphi_2\|^2_{L^2(T/2,3T/4;L^2(\Omega))}\right.\\ &\qquad\left. + \|\eta\psi_1\|^2_{L^2(0,3T/4;L^2(\Omega))} + \|\eta\psi_2\|^2_{L^2(0,3T/4;L^2(\Omega))}\right).\end{aligned}$$

From the definition of η and adding $\|\psi_j\|^2_{L^2(0,T/2;L^2(\Omega))}$ on both sides of the previous inequality we have

$$\begin{aligned}&\|\varphi_1(0)\|^2_{L^2(\Omega)} + \|\varphi_2(0)\|^2_{L^2(\Omega)} + \sum_{i=1}^{2}\|\varphi_j\|^2_{L^2(0,T/2;L^2(\Omega))} + \sum_{j=1}^{2}\|\psi_j\|^2_{L^2(0,T/2;L^2(\Omega))}\\ &\leq C\left(\sum_{j=1}^{2}\|\varphi_j\|^2_{L^2(T/2,3T/4;L^2(\Omega))} + \sum_{j=1}^{2}\|\psi_j\|^2_{L^2(T/2,3T/4;L^2(\Omega))} + \sum_{j=1}^{2}\|\psi_j\|^2_{L^2(0,T/2;L^2(\Omega))}\right).\end{aligned} \tag{33}$$

In order to eliminate the terms $\|\psi_j\|^2_{L^2(0,T/2;L^2(\Omega))}$ in the right hand side, we use standard energy estimates for the third and fourth equation in (24), thus

$$\iint_{\Omega\times(0,T/2)}(|\psi_1|^2+|\psi_2|^2)dxdt \le C\left(\frac{\alpha_1^2}{\mu_1^2}+\frac{\alpha_2^2}{\mu_2^2}\right)\iint_{\Omega\times(0,T/2)}|\varphi_1|^2dxdt. \tag{34}$$

Replacing (34) in (35) and since μ_i, $i=1,2$, are large enough we obtain

$$\begin{aligned}&\|\varphi_1(0)\|^2_{L^2(\Omega)}+\|\varphi_2(0)\|^2_{L^2(\Omega)}+\sum_{i=1}^{2}\|\varphi_j\|^2_{L^2(0,T/2;L^2(\Omega))}+\sum_{j=1}^{2}\|\psi_j\|^2_{L^2(0,T/2;L^2(\Omega))}\\ &\quad\le C\left(\sum_{j=1}^{2}\|\varphi_j\|^2_{L^2(T/2,3T/4;L^2(\Omega))}+\sum_{j=1}^{2}\|\psi_j\|^2_{L^2(T/2,3T/4;L^2(\Omega))}\right).\end{aligned} \tag{35}$$

Using (32) to estimate the first four terms in the right hand side of (35) and taking into account that the weight functions are bounded in $[0,3T/4]$ we have the estimate

$$\begin{aligned}&\|\varphi_1(0)\|^2_{L^2(\Omega)}+\|\varphi_2(0)\|^2_{L^2(\Omega)}+\int_0^{T/2}\!\!\int_\Omega e^{-2s\beta}\gamma^3\left(|\varphi_1|^2+|\varphi_2|^2\right)dxdt\\ &\quad+\int_0^{T/2}\!\!\int_\Omega e^{-2s\beta}\gamma^3\left(|\psi_1|^2+|\psi_2|^2\right)dxdt\le C\left(\iint_{\omega\times(0,T)}e^{-2s\beta}\gamma^7\left(|\varphi_1|^2+|\varphi_2|^2\right)dxdt\right).\end{aligned}$$

This estimate, together with (32), and the definitions of γ^* and β^* yield the desired inequality (31).

Now we conclude the proof of Proposition 1. To this end, define $\rho(t)=e^{s\beta^*}$. Thus, $\rho(t)$ is a non-decreasing strictly positive function blowing up at $t=T$. We obtain energy estimates with this new weight function for (θ_1^i,θ_2^i) solution to the third and fourth equation of system (20). More precisely,

$$\iint_Q \rho^{-2}(|\theta_1^i|^2+|\theta_2^i|^2)dxdt\le C\iint_{\omega_i\times(0,T)}\rho^{-2}|\varphi_1|^2dxdt,\quad i=1,2.$$

Since $e^{-2s\beta}\gamma^7\le C$ for all $(x,t)\in Q$ and noting that the right hand side of the previous inequality is comparable to the left hand side of inequality (31) up to a multiplicative constant, we obtain (21). This concludes the proof of Proposition 1.

4 Proof of Theorem 2

In this section, we present a Stackelberg-Nash strategy where the leader control acts only on the first equation of system (1). As mentioned before, one important subject in the controllability of non-scalar system is the possibility to control many equations with few controls.

Under the assumptions on the leader control, system (1) can be written as

$$\begin{cases} y_{1,t} - \Delta y_1 + a_{11}y_1 + a_{12}y_2 = h\chi_\omega + v^1\chi_{\omega_1} + v^2\chi_{\omega_2} \quad \text{in } Q, \\ y_{2,t} - \Delta y_2 + a_{21}y_1 + a_{22}y_2 = 0 \quad \text{in } Q, \\ y_j = 0 \text{ on } \Sigma, \ j = 1, 2, \\ y_j(x,0) = y_j^0(x) \text{ in } \Omega, \ j = 1, 2. \end{cases} \tag{36}$$

where $a_{ij} \in L^\infty(Q)$ and $y_j^0 \in L^2(\Omega)$ are given.

Recall that we consider the follower functionals

$$\widetilde{J}_i(h, v^1, v^2) = \frac{\alpha_i}{2}\iint_{\mathscr{O}_d\times(0,T)} |y_2 - y_{2,d}^i|^2 dxdt + \frac{\mu_i}{2}\iint_{\omega_i\times(0,T)} |v^i|^2 dxdt, \ i = 1, 2. \tag{37}$$

By optimizing only the second component of the solution to (36) in the follower step, we are able to obtain an observability inequality with only one observation term in the right-hand side.

Adapting the methods discussed in [5] or [17], we guarantee the existence and uniqueness of the Nash equilibrium if μ_i are large enough. Also, we can easily verify that the pair $(\bar{v}_1, \bar{v}_2)$ is a Nash equilibrium for (37) if and only if

$$\bar{v}_i = -\frac{1}{\mu_i} p_1^i, \ i = 1, 2,$$

where p_1^j, $j = 1, 2$, is a component of the solution to the coupled system

$$\begin{cases} y_{1,t} - \Delta y_1 + a_{11}y_1 + a_{12}y_2 = h\chi_\omega - \frac{1}{\mu_1} p_1^1 \chi_{\omega_1} - \frac{1}{\mu_2} p_1^2 \chi_{\omega_2} & \text{in } Q, \\ y_{2,t} - \Delta y_2 + a_{21}y_1 + a_{22}y_2 = 0 & \text{in } Q, \\ -p_{1,t}^i - \Delta p_1^i + a_{11}p_1^i + a_{21}p_2^i = 0 & \text{in } Q, \\ -p_{2,t}^i - \Delta p_2^i + a_{12}p_1^i + a_{22}p_2^i = \alpha_i \left(y_2 - y_{2,d}^i\right)\chi_{\mathscr{O}_d} & \text{in } Q, \\ y_j(0) = y_j^0, \ p_j^i(T) = 0, \quad y_j = p_j^i = 0 \text{ on } \Sigma, \ i, j = 1, 2. \end{cases} \tag{38}$$

Our task is now to establish an appropriate observability estimate for the solutions to the simplified adjoint system to (38), that is

$$\begin{cases} -\varphi_{1,t} - \Delta\varphi_1 + a_{11}\varphi_1 + a_{21}\varphi_2 = 0 & \text{in } Q, \\ -\varphi_{2,t} - \Delta\varphi_2 + a_{12}\varphi_1 + a_{22}\varphi_2 = \psi_2\chi_{\mathscr{O}_d} & \text{in } Q, \\ \psi_{1,t} - \Delta\psi_1 + a_{11}\psi_1 + a_{12}\psi_2 = -\left(\frac{\alpha_1}{\mu_1}\chi_{\omega_1} + \frac{\alpha_2}{\mu_2}\chi_{\omega_2}\right)\varphi_1 & \text{in } Q, \\ \psi_{2,t} - \Delta\psi_2 + a_{21}\psi_1 + a_{22}\psi_2 = 0 & \text{in } Q, \\ \varphi_j(T) = f_j,\ \psi_j(0) = 0 \text{ in } \Omega, \quad \varphi_j = \psi_j = 0 \text{ on } \Sigma,\ j = 1, 2, \end{cases} \tag{39}$$

where we have used the same change of variable $\psi_j = \alpha_1\theta_j^1 + \alpha_2\theta_j^2$ as in (24). Note that systems (39) and (24) are almost identical except for the right-hand side of the first equation.

We have the following result:

Proposition 4 *Under assumptions of Theorem 2. There exists a positive constant C such that the solution (φ, ψ) to (39) satisfies*

$$\begin{aligned} &I_3(s,\lambda;\varphi_1) + I_3(s,\lambda;\varphi_2) + I_3(s,\lambda;\psi_1) + I_3(s,\lambda;\psi_2) \\ &\qquad \le C \iint_{\omega\times(0,T)} e^{-2s\alpha} s^{31}\lambda^{32}\xi^{31}|\varphi_1|^2 dxdt, \end{aligned} \tag{40}$$

for any s and λ large enough and for every $(f_1, f_2) \in [L^2(\Omega)]^2$.

Proof The proof is similar to the proof of Proposition 2. We define $\tilde{\omega}_0 := \omega \cap \mathscr{O}_d$ and consider subsets $\tilde{\omega}_i$, $i = 1, 2, 3$, such that

$$\tilde{\omega}_3 \subset\subset \tilde{\omega}_2 \subset\subset \tilde{\omega}_1 \subset\subset \tilde{\omega}_0. \tag{41}$$

We apply Carleman inequality (23) to each equation in (39) with $m = 3$ and $\mathscr{B} = \tilde{\omega}_3$. Adding them up and arguing as in the proof of Proposition 2 we can use the parameters s and λ to absorb the lower order terms. More precisely, we obtain

$$\begin{aligned} &I_3(s,\lambda;\varphi_1) + I_3(s,\lambda;\varphi_2) + I_3(s,\lambda;\psi_1) + I_3(s,\lambda;\psi_2) \\ &\qquad \le C\left(\sum_{j=1}^{2} I_{3,\tilde{\omega}_3}(s,\lambda;\varphi_j) + \sum_{j=1}^{2} I_{3,\tilde{\omega}_3}(s,\lambda;\psi_j)\right), \end{aligned} \tag{42}$$

for all λ and s large enough.

The next step is to eliminate the local terms corresponding to ψ_1. Unlike the proof of Proposition 2, we cannot longer use the equation that satisfies φ_1 to estimate this term. We will use the sign condition (12) and the fourth equation of system (39) to estimate ψ_1 locally.

We set λ to a fixed value large enough. Given the sets (41), we consider functions $\zeta_k \in C^\infty(\mathbb{R}^N)$ verifying:

$$0 \le \zeta_k \le 1 \text{ in } \Omega, \quad \zeta_k \equiv 1 \quad \text{in } \tilde{\omega}_k, \quad \operatorname{supp} \zeta_k \subset \tilde{\omega}_{k-1}, \tag{43}$$

$$\frac{\Delta \zeta_k}{\zeta_k^{1/2}} \in L^\infty(\Omega), \quad \frac{\nabla \zeta_k}{\zeta_k^{1/2}} \in L^\infty(\Omega)^N, \quad k = 1, 2, 3. \tag{44}$$

We define $u_3 := e^{-2s\alpha} s^3 \lambda^4 \xi^3$. Recall that the coefficient a_{21} satisfies (12) and, for simplicity, assume that $a_{21} \ge a_0$ in $\omega \cap \mathscr{O}_d \times (0, T)$. We multiply the equation satisfied by ψ_2 in system (39) by $u_3 \zeta_3 \psi_1$ and integrate in Q. We obtain

$$\begin{aligned} a_0 \iint_{\tilde{\omega}_3 \times (0,T)} e^{-2s\alpha} s^3 \lambda^4 \xi^3 |\psi_1|^2 &\le \iint_Q u_3 \zeta_3 a_{21} |\psi_1|^2 \\ &= \iint_Q (-\psi_{2,t} + \Delta \psi_2 - a_{22} \psi_2) u_3 \zeta_3 \psi_1. \end{aligned} \tag{45}$$

Integrating by parts in the right-hand side of (49) and using (27), (43)–(44), it is not difficult to see that

$$a_0 \iint_{\tilde{\omega}_3 \times (0,T)} e^{-2s\alpha} s^3 \lambda^4 \xi^3 |\psi_1|^2 \le \varepsilon I_3(s, \lambda; \psi_1) + C_\varepsilon \iint_{\tilde{\omega}_2 \times (0,T)} e^{-2s\alpha} s^7 \lambda^8 \xi^7 |\psi_2|^2 \tag{46}$$

for any $\varepsilon > 0$. Choosing ε small enough, we obtain from (46) and (42)

$$\begin{aligned} I_3(s, \lambda; \varphi_1) + I_3(s, \lambda; \varphi_2) + I_3(s, \lambda; \psi_1) + I_3(s, \lambda; \psi_2) \\ \le C \left(\sum_{j=1}^{2} I_{3,\tilde{\omega}_3}(s, \lambda; \varphi_j) + I_{7,\tilde{\omega}_2}(s, \lambda; \psi_2) \right), \end{aligned} \tag{47}$$

We proceed to estimate the local term of ψ_2. Set $u_2 = e^{-2s\alpha} s^7 \lambda^8 \xi^7$ and multiply in $L^2(Q)$ the second equation in (39) by $u_2 \zeta_2 \psi_2$. We get

$$\begin{aligned} \iint_{\tilde{\omega}_2 \times (0,T)} e^{-2s\alpha} s^7 \lambda^8 \xi^7 |\psi_2|^2 &\le \iint_{\mathscr{O}_d \times (0,T)} u_2 \zeta_2 |\psi_2|^2 \\ &= \iint_Q (-\varphi_{2,t} - \Delta \varphi_2 + a_{12} \varphi_1 + a_{22} \varphi_2) u_2 \zeta_2 \psi_2. \end{aligned}$$

Proceeding as before, we can readily obtain

$$\begin{aligned} \iint_{\tilde{\omega}_2 \times (0,T)} e^{-2s\alpha} s^7 \lambda^8 \xi^7 |\psi_2|^2 &\le \varepsilon I_7(s, \lambda; \psi_2) + \varepsilon \iint_Q e^{-2s\alpha} s^3 \lambda^4 \xi^3 |\psi_1|^2 \\ &+ C_\varepsilon \iint_{\tilde{\omega}_1 \times (0,T)} e^{-2s\alpha} \left(s^{11} \lambda^{12} \xi^{11} |\varphi_1|^2 + s^{15} \lambda^{16} \xi^{15} |\varphi_2|^2 \right). \end{aligned}$$

Therefore, replacing the above expression in (47), we have

$$
\begin{aligned}
I_3(s,\lambda;\varphi_1)+I_3(s,\lambda;\varphi_2)+I_3(s,\lambda;\psi_1)+I_3(s,\lambda;\psi_2)&\\
\leq C\left(I_{11,\tilde{\omega}_1}(s,\lambda;\varphi_1)+I_{15,\tilde{\omega}_1}(s,\lambda;\varphi_2)\right).&
\end{aligned}
\tag{48}
$$

To estimate the local term of φ_2, we use again that a_{21} satisfies (12). Thus, multiplying the first equation in (39) by $u_1\zeta_1\varphi_2$

$$
\begin{aligned}
a_0\iint_{\tilde{\omega}_1\times(0,T)} e^{-2s\alpha}s^{15}\lambda^{16}\xi^{15}|\varphi_2|^2 &\leq \iint_Q u_1\zeta_1 a_{21}|\varphi_2|^2\\
&=\iint_Q(\varphi_{2,t}+\Delta\varphi_1-a_{11}\varphi_1)u_1\zeta_1\varphi_2
\end{aligned}
\tag{49}
$$

whence, integrating by parts, we obtain

$$
\begin{aligned}
a_0\iint_{\tilde{\omega}_1\times(0,T)} &e^{-2s\alpha}s^{15}\lambda^{16}\xi^{15}|\varphi_2|^2\\
&\leq \varepsilon I_3(s,\lambda;\varphi_2)+C_\varepsilon\iint_{\tilde{\omega}_0\times(0,T)} e^{-2s\alpha}s^{31}\lambda^{32}\xi^{31}|\varphi_1|^2.
\end{aligned}
$$

Finally, replacing the above estimate in (48) with $\varepsilon>0$ small enough and since $\tilde{\omega}_0\subset\omega$, we obtain the desired result. This concludes the proof of Proposition 4.

With the new Carleman estimate (40), we can obtain an observability inequality following the procedure of Sect. 3.3. Such inequality will only have φ_1 as an observation term in the right-hand side and will imply the null controllability of (38) with one leader control. This concludes the proof of Theorem 2.

5 Concluding Remarks

The first main result of this chapter can be easily extended to the control problem

$$
\begin{cases}
y_{1,t}-\Delta y_1+a_{11}y_1+a_{12}y_2=h_1\chi_{\omega_1}+v^1\chi_{\mathcal{O}_1}+v^2\chi_{\mathcal{O}_2} & \text{in } Q,\\
y_{2,t}-\Delta y_2+a_{21}y_1+a_{22}y_2=h_2\chi_{\omega_2} \quad \text{in } Q,\\
y_j(x,0)=y_{j,0} \text{ in } \Omega, \quad y_j=0 \text{ on } \Sigma,\ j=1,2,
\end{cases}
\tag{50}
$$

as long as $\omega_1\cap\omega_2\neq\emptyset$. Indeed, it is enough to consider a set $\omega_0\subset\subset\omega_1\cap\omega_2$ and then apply the results of this paper to this new set to obtain a hierarchic control result. However, the same is not true when $\omega_1\cap\omega_2=\emptyset$. The techniques shown in this paper fail to obtain an observability inequality as (21) since we cannot use Carleman estimates with different weights (related to ω_1 and ω_2) and eliminate all the local terms that appear on the right hand side. Indeed, to eliminate some of them we will need an upper estimation on the first Carleman weight by the second, and to eliminate the others we will need the contrary. This is due to the fact that we have a system of four equations fully coupled.

The hierarchic control is an interesting and challenging problem because there are many available configurations where the leader and follower controls may be placed, and several controllability constraints that may be imposed. As discussed in [5], some problems have been solved for the scalar problem, but other difficulties arise when dealing with coupled systems. Thus, the results are far from being complete.

Acknowledgements This work was partially supported by CONACyT and UNAM-DGAPA-PAPIIT IN102116.

References

1. Ammar-Khojda, F., Benabdallah, A., Dupaix, C., Kostin, I.: Null controllability of some systems of parabolic type by one control force. ESAIM Control Optim. Calc. Var. **11**(3), 426–448 (2005)
2. Ammar-Khojda, F., Benabdallah, A., Dupaix C., González-Burgos, M.: A generalization of the Kalman rank condition for time-dependent coupled linear parabolic systems. Differ. Equ. Appl. **1**(3), 427–457 (2009)
3. Ammar-Khojda, F., Benabdallah, A., González-Burgos, M., de Teresa, L.: Recent results on the controllability of linear coupled parabolic problems: a survey. Math. Control Relat. Fields **1**(3), 267–306 (2011)
4. Araruna, F.D., de Menezes, S.D.B., Rojas-Medar, M.A.: On the approximate controllability of Stackelberg-Nash strategies for linearized micropolar fluids. Appl. Math. Optim. **70**(3), 373–393 (2014)
5. Araruna, F.D., Fernández-Cara E., Santos, M.C.: Stackelberg-Nash exact controllability for linear and semilinear parabolic equations. ESAIM Control Optim. Calc. Var. **21**(3), 835–856 (2015)
6. Araruna, F. D., Fernández-Cara, E., Guerrero, S., Santos, M. C. New results on the Stackelberg-Nash exact control of linear parabolic equations. Systems Control Lett. **104**, 78–85 (2017).
7. Calsarava, B.M.R., Carreño, N., Cerpa, E.: Insensitizing controls for a phase field system. Nonlinear Anal. **143**, 120–137 (2016)
8. Chakrabarty, S.P., Hanson, F.B.: Optimal control of drug delivery to brain tumors for a distributed parameters model. In: Proceedings of the American Control Conference, pp. 973–978 (2005)
9. Corrias, L., Perthame, B., Zaag, H.: Global solutions of some chemotaxis and angiogenesis systems in high space dimensions. Milan J. Math. **72**, 1–28 (2004)
10. Díaz, J.I.: On the von Neumann problem and the approximate controllability of Stackelberg-Nash strategies for some environmental problems. Rev. Real Acad. Cienc. Exact. Fís. Natur. **96**(3), 343–356 (2002)
11. de Teresa, L.: Insensitizing controls for a semilinear heat equation. Commun. Partial Differ. Equ. **25**(1–2), 39–72 (2000)
12. Fernández-Cara, E., Guerrero, S.: Global Carleman inequalities for parabolic systems and applications to controllability. SIAM J. Control Optim. **45**(4), 1395–1446 (2006)

13. Fernández-Cara, E., Guerrero, S., Imanuvilov, O.Y., Puel, J.P.: Local exact controllability of the Navier-Stokes system. J. Math. Pures Appl. **83**(12), 1501–1542 (2004)
14. Fursikov, A., Imanuvilov, O.Y.: Controllability of Evolution Equations. Lecture Notes, Research Institute of Mathematics. Seoul National University, Seoul (1996)
15. González-Burgos, M., de Teresa, L.: Controllability results for cascade systems of *m* coupled parabolic PDEs by one control force. Port. Math. **67**(1), 91–113 (2010)
16. Guillén-González, F., Marques-Lopes, F., Rojas-Medar, M.: On the approximate controllability of Stackelberg-Nash strategies for Stokes equations. Proc. Am. Math. Soc. **141**(5), 1759–1773 (2013)
17. Hernández-Santamaría, V., de Teresa, L., Poznyak, A.: Hierarchic control for a coupled parabolic system. Port. Math. **73**(2), 115–137 (2016)
18. Imanuvilov, O.Y., Yamamoto, M.: Carleman inequalities for parabolic equations in Sobolev spaces of negative order and exact controllability for semilinear parabolic equations. Publ. Res. Inst. Math. Sci. **39**(2), 227–274 (2003)
19. Limaco, F., Clark, H.R., Medeiros, L.A.: Remarks on hierarchic control. J. Math. Anal. Appl. **359**(1), 368–383 (2009)
20. Lions, J.-L.: Hierarchic control. Indian Acad. Sci. Proc. Math. Sci. **104**(1), 295–304 (1994)
21. Lions, J.-L.: Some remarks on Stackelberg's optimization. Math. Models Methods Appl. Sci. **4**(4), 477–487 (1994)
22. Nagai, T., Senba, T., Suzuki, T.: Chemotactic collapse in a parabolic system of mathematical biology. Hiroshima Math. J. **30**(3), 463–497 (2000)
23. Nash, J.F.: Non-cooperative games. Ann. Math. **54**(2), 286–295 (1951)
24. von Stackelberg, H.: Marktform und Gleichgewicht. Springer, Vienna (1934)
25. Zabczyk, J.: Mathematical Control Theory: An Introduction. Birkhäuser, Boston (1992)

Local Null Controllability of the N-Dimensional Ladyzhenskaya-Smagorinsky with N-1 Scalar Controls

Dany Nina Huaman, Juan Límaco, and Miguel R. Nuñez Chávez

Dedicated to Prof. Enrique Fernández-Cara on the occasion of his 60th birthday.

Abstract This paper deals with the null controllability of a differential turbulence model of the Ladyzhenskaya-Smagorinsky kind. In the equations, we find local and nonlocal nonlinearities: the usual transport terms and a turbulent viscosity that depends on the global in space energy dissipated by the mean flow. We prove that the N-systems are locally null-controllable with N-1 scalar controls in an arbitrary control domain.

Keywords Ladyzhenskaya-Smagorinsky · Boussinesq · Null controllability · Carleman inequalities

1 Introduction and Main Results

Let $\Omega \subset \mathbb{R}^N$ $(N = 2$ or $3)$ be a non-empty bounded connected open set with boundary $\Gamma = \partial\Omega$ in the class C^∞. We fix $T > 0$ and denote by Q the cylinder $Q = \Omega \times (0, T)$ with lateral boundary $\Sigma = \partial\Omega \times (0, T)$. We also consider be a (small) non-empty open set $\omega \subset \Omega$ which is the control domain.

In the sequel, we denoted by $(\cdot, \cdot)$ and $\|\cdot\|$, respectively, the L^2 inner products and norms in Ω and Q. The symbol C will be used to design a generic positive constant.

D. N. Huaman · J. Límaco (✉) · M. R. Nuñez Chávez
Instituto de Matemática e Estatística, UFF, Niterói, RJ, Brazil
e-mail: jlimaco@vm.uff.br

A. Doubova et al. (eds.), *Recent Advances in PDEs: Analysis, Numerics and Control*, SEMA SIMAI Springer Series 17,
https://doi.org/10.1007/978-3-319-97613-6_8

We will be concerned with the null controllability of the following Ladyzhenskaya–Smagorinsky systems:

$$\begin{cases} y_t - \nabla \cdot ((\nu_0 + \nu_1 \, \|\nabla y\|^2) Dy) + (y \cdot \nabla) y + \nabla p = v 1_\omega & \text{in } Q, \\ \nabla \cdot y = 0 & \text{in } Q, \\ y = 0 & \text{on } \Sigma, \\ y(0) = y_0 & \text{in } \Omega, \end{cases} \tag{1.1}$$

and

$$\begin{cases} y_t - \nabla \cdot ((\nu_0 + \nu_1 \|\nabla y\|^2) Dy) + (y \cdot \nabla) y + \nabla p = v 1_\omega + \theta e_N & \text{in } Q, \\ \nabla \cdot y = 0 & \text{in } Q, \\ \theta_t - \nabla \cdot ((\nu_0 + \nu_1 \|\nabla y\|^2) \nabla \theta) + y \cdot \nabla \theta = v_0 1_\omega & \text{in } Q, \\ y = 0, \;\; \theta = 0 & \text{on } \Sigma, \\ y(0) = y_0, \;\; \theta(0) = \theta_0 & \text{in } \Omega. \end{cases} \tag{1.2}$$

Here, $y = y(x,t)$, $\theta = \theta(x,t)$ and $p = p(x,t)$ represent, respectively, the "averaged" velocity field, temperature and pressure of a turbulent fluid whose particles are in Ω during the time interval $(0, T)$; y_0 is the averaged velocity at time $t = 0$; 1_ω is the characteristic function of ω; ν_0 and ν_1 are positive constants and Dy stands for the symmetrized gradient of y: $Dy = \nabla y + \nabla^T y$.

On the other hand, $\omega \times (0, T)$ is the control domain and v and v_0 must be viewed as controls (averaged forces) acting on the systems.

The systems (1.1) and (1.2) are generalizations of the Navier-Stokes model and Boussinesq model, respectively. The controllability of Navier-Stokes system has been the objective of considerable work over the last years (see [5] and [12]), analogously to Boussinesq system (see [6] and [9]). Also, the controllability of N-dimensional Navier-Stokes with N-1 scalar controls has been studied in [6] and [3], analogously to Boussinesq system (see [2]). The first work in the study of the controllability of (1.1) was in the paper [7], where it has been studied the null controllability with N scalar controls and it also gives details of a numerical approximation. We will present some new results which show that the N-dimensional systems (1.1) and (1.2) can be controlled with N-1 scalar controls in $L^2(\omega \times (0, T))$ in an arbitrary control domain. For the system (1.2) there wasn't a result about controllability so far, then we got the first result controllability for (1.2).

The following vector spaces, usually in the context of incompressible fluids, will be used along the paper

$$H = \{w \in L^2(\Omega)^N : \nabla \cdot w = 0 \text{ in } \Omega, \, w \cdot \eta = 0 \text{ on } \partial\Omega\}$$

and

$$V = \{w \in H_0^1(\Omega)^N : \nabla \cdot w = 0 \text{ in } \Omega\}$$

We will denote by $A : D(A) \longrightarrow H$ the Stokes operator. By definition, one has $Aw = P(-\Delta w)$, where $P : L^2(\Omega)^N \longrightarrow H$ is the orthogonal projector and $D(A) = H^2(\Omega)^N \cap V$.

The system (1.1), for $N = 2$, $y_0 \in V$, $v \in L^2(\omega \times (0, T))^N$, possesses exactly one strong solution (y, p) with

$$y \in L^2(0, T; D(A)) \cap C([0, T]; V),\ \ y_t \in L^2(0, T; H).$$

For $N = 3$, this is also true if y_0 and v are sufficiently small in their respective spaces.

The system (1.2), for any $y_0 \in V$, $\theta_0 \in H_0^1(\Omega)$ and any $v \in L^2(\omega \times (0, T))^N$ sufficiently small in their respective spaces, possesses exactly one strong solution (y, p, θ), with

$$y \in L^2(0, T; D(A)) \cap C([0, T]; V),\ \ \ y_t \in L^2(0, T; H)$$

and

$$\theta \in L^2(0, T; H^2(\Omega) \cap H_0^1(\Omega)) \cap C([0, T]; H_0^1(\Omega)),\ \ \ \theta_t \in L^2(Q).$$

These assertions can be deduced arguing as in [11].

This paper concerns the local null controllability of the systems (1.1) and (1.2) at time $t = T$ with a reduced number of controls, we remove the geometric assumption on ω considered in [6]. The present work can be viewed as an extension of [7].

In this paper, the main results are the following:

Theorem 1.1 *Let $i \in \{1, \ldots, N\}$. Then, for every $T > 0$ and $\omega \subset \Omega$, there exists $\delta > 0$ such that, for every $y_0 \in V$ satisfying*

$$\|y_0\|_V \leq \delta$$

we can find a control $v \in L^2(\omega \times (0, T))^N$, with $v_i \equiv 0$ such that the corresponding solution y to (1.1) satisfies

$$y(T) = 0\ \ in\ \ \Omega.$$

i.e., the nonlinear system (1.1) is locally null controllable by means of N-1 scalar controls for an arbitrary control domain.

Analogously, we can to prove the local null controllability of system (1.2).

Theorem 1.2 *Let $i \in \{1, \ldots, N\}$. Then, for every $T > 0$ and $\omega \subset \Omega$, there exists $\delta > 0$ such that, for every $(y_0, \theta_0) \in V \times H_0^1(\Omega)$ satisfying*

$$\|(y_0, \theta_0)\|_{V \times H_0^1(\Omega)} < \delta$$

we can find controls $v \in L^2(\omega \times (0,T))^N$ *and* $v_0 \in L^2(\omega \times (0,T))$*, with* $v_i \equiv 0$ *and* $v_N \equiv 0$*, such that the corresponding solution* (y,θ) *to (1.2) satisfies*

$$y(T) = 0 \ \ and \ \ \theta(T) = 0 \ \ in \ \ \Omega.$$

i.e., the nonlinear system (1.2) is locally null controllable by means of N-1 scalar controls for an arbitrary control domain.

To prove Theorems 1.1 and 1.2, we follow a standard approach (see for instance [2, 3, 5, 6]), using an Inverse Mapping Theorem. We first deduce a null controllability result for linear systems associated to (1.1) and (1.2):

$$\begin{cases} y_t - \nu_0 \Delta y + \nabla p = v 1_\omega + f & \text{in } Q, \\ \nabla \cdot y = 0 & \text{in } Q, \\ y = 0 & \text{on } \Sigma, \\ y(0) = y_0 & \text{in } \Omega, \end{cases} \tag{1.3}$$

and

$$\begin{cases} y_t - \nu_0 \Delta y + \nabla p = f + v 1_\omega + \theta e_N & \text{in } Q, \\ \nabla \cdot y = 0 & \text{in } Q, \\ \theta_t - \nu_0 \Delta \theta = f_0 + v_0 1_\omega & \text{in } Q, \\ y = 0, \ \ \theta = 0 & \text{on } \ \Sigma, \\ y(0) = y_0, \ \ \theta(0) = \theta_0 & \text{in } \Omega. \end{cases} \tag{1.4}$$

This paper is organized as follows. In Sect. 2 we deal with null controllability results for the linear control systems (1.3) and (1.4) using a different Carleman estimate (see [2, 3]). The proof of Theorem 1.1 will be given in Sect. 3. Analogously, we will prove Theorem 1.2 in Sect. 4. Finally, Sect. 5 deals with some open questions.

2 Some Technical Results

2.1 Carleman Estimates

We will present the *Carleman inequalities* for the adjoint systems of (1.3) and (1.4), these are given by

$$\begin{cases} -\varphi_t - \nu_0 \Delta \varphi + \nabla \pi = g & \text{in } Q, \\ \nabla \cdot \varphi = 0 & \text{in } Q, \\ \varphi = 0 & \text{on } \Sigma, \\ \varphi(T) = \varphi^T & \text{in } \Omega, \end{cases} \tag{2.1}$$

and

$$\begin{cases} -\varphi_t - \nu_0 \Delta\varphi + \nabla\pi = g & \text{in } Q, \\ \nabla\cdot\varphi = 0 & \text{in } Q, \\ -\psi_t - \nu_0\Delta\psi = g_0 + \varphi e_N & \text{in } Q, \\ \varphi = 0, \;\; \psi = 0 & \text{on } \Sigma, \\ \varphi(T) = \varphi^T, \;\; \psi(T) = \psi^T & \text{in } \Omega. \end{cases} \tag{2.2}$$

In order to do so, we will need some (well-known) results from Fursikov and Imanuvilov [10], see also [8]. Also, it will be convenient to introduce a new nonempty open set ω_0, with $\omega_0 \subset\subset \omega$.

Lemma 2.1 *There exists a function $\eta \in C^2(\overline{\Omega})$ satisfying:*

$$\begin{cases} \eta(x) > 0, \; \forall\, x \in \Omega, \\ \eta(x) = 0, \forall\, x \in \partial\Omega, \\ |\nabla\eta(x)| > 0, \forall\, x \in \overline{\Omega}\setminus\omega_0. \end{cases}$$

Let also $l \in C^\infty[0, T]$ be a positive function satisfying

$$l(t) = \begin{cases} \nu_0 t, & t \in [0, T/4], \\ \nu_0 T - \nu_0 t, & t \in [3T/4, T], \end{cases}$$

and $l(t) \le l(T/2)$, for every $t \in [0, T]$. We introduce a function

$$\tilde{l}(t) = \begin{cases} \|l\|_\infty, & 0 \le t \le T/2, \\ l(t), & T/2 \le t \le T. \end{cases}$$

Then, for all $\lambda \ge 1$, we consider the following weight functions:

$$\beta(x,t) = \frac{\alpha_0(x)}{\tilde{l}^8(t)} = \frac{e^{2\lambda\|\eta\|_\infty} - e^{\lambda\eta(x)}}{\tilde{l}^8(t)}, \;\; \gamma(x,t) = \frac{e^{\lambda\eta(x)}}{\tilde{l}^8(t)},$$

$$\beta^*(t) = \max_{x\in\overline{\Omega}} \beta(x,t), \;\; \gamma^*(t) = \min_{x\in\overline{\Omega}} \gamma(x,t),$$

$$\hat{\beta}(t) = \min_{x\in\overline{\Omega}} \beta(x,t), \;\; \hat{\gamma}(t) = \max_{x\in\overline{\Omega}} \gamma(x,t).$$

There exists $\lambda_{00} > 0$ such that for every $\lambda \ge \lambda_{00}$, we have

$$5 \max_{x\in\overline{\Omega}} \alpha_0(x) < 6 \min_{x\in\overline{\Omega}} \alpha_0(x). \tag{2.3}$$

These exact weight functions were considered in [3].

Our Carleman estimates are given in the following Lemmas:

Lemma 2.2 *There exists a constant* $\lambda_0 > 0$*, such that for any* $\lambda > \lambda_0$ *there exist two constants* $C(\lambda) > 0$ *and* $s_0 > 0$ *such that for any* $i \in \{1, \ldots, N\}$*, and* $g \in L^2(Q)^N$ *and any* $\varphi^T \in H$*, the solution of (2.1) satisfies:*

$$\|\varphi(0)\|^2 + s^4 \iint_Q e^{-5s\beta^*}(\gamma^*)^4 |\varphi|^2 \, dxdt \leq C \left(\iint_Q e^{-3s\beta^*} |g|^2 \, dxdt + s^7 \sum_{j=1, j\neq i}^{N} \iint_{\omega\times(0,T)} e^{-2s\hat{\beta}-3s\beta^*} (\hat{\gamma})^7 \left|\varphi_j\right|^2 dxdt \right) \tag{2.4}$$

for every $s \geq s_0$.

Proof Let φ is the solution of (2.1). We denote $T_{\nu_0} = \nu_0 T$, $Q_{\nu_0} = \Omega \times [0, T_{\nu_0}]$, $\Sigma_{\nu_0} = \partial\Omega \times [0, T_{\nu_0}]$ and we consider the following functions:

$$\varphi_{\nu_0}(x, t) = \varphi(x, t/\nu_0), \;\; \pi_{\nu_0}(x, t) = \frac{1}{\nu_0}\pi(x, t/\nu_0) \;\text{ and } g_{\nu_0}(x, t) = \frac{1}{\nu_0} g(x, t/\nu_0).$$

Notice that φ_ν is the solution φ of:

$$\begin{cases} -(\varphi_{\nu_0})_t - \Delta\varphi_{\nu_0} + \nabla\pi_{\nu_0} = g_{\nu_0} & \text{in } Q, \\ \nabla \cdot \varphi_{\nu_0} = 0 & \text{in } Q, \\ \varphi_{\nu_0} = 0 & \text{on } \Sigma, \\ \varphi_{\nu_0}(T) = \varphi^T_{\nu_0} & \text{in } \Omega. \end{cases}$$

From Lemma 3.1 in [3], we obtain at once (2.4). □

Lemma 2.3 *Assume* $N = 3$*. There exists a constant* $\lambda_0 > 0$ *such that for any* $\lambda > \lambda_0$ *there exists two constants* $C(\lambda) > 0$ *and* $s_0(\lambda) > 0$ *such that for any* $j \in \{1, 2\}$*,* $g \in L^2(Q)^3$*,* $g_0 \in L^2(Q)$*,* $\varphi^T \in H$ *and* $\psi^T \in L^2(\Omega)$*, the solution* (φ, ψ) *of (2.2) satisfies*

$$\begin{aligned} &\iint_Q e^{-5s\beta^*}(\gamma^*)^4|\varphi|^2 dxdt + \iint_Q e^{-5s\beta^*}(\gamma^*)^5|\psi|^2 dxdt + \|\varphi(0)\|^2 + \|\psi(0)\|^2 \\ &\quad \leq C \left(\iint_Q e^{-3s\beta^*}(|g|^2 + |g_0|^2) dxdt + \iint_{\omega\times(0,T)} e^{-2s\hat{\beta}-3s\beta^*}(\hat{\gamma})^7|\varphi_j|^2 dxdt \right. \\ &\qquad \left. + \iint_{\omega\times(0,T)} e^{-4s\hat{\beta}-s\beta^*}(\hat{\gamma})^{\frac{49}{4}}|\psi|^2 dxdt \right) \end{aligned} \tag{2.5}$$

for every $s \geq s_0$.

Proof Analogously as in the proof of Lemma 2.2, making a change of variable and using Lemma 3.1 in [2], we obtain (2.5). □

Let us also state this result for $N = 2$.

Lemma 2.4 *Assume $N = 2$. There exists a constant $\lambda_0 > 0$, such that for any $\lambda > \lambda_0$ there exists two constants $C(\lambda) > 0$ and $s_0(\lambda) > 0$ such that for any $g \in L^2(Q)^2$, $g_0 \in L^2(Q)$, $\varphi^T \in H$ and $\psi^T \in L^2(\Omega)$, the solution (φ, ψ) of (2.2) satisfies*

$$\iint_Q e^{-5s\beta^*}(\gamma^*)^4|\varphi|^2 dxdt + \iint_Q e^{-5s\beta^*}(\gamma^*)^5|\psi|^2 dxdt + \|\varphi(0)\|^2 + \|\psi(0)\|^2 \tag{2.6}$$

$$\leq C\left(\iint_Q e^{-3s\beta^*}(|g|^2 + |g_0|^2)dxdt + \iint_{\omega\times(0,T)} e^{-4s\hat{\beta}-s\beta^*}(\hat{\gamma})^{\frac{49}{4}}|\psi|^2 dxdt\right)$$

for every $s \geq s_0$.

Proof Similarly to Lemma 2.3. □

2.2 Null Controllability of (1.3) and (1.4)

In order to simplify the notation, we fix $\lambda = \lambda_1 > \lambda_0$, $s = s_1 > s_0$ and we set

$$\begin{cases} \rho = e^{5s\beta^*/2}(\gamma^*)^{-2}, \ \rho_0 = e^{3s\beta^*/2}, \ \rho_1 = e^{s\hat{\beta}+3s\beta^*/2}(\hat{\gamma})^{-7/2}, \\ \mu = e^{3s\beta^*/2}(\widehat{\gamma})^{-1}, \ \xi = e^{3s\beta^*/2}(\widehat{\gamma})^{-2}, \\ \hat{\rho} = e^{5s\beta^*/2}(\gamma^*)^{-5/2}, \ \hat{\rho}_0 = e^{3s\hat{\beta}/2}, \ \hat{\rho}_1 = e^{2s\hat{\beta}+s\beta^*/2}\hat{\gamma}^{-49/8}, \\ \zeta = \hat{\rho}\,\bar{l}^{12}, \ \kappa = \hat{\rho}\,\bar{l}^{33/2}. \end{cases} \tag{2.7}$$

With Lemmas 2.2–2.4, we are able to show the null controllability of (1.3) and (1.4) for right-hand sides f and f_0 that decay sufficiently fast to zero as $t \to T$. More precisely, one has:

Proposition 2.1 *Let $i \in \{1, \ldots, N\}$. Assume that $y_0 \in H$ and $\rho f \in L^2(Q)^N$. Then, we can find a control-state pair (y, v) for (1.3) satisfying $v_i \equiv 0$ and*

$$\iint_Q \rho_0^2\,|y|^2\,dxdt + \sum_{\substack{j=1 \\ j\neq i}}^{N} \iint_{\omega\times(0,T)} \rho_1^2\left|v_j\right|^2 dxdt < \infty. \tag{2.8}$$

In particular, one has $y(T) = 0$ and $y \in L^2(0, T; V) \cap C([0, T]; H)$.

Proof The proof of this proposition is very similar to the one of Proposition 3.3 in [3] and Proposition 1 in [6]. □

Analogously, we have

Proposition 2.2 *Let $i \in \{1, \dots, N-1\}$. Assume $y_0 \in H$, $\theta_0 \in L^2(\Omega)$, $\rho f \in L^2(Q)^N$ and $\hat{\rho} f_0 \in L^2(Q)$. Then, we can find control-state (y, θ, v, v_0) for (1.4) satisfying $v_i \equiv v_N \equiv 0$ and*

$$\begin{aligned} &\iint_Q \rho_0^2 |y|^2 dxdt + \iint_Q \rho_0^2 |\theta|^2 dxdt \\ &+ \sum_{\substack{j=1 \\ j \neq i}}^{N-1} \iint_{\omega \times (0,T)} \rho_1^2 |v_j|^2 dxdt + \iint_{\omega \times (0,T)} \hat{\rho}_1^2 |v_0|^2 dxdt < +\infty. \end{aligned} \tag{2.9}$$

In particular, one has $y(T) = 0$, $\theta(T) = 0$, $y \in L^2(0, T; V) \cap C([0, T]; H)$ and $\theta \in L^2(0, T; H_0^1(\Omega)) \cap C([0, T]; L^2(\Omega))$.

Proof Similarly to Proposition 3.4 in [2]. □

2.3 Estimates for the States Solutions

The states solutions found in Propositions 2.1 and 2.2 satisfy some additional properties (those will be needed below, in Sects. 3 and 4). More precisely, we will show that ∇y, $\nabla \theta$, y_t, θ_t, Δy and $\Delta \theta$ belong to weighted L^2 spaces.

Proposition 2.3 *Let the hypotheses in Proposition 2.1 be satisfied and let (y, p, v, f) satisfies (1.3) and (2.8). Then one has*

$$\begin{aligned} \sup_{t \in [0,T]} \int_\Omega \mu^2 \, |y|^2 \, dx + \iint_Q \mu^2 \, |\nabla y|^2 \, dxdt \leq C \Bigg(\|y_0\|_H^2 + \iint_Q \rho_0^2 \, |y|^2 \, dxdt + \\ \iint_Q \rho^2 \, |f|^2 \, dxdt + \iint_{\omega \times (0,T)} \rho_1^2 \, |v|^2 \, dxdt \Bigg). \end{aligned} \tag{2.10}$$

Proof In view of the definitions of μ and ρ_0 in (2.7), one has:

$$|\mu \mu_t| \leq C \rho_0^2.$$

Let us multiply the PDE in (1.3) by $\mu^2 y$ and let us integrate in Ω. We obtain:

$$\mu^2 y (y_t - \nu_0 \Delta y + \nabla p) = (f + v 1_\omega) \mu^2 y.$$

$$\frac{1}{2} \frac{d}{dt} \int_\Omega \mu^2 \, |y|^2 \, dx - \int_\Omega \mu \mu_t \, |y|^2 \, dx + \nu_0 \int_\Omega \mu^2 \, |\nabla y|^2 \, dx = \int_\Omega \mu^2 f y dx + \int_\Omega \mu^2 v 1_\omega y dx.$$

Notice that

- $\left|\int_\Omega \mu^2 v 1_\omega y dx\right| \leq \frac{1}{2}\left(\int_\Omega \rho_1^2 |v|^2 1_\omega dx\right) + C\left(\int_\Omega \rho_0^2 |y|^2 dx\right).$
- $\left|\int_\Omega \mu^2 f y dx dt\right| \leq \frac{1}{2}\int_\Omega \rho^2 |f|^2 dx + C\int_\Omega \rho_0^2 |y|^2 dx.$
- $\left|\int_\Omega \mu\mu_t |y|^2 dx\right| \leq C\int_\Omega \rho_0^2 |y|^2 dx.$

Therefore,

$$\frac{1}{2}\frac{d}{dt}\int_\Omega \mu^2 |y|^2 dx + \nu_0 \int_\Omega \mu^2 |\nabla y|^2 dx \leq \frac{1}{2}\int_\omega \rho_1^2 |v|^2 dx + \frac{1}{2}\int_\Omega \rho^2 |f|^2 dx \\ + C\textstyle\int_\Omega \rho_0^2 |y|^2 dx.$$

We obtain at once (2.10). □

Proposition 2.4 *Under the hypotheses of Proposition 2.3 and let us assume that* $y_0 \in V$. *Then one has*

$$\sup_{t\in[0,T]}\int_\Omega \xi^2 |\nabla y|^2 dx + \iint_Q \xi^2\left(|y_t|^2 + |\Delta y|^2\right) dxdt \leq C\left(\|y_0\|_V^2 + \iint_Q \rho^2 |f|^2 dxdt \right. \\ \left. + \iint_Q \rho_0^2 |y|^2 dxdt + \iint_{\omega\times(0,T)} \rho_1^2 |v|^2 dxdt\right). \tag{2.11}$$

Proof From definition of ξ, we see that

$$|\xi\xi_t| \leq C\mu^2.$$

Let us multiply the PDE in (1.3) by $\xi^2 y_t$ and let us integrate in Ω. The following holds:

$$\int_\Omega \xi^2 |y_t|^2 dx + \frac{1}{2}\frac{d}{dt}\int_\Omega \nu_0 \xi^2 |\nabla y|^2 dx - \nu_0 \int_\Omega \xi\xi_t |\nabla y|^2 dx \\ = \textstyle\int_\Omega \xi^2 f y_t dx + \displaystyle\int_\Omega \xi^2 v 1_\omega y_t dx.$$

- $\left|\int_\Omega \xi^2 v 1_\omega y_t dx\right| \leq C\int_\omega \rho_1^2 |v|^2 dx + \frac{1}{8}\int_\Omega \xi^2 |y_t|^2 dx.$
- $\left|\int_\Omega \xi^2 f y_t dx\right| \leq C\int_\Omega \rho^2 |f|^2 dx + \frac{1}{8}\int_\Omega \xi^2 |y_t|^2 dx.$
- $\left|\int_\Omega \xi\xi_t |\nabla y|^2 dx\right| \leq C\int_\Omega \mu^2 |\nabla y|^2 dx.$

Therefore,

$$\frac{3}{4}\int_{\Omega}\xi^2|y_t|^2\,dx+\frac{1}{2}\frac{d}{dt}\int_{\Omega}\nu_0\xi^2|\nabla y|^2\,dx$$
$$\leq C\left(\int_{\Omega}\rho_1^2|v|^2 1_{\omega}dx+\int_{\Omega}\rho^2|f|^2\,dx+\int_{\Omega}\mu^2|\nabla y|^2\,dx\right).$$

Integrating in time from 0 to T, with $t\leq T$ and from (2.10), we have

$$\frac{3}{4}\int_0^t\int_{\Omega}\xi^2|y_t|^2\,dxds+\frac{1}{2}\nu_0\int_{\Omega}\xi^2(t)\,|\nabla y(t)|^2\,dx$$
$$\leq C\left(\|y_0\|_V^2+\iint_Q\rho^2|f|^2\,dxdt+\iint_Q\rho_1^2|v|^2 1_{\omega}dxdt+\iint_Q\rho_0^2|y|^2\,dxdt\right).$$

Consequently,

$$\iint_Q\xi^2|y_t|^2\,dxdt+\sup_{t\in[0,T]}\int_{\Omega}\xi^2|\nabla y|^2\,dx\leq C\left(\|y_0\|_V^2+\iint_Q\rho^2|f|^2\,dxdt+\right.$$
$$\left.\iint_Q\rho_1^2|v|^2 1_{\omega}dxdt+\iint_Q\rho_0^2|y|^2\,dxdt\right). \tag{2.12}$$

In order to estimates of $\xi^2|\nabla y|^2$ and $\xi^2|\Delta y|^2$, let us multiply the PDE in (1.3) by $\xi^2 Ay$ and integrate in Ω. Then, we have:

$$\int_{\Omega}\xi^2 Ayy_t dx-\nu_0\int_{\Omega}\xi^2 Ay\Delta y dx=\int_{\Omega}f\xi^2 Ay dx++\int_{\Omega}v1_{\omega}\xi^2 Ay dx. \tag{2.13}$$

Notice that $\|Ay\|_H^2\leq C\,\|\Delta y\|^2$, one has

- $\left|\int_{\Omega}\xi^2 v1_{\omega}Aydx\right|\leq\frac{1}{8}\int_{\Omega}\xi^2|\Delta y|^2\,dx+C\int_{\omega}\rho_1^2|v|^2\,dx.$
- $\left|\int_{\Omega}\xi^2 fAydx\right|\leq\frac{1}{8}\int_{\Omega}\xi^2|\Delta y|^2+C\int_{\Omega}\rho^2|f|^2\,dx.$
- $\left|\int_{\Omega}\xi^2 y_t Ay\right|\leq\frac{\nu_0}{8}\int_{\Omega}\xi^2|\Delta y|^2\,dx+C\int_{\Omega}\xi^2|y_t|^2\,dx.$
- $-\nu_0\int_{\Omega}\xi^2 Ay\Delta ydx=\nu_0\int_{\Omega}\xi^2|\Delta y|^2\,dx.$

Using the last equality in (2.13), we obtain that

$$\int_{\Omega}\xi^2|\Delta y|^2\,dx\leq C\left(\int_{\Omega}\rho^2|f|^2\,dx+\int_{\omega}\rho_1^2|v|^2\,dx+\int_{\Omega}\xi^2|y_t|^2\,dx\right).$$

Integrating in time from 0 to T and from Proposition 2.3, one has

$$\iint_Q \xi^2 \,|\Delta y|^2 \,dxdt \\ \le C\left(\|y_0\|_V^2 + \iint_Q \rho^2 \,|f|^2 \,dxdt + \iint_Q \rho_1^2 \,|v|^2 \,1_\omega dxdt + \iint_Q \rho_0^2 \,|y|^2 \,dxdt\right).$$

From this inequality and (2.12), we obtain at once (2.11). □

Now, for the system (1.4), we have

Proposition 2.5 *Let the hypotheses in Proposition 2.2 be satisfied and let* $(y, p, \theta, v, v_0, f, f_0)$ *satisfies (1.4) and (2.9). Then one has*

$$\begin{aligned}\sup_{[0,T]} \int_\Omega \zeta^2|y|^2dx + \iint_Q \zeta^2|\nabla y|^2dxdt \\ \le C \Bigg(\|y_0\|_H^2 + \iint_Q \rho^2|f|^2dxdt + \iint_Q \rho_0^2|y|^2dxdt \\ + \iint_Q \rho_0^2|\theta|^2dxdt + \iint_{\omega\times(0,T)} \rho_1^2|v|^2dxdt\Bigg)\end{aligned}$$

and

$$\begin{aligned}\sup_{[0,T]} \int_\Omega \zeta^2|\theta|^2dx + \iint_Q \zeta^2|\nabla\theta|^2dxdt \le C \Bigg(\|\theta_0\|^2 + \iint_Q \hat\rho^2|f_0|^2dxdt + \iint_Q \rho_0^2|y|^2dxdt \\ + \iint_Q \rho_0^2|\theta|^2dxdt + \iint_{\omega\times(0,T)} \hat\rho_1^2|v_0|^2dxdt\Bigg).\end{aligned}$$

Proof Similar to Proposition 2.3. □

Proposition 2.6 *Under the hypotheses in Proposition 2.5 and let us assume that* $(y_0, \theta_0) \in V \times H_0^1(\Omega)$. *Then one has*

$$\begin{aligned}\sup_{[0,T]} \int_\Omega \kappa^2|\nabla y|^2dx + \iint_Q \kappa^2(|y_t|^2 + |\Delta y|^2)dxdt \le\ C \Bigg(\|y_0\|_V^2 + \iint_Q \rho^2|f|^2dxdt \\ + \iint_Q \rho_0^2|y|^2dxdt + \iint_Q \rho_0^2|\theta|^2dxdt + \iint_{\omega\times(0,T)} \rho_1^2|v|^2dxdt\Bigg)\end{aligned}$$

and

$$\begin{aligned}\sup_{[0,T]} \int_\Omega \kappa^2|\nabla\theta|^2dx + \iint_Q \kappa^2(|\theta_t|^2 + |\Delta\theta|^2)dxdt \le\ C \Bigg(\|\theta_0\|_{H_0^1(\Omega)}^2 + \iint_Q \hat\rho^2|f_0|^2dxdt \\ + \iint_Q \rho_0^2|y|^2dxdt + \iint_Q \rho_0^2|\theta|^2dxdt + \iint_{\omega\times(0,T)} \hat\rho_1^2|v_0|^2dxdt\Bigg).\end{aligned}$$

Proof Similar to Proposition 2.4. □

3 The Proof of Theorem 1.1

In this Section, we will prove the local null controllability for the system (1.1). Let us set $Ly = y_t - \nu_0 \Delta y$ and introduce the space, for $N = 2$ or 3 and $i \in \{1, \ldots, N\}$,

$$E_N^i = \{(y, p, v) : \rho_0 y,\ \rho_1 v 1_\omega \in L^2(Q)^N,\ v_i = 0,\ y \in L^2(0, T; D(A)),$$
$$\rho(Ly + \nabla p - v1_\omega) \in L^2(Q)^N,\ y(0) \in V,\ p \in L^2(0, T; H^1(\Omega))\}.$$

It is clear that E_N^i is a Banach space with the following norm:

$$\|(y, p, v)\|^2_{E_N^i} = \|\rho_0 y\|^2_{L^2(Q)^N} + \|\rho_1 v 1_\omega\|^2_{L^2(Q)^N} + \|p\|^2_{L^2(0,T;H^1(\Omega))} +$$
$$\|y\|^2_{L^2(0,T;D(A))} + \|\rho(Ly + \nabla p - v1_\omega)\|^2_{L^2(Q)^N}.$$

Let us assume that $(y, p, v) \in E_N^i$. Then $y_t \in L^2(Q)^N$, whence $y \in L^2(0, T; D(A))$. One has that $y \in C([0, T]; V)$, in particular, we have $y(0) \in V$ and:

$$\|y(0)\|_V \le C\, \|(y, p, v)\|_{E_N^i},\ \forall\, (y, p, v) \in E_N^i.$$

Furthermore, in view of Propositions 2.3 and 2.4, one also has $\mu y \in L^\infty(0, T; H) \cap L^2(0, T; V)$ and $\xi y \in L^2(0, T; D(A)) \cap L^\infty(0, T; V)$ with

$$\begin{cases} \|\mu y\|^2_{L^\infty(0,T;H)} + \|\mu y\|^2_{L^2(0,T;V)} \le C\, \|(y, p, v)\|^2_{E_N^i}, \\ \|\xi y\|^2_{L^\infty(0,T;V)} + \|\xi y\|^2_{L^2(0,T;D(A))} \le C\, \|(y, p, v)\|^2_{E_N^i}. \end{cases} \tag{3.1}$$

Let us introduce also the space, Banach , $F_N = L^2\left(\rho^2; Q\right)^N \times V$ and the mapping $\mathcal{A} : E_N^i \longrightarrow F_N$, given by:

$$\mathcal{A}(y, p, v) = \Big(y_t - \nabla \cdot ((\nu_0 + \nu_1\, \|\nabla y\|^2) Dy) + (y \cdot \nabla) y + \nabla p - v1_\omega, y(0)\Big). \tag{3.2}$$

Notice that, this definition $\nabla \cdot \left((\nu_0 + \nu_1\, \|\nabla y\|^2) Dy\right)$ can be rewritten (using $\nabla \cdot y = 0$) in the form $\left(\nu_0 + \nu_1\, \|\nabla y\|^2\right) \Delta y$.

We will prove that there exists $\epsilon > 0$ such that, if $(f, y_0) \in F_N$ and $\|(f, y_0)\|_{F_N} \le \epsilon$, then the equation

$$\mathcal{A}(y, p, v) = (f, y_0),\ (y, p, v) \in E_N^i,$$

possesses at least one solution.

In particular, this shows that (1.1) is locally null controllable and, furthermore, the state-control can be chosen in E_N^i. We will apply the following version of *Liusternik's Inverse Mapping Theorem* in infinite dimensional spaces, that can be found for instance in [1]. In the following statement, $B_r(0)$ and $B_\epsilon(\xi_0)$ are open balls with radius r and ϵ, respectively.

Theorem 3.1 *Let Y and Z be Banach spaces and let $H : B_r(0) \subset Y \longrightarrow Z$ be a $\mathcal{C}^1$ mapping. Let us assume that the derivative $H'(0) : Y \longrightarrow Z$ is onto and let us set $H(0) = \xi_0$. Then there exist $\epsilon > 0$ and $W : B_\epsilon(\xi_0) \subset Z \longrightarrow Y$ and $k > 0$ satisfying:*

$$\begin{cases} W(z) \in B_r(0) \text{ and } H(W(z)) = z, \ \forall z \in B_\epsilon(\xi_0), \\ \|W(z)\|_Y \le k \|z - H(0)\|_Z, \ \forall\, z \in B_\epsilon(\xi_0). \end{cases}$$

In order to show that Theorem 3.1 can be applied, we will use several lemmas.

Lemma 3.1 *Let $\mathcal{A} : E^i_N \longrightarrow F_N$ be the mapping defined by (3.2). Then $\mathcal{A}$ is well defined and continuous.*

Proof For any $(y, p, v) \in E^i_N$

$$\iint_Q \rho^2 |\mathcal{A}_1(y,p,v)|^2\,dxdt = \iint_Q \rho^2 \Big| y_t - (\nu_0 + \nu_1 \|\nabla y\|^2)\Delta y + \nabla p + (y \cdot \nabla) y - v1_\omega \Big|^2 dxdt.$$

$$\begin{aligned} \iint_Q \rho^2 |\mathcal{A}_1(y,p,v)|^2\,dxdt \le{}& 3 \iint_Q \rho^2 |y_t - \nu_0 \Delta y + \nabla p - v1_\omega|^2\,dxdt \\ &+ 3 \iint_Q \rho^2 \nu_1^2 \|\nabla y\|^4 |\Delta y|^2\,dxdt \\ &+ 3 \iint_Q \rho^2 |(y\cdot\nabla)y|^2\,dxdt. \end{aligned}$$

$$\iint_Q \rho^2 |\mathcal{A}_1(y,p,v)|^2\,dxdt \le 3I_1 + 3I_2 + 3I_3.$$

Obviously

$$I_1 \le \|(y,p,v)\|^2_{E^i_N}. \tag{3.3}$$

Taking into account that

$\|\nabla w\|_{L^3} \le C \|\nabla w\|^{1/2} \|\Delta w\|^{1/2}$ and $\|(w\cdot\nabla)w\|^2 \le C \|w\|^2_{L^6} \|\nabla w\|^2_{L^3}$, for all $w \in D(A)$, and $H^1(\Omega) \hookrightarrow L^6(\Omega)$, we have:

$$\begin{aligned} I_3 &\le C \int_0^T \rho^2 \|(y\cdot\nabla)y\|^2\,dt \\ &\le C \int_0^T \rho^2 \|y\|^2_{L^6} \|\nabla y\|^2_{L^3}\,dt \\ &\le C \int_0^T \rho^2 \|\nabla y\|^2 \|\nabla y\|^2_{L^3}\,dt \\ &\le C \int_0^T \rho^2 \|\nabla y\|^3 \|\Delta y\|\,dt. \end{aligned} \tag{3.4}$$

From definition of μ and ξ in (2.7), we have

$$\rho^2 \leq C\mu\xi^3 \leq C\xi^6. \tag{3.5}$$

Combining (3.5) and (3.4), one has

$$\begin{aligned} I_3 &\leq C\int_0^T \mu\xi^3 \|\nabla y\|^3 \|\Delta y\| \, dt \\ &\leq C\left(\sup_{t\in[0,T]} \xi^2 \|\nabla y\|^2\right)\int_0^T \mu\xi \|\nabla y\| \|\Delta y\| \, dt \\ &\leq C\left(\sup_{t\in[0,T]} \xi^2 \|\nabla y\|^2\right)\left(\int_0^T \mu^2 \|\nabla y\|^2 \, dt\right)^{1/2}\left(\int_0^T \xi^2 \|\Delta y\|^2 \, dt\right)^{1/2}. \end{aligned}$$

Consequently,

$$I_3 \leq C\,\|(y, p, v)\|^4_{E^i_N}. \tag{3.6}$$

From (3.5), we have

$$\begin{aligned} I_2 &\leq C\int_0^T \xi^4 \|\nabla y\|^4 \xi^2 \|\Delta y\|^2 \, dt \\ &\leq C\left(\sup_{t\in[0,T]} \xi^2 \|\nabla y\|^2\right)^2\left(\iint_Q \xi^2 |\Delta y|^2 \, dx dt\right). \end{aligned}$$

From Proposition 2.4 and the inequality (3.1), one has:

$$I_2 \leq C\,\|(y, p, v)\|^6_{E^i_N}.$$

Finally, from this last inequality together with (3.3) and (3.6), we have that $\mathcal{A}_1(y, p, v) \in L^2\left(\rho^2; Q\right)^N$ and this concludes that $\mathcal{A}$ is well defined.

Furthermore, using similar arguments, it is easy to check the $\mathcal{A}$ continuous. □

Lemma 3.2 *The mapping $\mathcal{A} : E^i_N \longrightarrow F_N$ is continuously differentiable.*

Proof We will present the proof for $N = 3$ (the case $N = 2$ is similar). Let us first prove that $\mathcal{A}$ is Gâteaux derivative for all $(y, p, v) \in E^i_3$ and let us compute the $G\text{-}derivative$ of $\mathcal{A}$.

Let us fix $(y, p, v) \in E^i_3$ and let us take $(y', p', v') \in E^i_3$ and $\sigma > 0$. We have:

$$\begin{aligned} &\frac{1}{\sigma}\left[\mathcal{A}_1\left((y, p, v) - \sigma(y', p', v')\right) - \mathcal{A}_1(y, p, v)\right] = y'_t - (\nu_0 + \nu_1 \left\|\nabla(y + \sigma y')\right\|^2)\Delta y' \\ &-\frac{\nu_1}{\sigma}(\left\|\nabla(y + \sigma y')\right\| - \|\nabla y\|^2)\Delta y + \nabla p' - v' 1_\omega + (y' \cdot \nabla)y + (y \cdot \nabla)y' + \sigma(y' \cdot \nabla)y'. \end{aligned}$$

Let us introduce the linear mapping $D\mathcal{A} : E_3^i \longrightarrow F_3$, with $D\mathcal{A} = (D\mathcal{A}_1, D\mathcal{A}_2)$ where:

$$\begin{aligned} D\mathcal{A}_1(y', p', v') = y'_t - (\nu_0 + \nu_1 \, \|\nabla y\|^2)\Delta y' - 2\nu_1(\nabla y, \nabla y')\Delta y \\ \nabla p' - v' 1_\omega + (y' \cdot \nabla)y + (y \cdot \nabla)y'. \end{aligned}$$

$$D\mathcal{A}_2(y', p', v') = y'(\cdot, 0).$$

It becomes clear that $D\mathcal{A} \in \mathcal{L}(E_3^i, F_3)$. Furthermore,

$$\begin{gathered} \frac{1}{\sigma}\left[\mathcal{A}_1\left((y, p, v) + \sigma(y', p', v')\right) - \mathcal{A}_1(y, p, v)\right] \longrightarrow D\mathcal{A}_1(y', p', v') \\ \text{strong in } \; L^2(\rho^2; Q)^3, \text{ as } \sigma \longrightarrow 0. \end{gathered} \tag{3.7}$$

Indeed,

$$\begin{aligned} &\| \frac{1}{\sigma}\left[\mathcal{A}_1\left((y, p, v) + \sigma(y', p', v')\right) - \mathcal{A}_1(y, p, v)\right] - D\mathcal{A}_1(y', p', v')\|_{L^2(\rho^2; Q)^3} \\ &\qquad \le \|\nu_1(\left\|\nabla(y + \sigma y')\right\|^2 - \|\nabla y\|^2)\Delta y'\|_{L^2(\rho^2; Q)^3} \\ &\qquad + \|\frac{\nu_1}{\sigma}(\left\|\nabla(y + \sigma y')\right\|^2 - \|\nabla y\|^2)\Delta y - 2\nu_1(\nabla y, \nabla y')\Delta y\|_{L^2(\rho^2; Q)^3} \\ &\qquad + \|\sigma(y' \cdot \nabla)y'\|_{L^2(\rho^2; Q)^3}. \\ &\qquad = B_1 + B_2 + B_3. \end{aligned}$$

We see that for all $i \in \{1, 2, 3\}$, $B_i \to 0$, as $\sigma \to 0$. We have

$$B_1^2 = \nu_1^2 \iint_Q \rho^2 | \left\|\nabla(y + \sigma y')\right\|^2 - \|\nabla y\|^2 \, |^2 |\Delta y'|^2 dxdt \to 0.$$

$$B_2^2 = \nu_1^2 \iint_Q \rho^2 \rho^2 | \frac{1}{\sigma}(\left\|\nabla(y + \sigma y')\right\|^2 - \|\nabla y\|^2)\Delta y - 2(\nabla y, \nabla y')\Delta y|^2 dxdt \to 0.$$

and

$$B_3^2 = \sigma^2 \iint_Q \rho^2 \left|(y' \cdot \nabla)y'\right|^2 dxdt \longrightarrow 0.$$

as consequence of Proposition 2.4. Thus, (3.7) holds.

Therefore $\mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2)$ is Gâteaux-differentiable.

Let us check that $\mathcal{A} \in C^1(E_3^i, F_3)$ with $\mathcal{A}'(y, p, v) = D_G\mathcal{A}(y, p, v)$, *i.e*

$$\mathcal{A}'(y, p, v)(y', p', v') = D_G\mathcal{A}(y, p, v)(y', p', v').$$

But this last equality is equivalent to prove that there exists $\epsilon_n(y, p, v)$ such that

$$\left\|(D_G\mathcal{A}(y_n, p_n, v_n) - D_G\mathcal{A}(y, p, v))\,(y', p', v')\right\|_{F_3}^2 \leq \epsilon_n \left\|(y', p', v')\right\|_{E_3^i}^2 . \tag{3.8}$$

for all $(y', p', v') \in E_3^i$ and $\lim\limits_{n\longrightarrow\infty} \epsilon_n = 0$.

Let us prove (3.8)

$$\begin{aligned} D_G\mathcal{A}_1(y, p, v)(y', p', v') &= y_t' - (\nu_0 + \nu_1 \left\|\nabla y\right\|^2)\Delta y' - 2\nu_1(\nabla y, \nabla y')\Delta y + \\ &\quad \nabla p' - v' 1_\omega + (y' \cdot \nabla)y + (y \cdot \nabla)y'. \\ D_G\mathcal{A}_1(y_n, p_n, v_n)(y', p', v') &= y_t' - (\nu_0 + \nu_1 \left\|\nabla y_n\right\|^2)\Delta y' - 2\nu_1(\nabla y_n, \nabla y')\Delta y_n + \\ &\quad \nabla p' - v' 1_\omega + (y' \cdot \nabla)y_n + (y_n \cdot \nabla)y'. \end{aligned}$$

Then, we have,

$$\begin{aligned} (D_G\mathcal{A}_1(y, p, v) - D_G\mathcal{A}_1(y_n, p_n, v_n))\,(y', p', v') = {} & \nu_1(\|\nabla y\|^2 - \|\nabla y_n\|^2)\Delta y' \\ & -2\nu_1(\nabla y_n, \nabla y')\Delta y_n \\ & +2\nu_1(\nabla y, \nabla y')\Delta y + (y' \cdot \nabla)y_n \\ & +(y_n \cdot \nabla)y' - (y' \cdot \nabla)y - (y \cdot \nabla)y'. \end{aligned}$$

$$\begin{aligned} &\left\|(D_G\mathcal{A}_1(y_n, p_n, v_n) - D_G\mathcal{A}_1(y, p, v))\,(y', p', v')\right\|_{L^2(\rho^2; Q)^3}^2 \\ &\quad \leq 3\left\|(\nu_1 \left\|\nabla y\right\|^2 - \nu_1 \left\|\nabla y_n\right\|^2)\Delta y'\right\|_{L^2(\rho^2; Q)^3}^2 + \\ &\qquad 3\left\|-2\nu_1(\nabla y_n, \nabla y')\Delta y_n + 2\nu_1(\nabla y, \nabla y')\Delta y\right\|_{L^2(\rho^2; Q)^3}^2 + \\ &\qquad 3\left\|(y' \cdot \nabla)(y_n - y) + (y_n - y) \cdot \nabla y'\right\|_{L^2(\rho^2; Q)^3}^2 . \\ &\quad = 3D_{1,n} + 12D_{2,n} + 3D_{3,n}. \end{aligned}$$

Then, after some tedious but straightforward computations, we see that

$$D_{1,n} \leq \epsilon_{1,n} \left\|(y', p', v')\right\|_{E_3^i}^2 , \tag{3.9}$$

where

$$\epsilon_{1,n} = C \left\|(y_n, p_n, v_n) - (y, p, v)\right\|_{E_3^i}^2 \left(\left\|(y, p, v)\right\|_{E_3^i}^2 + \left\|(y_n, p_n, v_n)\right\|_{E_3^i}^2\right).$$

$$D_{2,n} \leq \epsilon_{2,n} \left\|(y', p', v')\right\|_{E_3^i}^2 , \tag{3.10}$$

where

$$\epsilon_{2,n} = C \, \|(y_n, p_n, v_n)\|^2_{E^i_3} \, \|(y_n, p_n, v_n) - (y, p, v)\|^2_{E^i_3} \\ + \|(y, p, v)\|^2_{E^i_3} \, \|(y_n, p_n, v_n) - (y, p, v)\|^2_{E^i_3} .$$

$$D_{3,n} \leq \epsilon_{3,n} \left\|(y', p', v')\right\|^2_{E^i_3} , \tag{3.11}$$

where

$$\epsilon_{3,n} = \|(y_n, p_n, v_n) - (y, p, v)\|^2_{E^i_3} .$$

From (3.9)–(3.11), we have $\lim_{n\longrightarrow\infty} \epsilon_{j,n} = 0$, for all $j \in \{1, 2, 3\}$ and this ends the proof. □

Lemma 3.3 *Let $\mathcal{A}$ be the mapping defined by (3.2). Then $\mathcal{A}'(0, 0, 0)$ is onto.*

Proof Let $(f, y_0) \in F_N$ from Proposition 2.1 we know that there exists (y, p, v) such that

$$\begin{cases} y_t - \nu_0 \Delta y + \nabla p = f + v 1_\omega & \text{in } Q, \\ \nabla \cdot y = 0 & \text{in } Q, \\ y = 0 & \text{on } \Sigma, \\ y(0) = y_0 & \text{in } \Omega, \end{cases}$$

satisfying $\rho_0 y, \rho_1 v 1_\omega \in L^2(Q)^N$, $y \in L^2(0, T; V) \cap C([0, T]; H)$, $y(0) \in V$, $\rho(Ly + \nabla p - v1_\omega) \in L^2(Q)^N$, $v_i = 0$ and the usual regularity results for the Stokes System (see [13]), we have $y \in L^2(0, T; D(A))$ and $p \in L^2(0, T; H^1(\Omega))$. Therefore $(y, p, v) \in E^i_N$ and $\mathcal{A}'(0, 0, 0)(y, p, v) = (f, y_0)$. This ends the proof. □

According to Lemmas 3.1–3.3, we notice that Liusternik's Theorem (Theorem 3.1) can be applied to the spaces E^i_N and F_N and to the mapping $\mathcal{A}$ introduced at the beginning of this Section. The consequence is that (1.1) is locally null-controllable, with triplets (y, p, v) in E^i_N.

4 The Proof of Theorem 1.2

Let us introduce the space

$$Y_N = \Big\{(y, p, v, \theta, v_0) : v_N \equiv 0,\ v_j \equiv 0,\ \text{for one } j < N;\ \rho_0 y, \rho_1 v 1_\omega \in L^2(Q)^N; \\ \rho_0\theta, \hat\rho_1 v_0 1_\omega \in L^2(Q);\ y \in L^2(0, T; D(A)), \theta \in L^2(0, T; H^2(\Omega) \cap H^1_0(\Omega)), \\ p \in L^2(0, T; H^1(\Omega));\ \rho(y_t - \nu_0 \Delta y + \nabla p - \theta e_N - v 1_\omega) \in L^2(Q)^N, \\ \hat\rho(\theta_t - \nu_0 \Delta\theta - v_0 1_\omega) \in L^2(Q)\Big\} .$$

It is clear that Y_N is a Hilbert space for the norm $\|\cdot\|_{Y_N}$, where

$$\begin{aligned}\|(y,p,v,\theta,v_0)\|^2_{Y_N} &= \|\rho_0 y\|^2_{L^2(Q)^N} + \|\rho_1 v\|^2_{L^2(\omega\times(0,T))^N} + \|\rho_0\theta\|^2_{L^2(Q)} + \|\hat\rho_1 v_0\|^2_{L^2(\omega\times(0,T))}\\ &\quad + \|y\|^2_{L^2(0,T;D(A))} + \|\theta\|^2_{L^2(0,T;H^2(\Omega)\cap H^1_0(\Omega))} + \|p\|^2_{L^2(0,T;H^1(\Omega))}\\ &\quad + \|\rho(y_t - \nu_0\Delta y + \nabla p - \theta e_N - v1_\omega)\|^2_{L^2(Q)^N}\\ &\quad + \|\hat\rho(\theta_t - \nu_0\Delta\theta - v_0 1_\omega)\|^2_{L^2(Q)}.\end{aligned}$$

Notice that, if $(y,p,v,\theta,v_0)\in Y_N$, then $y_t\in L^2(Q)^N$, $\theta_t\in L^2(Q)$, whence $y:[0,T]\mapsto V$ and $\theta:[0,T]\mapsto H^1_0(\Omega)$ are continuous and, in particular, we have $y(0)\in V$, $\theta(0)\in H^1_0(\Omega)$, and also

$$\|y(0)\|^2_V \le C\|(y,p,v,\theta,v_0)\|^2_{Y_N} \text{ and } \|\theta(0)\|^2_{H^1_0(\Omega)} \le C\|(y,p,v,\theta,v_0)\|^2_{Y_N}.$$

Furthermore, in view of Propositions 2.5 and 2.6, one also has

$$\|\zeta y\|^2_{L^2(0,T;V)\cap L^\infty(0,T;H)} + \|\zeta\theta\|^2_{L^2(0,T;H^1_0)\cap L^\infty(0,T;L^2)} \le C\|(y,p,v,\theta,v_0)\|^2_{Y_N}. \tag{4.1}$$

$$\|\kappa y\|^2_{L^2(0,T;D(A))\cap L^\infty(0,T;V)} + \|\kappa\theta\|^2_{L^2(0,T;H^1_0\cap H^2)\cap L^\infty(0,T;H^1_0)} \le C\|(y,p,v,\theta,v_0)\|^2_{Y_N}. \tag{4.2}$$

Let us introduce the Hilbert space

$$Z_N = L^2(\rho^2;Q)^N \times V \times L^2(\hat\rho^2;Q)\times H^1_0(\Omega),$$

and the mapping

$$\begin{aligned}&\mathcal{A}: Y_N \longrightarrow Z_N\\ &\quad\mathcal{A}(y,p,v,\theta,v_0) = (\mathcal{A}_1,\mathcal{A}_2,\mathcal{A}_3,\mathcal{A}_4)(y,p,v,\theta,v_0),\end{aligned}$$

where

$$\begin{aligned}&\mathcal{A}_1(y,p,v,\theta,v_0) = y_t - (\nu_0+\nu_1\|\nabla y\|^2)\Delta y + (y\cdot\nabla)y + \nabla p - v1_\omega - \theta e_N,\\ &\mathcal{A}_2(y,p,v,\theta,v_0) = y(0),\\ &\mathcal{A}_3(y,p,v,\theta,v_0) = \theta_t - (\nu_0+\nu_1\|\nabla y\|^2)\Delta\theta + y\cdot\nabla\theta - v_0 1_\omega,\\ &\mathcal{A}_4(y,p,v,\theta,v_0) = \theta(0).\end{aligned}$$

Note that, in view of (2.3) and (2.7), we have

$$\tilde\sigma^2 \le C\zeta\kappa^3 \le C\kappa^6,\ \ \hat\sigma^2 \le C\zeta\kappa^3,$$

and using (4.1), (4.2) and Proposition 2.2, we prove the following results

Lemma 4.1 *$\mathcal{A}$ is well defined and continuous.*

Lemma 4.2 *$\mathcal{A}$ is continuously differentiable.*

Lemma 4.3 *$\mathcal{A}'(0,0,0,0,0) : Y_N \to Z_N$ is onto.*

According to Lemmas 4.1–4.3, we can apply Liusternik's Theorem (Theorem 3.1), thus, there exists $\epsilon > 0$ and a mapping $W : B_\epsilon(0) \subset Z_N \to Y_N$ such that

$$W(z) \in B_r(0) \text{ and } \mathcal{A}(W(z)) = z, \quad \forall z \in B_\epsilon(0).$$

Taking $(0, y_0, 0, z_0) \in B_\epsilon(0)$ and $(y, p, v, \theta, v_0) = W(0, y_0, 0, z_0) \in Y_N$, we have

$$\mathcal{A}((y, p, v, \theta, v_0)) = (0, y_0, 0, z_0).$$

Therefore, (1.2) is locally null controllable at time $T > 0$.

5 Some Open Problems

Let us now indicate some open questions that arise naturally in the context of the results in this paper:

1. Is it possible the local exact controllability to the trajectories for the systems (1.1) and (1.2)?

   The main problem is to find a Carleman estimate for the follow system

$$\begin{cases} -\varphi_t - (\nu_0 + \nu_1 \|\nabla \overline{y}\|^2)\Delta\varphi + 2\nu_1 \Delta \overline{y} \int_\Omega \Delta \overline{y} \varphi dx' + (D\varphi)\overline{y} + \nabla \pi = g & \text{in } Q, \\ \nabla \cdot \varphi = 0 & \text{in } Q, \\ \varphi = 0 & \text{on } \Sigma, \\ \varphi(T) = \varphi^T & \text{in } \Omega, \end{cases}$$

   and this is very difficult using the conventional computations.
2. It is possible the null controllability of (1.1) in three dimensions with one scalar control?, in the case of system Navier-Stokes was solve for J.M. Coron and Pierre Lissy in [4], the general case remains open.
3. Finally, can we deduce the null controllability of (1.2) in N dimensions, with $N-1$ controls in the velocity and without controls in the temperature equation?

   This question is very difficult, because we need to obtain a new Carleman estimate that would make it possible, unfortunately in the adjoint system (2.2) we can not estimate the temperature in terms of the velocity, this is important in the computation of the desired result. That is why this result is in open.

References

1. Alekseev, V.M., Tikhomirov, V.M., Fomin, S.V.: Optimal Control. Translated from the Russian by V. M. Volosov. Contemporary Soviet Mathematics. Consultants Bureau, New York (1987)
2. Carreño, N.: Local controllability of the N-dimensional Boussinesq system with N-1 scalar controls in an arbitrary control domain. Math. Control Relat. Fields **2**(4), 361–382 (2012)
3. Carreño, N., Guerrero, S.: Local null controllability of the N-dimensional Navier-Stokes system with N-1 scalar controls in an arbitrary control domain. J. Math. Fluid Mech. **15**(1), 139–153 (2012)
4. Coron, J.-M., Lissy, P.: Local null controllability of the three-dimensional Navier-Stokes system with a distributed control having two vanishing components. Invent. Math. **198**(3), 833–880 (2014)
5. Fernández-Cara, E., Guerrero, S., Imanuvilov, O.Y., Puel, J.-P.: Local exact controllability of the Navier Stokes system. J. Math. Pures Appl. **83**, 1501–1542 (2004)
6. Fernández-Cara, E., Guerrero, S., Imanuvilov, O.Y., Puel, J.-P.: Some controllability for the N-dimensional Navier-Stokes and Boussinesq systems with N-1 scalar controls. SIAM J. Control Optim. **45**(1), 146–173 (2006)
7. Fernández-Cara, E., Límaco, J., de Menezes, S.B.: Theoretical and numerical local null controllability of a Ladyzhenskaya-Smagorinsky model of turbulence. J. Math. Fluid Mech. **17**(4), 669–698 (2015)
8. Fursikov, A., Imanuvilov, O.Y.: Controllability of Evolution Equations. Lecture Notes, vol. 34. Seoul National University, Seoul (1996)
9. Guerrero, S.: Local exact controllability to the trajectories of the Boussinesq system. Ann. I. H. Poincaré **23**, 29–61 (2006)
10. Imanuvilov, O.Y., Puel, J.-P.: Global Carleman estimates for weak elliptic non homogeneous Dirichlet problem. Int. Math. Res. Not. **16**, 883–913 (2003)
11. Lions, J.L.: Quelques Méthodes de Résolutions des Problèmes aux Limites non Linéaires. Dunod Gauthier-Villars, Paris (1969)
12. Puel, J.-P.: Controllability of Navier-Stokes Equations. Laboratoire de Mathematiques de Versailles (2012)
13. Temam, R.: Navier-Stokes Equations, Theory and Numerical Analysis. Studies in Applied Mathematics, vol. 2. North-Holland, Amsterdam (1977)

Numerical Estimations of the Cost of Boundary Controls for the Equation $y_t - \varepsilon y_{xx} + My_x = 0$ with Respect to ε

Arnaud Münch

Dedicated to Prof. Enrique Fernández-Cara on the occasion of his 60th birthday.

Abstract We numerically examine the cost of the null boundary control for the transport diffusion equation $y_t - \varepsilon y_{xx} + My_x = 0, x \in (0, L), t \in (0, T)$ with respect to the positive parameter ε. It is known that this cost is uniformly bounded with respect to ε if $T \geq T_M$ with $T_M \in [1, 2\sqrt{3}]L/M$ if $M > 0$ and if $T_M \in [2\sqrt{2}, 2(1+\sqrt{3})]L/|M|$ if $M < 0$. We propose a method to approximate the underlying observability constant and then conjecture, through numerical computations, the minimal time of controllability T_M leading to a uniformly bounded cost. Several experiments for $M \in \{-1, 1\}$ are performed and discussed.

Keywords Singular controllability · Lagrangian variational formulation · Numerical approximation

1 Introduction: Problem Statement

Let $L > 0, T > 0$ and $Q_T := (0, L) \times (0, T)$. This work is concerned with the null controllability problem for the parabolic equation

$$\begin{cases} y_t - \varepsilon y_{xx} + My_x = 0 & \text{in} \quad Q_T, \\ y(0, \cdot) = v, \;\; y(L, \cdot) = 0 & \text{on} \quad (0, T), \\ y(\cdot, 0) = y_0 & \text{in} \quad (0, L). \end{cases} \tag{1}$$

A. Münch (✉)
Laboratoire de Mathématiques Blaise Pascal, Université Clermont Auvergne, UMR CNRS 6620, Aubière, France
e-mail: arnaud.munch@uca.fr

A. Doubova et al. (eds.), *Recent Advances in PDEs: Analysis, Numerics and Control*, SEMA SIMAI Springer Series 17,
https://doi.org/10.1007/978-3-319-97613-6_9

Here we assume that $y_0 \in H^{-1}(0,L)$. $\varepsilon > 0$ is the diffusion coefficient while $M \in \mathbb{R}$ is the transport coefficient; $v = v(t)$ is the control (a function in $L^2(0,T)$) and $y = y(x,t)$ is the associated state. In the sequel, we shall use the following notations:

$$L_\varepsilon y := y_t - \varepsilon y_{xx} + My_x, \qquad L_\varepsilon^\star \varphi := -\varphi_t - \varepsilon\varphi_{xx} - M\varphi_x.$$

For any $y_0 \in H^{-1}(0,L)$ and $v \in L^2(0,T)$, there exists exactly one solution y to (1), with the regularity $y \in L^2(Q_T) \cap \mathcal{C}([0,T]; H^{-1}(0,L))$ (see for instance [12, Prop. 2.2]). Accordingly, for any final time $T > 0$, the associated null controllability problem at time $T > 0$ is the following: for each $y_0 \in H^{-1}(0,L)$, find $v \in L^2(0,T)$ such that the corresponding solution to (1) satisfies

$$y(\cdot,T) = 0 \text{ in } H^{-1}(0,L). \tag{2}$$

For any $T > 0$, $M \in \mathbb{R}$ and $\varepsilon > 0$, the null controllability for the parabolic type equation (1) holds true. We refer to [13] and [16] using Carleman type estimates. We therefore introduce the non-empty set of null controls

$$\mathcal{C}(y_0,T,\varepsilon,M) := \{(y,v) : v \in L^2(0,T);\ y \text{ solves } (1) \text{ and satisfies } (2)\}.$$

For $\varepsilon = 0$, the system (1) degenerates into a transport equation and is uniformly controllable as soon as T is large enough, according to the speed $|M|$ of transport, precisely as soon as $T \geq L/|M|$. On the other hand, for $\varepsilon > 0$, the asymptotic behavior of the null controls as $\varepsilon \to 0^+$ is less clear, depends on the sign of M, and has been the subject of several works in the last decade.

For any $\varepsilon > 0$, we define the cost of control by the following quantity:

$$K(\varepsilon,T,M) := \sup_{\|y_0\|_{L^2(0,L)}=1} \left\{ \min_{u \in \mathcal{C}(y_0,T,\varepsilon,M)} \|u\|_{L^2(0,T)} \right\}, \tag{3}$$

and denote by T_M the minimal time for which the cost $K(\varepsilon,T,M)$ is uniformly bounded with respect to the parameter ε. In other words, (1) is uniformly controllable with respect to ε if and only if $T \geq T_M$. In [9], J-M. Coron and S. Guerrero proved, using spectral arguments coupled with Carleman type estimates, that

$$T_M \in \begin{cases} [1,4.3]\dfrac{L}{M} & \text{if} \quad M > 0, \\[2ex] [2,57.2]\dfrac{L}{|M|} & \text{if} \quad M < 0. \end{cases}$$

The lower bounds are obtained using the initial condition $y_0(x) = \sin(\pi x/L)e^{\frac{Mx}{2\varepsilon}}$. The upper bounds are deduced from Carleman type inequalities for the adjoint solution. Then, using complex analysis arguments, O. Glass improved in [14] the previous estimations: precisely, he obtained that

$$T_M \in \begin{cases} [1, 4.2]\dfrac{L}{M} & \text{if} \quad M > 0, \\ [2, 6.1]\dfrac{L}{|M|} & \text{if} \quad M < 0. \end{cases}$$

These authors exhibit an exponential behavior of the L^2-norm of the controls with respect to ε. More recently, P. Lissy in [18, 19] yielded to the following conclusions:

$$T_M \in \begin{cases} [1, 2\sqrt{3}]\dfrac{L}{M} & \text{if} \quad M > 0, \\ [2\sqrt{2}, 2(1+\sqrt{3})]\dfrac{L}{|M|} & \text{if} \quad M < 0. \end{cases} \tag{4}$$

Remark that $2(1 + \sqrt{3}) \approx 5.46$. The second lower bound $2\sqrt{2}$ is obtained by considering again the initial data $y_0(x) = \sin(\pi x/L)e^{\frac{Mx}{2\varepsilon}}$.

The main goal of the present work is to approximate numerically the value of T_M, for both $M > 0$ and $M < 0$. This can be done by approximating the cost K for various values of ε and $T > 0$, the ratio L/M being fixed.

In Sect. 2, we reformulate the cost of control K as the solution of a generalized eigenvalue problem, involving the control operator. In Sect. 3, we adapt [21], present a robust method to approximate numerically the control of minimal L^2-norm and discuss some experiments, for a given initial data y_0. In Sect. 4, we solve at the finite dimensional level the related eigenvalue problem using the power iterate method: each iteration requires the resolution of a null controllability problem for (1). We then discuss some experiments with respect to ε and T for $L/M = 1$ and $L/M = -1$ respectively.

2 Reformulation of the Controllability Cost $K(\varepsilon, T, M)$

We reformulate the cost of control K as the solution of a generalized eigenvalues problem involving the control operator (named as the HUM operator by J.-L. Lions for wave type equations). From (3), we can write

$$K^2(\varepsilon, T, M) = \sup_{y_0 \in L^2(0,L)} \frac{(v, v)_{L^2(0,T)}}{(y_0, y_0)_{L^2(0,L)}}$$

where $v = v(y_0)$ is the null control of minimal $L^2(0, T)$-norm for (1) with initial data y_0 in $L^2(0, L)$. Let us recall that any null control for (1) satisfies the following characterization

$$(v, \varepsilon\varphi_x(0,\cdot))_{L^2(0,T)} + (y_0, \varphi(\cdot,0))_{L^2(0,L)} = 0, \tag{5}$$

for any φ solution of the adjoint problem

$$\begin{cases} -\varphi_t - \varepsilon\varphi_{xx} - M\varphi_x = 0 & \text{in} \quad Q_T, \\ \varphi(0,\cdot) = \varphi(L,\cdot) = 0 & \text{on} \quad (0,T), \\ \varphi(\cdot,T) = \varphi_T & \text{in} \quad (0,L), \end{cases} \tag{6}$$

where $\varphi_T \in H_0^1(0,L)$. In particular, the control of minimal L^2-norm is given by $v = \varepsilon\hat{\varphi}_x(0,\cdot)$ in $(0,T)$ where $\hat{\varphi}$ solves (6) associated to the initial $\hat{\varphi}_T$, solution of the extremal

$$\sup_{\varphi_T \in H_0^1(0,L)} J^\star(\varphi_T) := \frac{1}{2}\int_0^T (\varepsilon\varphi_x(0,\cdot))^2 dt + (y_0, \varphi(\cdot,0))_{L^2(0,L)}. \tag{7}$$

Taking $\varphi = \hat{\varphi}$ associated to $\hat{\varphi}_T$ in (5), we therefore have

$$(v,v)_{L^2(0,T)} = (v, \varepsilon\hat{\varphi}_x(0,t))_{L^2(0,T)} = -(y_0, \hat{\varphi}(\cdot,0))_{L^2(0,T)}. \tag{8}$$

Consequently, if we denote by $\mathcal{A}_\varepsilon : L^2(0,L) \to L^2(0,L)$ the control operator defined by $\mathcal{A}_\varepsilon y_0 := -\hat{\varphi}(\cdot,0)$, we finally obtain

$$K^2(\varepsilon,T,M) = \sup_{y_0 \in L^2(0,L)} \frac{(\mathcal{A}_\varepsilon y_0, y_0)_{L^2(0,L)}}{(y_0,y_0)_{L^2(0,L)}} \tag{9}$$

and conclude that $K^2(\varepsilon,T,M)$ is solution of the following generalized eigenvalue problem:

$$\sup\left\{\lambda \in \mathbb{R} : \exists\, y_0 \in L^2(0,L),\, y_0 \neq 0,\ \text{s.t. } \mathcal{A}_\varepsilon y_0 = \lambda y_0 \quad \text{in} \quad L^2(0,L)\right\}. \tag{10}$$

Remark 1 The controllability cost is related to the observability constant $C_{obs}(\varepsilon,T,M)$ which appears in the observability inequality for (6)

$$\|\varphi(\cdot,0)\|^2_{L^2(0,L)} \le C_{obs}(\varepsilon,T,M)\|\varepsilon\varphi_x(0,\cdot)\|^2_{L^2(0,T)}, \quad \forall\varphi_T \in H_0^1(0,L)\cap H^2(0,L)$$

defined by

$$C_{obs}(\varepsilon,T,M) = \sup_{\varphi_T \in H_0^1(0,L)} \frac{\|\varphi(\cdot,0)\|^2_{L^2(0,L)}}{\|\varepsilon\varphi_x(0,\cdot)\|^2_{L^2(0,T)}}. \tag{11}$$

Precisely, we get that $K(\varepsilon,T,M) = \sqrt{C_{obs}(\varepsilon,T,M)}$ (see [8], Remark 2.98).

Remark 2 We may reformulate as well the previous extremal problem over $H_0^1(0,L)$ (seen as the dual space of $H^{-1}(0,L) \ni y(\cdot,T)$) in term of a generalized eigenvalue problem; we proceed as follows.

We introduce the operators A_ε and B_ε given by

$$A_\varepsilon : \begin{array}{rcl} H_0^1(0,L) & \to & L^2(0,L) \\ \varphi_T & \mapsto & \varphi(\cdot,0) \end{array} \quad \text{and} \quad B_\varepsilon : \begin{array}{rcl} H_0^1(0,L) & \to & L^2(0,T) \\ \varphi_T & \mapsto & \varepsilon\varphi_x(0,\cdot), \end{array}$$

where φ solves (6). The adjoint operators $A_\varepsilon^\star$ and $B_\varepsilon^\star$ of A_ε and B_ε are given by:

$$A_\varepsilon^\star : \begin{array}{rcl} L^2(0,L) & \to & H^{-1}(0,L) \\ y_0 & \mapsto & y(T;y_0,0) \end{array} \quad \text{and} \quad B_\varepsilon^\star : \begin{array}{rcl} L^2(0,L) & \to & H^{-1}(0,L) \\ v & \mapsto & y(T;0,v), \end{array}$$

where $y(t;y_0,v)$ is the solution to (1) at time t for the initial data y_0 and the control v. With these notations, we may rewrite C_{obs} given by (11) as follows

$$\begin{aligned} C_{obs}(\varepsilon,T,M) &= \sup_{\varphi_T \in H_0^1(0,L)} \frac{(A_\varepsilon\varphi_T, A_\varepsilon\varphi_T)_{L^2(0,L)}}{(B_\varepsilon\varphi_T, B_\varepsilon\varphi_T)_{L^2(0,T)}} \\ &= \sup_{\varphi_T \in H_0^1(0,L)} \frac{((-\Delta^{-1})A_\varepsilon^\star A_\varepsilon\varphi_T, \varphi_T)_{H_0^1(0,L)}}{((-\Delta^{-1})B_\varepsilon^\star B_\varepsilon\varphi_T, \varphi_T)_{H_0^1(0,L)}} \end{aligned}$$

leading to an eigenvalue problem over $H_0^1(0,L)$.

Remark that the operator $B_\varepsilon^\star B_\varepsilon$ from $H_0^1(0,L)$ to $H^{-1}(0,L)$ associates to the initial state φ_T of (6) the final state $y(\cdot,T)$ of (1) with $y_0 = 0$ and $v = \varepsilon\varphi_x(0,\cdot)$. v is therefore the control of minimal $L^2(0,T)$-norm with drives the state y from 0 to the trajectory $y(\cdot,T)$. $B_\varepsilon^\star B_\varepsilon$ is the so-called HUM operator.

Remark 3 Actually, the supremum of $\varphi_T \in H_0^1(0,L)$ in (11) can be taken over $\varphi(\cdot,0) \in L^2(0,L)$ (or even over φ !) leading immediately to

$$C_{obs}(\varepsilon,T,M) = \sup_{\varphi(\cdot,0)\in L^2(0,L)} \frac{(\varphi(\cdot,0),\varphi(\cdot,0))}{(\mathcal{A}_\varepsilon^{-1}\varphi(\cdot,0),\varphi(\cdot,0))_{L^2(0,L)}}$$

in full agreement with (9) and the equality $K(\varepsilon,T,M) = \sqrt{C_{obs}(\varepsilon,T,M)}$.

Remark 4 The sup-inf problem (3) may be solved by a gradient procedure. Let us consider the Lagrangien $\mathcal{L} : L^2(0,L) \times \mathbb{R} \to \mathbb{R}$ defined by

$$\mathcal{L}(y_0,\mu) := \frac{1}{2}\|v(y_0)\|^2_{L^2(0,T)} + \frac{1}{2}\mu\Big(\|y_0\|^2_{L^2(0,L)} - 1\Big)$$

where $v(y_0)$ is the control of minimal L^2-norm associated to the initial data $y_0 \in L^2(0, L)$ and $\mu \in \mathbb{R}$ a lagrange multiplier to enforce the constraint $\|y_0\|_{L^2(0,L)} = 1$. $v(y_0)$ satisfies (8). The first variation of $\mathcal{L}$ is given by

$$D\mathcal{L}(y_0) \cdot \overline{y_0} = (\mu y_0 - \varphi(\cdot, 0), \overline{y_0})_{L^2(0,L)} = \Big((\mu \, Id + \mathcal{A}_\varepsilon) y_0, \overline{y_0}\Big)_{L^2(0,L)} \tag{12}$$

where φ solves (6)–(7). A maximizing sequence $\{y_0^k\}_{k\geq 1}$ can be constructed as follows: given $y_0^0 \in L^2(0, L)$ such that $\|y_0^0\|_{L^2(0,L)} = 1$, compute iteratively

$$y_0^{k+1} = y_0^k + \eta^k(\mu^k y_0^k - \varphi^k(\cdot, 0)), \quad k \geq 0$$

with $\eta^k > 0$ small enough and μ^k such that $\|y_0^{k+1}\|_{L^2(0,L)} = 1$, that is,

$$\mu^k = \frac{\theta^k - 1}{\eta^k}, \quad \theta^k = \eta^k (y_0^k, \varphi^k(\cdot, 0))_{L^2(0,L)} \pm \sqrt{1 + (\eta^k)^2 (y_0^k - \varphi^k(\cdot, 0), \varphi^k(\cdot, 0))_{L^2(0,L)}}.$$

Remark that (12) implies that the optimal initial data y_0 is proportional to the optimal terminal state $\varphi(\cdot, 0)$ of φ solution of (6)–(7). Then, from the characterization (8), the sequence μ^k satisfies $(v^k, v^k) + \mu^k (y_0^k, \varphi^k(\cdot, 0))_{L^2(0,L)} = 0$ and converges toward $-K^2(\varepsilon, T, M)$. Remark that μ^k defined above is always negative.

In order to solve the eigenvalue problem (10) and get the largest eigenvalue of the operator $\mathcal{A}_\epsilon$, we may employ the power iterate method (see [6]), which reads as follows: given $y_0^0 \in L^2(0, L)$ such that $\|y_0^0\|_{L^2(0,L)} = 1$, compute

$$\begin{cases} z_0^{k+1} = \mathcal{A}_\varepsilon y_0^k, \quad k \geq 0, \\ y_0^{k+1} = \dfrac{z_0^{k+1}}{\|z_0^{k+1}\|_{L^2(0,L)}}, \quad k \geq 0. \end{cases}$$

The real sequence $\{\|z_0^k\|_{L^2(0,L)}\}_{(k>0)}$ then converges to the eigenvalue with largest modulus of the operator $\mathcal{A}_\varepsilon$, so that

$$\sqrt{\|z_0^k\|_{L^2(0,L)}} \to K(\varepsilon, T, M) \quad \text{as} \quad k \to \infty.$$

The L^2 sequence $\{y_0^k\}_{(k\geq 0)}$ then converges toward the corresponding eigenvector. The first step requires to compute the image of the control operator $\mathcal{A}_\varepsilon$: this is done by determining the control of minimal L^2-norm, i.e. by solving the extremal problem (7) with y_0^k as initial condition for (1).

3 Approximation of the Control Problem

The generalized eigenvalue problem (10) involves the null control operator $\mathcal{A}_\varepsilon$ associated to (1). At the finite dimensional level, this problem can be solved by the way of the power iterate method, which requires at each iterates, the approximation of the null control of minimal L^2-norm for (1). We discuss in this section such approximation, the initial data y_0 in (1) being fixed.

The numerical approximation of null controls for parabolic equations is a not an easy task and has been first discussed in [4], and then in several works: we refer to the review [23]. Duality theory reduces the problem to the resolution of the unconstrained extremal problem (7). In view of the regularization character of the parabolic operator, the extremal problem (7) is ill-posed as the supremum is not reached in $H_0^1(0,L)$ but in a space, say $\mathcal{H}$, defined as the completion of $H_0^1(0,L)$ for the norm $\|\varphi_T\|_{\mathcal{H}} := \|\varepsilon\varphi_x(0,\cdot)\|_{L^2(0,T)}$, much larger than $H_0^1(0,L)$ and difficult to approximate. We refer to the review paper [23]. The usual "remedy" consists to enforce the regularity H_0^1 and replace (7) by

$$\min_{\varphi_T\in H_0^1(0,L)} J_\beta^\star(\varphi_T) := \frac{1}{2}\|\varepsilon\varphi_x(0,\cdot))\|^2_{L^2(0,T)} + (y_0,\varphi(\cdot,0))_{L^2(0,T)} + \frac{\beta}{2}\|\varphi_T\|^2_{H_0^1(0,L)} \tag{13}$$

for any $\beta > 0$ small. The resulting approximate control $v_\beta = \varepsilon\varphi_{\beta,x}(0,\cdot)$ leads to a state y_β solution of (1) satisfying the property

$$\|y_\beta(\cdot,T)\|_{H^{-1}(0,L)} \leq C\sqrt{\beta}\|y_0\|_{L^2(0,L)} \tag{14}$$

(for a constant $C > 0$ independent of β). This penalty method is discussed in [4] for the boundary controllability of the heat equation (for the distributed case, we refer to [2, 11, 15]). As in [4], problem (13) may be solved using a gradient iterative method: in view of the ill-posedness of (7), such method requires an increasing number of iterates to reach convergence as β goes to zero.

Moreover, in the context of the transport equation (1), it is necessary to take β small enough, in relation with the diffusion coefficient ε. Indeed, if $\beta > 0$ is fixed (independently of ε), then for $\varepsilon > 0$ small enough, the uncontrolled solution of (1) satisfies (14) as soon as $T \geq L/|M|$. In that case, problem (13) leads to the minimizer $\varphi_T = 0$ and then to the null control which is certainly not the optimal control we expect for negatives values of M (in view of (4))!

Therefore, as ε tends to 0, the occurrence of the transport term makes the approximation of the null control for (1) a challenging task. Consequently, instead of minimizing the functional $J^\star$ (or $J_\beta^\star$), we adapt [21] (devoted to the inner situation for $M = 0$ and $\varepsilon = 1$) and try to solve directly the corresponding optimality conditions. This leads to a mixed variational formulation (following the terminology used in [21]).

3.1 Mixed Variational Formulation

We introduce the linear space $\Phi^0 := \{\varphi \in C^2(\overline{Q_T}),\ \varphi = 0 \text{ on } \Sigma_T\}$. For any $\eta > 0$, we define the bilinear form

$$(\varphi, \overline{\varphi})_{\Phi^0} := \int_0^T \varepsilon\varphi_x(0,t)\,\varepsilon\overline{\varphi}_x(0,t)\,dt + \beta\big(\varphi(\cdot,T), \overline{\varphi}(\cdot,T)\big)_{H_0^1(0,L)} + \eta \iint_{Q_T} L^\star\varphi\, L^\star\overline{\varphi}\,dx\,dt, \quad \forall \varphi, \overline{\varphi} \in \Phi^0.$$

From the unique continuation property for the transport equation, this bilinear form defines for any $\beta \geq 0$ a scalar product. Let Φ_β be the completion of Φ^0 for this scalar product. We denote the norm over Φ_β by $\|\cdot\|_{\Phi_\beta}$ such that

$$\|\varphi\|^2_{\Phi_\beta} := \|\varepsilon\varphi_x(0,\cdot)\|^2_{L^2(0,T)} + \beta\|\varphi(\cdot,T)\|^2_{H_0^1(0,L)} + \eta\|L^\star\varphi\|^2_{L^2(Q_T)}, \quad \forall \varphi \in \Phi_\beta. \tag{15}$$

Finally, we define the closed subset W_β of Φ_β by $W_\beta = \{\varphi \in \Phi_\beta : L^\star\varphi = 0 \text{ in } L^2(Q_T)\}$ endowed with the same norm than Φ_β. Then, for any $r \geq 0$, we define the following extremal problem:

$$\min_{\varphi\in W_\beta} \hat{J}^\star_\beta(\varphi) := \frac{1}{2}\|\varepsilon\varphi_x(0,\cdot)\|^2_{L^2(0,T)} + \frac{\beta}{2}\|\varphi(\cdot,T)\|^2_{H_0^1(0,L)} + (y_0, \varphi(\cdot,0))_{L^2(0,L)} + \frac{r}{2}\|L^\star\varphi\|^2_{L^2(Q_T)}. \tag{16}$$

Standard energy estimates for (1) imply that, for any $\varphi \in W_\beta$, $\varphi(\cdot,0) \in L^2(0,L)$ so that the functional $\hat{J}^\star_\beta$ is well-defined over W_β. Moreover, since for any $\varphi \in W_\beta$, $\varphi(\cdot,T)$ belongs to $H_0^1(0,L)$, problem (16) is equivalent to the extremal problem (13). The main variable is now φ submitted to the constraint equality (in $L^2(Q_T)$) $L^\star\varphi = 0$, which is addressed through a Lagrange multiplier.

3.1.1 Mixed Formulation

We consider the following mixed formulation : find $(\varphi_\beta, \lambda_\beta) \in \Phi_\beta \times L^2(Q_T)$ solution of

$$\begin{cases} a_{\beta,r}(\varphi_\beta, \overline{\varphi}) + b(\overline{\varphi}, \lambda_\beta) = l(\overline{\varphi}), & \forall \overline{\varphi} \in \Phi_\beta \\ b(\varphi_\beta, \overline{\lambda}) = 0, & \forall \overline{\lambda} \in L^2(Q_T), \end{cases} \tag{17}$$

where

$$\begin{aligned} a_{\beta,r} : \Phi_\beta \times \Phi_\beta \to \mathbb{R}, \quad a_{\beta,r}(\varphi, \overline{\varphi}) &:= (\varepsilon\varphi_x(0,\cdot), \varepsilon\overline{\varphi}_x(0,\cdot))_{L^2(0,T)} \\ &\quad + \beta(\varphi(\cdot,T), \overline{\varphi}(\cdot,T))_{H_0^1(0,L)} \\ &\quad + r(L^\star\varphi, L^\star\overline{\varphi})_{L^2(Q_T)} \\ b : \Phi_\beta \times L^2(Q_T) \to \mathbb{R}, \quad b(\varphi,\lambda) &:= (L^\star\varphi,\, \lambda)_{L^2(Q_T)} \\ l : \Phi_\beta \to \mathbb{R}, \quad l(\varphi) &:= -(y_0, \varphi(\cdot,0))_{L^2(0,L)}. \end{aligned}$$

We have the following result:

Theorem 3.1 *Assume that $\beta > 0$ and $r \geq 0$.*

1. *The mixed formulation (17) is well-posed.*
2. *The unique solution $(\varphi_\beta, \lambda_\beta) \in \Phi_\beta \times L^2(Q_T)$ is the unique saddle-point of the Lagrangian $\mathcal{L}_{\beta,r} : \Phi_\beta \times L^2(Q_T) \to \mathbb{R}$ defined by*

$$\mathcal{L}_{\beta,r}(\varphi,\lambda) := \frac{1}{2}a_{\beta,r}(\varphi,\varphi) + b(\varphi,\lambda) - l(\varphi). \tag{18}$$

3. *The optimal function φ_β is the minimizer of $\hat{J}^\star_\beta$ over W_β while $\lambda_\beta \in L^2(Q_T)$ is the state of (1) in the weak sense.*

Proof The proof is very closed to the proof given in [21, Section 2.1.1]. The bilinear form $a_{\beta,r}$ is continuous, symmetric and positive over $\Phi_\beta \times \Phi_\beta$. The bilinear form b is continuous over $\Phi_\beta \times L^2(Q_T)$. Furthermore, for any $\beta > 0$, the continuity of the linear form l over Φ_β is deduced from the energy estimate:

$$\|\varphi(\cdot,0)\|^2_{L^2(0,L)} \leq C \iint_{Q_T} |L^\star\varphi|^2 dx\,dt + \|\varphi(\cdot,T)\|^2_{L^2(0,L)}, \quad \forall\varphi \in \Phi_\beta,$$

for some $C > 0$ so that $\|\varphi(\cdot,0)\|^2_{L^2(0,L)} \leq \max(C\eta^{-1}, \beta^{-1})\|\varphi\|^2_{\Phi_\beta}$. Therefore, the well-posedness of the mixed formulation is a consequence of the following properties (see [3]):

- $a_{\beta,r}$ is coercive on $\mathcal{N}(b)$, where $\mathcal{N}(b)$ denotes the kernel of b :

$$\mathcal{N}(b) := \{\varphi \in \Phi_\beta \ :\ b(\varphi,\lambda) = 0 \text{ for every } \lambda \in L^2(Q_T)\}.$$

- b satisfies the usual "inf-sup" condition over $\Phi_\beta \times L^2(Q_T)$: there exists $\delta > 0$ such that

$$\inf_{\lambda \in L^2(Q_T)} \sup_{\varphi \in \Phi_\beta} \frac{b(\varphi,\lambda)}{\|\varphi\|_{\Phi_\beta}\|\lambda\|_{L^2(Q_T)}} \geq \delta. \tag{19}$$

The first point follows from the definition. Concerning the inf-sup condition, for any fixed $\lambda^0 \in L^2(Q_T)$, we define the (unique) element φ^0 such that $L^\star\varphi^0 = \lambda^0$, $\varphi = 0$ on Σ_T and $\varphi^0(\cdot, T) = 0$ in $L^2(0, L)$. The function φ^0 is therefore solution of the backward transport equation with source term $\lambda^0 \in L^2(Q_T)$, null Dirichlet boundary condition and zero initial state. Moreover, since $\lambda^0 \in L^2(Q_T)$, the following estimate proved in the Appendix A of [9] (more precisely, we refer to the inequality (94))

$$\varepsilon\|\varphi^0_x(0,\cdot)\|_{L^2(0,T)} \le C_{L,T,M}\|\lambda^0\|_{L^2(Q_T)}$$

for a constant $C_{L,T,M} > 0$ independent of ε, implies that $\varphi^0 \in \Phi_\beta$. In particular, we have $b(\varphi^0, \lambda^0) = \|\lambda^0\|^2_{L^2(Q_T)}$ and

$$\begin{aligned}\sup_{\varphi\in\Phi_\beta} \frac{b(\varphi, \lambda^0)}{\|\varphi\|_{\Phi_\beta}\|\lambda^0\|_{L^2(Q_T)}} &\ge \frac{b(\varphi^0, \lambda^0)}{\|\varphi^0\|_{\Phi_\beta}\|\lambda^0\|_{L^2(Q_T)}}\\ &= \frac{\|\lambda^0\|^2_{L^2(Q_T)}}{\left(\|\varepsilon\varphi^0_x(0,\cdot)\|^2_{L^2(0,T)} + \eta\|\lambda^0\|^2_{L^2(Q_T)}\right)^{\frac{1}{2}}\|\lambda^0\|_{L^2(Q_T)}}.\end{aligned}$$

Combining the above two inequalities, we obtain

$$\sup_{\varphi^0\in\Phi_\beta} \frac{b(\varphi^0, \lambda^0)}{\|\varphi^0\|_{\Phi_\beta}\|\lambda^0\|_{L^2(Q_T)}} \ge \frac{1}{\sqrt{C^2_{L,T,M} + \eta}} \tag{20}$$

and, hence, (19) holds with $\delta = \left(C^2_{L,T,M} + \eta\right)^{-1/2}$.

The second point is due to the symmetry and to the positivity of the bilinear form $a_{\beta,r}$. Concerning the third point, the equality $b(\varphi_\beta, \overline{\lambda}) = 0$ for all $\overline{\lambda} \in L^2(Q_T)$ implies that $L^\star\varphi_\beta = 0$ as an $L^2(Q_T)$-function, so that if $(\varphi_\beta, \lambda_\beta) \in \Phi_\beta \times L^2(Q_T)$ solves the mixed formulation, then $\varphi_\beta \in W_\beta$ and $\mathcal{L}_\beta(\varphi_\beta, \lambda_\beta) = \hat{J}^\star_\beta(\varphi_\beta)$. Finally, the first equation of the mixed formulation (taking $r = 0$) reads as follows:

$$\begin{aligned}\int_0^T \varepsilon(\varphi_\beta)_x(0,t)\,\varepsilon\overline{\varphi}_x(0,t)dt + \beta\big(\varphi_\beta(\cdot,T), \overline{\varphi}(\cdot,T)\big)_{H^1_0(0,L)}\\ - \iint_{Q_T} L^\star\overline{\varphi}\,\lambda_\beta\,dx\,dt = l(\overline{\varphi}), \quad \forall\overline{\varphi} \in \Phi_\beta,\end{aligned}$$

or equivalently, since the control is given by $v_\beta := \varepsilon(\varphi_\beta)_x(0,\cdot)$,

$$\begin{aligned}\int_0^T v_\beta(t)\,\varepsilon\overline{\varphi}_x(0,t)\,dt + \beta(\varphi_\beta(\cdot,T), \overline{\varphi}(\cdot,T))_{H^1_0(0,L)}\\ - \iint_{Q_T} L^\star\overline{\varphi}\,\lambda_\beta\,dx\,dt = l(\overline{\varphi}), \quad \forall\overline{\varphi} \in \Phi_\beta.\end{aligned}$$

But this means that $\lambda_\beta \in L^2(Q_T)$ is solution of (1) in the transposition sense. Since $y_0 \in L^2(0, L)$ and $v_\beta \in L^2(0, T)$, λ_β coincides with the unique weak solution to (1) such that $-\Delta^{-1}\lambda_\beta(\cdot, T) + \beta\varphi_\beta(\cdot, T) = 0$. □

3.1.2 Minimization with Respect to the Lagrange Multiplier

The augmented mixed formulation (17) allows to solve simultaneously the dual variable φ_β, argument of the conjugate functional (16), and the Lagrange multiplier λ_β, qualified as the primal variable of the problem.

Assuming that the augmentation parameter r is strictly positive, we derive the corresponding extremal problem involving only the variable λ_β. For any $r > 0$, let the linear operator $\mathcal{A}_{\beta,r}$ from $L^2(Q_T)$ into $L^2(Q_T)$ be defined by $\mathcal{A}_{\beta,r}\lambda := L^\star\varphi$ where $\varphi = \varphi(\lambda) \in \Phi_\beta$ is the unique solution to

$$a_{\beta,r}(\varphi, \overline{\varphi}) = b(\overline{\varphi}, \lambda), \quad \forall\overline{\varphi} \in \Phi_\beta. \tag{21}$$

For any $r > 0$, the form $a_{\beta,r}$ defines a norm equivalent to the norm on Φ_β (see (15)), so that (21) is well-posed. The following crucial lemma holds true.

Lemma 3.1 *For any $r > 0$, the operator $\mathcal{A}_{\beta,r}$ is a strongly elliptic, symmetric isomorphism from $L^2(Q_T)$ into $L^2(Q_T)$.*

It allows to get the following proposition which permits to replace the minimization of J_β over W_β to the minimization of the functional $J^{\star\star}_{\beta,r}$ over $L^2(Q_T)$, which is a space much easier to approximate than W_β.

Proposition 3.1 *For any $r > 0$, let $\varphi^0 \in \Phi_\beta$ be the unique solution of*

$$a_{\beta,r}(\varphi^0, \overline{\varphi}) = l(\overline{\varphi}), \quad \forall\overline{\varphi} \in \Phi_\beta$$

and let $J^{\star\star}_{\beta,r} : L^2(Q_T) \to L^2(Q_T)$ be the functional defined by

$$J^{\star\star}_{\beta,r}(\lambda) := \frac{1}{2}(\mathcal{A}_{\beta,r}\lambda,\ \lambda)_{L^2(Q_T)} - b(\varphi^0, \lambda).$$

The following equality holds:

$$\sup_{\lambda\in L^2(Q_T)} \inf_{\varphi\in\Phi_\beta} \mathcal{L}_{\beta,r}(\varphi, \lambda) = -\inf_{\lambda\in L^2(Q_T)} J^{\star\star}_{\beta,r}(\lambda) \ + \ \mathcal{L}_{\beta,r}(\varphi^0, 0).$$

We refer to [21, section 2.1], for the proof in the case $M = 0$.

Remark 5 By introducing appropriate weights functions (vanishing at the time $t = T$) leading to optimal L^2-weighted controls vanishing at time T, we may consider the case $\beta = 0$. We refer to [21, section 2.3].

3.2 Numerical Approximation

We now turn to the discretization of the mixed formulation (17) assuming $r > 0$. We follow [21] for which we refer for the details. Let then $\Phi_{\beta,h}$ and $M_{\beta,h}$ be two finite dimensional spaces parametrized by the variable h such that, for any $\beta > 0$,

$$\Phi_{\beta,h} \subset \Phi_\beta, \quad M_{\beta,h} \subset L^2(Q_T), \quad \forall h > 0.$$

Then, we can introduce the following approximated problems: find $(\varphi_h, \lambda_h) \in \Phi_{\beta,h} \times M_{\beta,h}$ solution of

$$\begin{cases} a_{\beta,r}(\varphi_h, \overline{\varphi}_h) + b(\overline{\varphi}_h, \lambda_h) = l(\overline{\varphi}_h), & \forall \overline{\varphi}_h \in \Phi_{\beta,h} \\ b(\varphi_h, \overline{\lambda}_h) = 0, & \forall \overline{\lambda}_h \in M_{\beta,h}. \end{cases} \tag{22}$$

The well-posedness of this mixed formulation is a consequence of two properties: the first one is the coercivity of the form $a_{\beta,r}$ on the subset $\mathcal{N}_h(b) = \{\varphi_h \in \Phi_{\beta,h}; b(\varphi_h, \lambda_h) = 0 \quad \forall \lambda_h \in M_{\beta,h}\}$. Actually, from the relation

$$a_{\beta,r}(\varphi, \varphi) \geq C_{r,\eta} \|\varphi\|^2_{\Phi_\beta}, \quad \forall \varphi \in \Phi_\beta,$$

where $C_{r,\eta} = \min\{1, r/\eta\}$, the form $a_{\beta,r}$ is coercive on the full space Φ_β, and so *a fortiori* on $\mathcal{N}_h(b) \subset \Phi_{\beta,h} \subset \Phi_\beta$. The second property is a discrete inf-sup condition:

$$\delta_{r,h} := \inf_{\lambda_h \in M_{\beta,h}} \sup_{\varphi_h \in \Phi_{\beta,h}} \frac{b(\varphi_h, \lambda_h)}{\|\varphi_h\|_{\Phi_{\beta,h}} \|\lambda_h\|_{M_{\beta,h}}} > 0 \quad \forall h > 0. \tag{23}$$

Let us assume that this property holds. Consequently, for any fixed $h > 0$, there exists a unique couple (φ_h, λ_h) solution of (22). The property (23) is in general difficult to prove and strongly depends on the choice made for the approximated spaces $M_{\beta,h}$ and $\Phi_{\beta,h}$. We shall analyze numerically this property in the next section.

Remark 6 For $r = 0$, the discrete formulation (22) is not well-posed over $\Phi_{\beta,h} \times M_{\beta,h}$ because the form $a_{\beta,r=0}$ is not coercive over the discrete kernel of b: the equality $b(\lambda_h, \varphi_h) = 0$ for all $\lambda_h \in M_{\beta,h}$ does not imply that $L^\star \varphi_h$ vanishes. The term $r\|L^\star \varphi_h\|^2_{L^2(Q_T)}$ is a numerical stabilization term: for any $h > 0$, it ensures the uniform coercivity of the form $a_{\beta,r}$ and vanishes at the limit in h. We also emphasize that this term is not a regularization term as it does not add any regularity to the solution φ_h.

The finite dimensional and conformal space $\Phi_{\beta,h}$ must be chosen such that $L^\star \varphi_h$ belongs to $L^2(Q_T)$ for any $\varphi_h \in \Phi_{\beta,h}$. This is guaranteed as soon as φ_h possesses second-order derivatives in $L^2(Q_T)$. Any conformal approximation based on standard triangulation of Q_T achieves this sufficient property as soon as it is generated by spaces of functions continuously differentiable with respect to the variable x and spaces of continuous functions with respect to the variable t.

We introduce a triangulation $\mathcal{T}_h$ such that $\overline{Q_T} = \cup_{K\in\mathcal{T}_h} K$ and we assume that $\{\mathcal{T}_h\}_{h>0}$ is a regular family. Then, we introduce the space $\Phi_{\beta,h}$ as follows:

$$\Phi_{\beta,h} = \{\varphi_h \in C^1(\overline{Q_T}) : \varphi_h|_K \in \mathbb{P}(K) \quad \forall K \in \mathcal{T}_h,\ \varphi_h = 0 \text{ on } \Sigma_T\} \tag{24}$$

where $\mathbb{P}(K)$ denotes an appropriate space of polynomial functions in x and t. In this work, we consider for $\mathbb{P}(K)$ the so-called *Bogner-Fox-Schmit* (BFS for short) C^1-element defined for rectangles. In the one dimensional setting (in space), $\mathbb{P}(K) = (\mathbb{P}_{3,x} \otimes \mathbb{P}_{3,t})(K)$ where $\mathbb{P}_{r,\xi}$ is the space of polynomial functions of order r in the variable ξ.

We also define the finite dimensional space

$$M_{\beta,h} = \{\lambda_h \in C^0(\overline{Q_T}) : \lambda_h|_K \in \mathbb{Q}(K) \quad \forall K \in \mathcal{T}_h\},$$

where $\mathbb{Q}(K)$ denotes the space of affine functions both in x and t on the element K. In the one dimensional setting in space, K is a rectangle and we simply have $\mathbb{Q}(K) = (\mathbb{P}_{1,x} \otimes \mathbb{P}_{1,t})(K)$.

The resulting approximation is conformal: for any $h > 0$, $\Phi_{\beta,h} \subset \Phi_\beta$ and $M_{\beta,h} \subset L^2(Q_T)$.

Let $n_h = \dim \Phi_{\beta,h}$, $m_h = \dim M_{\beta,h}$ and let the real matrices $A_{\beta,r,h} \in \mathbb{R}^{n_h,n_h}$, $B_h \in \mathbb{R}^{m_h,n_h}$, $J_h \in \mathbb{R}^{m_h,m_h}$ and $L_h \in \mathbb{R}^{n_h}$ be defined by

$$\begin{cases} a_{\beta,r}(\varphi_h, \overline{\varphi_h}) =< A_{\beta,r,h}\{\varphi_h\}, \{\overline{\varphi_h}\} >_{\mathbb{R}^{n_h},\mathbb{R}^{n_h}} & \forall \varphi_h, \overline{\varphi_h} \in \Phi_{\beta,h}, \\ b(\varphi_h, \lambda_h) =< B_h\{\varphi_h\}, \{\lambda_h\} >_{\mathbb{R}^{m_h},\mathbb{R}^{m_h}} & \forall \varphi_h \in \Phi_{\beta,h} \lambda_h \in M_{\beta,h}, \\ \displaystyle\iint_{Q_T} \lambda_h \overline{\lambda_h}\, dx\, dt =< J_h\{\lambda_h\}, \{\overline{\lambda_h}\} >_{\mathbb{R}^{m_h},\mathbb{R}^{m_h}} & \forall \lambda_h, \overline{\lambda_h} \in M_{\beta,h}, \\ l(\varphi_h) =< L_h, \{\varphi_h\} > & \forall \varphi_h \in \Phi_{\beta,h}, \end{cases}$$

where $\{\varphi_h\} \in \mathbb{R}^{n_h}$ denotes the vector associated to φ_h and $< \cdot, \cdot >_{\mathbb{R}^{n_h},\mathbb{R}^{n_h}}$ the usual scalar product over $\mathbb{R}^{n_h}$. With these notations, Problem (22) reads as follows: find $\{\varphi_h\} \in \mathbb{R}^{n_h}$ and $\{\lambda_h\} \in \mathbb{R}^{m_h}$ such that

$$\begin{pmatrix} A_{\beta,r,h} & B_h^T \\ B_h & 0 \end{pmatrix}_{\mathbb{R}^{n_h+m_h,n_h+m_h}} \begin{pmatrix} \{\varphi_h\} \\ \{\lambda_h\} \end{pmatrix}_{\mathbb{R}^{n_h+m_h}} = \begin{pmatrix} L_h \\ 0 \end{pmatrix}_{\mathbb{R}^{n_h+m_h}}.$$

3.2.1 The Discrete inf-sup Test

Before to discuss some numerical experiments, we numerically test the discrete inf-sup condition (23). Taking $\eta = r > 0$ so that $a_{\beta,r}(\varphi, \overline{\varphi}) = (\varphi, \overline{\varphi})_{\Phi_\beta}$ exactly for all $\varphi, \overline{\varphi} \in \Phi_\beta$, it is readily seen (see for instance [5]) that the discrete inf-sup constant

Table 1 $\delta_{\beta,r,h}$ w.r.t. h and r; $\varepsilon = 10^{-1}$—$\beta = 10^{-16}$—$M = 1$

r	10	1	0.1	h	h^2
$h = 1/80$	0.315	0.919	1.909	2.359	2.535
$h = 1/160$	0.313	0.923	1.94	2.468	2.599
$h = 1/320$	0.313	0.927	1.969	2.548	2.658

Table 2 $\delta_{\beta,r,h}$ w.r.t. h and r; $\varepsilon = 10^{-2}$—$\beta = 10^{-16}$—$M = 1$

r	10	1	0.1	h	h^2
$h = 1/80$	0.311	0.961	2.423	3.64	4.473
$h = 1/160$	0.316	0.967	2.492	4.06	4.692
$h = 1/320$	0.316	0.971	2.545	4.406	4.916

Table 3 $\delta_{\beta,r,h}$ w.r.t. h and r; $\varepsilon = 10^{-3}$—$\beta = 10^{-16}$—$M = 1$

r	10	1	0.1	h	h^2
$h = 1/80$	0.310	0.942	2.121	3.412	6.012
$h = 1/160$	0.310	0.987	2.435	4.012	5.944
$h = 1/320$	0.310	0.969	2.544	4.561	5.756

satisfies

$$\delta_{\beta,r,h} = \inf\left\{\sqrt{\delta} : B_h A_{\beta,r,h}^{-1} B_h^T\{\lambda_h\} = \delta\, J_h\{\lambda_h\}, \quad \forall\,\{\lambda_h\} \in \mathbb{R}^{m_h} \setminus \{0\}\right\}. \tag{25}$$

The matrix $B_h A_{\beta,r,h}^{-1} B_h^T$ enjoys the same properties than the matrix $A_{\beta,r,h}$: it is symmetric and positive definite so that the scalar $\delta_{\beta,r,h}$ defined in term of the (generalized) eigenvalue problem (25) is strictly positive. This eigenvalue problem is solved using the power iterate algorithm (assuming that the lowest eigenvalue is simple): for any $\{v_h^0\} \in \mathbb{R}^{n_h}$ such that $\|\{v_h^0\}\|_2 = 1$, compute for any $n \geq 0$, $\{\varphi_h^n\} \in \mathbb{R}^{n_h}$, $\{\lambda_h^n\} \in \mathbb{R}^{m_h}$ and $\{v_h^{n+1}\} \in \mathbb{R}^{m_h}$ iteratively as follows:

$$\begin{cases} A_{\beta,r,h}\{\varphi_h^n\} + B_h^T\{\lambda_h^n\} = 0 \\ B_h\{\varphi_h^n\} = -J_h\{v_h^n\} \end{cases}, \quad \{v_h^{n+1}\} = \frac{\{\lambda_h^n\}}{\|\{\lambda_h^n\}\|_2}.$$

The scalar $\delta_{\beta,r,h}$ defined by (25) is then given by $\delta_{\beta,r,h} = \lim_{n\to\infty}(\|\{\lambda_h^n\}\|_2)^{-1/2}$.

We now reports some numerical values of $\delta_{\beta,r,h}$ with respect to h for the C^1-finite element introduced in Sect. 3.2. We use the value $T = 1$ and $\beta = 10^{-16}$. Tables 1, 2 and 3 provides the value of $\delta_{\beta,r,h}$ with respect to h and r for $M = 1$ for $\varepsilon = 10^{-1}, 10^{-2}$ and $\varepsilon = 10^{-3}$ respectively. For a fixed value of the parameter ε, we observe as in [21], that the inf sup constant increases as $r \to 0$ and behaves like $\delta_{\beta,r,h} \approx r^{-1/2}$, and more importantly, is bounded by below uniformly with respect to h. This key property is preserved as the parameter ε decreases, in agreement with the estimate (20) uniform with respect to ε.

The case $M = -1$ is reported in Tables 4, 5 and 6. The same behavior is observed except that we note larger values of the inf-sup constant.

Table 4 $\delta_{\beta,r,h}$ for $\varepsilon = 10^{-1}$—$\beta = 10^{-16}$—$M = -1$

r	10	1	0.1	h	h^2
$h = 1/80$	0.3161	0.997	2.663	4.358	5.069
$h = 1/160$	0.316	0.9805	2.673	4.69	5.139
$h = 1/320$	0.3162	0.9801	2.653	4.172	5.171

Table 5 $\delta_{\beta,r,h}$ for $\varepsilon = 10^{-2}$—$\beta = 10^{-16}$—$M = -1$

r	10	1	0.1	h	h^2
$h = 1/80$	0.316	0.997	3.109	7.562	13.936
$h = 1/160$	0.3161	0.9997	3.086	9.433	14.101
$h = 1/320$	0.316	0.9809	3.086	11.101	14.140

Table 6 $\delta_{\beta,r,h}$ for $\varepsilon = 10^{-3}$—$\beta = 10^{-16}$—$M = -1$

r	10	1	0.1	h	h^2
$h = 1/80$	0.302	0.9129	2.887	8.16	39.09
$h = 1/160$	0.301	0.957	3.022	12.14	43.08
$h = 1/320$	0.301	0.981	3.084	16.61	44.29

Consequently, we may conclude that the finite approximation we have used "passes" the discrete inf-sup test. Such property together with the uniform coercivity of the form $a_{\beta,r}$ then imply the convergence of the approximation sequence (φ_h, λ_h), unique solution of (22). As the matter of fact, the use of stabilization technics (so as to enrich the coercivity of the saddle point problem) introduced and analyzed in a closed context in [20, 22] is not necessary here. We emphasize that for $\beta = 0$ (or $\beta \to 0$ as $h \to 0$), the convergence of the approximation v_h is still an open issue. For $\beta = 0$, the convergence is guarantees if a vanishing weight is introduced, see [11]. This however leads to a different control and therefore a different definition of the cost of control $K(\varepsilon, T, M)$.

The choice of r affects the convergence of the sequences φ_h and λ_h with respect to h and may be very important here, in view of the sensitivity of the boundary control problem with respect to ε. Recall from Theorem 3.1, that for any $r \geq 0$, the multiplier λ coincides with the controlled solution. At the finite dimensional level of the mixed formulation (22) where r must be strictly positive, this property is lost for any h fixed: the non zero augmentation term $r\|L^\star\varphi_h\|_{L^2(Q_T)}$ introduces a small perturbation and requires to take $r > 0$ small (in order that the approximation λ_h be closed to the controlled solution y). In the sequel, the value $r = h^2$ is used.

3.3 Numerical Experiments

We discuss some experiments for both $M = 1$ and $M = -1$ respectively and several values of ε. We consider a fixed data, independent of the parameter ε: precisely, we take $y_0(x) = \sin(\pi x)$ for $x \in (0, L)$ and $L = 1$.

We consider regular but non uniform rectangular meshes refined near the four edges of the space-time domain Q_T. More precisely, we refine at the edge $\{x =$

Table 7 Approximation $\|y_h(\cdot, T)\|_{H^{-1}(0,L)}$ w.r.t. T and ε for $y_0(x) = \sin(\pi x)$—$M = L = 1$

ε	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
$T = 0.9L/\|M\|$	2.20×10^{-2}	7.45×10^{-4}	2.76×10^{-3}	2.20×10^{-3}	2.15×10^{-3}
$T = L/\|M\|$	1.58×10^{-2}	2.67×10^{-3}	1.72×10^{-4}	9.76×10^{-6}	3.07×10^{-7}
$T = 1.1L/\|M\|$	1.12×10^{-2}	8.13×10^{-4}	1.15×10^{-6}	1.63×10^{-19}	8.62×10^{-20}

$1\} \times (0, T)$ to capture the boundary layer of length ε which appear for the variable λ_h when M is positive (see [1]), at the edge $\{x = 0\} \times (0, T)$ to approximate correctly the "control" function given by $v_h := \varepsilon\varphi_{h,x}$, and finally at $(0, L) \times \{0, T\}$ to represent correctly the initial condition and final condition. Precisely, let $p : [0, L] \to [0, L]$ be the polynomial of degree 3 such that $p(0) = 0, p'(0) = \eta_1, p'(L) = \eta_2$ and $p(L) = L$ for some fixed $\eta_1, \eta_2 > 0$. The $[0, L]$ interval is then discretized as follows:

$$\begin{cases} [0, L] = \cup_{j=0}^{J}[y_j, y_{j+1}], \\ y_0 = 0, y_j - y_{j-1} = p(x_j) - p(x_{x_{j-1}}), \quad j = 1, \cdots, J+1 \end{cases} \tag{26}$$

where $\{x_j\}_{j=0,\cdots,J+1}$ is the uniform discretization of $[0, L]$ defined by $x_j = jh, j = 0, \cdot, J + 1, h = L/(J + 1)$. Small values for η_1, η_2 lead to a refined discretization $\{y_j\}_{j=0,\cdots,J+1}$ at $x = 0$ and $x = L$. The same procedure is used for the time discretization of $[0, T]$. In the sequel, we use $\eta_1 = \eta_2 = 10^{-3}$.

Preliminary, Table 7 gives some values of the H^{-1}-norm of the uncontrolled solution of (1) at time T associated to $y_0(x) = \sin(\pi x)$. We take $L = |M| = 1$. A time-marching approximation scheme is used with a very fine discretization both in time and space. As expected, for T greater than $L/|M|$, the norm $\|y(\cdot, T)\|_{H^{-1}(0,1)}$ decreases as ε goes to zero. For $T = L/M$, we observe that $\|y(\cdot, T)\|_{H^{-1}(0,1)} = O(\varepsilon)$ while for T strictly greater than $L/|M|$, the decrease to zero as $\varepsilon \to 0$ is faster.

We first discuss the case $M = 1$. As ε goes to 0^+, a boundary layer appears for the approximation λ_h at $x = 1$. The profile of the solution takes along the normal the form $(1 - e^{\frac{-M(1-x)}{\varepsilon}})$ and is captured with a locally refined mesh (we refer to [1]). Tables 8, 9 and 10 reports some numerical norms for $\epsilon = 10^{-1}, 10^{-2}$ and 10^{-3} respectively. These results are obtained by minimizing the functional $J^{\star\star}_{\beta,r}$ over $M_{\beta,h}$ defined in Proposition 3.1. The minimization of $J^{\star\star}_{\beta,r}$ of M_h is performed using the conjugate gradient algorithm: the stopping criterion is $\|g_h^n\|_{L^2(Q_T)} \leq 10^{-6}\|g_h^0\|_{L^2(Q_T)}$ where g_h^n is the residus at the iterate n. The algorithm is initialized with $\lambda_h^0 = 0$. We refer to [21] for the details.

We take $\beta = 10^{-16}$ and $r = h^2$ for the augmentation parameter leading to an appropriate approximation of the controlled solution y by the function λ_h: in particular, the optimality condition $\lambda_h(0, \cdot) - \varepsilon\varphi_{h,x}(0, \cdot) = 0$ is well respected in $L^2(0, T)$. The convergence of $\sqrt{r}\|L^\star\varphi_h\|_{L^2(Q_T)}$ (close to $\|L^\star\varphi_h\|_{L^2(H^{-1})}$ and actually sufficient to describe the solution of (1), see [7]) is also observed. As usual, we observe a faster convergence for the norm $\|\lambda_h\|_{L^2(Q_T)}$ than for the norm

Table 8 Mixed formulation (17)—$r = h^2$; $\varepsilon = 10^{-1}$; $\beta = 10^{-16}$—$M = L = 1$

h	$1/80$	$1/160$	$1/320$	$1/640$
$\sqrt{r}\|L^\star\varphi_h\|_{L^2(Q_T)}$	7.76×10^{-2}	3.01×10^{-2}	1.12×10^{-2}	7.12×10^{-3}
$\frac{\|\varepsilon\varphi_x(0,\cdot)-\lambda_h(0,\cdot)\|_{L^2(0,T)}}{\|\lambda_h(0,\cdot)\|_{L^2(0,T)}}$	1.06×10^{-2}	4.45×10^{-3}	1.97×10^{-3}	7.61×10^{-4}
$\|v_h\|_{L^2(0,T)}$	0.324	0.357	0.3877	0.3912
$\|\lambda_h\|_{L^2(Q_T)}$	0.367	0.366	0.362	0.363
$\|\lambda_h(\cdot,T)\|_{H^{-1}(0,T)}$	4.47×10^{-6}	9.59×10^{-7}	2.03×10^{-7}	1.01×10^{-7}
♯ CG iterate	76	117	175	231

Table 9 Mixed formulation (17)—$r = h^2$; $\varepsilon = 10^{-2}$; $\beta = 10^{-16}$—$M = L = 1$

h	$1/80$	$1/160$	$1/320$	$1/640$
$\sqrt{r}\|L^\star\varphi_h\|_{L^2(Q_T)}$	5.86×10^{-1}	2.43×10^{-1}	1.41×10^{-1}	9.12×10^{-2}
$\frac{\|\varepsilon\varphi_x(0,\cdot)-\lambda_h(0,\cdot)\|_{L^2(0,T)}}{\|\lambda_h(0,\cdot)\|_{L^2(0,T)}}$	2.5×10^{-2}	1.24×10^{-2}	6.04×10^{-3}	2.89×10^{-3}
$\|v_h\|_{L^2(0,T)}$	1.391	2.392	2.929	3.316
$\|\lambda_h\|_{L^2(Q_T)}$	0.518	0.6001	0.789	0.832
$\|\lambda_h(\cdot,T)\|_{H^{-1}(0,T)}$	5.46×10^{-6}	3.56×10^{-6}	8.77×10^{-7}	6.12×10^{-8}
♯ CG iterate	53	93	155	181

Table 10 Mixed formulation (17)—$r = h^2$; $\varepsilon = 10^{-3}$; $\beta = 10^{-16}$—$M = L = 1$

h	$1/80$	$1/160$	$1/320$	$1/640$
$\sqrt{r}\|L^\star\varphi_h\|_{L^2(Q_T)}$	1.75×10^{-1}	1.01×10^{-1}	8.51×10^{-2}	6.91×10^{-2}
$\frac{\|\varepsilon\varphi_x(0,\cdot)-\lambda_h(0,\cdot)\|_{L^2(0,T)}}{\|\lambda_h(0,\cdot)\|_{L^2(0,T)}}$	4.87×10^{-2}	2.43×10^{-2}	1.3×10^{-4}	7.19×10^{-5}
$\|v_h\|_{L^2(0,T)}$	0.231	0.713	0.855	0.911
$\|\lambda_h\|_{L^2(Q_T)}$	0.498	0.5015	0.5210	0.5319
$\|\lambda_h(\cdot,T)\|_{H^{-1}(0,T)}$	1.17×10^{-6}	3.69×10^{-7}	1.20×10^{-7}	8.12×10^{-8}
♯ CG iterate	29	68	129	151

$\|v_h\|_{L^2(0,T)}$. From $\varepsilon = 10^{-1}$ to 10^{-3}, we also clearly observe a deterioration of the convergence order with respect to h.

For $h = 1/320$, Figs. 1, 2 and 3 depict the function $\lambda_h(\cdot,t)$, approximation of the control v, for $t \in (0,T)$, $T = 1$ for $\varepsilon = 10^{-1}$, $\varepsilon = 10^{-2}$ and $\varepsilon = 10^{-3}$ respectively. For large values of the diffusion coefficient ε, for instance $\varepsilon = 10^{-1}$, the transport term has a weak influence: the control of minimal L^2-norm is similar to the corresponding control for the heat equation and oscillates near the controllability time. On the contrary, for ε small, typically $\varepsilon = 10^{-3}$, the solution—mainly driven by the transport term—is transported along a direction closed to $(1, 1/M) = (1, 1)$, so that at time $T = 1/M$, is mainly distributed in the neighborhood of $x = 1$. Consequently, the control (of minimal L^2-norm) acts mainly at the beginning of the

Fig. 1 Approximation $\lambda_h(0,t)$ of the control w.r.t. $t \in [0,T]$ for $\varepsilon = 10^{-1}$ and $T = L = M = 1$; $r = h^2$—$h = 1/320$

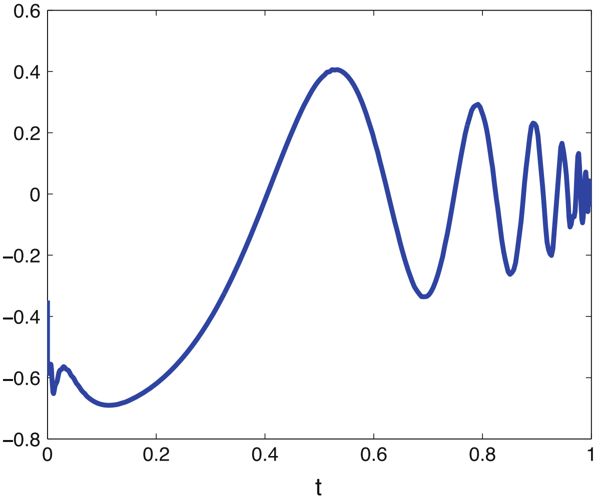
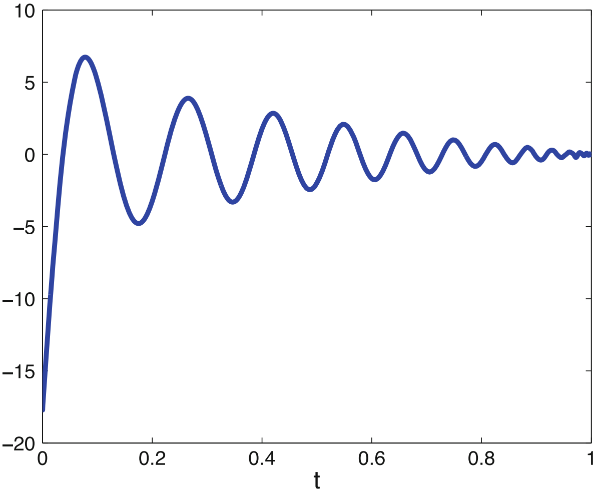


Fig. 2 Approximation $\lambda_h(0,t)$ of the control w.r.t. $t \in [0,T]$ for $\varepsilon = 10^{-2}$ and $T = L = M = 1$; $r = h^2$—$h = 1/320$



time interval, so as to have an effect, at time T, in the neighborhood of $x = 1$. We observe a regular oscillatory and decreasing behavior of the controls.

Let us now discuss the case $M = -1$. This negative case is *a priori* “simpler” since there is no more boundary layer at $x = 1$: the solution is somehow “absorbed” by the control at the left edge $x = 0$. Tables 11, 12 and 13 give some numerical values with respect to h for $\varepsilon = 10^{-1}, 10^{-2}$ and 10^{-3}. Concerning the behavior of the approximation with respect to h, similar remarks (than for $M = 1$) can be made: the notable difference is a lower rate of convergence, probably due to the singularity of the controls we obtain. Precisely, for the same data as in the case $M = 1$, Figs. 4, 5 and 6 depicts the “control” function $\lambda_h(0,t)$ for $t \in (0,T)$,

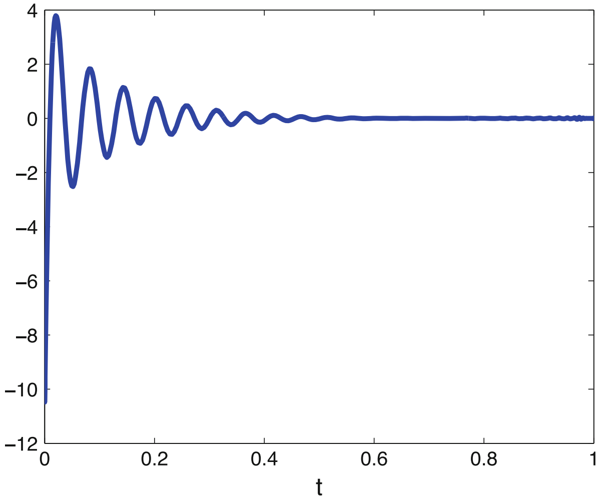


Fig. 3 Approximation $\lambda_h(0,t)$ of the control w.r.t. $t \in [0,T]$ for $\varepsilon = 10^{-3}$ and $T = L = M = 1$; $r = h^2$—$h = 1/320$

Table 11 Mixed formulation (17)—$r = h^2$; $\varepsilon = 10^{-1}$; $\beta = 10^{-16}$—$M = -1$

h	1/80	1/160	1/320	1/640
$\sqrt{r}\|L^\star\varphi_h\|_{L^2(Q_T)}$	1.51	0.731	0.231	0.101
$\frac{\|\varepsilon\varphi_x(0,\cdot)-\lambda_h(0,\cdot)\|_{L^2(0,T)}}{\|\lambda_h(0,\cdot)\|_{L^2(0,T)}}$	9.19×10^{-3}	3.87×10^{-3}	1.61×10^{-3}	1.12×10^{-3}
$\|v_h\|_{L^2(0,T)}$	28.16	39.26	49.96	52.03
$\|\lambda_h\|_{L^2(Q_T)}$	5.74	7.96	9.05	10.12
$\|\lambda_h(\cdot,T)\|_{H^{-1}(0,T)}$	8.35×10^{-4}	1.82×10^{-4}	3.97×10^{-5}	1.12×10^{-5}
♯ CG iterate	48	80	129	157

Table 12 Mixed formulation (17)—$r = h^2$; $\varepsilon = 10^{-2}$; $\beta = 10^{-16}$—$M = -1$

h	1/80	1/160	1/320	1/640
$\sqrt{r}\|L^\star\varphi_h\|_{L^2(Q_T)}$	5.291	2.134	1.213	0.591
$\frac{\|\varepsilon\varphi_x(0,\cdot)-\lambda_h(0,\cdot)\|_{L^2(0,T)}}{\|\lambda_h(0,\cdot)\|_{L^2(0,T)}}$	5.27×10^{-4}	2.08×10^{-2}	8.05×10^{-3}	5.01×10^{-3}
$\|v_h\|_{L^2(0,T)}$	250.54	457.78	666.902	712.121
$\|\lambda_h\|_{L^2(Q_T)}$	6.76	10.05	13.111	15.301
$\|\lambda_h(\cdot,T)\|_{H^{-1}(0,T)}$	1.54×10^{-3}	2.08×10^{-3}	1.71×10^{-3}	6.12×10^{-4}
♯ CG iterate	22	41	79	101

$T = 1$ for $\varepsilon = 10^{-1}$, $\varepsilon = 10^{-2}$ and $\varepsilon = 10^{-3}$ respectively. The behavior of the control is quite different from the previous case. For ε large, typically $\varepsilon = 10^{-1}$, the control is again similar to the control we observe for the heat equation, with an oscillatory behavior at the final time. We observe however that the corresponding norm is significantly larger that for the case $M = 1$: this is due to the fact, that for

Table 13 Mixed formulation (17)—$r = h^2$; $\varepsilon = 10^{-3}$; $\beta = 10^{-16}$—$M = -1$

h	1/80	1/160	1/320	1/640
$\sqrt{r}\Vert L^\star\varphi_h\Vert_{L^2(Q_T)}$	7.12	2.14	1.31	0.59
$\frac{\Vert\varepsilon\varphi_x(0,\cdot)-\lambda_h(0,\cdot)\Vert_{L^2(0,T)}}{\Vert\lambda_h(0,\cdot)\Vert_{L^2(0,T)}}$	2.87×10^{-1}	7.76×10^{-2}	4.31×10^{-2}	2.12×10^{-2}
$\Vert v_h\Vert_{L^2(0,T)}$	0.281×10^{-1}	2.35	18.98	21.23
$\Vert\lambda_h\Vert_{L^2(Q_T)}$	4.97×10^{-1}	5.01×10^{-1}	6.38×10^{-1}	7.23×10^{-1}
$\Vert\lambda_h(\cdot,T)\Vert_{H^{-1}(0,T)}$	2.03×10^{-5}	3.28×10^{-5}	6.01×10^{-5}	8.01×10^{-5}
♯ CG iterate	7	11	23	26

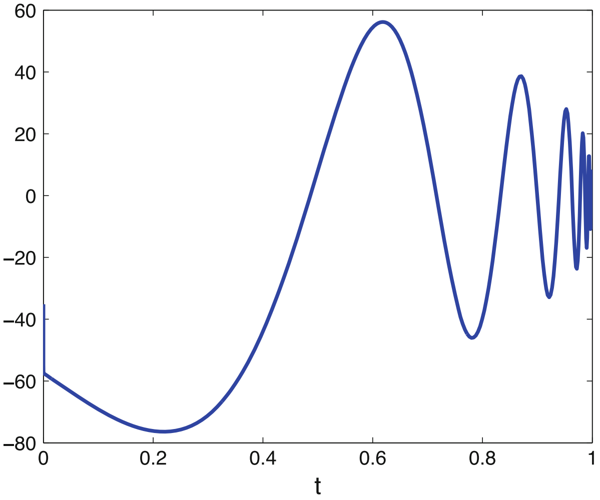


Fig. 4 Approximation $\lambda_h(0,t)$ of the control w.r.t. $t \in [0,T]$ for $\varepsilon = 10^{-1}$ and $T = L = -M = 1$; $r = h^2$—$h = 1/320$

$M < 0$, the transport term "pushes" the solution toward $x = 0$ where the control acts: this reduces the effect of the control which therefore must be stronger. For ε small, the solution is mainly transported along the direction $(1, 1/M) = (1, -1)$ so that at time T, the solution is mainly concentrated in the neighborhood of $x = 0$. For this reason, the control mainly acts at the end of the time interval: any action of the control not concentrated at the end of the time interval would be useless because pushed back to the edge $x = 0$ and will produce a larger L^2-norm. As ε goes to zero, the control is getting concentrated at the terminal time with an oscillatory behavior and large amplitudes. This fact may explain why the behavior of the cost of control with respect to ε observed in [9, 14, 18] is singular for negatives values of M. For $M > 0$, the transport term "helps" the control to act on the edge $x = 1$ while for $M < 0$, the transport term is against the control and reduces its action. For this reason, the numerical approximation of controls for $M = -1$ is definitively more involved and requires to take a very fine discretization, which will then imply a large number of CG iterates.

Fig. 5 Approximation $\lambda_h(0,t)$ of the control w.r.t. $t \in [0,T]$ for $\varepsilon = 10^{-2}$ and $T = L = -M = 1$; $r = h^2$—$h = 1/320$

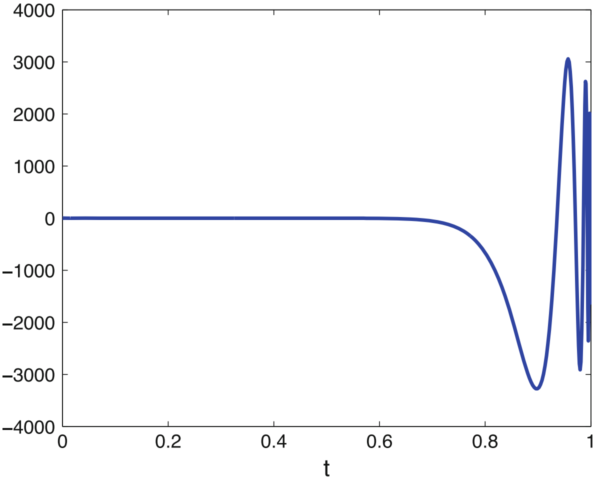
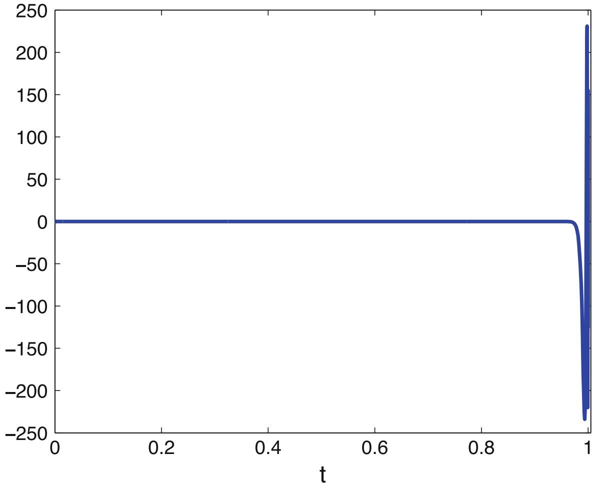


Fig. 6 Approximation $\lambda_h(0,t)$ of the control w.r.t. $t \in [0,T]$ for $\varepsilon = 10^{-3}$ and $T = L = -M = 1$; $r = h^2$—$h = 1/320$



We also observe, both for $M = 1$ and $M = -1$, that from $\varepsilon = 10^{-2}$ to $\varepsilon = 10^{-3}$, the L^2-norm $\|v_\varepsilon\|_{L^2(0,T)}$ decreases. Very likely, as ε goes to zero, this norm goes to zero. This does not contradict the theoretical results and is due to the fact that the initial condition we have taken here is independent of ε. In other words, the optimal problem (3) of control is not obtained for $y_0(x) = \sin(\pi x)$ nor by any initial condition independent of the parameter ε. This fact is proven in [1]. We remind that the initial condition $y_0(x) = e^{\frac{Mx}{2\varepsilon}} \sin(\pi x)$ is used in [9, 19].

4 Numerical Approximation of the Cost of Control

We now turn to the numerical approximation of the cost of control $K(\varepsilon, T, M)$ defined by (3). Precisely, we address numerically the resolution of the generalized eigenvalue problem (10):

$$\sup\left\{\lambda \in \mathbb{R} : \exists\, y_0 \in L^2(0,L),\, y_0 \neq 0,\ \text{s.t. } \mathcal{A}_\varepsilon y_0 = \lambda y_0 \quad \text{in} \quad L^2(0,L)\right\}.$$

Let V_h be a conformal approximation of the space $L^2(0,L)$ for all $h > 0$. We have then face to the following finite dimensional eigenvalues problem:

$$\sup\left\{\lambda \in \mathbb{R} : \exists\, y_{0,h} \in V_h,\, y_{0,h} \neq 0,\ \text{s.t. } \mathcal{A}_\varepsilon y_{0,h} = \lambda y_{0,h} \quad \text{in} \quad V_h\right\}.$$

$\mathcal{A}_\varepsilon y_{0,h}$ in $L^2(0,L)$ is defined as $-\varphi_h(\cdot,0)$ where $\varphi_h \in \Phi_{\beta,h}$ solves the variational formulation (22). Consequently, from the definition of $\Phi_{\beta,h}$ in (24), the space V_h is the set of C^1-functions and piecewise polynomial of order 3:

$$V_h = \left\{y_{0,h} \in C^1([0,L]) : y_{0,h}|_K \in \mathbb{P}_{3,x} \quad \forall K \in T_h\right\}$$

where T_h is the triangulation of $[0,L]$ defined by (26).

This kind of finite dimensional eigenvalue problems may be solved using the power iterate method (see [6]): the algorithm is as follows: given $y^0_{0,h} \in L^2(0,L)$ such that $\|y^0_{0,h}\|_{L^2(0,L)} = 1$, compute for all $k \geq 0$,

$$\begin{cases} z^k_{0,h} = \mathcal{A}_\varepsilon y^k_{0,h}, \quad k \geq 0, \\[2mm] y^{k+1}_{0,h} = \dfrac{z^k_{0,h}}{\|z^k_{0,h}\|_{L^2(0,L)}}, \quad k \geq 0. \end{cases}$$

The real sequence $\{\|z^k_{0,h}\|_{L^2(0,L)}\}$ then converges to the eigenvalue with largest modulus of the operator $\mathcal{A}_\varepsilon$, so that

$$\sqrt{\|z^k_{0,h}\|_{L^2(0,1)}} \to K(\varepsilon, T, M, L) \quad \text{as} \quad k \to \infty.$$

$\{y^k_{0,h}\}_{k>0}$ converges to the corresponding eigenvectors. The first step requires to compute the image of the control operator $\mathcal{A}_\varepsilon$: this is done by solving the mixed formulation (22) taking $y^k_{0,h}$ as initial condition for (1).

The algorithm is stopped as soon as the sequence $\{z_{0,h}^k\}_{k\geq 0}$ satisfies

$$\frac{\left|\|z_{0,h}^k\|_{L^2(0,L)} - \|z_{0,h}^{k-1}\|_{L^2(0,L)}\right|}{\|z_{0,h}^{k-1}\|_{L^2(0,1)}} \leq 10^{-3}, \tag{27}$$

for some $k > 0$.

We now report the numerical values for $L = 1$ and $M = \pm 1$. We initialize the algorithm with

$$y_0^0(x) = \frac{e^{-\frac{Mx}{2\epsilon}} \sin(\pi x)}{\|e^{-\frac{Mx}{2\epsilon}} \sin(\pi x)\|_{L^2(0,L)}}, \quad x \in (0, L).$$

4.1 Cost of Control in the Case $M = 1$

Table 14 in the Appendix reports the approximations obtained of the cost of control $K(\varepsilon, T, M)$ for $M = 1$ with respect to T and ε. They corresponds to the discretisation $h = 1/320$. As expected, for T strictly lower than $L/M = 1$, here $T = 0.95$ and $T = 0.99$, we obtain that the cost $K(\varepsilon, T, M)$ blows up as ε goes to zero. This is in agreement with the fact, that for $T < L/M$, the system (1) is not uniformly controllable with respect to the initial data y_0 and ε. Figure 7 displays the approximations with respect to ε for $T = 0.95$. On the other hand, for T larger than $L/M = 1$, we observe that the numerical approximation of $K(\varepsilon, T, M)$ is bounded with respect to ε. More precisely, the cost is not monotonous with respect to ε as it reaches a maximal value for $\varepsilon \approx 1.75 \times 10^{-3}$ for $T = 1$ and $\varepsilon \approx 6 \times 10^{-3}$ for

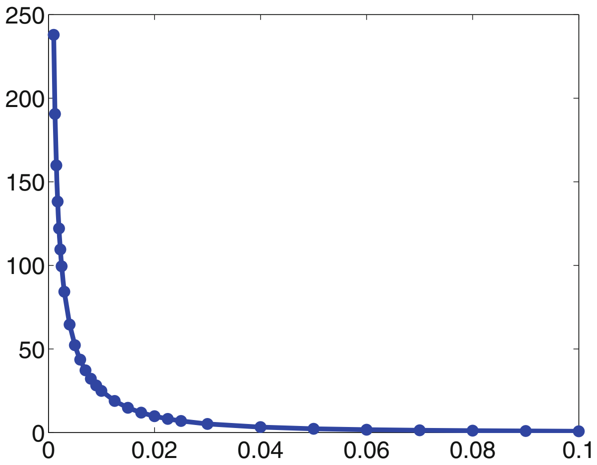


Fig. 7 Cost of control $K(\varepsilon, T, M)$ w.r.t. $\varepsilon \in [10^{-3}, 10^{-1}]$ for $T = 0.95L/M$ and $L = M = 1$; $r = h^2$—$h = 1/320$

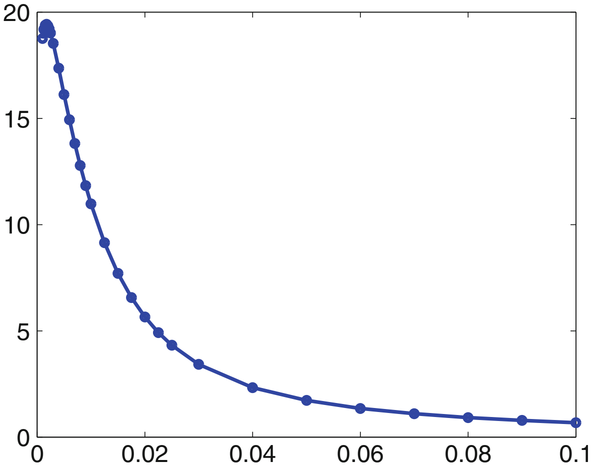


Fig. 8 Cost of control $K(\varepsilon, T, M)$ w.r.t. $\varepsilon \in [10^{-3}, 10^{-1}]$ for $T = L/M$ and $L = M = 1$; $r = h^2$—$h = 1/320$

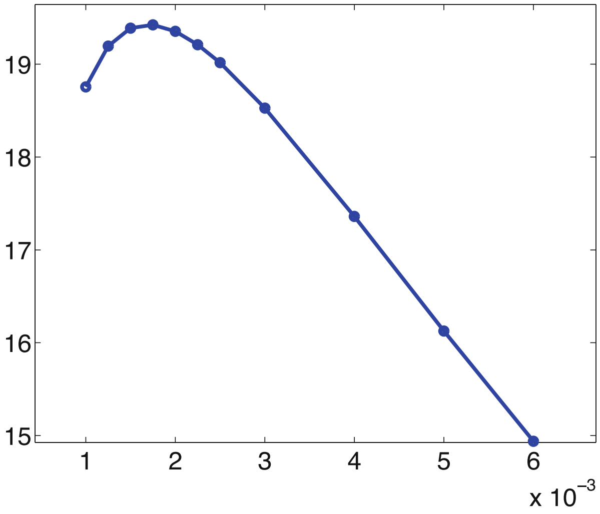


Fig. 9 Cost of control $K(\varepsilon, T, M)$ w.r.t. $\varepsilon \in [10^{-3}, 6\times 10^{-3}]$ for $T = 0.95L/M$ and $L = M = 1$; $r = h^2$—$h = 1/320$

$T = 1.05$ (see Figs. 8 and 10). Figure 9 is a zoom of Fig. 10 in the case $T = 1$ for the smallest values of the diffusion coefficient ε.

Figure 11 displays the approximation of the initial data $y_0 \in L^2(0, L)$ solution of the optimal problem (9) for $T = 1$ and $\varepsilon = 10^{-1}, 10^{-2}$ and 10^{-3}. As ε

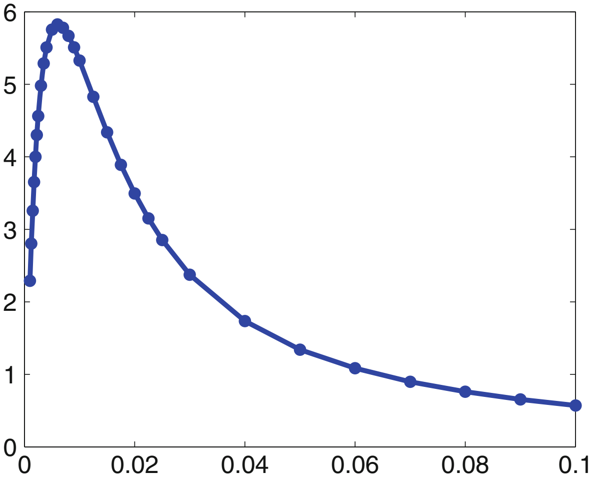


Fig. 10 Cost of control $K(\varepsilon, T, M)$ w.r.t. $\varepsilon \in [10^{-3}, 10^{-1}]$ for $T = 1.05L/M$ and $L = M = 1$; $r = h^2$—$h = 1/320$

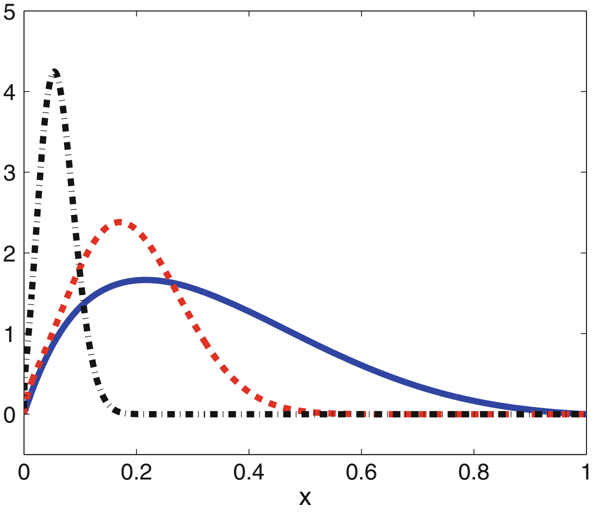


Fig. 11 The optimal initial condition y_0 in $(0, L)$ for $\varepsilon = 10^{-1}$ (full line), $\varepsilon = 10^{-2}$ (dashed line) and $\varepsilon = 10^{-3}$ (dashed-dotted line) and $T = M = L = 1$; $r = h^2$—$h = 1/320$

decreases, the optimal initial condition y_0 with $\|y_0\|_{L^2(0,L)} = 1$ gets concentrated as $x = 0$. Again, this is in agreement with the intuition since such condition produces (in the uncontrolled situation) larger values of $\|y(\cdot, T)\|_{H^{-1}(0,L)}$.

It should be noted however that the solutions we get are different from $e^{-\frac{Mx}{2\epsilon}}\sin(\pi x)/\|e^{-\frac{Mx}{2\epsilon}}\sin(\pi x)\|_{L^2(0,L)}$. Moreover, they are apparently independent of the controllability time T (at least for the values of T closed to $1/M$ we have used). Remark also that the initial data $y_0(x) = e^{\frac{Mx}{2\epsilon}}\sin(\pi x)/\|e^{\frac{Mx}{2\epsilon}}\sin(\pi x)\|_{L^2(0,L)}$ highlighted in [9, 19] leads to a lower numerical value of $\|v_h\|_{L^2(0,L)}$.

For each values of ε and T, the convergence of the power iterate algorithm is fast: the stopping criterion (27) is reached in less than 5 iterates.

Remark 7 In [9], Theorem 2, the following estimate is obtained for all $(\varepsilon, T, M) \in]0, \infty[$ and $L = 1$:

$$K(\varepsilon, T, M) \geq C_1 \frac{\varepsilon^{-3/2}T^{-1/2}M^2}{1 + M^3\varepsilon^{-3}} \exp\left(\frac{M}{2\varepsilon}(1 - TM) - \pi^2\varepsilon T\right) := C_1 f(\varepsilon, T, M)$$

for a positive constant C_1. This estimate is in agreement with the behavior we observe with respect to ε and T in the previous figures. For $T = 0.95/M$, the function f increases as $\varepsilon \to 0$, while for $T \geq 1/M$, f increases, reaches a unique maximum and then decreases to 0 as ε goes to zero.

4.2 Controllability Cost in the Case $M = -1$

Table 15 in the Appendix reports the approximation obtained of the cost of control $K(\varepsilon, T, M)$ for $M = -1$ and $T = 1/|M|$ with respect to $\varepsilon \in [10^{-3}, 10^{-1}]$. With respect to the positive case, the notable difference is the amplitude of the cost, as expected much larger, since the transport term now acts “against” the control. For instance, for $\varepsilon = 10^{-3}$, we obtain $K(\varepsilon, T, M) \approx 18.7555$ for $M = 1$ and $K(\varepsilon, T, M) \approx 1.0718 \times 10^4$ for $M = -1$. Moreover, the corresponding optimal initial condition y_0 is supported as $\varepsilon \to 0$ at the right extremity $x = 1$ (see Fig. 12) leading to a corresponding control localized at $t = T = 1/|M|$, with very large amplitude and oscillations, as shown on Fig. 13 for $\varepsilon = 10^{-3}$. Such oscillations are difficult to capture numerically and are very sensitive to the discretization used. On the other hand, we observe, as for $M = 1$, that the cost $K(\varepsilon, T, M)$ does not blow up as $\varepsilon \to 0$, in contradiction with the theoretical results from [9, 19]. The discretization used is not fine enough here to capture the highly oscillatory behavior of the control near the controllability time T (in contrast to the positive case) and very likely leads to an uncorrect approximation of the controls. For T lower than $1/|M|$, as expected, we observe that the cost blows up, while for T strictly greater than $1/|M|$, the cost decreases to zero with ε (Fig. 14).

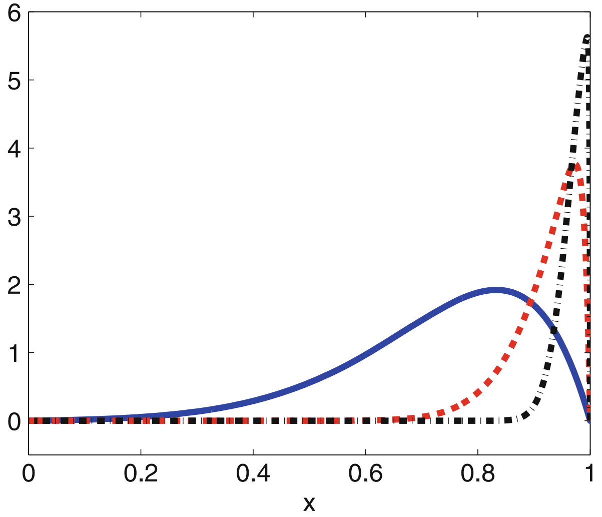


Fig. 12 The optimal initial condition y_0 in $(0, L)$ for $\varepsilon = 10^{-1}$ (full line), $\varepsilon = 10^{-2}$ (dashed line) and $\varepsilon = 10^{-3}$ (dashed-dotted line) and $T = -M = L = 1$; $r = h^2$—$h = 1/320$

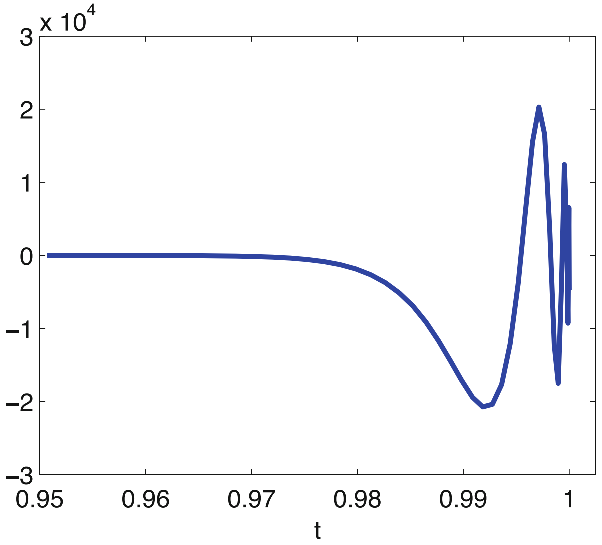


Fig. 13 Approximation $\lambda_h(0, t)$ of the corresponding control w.r.t. $t \in [0, T]$ for $\varepsilon = 10^{-3}$ and $T = L = -M = 1$; $r = h^2$—$h = 1/320$

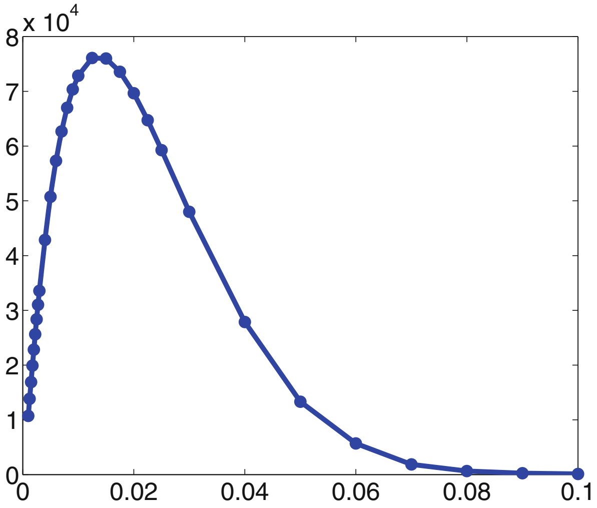


Fig. 14 Cost of control $K(\varepsilon, T, M)$ w.r.t. $\varepsilon \in [10^{-3}, 10^{-1}]$ for $T = L/|M|$ and $L = -M = 1$; $r = h^2$—$h = 1/320$

5 Concluding Remarks and Perspectives

We have presented a direct method to approximate the cost of control associated to the equation $y_t - \varepsilon y_{xx} + My_x = 0$. For $M > 0$, the "worst" initial data we observe are concentrated at $x = 0$ leading to a control distributed at the beginning of the time interval, and vanishing as $t \to T$. In this case, controls v are smooth and easily approximated. Vanishing exponentially weighs as considered in [21] leading to strong convergent results (w.r.t. h) are not necessary here. Consequently, for $M > 0$, we are confident with the numerical approximation obtained and may conjecture that the minimal time of uniform controllability w.r.t. ε is $T_M = L/M$. The situation is much more singular for $M < 0$ for which the transport term acts "against" the control. The optimal initial data are now concentrated as the right extremity leading to a highly singular controls at the end of the time interval. Such controls, similar to the controls we observed for the heat equation (see [23]) are difficult to approximate. The strong convergent approximation of controls w.r.t. h is still open in such situations. Let us comment possible perspectives to improve the resolution of this singular controllability problem.

(a) A way to recover a strong convergent approximation with respect to h is to force the control to vanish exponentially as time T of the form $v(t) := \varepsilon\rho^{-2}(t)\varphi_x(0, t)$, with $\rho(t) := O(e^{1/(T-t)})$. Remark that this modifies the cost of control as follows:

$$K_\rho(\varepsilon, T, M) := \sup_{\|y_0\|_{L^2(0,L)}=1} \left\{ \min_{u \in \mathcal{C}(y_0,T,\varepsilon,M)} \|\rho\, u\|_{L^2(0,T)} \right\},$$

larger than $K(\varepsilon, T, M)$ leading a priori to an upper bound $T_{M,\rho}$ of T_M. Since ρ^{-1} vanishes only at time T, we suspect that the minimal time of uniform controllability $T_{M,\rho}$ coincides with T_M.

(b) Even if the introduction of weights like ρ improves the numerical stability of the mixed formulation (22), it seems quite impossible to consider values of T far from $L/|M|$: for instance, for $T = 2\sqrt{2}$ exhibited in [19] (see (4)), the norm $\|y(\cdot, T)\|_{H^{-1}(0,L)}$ is the uncontrolled situation, is for $\varepsilon = 10^{-2}$, about 3.33×10^{-17}. Consequently, when the double precision is used, we achieve "numerically" zero. Resolution of (22) would then lead to $v := 0$ on $(0, T)$! A possible way to avoid such pathologies is to preliminary consider a change of variables. We may write the solution y as follows, for any $\alpha, \gamma \in \mathbb{R}$,

$$y(x,t) = e^{\frac{M\alpha x}{2\epsilon}} e^{-\frac{\gamma M^2 t}{4\varepsilon}} z(x,t)$$

leading to

$$L_\varepsilon y := e^{\frac{M\alpha x}{2\epsilon}} e^{-\frac{\gamma M^2 t}{4\varepsilon}} \left(z_t - \varepsilon z_{xx} + M(1-\alpha) z_x - \frac{M^2}{4\varepsilon}(\gamma + \alpha^2 - 2\alpha) z \right).$$

Remark that $y(\cdot, T) = 0$ if and only if $z(\cdot, T) = 0$. Taking $1 - \alpha$ small and $\frac{M^2}{4\varepsilon}(\gamma + \alpha^2 - 2\alpha) \geq 0$ allows to reduce the dissipation of the solution at time T as $\varepsilon \to 0$ and therefore avoid the zero numeric effect. For instance, for $\alpha = \gamma = 1$, z solves $z_t - \varepsilon z_{xx} = 0$. Within this change of variable, the cost of control is

$$K^2(\varepsilon, T, M) = \sup_{z_0 \in L^2(0,L)} \frac{(\mathcal{A}_\varepsilon z_0, z_0)}{(e^{\frac{M\alpha x}{\varepsilon}} z_0, z_0)}$$

where $\mathcal{A}_\varepsilon$ is the control operator defined by $\mathcal{A}_\varepsilon : z_0 \to -w(\cdot, 0) \in L^2(0, L)$; here w solves the adjoint problem

$$\begin{cases} -w_t - \varepsilon w_{xx} - M(1-\alpha) w_x - \frac{M^2}{4\varepsilon}(\gamma + \alpha^2 - 2\alpha) w = 0 & \text{in} \quad Q_T, \\ w(0, \cdot) = w(L, \cdot) = 0 & \text{on} \quad (0, T), \\ w(\cdot, T) = w_T & \text{in} \quad (0, L), \end{cases}$$

with $w_T \in H_0^1(0, L)$ the minimizer of the functional

$$J^\star(w_T) := \frac{1}{2} \int_0^T \varepsilon^2 e^{\frac{\gamma M^2 t}{2\varepsilon}} w_x^2(0, t) dt + (z_0, w(\cdot, 0))_{L^2(0,L)}.$$

The corresponding control of minimal $L^2(e^{-\frac{\gamma M^2 t}{4\varepsilon}})$ norm for the variable z is given by $v_{\varepsilon,z} := \varepsilon e^{\frac{\gamma M^2 t}{2\varepsilon}} w_x(\cdot, t)$. The optimality conditions for $J^\star$ lead to a mixed formulation similar to (17). The introduction of appropriate parameters

α and γ allows to avoid the effect of the transport term; on the other hand, the change of variables make appear explicitly in the formulation exponential functions which may leads to numerical overflow for small values of ε.

(c) Another numerical strategy, employed in [23], is to use a spectral expansion of the adjoint solution φ of (6):

$$\varphi(x,t) = e^{-\frac{Mx}{2\varepsilon}} \sum_{k>0} \alpha_k e^{-\lambda_{\varepsilon,k}(T-t)} \sin(k\pi x), \quad \lambda_{\varepsilon,k} := \varepsilon k^2\pi^2 + \frac{M^2}{4\varepsilon}$$

with $\{\alpha_k\}_{k>0} \in L(\varepsilon, M, T)$ such that $\varphi(\cdot, 0)$ is in $L^2(0, L)$, equivalently

$$L(\varepsilon, M, T) := \Bigg\{ \{\alpha_p\}_{p>0} \in \mathbb{R}, \sum_{p,q\geq 0} \alpha_p\alpha_q e^{-(\lambda_{\varepsilon,k}+\lambda_{\varepsilon,p})T} \times \frac{32\varepsilon^3 M(p\pi)(q\pi)(1-e^{-\frac{M}{\varepsilon}}(-1)^{p+q})}{(a_{p,q}^2 - b_{p,q}^2)}) < \infty \Bigg\}$$

with $a_{p,q} := 4(M^2 + \varepsilon^2((p\pi)^2 + (q\pi)^2))$ and $b_{p,q} := 8\varepsilon^2(p\pi)(q\pi)$. The characterization (8) of the control with $v_\varepsilon = \varepsilon\varphi_x(0, \cdot)$ then rewrites as follows: find $\{\alpha_k\}_{k\geq 1} \in L(\varepsilon, M, T)$ such that

$$\varepsilon^2 \sum_{k,p\geq 1} \alpha_k \overline{\alpha}_p (k\pi)(p\pi) \frac{1-e^{-(\lambda_{\varepsilon,p}+\lambda_{\varepsilon,k})T}}{\lambda_{\varepsilon,p}+\lambda_{\varepsilon,k}} + \sum_{k\geq 1} \overline{\alpha}_k e^{-\lambda_{\varepsilon,k}T} \sum_{p\geq 1} \beta_p M_{p,k} = 0, \quad \forall \{\overline{\alpha_k}\}_{k\geq 1} \in L(\varepsilon, M, T), \tag{28}$$

with $y_0(x) := \sum_{p>0} \beta_p \sin(p\pi x)$ and $M_{p,q} := \int_0^1 e^{-\frac{Mx}{2\varepsilon}} \sin(p\pi x) \sin(q\pi x) dx$. The use of symbolic computations with large digit numbers may allow to solve (28) with robustness.

(d) At last, it seems interesting to perform as well an asymptotic analysis of the system of optimality (17) with respect to ε, in the spirit of [17]. This may allow to replace the direct resolution of (17) by the resolution of a sequel of simpler optimality systems independent of ε. This analysis is investigated in [1].

Eventually, we also mention that similar methods can be used to consider the case $M = 0$ in (3) in order to examine precisely the evolution of the cost of control for the heat equation when the controllability time T goes to zero. Precisely, the change of variable $\tilde{t} := \varepsilon t$ in (1) leads to the equation $\tilde{y}_{tt} - \tilde{y}_{xx} = 0$ over $(0, L) \times (0, \varepsilon T)$. This case, easier than the case considered in this work, is still open in the literature and is numerically discussed in [10].

Appendix

See Tables 14 and 15.

Table 14 Cost of control $K(\varepsilon, T, M)$ for $L = M = 1$ with respect to T and ε;—$h = 1/320$—$r = h^2$—$\beta = 10^{-16}$

ε	$T = 0.95$	$T = 0.99$	$T = 1$	$T = 1.05$
10^{-3}	237.877	30.4972	18.7555	2.2915
1.25×10^{-3}	190.574	29.7622	19.1953	2.8028
1.5×10^{-3}	159.813	29.0015	19.3883	3.2556
1.75×10^{-3}	138.166	28.2446	19.4234	3.6529
2×10^{-3}	122.044	27.4997	19.3540	4.0005
2.25×10^{-3}	109.519	26.7745	19.2093	4.3013
2.5×10^{-3}	99.476	26.0722	19.0163	4.5623
3×10^{-3}	84.250	24.7318	18.5275	4.9814
4×10^{-3}	64.648	22.3060	17.3600	5.5078
5×10^{-3}	52.289	20.1837	16.1269	5.7530
6×10^{-3}	43.650	18.3289	14.9392	5.8259
7×10^{-3}	37.213	16.6883	13.8166	5.7787
8×10^{-3}	32.198	15.2461	12.7839	5.6683
9×10^{-3}	28.210	13.9660	11.8380	5.5099
10^{-2}	24.934	12.8331	10.9763	5.3276
1.25×10^{-2}	18.898	10.5015	9.1493	4.8282
1.5×10^{-2}	14.810	8.7281	7.7087	4.3378
1.75×10^{-2}	11.913	7.3526	6.5694	3.8897
2×10^{-2}	9.784	6.2780	5.6566	3.4943
2.25×10^{-2}	8.176	5.4196	4.9210	3.1506
2.5×10^{-2}	6.937	4.7293	4.3237	2.8534
3×10^{-2}	5.180	3.7047	3.4240	2.3744
4×10^{-2}	3.264	2.4895	2.3297	1.7350
5×10^{-2}	2.294	1.8261	1.7304	1.3416
6×10^{-2}	1.736	1.4209	1.3522	1.0848
7×10^{-2}	1.376	1.1510	1.1030	0.8978
8×10^{-2}	1.113	0.9596	0.9223	0.7612
9×10^{-2}	0.0952	0.8130	0.7865	0.6554
10^{-1}	0.08175	0.7075	0.6808	0.5711

Table 15 Cost of control $K(\varepsilon, T, M)$ for $L = -M = 1$ with respect to T and ε; $h = 1/320$—$r = h^2$—$\beta = 10^{-16}$

ε	$T = 1$
10^{-3}	10,718.0955936799
1.25×10^{-3}	13,839.4039394749
1.5×10^{-3}	16,903.9918205099
1.75×10^{-3}	19,898.1360771887
2×10^{-3}	22,812.2634798022
2.25×10^{-3}	25,638.7601386909
2.5×10^{-3}	28,375.3693789053
2.75×10^{-3}	31,021.5479842987
3×10^{-3}	33,575.948263826
4×10^{-3}	42,871.1424334121
5×10^{-3}	50,751.4443114544
6×10^{-3}	57,316.7716579456
7×10^{-3}	62,692.7273334616
8×10^{-3}	66,997.3602057935
9×10^{-3}	70,350.3966144308
10^{-2}	72,862.0738060569
1.25×10^{-2}	76,089.8839137614
1.5×10^{-2}	75,988.4041456468
1.75×10^{-2}	73,579.1022138189
2×10^{-2}	69,647.3042543371
2.25×10^{-2}	64,735.7778969391
2.5×10^{-2}	59,254.0430977822
3×10^{-2}	47,994.1519570731
4×10^{-2}	27,872.8642664892
5×10^{-2}	13,312.4452504554
6×10^{-2}	5687.69600914237
7×10^{-2}	1864.72524997867
8×10^{-2}	648.702980070232
9×10^{-2}	264.559407164062
10^{-1}	123.306947646919

References

1. Amirat, Y., Munch, A.: On the controllability of an advection-diffusion equation with respect to the diffusion parameter: asymptotic analysis and numerical simulations. Acta Math. Appl. Sin. (to appear)
2. Boyer, F.: On the penalised HUM approach and its applications to the numerical approximation of null-controls for parabolic problems. In: CANUM 2012, 41e Congrès National d'Analyse Numérique. ESAIM Proceedings, vol. 41, pp. 15–58. EDP Science, Les Ulis (2013)
3. Brezzi, F., Fortin, M.: Mixed and Hybrid Finite Element Methods. Springer Series in Computational Mathematics, vol. 15. Springer, New York (1991)
4. Carthel, C., Glowinski, R., Lions, J.-L.: On exact and approximate boundary controllabilities for the heat equation: a numerical approach. J. Optim. Theory Appl. **82**, 429–484 (1994)
5. Chapelle, D., Bathe, K.-J.: The inf-sup test. Comput. Struct. **47**, 537–545 (1993)

6. Chatelin, F.: Eigenvalues of Matrices. Classics in Applied Mathematics, vol. 71. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2012). With exercises by Mario Ahués and the author, Translated with additional material by Walter Ledermann, Revised reprint of the 1993 edition [MR1232655]
7. Cîndea, N., Münch, A.: A mixed formulation for the direct approximation of the control of minimal L^2-norm for linear type wave equations. Calcolo **52**, 245–288 (2015)
8. Coron, J.-M.: Control and Nonlinearity. Mathematical Surveys and Monographs, vol. 136. American Mathematical Society, Providence (2007)
9. Coron, J.-M., Guerrero, S.: Singular optimal control: a linear 1-D parabolic-hyperbolic example. Asymptot. Anal. **44**, 237–257 (2005)
10. Duprez, M., Münch, A.: Numerical estimations of the cost of boundary controls for the one dimensional heat equation (in preparation)
11. Fernández-Cara, E., Münch, A.: Numerical exact controllability of the 1D heat equation: duality and Carleman weights. J. Optim. Theory Appl. **163**, 253–285 (2014)
12. Fernández-Cara, E., González-Burgos, M., de Teresa, L.: Boundary controllability of parabolic coupled equations. J. Funct. Anal. **259**, 1720–1758 (2010)
13. Fursikov, A.V., Imanuvilov, O.Y.: Controllability of Evolution Equations. Lecture Notes Series, vol. 34. Seoul National University Research Institute of Mathematics Global Analysis Research Center, Seoul (1996)
14. Glass, O.: A complex-analytic approach to the problem of uniform controllability of a transport equation in the vanishing viscosity limit. J. Funct. Anal. **258**, 852–868 (2010)
15. Labbé, S., Trélat, E.: Uniform controllability of semidiscrete approximations of parabolic control systems. Syst. Control Lett. **55**, 597–609 (2006)
16. Lebeau, G., Robbiano, L.: Contrôle exact de l'équation de la chaleur. Commun. Partial Differ. Equ. **20**, 335–356 (1995)
17. Lions, J.-L.: Perturbations singulières dans les problèmes aux limites et en contrôle optimal. Lecture Notes in Mathematics, vol. 323. Springer, Berlin (1973)
18. Lissy, P.: A link between the cost of fast controls for the 1-d heat equation and the uniform controllability of a 1-d transport-diffusion equation. C.R. Math. **350**, 591–595 (2012)
19. Lissy, P.: Explicit lower bounds for the cost of fast controls for some 1-D parabolic or dispersive equations, and a new lower bound concerning the uniform controllability of the 1-D transport-diffusion equation. J. Differ. Equ. **259**, 5331–5352 (2015)
20. Montaner, S., Munch, A.: Approximation of controls for the linear wave equation: a first order mixed formulation (submitted)
21. Münch, A., Souza, D.: A mixed formulation for the direct approximation of L^2-weighted controls for the linear heat equation. Adv. Comput. Math. **42**, 85–125 (2016)
22. Münch, A., Souza, D.A.: Inverse problems for linear parabolic equations using mixed formulations – Part 1: theoretical analysis. J. Inverse Ill-Posed Probl. **25**, 445–468 (2017)
23. Münch, A., Zuazua, E.: Numerical approximation of null controls for the heat equation: ill-posedness and remedies. Inverse Prob. **26**, 085018, 39 (2010)

Control of Random PDEs: An Overview

Francisco J. Marín, Jesús Martínez-Frutos, and Francisco Periago

Dedicated to Prof. Enrique Fernández-Cara on the occasion of his 60th birthday.

Abstract This work reviews theoretical and numerical concepts in the emergent field of optimal control of partial differential equations under uncertainty. The following topics are considered: uncertainty modelling in control problems using probabilistic tools, variational formulation of partial differential equations with random inputs, robust and risk averse formulations of optimal control problems, and numerical resolution methods. The exposition is focused on running the path starting from uncertainty modelling and ending in the practical implementation of numerical schemes for the numerical approximation of the considered problems. To this end, a selected number of illustrative examples is analysed.

Keywords Uncertainty quantification · Partial differential equations with random inputs · Stochastic expansion methods · Robust optimal control · Risk averse control

1 Introduction: Some Motivating Examples

Both the theory and numerical resolution of optimal control problems of systems governed by—*deterministic*—Partial Differential Equations (PDEs) are very well

F. J. Marín · F. Periago (✉)
Departamento de Matemática Aplicada y Estadística, Universidad Politécnica de Cartagena, Cartagena, Spain
e-mail: francisco.marin@upct.es; f.periago@upct.es

J. Martínez-Frutos
Departamento de Estructuras y Construcción, Universidad Politécnica de Cartagena, Cartagena, Spain
e-mail: jesus.martinez@upct.es

A. Doubova et al. (eds.), *Recent Advances in PDEs: Analysis, Numerics and Control*, SEMA SIMAI Springer Series 17,
https://doi.org/10.1007/978-3-319-97613-6_10

established and several textbooks that provide an introduction to the fundamental concepts of the mathematical theory are available in the literature (e.g. [16, 27]).

However, the topic of optimal control problems constrained by stochastic (or better named, *random*) PDEs is still in its infancy. Although it is difficult to fix the beginning of this subject, at least in what concerns the techniques and tools to be presented in the current work, the papers [3, 9, 12, 26] may be considered as pioneering.

This section is mainly devoted to introduce some motivating examples, which shall be analysed later on. Section 2 is concerned with existence of solutions for random PDEs and for some classes of optimal control problems constrained by random PDEs. Section 3 focuses on the numerical resolution of these optimal control problems. Some final remarks and challenging problems complete this work.

1.1 Uncertainty Is Almost Everywhere

Predictions obtained from mathematical models of physical, biological or economical systems always involve errors. For instance:

- *model errors,* which are due to simplifications of the mathematical model,
- *numerical errors,* which come from the numerical resolution method,
- *data errors,* which are due to a limited knowledge of the system's parameters, such as its geometry, initial and/or boundary conditions, external forces and material properties (diffusion coefficients, elasticity modulus, etc.).

Some of the above errors may be reduced (of course, not completely removed). This is the so-called *epistemic or systematic uncertainty.* However, there are other sources of randomness that are intrinsic to the system itself, and hence, cannot be reduced. Uncertainty Principle in Quantum Physics is a relevant example. This type of uncertainty is referred in the literature as to *aleatoric or statistical uncertainty.*

Consequently, if one aims at obtaining more reliable numerical predictions from mathematical models, then these should account for uncertainty. The question is:

1.2 How to Model Uncertainty in PDEs-Based Models?

The answer to this question depends on the a priori information available about the uncertain inputs. When statistical information is available, it is natural to model uncertainty by using probabilistic tools. On the contrary, in the absence of statistical information, non-probabilistic methods such as interval sets, convex modelling or fuzzy sets may be used. This work is focused on a probabilistic framework for uncertainty quantification and it is restricted to data error, i.e., to uncertainty in the input data of a PDEs-based model.

Before describing some illustrative examples in the framework of optimal control theory, let us introduce some notation. $D \subset \mathbb{R}^d$, $d = 1, 2$ or 3 in applications, denotes a bounded domain with smooth boundary ∂D. $(\Omega, \mathscr{F}, \mathbb{P})$ stands for a complete probability space. The sample space Ω is the set of all possible outcomes (e.g., all possible measures of a physical parameter), $\mathscr{F}$ is the σ—algebra of events (hence, subsets of Ω to which probabilities may be assigned), and $\mathbb{P} : \mathscr{F} \to [0, 1]$ is a probability measure.

Example 1 (Laplace-Poisson Equation) Consider the following control system for the Laplace-Poisson equation

$$\begin{cases} -\text{div}\,(a\nabla y) = 1_{\mathscr{O}} u, & \text{in } D \\ +\text{boundary conditions}, & \text{on } \partial D, \end{cases} \tag{1}$$

where y is the state variable and u is the control function, which acts on the spatial region $\mathscr{O} \subset D$. As usual, $1_{\mathscr{O}}$ stands for the characteristic function of $\mathscr{O}$, and the gradient operator ∇ involves derivatives only w.r.t. the spatial variable $x \in D$.

For the case of a steady-state, single-phase groundwater flow, a is the hydraulic conductivity field, u is the source term (due to recharge, pumping or injecting), and y is the so-called hydraulic head. In most aquifers, a is highly variable and never perfectly known. More precisely, experimental data reported in [29] show that, at each location $x \in D$, the hydraulic conductivity $a\,(x)$ follows a log-normal distribution. Since only limited measurements are available at a few locations, there is uncertainty about the conductivity values at points between sparse measurements. Hence, it is natural to consider the coefficient a as a random function $a = a\,(x, \omega)$, where ω denotes an elementary random event. As a consequence, the hydraulic head, solution to (1), becomes a random space function $y = y\,(x, \omega)$.

A typical optimal control problem that arises in this context aims to find the control $u = u\,(x)$, whose associated state $y = y\,(u)$ is the best approximation (in a least-squares sense) to a desired target $y_d\,(x)$ in D. Hence, the cost functional

$$\frac{1}{2}\int_D |y\,(x, \omega) - y_d\,(x)\,|^2\,dx + \frac{\gamma}{2}\int_{\mathscr{O}} u^2\,(x)\,dx \tag{2}$$

is introduced. The second term in (2) is a measure of the energy cost needed to implement the control u. It is observed that (2) depends on each realization $\omega \in \Omega$. Hence, a control u that minimizes (2) also depends on ω . Of course, one is typically interested in a control u, *independent of* ω, which minimizes (in some sense) the distance between $y\,(x, \omega)$ and $y_d\,(x)$. At first glance, it is natural to consider the average, w.r.t. ω, of the functional (2). The problem is then formulated as:

$$\begin{cases} \text{Minimize in } u : J\,(u) = \frac{1}{2}\int_{\Omega}\int_D |y\,(x, \omega) - y_d\,(x)\,|^2\,dx\,d\mathbb{P}\,(\omega) + \frac{\gamma}{2}\int_{\mathscr{O}} u^2\,(x)\,dx \\ \text{subject to} \\ \qquad -\text{div}\,(a\,(x, \omega)\,\nabla y\,(x, \omega)) = 1_{\mathscr{O}} u\,(x)\,, \quad \text{in} \quad D \times \Omega \\ \qquad +\text{boundary conditions}, \qquad\qquad\qquad \text{on} \quad \partial D \times \Omega. \end{cases} \tag{3}$$

Example 2 (Optimal Control for the Heat Equation) Consider the following control system for the transient heat equation:

$$\begin{cases} y' - \operatorname{div}(a\nabla y) = 0, & \text{in } (0,T) \times D \times \Omega \\ a\nabla y \cdot n = 0, & \text{on } (0,T) \times \partial D_0 \times \Omega \\ a\nabla y \cdot n = \alpha\,(u-y)\,, & \text{on } (0,T) \times \partial D_1 \times \Omega, \\ y\,(0) = y^0, & \text{in } D \times \Omega \end{cases} \tag{4}$$

where the boundary of the spatial domain $D \subset \mathbb{R}^d$ is decomposed into two disjoint parts $\partial D = \partial D_0 \cup \partial D_1$, and n is the unit outward normal vector to ∂D. Here and throughout the text, y' is the partial derivative of y w.r.t. the time variable $t \in (0,T)$, and the divergence (div) operator involves only derivatives w.r.t. the spatial variable $x \in D$.

As reported in [5], in addition to randomness in the thermal conductivity coefficient $a = a\,(x,\omega)$, the initial temperature y^0 and the convective heat transfer coefficient α are very difficult to measure in practise. Hence, both are affected by a certain amount of uncertainty, i.e., $y^0 = y^0\,(x,\omega)$ and $\alpha = \alpha\,(x,\omega)$. In real applications α may also depend on the time variable $t \in [0,T]$ but, for the sake of simplicity, here it is assumed to be stationary.

Suppose that a desired temperature $y_d\,(x)$ is given and that it is aimed to choose a control u, which must be applied through ∂D_1, such that its associated $y\,(u)$ be closer as possible to y_d at time T. Similarly to problem (3), one may consider the averaged distance between $y\,(T,x,\omega)$ and $y_d\,(x)$ as the cost functional to be minimized. Another possibility is to minimize the distance between the mean temperature of the body occupying the region D and y_d. However, if only the mean of $y\,(T)$ is considered, then there is no control on the dispersion of $y\,(T)$. Consequently, if the dispersion of $y\,(T)$ is large, then minimizing the expectation of $y\,(T)$ is useless because the probability of $y\,(T,\omega)$ of being close to its average is small. It is then convenient to minimize not only the expectation but also a measure of dispersion such as the variance. Thus, the optimal control problem reads as follows:

$$\begin{cases} \text{Minimize in } u : J\,(u) = \int_D \left(\int_\Omega y(T)\, d\mathbb{P}(\omega) - y_d \right)^2 dx + \frac{\gamma}{2} \int_D \operatorname{Var}\,(y(T))\; dx \\ \text{subject to} \\ \qquad\qquad y = y\,(u) \quad \text{solves (4)}, \end{cases} \tag{5}$$

where $\gamma \geq 0$ is a weighting parameter, and

$$\operatorname{Var}\,(y(T,x)) = \int_\Omega y^2(T,x,\omega)\, dP(\omega) - \left(\int_\Omega y(T,x,\omega)\, dP(\omega) \right)^2$$

is the variance of $y\,(T,x,\cdot)$.

Example 3 (Piezoelectric Control of the Beam Equation) The small random vibrations of a thin, uniform, hinged beam of length L, driven by a piezoelectric actuator located along the random interval $(\xi(\omega), \eta(\omega)) \subset (0, L)$ may be described by the system:

$$\begin{cases} y'' + [Dy_{xx}]_{xx} = v\left[\delta_{\eta(\omega)} - \delta_{\xi(\omega)}\right]_x, & \text{in } (0,T) \times (0,L) \times \Omega \\ y(0) = y_{xx}(0) = y(L) = y_{xx}(L) = 0, & \text{on } (0,T) \times \Omega \\ y(0) = y^0, \quad y'(0) = y^1, & \text{in } (0,L) \times \Omega, \end{cases} \tag{6}$$

where the beam flexural stiffness $D = EI$, is assumed to depend on both $x \in (0, L)$ and $\omega \in \Omega$. As usual, E denotes Young's modulus and I is the area moment of inertia of the beam's cross-section. In system (6), $\xi(\omega)$ and $\eta(\omega)$ stand for the extremes of the actuator. The dependence of these two points on a random event ω indicates that there is some uncertainty in the location of the actuator. $\delta_{x_0} = \delta_{x_0}(x)$ is the Dirac mass at the spatial point $x_0 \in (0, L)$. The random output $y(t, x, \omega)$ represents vertical displacement at time t. Since physical controller devices are affected by uncertainty, it is realistic to decompose the control variable into an unknown deterministic and a known stochastic components. Moreover, it is reasonable to consider the stochastic part to be modulated by the deterministic one. Thus, the function $v = v(t, \omega)$, which appears in (6), takes the form

$$v(t, \omega) = u(t)\left(1 + \hat{u}(\omega)\right), \tag{7}$$

where $u : (0, T) \to \mathbb{R}$ is the (*unknown*) deterministic control and $\hat{u}$ is a (*known*) zero-mean random variable which accounts for uncertainty in the controller device. See Fig. 1 for the problem configuration.

For each random event $\omega \in \Omega$ and each control $u \in L^2(0, T)$, a measure of the energy of the beam vibrations at a control time $T > 0$ is expressed by

$$J(u, \omega) = \frac{1}{2}\left(\|y(T, \omega)\|_H^2 + \|y'(T, \omega)\|_{V^\star}^2\right). \tag{8}$$

where $H = L^2(0, L)$, $V = H^2(0, L) \cap H_0^1(0, L)$ and $V^\star$ is the topological dual of V.

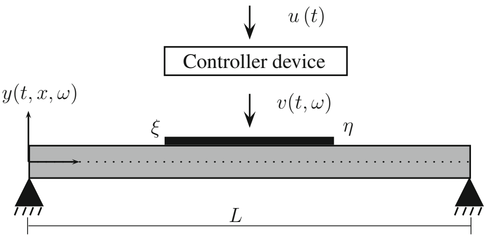


Fig. 1 Example 3: problem configuration

Assume that we are interested in computing a control u for rare occurrences of the random inputs (which therefore have a very little impact on the mean and variance of J), but which could have catastrophic consequences. For that purpose, the following risk averse cost functional is considered:

$$J_{\varepsilon}\left(u\right)=\mathbb{P}\left\{\omega\in\Omega:J\left(u,\omega\right)>\varepsilon\right\}, \tag{9}$$

where $\varepsilon > 0$ is a prescribed threshold parameter. The corresponding control problem is formulated as

$$\begin{cases}\text{Minimize in } u: J_{\varepsilon}\left(u\right)=\mathbb{P}\left\{\omega\in\Omega:J\left(u,\omega\right)>\varepsilon\right\}\\ \text{subject to}\\ \qquad\qquad y=y\left(u\right)\quad\text{solves }\ (6).\end{cases} \tag{10}$$

2 Existence of Solutions for Random PDEs and for Its Associated Optimal Control Problems

This section briefly reviews existence theory for solutions of random PDEs and its related robust and risk averse optimal control problems.

2.1 Variational Formulation of Random PDEs

In this section, we study separately the cases of elliptic and evolutionary PDEs.

The Elliptic Case Consider the elliptic problem

$$\begin{cases}-\operatorname{div}\left(a(x,\omega)\nabla y(x,\omega)\right)=f\left(x,\omega\right) & \text{in } \ D\times\Omega\\ y(x,\omega)=0 & \text{on } \ \partial D\times\Omega.\end{cases} \tag{11}$$

The following hypotheses on the input data are assumed to hold:

(A1) $a=a(x,\omega)\in L^{\infty}_{\mathbb{P}}\left(\Omega;L^{\infty}(D)\right)$ and there exist $a_{\min}, a_{\max}>0$ such that

$$0<a_{\min}\leq a(x,\omega)\leq a_{\max}<\infty\quad\text{a. e. } x\in D \text{ and } \mathbb{P}-\text{ a. s. }\omega\in\Omega,$$

(A2) $f\in L^{2}_{\mathbb{P}}\left(\Omega;L^{2}(D)\right)$.

The variational formulation of problem (11) is: find $y\in L^{2}_{\mathbb{P}}\left(\Omega;H^{1}_{0}(D)\right)$ such that

$$\int_{\Omega}\int_{D}a\nabla y\nabla v\,dxd\mathbb{P}(\omega)=\int_{\Omega}\int_{D}fv\,dxd\mathbb{P}(\omega)\quad\forall v\in L^{2}_{\mathbb{P}}\left(\Omega;H^{1}_{0}(D)\right). \tag{12}$$

A straightforward application of Lax-Milgram's lemma ensures the existence and uniqueness of solution to (12). Moreover,

$$\|y\|_{L^2_{\mathbb{P}}(\Omega;H^1_0(D))} \leq C\,(D, a_{\min})\,\|f\|_{L^2_{\mathbb{P}}(\Omega;L^2(D))}. \tag{13}$$

The Case of Evolution PDEs As an illustration of the case of time-dependent PDEs, the following parabolic problem is considered:

$$\begin{cases} y'(t,x,\omega) - \operatorname{div}\left(a(x,\omega)\nabla y(t,x,\omega)\right) = f\,(t,x,\omega)\,, & \text{in} \quad (0,T)\times D\times\Omega \\ y(t,x,\omega) = 0, & \text{on} \quad (0,T)\times\partial D\times\Omega \\ y\,(0,x,\omega) = y^0\,(x,\omega)\,, & \text{in} \quad D\times\Omega, \end{cases} \tag{14}$$

where a satisfies (A1), $y^0 \in L^2_{\mathbb{P}}\left(\Omega; L^2(D)\right)$, and $f \in L^2\left(0,T; L^2_{\mathbb{P}}\left(\Omega; L^2(D)\right)\right)$.

Consider the spaces $V = H^1_0\,(D)$, $V_{\mathbb{P}} = L^2_{\mathbb{P}}\,(\Omega; V)$, $V^\star_{\mathbb{P}} = L^2_{\mathbb{P}}\,(\Omega; V^\star)$ and

$$W_{\mathbb{P}}\,(0,T) = \left\{y \in L^2\,(0,T; V_{\mathbb{P}}) : y' \in L^2\left(0,T; V^\star_{\mathbb{P}}\right)\right\}.$$

Due to the tensor product structure of Bochner spaces, $W_{\mathbb{P}}\,(0,T)$ may be identified with the spaces

$$\begin{aligned} W_{\mathbb{P}}\,(0,T) &\simeq L^2_{\mathbb{P}}\,(\Omega)\otimes\left[\left(L^2\,(0,T)\otimes V\right)\cap\left(H^1\,(0,T)\otimes V^\star\right)\right] \\ &\simeq L^2_{\mathbb{P}}\left(\Omega; L^2\,(0,T;V)\cap H^1\,(0,T;V^\star)\right). \end{aligned}$$

The continuous injection $W_{\mathbb{P}}\,(0,T) \hookrightarrow C\left(0,T; L^2_{\mathbb{P}}\left(\Omega; L^2(D)\right)\right)$ also holds [8, 17, 28].

The variational formulation of problem (14) intends to find $y \in W_{\mathbb{P}}\,(0,T)$ such that

$$\int_0^T \left(y', v\right)_{V^\star_{\mathbb{P}}, V_{\mathbb{P}}}\,dt + \int_0^T\int_\Omega\int_D a\nabla y\nabla v\,dx d\mathbb{P}(\omega)dt = \int_0^T\int_\Omega\int_D f v\,dx d\mathbb{P}(\omega)dt,$$

for all $v \in L^2\,(0,T;V_{\mathbb{P}})$, and, in addition, $y\,(0) = y^0$.

Existence and uniqueness of solutions follows from the Galerkin method. Indeed, assumption (A1) ensures that the bilinear form

$$V_{\mathbb{P}}\times V_{\mathbb{P}} \ni (\varphi,\psi) \mapsto \int_\Omega\int_D a(x,\omega)\nabla\varphi\nabla\psi\,dx\mathbb{P}(\omega),$$

is continuous and $V_{\mathbb{P}}$-elliptic. The space $(\Omega, \mathscr{F}, \mathbb{P})$ is assumed to be separable so that $L^2_{\mathbb{P}}\,(\Omega)$ so is [10]. Hence, as $V_{\mathbb{P}}$ is isomorphic to $L^2_{\mathbb{P}}\,(\Omega)\otimes V$ [15, Chap. 1], given the orthonormal bases $\{\phi_i\}_{i\geq 1}$ and $\left\{\psi_j\right\}_{j\geq 1}$ of $L^2_{\mathbb{P}}\,(\Omega)$ and V, respectively, the set $\left\{\phi_i\otimes\psi_j\right\}_{i,j\geq 1}$ is an orthonormal basis of $L^2_{\mathbb{P}}\,(\Omega)\otimes V$. As in the deterministic case (see [17, Th. 8.1 and 8.2] or [28]), it is proved that there exists a unique weak solution of (14). Moreover,

$$\|y\|_{W_{\mathbb{P}}(0,T)} \leq C \left(\|y^0\|_{L^2_{\mathbb{P}}(\Omega;L^2(D))} + \|f\|_{L^2(0,T;L^2_{\mathbb{P}}(\Omega;L^2(D)))} \right).$$

Before proceeding, the following remark on assumptions (A1) and (A2) is in order.

Remark 1 In most applications, Gaussian fields are the model of choice to represent uncertain parameters which show a spatial correlation. The reason for this choice is Central Limit Theorem, which supports the so-called *additive hypothesis of small errors: if random variation is the sum of many small errors, then a normal distribution is the result*. Moreover, for numerical simulation purposes, Gaussian fields are typically approximated by truncated Karhunen-Loève (KL) expansions [18] of the form

$$f\left(x,\omega\right) \approx \sum_{n=1}^{N} \sqrt{\lambda_n} b_n\left(x\right) \xi_n\left(\omega\right),$$

where ξ_n are independent and identically distributed standard Gaussian variables. This choice, which is appropriate for forcing terms and initial conditions, is no longer suitable for random fields which are positive in nature, as the coefficient $a\left(x,\omega\right)$ which appears in the principal part of elliptic differential operators. In fact, as indicated in Example 1, experimental data reveal that $a\left(x,\omega\right)$ is often log-normal distributed. The reason is again Central Limit Theorem, but in its multiplicative version, which supports the *multiplicative hypothesis of small errors: if random variation is the product of many small errors, then a log-normal distribution is the result*. Note that the coefficient a typically collects the product of several random parameters (see, e.g., Example 3). Hence, $a\left(x,\omega\right)$ is approximated as

$$a\left(x,\omega\right) \approx e^{\mu(x)+\sigma(x)\sum_{n=1}^{N} \sqrt{\lambda_n} b_n(x) \hat{\xi}_n(\omega)},$$

where $\mu\left(x\right)$ and $\sigma\left(x\right)$ are, respectively, scale and shape parameters. However, if $\hat{\xi}_n\left(\omega\right)$ are Gaussian variables, this representation is not completely satisfactory from a mathematical point of view because, in such a case, a may approximate to zero or to $+\infty$. To overcome this technical difficulty, $\hat{\xi}_n$ is a truncated Gaussian with probability density function

$$\rho\left(s\right) = \begin{cases} \frac{e^{-s^2/2}}{\Phi(d)-\Phi(-d)}, & -d \leq s \leq d \\ 0, & \text{otherwise.} \end{cases}$$

Here $\Phi\left(s\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{s} e^{-r^2/2}\, dr$ is the cumulative density function of the normal distribution.

As a conclusion, assumptions (A1) and (A2) are suitable both from a mathematical point of view and for applications.

2.2 Existence of Solutions for Robust and Risk Averse Control Problems

As representative of robust and risk averse control problems, the two following control problems are considered:

$$(P)\begin{cases}\text{Minimize in } u: J(u)=\int_{\Omega}\int_{D}|y(x,\omega)-y_d(x)|^2\,dx\,d\mathbb{P}(\omega)+\frac{\gamma}{2}\int_{\mathcal{O}}u^2(x)\,dx\\ \text{subject to}\\ \qquad y=y(u)\quad\text{solves }(11),\text{ with } f=1_{\mathcal{O}}u,\quad u\in\mathcal{U}_{ad}\end{cases}$$

and

$$(P_\varepsilon)\begin{cases}\text{Minimize in } u: J_\varepsilon(u)=\mathbb{P}\left\{\omega\in\Omega: I(u,\omega):=\frac{1}{2}\|y(\omega)-y_d\|^2_{L^2(D)}\geq\varepsilon\right\}\\ \text{subject to}\\ \qquad y=y(u)\quad\text{solves }(11),\text{ with } f=1_{\mathcal{O}}u,\quad u\in\mathcal{U}_{ad},\end{cases}$$

where $y_d\in L^2(D)$, $\gamma\geq 0$, $\varepsilon>0$ and $\mathcal{U}_{ad}=\left\{u\in L^2(D): |u(x)|\leq M \text{ a.e. } x\in D\right\}$, with $M>0$.

Existence of solutions for the robust optimal control problem (P) and the risk averse problem (P_ε) is established next.

Theorem 1 *Problems (P) and (P_ε) have, at least, one solution.*

Proof Let us start with problem (P). Because of estimate (13), the control-to-state linear operator $S: L^2(D)\to L^2_{\mathbb{P}}\left(\Omega; L^2(D)\right)$, $\quad u\mapsto y(u)$ is continuous. Moreover, the set of admissible controls $\mathcal{U}_{ad}$ is bounded, closed and convex. The existence of a solution for (P) follows from [27, Th. 2.14]. The solution is unique if $\gamma>0$ since, in that case, J is strictly convex.

As for problem (P_ε), let u_n be a minimizing sequence. Up to a subsequence, still labelled by n, $u_n\rightharpoonup u$ weakly in $L^2(D)$. For a fixed $\omega\in\Omega$, since $I(\cdot,\omega)$ is continuous and convex, $I(u,\omega)\leq\liminf_{n\to\infty}I(u_n,\omega)$. The cost functional J_ε may be expressed in integral form as $J_\varepsilon(u)=\int_\Omega H(I(u,\omega)-\varepsilon)\,d\mathbb{P}(\omega)$ (H being the Heaviside function). Hence, by Fatou's lemma,

$$\int_\Omega\liminf_{n\to\infty}H(I(u_n,\omega)-\varepsilon)\,d\mathbb{P}(\omega)\leq\liminf_{n\to\infty}\int_\Omega H(I(u_n,\omega)-\varepsilon)\,d\mathbb{P}(\omega).$$

By the lower-semicontinuity of the Heaviside function,

$$H(I(u,\omega)-\varepsilon)\leq\liminf_{n\to\infty}H(I(u_n,\omega)-\varepsilon)).$$

Thus,

$$\begin{aligned}\int_\Omega H\left(I(u)-\varepsilon\right)\,d\mathbb{P}(\omega) &\le \int_\Omega \liminf_{n\to\infty} H\left(I(u_n)-\varepsilon\right)\,d\mathbb{P}(\omega)\\ &\le \liminf_{n\to\infty}\int_\Omega H\left(I(u_n)-\varepsilon\right)\,d\mathbb{P}(\Omega).\end{aligned}$$

□

Remark 2 Using similar ideas as in the preceding theorem, existence of solutions for the evolution control problems considered in Examples 2 and 3 above may be proved [19, 20, 22]. Existence of solution for robust optimal control problems, similar to problem (P), may be also obtained by using a stochastic saddle point formulation [3].

3 Numerical Resolution of Robust and Risk Averse Optimal Control Problems

This section reviews (some of) the most popular numerical methods that are currently being used to solve robust and risk averse control problems. To illustrate these methods, the chosen model is the Bernoulli-Euler beam system

$$\begin{cases} y''(t,x,\omega)+\left[D(x,\omega)\,y_{xx}(t,x,\omega)\right]_{xx}=f(t,x,\omega)\,, & \text{in } (0,T)\times(0,L)\times\Omega\\ y(t,0,\omega)=y_{xx}(t,0,\omega)=y(t,L,\omega)=y_{xx}(t,L,\omega)=0, & \text{on } (0,T)\times\Omega\\ y(0,x,\omega)=y^0(x,\omega)\,,\quad y'(0,x,\omega)=y^1(x,\omega)\,, & \text{in } (0,L)\times\Omega. \end{cases} \tag{15}$$

The cases of robust and risk averse controls will be analysed separately.

3.1 Numerical Approximation of Robust Optimal Control Problems

Assume that the control function takes the form $f(t,x,\omega)=1_{\mathscr{O}}u(t,x)$, where $\mathscr{O}$ is a measurable subset of the physical domain $(0,L)$, and the cost functional is

$$\begin{aligned}J(u)=&\tfrac{\alpha_1}{2}\int_0^L\left(\int_\Omega y(T,x,\omega)\,d\mathbb{P}(\omega)\right)^2dx+\tfrac{\alpha_2}{2}\int_0^L\left(\int_\Omega y'(T,x,\omega)\,d\mathbb{P}(\omega)\right)^2dx\\ &+\tfrac{\beta_1}{2}\int_0^L\int_\Omega y^2(T,x,\omega)\,d\mathbb{P}(\omega)dx+\tfrac{\beta_2}{2}\int_0^L\int_\Omega\left(y'\right)^2(T,x,\omega)\,d\mathbb{P}(\omega)dx\\ &+\tfrac{\gamma}{2}\int_0^T\int_{\mathscr{O}}u^2(t,x)\,dxdt,\end{aligned} \tag{16}$$

with $\alpha_1,\alpha_2>0$, $\beta_1,\beta_2\ge 0$, and y a solution to (15).

Both, methods based on first order optimality conditions [3, 26] and gradient-based minimization algorithms [19] may be used. The latter type is considered next. Following the same lines as in the deterministic case, the reduced gradient of the cost functional (16), denoted by $J'(u)$, may be computed by using the formal Lagrangian method, which leads to

$$J'(u) = \gamma u(t,x) - \int_\Omega p(t,x,\omega)\, d\mathbb{P}(\omega), \quad (t,x) \in (0,T)\times \mathscr{O},$$

where p solves the backward in time adjoint problem

$$\begin{cases} p''(t,x,\omega) + [D(x,\omega)p_{xx}(t,x,\omega)]_{xx} = 0, & \text{in } (0,T)\times(0,L)\times\Omega \\ p(t,0,\omega) = p_{xx}(t,0,\omega) = p(t,L,\omega) = p_{xx}(t,L,\omega) = 0, & \text{on } (0,T)\times\Omega \\ p(T,x,\omega) = -\beta_2 y'(T,x,\omega) - \alpha_2 \int_\Omega y'(T,x,\omega)\, d\mathbb{P}(\omega), & \text{in } (0,L)\times\Omega \\ p'(T,x,\omega) = \beta_1 y(T,x,\omega) + \alpha_1 \int_\Omega y(T,x,\omega)\, d\mathbb{P}(\omega) & \text{in } (0,L)\times\Omega. \end{cases} \tag{17}$$

It is observed that the main difficulty arises in the numerical approximation of statistics (in this case, first and second order moments) associated to solutions of the direct and adjoint systems (15) and (17). Apart from the classical Monte Carlo method, stochastic finite element methods [7] and, more recently, stochastic collocation methods [1], reduced basis methods [11] and combination of them [2] are being developed to efficiently solve these problems. All these methods are based on the following

Finite dimensional noise assumption

The random inputs of the PDE depend on a finite number of uncorrelated real random variables

$$\xi(\omega) = (\xi_1(\omega), \cdots, \xi_N(\omega)).$$

Note that this assumption is in agreement with the representation of the random inputs of a PDE as indicated in Remark 1. This finite dimensional noise assumption lets transform the random PDE (15) into a deterministic PDE with a finite dimensional parameter.

From Random PDEs to Parametric Deterministic PDEs According to Dood-Dynkin's lemma [18, lemmas 4.46 and 9.40], the solution $y(t,x,\omega) = y(t,x,\xi(\omega))$ of (15) is measurable w.r.t. the σ-algebra generated by ξ. Denoting by $\Gamma_n = \xi_n(\Omega)$ the image space of ξ_n, and by $\Gamma = \prod_{n=1}^N \Gamma_n \subset \mathbb{R}^N$, the abstract probability space $(\Omega, \mathscr{F}, \mathbb{P})$ is mapped to $(\Gamma, \mathscr{B}(\Gamma), \rho(z)\, dz)$, where $\mathscr{B}(\Gamma)$ is the σ-algebra of Borel sets on Γ and $\rho : \Gamma \to \mathbb{R}$ is the joint probability density function of ξ, which is assumed to exist. As usual, dz is the Lebesgue measure.

Thus, the random PDE (15) is transformed into the deterministic PDE with an N-dimensional parameter $z \in \Gamma$

$$\begin{cases} y''(t,x,z) + \left[D(x,z)(x,z)\, y_{xx}(t,x,z) \right]_{xx} = f(t,x,z), & \text{in } (0,T) \times (0,L) \times \Gamma \\ y(t,0,z) = y_{xx}(t,0,z) = y(t,L,z) = y_{xx}(t,L,z) = 0, & \text{on } (0,T) \times \Gamma \\ y(0,x,z) = y^0(x,z), \quad y'(0,x,z) = y^1(x,z), & \text{in } (0,L) \times \Gamma. \end{cases} \tag{18}$$

The same argument applies for the adjoint system (21).

The discretization process to approximate the first and second order moments of the solutions to (18) involves: (a) time discretization, e.g. by using Newmark scheme, finite element approximation in space, e.g., with cubic Hermite finite elements, and discretization in the random domain Γ. As indicated above, several methods may be used for discretization in the random domain (see [6] for a recent review). Since, in this case, the problem is smooth w.r.t. the random parameter, an adaptive, anisotropic, sparse grid collocation method is well suited (see [4] for a comparison study with reduced basis methods). Roughly speaking, the situation is that if the number of terms N in a KL representation of a random field, as in Remark 1, is relatively large, then full tensor product rules lead to an unaffordable computational problem. This is the well-known *curse of dimensionality* phenomenon. Then, a sparse grid is used instead. In addition, random variables ξ_n in KL expansion do not weight equally because they are multiplied by a decreasing sequence of positive numbers $\sqrt{\lambda_n}$. Anisotropic grids are able to keep accuracy with a lower computational cost [24]. Finally, the level of the chosen quadrature rule (e.g., Smolyak's rule) in the probability space should be adaptively chosen as to comply with a prescribed accuracy level. A suitable criterion for the problem under consideration is that the relative error in computing first and second order moments of solutions to (18) in two consecutive quadrature levels be smaller than a prescribed tolerance. We refer the reader to [19, 22, 24] for a detailed description of this algorithm.

Numerical simulation results are presented in Fig. 2 for the following data: $L = 1$, $\mathscr{O} = (0.2, 0.8)$, $T = 0.5$, $\Gamma = [-3, 3]^6$, $y^0(x) = \sin(\pi x)$, $y^1(x) = 0$,

$$D(x,z) = e^{-0.04 + 0.283 \sum_{n=1}^{6} \sqrt{\lambda_n} b_n(x) z_n}, \quad z = (z_1, \cdots, z_6) \in \Gamma,$$

where $\{\lambda_n, b_n(x)\}$ are the eigenpairs associated to a Gaussian random field with isotropic exponential covariance function $C(x_1, x_2) = e^{-|x_1 - x_2|/0.4}$. A non-nested quadrature rule, whose collocation nodes are determined by the roots of Hermite polynomials, is used. The nodes and weights for the anisotropic sparse grid are computed adaptively as described above. We refer to [22] for more details.

It is observed that the optimal control obtained minimizing only the mean of the state variable (Fig. 2b) is quite similar to the one obtained in the deterministic case (Fig. 2a). Here, deterministic problem means that the random input datum $D(x,z)$ is replaced by its mean value and the cost functional to be minimized is the one

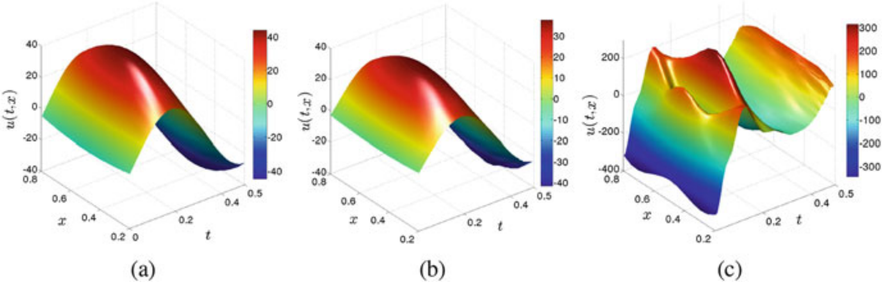


Fig. 2 $\gamma = 10^{-6}$. Optimal controls for: (**a**) deterministic problem, (**b**) $\alpha_1 = \alpha_2 = 1$, $\beta_1 = \beta_2 = 0$ and (**c**) $\alpha_1 = \alpha_2 = 1$, $\beta_1 = \beta_2 = 1$

given by (8), without the ω dependence. However, the optimal control including the second raw moment in the cost functional (Fig. 2c) shows a very different behavior. Both the similarity between Fig. 2a and b, and the differences between these two figures and Fig. 2c may be explained by looking at the first and second order statistical moments of the underlying uncontrolled solution (i.e., the solution of (18) with $f = 0$), and to the uncontrolled solution of its associated deterministic problem. For more details on this passage we refer to [19].

3.2 Numerical Approximation of Risk Averse Control Problems

The numerical resolution of the risk averse control problem presented in Example 3 is discussed next. The main difficulty is that the cost functional (9) is discontinuous. To overcome this difficulty, $J_\varepsilon(u)$, as given by (9), is approximated by

$$J_\varepsilon^\alpha(u) = \int_\Omega \left(1 + e^{-\frac{2}{\alpha}(J(u,\omega)-\varepsilon)}\right)^{-1} d\mathbb{P}(\omega), \tag{19}$$

where $0 < \alpha < 1$. Then, the gradient of this approximated cost functional may be computed. Precisely,

$$\left(J_\varepsilon^\alpha\right)'(u) = \int_\Omega \left[p_x(t, \eta(\omega), \omega) - p_x(t, \xi(\omega), \omega)\right] d\mathbb{P}(\omega), \tag{20}$$

where the $p = p(t, x, \omega)$ solves the adjoint system

$$\left\{\begin{array}{ll} p''(t,x,\omega) + [D(x,\omega)p_{xx}(t,x,\omega)]_{xx} = 0, & \text{in } (0,T)\times(0,L)\times\Omega \\ p(t,0,\omega) = p_{xx}(t,0,\omega) = p(t,L,\omega) = p_{xx}(t,L,\omega) = 0, & \text{on } (0,T)\times\Omega \\ p(T,x,\omega) = -C(\omega)y'(T,x,\omega), & \text{in } (0,L)\times\Omega \\ p'(T,x,\omega) = C(\omega)y(T,x,\omega), & \text{in } (0,L)\times\Omega, \end{array}\right. \tag{21}$$

with

$$C(\omega) = \frac{2}{\alpha} e^{-\frac{2}{\alpha}(J(u,\omega)-\varepsilon)} \left(1 + e^{-\frac{2}{\alpha}(J(u,\omega)-\varepsilon)}\right)^{-2}. \tag{22}$$

However, for α small, (19) and (22) are still, numerically, discontinuous. As a consequence, stochastic collocation methods should not be used, for numerical approximation in the random domain, because if a few collocation points are located in the unknown discontinuity, then the approximation could be very poor. On the contrary, the use of a direct Monte Carlo (MC) method requires the numerical resolution of (6) and (21) at a large number of sampling points $\omega_k \in \Omega$ and at each step of the descent method. This makes MC method unaffordable from a computational point of view, at least in optimization. A possible remedy is to use Monte Carlo in combination with a polynomial chaos (PC) approach for uncertainty propagation. More precisely, $y(t, x, \omega)$ and $p(t, x, \omega)$ are approximated with a PC expansion. Then, to compute the cost functional (19), the random variable (22) and the gradient (20), MC is applied to those approximations. Next, more details are provided.

For the sake of clarity, since the main goal of this subsection is to illustrate a way to deal with random discontinuous control problems, let us assume that uncertainty in problem (10) only appears in the location of the piezoelectric actuator, i.e., $\xi(\omega) = \xi_0 + X(\omega)$ and $\eta(\omega) = \eta_0 + X(\omega)$ with $\xi_0, \eta_0 \in (0, L)$ and $X : \Omega \to \mathbb{R}$ a random variable. The more general case in which uncertainty is modelled by truncated KL expansions of random fields is treated in a similar manner.

PC Expansion for Uncertainty Propagation Let $\{\psi_r(z)\}_{r=1}^{\infty}$ be an orthonormal basis of $L^2_\rho(\Gamma)$ composed of a suitable class of orthonormal polynomials. As usual, $\Gamma = X(\Omega)$ and ρ is the probability density function of the random variable X. For a positive integer $\ell \in \mathbb{N}_+$, consider the finite dimensional space

$$\mathbb{P}_\ell(\Gamma) = \text{ span } \{\psi_r(z), \quad 0 \le r \le \ell\}.$$

An approximated solution $y_\ell(t, x, z) \in L^2((0, T); V) \otimes \mathbb{P}_\ell(\Gamma)$ of (6) is expressed in the form

$$y_\ell(t, x, z) = \sum_{0 \le r \le \ell} \hat{y}_r(t, x)\, \psi_r(z), \quad \hat{y}_r \in L^2((0, T); V), \tag{23}$$

where due to the orthonormality of $\{\psi_r(z)\}_{0 \le r \le \ell}$,

$$\hat{y}_r(t, x) = \int_\Gamma y_\ell(t, x, z)\, \psi_r(z)\, \rho(z)\, dz. \tag{24}$$

This latter integral may be numerically approximated by using a stochastic collocation method. This requires the knowledge of $y_\ell(t, x, z_k)$, where $z_k \in \Gamma$ are sampling nodes, which are computed by using the Newmark method for time discretization and cubic Hermite finite elements in the spatial domain. A very important advantage of this approach is that the automatic parallelization of the collocation method enables to reduce the computational cost in a very significant way.

It remains to analyse how to choose ℓ. Since the goal is to minimize the cost functional (19), ℓ is adaptively chosen as to comply with a prescribed accuracy level δ for that functional. This is done as follows:

(i) *Initialization:* (a) Compute an approximated solution $y_{\mathrm{MC}}(t,x,z)$ of (6) by applying Monte Carlo method directly on (6). This Monte Carlo solution (which plays the role of *exact solution*) is then used to obtain an approximation $J^{\alpha}_{\varepsilon,\mathrm{MC}}(u)$ of $J^{\alpha}_{\varepsilon}(u)$, also by using MC method. The control $u(t)$, which is used here, is the optimal control of the deterministic problem, where the random input parameters of (6) are replaced by its nominal (or mean) value and the considered cost criterion is

$$J_d(u) = \frac{1}{2}\left(\|y(T)\|^2_H + \|y'(T)\|^2_{V^\star}\right).$$

Note that these computations, although computationally expensive, are performed just only once at the beginning of the optimization algorithm.
(b) Initialize $\ell = 1$.

(ii) *Construction:* For the selected ℓ, compute $y_\ell(t,x,z)$ by using (23) and (24). With this approximated solution, compute an approximation $J^{\alpha}_{\varepsilon,\ell}(u)$ of $J^{\alpha}_{\varepsilon}(u)$ by using Monte Carlo sampling, where samples are applied to the PC solution (23).

(iii) *Verification:* Check if the stopping criterion

$$\frac{|J^{\alpha}_{\varepsilon,\ell}(u) - J^{\alpha}_{\varepsilon,\mathrm{MC}}(u)|}{J^{\alpha}_{\varepsilon,\mathrm{MC}}(u)} \le \delta. \tag{25}$$

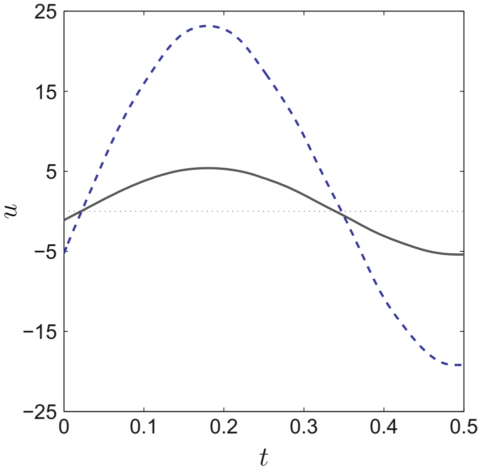
is satisfied. If (25) does not hold, then increase ℓ in one unit, i.e., $\ell = \ell + 1$, and go to (ii).

A proof of the convergence of the proposed adaptive algorithm is an open problem, but numerical simulation results, which shall be presented here below, suggest that convergence holds.

MC Sampling for the Numerical Approximation of the Cost Functional At each iteration of the descent algorithm, the cost functional $J^{\alpha}_{\varepsilon}(u)$ and the gradient (20) are numerically approximated using Monte Carlo integration. This is not a daunting task because MC samples are applied to the approximated solution $y_\ell(t,x,z)$, for which the explicit representation (23) is available.

As indicated in [13] in a similar context, the choice of the parameter α used in (19) is a very delicate issue because if a few number of sampling points are in the unknown transition regions, then (20) may be equal to zero (hence providing

Fig. 3 Deterministic control (continuous line) and risk averse control (dashed line)



no descent direction for the gradient method). To overcome this difficulty, α is adaptively chosen at each iterate (see [13, 20] for details).

The methods described in this subsection have been implemented in the following numerical experiment, where the goal is to analyze numerically the influence of small errors in the location of the piezoelectric actuator. The spatial domain is $(0, 1)$ and the control time is $T = 0.5$. It is assumed that the initial conditions $\left(y^0(x), y^1(x)\right) = (\sin(\pi x), 0)$, the flexural rigidity $D = 1$ and the control function $v(t, \omega) = u(t)$ are unaffected by uncertainty. Uncertainty in the location of the piezoelectric actuator is modelled as described above, with $X = \mathscr{U}(-0.1, 0.1)$, a uniformly distributed random variable, which corresponds to 10% of error in the location of the piezoelectric actuator. Finally, $\varepsilon = \alpha = 0.1$ and $\ell = 3$. Figure 3 plots *preliminary results* for the deterministic control and the risk averse control obtained after convergence of the algorithm. The values of the cost functional (19) for these two controls are: J_ε^α (deterministic control) $= 1$ and J_ε^α (risk averse control) $= 0.1218$, which indicates that the deterministic control has a very poor performance.

4 Further Comments and Challenging Problems

Although presented in the framework of optimal control for random PDEs, the methods described in this work may be applied to other types of optimization problems, e.g., for shape and topology optimization under uncertainty [21, 23].

The two following computational challenges arise in this context: (1) *curse of dimensionality*, meaning that when the number of random variables, which appear in the uncertain parameters, is large, the number of collocation nodes grows exponentially so that the problem is computationally unaffordable. (2) this issue is exacerbated when the solution of the underlying PDE is expensive (e.g., multi-physics and/or multiscale problems) or when the stochastic PDE is involved within optimization processes. A number of remedies are being proposed to overcome these two difficulties. For instance, Analysis of Variance (ANOVA) based techniques, High-dimensional Model representation (HDMR) or a combination of them, aim at detecting the most influential random variables (or the interaction between some of them) in the mathematical model (see [14] and the references therein). To alleviate the computational burden of multiphysics or multiscale problems, model order reduction methods (such as Reduced Basis methods or Proper-Orthogonal Decomposition (POD) methods) have been receiving an increasing interest in the last decades (see, e.g., [2, 25]).

Despite these theoretical and numerical developments, the (very important in applications) topic of control of random PDEs may be still considered to be in its infancy and further research is needed to better understand and more efficiently solve this type of problems.

Acknowledgements Work supported by projects MTM2013-47053 from Ministerio de Economía y Competitividad (Spain), the AEI/FEDER and UE under the contract DPI2016-77538-R, and 19274/PI/14 from Fundación Séneca (Agencia de Ciencia y Tecnología de la Región de Murcia, Spain). The first author was also supported by a Ph.D. grant from Fundación Séneca.

References

1. Babuška, I., Nobile, F., Tempone, R.: A stochastic collocation method for elliptic partial differential equations with random input data. SIAM Rev. **52**(2), 317–355 (2010)
2. Chen, P., Quarteroni, A.: A new algorithm for high-dimensional uncertainty quantification based on dimension-adaptive sparse grid approximation and reduced basis methods. J. Comput. Phys. **298**, 176–193 (2015)
3. Chen, P., Quarteroni, A., Rozza, G.: Stochastic optimal Robin boundary control problems of advection-dominated elliptic equations. SIAM J. Numer. Anal. **51**(5), 2700–2722 (2013)
4. Chen, P., Quarteroni, A., Rozza, G.: Comparison between reduced basis and stochastic collocation methods for elliptic problems. J. Sci. Comput. **59**(1), 187–216 (2014)
5. Chiba, R.: Stochastic analysis of heat conduction and thermal stresses in solids: a review. In: Kazi, S.N. (ed.) Heat Transfer Phenomena and Applications. InTech, London (2012)
6. Cohen, A., DeVore, R.: Approximation of high-dimensional parametric PDEs. Acta Numer. **24**, 1–159 (2015)
7. Ghanem, R.G., Spanos, P.D.: Stochastic Finite Elements. A Spectral Approach. Springer, Berlin (1981)
8. Gittelson, C.J., Andreev, R., Schwab, C.: Optimality of adaptive Galerkin methods for random parabolic partial differential equations. J. Comput. Appl. Math. **263**, 189–201 (2014)
9. Gunzburger, M.D., Lee, H.-C., Lee, J.: Error estimates of stochastic optimal Neumann boundary control problems. SIAM J. Numer. Anal. **49**(4), 1532–1552 (2011)
10. Halmos, P.: Measure Theory. Graduate Texts in Mathematics, vol. 18. Springer, New York (1970)
11. Hesthaven, J.S., Rozza, G., Stamm, B.: Certified Reduced Basis Methods for Parametrized Partial Differential Equations. Springer Briefs in Mathematics. BCAM, Bizkaia (2016)
12. Hou, L.S., Lee, J., Manouzi, H.: Finite element approximations of stochastic optimal control problems constrained by stochastic elliptic PDEs. J. Math. Anal. Appl. **384**(1), 87–103 (2011)
13. Keshavarzzadeh, V., Meidani, H., Tortorelli, A.: Gradient based design optimization under uncertainty via stochastic expansion methods. Comput. Methods Appl. Mech. Eng. **3016**, 47–76 (2016)

14. Labovsky, A., Gunzburger, M.: An efficient and accurate method for the identification of the most influential random parameters appearing in the input data for PDEs. SIAM/ASA J. Uncertain. Quantif. **2**(1), 82–105 (2014)
15. Light, W.A., Cheney, E.W.: Approximation Theory in Tensor Product Spaces. Lecture Notes in Mathematics, vol. 1169. Springer, New York (1985)
16. Lions, J.L.: Optimal Control of Systems Governed by Partial Differential Equations. Springer, Berlin (1971)
17. Lions, J.L., Magenes, E.: Non-Homogeneous Boundary Value Problems and Applications, vol. I. Springer, New York (1972)
18. Lord, G.J., Powell, C.E., Shardlow, T.: An Introduction to Computational Stochastic PDEs. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge (2014)
19. Marín, F.J., Martínez-Frutos, J., Periago, F.: Robust averaged control of vibrations for the Bernoulli-Euler beam equation. J. Optim. Theory Appl. **174**(2), 428–454 (2017)
20. Marín, F.J., Martínez-Frutos, J., Periago, F.: A polynomial chaos-based approach to risk-averse piezoelectric control of random vibrations of beams. Int. J. Numer. Methods Eng. 1–18 (2018). https://doi.org/10.1002/nme.5823
21. Martínez-Frutos, J., Kessler, M., Periago, F.: Robust optimal shape design for an elliptic PDE with uncertainty in its input data. ESAIM Control Optim. Calc. Var. **21**(4), 901–923 (2015)
22. Martínez-Frutos, J., Kessler, M., Münch, A., Periago, F.: Robust optimal Robin boundary control for the transient heat equation with random input data. Int. J. Numer. Methods Eng. **108**(2), 116–135 (2016)
23. Martínez-Frutos, J., Herrero-Pérez, D., Kessler, M., Periago, F.: Robust shape optimization of continuous structures via the level set method. Comput. Methods Appl. Mech. Eng. **305**, 271–291 (2016)
24. Nobile, F., Tempone, R., Webster, C.G.: An anisotropic sparse grid stochastic collocation method for elliptic partial differential equations with random input data. SIAM J. Numer. Anal. **46**(5), 2411–2442 (2008)
25. Nouy, A.: Recent developments in spectral stochastic methods for the numerical solution of stochastic partial differential equations. Arch. Comput. Meth. Eng. **16**(3), 251–285 (2009)
26. Rosseel, E., Wells, G.N.: Optimal control with stochastic PDE constrains and uncertain controls. Comput. Methods Appl. Mech. Eng. **213/216**, 152–167 (2012)
27. Tröltzsch, F.: Optimal Control of Partial Differential Equations: Theory, Methods and Applications. Graduate Studies in Mathematics, vol. 112. AMS, Providence (2010)
28. Wloka, J.: Partial Differential Equations. Cambridge University Press, Cambridge (1987)
29. Zhang, D.: Stochastic Methods for Flow in Porous Media: Coping with Uncertainties. Academic Press, San Diego (2002)

The Dubovitskii and Milyutin Formalism Applied to an Optimal Control Problem in a Solidification Model

Aníbal Coronel, Francisco Guillén-González, Francisco Marques-Lopes, and Marko Rojas-Medar

Dedicated to Prof. Enrique Fernández-Cara on the occasion of his 60th birthday.

Abstract In this paper we study an optimal control problem in a physical system governed by a solidification model. The solidification system is given by a nonlinear parabolic PDE system of two equations for the unknowns the (reduced) temperature and a phase field function, with a temperature source term. The optimal control problem is defined via the source term as the control function and the objective functional given by the comparison in L^{2n}-norms of the real state with a given target state and the cost of the control. The main results of the paper are the existence of a global optimal solution via a minimizing sequence, and the first-order necessary conditions for local optimal solutions, by means of the application of the Dubovitskii and Milyutin formalism.

Keywords Optimal control · Parabolic systems · Diffuse-interface phase field · Solidification · Optimality conditions

A. Coronel
GMA, Dpto. de Ciencias Básicas, Fac. de Ciencias, Universidad del Bío-Bío, Chillán, Chile
e-mail: acoronel@ubiobio.cl

F. Guillén-González
Dpto. EDAN and IMUS, Fac. de Matemáticas, Universidad de Sevilla, Sevilla, Spain
e-mail: guillen@us.es

F. Marques-Lopes
Dpto. de Matemática, UFPA, Belém, PA, Brazil
e-mail: fpmolopes@ufpa.br

M. Rojas-Medar (✉)
Instituto de Alta Investigación, Universidad de Tarapacá, Arica, Chile

A. Doubova et al. (eds.), *Recent Advances in PDEs: Analysis, Numerics and Control*, SEMA SIMAI Springer Series 17,
https://doi.org/10.1007/978-3-319-97613-6_11

1 Introduction

In the recent decades, has emerged an increasing progress on the research area of optimization of physical phenomena and industrial processes modeled by partial differential equations. For instance, fluid flows among others are applied in aviation and space technology or melting-solidification processes, see [16] and references therein. In particular, in this paper, we are concerned with the melting-solidification phenomena, which is used in several industrial processes, for instance in manufacturing of crystals, metal casting, electronic printed circuit production or oxygen free copper production. The melting-solidification process is interesting since permits the production of high quality solids by using appropriately the phase change of the material produced by decreasing the temperature or making compression of the material.

The first attempts for modeling solidification were proposed by Lamé and Clayperon [12] and by Stefan [15], nowadays known as the classical Stefan Problem. Later, the diffuse-interface phase field problems modeling the solidification were introduced by Fix [6] and Caginalp [4] and after them several authors have applied this approach and make some improvements, see for instance [2, 13].

In this paper we consider the model introduced in [4]. Indeed, in order to precise this mathematical model, we consider a domain $\Omega \subset \mathbb{R}^3$ which contains the material in the liquid phase and the solid formation occurrences by cooling in a control subdomain $\omega_c \subset \Omega$. If the domain is modeled by the region Ω with boundary $\partial\Omega$, during a finite time interval $(0, T)$, then by considering the assumptions and simplifications given on [4], the solidification process occurring in the time-space cylindrical domain

$$Q := \Omega \times (0, T) \quad \text{with} \quad \Sigma := \partial\Omega \times (0, T)$$

can be governed by the following PDE system

$$\partial_t u + \ell \partial_t \phi - \Delta u = f\chi_{\omega_c}, \quad \text{in } Q, \tag{1a}$$

$$\partial_t \phi - \Delta\phi + F(\phi) = u, \quad \text{in } Q, \tag{1b}$$

$$u(0) = u_0, \quad \phi(0) = \phi_0, \quad \text{in } \Omega, \tag{1c}$$

$$\frac{\partial u}{\partial \mathbf{n}} = 0, \quad \frac{\partial \phi}{\partial \mathbf{n}} = 0, \quad \text{on } \Sigma, \tag{1d}$$

where $\ell > 0$ is the latent heat constant, $\mathbf{n} = \mathbf{n}(x)$ is the outward unit normal vector to Ω at the point $x \in \partial\Omega$ and the nonlinear function $F : \mathbb{R} \to \mathbb{R}$ is defined as

$$F(\phi) = \phi^3 - a\phi - b\phi^2 \text{ with } a, b : \Omega \to \mathbb{R}_+ \text{ some given functions.} \tag{1e}$$

Here $u = u(x,t)$ models the (reduced) temperature of the material, $\phi = \phi(x,t)$ is the phase-field function used to identify the level of crystallization and $f = f(x,t)$ is a temperature source term. Finally, (u_0, ϕ_0) is the initial data.

On the other hand, various control and inverse problems related with the solidification have been proposed and studied. Firstly, Hoffman and Jiang [9] obtained some results for an optimal control problem with respect to the source function f as control and an appropriate cost functional. Later on, Wang and Wang in [17] considered the time optimal control of the solidification system (1) obtaining some local exact controllability results, the existence of optimal controls and the associated first-order necessary optimality conditions. For the controllability of the solidification system (1) we refer also to [1, 5]. More recently, Gnanavel et al. in [8] have investigated an inverse problem of simultaneous reconstruction of two coefficients in the one-dimensional version of the solidification system (1) considering overdetermined final time data for the reduced temperature and the phase field function.

Let us consider that ω_c and ω_d are two open subsets of Ω (the control and observability domains, respectively). In this paper, we focus on the analysis of a distributed optimal control problem defined by assuming that the source term f is the partially distributed control acting in the temperature equation (1) in $Q_c = \omega_c \times (0,T)$ which wants to furnish states (u,ϕ) near of a given target states (u_d, ϕ_d) in $Q_d = \omega_d \times (0,T)$. Indeed, in order to precise the definition of this problem we introduce some notations. Let us consider the following notation for some appropriate Lebesgue and Sobolev spaces

$$\begin{aligned}
&W^{2,1}_{q,n}(Q) := \left\{ u \in W^{2,1}_q(Q);\ \frac{\partial u}{\partial \mathbf{n}} = 0 \ \text{ on } \Sigma \right\},\\
&E_u := W^{2,1}_{q,n}(Q) \cap L^{2k}(Q_d), \quad E_\phi := W^{2,1}_{\overline{q},n}(Q) \cap L^{2m}(Q_d),\\
&E_f := L^{2s}(Q_c),\\
&E := E_u \times E_\phi \times E_f,
\end{aligned} \tag{2}$$

where $k, m, s \in \mathbb{N}$, $q = 2s$ and $\overline{q} = \overline{q}(s)$ is given in (14). In fact, if $s = 1$ then $\overline{q} = \overline{2} = 10$ and if $s \geq 2$ then $\overline{q} = +\infty$. Moreover, we introduce the functional $J : E \to \mathbb{R}$ as

$$\begin{aligned}
J(u,\phi,f) = \frac{\alpha}{2}\int_0^T\int_{\omega_d} |u - u_d|^{2k} dxdt + \frac{\beta}{2}\int_0^T\int_{\omega_d} |\phi - \phi_d|^{2m} dxdt\\
+\frac{\mu}{2}\int_0^T\int_{\omega_c} |f|^{2s} dxdt,
\end{aligned} \tag{3}$$

where $\mu > 0, \alpha > 0, \beta > 0$ are constants, $u_d, \phi_d \in L^\infty(Q_d)$ are given functions, $f \in E_f = L^{2s}(Q_c)$ is the control which acts on the distributed region ω_c and the state $(u,\phi) \in E_u \times E_\phi$ is the (strong) solution of (1) associated to f.

Thus, the **optimal control problem** is defined as follows:

$$\left.\begin{array}{l}\text{Given the desired states}(u_d, \phi_d)\text{, find the control} f \in E_f \\ \text{and the state variables } (u, \phi) \in E_u \times E_\phi \text{ such that the functional} \\ \text{defined on (3) and subject to (1) is minimized,}\end{array}\right\}$$

or equivalently as the generic optimization problem

$$\min_{(u,\phi,f)\in\mathscr{U}_{ad}} J(u, \phi, f) \tag{4}$$

where the admissible set is

$$\mathscr{U}_{ad} = \{(u, \phi, f) \in E \; ; \; M(u, \phi, f) = 0\}$$

which states for the equality constraint with the operator

$$M : E \to \overline{E} := L^q(Q) \times L^{\overline{q}}(Q) \times W^2_{\infty,n}(\Omega) \times W^2_{\infty,n}(\Omega)$$

where

$$W^2_{\infty,n}(\Omega) := \left\{ u \in W^2_\infty(\Omega); \; \frac{\partial u}{\partial \mathbf{n}} = 0 \; \text{ on } \partial\Omega \right\}$$

and M is defined by:

$$M(u, \phi, f) = (\psi_1, \psi_2, \psi_3, \psi_4)$$

if and only if $(u, \phi, f) \in E$ and satisfies

$$u_t + \ell\phi_t - \Delta u - f\chi_{\omega_c} = \psi_1, \qquad \text{in } Q, \tag{5a}$$

$$\phi_t - \Delta\phi - F(\phi) - u = \psi_2, \qquad \text{in } Q, \tag{5b}$$

$$u(0) - u_0 = \psi_3, \; \phi(0) - \phi_0 = \psi_4, \qquad \text{in } \Omega, \tag{5c}$$

$$\frac{\partial u}{\partial \mathbf{n}} = \frac{\partial \phi}{\partial \mathbf{n}} = 0, \qquad \text{on } \Sigma. \tag{5d}$$

Note that, the functional given in (3) consists of two parts. First a comparison of a real state (u, ϕ) and a given ideal state (u_d, ϕ_d) on the observability region ω_d which is given by the first and second integrals in (3). Second, the cost of the control f given by the third integral in (3).

Definition 1 $(u, \phi, f) \in \mathscr{U}_{ad}$ is called a local optimal solution of (4), if there exists $\varepsilon > 0$ such that for all $(\bar{u}, \bar{\phi}, \bar{f}) \in \mathscr{U}_{ad}$ satisfying the inequality

$$\|\bar{u} - u\|_{E_u} + \|\bar{\phi} - \phi\|_{E_\phi} + \|\bar{f} - f\|_{E_f} \leq \varepsilon,$$

we have that $J(u, \phi, f) \leq J(\bar{u}, \bar{\phi}, \bar{f})$.

Moreover, $(u, \phi, f) \in \mathscr{U}_{ad}$ is called a global optimal solution of (4) if $J(u, \phi, f) \leq J(\bar{u}, \bar{\phi}, \bar{f})$ for all $(\bar{u}, \bar{\phi}, \bar{f}) \in \mathscr{U}_{ad}$.

The main results of the paper are given in the following theorems:

Theorem 1 (Existence of Global Optimal Solution) *Let $\Omega \subset \mathbb{R}^3$ be a bounded domain with C^2 boundary. Consider $F(\phi)$ given in (1e) with $a, b \in L^\infty(\Omega)$, $u_0, \phi_0 \in W^2_{\infty,n}(\Omega)$ and $u_d, \phi_d \in L^\infty(Q_d)$. Then the control problem (4) has a global optimal solution $(u, \phi, f) \in \mathscr{U}_{ad}$ in the sense of Definition 1.*

Theorem 2 (Optimal Control Depends on the Adjoint Problem) *Assume*

$$k \in \begin{cases} \{1, 2, 3, 4, 5\}, & \textit{if } s = 1, \\ \mathbb{N}, & \textit{if } s \geq 2. \end{cases} \tag{6}$$

If $(u, \phi, f) \in \mathscr{U}_{ad}$ is a local optimal solution of problem (4), and there exists $(v, \psi) \in W^{2,1}_{2,n}(Q) \times W^{2,1}_{2,n}(Q)$ a solution of the adjoint system

$$-\partial_t v - \Delta v - \psi = g_1(u), \qquad \textit{in } Q, \tag{7a}$$

$$-\partial_t \psi - \ell \partial_t v - \Delta \psi - F'(\phi)\psi = g_2(\phi), \qquad \textit{in } Q, \tag{7b}$$

$$v(T) = \psi(T) = 0, \qquad \textit{in } \Omega, \tag{7c}$$

$$\frac{\partial v}{\partial \mathbf{n}} = \frac{\partial \psi}{\partial \mathbf{n}} = 0, \qquad \textit{on } \Sigma, \tag{7d}$$

with

$$g_1(u) = \alpha k (u - u_d)^{2k-1} \chi_{\omega_d} (= J'_u), \quad g_2(\phi) = \beta m (\phi - \phi_d)^{2m-1} \chi_{\omega_d} (= J'_\phi), \tag{8}$$

then, the control $f = f(v)$ is given by

$$f = \left(-\frac{1}{s\mu} v\right)^{1/(2s-1)} \chi_{\omega_c}. \tag{9}$$

Notice that hypothesis (6) will used to obtain that $W^{2,1}_{q=2s}(Q) \subset L^{2k}(Q)$, and then $E_u = W^{2,1}_{q,n}(Q)$, see Lemma 4 below. On the other hand, $W^{2,1}_{\overline{q}}(Q) \subset L^\infty(Q) \subset L^{2m}(Q)$ for any $m \in \mathbb{N}$, and then $E_\phi = W^{2,1}_{\overline{q},n}(Q)$.

Theorem 3 (Existence, Uniqueness and Continuous Dependence of the Adjoint Problem) *Assume*

$$k \in \begin{cases} \{1, 2, 3\}, & \text{if } s = 1, \\ \mathbb{N}, & \text{if } s \geq 2. \end{cases} \tag{10}$$

For any $(u, \phi) \in E_u \times E_\phi$, there exist a unique pair of functions $(v, \psi) \in W^{2,1}_{2,n}(Q) \times W^{2,1}_{2,n}(Q)$ satisfying the adjoint system (7). Moreover, the following estimate holds

$$\|v, \psi\|_{W^{2,1}_2(Q)} \leq C \tag{11}$$

with $C > 0$ depending on $\|\phi\|_{L^\infty(Q)}$, $\|g_1(u), g_2(\phi)\|_{L^2(Q)}$, $\|a, b\|_{L^\infty(\Omega)}$, T and ℓ.

Notice that hypothesis (10) is more restricted than (6) and is used to obtain that $g_1(u) \in L^2(Q)$. In fact, if $s = 1$ i.e. $q = 2$, then $u \in E_u \subset L^{10}(Q)$ (see Lemma 1 below), hence $g_1(u) \in L^p(Q)$ for $p = 10$ if $k = 1$, $p = 10/3$ if $k = 2$, and $p = 2$ if $k = 3$. On the other hand, in any case $\phi \in E_\phi \subset L^\infty(Q)$, hence $g_2(\phi) \in L^\infty(Q)$.

As a result of these three theorems, we arrive at the following

Corollary 1 (Optimality System for Local Optimal Solution) *Under hypotheses of Theorem 3, if $(u, \phi, f) \in \mathscr{U}_{ad}$ is a local optimal solution of problem (4), then there exist a unique pair of functions $(v, \psi) \in W^{2,1}_{2,n}(Q) \times W^{2,1}_{2,n}(Q)$ such that (u, ϕ, v, ψ) solves the following coupled system:*

$$\begin{aligned} \partial_t u + \ell \partial_t \phi - \Delta u &= \left(-\tfrac{1}{s\mu} v\right)^{1/(2s-1)} \chi_{\omega_c}, \quad \text{in } Q, \\ \partial_t \phi - \Delta \phi + F(\phi) &= u, \quad \text{in } Q, \\ -\partial_t v - \Delta v - \psi &= g_1(u), \quad \text{in } Q, \\ -\partial_t \psi - \ell \partial_t v - \Delta \psi - F'(\phi)\psi &= g_2(\phi), \quad \text{in } Q, \\ u(0) = u_0, \ \ \phi(0) = \phi_0, \ \ v(T) = \psi(T) &= 0, \quad \text{in } \Omega, \\ \frac{\partial u}{\partial \mathbf{n}} = \frac{\partial \phi}{\partial \mathbf{n}} = \frac{\partial v}{\partial \mathbf{n}} = \frac{\partial \psi}{\partial \mathbf{n}} &= 0, \quad \text{on } \Sigma, \end{aligned} \tag{12}$$

with $g_1(u)$, $g_2(\phi)$ are given in (8).

The rest of the paper is organized as follows. In Sect. 2 we recall some preliminary concepts and results, and a reformulation of the optimal control problem in a generic form and the specific calculus of cones is given in Sect. 3. In Sects. 4, 5 and 6 we prove the Theorems 1, 2 and 3, respectively.

2 Preliminaries

2.1 Embeddings in Time-Spatial Dependent Sobolev Spaces

Lemma 1 (Continuous Embeddings, [11, Lemma 3.3, pp. 80]) *Let Ω be a bounded domain of $\mathbb{R}^3$ with boundary $\partial\Omega$ sufficiently smooth (with the cone property). Then, the embedding*

$$W_p^{2,1}(Q) \subset L^{p_*}(Q)$$

is continuous and there exist a constant $C > 0$ depending only on p and Ω such that $\|u\|_{L^{p_*}(Q)} \leq C\,\|u\|_{W_p^{2,1}(Q)}$, with $p_* = p_*(p)$ defined as follows

$$p_* = \begin{cases} \infty, & \text{if } p > 5/2, \\ \text{any real} \geq 1, & \text{if } p = 5/2, \\ \left(\dfrac{1}{p} - \dfrac{2}{5}\right)^{-1}, & \text{if } p < 5/2. \end{cases} \tag{13}$$

Lemma 2 (Compact Embeddings, [14]) *Let X, B and Y be Banach spaces such that $X \to B \to Y$ are continuous embeddings and $X \to B$ is compact. Then, the following embeddings are compacts:*

$$L^q(0,T;X) \cap \left\{\phi;\ \frac{\partial\phi}{\partial t} \in L^1(0,T;Y)\right\} \to L^q(0,T;B) \quad \text{with} \quad 1 \leq q \leq \infty, \quad \text{and}$$

$$L^\infty(0,T;X) \cap \left\{\phi;\ \frac{\partial\phi}{\partial t} \in L^r(0,T;Y)\right\} \to C([0,T];B) \quad \text{with} \quad 1 < r \leq \infty.$$

2.2 Results of the Solidification Model (1)

The mathematical analysis of well-posedness for system (1), was introduced in [9], where the authors established the following existence and stability results:

Theorem 4 *Assume that $\Omega, a, b, u_0, \phi_0$ satisfy the hypothesis of Theorem 1. For any $f \in L^q(Q_c)$ with $q \geq 2$, there exist a unique solution $(u,\phi) \in W_q^{2,1}(Q) \times W_{\overline{q}}^{2,1}(Q)$ of the solidification system (1) with*

$$\overline{q} = \begin{cases} \left(\dfrac{1}{q} - \dfrac{2}{5}\right)^{-1}, & \text{if } q \in [2, 5/2), \\ \text{any real} \geq 1, & \text{if } q \in [5/2, \infty). \end{cases} \tag{14}$$

Moreover, the estimate

$$\|u\|_{W_q^{2,1}(Q)} + \|\phi\|_{W_{\overline{q}}^{2,1}(Q)} \leq C\big(\|u_0\|_{W_\infty^{2,1}(Q)} + \|\phi_0\|_{W_\infty^{2,1}(Q)} + \|f\|_{L^q(Q_c)} \big), \tag{15}$$

holds with $C > 0$ depending only on $\|a\|_{L^\infty(Q)}$, $\|b\|_{L^\infty(Q)}$, T and ℓ.

Note that relation (14) becomes from (13)

Theorem 5 *Assume that $\Omega, a, b, u_0, \phi_0$ satisfy the hypotheses of Theorem 1. For each $i \in \{1, 2\}$ consider any $f_i \in L^q(Q_c)$ with $q \geq 2$ and denote by $(u_i, \phi_i) \in W_q^{2,1}(Q) \times W_{\overline{q}}^{2,1}(Q)$ the corresponding solutions of the solidification system (1) with $f = f_i$. Then, the following estimate holds*

$$\|u_1 - u_2\|_{W_q^{2,1}(Q)} + \|\phi_1 - \phi_2\|_{W_{\overline{q}}^{2,1}(Q)} \leq C\|f_1 - f_2\|_{L^q(Q_c)}, \tag{16}$$

where C is a constant depending on $\|u_i\|_{W_q^{2,1}(Q)}$, $\|\phi_i\|_{W_{\overline{q}}^{2,1}(Q)}$, $i \in \{1, 2\}$, and $\overline{q}$ is given in (14). In particular, fixed f the solution (u, ϕ) of the solidification system (1) is unique.

The proof of Theorems 4 and 5, in a broad sense, are based on the Leray-Schauder fixed point theorem using the following global in time a priori estimate (deduced testing (1a) by u and (1b) by $\ell\, \partial_t\phi$ hence terms $\int_0^T \int_\Omega \partial_t\phi\, u\, dxdt$ cancel)

$$u \in L^\infty(0, T; L^2(\Omega)) \cap L^2(0, T; H^1(\Omega)),$$

$$\phi \in L^\infty(0, T; H^1(\Omega)), \quad \partial_t\phi \in L^2(0, T; L^2(\Omega)),$$

and standard bootstrap technique jointly to the L^p-regularity for second-order parabolic problems (see for instance [11, Theorem 9.1, pp. 341]), firstly for the ϕ-problem (1b) (looking at the cubic term ϕ^3 of $F(\phi)$ as a dissipative term with respect to the L^p-estimates) and afterwards for the u-problem (1a), one can arrive at $(u, \phi) \in W_{q,n}^{2,1}(Q) \times W_{\overline{q},n}^{2,1}(Q)$.

2.3 The Dubovitskii and Milyutin Formalism

In order to introduce the main concepts and results related with the Dubovitskii and Milyutin formalism we consider the optimization problem

$$\left.\begin{array}{l} \min\limits_{x\in\mathscr{Q}} J(x), \quad \mathscr{Q} = \bigcap\limits_{i=1}^{n+1} \mathscr{Q}_i, \\ \text{int}(\mathscr{Q}_i) \neq \emptyset,\ i = 1, \ldots, n, \quad \text{(inequality restrictions)} \\ \mathscr{Q}_{n+1} = \{x \in X \ :\ M(x) = 0\}, \quad \text{(equality restriction)} \end{array}\right\} \tag{17}$$

where $J : X \to \mathbb{R}$ is a functional and $M : X \to Y$ an operator, with X and Y Banach spaces.

For more details consult the book of Girsanov [7] (see also the Kotarski's work [10]).

Definition 2 Let X be a Banach space and $J : X \to \mathbb{R}$ a functional. The vector $h \in X$ is called a **descent direction** of the functional J at the point $x_0 \in X$ if there is a neighborhood U of h and a strictly positive number $\alpha = \alpha(J, x_0, h) > 0$ such that for all $\varepsilon \in (0, \varepsilon_0)$ and any $\overline{h} \in U$, $J(x_0 + \varepsilon\overline{h}) \leq J(x_0) - \varepsilon\alpha$. Moreover, it is called J **regularly decreasing at $\mathbf{x_0} \in \mathbf{X}$** if the set of all descent directions at x_0 is a convex set.

Definition 3 Let $\mathscr{Q}_i$ be a set giving an inequality restriction $(\text{int}(\mathscr{Q}_i) \neq \emptyset)$. Then $h \in X$ is called a **feasible direction** for $\mathscr{Q}_i$ at $x_0 \in X$ if there is a neighborhood U of h such that for all $\varepsilon \in (0, \varepsilon_0)$ and any $\overline{h} \in U$ the vectors $x_0 + \varepsilon\overline{h} \in \mathscr{Q}_i$. Moreover, it is called the restriction $\mathscr{Q}_i$ **regular at $\mathbf{x_0} \in \mathbf{X}$** if the set of feasible directions for $\mathscr{Q}_i$ at x_0 is a convex set.

Definition 4 Let X be a Banach space. The vector $h \in X$ is called a **tangent direction** to the restriction $\mathscr{Q}_i$ at x_0 if for any $\varepsilon \in (0, \varepsilon_0)$ there is a point $x(\varepsilon) \in \mathscr{Q}_i$ such that if we put $x(\varepsilon) = x_0 + \varepsilon h + r(\varepsilon)$, we have that $r(\varepsilon) \in X$ is such that for a neighborhood U of zero, $[r(\varepsilon)]^{-1} \in U$ for any $\varepsilon > 0$ small enough, or equivalently $\|r(\varepsilon)\| = o(\varepsilon)$. Moreover, it is called a **tangent space** if the set of all tangent directions is a vectorial subspace, and the inequality restriction $\mathscr{Q}_\mathbf{i}$ **is regular at $\mathbf{x_0}$** if the set of all tangent directions for $\mathscr{Q}_i$ at x_0 is a convex set.

Definition 5 Let X be a Banach space. A set $K \subset X$ is called a **cone with vertex at zero** if $\lambda x \in K$ for all $\lambda > 0$ and $x \in K$. Moreover, the called **dual cone** for K is denoted by K^* and is defined as $K^* = \left\{\varphi \in X^* \; ; \; \varphi(x) \geq 0, \quad \forall x \in K\right\}$.

Proposition 1 *The descent, feasible and tangent directions generate cones with vertex at zero. Moreover, the cones generated by the descent and feasible directions are open sets.*

Theorem 6 (Dubovitskii and Milyutin) *Consider the optimization problem* (17). *Assume that J has a local minimum at $x_0 \in \mathscr{Q} = \bigcap_{i=1}^{n+1} \mathscr{Q}_i$, J is regularly decreasing at x_0, with descent directions cone K_0, $\mathscr{Q}_i, i = 1, \ldots, n$, are regular at x_0, with feasible directions cone K_i, and $\mathscr{Q}_{n+1}$ is regular at x_0, with tangent directions cone K_{n+1}. Then, there exist $n+1$ continuous linear functionals $G_i \in K_i^*$, not all identically zero, such that*

$$\sum_{i=1}^{n+1} G_i = 0.$$

2.4 Some Concepts of Differential Calculus

We recall some definitions and concepts of differential calculus on Banach Spaces. For more details consult the book of Brezis [3]

Let X and Y be normed vector spaces, U a neighborhood of $x_0 \in X$, and an application $F : U \subset X \to Y$.

Definition 6 We say that F has a **derivative in the direction h** $\in$ **X at x$_0$** if there exists the Y-limit

$$\lim_{\varepsilon \to 0^+} \frac{F(x_0 + \varepsilon h) - F(x_0)}{\varepsilon} := F'(x_0, h) \in Y.$$

Definition 7 Consider that $F'(x_0, h)$ exists for all $h \in X$. The application $\delta F(x_0, \cdot) : X \to Y$ defined by $\delta F(x_0, h) = F'(x_0, h)$ is called the **first variation of F at x$_0$**.

Definition 8 Consider that F has a first variation at x_0 and there exists a linear continuous operator $\Lambda \in \mathscr{L}(X, Y)$ such that $\delta F(x_0, h) = \Lambda h$. Then the operator Λ is called the **Gâteaux derivative** of F at x_0 and is denoted by $F'_G(x_0)$. Thus $F'_G(x_0) \in \mathscr{L}(X, Y)$ such that

$$F(x_0 + \varepsilon h) = F(x_0) + \varepsilon F'_G(x_0)h + o(\varepsilon),$$

is satisfied for each $h \in X$ when $\varepsilon \downarrow 0$.

Definition 9 We say that F is **Fréchet differentiable** at x_0, if at some neighborhood of x_0 the following relation holds

$$\begin{aligned} &F(x_0 + h) = F(x_0) + \Lambda h + \alpha(h)\|h\|_X, \\ &\text{with} \quad \Lambda \in \mathscr{L}(X, Y) \quad \text{and} \quad \lim_{\|h\|_X \to 0} \|\alpha(h)\|_Y = \|\alpha(0)\|_Y = 0. \end{aligned}$$

Moreover, the operator Λ is called the Fréchet derivative (or briefly the derivative) of the application F in x_0 and is denoted by $F'(x_0)$.

Definition 10 We say that F is **strictly differentiable at x$_0$** if there exists $\Lambda \in \mathscr{L}(X, Y)$ such that, for all $\varepsilon > 0$ there exists $\delta > 0$ such that the following inequality holds

$$\|F(x_1) - F(x_2) - \Lambda(x_1 - x_2)\| \le \varepsilon \|x_1 - x_2\|$$

for all $x_1, x_2 \in X$ satisfying the restrictions $\|x_1 - x_0\| \le \delta$ and $\|x_2 - x_0\| \le \delta$.

Remark 1 If F is Gâteaux differentiable in each $x \in U$ and the application $x \in U \mapsto F'_G(x) \in \mathscr{L}(X, Y)$ is continuous in x_0, then F is strictly differentiable in U.

2.5 Some Generic Results for Explicit Calculus of Cones

For more details consult the book of Girsanov [7].

Theorem 7 (Descent Cones) *Let X be a real Banach space and $J : X \to \mathbb{R}$ a functional. Assume that J satisfies the Lipschitz condition in a neighborhood of $x_0 \in X$ ($|J(x_1) - J(x_2)| \le L(x_0)\|x_1 - x_2\|_X$ for any $x_1, x_2 \in U(x_0)$), and it is directional differentiable at x_0 for all directions $h \in X$ ($\exists\, J'(x_0, h)$ for all $h \in X$). If moreover $J'(x_0, h)$ is a convex function of h, then J is regularly decreasing at x_0 and the cone of descent directions K_0 is given by*

$$K_0 = \Big\{h \in X \; ; \; J'(x_0, h) < 0\Big\}.$$

Corollary 2 *Let X be a real Banach space and $J : X \to \mathbb{R}$ a functional. Then*

(i) If J is a convex and continuous functional, then J is regularly decreasing for all $x_0 \in X$ and K_0 is given by $K_0 = \Big\{h \in X \; ; \; J'(x_0, h) < 0\Big\}$.

(ii) If J is Fréchet differentiable, then J is regularly decreasing for all $x_0 \in X$ and K_0 is given by $K_0 = \Big\{h \in X \; ; \; J'(x_0)h < 0\Big\}$.

Theorem 8 (Lyusternik; Tangent Cone) *Let X and Y be two real Banach spaces, U a neighborhood of $x_0 \in X$, $M : U \subset X \to Y$ a map such that $M(x_0) = 0$. If M is strictly differentiable at x_0 and $M'(x_0)X \equiv Y$ ($M'(x_0)$ is an epimorphism), then the set $\mathcal{Q} = \{x \in X \; ; \; M(x) = 0\}$ is regular at x_0 and its tangent space at x_0 is given by*

$$T_{x_0}(\mathcal{Q}) = \ker M'(x_0) = \Big\{h \in X \; ; \; M'(x_0)h = 0\Big\}.$$

Theorem 9 (Dual Cone) *If the cone K is a vectorial subspace of the normed space X, then its dual cone is $K^* = \Big\{G \in X' \; ; \; G(h) = 0, \; \forall h \in K\Big\}$.*

3 Reformulation of the Optimal Control Problem and Specific Calculus of Cones

First of all, we can identify our particular problem (4) with the generic problem (17), taking

$$X = E, \quad Y = \overline{E}, \quad \mathcal{Q}_i = \emptyset \quad (i = 1, \cdots, n), \quad \mathcal{Q}_{n+1} = \mathcal{U}_{ad}.$$

By applying Corollary 2 we can prove the following Lemma.

Lemma 3 *Consider the spaces given in* (2)*,* $E' = E'_u \times E'_\phi \times \mathscr{U}'_c$ *the dual of* E*,* J *the functional defined in* (3)*, and take into account that the derivative of the functional* $J(\cdot,\cdot,\cdot)$ *at* $(u,\phi,f) \in E$ *in the direction* $(\bar{u},\bar{\phi},\bar{f}) \in E$ *is given by*

$$J'(u,\phi,f)(\bar{u},\bar{\phi},\bar{f}) = \alpha k \int_0^T \int_{\omega_d} (u-u_d)^{2k-1} \bar{u}\, dxdt \\ + \beta m \int_0^T \int_{\omega_d} (\phi-\phi_d)^{2m-1} \bar{\phi}\, dxdt + \mu s \int_0^T \int_{\omega_c} (f)^{2s-1} \bar{f}\, dxdt.$$

Then, the following sets

$$DC(J,(u,\phi,f)) = \Big\{ (\bar{u},\bar{\phi},\bar{f}) \in E\ ;\ J'(u,\phi,f)(\bar{u},\bar{\phi},\bar{f}) < 0 \Big\},$$

$$[DC(J)]^* = \Big\{ G \in E'\ ;\ G(\bar{u},\bar{\phi},\bar{f}) = -\lambda J'(u,\phi,f)(\bar{u},\bar{\phi},\bar{f})\ \textit{for some}\ \lambda \geq 0 \Big\}$$

are the cone of descent directions for the functional $J(\cdot,\cdot,\cdot)$ *at the point* (u,ϕ,f) *and its dual cone, respectively.*

Concerning to differentiability of the map $M : E \to \overline{E}$ (see (5)) defining the equality restriction given in problem (4), we have the following result.

Lemma 4 *Consider the application* M *defined in* (5)*. Then, the following assertions are valid:*

(i) The application $M : E \to \overline{E}$ *is Gâteaux differentiable and the Gâteaux derivative of* M *in* $(u,\phi,f) \in E$ *is* $M'_G(u,\phi,f)(\bar{u},\bar{\phi},\bar{f}) = (\bar{\psi}_1,\bar{\psi}_2,\bar{\psi}_3,\bar{\psi}_4) \in \overline{E}$ *where*

$$\partial_t \bar{u} + \ell \partial_t \bar{\phi} - \Delta \bar{u} - \bar{f} \chi_{\omega_c} = \bar{\psi}_1, \qquad \textit{in } Q, \tag{18a}$$

$$\partial_t \bar{\phi} - \Delta \bar{\phi} - \bar{u} - F'(\phi)\bar{\phi} = \bar{\psi}_2, \qquad \textit{in } Q, \tag{18b}$$

$$\bar{u}(0) = \bar{\psi}_3,\ \bar{\phi}(0) = \bar{\psi}_4, \qquad \textit{in } \Omega, \tag{18c}$$

$$\frac{\partial \bar{u}}{\partial \mathbf{n}} = \frac{\partial \bar{\phi}}{\partial \mathbf{n}} = 0, \qquad \textit{on } \Sigma. \tag{18d}$$

(ii) Under hypothesis (6) *(that is* $k \leq 5$ *if* $s = 1$*), the application* $M(\cdot,\cdot,\cdot)$ *is strictly differentiable and the linear operator* $M'(u,\phi,f) = M'_G(u,\phi,f)$ *is surjective (i.e.* $M'(u,\phi,f)E \equiv \overline{E}$*).*

Proof

(i) By definition of the application M and using (18), we deduce that for any $(u, \phi, f), (\bar{u}, \bar{\phi}, \bar{f}) \in E$,

$$\frac{1}{\varepsilon}\Big[M(u+\varepsilon\bar{u}, \phi+\varepsilon\bar{\phi}, f+\varepsilon\bar{f}) - M(u, \phi, f)\Big]$$
$$= \Big(\partial_t\bar{u} + \ell\partial_t\bar{\phi} - \Delta\bar{u} - \bar{f}\chi_{\omega_c},\ \partial_t\bar{\phi} - \Delta\bar{\phi} - \frac{F(\phi+\varepsilon\bar{\phi}) - F(\phi)}{\varepsilon} - \bar{u},\ \bar{u}(0),\ \bar{\phi}(0)\Big)$$
$$= \Big(\bar{\psi}_1,\ \bar{\psi}_2 + F'(\phi)\bar{\phi} - \frac{F(\phi+\varepsilon\bar{\phi}) - F(\phi)}{\varepsilon},\ \bar{\psi}_3, \bar{\psi}_4\Big).$$

Then

$$\lim_{\varepsilon\to 0^+}\left\|\frac{M(u+\varepsilon\bar{u}, \phi+\varepsilon\bar{\phi}, f+\varepsilon\bar{f}) - M(u, \phi, f)}{\varepsilon} - (\bar{\psi}_1, \bar{\psi}_2, \bar{\psi}_3, \bar{\psi}_4)\right\|_{\overline{E}}$$
$$= \lim_{\varepsilon\to 0^+}\left\|F'(\phi)\bar{\phi} - \frac{F(\phi+\varepsilon\bar{\phi}) - F(\phi)}{\varepsilon}\right\|_{L^{\overline{q}}(Q)} = 0,$$

owing to $\phi, \bar{\phi} \in E_\phi \subset L^\infty(Q)$. Thus, by the definition of Gâteaux derivative, we conclude the proof of item *(i)*.

(ii) By using the Remark 1, for proving that $M(\cdot, \cdot, \cdot)$ is strictly differentiable is enough to check that the application $M'_G : (u, \phi, f) \in E \mapsto M'_G(u, \phi, f) \in \mathscr{L}(E, \overline{E})$ is continuous. We note that the continuity of the application M'_G is a direct consequence of the continuity of the map $\phi \in W^{2,1}_{\overline{q}}(Q) \mapsto F'(\phi)\bar{\phi} \in L^{\overline{q}}(Q)$. Finally, for any $(\bar{\psi}_1, \bar{\psi}_2, \bar{\psi}_3, \bar{\psi}_4) \in \overline{E}$, one has the existence of solutions $(\bar{u}, \bar{\phi}, \bar{f}) \in E$ of system (18), because if we follow the argument of Theorem 4, for any $\bar{f} \in L^q(Q_c)$, there exists a unique $(\bar{u}, \bar{\phi}) \in W^{2,1}_{q,n}(Q) \times W^{2,1}_{\overline{q},n}(Q)$ such that $(\bar{u}, \bar{\phi}, \bar{f})$ solves system (18). Finally, owing to hypothesis $k \leq 5$ if $s = 1$ and the embeddings of Lemma 1, one has $\bar{u} \in L^{2k}(Q)$ (always $\bar{\phi} \in L^\infty(Q) \subset L^{2m}(Q)$), hence $(\bar{u}, \bar{\phi}, \bar{f}) \in E_u \times E_\phi \times E_f = E$. Then, the operator $M'(u, \phi, f) : E \to \overline{E}$ is surjective. □

Then using Lemma 4, Theorems 8 and 9 we prove the following result.

Lemma 5 *Under hypothesis* (6) *($k \leq 5$ if $s = 1$), the following sets*

$$TC(\mathscr{U}_{ad}, (u, \phi, f)) = \Big\{(\bar{u}, \bar{\phi}, \bar{f}) \in E\ ;\ M'(u, \phi, f)(\bar{u}, \bar{\phi}, \bar{f}) = 0\Big\},$$
$$[TC(\mathscr{U}_{ad})]^* = \Big\{G \in E'\ ;\ G(\bar{u}, \bar{\phi}, \bar{f}) = 0,\ \forall(\bar{u}, \bar{\phi}, \bar{f}) \in TC(\mathscr{U}_{ad}, (u, \phi, f))\Big\},$$

are the tangent and dual cones to the set $\mathscr{U}_{ad}$ in $(u, \phi, f) \in \mathscr{U}_{ad}$, respectively. Moreover $TC(\mathscr{U}_{ad}, (u, \phi, f))$ is a vectorial subspace.

4 Proof of Theorem 1

First we prove the non-triviality condition $\mathcal{U}_{ad} \neq \emptyset$. From Theorem 4, for all $f \in L^q(Q_c)(q = 2s \geq 2)$ there exists a unique solution $(u, \phi) \in W_q^{2,1}(Q) \times W_{\overline{q}}^{2,1}(Q)$ of the system (1) with $\overline{q}$ given by (14). Taking $f \equiv 0$, there exist $(\widetilde{u}, \widetilde{\phi}) \in W_q^{2,1}(Q) \times W_{\overline{q}}^{2,1}(Q)$ such that $(\widetilde{u}, \widetilde{\phi}, 0)$ solves (1), for any $q, \overline{q} < +\infty$, i.e. $M(\widetilde{u}, \widetilde{\phi}, 0) = 0$. Moreover, from (15)

$$\|\widetilde{u}\|_{W_q^{2,1}(Q)} + \|\widetilde{\phi}\|_{W_{\overline{q}}^{2,1}(Q)} \leq C\big(\|u_0\|_{W_\infty^{2,1}(Q)} + \|\phi_0\|_{W_\infty^{2,1}(Q)}\big),$$

In particular, from embeddings given in Lemma 1, $(\widetilde{u}, \widetilde{\phi}) \in L^\infty(Q) \times L^\infty(Q)$. Then, $J(\widetilde{u}, \widetilde{\phi}, 0) < \infty$ and therefore $\mathcal{U}_{ad} \neq \emptyset$.

Now, we use the classical minimizing sequence technique. Let us consider a sequence $\{(u_n, \phi_n, f_n)\} \subset \mathcal{U}_{ad}$ such that

$$\lim_{n\to\infty} J(u_n, \phi_n, f_n) = \inf_{(u,\phi,f)\in\mathcal{U}_{ad}} J(u, \phi, f).$$

From the fact that $J(u_n, \phi_n, f_n) \leq C$ and the definition of $J(\cdot, \cdot, \cdot)$, we follow that

$$\|u_n\|_{L^{2k}(Q_d)} + \|\phi_n\|_{L^{2m}(Q_d)} + \|f_n\|_{L^{2s}(Q_c)} \leq C$$

and from inequality (15) given in Theorem 4, we have the estimate

$$\|u_n\|_{W_q^{2,1}(Q)} + \|\phi_n\|_{W_{\overline{q}}^{2,1}(Q)} \leq C$$

where $q = 2s$ and $\overline{q} = 10$ if $s = 1$ or $\overline{q} < \infty$ otherwise. Thus, possibly selecting a subsequence of $\{(u_n, \phi_n, f_n)\}$, we deduce that

$$\begin{aligned} f_n &\rightharpoonup f \quad \text{weakly in } L^{2s}(Q_c),\\ u_n &\rightharpoonup u \quad \text{weakly in } W_q^{2,1}(Q) \cap L^{2k}(Q_d),\\ \phi_n &\rightharpoonup \phi \quad \text{weakly in } W_{\overline{q}}^{2,1}(Q) \cap L^{2m}(Q_d). \end{aligned} \tag{19}$$

By applying compactness given in Lemma 2, we have in particular that

$$\phi_n \to \phi \text{ strongly in } L^p(Q), \quad \forall\, p \leq \infty, \tag{20}$$

hence, in particular, $F(\phi_n) \to F(\phi)$ strongly in $L^p(Q)$ for any $p < \infty$. Now, we write the system (1) for each term of the sequence $\{(u_n, \phi_n, f_n)\}$ and multiply the equations by appropriate test functions and integrate by parts. The convergence relations (19) and (20) implies that we can take the limit on n and conclude that

(u, ϕ, f) satisfies (1), hence $(u, \phi, f) \in \mathcal{U}_{ad}$. Now, using that the functional $J(\cdot, \cdot, \cdot)$ is weakly lower semi-continuous in $L^{2k}(Q_d) \times L^{2m}(Q_d) \times L^{2s}(Q_c)$, we deduce that

$$J(u, \phi, f) \leq \lim \inf_{n \to \infty} J(u_n, \phi_n, f_n)$$

hence (u, ϕ, f) is a global optimal solution and the proof is finished. □

5 Proof of Theorem 2

If $(u, \phi, f) \in \mathcal{U}_{ad}$ is a local optimal solution of problem (4), we have that

$$DC(J, (u, \phi, f)) \cap TC(\mathcal{U}_{ad}, (u, \phi, f)) = \emptyset.$$

Then by the Dubovitskii and Milyutin Theorem (see Theorem 6), we have that there exist continuous functionals $G_1 \in [DC(J)]^*$ and $G_2 \in [TC(\mathcal{U}_{ad})]^*$, not both identically zero, such that satisfy the Euler-Lagrange equation

$$G_1 + G_2 = 0. \tag{21}$$

Let $\bar{f} \in L^{2s}(Q_c)$ be and arbitrary control and $(\bar{u}, \bar{\phi}) \in E_u \times E_\phi$ the solution of the following system (the homogeneous problem of (18))

$$\partial_t \bar{u} + \ell \partial_t \bar{\phi} - \Delta \bar{u} = \bar{f} \chi_{\omega_c}, \qquad \text{in } Q; \tag{22a}$$

$$\partial_t \bar{\phi} - \Delta \bar{\phi} - F'(\phi) \bar{\phi} = \bar{u}, \qquad \text{in } Q; \tag{22b}$$

$$\bar{u}(0) = 0, \ \bar{\phi}(0) = 0, \qquad \text{in } \Omega; \tag{22c}$$

$$\frac{\partial \bar{u}}{\partial \mathbf{n}} = \frac{\partial \bar{\phi}}{\partial \mathbf{n}} = 0, \qquad \text{on } \Sigma. \tag{22d}$$

In this case we have that $(\bar{u}, \bar{\phi}, \bar{f}) \in TC(\mathcal{U}_{ad}, (u, \phi, f))$ and consequently $G_2(\bar{u}, \bar{\phi}, \bar{f}) = 0$. From (21) and Lemma 3 we follow that there exists some $\lambda \geq 0$ such that

$$0 = G_1(\bar{u}, \bar{\phi}, \bar{f}) = -\lambda \alpha k \int_0^T \int_{\omega_d} (u - u_d)^{2k-1} \bar{u} \, dx dt$$

$$-\lambda \beta m \int_0^T \int_{\omega_d} (\phi - \phi_d)^{2m-1} \bar{\phi} \, dx dt - \lambda \mu s \int_0^T \int_{\omega_c} (f)^{2s-1} \bar{f} \, dx dt. \tag{23}$$

We note that λ is strictly positive, since if we assume that $\lambda = 0$ we have that $G_1 \equiv 0$ and from Eq. (21) we deduce that $G_2 \equiv 0$, which is a contradiction with Dubovitskii and Milyutin Theorem 6. In particular, by dividing (21) by λ, we can fix $\lambda = 1$.

Let $(v, \psi) \in W^{2,1}_{2,n}(Q) \times W^{2,1}_{2,n}(Q)$ be a solution of the adjoint system (7). Summing Eqs. (7a) and (7b), we obtain

$$\begin{aligned}
&-\alpha k \int_0^T \int_{\omega_d} (u-u_d)^{2k-1}\, \bar{u}\, dxdt - \beta m \int_0^T \int_{\omega_d} (\phi-\phi_d)^{2m-1}\, \bar{\phi}\, dxdt \\
&= \int_0^T \int_{\Omega} (\partial_t v + \Delta v + \psi)\, \bar{u}\, dxdt + \int_0^T \int_{\Omega} (\partial_t \psi + l\partial_t v + \Delta\psi + F'(\phi)\psi)\, \bar{\phi}\, dxdt.
\end{aligned} \tag{24}$$

Now, integrating by parts, noticing that $\bar{u}(0) = \bar{\phi}(0) = v(T) = \psi(T) = 0$ in Ω and $\frac{\partial \bar{u}}{\partial \mathbf{n}} = \frac{\partial \bar{\phi}}{\partial \mathbf{n}} = \frac{\partial v}{\partial \mathbf{n}} = \frac{\partial \psi}{\partial \mathbf{n}} = 0$ on Σ, we obtain

$$\begin{aligned}
&-\alpha k \int_0^T \int_{\omega_d} (u-u_d)^{2k-1}\, \bar{u}\, dxdt - \beta m \int_0^T \int_{\omega_d} (\phi-\phi_d)^{2m-1}\, \bar{\phi}\, dxdt \\
&= -\int_0^T \int_{\Omega} v\,(\partial_t \bar{u} - \Delta\bar{u} + l\partial_t\bar{\phi})\, dxdt - \int_0^T \int_{\Omega} \psi\,(\partial_t\bar{\phi} - \bar{u} - \Delta\bar{\phi} - F'(\phi)\bar{\phi})\, dxdt.
\end{aligned} \tag{25}$$

Comparing (25) with the system (22), we have

$$\begin{aligned}
-\alpha k \int_0^T \int_{\omega_d} (u-u_d)^{2k-1}\, \bar{u}\, dxdt - \beta m \int_0^T \int_{\omega_d} (\phi-\phi_d)^{2m-1}\, \bar{\phi}\, dxdt& \\
= -\int_0^T \int_{\omega_c} v\, \bar{f}\, dxdt.&
\end{aligned} \tag{26}$$

From (26) and (23) (for $\lambda = 1$), we deduce that

$$\mu s \int_0^T \int_{\omega_c} (f)^{2s-1}\, \bar{f}\, dxdt = \int_0^T \int_{\omega_c} (-v)\, \chi_{\omega_c}\, \bar{f}\, dxdt.$$

Now, since $\bar{f} \in L^q(Q_c)$ is arbitrary we deduce $\mu s\,(f)^{2s-1} = -v\chi_{\omega_c}$, hence (9) holds and the proof is finished. □

6 Proof of Theorem 3

The proof of Theorems 3, in a broad sense, is based on the Leray-Schauder fixed point theorem using the global in time a priori estimate (11), which can be deduced testing (7a) by $-\ell\,\partial_t v$ and (7b) by ψ (hence terms $\int_\Omega \partial_t v\,\psi\,dxdt$ cancel) and standard techniques of L^p-regularity for parabolic equations.

The remained proof will be split in two parts: existence and uniqueness.

6.1 Existence

To prove the existence of a solution for the adjoint problem (7) we apply the Leray-Schauder theorem. Indeed, let us consider the Banach space $Z = L^2(Q) \times L^2(Q)$ and we define the family of applications $T : Z \to Z$ as follows

$$T(\widehat{v}, \widehat{\psi}) = (v, \psi), \tag{27}$$

where (v, ψ) is the solution of the following auxiliary (decoupled) problem

$$-\partial_t v - \Delta v + v = \widehat{\psi} + g_1(u) + \widehat{v}, \qquad \text{in } Q, \tag{28a}$$

$$-\partial_t \psi - \Delta\psi + \psi = g_2(\phi) + \ell\partial_t v - F'(\phi)\widehat{\psi} + \widehat{\psi}, \qquad \text{in } Q, \tag{28b}$$

$$v(T) = \psi(T) = 0, \qquad \text{in } \Omega, \tag{28c}$$

$$\frac{\partial v}{\partial \mathbf{n}} = \frac{\partial \psi}{\partial \mathbf{n}} = 0, \qquad \text{on } \Sigma. \tag{28d}$$

Now, we are going to prove that $T : Z \to Z$ is well defined and satisfy the hypothesis of Leray-Schauder theorem, hence we can conclude that T has a fixed point (v, ψ), which is a solution of (7).

(i) T is well defined. By using Theorem 4 with $q = 2s$, the embeddings given in Lemma 1 and constraint $k \le 3$ if $s = 1$ given in (10), we have:

$$s = 1 : u \in W_2^{2,1}(Q) \quad \Rightarrow \quad u \in L^{10}(Q)$$
$$\Rightarrow \quad g_1(u) \sim (u - u_d)^{2k-1} \in L^{\frac{10}{2k-1}}(Q) \subset L^2(Q), \text{ if } k = 1, 2, 3, \tag{29}$$

$$s \ge 2 : u \in W_{2s}^{2,1}(Q) \;\; (2s \ge 4) \quad \Rightarrow \quad u \in L^\infty(Q),$$
$$\Rightarrow \quad g_1(u) \sim (u - u_d)^{2k-1} \in L^\infty(Q), \quad \forall k \ge 1. \tag{30}$$

Thus, in both cases ((29) and (30)) one has

$$g_1(u) \in L^2(Q). \tag{31}$$

Then, by applying the L^2-regularity of parabolic equations, we follow that there is a unique $v \in W_2^{2,1}(Q)$, solution of the backward problem defined by Eq. (28a) with end condition (28c) and boundary condition (28d). Moreover, one has the estimate

$$\|v\|_{W_2^{2,1}(Q)} \leq C\, \|\widehat{\psi} + g_1(u) + \widehat{v}\|_{L^2(Q)}. \tag{32}$$

On the other hand, since $\phi \in W_{\overline{q}}^{2,1}(Q)$, Lemma 1 implies that $\phi \in L^\infty(Q)$. Then, $F'(\phi) \in L^\infty(Q)$ and

$$g_2(\phi) + \ell\partial_t v - F'(\phi)\widehat{\psi} + \widehat{\psi} \in L^2(Q). \tag{33}$$

Thus, by applying again the L^2-regularity for parabolic equations [11, Theorem 9.1, pp. 341] now to the backward problem defined by (28b), (28c) and (28d), there exits a unique solution $\psi \in W_2^{2,1}(Q)$. Moreover, one has the estimate

$$\|\psi\|_{W_2^{2,1}(Q)} \leq C\, \|g_2(\phi) + \ell\partial_t v - F'(\phi)\widehat{\psi} + \widehat{\psi}\|_{L^2(Q)}, \tag{34}$$

which is bounded owing to (32). Therefore, T is well defined.

(ii) *The application $T : Z \to Z$ is continuous.* For each $i \in \{1, 2\}$, let us consider $(\widehat{v_i}, \widehat{\psi_i}) \in Z(= L^2(Q) \times L^2(Q))$, and denote by $(v_i, \psi_i) = T_\varepsilon(\widehat{v_i}, \widehat{\psi_i})$ the corresponding solution of the system (28) and by $(\overline{v}, \overline{\psi})$ its differences, i.e. $\overline{v} = v_1 - v_2$ and $\overline{\psi} = \psi_1 - \psi_2$. Then, making differences in system (28) we deduce that

$$-\partial_t\overline{v} - \Delta\overline{v} + \overline{v} =(\widehat{\psi}_1 - \widehat{\psi}_2) + (\widehat{v}_1 - \widehat{v}_2), \qquad \text{in } Q, \tag{35a}$$

$$-\partial_t\overline{\psi} - \Delta\overline{\psi} + \overline{\psi} =\ell\partial_t\overline{v} - F'(\phi)(\widehat{\psi}_1 - \widehat{\psi}_2) + (\widehat{\psi}_1 - \widehat{\psi}_2), \qquad \text{in } Q, \tag{35b}$$

$$\overline{v}(T) = \overline{\psi}(T) =0, \qquad \text{in } \Omega, \tag{35c}$$

$$\frac{\partial\overline{v}}{\partial\mathbf{n}} = \frac{\partial\overline{\psi}}{\partial\mathbf{n}} =0, \qquad \text{on } \Sigma. \tag{35d}$$

Considering the system (35) and the same decoupled argument done in the previous part *(i)*, we can obtain the estimates

$$\|v_1 - v_2\|_{W_2^{2,1}(Q)} \leq C(\|\widehat{\psi}_1 - \widehat{\psi}_2\|_{L^2(Q)} + \|\widehat{v}_1 - \widehat{v}_2\|_{L^2(Q)})$$

and

$$\begin{aligned}\|\psi_1 - \psi_2\|_{W_2^{2,1}(Q)} &\leq C(\|\partial_t v_1 - \partial_t v_2\|_{L^2(Q)} + \|\widehat{\psi}_1 - \widehat{\psi}_2\|_{L^2(Q)}) \\ &\leq C(\|\widehat{v}_1 - \widehat{v}_2\|_{L^2(Q)} + \|\widehat{\psi}_1 - \widehat{\psi}_2\|_{L^2(Q)}).\end{aligned}$$

Hence $T(\cdot, \cdot)$ is a continuous operator.

(iii) *The operator $T : Z \to Z$ is compact.* This fact is consequence of estimates (32) and (34) in $W_2^{2,1}(Q)$ and the compact embedding of $W_2^{2,1}(Q)$ in $L^2(Q)$ (see Lemma 2).

(iv) *The set of fixed points of εT for any $\varepsilon \in (0, 1]$ is bounded in Z.* Let (v, ψ) be a fixed point of εT. Then, (v, ψ) is a solution of the following ε-dependent adjoint problem

$$-\partial_t v - \Delta v + v = \varepsilon\psi + \varepsilon\, g_1(u) + \varepsilon\, v, \qquad \text{in } Q, \tag{36a}$$

$$-\partial_t \psi - \ell\partial_t v - \Delta\psi + \psi = \varepsilon\, g_2(\phi) - \varepsilon\, F'(\phi)\psi + \varepsilon\, \psi, \qquad \text{in } Q, \tag{36b}$$

$$v(T) = \psi(T) = 0, \qquad \text{in } \Omega, \tag{36c}$$

$$\frac{\partial v}{\partial \mathbf{n}} = \frac{\partial \psi}{\partial \mathbf{n}} = 0, \qquad \text{on } \Sigma. \tag{36d}$$

Similar estimates given in the proof of Theorems 4, testing (36a) by $-\ell\, \partial_t v$ and (36b) by ψ (again the terms $\int_\Omega \partial_t v\, \psi$ cancel), and standard techniques of L^p-regularity for parabolic equations, imply that

$$\|v\|_{W_2^{2,1}(Q)} + \|\psi\|_{W_2^{2,1}(Q)} \leq C. \tag{37}$$

By the embedding of $W_2^{2,1}(Q)$ in $L^2(Q)$ (Lemma 1) we deduce that the set of all possible fixed points of εT is bounded in $Z = L^2(Q)$.

Therefore, by *(i)–(iv)* the operators T satisfies the hypotheses of Leray-Schauder theorem, then there exist $(v, \psi) \in W_2^{2,1}(Q) \times W_2^{2,1}(Q)$ a solution of

$$(v, \psi) = T(v, \psi). \tag{38}$$

Thus, (v, ψ) is also a solution of the adjoint system (7).

6.2 Uniqueness

For each $i \in \{1, 2\}$, let us consider $(v_i, \psi_i) \in W_2^{2,1}(Q) \times W_2^{2,1}(Q)$ two solutions of the system (7) and denote by $(\overline{v}, \overline{\psi})$ its differences, i.e. $\overline{v} = v_1 - v_2$ and $\overline{\psi} = \psi_1 - \psi_2$. It suffices to prove that $(\overline{v}, \overline{\psi}) = (0, 0)$. For this, making differences in system (7) we deduce that

$$-\partial_t\overline{v} - \Delta\overline{v} - \overline{\psi} = 0, \quad \text{in } Q, \tag{39a}$$

$$-\partial_t\overline{\psi} - \ell\partial_t\overline{v} - \Delta\overline{\psi} + F'(\phi)\overline{\psi} = 0, \quad \text{in } Q, \tag{39b}$$

$$\overline{v}(T) = \overline{\psi}(T) = 0, \quad \text{in } \Omega, \tag{39c}$$

$$\frac{\partial\overline{v}}{\partial\mathbf{n}} = \frac{\partial\overline{\psi}}{\partial\mathbf{n}} = 0. \quad \text{on } \Sigma. \tag{39d}$$

Testing (39a) by $-\ell\,\partial_t\overline{v}$ and (39b) by $\overline{\psi}$ (again the terms $\int_\Omega \partial_t\overline{v}\,\overline{\psi}$ cancel), one has

$$\frac{\ell}{2}\|\partial_t\overline{v}\|^2_{L^2(\Omega)} + \|\nabla\overline{\psi}\|^2_{L^2(\Omega)} - \frac{1}{2}\frac{d}{dt}(\ell\|\overline{v}\|^2_{H^1(\Omega)} + \|\overline{\psi}\|^2_{L^2(\Omega)}) \le \frac{\ell}{2}\|\overline{\psi}\|^2_{L^2(\Omega)} + C\|\overline{\psi}\|^2_{L^2(\Omega)}$$

Therefore, the Gronwall's lemma implies uniqueness.

Acknowledgements A. Coronel thanks the support of research project DIUBB GI 172409/C at Universidad del Bío-Bío, Chile. F.Guillen-Gonzalez has been partially supported by project MTM2015-69875-P, Spain. M.A. Rojas-Medar has been partially supported by project MTM2012-32325, Spain, and Grant 1120260, Fondecyt-Chile.

References

1. Araruna, F.D., Boldrini, J.L., Calsavara, B.M.R.: Optimal control and controllability of a phase field system with one control force. Appl. Math. Optim. **70**(3), 539–563 (2014)
2. Boldrini, J.L., Caretta, B.M.C., Fernández-Cara, E.: Some optimal control problems for a two-phase field model of solidification. Rev. Mat. Complut. **23**(1), 49–75 (2010)
3. Brezis, H.: Analyse fonctionnelle: theórie et applications. Collection Mathématiques appliqués pour la maitrise. Dunod, Paris (1987)
4. Caginalp, G.: Analysis of a phase field model of a free boundary. Arch. Ration. Mech. Anal. **92**, 205–245 (1986)
5. Colli, P., Gilardi, G., Marinoschi, G., Rocca, E.: Optimal control for a phase field system with a possibly singular potential. Math. Control Relat. Fields **6**(1), 95–112 (2016)
6. Fix, G.J.: Phase field models for free boundary problems. In: Fasano, A., Primicerio, M. (eds.) Free Boundary Problems: Theory and Applications, vol. II. Pitman Research Notes in Mathematics Series, vol. 79, pp. 580–589. Longman, London (1983)
7. Girsanov, I.V.: Lectures on Mathematical Theory of Extremum Problems. Lectures Notes in Economics and Mathematical Systems, vol. 67. Springer, New York (1972)

8. Gnanavel, S., Barani Balan, N., Balachandran, K.: Simultaneous identification of two time independent coefficients in a nonlinear phase field system. J. Optim. Theory Appl. **160**(3), 992–1008 (2014)
9. Hoffman, K.H., Jiang, L.: Optimal control of a phase field model for solidification. Numer. Funct. Anal. Optim. **13**, 11–27 (1992)
10. Kotarski, W.: Characterization of Pareto optimal points in problems with multi-equality constraints. Optimization **20**, 93–106 (1989)
11. Ladyẑenskaja, O.A., Solonnikov, V.A., Ural´Ceva, N.N.: Linear and Quasilinear Equations of Parabólic Type, vol. 23. American Mathematical Society, Providence (1968)
12. Lamé, G., Clapeyron, B.P.: Mémoire sur la solidification par refroidissement d'un globe solide. Ann. Chem. Phys. **47**, 250–256 (1831)
13. Miranville, A.: Some mathematical models in phase transition. Discrete Contin. Dyn. Syst. Ser. S **7**(2), 271–306 (2014)
14. Simon, J.: Compact sets in the space $L^p(O, T; B)$. Ann. Mat. Pura Appl. (4) **146**, 65–96 (1987)
15. Stefan, J.: Uber einige Probleme der Theorie der Warmeleitung. S.-B. Wien Akad. Mat. Natur. **98**, 173–484 (1889)
16. Tröltzsch, F.: Optimal Control of Partial Differential Equations. Theory, Methods and Applications. Translated from the 2005 German original by Jürgen Sprekels. Graduate Studies in Mathematics, vol. 112. American Mathematical Society, Providence (2010)
17. Wang, L., Wang, G.: The optimal time control of a phase-field system. SIAM J. Control Optim. **42**(4), 1483–1508 (2003)

Local Regularity for Fractional Heat Equations

Umberto Biccari, Mahamadi Warma, and Enrique Zuazua

Dedicated to Prof. Enrique Fernández-Cara on the occasion of his 60th birthday.

Abstract We prove the maximal local regularity of weak solutions to the parabolic problem associated with the fractional Laplacian with homogeneous Dirichlet boundary conditions on an arbitrary bounded open set $\Omega \subset \mathbb{R}^N$. Proofs combine classical abstract regularity results for parabolic equations with some new local regularity results for the associated elliptic problems.

Keywords Fractional Laplacian · Heat equation · Dirichlet boundary condition · Weak solutions · Local regularity

2010 *Mathematics Subject Classification* 35B65, 35R11, 35S05

U. Biccari
DeustoTech, University of Deusto, Bilbao, Basque Country, Spain

Facultad de Ingeniería, Universidad de Deusto, Bilbao, Basque Country, Spain

M. Warma
University of Puerto Rico (Rio Piedras Campus), College of Natural Sciences, Department of Mathematics, San Juan, PR, USA
e-mail: mahamadi.warma1@upr.edu

E. Zuazua (✉)
DeustoTech, University of Deusto, Bilbao, Basque Country, Spain

Facultad de Ingeniería, Universidad de Deusto, Bilbao, Basque Country, Spain

Departamento de Matemáticas, Universidad Autónoma de Madrid, Madrid, Spain

Sorbonne Universites, UPMC Univ Paris 06, CNRS UMR 7598, Laboratoire Jacques-Louis Lions, Paris, France
e-mail: enrique.zuazua@deusto.es; enrique.zuazua@uam.es

A. Doubova et al. (eds.), *Recent Advances in PDEs: Analysis, Numerics and Control*, SEMA SIMAI Springer Series 17,
https://doi.org/10.1007/978-3-319-97613-6_12

1 Introduction

The aim of the present paper is to study the local regularity of weak solutions to the following parabolic problem

$$\begin{cases} u_t + (-\Delta)^s u = f & \text{in } \Omega \times (0,T) =: \Omega_T, \\ u \equiv 0 & \text{on } (\mathbb{R}^N \setminus \Omega) \times (0,T), \\ u(\cdot,0) \equiv 0 & \text{in } \Omega, \end{cases} \tag{1.1}$$

where $\Omega \subset \mathbb{R}^N$ is an arbitrary bounded open set, f is a given distribution and, for all $s \in (0,1)$, $(-\Delta)^s$ denotes the fractional Laplace operator, which is defined as the following singular integral

$$(-\Delta)^s u(x) := C_{N,s}\,\text{P.V.}\int_{\mathbb{R}^N} \frac{u(x)-u(y)}{|x-y|^{N+2s}}\,dy, \quad x \in \mathbb{R}^N. \tag{1.2}$$

In (1.2), $C_{N,s}$ is a normalization constant given by

$$C_{N,s} := \frac{s2^{2s}\Gamma\left(\frac{2s+N}{2}\right)}{\pi^{\frac{N}{2}}\Gamma(1-s)},$$

Γ being the usual Gamma function.

We are interested in analyzing the local regularity for solutions to the parabolic problem (1.1).

We first introduce the functional setting. Given $\Omega \subset \mathbb{R}^N$, an arbitrary open set, for $p \in (1,\infty)$ and $s \in (0,1)$, we denote by

$$W^{s,p}(\Omega) := \left\{ u \in L^p(\Omega) : \int_\Omega\int_\Omega \frac{|u(x)-u(y)|^p}{|x-y|^{N+ps}}dxdy < \infty \right\},$$

the fractional order Sobolev space endowed with the norm

$$\|u\|_{W^{s,p}(\Omega)} := \left(\int_\Omega |u|^p\,dx + \int_\Omega\int_\Omega \frac{|u(x)-u(y)|^p}{|x-y|^{N+ps}}dxdy \right)^{\frac{1}{p}}.$$

We let

$$W_0^{s,p}(\overline{\Omega}) := \left\{ u \in W^{s,p}(\mathbb{R}^N) : \ u = 0 \ \text{ on } \ \mathbb{R}^N \setminus \Omega \right\},$$

and we shall denote by $W^{-s,2}(\overline{\Omega})$ the dual of the Hilbert space $W_0^{s,2}(\overline{\Omega})$, that is, $W^{-s,2}(\overline{\Omega}) := (W_0^{s,2}(\overline{\Omega}))^\star$. The following continuous embeddings hold

$$W_0^{s,2}(\overline{\Omega}) \hookrightarrow L^2(\Omega) \hookrightarrow W^{-s,2}(\overline{\Omega}).$$

Next, if $s > 1$ is not an integer, we write $s = m + \sigma$ where m is an integer and $0 < \sigma < 1$. In this case

$$W^{s,p}(\Omega) := \Big\{u \in W^{m,p}(\Omega) : \ D^\alpha u \in W^{\sigma,p}(\Omega) \ \text{ for any } \ \alpha \ \text{ such that } \ |\alpha| = m\Big\}.$$

Then $W^{s,p}(\Omega)$ is a Banach space with respect to the norm

$$\|u\|_{W^{s,p}(\Omega)} := \left(\|u\|^p_{W^{m,p}(\Omega)} + \sum_{|\alpha|=m} \|D^\alpha u\|^p_{W^{\sigma,p}(\Omega)}\right)^{\frac{1}{p}}.$$

If $s = m$ is an integer, then $W^{s,p}(\Omega)$ coincides with the classical integral order Sobolev space $W^{m,p}(\Omega)$.

We also recall the following definition of the Besov space $B^s_{p,q}$, according to [16, Chapter V, Section 5.1, Formula (60)]:

$$B^s_{p,q}(\mathbb{R}^N) := \left\{u \in L^p(\mathbb{R}^N) : \left(\int_{\mathbb{R}^N} \frac{\|u(x+y) - u(y)\|^q_{L^p(\mathbb{R}^N)}}{|y|^{N+qs}}\, dy\right)^{\frac{1}{q}} < \infty\right\},$$
$$1 \le p, q \le \infty, 0 < s < 1. \tag{1.3}$$

Notice that, when $p = q$, we have $B^s_{p,p}(\mathbb{R}^N) = W^{s,p}(\mathbb{R}^N)$. Finally, we recall the definition of the following potential space

$$\mathscr{L}^p_{2s}(\mathbb{R}^N) := \Big\{u \in L^p(\mathbb{R}^N) : \ (-\Delta)^s u \in L^p(\mathbb{R}^N)\Big\}, \quad 1 \le p \le \infty, \quad s \ge 0, \tag{1.4}$$

introduced, for example, in [16, Chapter V, Section 3.3, Formula (38)]. Note that this same space is sometimes denoted as $H^s_p(\mathbb{R}^N)$ (see, e.g., [18, Section 1.3.2]). Here we adopt the notation $\mathscr{L}^p_{2s}(\mathbb{R}^N)$.

Let us now introduce the notion of solution that we shall consider. Following [12], we first consider weak solutions of (1.1) defined as follows.

Definition 1.1 Let $f \in L^2((0,T); W^{-s,2}(\overline{\Omega}))$. We say that $u \in L^2((0,T); W^{s,2}_0(\overline{\Omega})) \cap C([0,T]; L^2(\Omega))$ with $u_t \in L^2((0,T); W^{-s,2}(\overline{\Omega}))$ is a finite energy solution to the parabolic problem (1.1), if the identity

$$\int_0^T \int_\Omega u_t w\, dxdt + \frac{C_{N,s}}{2} \int_0^T \int_{\mathbb{R}^N} \int_{\mathbb{R}^N} \frac{(u(x) - u(y))(w(x) - w(y))}{|x - y|^{N+2s}}\, dxdydt$$
$$= \int_0^T \langle f, v\rangle_{W^{-s,2}(\overline{\Omega}), W^{s,2}_0(\overline{\Omega})}\, dt, \tag{1.5}$$

holds, for any $w \in L^2((0,T); W_0^{s,2}(\overline{\Omega}))$, where $\langle\cdot,\cdot\rangle_{W^{-s,2}(\overline{\Omega}),W_0^{s,2}(\overline{\Omega})}$ denotes the duality pairing between $W^{-s,2}(\overline{\Omega})$ and $W_0^{s,2}(\overline{\Omega})$.

Remark 1.2 We observe the following facts.

(a) According to [12, Theorem 10], if $u \in L^2((0,T); W_0^{s,2}(\overline{\Omega}))$ and $u_t \in L^2((0,T); W^{-s,2}(\overline{\Omega}))$, then $u \in C([0,T]; L^2(\Omega))$. Thus the identity $u(\cdot,0) = 0$ makes sense in $L^2(\Omega)$.

(b) When considering right hand side terms $f \in L^p((0,T); L^p(\Omega)) = L^p(\Omega \times (0,T))$ with $p \geq 2$, since we have the continuous embedding $L^p(\Omega \times (0,T)) \hookrightarrow L^2((0,T); W^{-s,2}(\overline{\Omega}))$, this notion of weak finite energy solution suffices.

(c) When $f \in L^p(\Omega \times (0,T))$ with $1 \leq p < 2$, the regularity of the right hand side term does not suffice to define weak finite energy solutions as above. We shall rather consider those defined by duality or transposition.

Duality or transposition solutions of (1.1) are given by duality with respect to the following class of test functions

$$\mathcal{P}(\Omega_T) = \Big\{\phi(\cdot,t) \in C^1((0,T), C_0^{\beta}(\Omega)) : \phi \text{ is a solution to Problem (P)}\Big\},$$

where

$$(P) = \begin{cases} -\phi_t + (-\Delta)^s\phi = \psi & \text{in } \Omega \times (0,T) =: \Omega_T, \\ \phi \equiv 0 & \text{on } (\mathbb{R}^N \setminus \Omega) \times (0,T), \\ \phi(\cdot,T) \equiv 0 & \text{in } \Omega \end{cases}$$

for $\psi \in C_0^{\infty}(\Omega_T)$.

Definition 1.3 Let $f \in L^1(\Omega \times (0,T))$. We say that $u \in C([0,T]; L^1(\Omega))$ is a weak duality or transposition solution to the parabolic problem (1.1), if the identity

$$\int_0^T \int_\Omega u\psi\, dxdt = \int_0^T \int_\Omega f\phi\, dxdt \tag{1.6}$$

holds, for any $\phi \in \mathcal{P}(\Omega_T)$ and $\psi \in C_0^{\infty}(\Omega_T)$.

Remark 1.4 The existence and uniqueness of finite energy weak solutions or the duality/transposition ones (depending on the regularity imposed on the right hand side term f) to problem (1.1) is guaranteed by Leonori et al. [12, Theorem 26] and [12, Theorem 28], respectively. If $f \in L^p(\Omega \times (0,T))$, with $p \geq 2$, finite energy solutions of (1.1) will be considered while, if $1 < p < 2$, solutions will be understood in the sense of duality/transposition. In both cases we shall refer to them as weak solutions.

Our first regularity result concerns the case $p = 2$. It reads as follows:

Theorem 1.5 *Assume $f \in L^2(\Omega \times (0,T))$ and let $u \in L^2((0,T); W_0^{s,2}(\overline{\Omega})) \cap C([0,T]; L^2(\Omega))$ with $u_t \in L^2((0,T); W^{-s,2}(\overline{\Omega}))$ be the unique finite energy solution of system (1.1). Then*

$$u \in L^2((0,T); W^{2s,2}_{\mathrm{loc}}(\Omega)) \cap L^\infty((0,T); W_0^{s,2}(\overline{\Omega})) \quad \textit{and} \quad u_t \in L^2(\Omega \times (0,T)).$$

Theorem 1.5 can be extended to the L^p-setting as follows.

Theorem 1.6 *Let $1 < p < \infty$ and $f \in L^p(\Omega \times (0,T))$. Then, problem (1.1) has a unique weak solution $u \in C([0,T]; L^p(\Omega))$ such that $u \in L^p\Big((0,T); \mathscr{L}^p_{2s,\mathrm{loc}}(\Omega)\Big)$ and $u_t \in L^p(\Omega \times (0,T))$. As a consequence we have the following result.*

(a) If $1 < p < 2$ and $s \neq 1/2$, then $u \in L^p\Big((0,T); B^{2s}_{p,2,\mathrm{loc}}(\Omega)\Big)$.

(b) If $1 < p < 2$ and $s = 1/2$, then $u \in L^p\Big((0,T); W^{2s,p}_{\mathrm{loc}}(\Omega)\Big) = L^p\Big((0,T); W^{1,p}_{\mathrm{loc}}(\Omega)\Big)$.

(c) If $2 \le p < \infty$, then $u \in L^p\Big((0,T); W^{2s,p}_{\mathrm{loc}}(\Omega)\Big)$.

In Theorem 1.6, with $(\mathscr{L}^p_{2s})_{\mathrm{loc}}(\Omega)$ we indicate the potential space

$$(\mathscr{L}^p_{2s})_{\mathrm{loc}}(\Omega) := \Big\{u \in L^p(\Omega) : u\eta \in \mathscr{L}^p_{2s}(\mathbb{R}^N) \ \text{ for any test function } \eta \in \mathcal{D}(\Omega)\Big\}. \tag{1.7}$$

Analogously, with $B^{2s}_{p,2,\mathrm{loc}}(\Omega)$ we denote the Besov space

$$B^{2s}_{p,2,\mathrm{loc}}(\Omega) := \Big\{u \in L^p(\Omega) : u\eta \in B^{2s}_{p,2}(\mathbb{R}^N) \ \text{ for any test function} \eta \in \mathcal{D}(\Omega)\Big\}. \tag{1.8}$$

Moreover, our results guarantee that when the right hand side belongs to $L^p(\Omega \times (0,T))$ for $2 \le p < \infty$ and for $1 < p < 2$, $s = 1/2$, then the corresponding solution gains locally the maximum possible regularity, that is, it gains one time derivative and up to $2s$ space derivatives, locally, in $L^p(\Omega)$. For $1 < p < 2$ and $s \neq 1/2$, instead, the local regularity is obtained in the Besov space $B^{2s}_{p,2,\mathrm{loc}}(\Omega)$, which is strictly larger than $W^{2s,p}_{\mathrm{loc}}(\Omega)$.

For the classical Laplace operator (which corresponds to the case $s = 1$), this kind of results are standard, see e.g., [3, Theorem X.12], [7, Section 9], [10, Section 4.1]. Also, we recall [11, Theorem 1] for a more general result in an abstract setting.

Theorems 1.5 and 1.6 are natural extensions of analogous results of local regularity for the elliptic problem associated to the fractional Laplacian on a bounded domain, which have been obtained recently in [1, 2].

In the recent years, research on regularity of heat equations involving non-local terms has been very active. For instance, Hölder regularity was proved in [6, 9]. Boundary regularity has also been analyzed showing that, if $f = 0$ and taking initial data $u(\cdot, 0) = u_0 \in L^2(\Omega)$, the corresponding solution to (1.1) is such that $u(\cdot, t)$ belongs to $C^s(\mathbb{R}^N)$ for all $t > 0$ and satisfies $u(\cdot, t)/\rho^s \in C^{s-\varepsilon}(\Omega)$ for any $\varepsilon > 0$, $\rho(x) = \text{dist}(x, \partial\Omega)$ being the distance to the boundary function. Concerning regularity in the Sobolev setting, we refer instead to [12, Theorem 26], where it has been proved the existence of a finite energy solution to (1.1), according to Definition 1.3 above. However, to the best of our knowledge, our Theorems 1.5 and 1.6 providing maximal space-time local regularity are new.

The controllability of parabolic equations involving non-local terms has also been investigated. We refer for instance to [5] where null controllability issues were addressed for heat equations involving non-local lower order terms. On the other hand, [13, 14] dealt with the control of heat equations involving the *spectral* fractional Laplacian (see [13, Section 1] for the definition of this operator), proving that null controllability holds for $s > 1/2$, while for $s \leq 1/2$ the equation fails to be controllable. Notice that this operator does not coincide with (1.2).

The present paper is organized as follows. In Sect. 2, we will recall the sharp local regularity results obtained in [1, 2] for the elliptic problems associated to the fractional Laplacian. These results will be necessary in the proof of Theorems 1.5 and 1.6. In Sect. 3, we give the proof of Theorem 1.5, using the corresponding result for the classical Laplace operator in [4, Section 7.1.3, Theorem 5], employing a cut-off argument and using [2, Theorem 1.2]. In Sect. 4 we give the proof of Theorem 1.6 by applying the results contained in [11]. Finally, in Sect. 5, we present some open problems and perspectives that are closely related to our work.

2 Regularity Results for the Elliptic Problem

In this section, we recall some regularity results for weak solutions to the elliptic problem associated to the fractional Laplacian on a bounded open set. These results have been recently obtained in [1, 2], and they will be fundamental in the proof of Theorems 1.5 and 1.6. Therefore, throughout this section we are going to consider the following elliptic problem

$$\begin{cases} (-\Delta)^s u = f & \text{in } \Omega, \\ u \equiv 0 & \text{on } \mathbb{R}^N \setminus \Omega. \end{cases} \tag{2.1}$$

Let us start by recalling the definition of a weak solution, according to [2, 12].

Definition 2.1 Let $f \in W^{-s,2}(\overline{\Omega})$. A function $u \in W_0^{s,2}(\overline{\Omega})$ is said to be a finite energy solution to the Dirichlet problem (2.1) if for every $v \in W_0^{s,2}(\overline{\Omega})$, the equality

$$\frac{C_{N,s}}{2} \int_{\mathbb{R}^N} \int_{\mathbb{R}^N} \frac{(u(x)-u(y))(v(x)-v(y))}{|x-y|^{N+2s}}\, dxdy = \langle f, v \rangle_{W^{-s,2}(\overline{\Omega}), W_0^{s,2}(\overline{\Omega})} \tag{2.2}$$

holds.

We notice that, when $f \in L^p(\Omega)$ with $1 < p < 2$ and it does not belong to $W^{-s,2}(\overline{\Omega})$, it is not natural to consider finite energy solutions for the problem (2.1). As for the parabolic problem above, we shall introduce an alternative notion of solution. This will be given by duality with respect to the following class of test functions:

$$\mathcal{T}(\Omega) = \Big\{ \phi : (-\Delta)^s \phi = \psi \quad \text{in } \Omega, \ \phi = 0 \quad \text{in } \mathbb{R}^N \setminus \Omega, \ \psi \in C_0^\infty(\Omega) \Big\}.$$

Definition 2.2 Let $f \in L^1(\Omega)$. We say that $u \in L^1(\Omega)$ is a weak duality or transposition solution to (2.1) if the equality

$$\int_\Omega u\psi\, dx = \int_\Omega f\phi\, dx,$$

holds for any $\phi \in \mathcal{T}(\Omega)$ and $\psi \in C_0^\infty(\Omega)$.

The existence and uniqueness of finite energy weak solutions or the duality/transposition ones (depending on the regularity imposed on the right hand side term f) to problem (2.1) are guaranteed by Leonori et al. [12, Theorem 12] and [12, Theorem 23], respectively. If $f \in L^p(\Omega)$, with $p \geq 2$, finite energy solutions of (2.1) will be considered while, if $1 < p < 2$, solutions will be understood in the sense of duality/transposition. In both cases, we shall refer to them as weak solutions. Moreover, we notice that, according to Definition 2.2, duality solutions do not require that f belongs to the dual space $W^{-s,2}(\overline{\Omega})$. Finally, we also notice that, if $f \in L^p(\Omega)$ with $p \geq 2$, we have the continuous embedding $L^p(\Omega) \hookrightarrow L^2(\Omega) \hookrightarrow W^{-s,2}(\overline{\Omega})$, meaning that the property $f \in W^{-s,2}(\overline{\Omega})$ is automatically guaranteed.

Concerning the regularity of the solutions to (2.1), the following result has been proved in [1, 2].

Theorem 2.3 (**L^p-Local Elliptic Regularity**) *Let $1 < p < \infty$. Given $f \in L^p(\Omega)$, let u be the unique weak solution to the Dirichlet problem (2.1). Then $u \in \mathscr{L}^p_{2s,\mathrm{loc}}(\Omega)$. As a consequence we have the following result.*

1. *If $1 < p < 2$ and $s \neq 1/2$, then $u \in B^{2s}_{p,2,\mathrm{loc}}(\Omega)$.*
2. *If $1 < p < 2$ and $s = 1/2$, then $u \in W^{2s,p}_{\mathrm{loc}}(\Omega) = W^{1,p}_{\mathrm{loc}}(\Omega)$.*
3. *If $2 \leq p < \infty$, then $u \in W^{2s,p}_{\mathrm{loc}}(\Omega)$.*

The proof of Theorem 2.3 requires a cut-off argument that allows us to reduce the problem to the whole space case, for which the result is already known. In particular, we have the following.

Theorem 2.4 *Let* $1 < p < \infty$. *Given* $F \in L^p(\mathbb{R}^N)$, *let* u *be the unique weak solution to the fractional Poisson type equation*

$$(-\Delta)^s u = F \quad \textit{in} \;\; \mathbb{R}^N. \tag{2.3}$$

Then $u \in \mathscr{L}^p_{2s}(\mathbb{R}^N)$. *As a consequence we have the following.*

(a) If $1 < p < 2$ *and* $s \neq 1/2$, *then* $u \in B^{2s}_{p,2}(\mathbb{R}^N)$.
(b) If $1 < p < 2$ *and* $s = 1/2$, *then* $u \in W^{2s,p}(\mathbb{R}^N) = W^{1,p}(\mathbb{R}^N)$.
(c) If $2 \leq p < \infty$, *then* $u \in W^{2s,p}(\mathbb{R}^N)$.

Theorem 2.4 is a classical result whose proof can be done by combining several results on singular integrals and Fourier transform contained in [16, Chapter V]. See also [1, 2]. In particular:

- If $1 < p < 2$ and $s \neq 1/2$, then the result follows from [16, Chapter V, Section 5.3, Theorem 5(B)], which provides the inclusion $\mathscr{L}^p_{2s}(\mathbb{R}^N) \subset B^{2s}_{p,2}(\mathbb{R}^N)$. Moreover, an explicit counterexample showing that sharper inclusions are not possible has been given in [16, Chapter V, Section 6.8].
- If $1 < p < 2$ and $s = 1/2$, then applying [16, Chapter V, Section 3.3, Theorem 3] we have $\mathscr{L}^p_{2s}(\mathbb{R}^N) = \mathscr{L}^p_1(\mathbb{R}^N) = W^{1,p}(\mathbb{R}^N)$.
- If $2 \leq p < \infty$, then [16, Chapter V, Section 5.3, Theorem 5(A)] yields $u \in B^{2s}_{p,p}(\mathbb{R}^N)$ and this latter space, by definition, coincides with $W^{2s,p}(\mathbb{R}^N)$ (see, e.g., [16, Chapter V, Section 5.1, Formula (60)]).

While developing the cut-off argument that we mentioned above, as an intermediate step we need to show that $u \in W^{s,p}(\Omega)$. Notice that, for $p \geq 2$, this is true for all weak solutions to (1.1) by classical embedding results. When $1 < p < 2$, instead, according to [12, Theorem 23], weak duality solutions to (2.1) are such that

$$(-\Delta)^{\frac{s}{2}} u \in L^p(\Omega), \quad \forall\, p \in (1, N/(N-s)) \tag{2.4}$$

an this implies that $u \in W^{s,p}(\Omega)$ too.

Proof of Theorem 2.3 For the sake of completeness we include the proof.

We start by noticing that, assuming $f \in L^p(\Omega)$, $1 < p < \infty$, we have that (2.1) has a unique weak solution u (either the finite-energy or the duality one) and that, from the discussion above, we have $u \in W^{s,p}(\Omega)$. In particular, $u \in L^p(\Omega)$.

As we have mentioned above, our strategy is based on a cut-off argument that will allow us to show that the solutions of the fractional Dirichlet problem in Ω, after cut-off, are solutions of the elliptic problem on the whole space $\mathbb{R}^N$, for which Theorem 2.4 holds. For this purpose, given ω and $\widetilde{\omega}$ two open subsets of the domain

Ω such that $\widetilde{\omega} \Subset \omega \Subset \Omega$, we introduce a cut-off function $\eta \in \mathcal{D}(\omega)$ such that

$$\begin{cases} \eta(x) \equiv 1 & \text{if } \; x \in \widetilde{\omega} \\ 0 \leq \eta(x) \leq 1 & \text{if } \; x \in \omega \setminus \widetilde{\omega} \\ \eta(x) = 0 & \text{if } \; x \in \mathbb{R}^N \setminus \omega. \end{cases} \tag{2.5}$$

Let ω and $\eta \in \mathcal{D}(\omega)$ be respectively the set and the cut-off function constructed in (2.5). We consider the function $u\eta \in W^{s,p}(\mathbb{R}^N)$ and we have that $(-\Delta)^s(u\eta)$ is given by (see, e.g., [2, Proposition 1.5] or [15])

$$(-\Delta)^s(u\eta) = \eta f + u(-\Delta)^s\eta - I_s(u,\eta), \tag{2.6}$$

where $I_s(u,\eta)$ is a remainder term which is given by

$$I_s(u,\eta)(x) := C_{N,s} \int_{\mathbb{R}^N} \frac{(u(x)-u(y))(\eta(x)-\eta(y))}{|x-y|^{N+2s}}\, dy, \quad x \in \mathbb{R}^N. \tag{2.7}$$

Let ω_1, ω_2 be open sets such that

$$\overline{\omega} \subset \omega_1 \subset \overline{\omega}_1 \subset \omega_2 \subset \overline{\omega}_2 \subset \Omega. \tag{2.8}$$

Since the function η and the set ω in (2.5) are arbitrary, it follows that $u \in W^{s,p}(\omega_2)$. Thus we have $u \in W^{s,p}(\omega_2) \cap L^p(\Omega)$. Let

$$g := u(-\Delta)^s\eta - I_s(u,\eta).$$

We now claim that $g \in L^p(\mathbb{R}^N)$ and there exists a constant $C > 0$ such that

$$\|g\|_{L^p(\mathbb{R}^N)} \leq C \left(\|u\|_{W^{s,p}(\omega_2)} + \|u\|_{L^p(\Omega)} \right). \tag{2.9}$$

Indeed, it is clear that g is defined on all $\mathbb{R}^N$. Moreover

$$\|u(-\Delta)^s\eta\|^p_{L^p(\mathbb{R}^N)} = \int_\Omega |u(-\Delta)^s\eta|^p\, dx \leq \|(-\Delta)^s\eta\|^p_{L^\infty(\Omega)} \|u\|^p_{L^p(\Omega)}. \tag{2.10}$$

For estimating the term I_s, we use the decomposition

$$\begin{aligned} I_s(u,\eta)(x) :=& C_{N,s} \int_{\mathbb{R}^N} \frac{(u(x)-u(y))(\eta(x)-\eta(y))}{|x-y|^{N+2s}}\, dy \\ =& C_{N,s} \int_{\omega_1} \frac{(u(x)-u(y))(\eta(x)-\eta(y))}{|x-y|^{N+2s}}\, dy \\ &+ C_{N,s}\eta(x) \int_{\mathbb{R}^N \setminus \omega_1} \frac{u(x)-u(y)}{|x-y|^{N+2s}}\, dy = \mathbb{I}_1(x) + \mathbb{I}_2(x), \quad x \in \mathbb{R}^N, \end{aligned}$$

where we have set

$$\mathbb{I}_1(x) := C_{N,s}\int_{\omega_1}\frac{(u(x)-u(y))(\eta(x)-\eta(y))}{|x-y|^{N+2s}}\,dy,\quad x\in\mathbb{R}^N,$$

and

$$\mathbb{I}_2(x) := C_{N,s}\eta(x)\int_{\mathbb{R}^N\setminus\omega_1}\frac{u(x)-u(y)}{|x-y|^{N+2s}}\,dy,\quad x\in\mathbb{R}^N.$$

Let $p' := p/(p-1)$. Using the Hölder inequality, we get that for a.e. $x\in\mathbb{R}^N$,

$$|\mathbb{I}_1(x)|\le C_{N,s}\left(\int_{\omega_1}\frac{|u(x)-u(y)|^p}{|x-y|^{N+sp}}\,dy\right)^{\frac{1}{p}}\left(\int_{\omega_1}\frac{|\eta(x)-\eta(y)|^{p'}}{|x-y|^{N+sp'}}\,dy\right)^{\frac{1}{p'}}. \tag{2.11}$$

Let $x\in\omega_1$ be fixed and $R>0$ such that $\omega_1\subset B(x,R)$. Using the Lipschitz continuity of the function η, we obtain that there exists constant $C>0$ such that

$$\int_{\omega_1}\frac{|\eta(x)-\eta(y)|^{p'}}{|x-y|^{N+sp'}}\,dy\le C\int_{\omega_1}\frac{dy}{|x-y|^{N+sp'-p'}}\le C\int_{B(x,R)}\frac{dy}{|x-y|^{N+sp'-p'}}\le C. \tag{2.12}$$

In what follows, we will employ the following estimate. Let $A\subset\mathbb{R}^N$ be a bounded set and $B\subset\mathbb{R}^N$ an arbitrary set. Then there exists a constant $C>0$ (depending on A and B) such that

$$|x-y|\ge C(1+|y|),\;\;\forall\,x\in A,\;\forall\,y\in\mathbb{R}^N\setminus B,\;\operatorname{dist}(A,\mathbb{R}^N\setminus B)=\delta>0. \tag{2.13}$$

Now, using (2.11), (2.12) and (2.13), we get

$$\begin{aligned}\int_{\mathbb{R}^N}|\mathbb{I}_1(x)|^p\,dx\le& C\left(\int_{\omega_2}\int_{\omega_1}\frac{|u(x)-u(y)|^p}{|x-y|^{N+sp}}\,dydx+\int_{\mathbb{R}^N\setminus\omega_2}\int_{\omega_1}\frac{|u(x)-u(y)|^p}{|x-y|^{N+sp}}\,dydx\right)\\ \le& C\left(\|u\|^p_{W^{s,p}(\omega_2)}+\int_{\mathbb{R}^N\setminus\omega_2}\int_{\omega_1}\frac{|u(x)|^p+|u(y)|^p}{(1+|x|)^{N+sp}}\,dydx\right)\\ \le& C\left(\|u\|^p_{W^{s,p}(\omega_2)}+\|u\|^p_{L^p(\Omega)}\right),\end{aligned} \tag{2.14}$$

where we have also used that $u=0$ on $\mathbb{R}^N\setminus\Omega$. Recall that $\mathbb{I}_2=0$ on $\mathbb{R}^N\setminus\omega$. Then using the Hölder inequality, we get that

$$|\mathbb{I}_2(x)|^p\le C\left(\int_{\mathbb{R}^N\setminus\omega_1}\frac{\eta^{p'}(x)dy}{|x-y|^{N+sp'}}\right)^{p-1}\int_{\mathbb{R}^N\setminus\omega_1}\frac{|u(x)-u(y)|^p}{|x-y|^{N+sp}}\,dy. \tag{2.15}$$

For any $y \in \mathbb{R}^N \setminus \omega_1$, we have that

$$\frac{\eta^{p'}(x)}{|x-y|^{N+sp'}} = \frac{\chi_{\overline{\omega}}(x)\eta^{p'}(x)}{|x-y|^{N+sp'}} \leq \chi_{\overline{\omega}}(x)\eta^{p'}(x) \sup_{x\in\overline{\omega}} \frac{1}{|x-y|^{N+sp'}}.$$

So there exists a constant $C > 0$ such that

$$\int_{\mathbb{R}^N\setminus\omega_1} \frac{\eta^{p'}(x)dy}{|x-y|^{N+sp'}} \leq \chi_{\overline{\omega}}(x)\eta^{p'}(x) \int_{\mathbb{R}^N\setminus\omega_1} \frac{dy}{\operatorname{dist}(y,\partial\overline{\omega})^{N+sp'}} \leq C\chi_{\overline{\omega}}(x)\eta^{p'}(x). \tag{2.16}$$

In (2.16) we have also used that the integral is finite which follows from the fact that $\operatorname{dist}(\partial\omega_1, \partial\overline{\omega}) \geq \delta > 0$ together with the fact that $\operatorname{dist}(y, \partial\overline{\omega})$ grows linearly as y tends to infinity and $N + sp' > N$.

Since $\chi_{\overline{\omega}}\eta^{p'} \in L^\infty(\omega)$, and using (2.15), (2.16) and (2.13), we also get that there exists a constant $C > 0$ such that

$$\begin{aligned}\int_{\mathbb{R}^N} |\mathbb{I}_2(x)|^p\,dx = \int_\omega |\mathbb{I}_2(x)|^p\,dx &\leq C\int_\omega\int_{\mathbb{R}^N\setminus\omega_1} \frac{|u(x)-u(y)|^p}{|x-y|^{N+sp}}\,dydx \\ &\leq C\int_\omega\int_{\mathbb{R}^N\setminus\omega_1} \frac{|u(x)|^p+|u(y)|^p}{(1+|y|)^{N+sp}}\,dydx \leq C\|u\|^p_{L^p(\Omega)},\end{aligned} \tag{2.17}$$

where we have used again that $u = 0$ on $\mathbb{R}^N \setminus \Omega$. Estimate (2.9) follows from (2.10), (2.14), (2.17) and we have shown the claim. We therefore proved that ηu is a weak solution to the Poisson equation (2.3) with F given by $F = \eta f + g$. Since $F \in L^p(\mathbb{R}^N)$, it follows from Theorem 2.4 that $\eta u \in \mathscr{L}^p_{2s}(\mathbb{R}^N)$. We have shown that $u \in (\mathscr{L}^p_{2s})_{\mathrm{loc}}(\Omega)$. As a consequence we have the following results.

(a) If $1 < p < 2$ and $s \neq 1/2$, then $\eta u \in B^{2s}_{p,2}(\mathbb{R}^N)$, hence $u \in B^{2s}_{p,2,\mathrm{loc}}(\Omega)$.

(b) If $1 < p < 2$ and $s = 1/2$, then $\eta u \in W^{2s,p}(\mathbb{R}^N) = W^{1,p}(\mathbb{R}^N)$, hence $u \in W^{2s,p}_{\mathrm{loc}}(\Omega) = W^{1,p}_{\mathrm{loc}}(\Omega)$.

(c) If $2 \leq p < \infty$, then $\eta u \in W^{2s,p}(\mathbb{R}^N)$, hence $u \in W^{2s,p}_{\mathrm{loc}}(\Omega)$.

The proof is finished. □

We conclude this section mentioning that Theorem 2.3 can be proved also using techniques from pseudo-differential calculus (see, e.g., [8, Section 7] or [17, Chapter XI, Theorem 2.5]). Our approach is different and provides a proof based on basic estimates of solutions of general elliptic operators.

3 Proof of Theorem 1.5

The proof of Theorem 1.5 employs a cut-off argument, as in Theorem 2.3. In particular:

- Firstly, we treat the case $\Omega = \mathbb{R}^N$, adapting the proof in [4, Section 7.1.3, Theorem 5] for the classical Laplace operator.
- The case of a general Ω is reduced to the previous one applying a cut-off argument.

3.1 The $W^{2s,2}$-Regularity on $\mathbb{R}^N$

In this Section, we prove the $W^{2s,2}$-regularity result in the case where Ω is the whole space $\mathbb{R}^N$. We will adapt the proof presented in [4, Section 7.1.3, Theorem 5] for the local case.

Theorem 3.1 *Assume $f \in L^2(\mathbb{R}^N \times (0,T))$ and let $u \in L^2((0,T);W^{s,2}(\mathbb{R}^N)) \cap C([0,T];L^2(\mathbb{R}^N))$ with $u_t \in L^2((0,T);W^{-s,2}(\mathbb{R}^N))$ be the unique finite energy solution of the system*

$$\begin{cases} u_t + (-\Delta)^s u = f & \textit{in } \mathbb{R}^N \times (0,T), \\ u(\cdot,0) \equiv 0 & \textit{on } \mathbb{R}^N. \end{cases} \tag{3.1}$$

Then

$$u \in L^2((0,T);W^{2s,2}(\mathbb{R}^N)) \cap L^\infty((0,T);W^{s,2}(\mathbb{R}^N)), \quad u_t \in L^2(\mathbb{R}^N \times (0,T)).$$

Proof First of all, we notice that the function $v := ue^{-t}$ solves the system

$$\begin{cases} v_t + (-\Delta)^s v + v = g & \text{in } \mathbb{R}^N \times [0,T], \\ v(\cdot,0) \equiv 0 & \text{on } \mathbb{R}^N, \end{cases} \tag{3.2}$$

with $g := fe^{-t} \in L^2(\mathbb{R}^N \times (0,T))$. Now, multiplying (3.2) by v_t and integrating by parts over $\mathbb{R}^N$ we obtain that

$$(v_t, v_t) + B[v, v_t] + (v, v_t) = (g, v_t),$$

where $(\cdot,\cdot)$ is the classical scalar product on $L^2(\mathbb{R}^N)$, while with $B[\cdot,\cdot]$ we indicated the bilinear form

$$B[\phi,\psi] := \frac{C_{N,s}}{2} \int_{\mathbb{R}^N} \int_{\mathbb{R}^N} \frac{(\phi(x)-\phi(y))(\psi(x)-\psi(y))}{|x-y|^{N+2s}}\, dxdy.$$

Moreover, we observe that

$$B[v, v_t] = \frac{1}{2}\frac{d}{dt}B[v, v] \text{ and } (v, v_t) = \frac{1}{2}\frac{d}{dt}(v, v).$$

Hence, using Young's inequality we have that, for every $\varepsilon > 0$,

$$\|v_t\|^2_{L^2(\mathbb{R}^N)} + \frac{1}{2}\frac{d}{dt}\Big(B[v, v] + (v, v)\Big) = (g, v_t) \le \frac{C}{\varepsilon}\|g\|^2_{L^2(\mathbb{R}^N)} + \varepsilon\|v_t\|^2_{L^2(\mathbb{R}^N)}.$$

Choosing $\varepsilon \le 1$ and integrating in time we find that

$$\int_0^T \|v_t\|^2_{L^2(\mathbb{R}^N)}\,dt + \sup_{t\in[0,T]}\Big(B[v(t), v(t)] + (v(t), v(t))\Big) \le C\int_0^T \|g\|^2_{L^2(\mathbb{R}^N)}\,dt,$$

which implies that

$$\|v_t\|^2_{L^2(\mathbb{R}^N\times(0,T))} + \|v\|_{L^\infty((0,T),W^{s,2}(\mathbb{R}^N))} \le C\|g\|^2_{L^2(\mathbb{R}^N\times(0,T))}.$$

Therefore,

$$v \in L^\infty((0, T); W^{s,2}(\mathbb{R}^N)), \quad v_t \in L^2(\mathbb{R}^N \times (0, T))$$

and, by definition, u has the same regularity too. Finally, the $W^{2s,2}$ regularity for u in the space variable is obtained in the following way. From (3.1) we have that $(-\Delta)^s u = f - u_t \in L^2(\mathbb{R}^N \times (0, T))$. Hence, a. e. $t \in (0, T)$, we have that $(-\Delta)^s u(\cdot, t) = h(\cdot, t) \in L^2(\mathbb{R}^N)$ and, applying the regularity results for the elliptic case (see Theorem 2.4) we get that $u(\cdot, t) \in W^{2s,2}(\mathbb{R}^N)$ a. e. $t \in (0, T)$. Furthermore $u \in L^2((0, T); W^{2s,2}(\mathbb{R}^N))$ and the proof is finished. □

3.2 The $W^{2s,2}_{loc}$-Regularity in Ω

Proof of Theorem 1.5 As we have mentioned above, our strategy is based on a cut-off argument that will allow us to show that solutions of the fractional parabolic problem in Ω, after cut-off, are solutions of a problem on the whole space $\mathbb{R}^N$, for which Theorem 3.1 holds.

Let $f \in L^2(\Omega \times (0, T))$ and $u \in L^2((0, T); W^{s,2}_0(\overline{\Omega})) \cap C([0, T]; L^2(\Omega))$ with $u_t \in L^2((0, T); W^{-s,2}(\overline{\Omega}))$ be the unique finite energy solution to the system (1.1).

Let ω and $\eta \in \mathcal{D}(\omega)$ be respectively the set and the cut-off function constructed in (2.5). We consider the function $v := u\eta$ and we write the equation satisfied by v. Recall from (2.6) that the fractional Laplacian of v is given by

$$(-\Delta)^s v = (-\Delta)^s(u\eta) = u(-\Delta)^s\eta + \eta(-\Delta)^s u - I_s(u, \eta),$$

where the remainder term I_s has been defined in (2.7). Then, v is a solution to the following problem on $\mathbb{R}^N$:

$$\begin{cases} v_t + (-\Delta)^s v = F & \text{in } \mathbb{R}^N \times (0, T), \\ v(\cdot, 0) \equiv 0 & \text{on } \mathbb{R}^N, \end{cases} \tag{3.3}$$

with $F = \eta f + u(-\Delta)^s \eta - I_s(u, \eta)$.

Following the proof of [2, Theorem 1.2], we can show that $F \in L^2(\mathbb{R}^N \times (0, T))$. Hence, from Theorem 3.1 we obtain that

$$v \in L^2((0, T); W^{2s,2}(\mathbb{R}^N)) \cap L^\infty((0, T); W^{s,2}(\mathbb{R}^N)), \quad v_t \in L^2(\mathbb{R}^N \times (0, T)).$$

This implies that $u \in L^2((0, T); W^{2s,2}_{\text{loc}}(\Omega)) \cap L^\infty((0, T); W^{s,2}_0(\overline{\Omega}))$ and $u_t \in L^2(\Omega \times (0, T))$. The proof is finished. □

4 Proof of Theorem 1.6

In this section, we prove the local regularity for the solutions to the parabolic problem (1.1), corresponding to a right hand side $f \in L^p(\Omega \times (0, T))$, with $1 < p < \infty$.

First of all, notice that the following discussion also applies to the case $p = 2$. This special case has already been treated in the previous section, and there the proof of our local regularity Theorem 1.5 has been developed taking advantage of the Hilbert structure of the spaces $L^2(\Omega)$ and $L^2(\Omega \times (0, T))$.

Clearly that strategy cannot be extended to the general L^p setting, and we have to adopt a different approach. This approach relies on an abstract result due to Lamberton [11]. In particular, the proof of Theorem 1.6 will be a direct consequence of [11, Theorem 1]. For the sake of completeness, we recall its statement here.

Theorem 4.1 *Let (Ω, Σ, m) be a measure space and let A be the generator of a strongly continuous semigroup of linear operators $(\mathbb{T}_t)_{t \geq 0}$ on $L^2(\Omega, \Sigma, m)$ satisfying the following hypothesis:*

(a) The semigroup $(\mathbb{T}_t)_{t \geq 0}$ is analytic and bounded on $L^2(\Omega, \Sigma, m)$.
(b) For every $p \in [1, \infty]$ and $\phi \in L^p(\Omega) \cap L^2(\Omega)$ we have the estimate

$$\|\mathbb{T}_t \phi\|_{L^p(\Omega)} \leq \|\phi\|_{L^p(\Omega)}, \textit{ for all } t \geq 0.$$

Let $p \in (1, \infty)$. If $f \in L^p(\Omega \times (0, T))$, then the system

$$\begin{cases} u_t - Au = f, & t \in (0, T) \\ u(0) = 0 \end{cases}$$

admits a solution $u \in C([0,T]; L^p(\Omega))$, *such that* $u_t,\ Au \in L^p(\Omega \times (0,T))$.

Proof of Theorem 1.6 First of all notice that the operator $A = -(-\Delta)^s$ with domain

$$\mathcal{D}(A) = \left\{ u \in W_0^{s,2}(\overline{\Omega}), \quad (-\Delta)^s u \in L^2(\Omega) \right\} \tag{4.1}$$

is the generator of a submarkovian strongly continuous semigroup $(\mathbb{T}_t)_{t\geq 0}$ which is also ultracontractive (see, e.g., [2, Lemma 2.4]). Let $f \in L^p(\Omega \times (0,T))$ and let u be the corresponding weak solution to the system (1.1). Then, it follows from Theorem 4.1 that $u_t, (-\Delta)^s u \in L^p(\Omega \times (0,T))$. In particular we have that $(-\Delta)^s u(\cdot,t) = (f - u_t)(\cdot,t) \in L^p(\Omega)$ a. e. $t \in (0,T)$ and, according to Theorem 2.3, this implies that $u(\cdot,t) \in \mathscr{L}^p_{2s,\mathrm{loc}}(\Omega)$ a. e. $t \in (0,T)$. Therefore, for all $t \in (0,T)$ we have the following results.

(i) If $1 < p < 2$ and $s \neq 1/2$, then $u(\cdot,t) \in B^{2s}_{p,2,\mathrm{loc}}(\Omega)$, a. e. $t \in (0,T)$.
(ii) If $1 < p < 2$ and $s = 1/2$, then $u(\cdot,t) \in W^{2s,p}_{\mathrm{loc}}(\Omega) = W^{1,p}_{\mathrm{loc}}(\Omega)$, a. e. $t \in (0,T)$.
(iii) If $2 \leq p < \infty$, then $u(\cdot,t) \in W^{2s,p}_{\mathrm{loc}}(\Omega)$, a. e. $t \in (0,T)$.

Consequently:

(a) If $1 < p < 2$ and $s \neq 1/2$, then $u \in L^p((0,T); B^{2s}_{p,2,\mathrm{loc}}(\Omega))$.
(b) If $1 < p < 2$ and $s = 1/2$, then $u \in L^p((0,T); W^{2s,p}_{\mathrm{loc}}(\Omega)) = L^p((0,T); W^{1,p}_{\mathrm{loc}}(\Omega))$.
(c) If $2 \leq p < \infty$, then $u \in L^p((0,T); W^{2s,p}_{\mathrm{loc}}(\Omega))$.

The proof of the theorem is finished. □

We conclude this section with the following remark.

Remark 4.2 Recall that we have said in the proof of Theorem 1.6 that the operator $A = -(-\Delta)^s$ with domain given by (4.1) generates a strongly continuous submarkovian semigroup $(\mathbb{T}_t)_{t\geq 0}$ on $L^2(\Omega)$ and the semigroup is analytic and ultracontractive. This implies that the semigroup can be extended to contraction semigroups on $L^p(\Omega)$ for all $p \in [1,\infty]$ and each semigroup is strongly continuous if $p \in [1,\infty)$ and bounded analytic if $p \in (1,\infty)$. Let A_p denote the generator of the semigroup on $L^p(\Omega)$ for $p \in [1,\infty]$ so that A_2 coincides with A. By Theorem 4.1 if $1 < p < \infty$ and $f \in L^p(\Omega \times (0,T))$, then the unique solution $u \in C([0,T]; L^p(\Omega))$ of the system (1.1) has the following regularity:

$$u \in L^p((0,T); D(A_p)).$$

This trivially implies that $A_p u \in L^p(\Omega \times (0,T))$ and $u_t \in L^p(\Omega \times (0,T))$. Our contribution in the present paper was to show that $D(A_p) \subset \mathscr{L}^p_{2s,\mathrm{loc}}(\Omega)$ for every $1 < p < \infty$.

5 Open Problems and Perspectives

In the present paper we proved that weak solutions to the parabolic problem for the fractional Laplacian, with a non-homogeneous right-hand side $f \in L^p(\Omega \times (0,T))$ $(1 < p < \infty)$ and zero initial datum, belong to $L^p((0,T); \mathscr{L}^p_{2s,\mathrm{loc}}(\Omega))$. The following comments are worth considering.

(a) A natural interesting extension of our result would be the analysis of the global maximal regularity in space for weak solutions to (1.1). The problem is delicate however.

Indeed, already at the elliptic level, we know that even if Ω has a smooth boundary, then the global maximal regularity up to the boundary does not hold. To be more precise, assume that Ω has a smooth boundary, $f \in L^p(\Omega)$ $(1 < p < \infty)$ and let u be the associated weak solution to the Dirichlet problem (2.1). It is known that, if $p \geq 2$, then u does not always belongs to $W^{2s,p}(\Omega)$ and, if $1 < p < 2$, then u does not always belong to $B^{2s}_{p,2}(\Omega)$. This shows that in general, the corresponding weak solution u to the parabolic system (1.1) does not always belong to $L^p((0,T); W^{2s,p}(\Omega))$ if $p \geq 2$ and does not always belong to $L^p((0,T); B^{2s}_{p,2}(\Omega))$ if $1 < p < 2$.

On the one hand, Theorem 4.1 shows that $u \in L^p((0,T); D(A_p))$, that is, in particular $u(\cdot,t) \in D(A_p)$ for a.e. $t \in (0,T)$. On the other hand, according to the discussions given in [2, Section 5], at least if Ω has a sufficiently smooth boundary, one has that $u(\cdot,t) = \rho^s v(\cdot,t)$ where $v(\cdot,t)$ is a regular function up to the boundary. Here, $\rho(x) := \mathrm{dist}(x, \partial\Omega)$ for $x \in \Omega$. In addition one could expect that $\rho^{-s}u, \rho^{1-s}u \in L^p((0,T); L^p((0,T); W^{2s,p}(\Omega))$ if $2 \leq p < \infty$, and $\rho^{-s}u, \rho^{1-s}u \in L^p((0,T); L^p((0,T); B^{2s}_{p,2}(\Omega))$ if $1 < p < 2$. This constitutes an interesting open problem. We refer to [2, Section 5] for a more complete discussion on related topics and the difficulties that it raises.

(b) It would be interesting to consider the case of a non-zero initial datum in Eq. (1.1). In the Hilbert space framework, i.e. when working in the $L^2(\Omega)$ setting, the strategy of Sect. 3 can be extended to deal with initial data in $W^{s,2}_0(\overline{\Omega})$. To the best of our knowledge, the corresponding analogous result in the case $p \in (1,\infty)$, $p \neq 2$, is still unknown.

Acknowledgments The work of Umberto Biccari was partially supported by the Advanced Grant DYCON (Dynamic Control) of the European Research Council Executive Agency, by the MTM2014-52347 and MTM2017-92996 Grants of the MINECO (Spain) and by the Air Force Office of Scientific Research under the Award No: FA9550-15-1-0027. The work of Mahamadi Warma was partially supported by the Air Force Office of Scientific Research under the Award No: FA9550-15-1-0027. The work of Enrique Zuazua was partially supported by the Advanced Grant DYCON (Dynamic Control) of the European Research Council Executive Agency, FA9550-15-1-0027 of AFOSR, FA9550-14-1-0214 of the EOARD-AFOSR, the MTM2014-52347 and MTM2017-92996 Grants of the MINECO (Spain) and ICON of the French ANR

References

1. Biccari, U., Warma, M., Zuazua, E.: Addendum: local elliptic regularity for the Dirichlet fractional Laplacian. Adv. Nonlinear Stud. **17**, 837–839 (2017)
2. Biccari, U., Warma, M., Zuazua, E.: Local elliptic regularity for the Dirichlet fractional Laplacian. Adv. Nonlinear Stud. **17**, 387–409 (2017)
3. Brezis, H.: Functional Analysis, Sobolev Spaces and Partial Differential Equations. Springer, New York (2010)
4. Evans, L.C.: Partial Differential Equations. American Mathematical Society, Providence (2010)
5. Fernández-Cara, E., Lü, Q., Zuazua, E.: Null controllability of linear heat and wave equations with nonlocal in space terms. SIAM J. Control Optim. **54**, 2009–2019 (2016)
6. Fernández-Real, X., Ros-Oton, X.: Boundary regularity for the fractional heat equation. Rev. R. Acad. Cienc. Exactas Fís. Nat. Ser. A Math. RACSAM **110**, 49–64 (2016)
7. Grisvard, P.: Équations différentielles abstraites. Ann. Sci. École Norm. Sup. **2**, 311–395 (1969)
8. Grubb, G.: Fractional Laplacians on domains, a development of Hörmander's theory of μ-transmission pseudodifferential operators. Adv. Math. **268**, 478–528 (2015)
9. Kassmann, M., Schwab, R.W.: Regularity results for nonlocal parabolic equations. Riv. Mat. Univ. Parma. **5**(1), 183–212 (2014)
10. Ladyzhenskaya, O.A., Solonnikov, V.A., Uraltseva, N.N.: Linear and Quasi-Linear Equations of Parabolic Type. American Mathematical Society, Providence (1968)
11. Lamberton, D.: Equations d'évolution linéaires associées à des semi-groupes de contractions dans les espaces L^p. J. Funct. Anal. **72**, 252–262 (1987)
12. Leonori, T., Peral, I., Primo, A., Soria, F.: Basic estimates for solutions of a class of nonlocal elliptic and parabolic equations. Discrete Contin. Dyn. Syst. **35**, 6031–6068 (2015)
13. Micu, S., Zuazua, E.: On the controllability of a fractional order parabolic equation. SIAM J. Control Optim. **44**, 1950–1972 (2006)
14. Miller, L.: On the controllability of anomalous diffusions generated by the fractional Laplacian. Math. Control Signals Syst. **18**, 260–271 (2006)
15. Ros-Oton, X., Serra, J.: The Dirichlet problem for the fractional Laplacian: regularity up to the boundary. J. Math. Pures Appl. **101**, 275–302 (2014)
16. Stein, E.: Singular Integrals and Differentiability Properties of Functions. Princeton Mathematical Series. Princeton University Press, Princeton (1970)
17. Taylor, M.E.: Pseudodifferential Operators. Princeton Mathematical Series, vol. 4. Princeton University Press, Princeton (1981)
18. Triebel, H.: Theory of Function Spaces II. Monographs in Mathematics, vol. 84. Birkhäuser, Basel (1992)