Jindřich Nečas Center for Mathematical Modeling
Lecture notes

Volume 7

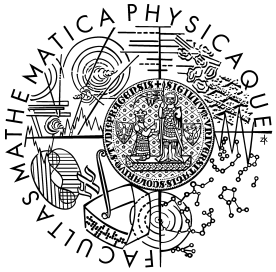
Topics in mathematical modeling and analysis

Volume edited by P. Kaplický

matfyzpress
VYDAVATELSTVÍ MATEMATICKO-FYZIKÁLNÍ FAKULTY
UNIVERZITY KARLOVY V PRAZE

Jindřich Nečas Center for Mathematical Modeling
Lecture notes

Volume 7



matfyzpress
VYDAVATELSTVÍ
MATEMATICKO-FYZIKÁLNÍ FAKULTY
UNIVERZITY KARLOVY V PRAZE

Jindřich Nečas Center for Mathematical Modeling
Lecture notes

Topics in mathematical modeling and analysis

Chérif Amrouche
Laboratoire de Mathématiques et de leurs Applications
Bâtiment IPRA - Université de Pau et des Pays de l'Adour
Avenue de l'Université - BP 1155
64013 PAU CEDEX
France

David E. Edmunds
Department of Mathematics
University of Sussex
Brighton BN1 9QH
United Kingdom

Enrique Fernández-Cara
Departamento Ecuaciones Diferenciales y Análisis Numérico
Universidad de Sevilla
Tarfia s/n, 41012 Sevilla
Spain

Anna Hundertmark-Zaušková
Institute of Mathematics
Johannes Gutenberg University
Staudingerweg 9, 55099 Mainz
Germany

Ron Kerman
Department of Mathematics
Brock University
500 Glenridge Avenue
St. Catharines, ON
L2S 3A1, Canada

Mária Lukáčová-Medviďová
Institute of Mathematics
Johannes Gutenberg University
Staudingerweg 9, 55099 Mainz
Germany

Gabriela Rusnáková
Institute of Mathematics
Pavol Jozef Šafárik University
Jesenná 5, 040 01 Košice
Slovakia

Arghir Zarnescu
Mathematical Institute
University of Oxford
24-29 St. Giles'
Oxford OX1 3LB
United Kingdom

Volume edited by P. Kaplický

2000 *Mathematics Subject Classification. 35-06*

Key words and phrases. weighted Sobolev spaces, Orlicz spaces, generalized trigonometric functions, optimal control, fluid-structure interaction, liquid crystals

Abstract. The text provides a record of lectures given by the visitors of the Jindřich Nečas Center for Mathematical Modeling during academic year 2010/2011.

ISBN 978-80-7378-196-5

Preface

This volume consists of six contributions based on the minicourses delivered by Chérif Amrouche (Pau, France), David E. Edmunds (Sussex, United Kingdom), Enrique Fernández-Cara (Seville, Spain), Ron Kerman (St. Catharines, Canada), Mária Lukáčová-Medviďová[1] (Mainz, Germany) and Arghir Zarnescu (Oxford, United Kingdom) during their visits of the Jindřich Nečas Center for Mathematical Modeling in the years 2010 and 2011.

The topics presented in their contributions document broad extent of research activities of the Nečas Center; particular themes range from real analysis and function spaces on one end to modeling of anisotropic liquids on the other end, whereas the central topic—the development of the mathematical theory for classes of nonlinear partial differential equations—is a common feature to all of them.

Understanding optimal properties of function spaces and their role in the analysis of elliptic PDEs are the shared characteristics of the contributions *Weighted Sobolev spaces and elliptic problems in half-space* by Ch. Amrouche and *Analysis in Orlicz spaces* by R. Kerman. A generalization of classical and powerful spectral theory in the Hilbert spaces setting to the setting given by reflexive Banach spaces with strictly convex duals and a link of such a generalized framework to the eigenvalue problem for p-Laplacian is the central theme of the paper *Generalised trigonometric functions, compact operators and the p-Laplacian* by D. E. Edmunds. The other two papers—*The control of PDEs: some basic concepts, recent results and open problems* by E. Fernández-Cara and *Fluid-structure interaction for shear-dependent non-Newtonian fluids* by A. Hundertmark-Zaušková, M. Lukáčová-Medviďová and G. Rusnáková—focus on two important areas: control theory and fluid-solid interactions. The analysis of problems generated in these fields is connected with tremendous amount of challenging questions. Finally, in the paper *Topics in the Q-tensor theory of liquid crystals*, A. Zarnescu concentrates on those aspects of the mathematical theory for flows of anisotropic fluids in which tensorial features considered in the model are significant and physically relevant.

We would like to use this opportunity and thank all contributors for their enthusiasm both in delivering the lectures and in the preparation of lecture notes that nicely survey the current state of knowledge in their research fields and are plentiful sources of challenging open problems.

Since this is the last of seven volumes whose production was funded by the project *LC06052 Jindřich Nečas Center for Mathematical Modeling*, we wish to summarize main accomplishments achieved within the Center.

[1] Mária Lukáčová-Medviďová wrote her paper jointly with Anna Hundertmark-Zaušková and Gabriela Rusnáková.

The project LC06052, financed by the Ministry of Education, Youth and Sports, started its activities in March 2006 and ended in December 2011. During this almost six years long period, the program of the Center has brought several innovative manners to the mathematical culture in the Czech Republic. Let us recall that after the massive emigration of mathematicians and other scientists during the last decade of the 20th century (emigration is meant in the generalized sense and include those researchers in mathematics who worked abroad regularly for a significant period of the year or for several years in length) and after the leave of plenty of high-level mathematicians-septuagenarians at the beginning of the 21th century, the Nečas Center came with a so far rare program of inviting internationally recognized scholars for month stays at our institutions and for delivering the series of lecturers on topics that have not been well developed in the Czech mathematical community. In addition, the Center opened six positions for young researchers that has been filled either by gifted postdocs from abroad (selected by the Steering Committee) or by the best Ph.D. graduates from domestic institutions. The main advantage of this kind of position in contrast to the position of Assistant Professor consists in the reduced amount of teaching duties and more time for doing research.

During its six years of existence, the Nečas Center has been internationally recognized as a quickly developing and cultivating research medium, setting up three to five specialized lectures or seminars weekly, completing classical as well as recent textbooks and monographs to the libraries shells, developing and administrating computer cluster, publishing its own Lecture Notes, organizing periodically international scientific schools (Mathematical Theory in Fluid Mechanics, EVEQ, Spring schools on Function spaces, Czech–Japanese Seminars on Applied Mathematics), specialized workshops (such as Workshop on Geomaterials, Workshops on Scientific Computing, Workshop On Fluid Structure Interaction Problems, Workshop On Analysis of Evolutionary PDEs in Fluids, Workshop Spaces Between Us, Workshop Optimization with PDE Constraints, Colloquium Praha–Bratislava, Workshop Analysis of Multiphase Problems, Workshop Multicomponent and Multiphase Materials, etc.) and many miniworkshops and minisymposia. The Center regularly and spontaneously organized many research meetings involving appropriate visiting scholars, researchers of the Center and Ph.D. and M.S. students.

To be more specific, the list of the Nečas Center visitors invited to the position of Senior Lecturer includes Herbert Amann (University of Zurich), Chérif Amrouche (Pau University), Dorin Bucur (Université de Savoie), Andrea Cianchi (Universitá di Firenze), David E. Edmunds (Sussex University), Reinhardt Farwig (Technische Universität Darmstadt), Enrique Fernández-Cara (Sevilla University), Jens Frehse (Universität Bonn), Maurizio Grasselli (Politecnico di Milano), Martin Hairer (University of Warwick), Matthias Hieber (Technische Universität Darmstadt), Hishida Toshiaki (Nagoya University), Robert Holub (Colorado School of Mines, Golden), Willi Jäger (Heidelberg University), Ron Kerman (Brook University), Masato Kimura (Kyushu University Fukuoka), Dietmar Kröner (Albert-Ludwigs-Universität Freiburg), Philippe Laurençot (Université Paul Sabatier, Toulouse), Roger Lewandowski (Université De Rennes), Gert Lube (Universität Göttingen), Mária Lukáčová-Medviďová (Universität Mainz), Gerard Meurant (Paris), Alexander Mielke (Weierstrass Institute for Applied Analysis and Stochastics Berlin), Karol Mikula (Slovenská technická univerzita Bratislava), Tetsuro Myiakawa

(Kanazawa University), Antonín Novotný (University of Toulon), Patrick Rabier (University of Pittsburg), Hans-Görg Roos (Dresden University), Roman Shvydkoy (University of Illinois, Chicago), Endre Süli (Oxford Univerzity), Daniel Ševčovič (Univerzita Komenského Bratislava), Peter Takáč (Rostock University), Werner Varnhorn (Universität Kassel), Wolfgang Wendland (Universität Stuttgart), Jörg Wolf (Humboldt Universität Berlin), Shigetoshi Yazaki (University of Miyazaki) and Ping Zhang (Chinese Academy of Sciences).

The list of foreign postdocs who took the position of Junior Researchers in the Center include, among many others, Helmut Abels (currently full professor at University of Regensburg), Tomasz Cieslak, Lars Diening (currently full professor in München), Luisa Consiglieri, Pierre-Etienne Druet, Jan Schneider, Yutaka Terasawa, Guiseppe Tomasetti, Riccarda Rossi, Vladislav Mantič, Sören Bartels, Adrian Muntean, Mathieu Hillaret, Mark Steinhauer (currrently professor in Koblenz), Yongzhong Sun, Timofei Silkhin, Julia Namlyeyeva, Trygve Karper, Marita Thomas, Daniel Wachsmut, among others. At least for the period of several months these positions were gradually occupied by Jaroslav Hron, Miroslav Bulíček, Iveta Hnětýnková, Václav Kučera, Jiří Mikyška, Tomáš Oberhuber, Jan Stebel, Vít Průša, Martin Lanzendörfer, Miloslav Vlasák and Ondřej Souček.

The activities of the Nečas Center have been in particular an inspiration for a young generation of Polish mathematicians (P. Gwiazda, P. Mucha and A. Świerczewska-Gwiazda), the Center has strengthened the scientific as well as educational cooperation between Prague mathematical school and schools in Heidelberg and Darmstadt, the own existence of the Center further enhanced joint research programs with colleagues in Berlin, Regensburg, Paris, Toulon, Toulouse, Rome, Firenze, etc., and with members of Japanese and Chinese mathematical and engineering circles. Just to illustrate the atmosphere, after spending a month in the Nečas Center, Endré Süli writes: "I have been hugely impressed by the very high standard of scientific life in the Nečas Center. The existence of this center is vital for the future of Applied Mathematics in the Czech Republic and in particular in training future generations of excellent M.Sc. and Ph.D. students. Given the importance of the field of Applied Mathematics in a wide range of application areas, including industry, economics, the life sciences, physical, chemical and engineering sciences, and the Nečas Center's clear commitment to the understanding of difficult mathematical problems with direct relevance to applications, the continued existence of the Nečas Center seems paramount to me."

We wish to thank all visitors, to the members of the Steering Committee as well as particular members of research teams who contributed to whatever activities of the Center. In particular, we wish to underline the names of those who help us, in an extraordinary way, in advertising the Center abroad and in forming research directions. These are Willi Jäger, Antonín Novotný, K. R. Rajagopal, Zdeněk Strakoš and Daniel Ševčovič.

The Nečas Center, center for fundamental research, was focused on six followed-up, nevertheless different main goals that share a common feature, namely, a goal-oriented mergence of modern theoretical, numerical and computer approaches towards a solution of problems in particular thematic topics. The first goal was focused on problems describing mechanical processes in a single continuum, the

second and third goals consisted of solving problems involving in addition both thermal and chemical processes. The fourth goal concentrated on the theory of interacting continua (mixture theory, multiphase and multicomponent materials, role of boundary conditions). The fifth goal was to investigate the links between full models and their approximations (based on nondimensional analysis and model reduction). The last goal then integrated all the previous goals in a synthetic manner and culminated, in particular, by the preparation of several multidisciplinary projects based on these broad grounds. The output of the research is formed mostly by scientific papers and developed software. More than 200 papers supported by the project LC06052 were written in cooperation between members of various research units or in cooperation with visiting foreign guests.

Tremendous amount of activities and scientific results obtained in the Center, and its fundamental role in the growth of a new promising generation of young researchers suggest to evaluate this project as very successful. How successful the project of the Nečas Center was will be however fully evident much later. We still hope that we will be capable of establishing the Nečas Center as a research unit connecting groups in the Academy of Sciences, Czech Technical University, Charles University and others with similar visions and complementary expertise. In such a center—that may exist independently of any particular project—we wish to carry on and further enhance our experience built in the Center following the aim to establish a multidisciplinary mathematical environment opened towards cooperation with researchers in the areas of natural and life sciences.

Prague, 31 January 2012
Michal Beneš
Eduard Feireisl
Josef Málek

Contents

Part 1

Weighted Sobolev spaces and elliptic problems in the half-space

Chérif Amrouche

2000 *Mathematics Subject Classification.* 35D05, 35D10, 35J05, 35J25, 35J55, 76D03, 76D07

Key words and phrases. Laplace equation, biharmonic problem, Stokes problem, weighted Sobolev spaces, traces, reflection principle, half-space

Abstract. In this paper, we study the Laplace equation in the whole space and we give some properties concerning the functions which belong to weighted Sobolev spaces and their traces. We prove in L^p theory, with $1 < p < \infty$ some existence and uniqueness results concerning the Laplace equation, the biharmonic problem and the Stokes problem in the half-space $\mathbb{R}^N_+$, with $N \geqslant 2$ and with Dirichlet boundary conditions or with Navier boundary conditions in the case of Stokes problem. Finally, we investigate the case of data in L^1.

Contents

CHAPTER 1

Weighted Sobolev spaces and elliptic problems in the half-space

1. Introduction

We will study here the following elliptic problems: the Laplace equation with Dirichlet or Neumann boundary condition:

$$(L)\ \begin{cases} -\Delta u = f & \text{in } \mathbb{R}^N_+ = \{x \in \mathbb{R}^N;\ x_N > 0\}, \\ u = g_0 \quad \text{or} \quad \partial_N u = g_1 & \text{on } \Gamma \equiv \mathbb{R}^{N-1} \times \{0\}, \end{cases}$$

the biharmonic problem with Dirichlet boundary condition:

$$(B)\ \begin{cases} \Delta^2 u = f & \text{in } \mathbb{R}^N_+, \\ u = g_0 \quad \text{and} \quad \partial_N u = g_1 & \text{on } \Gamma, \end{cases}$$

the Stokes system with Dirichlet or Navier boundary condtion

$$(S)\ \begin{cases} -\Delta \boldsymbol{u} + \nabla \pi = \boldsymbol{f} \quad \text{and} \quad \operatorname{div} \boldsymbol{u} = h & \text{in } \mathbb{R}^N_+, \\ \boldsymbol{u} = \boldsymbol{g} \ \text{ or } \ u_N = g_N \text{ and } \ \partial_N \boldsymbol{u}' = \boldsymbol{g}' & \text{on } \Gamma. \end{cases}$$

We are also interested in Laplace's system with right hand side in L^1:

$$(\mathbf{L})\ \begin{cases} -\Delta \boldsymbol{u} = \boldsymbol{f} & \text{in } \mathbb{R}^N_+, \\ \boldsymbol{u} = \boldsymbol{g} \ \text{ or } \ \partial_N \boldsymbol{u} = \boldsymbol{g} & \text{on } \Gamma. \end{cases}$$

Question 1 : What is the functional setting to use Lax-Milgram Lemma for the problem:

$$-\Delta u = f \text{ in } \mathbb{R}^N_+,\ u = g \ \text{ on } \Gamma \quad ?$$

Considering the case $g = 0$, it is then clear that the solution must satisfy:

$$\nabla u \in L^2(\mathbb{R}^N_+).$$

We observe that the solution does not belong to $L^2(\mathbb{R}^N_+)$, hence the classical Sobolev space $H^1(\mathbb{R}^N_+)$ is not suited. Also, the trace of function such that $\nabla u \in L^2(\mathbb{R}^N_+)$ is not in $H^{1/2}(\mathbb{R}^{N-1})$.

Question 2 : What about the case $f \in L^1(\mathbb{R}^N_+)$ or $g \in L^1(\Gamma)$?

The Stokes problem in the half-space was studied by [4, 8, 10, 14, 16, 17, 19, 20, 21]. More recently, Laplace's equation, with estimates for L^1 vector fields was studied by [5, 11, 12, 13].

2. Laplace equation in $\mathbb{R}^N$

We will consider the following Laplace equation (see [2] for more details):

$$-\Delta u = f \quad \text{in } \mathbb{R}^N. \tag{1.1}$$

Variational Formulation: Find $u \in V$ such that

$$\forall v \in V, \quad \int_{\mathbb{R}^N} \nabla u \cdot \nabla v = < f, v >,$$

where

$$V = \{v \in \mathcal{D}'(\mathbb{R}^N);\ \nabla v \in L^2(\mathbb{R}^N)\}.$$

But we must bring V with an Hilbertian space structure.

LEMMA 1.1. *Let $1 < p < \infty$ and $v \in \mathcal{D}'(\mathbb{R}^N)$ such that $\nabla v \in L^p(\mathbb{R}^N)$.*

(i) If $p < N$, then there exists a unique constant K such that $\frac{v+K}{|x|} \in L^p(\mathbb{R}^N)$, with the estimate

$$\|\frac{v+K}{|x|}\|_{L^p(\mathbb{R}^N)} \leq C\|\nabla v\|_{L^p(\mathbb{R}^N)}$$

(ii) If $p > N$, then $\frac{v}{|x|} \in L^p(\mathbb{R}^N)$, with the estimate

$$\inf_{\mu\in\mathbb{R}}\|\frac{v+\mu}{|x|}\|_{L^p(\mathbb{R}^N)} \leq C\|\nabla v\|_{L^p(\mathbb{R}^N)}$$

(iii) If $p = N$, then $\frac{v}{|x|\ln(1+|x|)} \in L^p(\mathbb{R}^N)$, with the estimate

$$\inf_{\mu\in\mathbb{R}}\|\frac{v+\mu}{|x|\ln(1+|x|)}\|_{L^p(\mathbb{R}^N)} \leq C\|\nabla v\|_{L^p(\mathbb{R}^N)}$$

where $C > 0$ is a constant depending only on p and N.

PROOF. Because, we are interested here mainly in the behavior at the infinity, we will replace $|x| = r$ by $1 + r$. We set

$$W_0^{1,p}(\mathbb{R}^N) = \{u \in \mathcal{D}'(\mathbb{R}^N), \frac{u}{w_0} \in L^p(\mathbb{R}^N), \nabla u \in \mathbf{L}^p(\mathbb{R}^N)\},$$

where

$$w_0 = 1 + r \quad \text{if } p \neq N, \quad w_0 = (1 + r)\ \ln(2 + r) \quad \text{if } p = N.$$

We use then the Hardy inequalities: for any $v \in W_0^{1,p}(\mathbb{R}^N)$,

$$\|\frac{v}{w_0}\|_{L^p(\mathbb{R}^N)} \leq C\|\nabla v\|_{L^p(\mathbb{R}^N)} \quad \text{if} \quad p < N,$$

and

$$\mathrm{Inf}_{\mu\in\mathbb{R}}\|\frac{v+\mu}{w_0}\|_{L^p(\mathbb{R}^N)} \leq C\|\nabla v\|_{L^p(\mathbb{R}^N)} \quad \text{if} \quad p \geq N.$$

We deduce that the range of the gradient operator:

$$\mathrm{grad}:\ W_0^{1,p}(\mathbb{R}^N) \mapsto \mathbf{L}^p(\mathbb{R}^N) \perp \mathbf{H}_{p'}(\mathbb{R}^N),$$

with

$$\mathbf{H}_{p'}(\mathbb{R}^N) = \left\{\boldsymbol{\varphi} \in \mathbf{L}^{p'}(\mathbb{R}^N);\ \mathrm{div}\,\boldsymbol{\varphi} = 0\right\},$$

is a closed subspace of $\mathbf{L}^{p'}(\mathbb{R}^N)$. Here, $\mathbf{L}^p(\mathbb{R}^N) \perp \mathbf{H}_{p'}(\mathbb{R}^N)$ denotes the subspace of functions $\boldsymbol{f}$ in $\mathbf{L}^p(\mathbb{R}^N)$ which satisfy $< \boldsymbol{f},\ \boldsymbol{v} >= 0$ for any $\boldsymbol{v} \in \mathbf{H}_{p'}(\mathbb{R}^N)$. Let $v \in \mathcal{D}'(\mathbb{R}^N)$ such that $\nabla v \in L^p(\mathbb{R}^N)$. By the density of

$$\boldsymbol{\mathcal{V}}(\mathbb{R}^N) \ = \{\, \boldsymbol{\varphi} \in \boldsymbol{\mathcal{D}}(\mathbb{R}^N);\ \ \mathrm{div}\, \boldsymbol{\varphi} = 0 \,\}$$

in the space $\mathbf{H}_{p'}(\mathbb{R}^N)$, we deduce that $\nabla v \in L^p(\mathbb{R}^N) \perp \mathbf{H}_{p'}(\mathbb{R}^N)$ and then there exists $w \in W_0^{1,p}(\mathbb{R}^N)$ such that $\nabla w = \nabla v$ and then there exists a unique constant K such that $w = v + K$. □

Return to the variational formulation of the Laplace equation. The good choice of the space V is given by the space $W_0^{1,2}(\mathbb{R}^N)$. We denote by

$$W_0^{-1,p'}(\mathbb{R}^N) = [W_0^{1,p}(\mathbb{R}^N)]'.$$

For the general case $1 < p < \infty$, we have then

THEOREM 1.2. *Let $f \in W_0^{-1,p}(\mathbb{R}^N)$, with $1 < p < \infty$, satisfying the compatibility condition*

$$< f, 1 >= 0 \quad \text{if} \quad p \leq \frac{N}{N-1}. \tag{1.2}$$

Then, there exists a unique $u \in W_0^{1,p}(\mathbb{R}^N)$, up to a constant if $p \geq N$, satisfying (1.1) *and the corresponding estimate. Moreover, if $p < N$, then $u = E \star f$, where E is the fundamental solution of the Laplacian.*

PROOF. Thanks to Hardy inequality, we write $f = \mathrm{div}\ \boldsymbol{v}$, with $\boldsymbol{v} \in \boldsymbol{L}^p(\mathbb{R}^N)$. Let $\boldsymbol{v}_m \in \boldsymbol{\mathcal{D}}(\mathbb{R}^N)$ such that $\boldsymbol{v}_m \longrightarrow \boldsymbol{v}$ in $\boldsymbol{L}^p(\mathbb{R}^N)$. Setting $f_m = \mathrm{div}\ \boldsymbol{v}_m$ and $\psi_m = E \star f_m$. We have for any $\varphi \in \mathcal{D}(\mathbb{R}^N)$,

$$< \partial_i \psi_m,\ \varphi > = - < E \star f_m,\ \partial_i \varphi > = < \boldsymbol{v}_m,\ \nabla \partial_i (E \star \varphi) > .$$

According to Calderon-Zygmund inequality, we obtain:

$$\begin{aligned} | < \partial_i \psi_m,\ \varphi > | &\leq \|\boldsymbol{v}_m\|_{\boldsymbol{L}^p(\mathbb{R}^N)} \|\nabla \partial_i (E \star \varphi)\|_{L^{p'}(\mathbb{R}^N)} \\ &\leq C \|\boldsymbol{v}_m\|_{\boldsymbol{L}^p(\mathbb{R}^N)} \|\Delta (E \star \varphi)\|_{L^{p'}(\mathbb{R}^N)} \\ &\leq C \|f\|_{W_0^{-1,p}(\mathbb{R}^N)} \|\varphi\|_{L^{p'}(\mathbb{R}^N)}, \end{aligned}$$

so that $\nabla \psi_m$ is bounded in$\boldsymbol{L}^p(\mathbb{R}^N)$ and there exists a sequence c_m such that $\psi_m + c_m$ converges weakly to some function u in $W_0^{1,p}(\mathbb{R}^N)$ and $-\Delta u = f$. □

To study some regularity results, we need to define the following weighted spaces:

$$W_1^{2,p}(\Omega) = \{u \in \mathcal{D}'(\Omega), \frac{u}{w_0} \in L^p(\Omega), \nabla u \in \mathbf{L}^p(\Omega), w_2 D^2 u \in \mathbf{L}^p(\Omega)\},$$

where

$$w_0 = \ 1 + r \quad \text{if } p \neq N, \quad w_0 = \ (1+r)\ \ln(2+r) \quad \text{if } p = N, \quad w_2 = \ 1 + r.$$

and

$$W_1^{0,p}(\mathbb{R}^N) = \{u \in \mathcal{D}'(\mathbb{R}^N),\ (1+r)u \in L^p(\mathbb{R}^N)\}.$$

Note that $W_1^{2,p}(\Omega) \hookrightarrow W_0^{1,p}(\mathbb{R}^N)$ and

$$W_1^{0,p}(\mathbb{R}^N) \hookrightarrow W_0^{-1,p}(\mathbb{R}^N) \quad \Leftrightarrow \qquad N \neq p'$$

COROLLARY 1.3. *Let*

$$f \in W_1^{0,p}(\mathbb{R}^N) \quad \text{if } N \neq p' \quad \text{or} \quad f \in W_0^{-1,p}(\mathbb{R}^N) \cap W_1^{0,p}(\mathbb{R}^N) \quad \text{if } N = p'$$

and satisfy the compatibility condition (1.2). *Then, the solutions given by the previous theorem belong to* $W_1^{2,p}(\mathbb{R}^N)$ *and we have the following weighted Calderon-Zygmund inequality: for any* $\varphi \in \mathcal{D}(\mathbb{R}^N)$

$$\|(1+|x|)\frac{\partial^2 \varphi}{\partial x_i \partial x_j}\|_{L^p(\mathbb{R}^N)} \leq \|(1+|x|)\Delta\varphi\|_{L^p(\mathbb{R}^N)}$$

if and only if $N \neq p'$.

To find another strong solutions, with another behavior at the infinity, we set

$$W_0^{2,p}(\mathbb{R}^N) = \{u \in \mathcal{D}'(\Omega), \frac{u}{w_0} \in L^p(\Omega), \frac{\nabla u}{w_1} \in \mathbf{L}^p(\Omega), D^2 u \in \mathbf{L}^p(\Omega)\},$$

where

$$w_0 = (1+r)^2 \text{ if } p \notin \{\frac{N}{2}, N\}, \; w_0 = (1+r)^2 \ln(2+r) \text{ otherwise},$$

$$w_1 = 1+r \quad \text{if } p \neq N, \quad w_1 = (1+r)\ln(2+r) \quad \text{if } p = N.$$

THEOREM 1.4.

(i) Let $f \in L^p(\mathbb{R}^N)$. *Then, there exists a unique* $u \in W_0^{2,p}(\mathbb{R}^N)$, *up to a polynomial of degre less or equal 1 if* $p \geq N$, *up to a constant if* $\frac{N}{2} \leq p < N$, *satisfying* (1.1). *Moreover, if* $p < \frac{N}{2}$, *then* $u = E \star f$, *where* E *is the fundamental solution of the Laplacian.*

(ii) Let $f \in W_0^{-2,p}(\mathbb{R}^N)$ *and satisfy the compatibility condition*

$$< f, 1 >=< f, x_i >= 0 \text{ if } p \leq \frac{N}{N-1} \text{ and } < f, 1 >= 0 \text{ if } \frac{N}{N-1} < p \leq \frac{N}{N-2}.$$

Then, there exists a unique solution $u \in L^p(\mathbb{R}^N)$ *satisfying* (1.1).

3. Functional spaces

More generally, we define the following weighted Sobolev spaces (see [2] for more details). Let $m \in \mathbb{N}$, $\alpha, \beta \in \mathbb{R}$, $p \in]1, \infty[$, Ω any open set of $\mathbb{R}^N$,

$$W_{\alpha,\beta}^{m,p}(\Omega) = \Big\{ u \in \mathcal{D}'(\Omega) \; ; \; 0 \leqslant |\lambda| \leqslant k, \; \varrho^{\alpha-m+|\lambda|} (\lg \varrho)^{\beta-1} \partial^\lambda u \in L^p(\Omega) \; ;$$

$$k+1 \leqslant |\lambda| \leqslant m, \; \varrho^{\alpha-m+|\lambda|} (\lg \varrho)^{\beta} \partial^\lambda u \in L^p(\Omega) \Big\},$$

where $\varrho = (1+|x|^2)^{1/2}$, $\lg \varrho = \ln(1+\varrho^2)$, $\lambda \in \mathbb{N}^N$,

$$\text{and} \quad k = \begin{cases} -1 & \text{if } \frac{N}{p} + \alpha \notin \{1, \ldots, m\}, \\ m - \frac{N}{p} - \alpha & \text{if } \frac{N}{p} + \alpha \in \{1, \ldots, m\}. \end{cases}$$

If $\beta = 0$, we simply denote this space by $W_\alpha^{m,p}(\Omega)$ and we set

$$\mathring{W}_{\alpha,\beta}^{m,p}(\Omega) = \overline{\mathcal{D}(\Omega)}^{\|\cdot\|_{W_{\alpha,\beta}^{m,p}(\Omega)}} \text{ and } W_{-\alpha,-\beta}^{-m,p'}(\Omega) = \Big(\mathring{W}_{\alpha,\beta}^{m,p}(\Omega) \Big)'.$$

EXAMPLE 1.5 (Examples of Weighted Sobolev Spaces).

$$W_0^{1,p}(\Omega) = \{u \in \mathcal{D}'(\Omega), \frac{u}{w_0} \in L^p(\Omega), \nabla u \in \mathbf{L}^p(\Omega)\},$$

where

$$w_0 = 1 + r \quad \text{if } p \neq N, \quad w_0 = (1+r)\ \ln(2+r) \quad \text{if } p = N.$$

$$W_0^{2,p}(\Omega) = \{u \in \mathcal{D}'(\Omega), \frac{u}{w_0} \in L^p(\Omega), \frac{\nabla u}{w_1} \in \mathbf{L}^p(\Omega), D^2u \in \mathbf{L}^p(\Omega)\},$$

where

$$w_0 = (1+r)^2 \ \text{ if } p \notin \{\frac{N}{2}, N\},\ w_0 = (1+r)^2\ \ln(2+r), \quad \text{otherwise},$$

and

$$w_1 = 1 + r \quad \text{if } p \neq N, \quad w_1 = (1+r)\ \ln(2+r) \quad \text{if } p = N.$$

$$W_1^{2,p}(\Omega) = \{u \in \mathcal{D}'(\Omega), \frac{u}{w_0} \in L^p(\Omega), \nabla u \in \mathbf{L}^p(\Omega), w_2 D^2u \in \mathbf{L}^p(\Omega)\},$$

where

$$w_0 = 1 + r \quad \text{if } p \neq N, \quad w_0 = (1+r)\ \ln(2+r) \quad \text{if } p = N,$$
$$w_2 = 1 + r.$$

4. Spaces of traces and inequalities

In this section, we give some properties concerning the functions which belong to weighted Sobolev spaces and their traces (see [1, 3]).

4.1. Case $W_0^{1,2}(\mathbb{R}^N_+)$. We consider the space:

$$W_0^{1/2,2}(\mathbb{R}^{N-1}) = \{u \in \mathcal{D}'(\mathbb{R}^{N-1})\ ;\ \frac{u}{\omega'^{1/2}} \in L^2(\mathbb{R}^{N-1}),$$
$$\int_{\mathbb{R}^{N-1}\times\ \mathbb{R}^{N-1}} \frac{|u(x') - u(y')|^2}{|x' - y'|^N} dx'dy' < \infty\},$$

with $\omega' = \varrho' = (1 + |x'|^2)^{1/2}$ if $N \geq 3$ and $\omega' = \varrho'(\lg\varrho')^2$ if $N = 2$. It is a reflexive Banach for the norm:

$$||u||_{W_0^{1/2,2}(\mathbb{R}^{N-1})} = \Big(\int_{\mathbb{R}^{N-1}} \frac{|u(x')|^2}{\omega'} dx' + \int_{\mathbb{R}^{N-1}\times\ \mathbb{R}^{N-1}} \frac{|u(x') - u(y')|^2}{|x' - y'|^N} dx'dy' \Big)^{1/2}.$$

Its semi-norm:

$$|u|_{W_0^{1/2,2}(\mathbb{R}^{N-1})} = \Big(\int_{\mathbb{R}^{N-1}\times\ \mathbb{R}^{N-1}} \frac{|u(x') - u(y')|^2}{|x' - y'|^N} dx'dy' \Big)^{1/2}.$$

LEMMA 1.6. *For any $N \geq 2$ and $u \in \mathcal{D}(\mathbb{R}^{N-1})$, we have the relation*

$$\int_{\mathbb{R}^{N-1}\times\ \mathbb{R}^{N-1}} \frac{|u(x') - u(y')|^2}{|x' - y'|^N} dx'dy' = C_N \int_{\mathbb{R}^{N-1}} |\xi'||\widehat{u}(\xi')|^2 d\xi'$$

On $\mathcal{D}(\mathbb{R}^{N-1})$, the semi-norms $|\ \cdot\ |_{W_0^{1/2,2}(\mathbb{R}^{N-1})}$ and $|\ \cdot\ |_{H^{1/2}(\mathbb{R}^{N-1})}$ are equivalent, where $H^{1/2}(\mathbb{R}^{N-1})$ is the classical Sobolev space which corresponds to the traces of functions in $H^1(\mathbb{R}^N_+)$.

By Fourier's transform, it is easy to prove the following lemma:

LEMMA 1.7. *For any $N \geq 2$ and $u \in \mathcal{D}(\overline{\mathbb{R}^N_+})$, we have the inequality*

$$\int_{\mathbb{R}^{N-1}} |\xi'|\ |\widehat{u}(\xi',0)|^2\ d\xi' \leq C \int_{\mathbb{R}^N_+} |\nabla u|^2 dx,$$

that means that

$$\forall u \in \mathcal{D}(\overline{\mathbb{R}^N_+}), \quad |\ \gamma_0(u)\ |_{W_0^{1/2,2}(\mathbb{R}^{N-1})} \leq C|u|_{W_0^{1,2}(\mathbb{R}^N_+)}.$$

Recall now the Pitt's Inequality. For any $N \geq 3$ and $u \in \mathcal{D}(\mathbb{R}^{N-1})$, we have :

$$\int_{\mathbb{R}^{N-1}} \frac{|u(\xi')|^2}{|\xi'|}\ d\xi' \leq C \int_{\mathbb{R}^{N-1}} |\xi'|\ |\widehat{u}(\xi')|^2\ d\xi'.$$

If $N = 2$, we can prove the following lemma:

LEMMA 1.8. *For any $u \in \mathcal{D}(]0,\infty[)$, we have :*

$$\int_0^\infty \frac{|u(t)|^2}{t\ln^2(2+t)} dt \leq C \int_0^\infty \int_0^\infty \frac{|u(t)-u(\tau)|^2}{|t-\tau|^2} dt d\tau.$$

As consequence of the previous lemmas, we can prove

THEOREM 1.9.

(i) If $N \geq 3$, for any $u \in W_0^{1,2}(\mathbb{R}^N_+)$, we have:

$$\begin{aligned}\int_{\mathbb{R}^{N-1}} \frac{|\gamma_0 u(x')|^2}{|x'|} dx' &\leq C_1 |\ \gamma_0 u\ |_{W_0^{1/2,2}(\mathbb{R}^{N-1})} \\ &\leq C_2 |u|_{W_0^{1,2}(\mathbb{R}^N_+)}.\end{aligned}$$

(ii) If $N = 2$, for any $u \in W_0^{1,2}(\mathbb{R}^2_+)$, we have:

$$\begin{aligned}\inf_{K\in\mathbb{R}} \int_{\mathbb{R}} \frac{|\gamma_0 u(x_1)+K|^2}{|x_1|\ln^2(2+|x_1|)} dx_1 &\leq C_1 |\ \gamma_0 u\ |_{W_0^{1/2,2}(\mathbb{R}^{N-1})} \\ &\leq C_2 |u|_{W_0^{1,2}(\mathbb{R}^N_+)}.\end{aligned}$$

(iii) If $N \geq 2$, the mapping $\gamma_0 : u \in \mathcal{D}(\overline{\mathbb{R}^N_+}) \to \gamma_0 u \in \mathcal{D}(\mathbb{R}^{N-1})$ can be extended to a linear continuous mapping denoted by γ_0 from $W_0^{1,2}(\mathbb{R}^N_+)$ into $W_0^{1/2,2}(\mathbb{R}^{N-1})$.

The lifting operator in $W_0^{1/2,2}(\mathbb{R}^{N-1})$ is given by the following lemma.

LEMMA 1.10. *Let $N \geq 2$ and $g \in W_0^{1/2,2}(\mathbb{R}^{N-1})$. Then, there exists $u = Rg \in W_0^{1,2}(\mathbb{R}^N_+)$ such that $u = g$ on $\mathbb{R}^{N-1}$ and*

$$\begin{aligned}&||u||_{W_0^{1,2}(\mathbb{R}^N_+)} \leq C|g|_{W_0^{1/2,2}(\mathbb{R}^{N-1})} \quad \text{if } N \geq 3, \\ &\inf_{K\in\mathbb{R}} ||u+K||_{W_0^{1,2}(\mathbb{R}^2_+)} \leq C|g|_{W_0^{1/2,2}(\mathbb{R}^{N-1})} \quad \text{if } N = 2.\end{aligned}$$

A variant of the last inequality:

LEMMA 1.11. *For any $g \in W_0^{1/2,2}(\mathbb{R})$ such that*

$$\int_{-1}^1 g(t)dt = 0,$$

there exists $u = Rg \in W_0^{1,2}(\mathbb{R}^2_+)$ *such that* $u = g$ *on* $\mathbb{R}$ *and*

$$||u||_{W_0^{1,2}(\mathbb{R}^2_+)} \leq C||g||_{W_0^{1/2,2}(\mathbb{R})}.$$

We can now summarize:

THEOREM 1.12. *Let* $N \geq 2$. *The mapping*

$$\gamma_0 \ : \ u \in W_0^{1,2}(\mathbb{R}^N_+) \to \gamma_0 u \in W_0^{1/2,2}(\mathbb{R}^{N-1})$$

is continuous, surjective and $\operatorname{Ker}\gamma_0 = \overset{\circ}{W}{}_0^{1,2}(\mathbb{R}^N_+)$.

4.2. Case $W_0^{1,p}(\mathbb{R}^N_+)$. We consider the space:

$$W_0^{1-1/p,p}(\mathbb{R}^{N-1}) = \Big\{u \in \mathcal{D}'(\mathbb{R}^{N-1}) \ ; \ \frac{u}{\omega'^{1-1/p}} \in L^p(\mathbb{R}^{N-1}),$$
$$\int_{\mathbb{R}^{N-1}\times\ \mathbb{R}^{N-1}} \frac{|u(x') - u(y')|^p}{|x' - y'|^{N+p-2}} dx'dy' < \infty\Big\},$$

with

$$\omega' = \varrho' = (1 + |x'|^2)^{1/2} \quad \text{if } N \neq p \quad \text{and} \quad \omega' = \varrho'(\lg\varrho')^2 \quad \text{if } N = p.$$

It is a reflexive Banach with the norm $||u||_{W_0^{1-1/p,p}(\mathbb{R}^{N-1})}$:

$$\Big(\int_{\mathbb{R}^{N-1}} \frac{|u(x')|^p}{\omega'^{p-1}} dx' + \int_{\mathbb{R}^{N-1}\times\ \mathbb{R}^{N-1}} \frac{|u(x') - u(y')|^p}{|x' - y'|^{N+p-2}} dx'dy'\Big)^{1/p}.$$

Its semi-norm is defined by

$$|u|_{W_0^{1-1/p,p}(\mathbb{R}^{N-1})} = \Big(\int_{\mathbb{R}^{N-1}\times\mathbb{R}^{N-1}} \frac{|u(x') - u(y')|^p}{|x' - y'|^{N+p-2}}\Big)^{1/p}$$

LEMMA 1.13. *Let* $N \geq 2$ *and* $1 < p < \infty$. *Then for any* $u \in \mathcal{D}(\overline{\mathbb{R}^N_+})$, *we have the inequality*

$$\int_{\mathbb{R}^{N-1}\times\ \mathbb{R}^{N-1}} \frac{|u(x',0) - u(y',0)|^p}{|x' - y'|^{N+p-2}} dx'dy' \leq C \int_{\mathbb{R}^N_+} |\nabla u|^p dx,$$

that means that the semi-norm in $\mathbb{R}^{N-1}$ *is controlled by the semi-norm in the whole space.*

Recall now the Pitt's Inequality. For any $N \geq 3$, $1 < p < N$ and $u \in \mathcal{D}(\mathbb{R}^{N-1})$, we have the inequality

$$\int_{\mathbb{R}^{N-1}} \frac{|u(\xi')|^p}{|\xi'|^{p-1}}\ d\xi' \leq C \int_{\mathbb{R}^{N-1}} |\xi'|^{pN-2N+1}\ |\widehat{u}(\xi')|^p\ d\xi'.$$

COROLLARY 1.14. *For any* $N \geq 2$, $u \in \mathcal{D}(\overline{\mathbb{R}^N_+})$ *and* $1 < p \leq 2$, *we have the inequalities:*

$$\int_{\mathbb{R}^{N-1}} \frac{|u(\xi',0)|^p}{|\xi'|^{p-1}} d\xi' \leq C_1 \int_{\mathbb{R}^{N-1}\times\ \mathbb{R}^{N-1}} \frac{|u(x',0) - u(y',0)|^p}{|x' - y'|^{N+p-2}}$$
$$\leq C_2 \int_{\mathbb{R}^N_+} |\nabla u|^p dx,$$

that means that

$$|\gamma_0 u|_{W^{0,p}_{1/p-1}(\mathbb{R}^{N-1})} \leq C_1 |\gamma_0 u|_{W^{1-1/p,p}_0(\mathbb{R}^{N-1})} \leq C_2 |u|_{W^{1,p}_0(\mathbb{R}^N_+)}$$

What is the analogue of this result when $2 < p < N$? We will use an interpolation argument.

PROPOSITION 1.15. *For any $1 < p < N$ and any $u \in W_0^{1-1/p,p}(\mathbb{R}^{N-1})$, we have the estimates*

$$\int_{\mathbb{R}^{N-1}} \frac{|u(x)|^p}{|x|^{p-1}}\, dx \leq C |u|_{W_0^{1-1/p,p}(\mathbb{R}^{N-1})}.$$

REMARK 1.16. *See [18] for an another proof.*

COROLLARY 1.17. *Let $1 < p < N$. Then for any $u \in W_0^{1,p}(\mathbb{R}^N_+)$, we have the estimates:*

$$|\gamma_0 u|_{W^{0,p}_{1/p-1}(\mathbb{R}^{N-1})} \leq C_1 |\gamma_0 u|_{W^{1-1/p,p}_0(\mathbb{R}^{N-1})} \leq C_2 |u|_{W^{1,p}_0(\mathbb{R}^N_+)}$$

When the exponent p is greater or equal to the dimension N, we have:

LEMMA 1.18.

(i) Let $p > N$. Then for any $u \in W_0^{1,p}(\mathbb{R}^N_+)$, we have the estimates:

$$\inf_{K\in\mathbb{R}} |\gamma_0 u + K|_{W^{0,p}_{1/p-1}(\mathbb{R}^{N-1})} \leq C |u|_{W^{1,p}_0(\mathbb{R}^N_+)}.$$

(ii) If $p = N$ and $u \in W_0^{1,N}(\mathbb{R}^N_+)$, we have the estimates:

$$\inf_{K\in\mathbb{R}} \int_{\mathbb{R}^{N-1}} \frac{|\gamma_0 u(x') + K|^N}{|x'|^{N-1} \ln^N(2 + |x'|)} dx' \leq C |u|_{W_0^{1,N}(\mathbb{R}^N_+)}.$$

As for the Hilbertian case, we have:

THEOREM 1.19. *Let $N \geq 2$. The following mapping*

$$\gamma_0 \ : \ u \in \mathcal{D}(\overline{\mathbb{R}^N_+}) \to \gamma_0 u \in \mathcal{D}(\mathbb{R}^{N-1})$$

can be extended to a linear continuous mapping denoted also γ_0 from $W_0^{1,p}(\mathbb{R}^N_+)$ into $W_0^{1-1/p,p}(\mathbb{R}^{N-1})$. Moreover γ_0 is onto and $\operatorname{Ker}\gamma_0 = \overset{\circ}{W}{}_0^{1,p}(\mathbb{R}^N_+)$.

4.3. General Case. We consider the following spaces of traces.

For any $\sigma \in]0,1[$,

$$W^{\sigma,\,p}_\alpha(\mathbb{R}^N) = \Big\{ u \in \mathcal{D}'(\mathbb{R}^N) \ ; \ \omega^{\alpha-\sigma} u \in L^p(\mathbb{R}^N), \\ \int_{\mathbb{R}^N\times\mathbb{R}^N} \frac{|\varrho^\alpha(x)\, u(x) - \varrho^\alpha(y)\, u(y)|^p}{|x-y|^{N+\sigma p}}\, dx\, dy < \infty \Big\},$$

where $\omega = \varrho$ if $N/p + \alpha \neq \sigma$ and $\omega = \varrho\,(\lg\,\varrho)^{1/(\sigma-\alpha)}$ if $N/p + \alpha = \sigma$.

For any $s \in \mathbb{R}^+$,

$$W^{s,\,p}_\alpha(\mathbb{R}^N) = \Big\{ u \in \mathcal{D}'(\mathbb{R}^N) \ ; \ 0 \leqslant |\lambda| \leqslant k, \ \varrho^{\alpha-s+|\lambda|}\,(\lg\,\varrho)^{-1}\,\partial^\lambda u \in L^p(\mathbb{R}^N) \ ; \\ k+1 \leqslant |\lambda| \leqslant [s]-1, \ \varrho^{\alpha-s+|\lambda|}\,\partial^\lambda u \in L^p(\mathbb{R}^N) \ ; \ |\lambda| = [s], \ \partial^\lambda u \in W^{\sigma,\,p}_\alpha(\mathbb{R}^N) \Big\},$$

where $k = s - N/p - \alpha$ if $N/p + \alpha \in \{\sigma, \dots, \sigma + [s]\}$, with $\sigma = s - [s]$ and $k = -1$ otherwise.

We have the following general result (see [3]):

LEMMA 1.20. *The mapping* $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_{m-1}) : \mathcal{D}(\overline{\mathbb{R}^N_+}) \longrightarrow \mathcal{D}(\mathbb{R}^{N-1})^m$, *can be extended by continuity to a linear and continuous mapping:*

$$\gamma : W^{m,p}_\alpha(\mathbb{R}^N_+) \longrightarrow \prod_{j=0}^{m-1} W^{m-j-1/p,p}_\alpha(\mathbb{R}^{N-1}).$$

Moreover, γ *is onto and* $\operatorname{Ker}\gamma = \overset{\circ}{W}{}^{m,p}_\alpha(\mathbb{R}^N_+)$.

Let us now introduce the following spaces of polynomials. For $q \in \mathbb{Z}$, $\mathcal{P}_q$ is the space of polynomials of degree $\leqslant q$, $\mathcal{P}^\Delta_q$ the subspace of harmonic polynomials of $\mathcal{P}_q$, $\mathcal{P}^{\Delta^2}_q$ the subspace of biharmonic polynomials of $\mathcal{P}_q$, $\mathcal{A}^\Delta_q$ is the subspace of polynomials $\mathcal{P}^\Delta_q$, odd with respect x_N, $\mathcal{N}^\Delta_q$ is the subspace of polynomials $\mathcal{P}^\Delta_q$, even with respect x_N. Observe that $\varphi \in \mathcal{A}^\Delta_q \Leftrightarrow \varphi(x', 0) = 0$ and $\varphi \in \mathcal{N}^\Delta_q \Leftrightarrow \partial_N\varphi(x', 0) = 0$. For all $s \in \mathbb{R}$, $[s]$ denotes a real part of s.

5. The Dirichlet problem for the Laplacian in $\mathbb{R}^N_+$

We consider the following Dirichlet problem for the Laplacian in the half-space (see [3]):

$$(P_D) \qquad \Delta u = f \quad \text{in } \mathbb{R}^N_+ \quad u = g \quad \text{on } \Gamma := \mathbb{R}^{N-1} \times \{0\}.$$

We recall that A^Δ_q is the subspace of polynomials Q of P^Δ_q satisfying $Q(x', 0) = 0$.

THEOREM 1.21. *Let* $\ell \geq 0$ *be an integer and* $\frac{N}{p'} \notin \{1, \dots, \ell\}$ *with the convention this set is empty if* $\ell = 0$. *For any* f *in* $W^{-1,p}_\ell(\mathbb{R}^N_+)$ *and* g *in* $W^{\frac{1}{p'},p}_\ell(\Gamma)$ *satisfying the compatibility condition*

$$\forall \varphi \in A^\Delta_{[\ell+1-\frac{N}{p'}]} \qquad < f, \varphi >_{W^{-1,p}_\ell \times W^{1,p'}_{-l}} = < g, \frac{\partial \varphi}{\partial x_N} >_\Gamma,$$

where $< \cdot, \cdot >_\Gamma$ *denote the duality brackets between* $W^{\frac{1}{p'},p}_\ell(\Gamma)$ *and* $W^{-\frac{1}{p'},p}_{-\ell}(\Gamma)$, *problem* (P_D) *has a unique solution* $u \in W^{1,p}_\ell(\mathbb{R}^N_+)$ *and*

$$||u||_{W^{1,p}_\ell(\mathbb{R}^N_+)} \leq C(||f||_{W^{-1,p}_\ell(\mathbb{R}^N_+)} + ||g||_{W^{\frac{1}{p'},p}_\ell(\Gamma)}).$$

PROOF. Note that the kernel of the operator

$$(-\Delta, \gamma_0) : W^{1,p}_\ell(\mathbb{R}^N_+) \to W^{-1,p}_\ell(\mathbb{R}^N_+) \times W^{\frac{1}{p'},p}_\ell(\Gamma)$$

is precisely $A^\Delta_{[-\ell+1-N/p']}$ for any integer ℓ and $A^\Delta_{[-\ell+1-\frac{N}{p'}]} = \{0\}$ if $\ell \geq 0$. Let $u_g \in W^{1,p}_\ell(\mathbb{R}^N_+)$ be a lifting function of g:

$$u_g = g \text{ on } \Gamma \text{ and } ||u_g||_{W^{1,p}_\ell(R^N_+)} \leq C_1 ||g||_{W^{\frac{1}{p'},p}_\ell(\Gamma)}.$$

Then, $(\mathcal{P}_D)$ is equivalent to the problem

$$-\Delta v = f + \Delta u_g \quad \text{in} \quad \mathbb{R}^N_+, \quad v = 0 \quad \text{on} \quad \Gamma.$$

Let $h = f + \Delta u_g$. For any $\varphi \in W^{1,p'}_{-\ell}(\mathbb{R}^N)$, we set

$$\sqcap\varphi(x', x_N) = \varphi(x', x_N) - \varphi(x', -x_N) \quad \text{if} \quad x_N > 0.$$

It is clear that $\sqcap\varphi \in \overset{\circ}{W}{}^{1,p'}_{-\ell}(\mathbb{R}^N_+)$. Then h can be extended to $h_\pi \in W^{-1,p}_l(\mathbb{R}^N)$ defined by

$$\varphi \in W^{1,p'}_{-\ell}(\mathbb{R}^N),\ h_\pi(\varphi) = < h, \sqcap\varphi >_{W^{-1,p}_\ell(\mathbb{R}^N_+)\times W^{1,p'}_{-\ell}(\mathbb{R}^N_+)}.$$

Moreover,

$$||h_\pi||_{W^{-1,p}_\ell(\mathbb{R}^N)} = ||h||_{W^{-1,p}_\ell(\mathbb{R}^N_+)}.$$

Let $q \in P^\Delta_{[\ell+1-N/p']}$. Then,

$$q = r + s,\ r \in A^\Delta_{[\ell+1-N/p']} \quad \text{and } s \in N^\Delta_{[\ell+1-N/p]}.$$

So,

$$< h_\pi, q >= < f + \Delta u_g, r >_{W^{-1,p}_\ell(\mathbb{R}^N_+)\times W^{1,p'}_{-\ell}(\mathbb{R}^N_+)}$$

and applying the Green formula:

$$< \Delta u_g, r >= -\int_{\mathbb{R}^N_+} \nabla u_g \cdot \nabla r dx = -\Big\langle g, \frac{\partial r}{\partial x_N} \Big\rangle_{W^{\frac{1}{p'},p}_\ell(\Gamma)\times W^{\frac{-1}{p'},p'}_{-\ell}(\Gamma)}$$

(note that $\Delta r = 0$ in $\mathbb{R}^N_+$ and $r = 0$ on Γ). Then, $h_\pi \in W^{-1,p}_\ell(\mathbb{R}^N)$ and

$$\forall q \in P^\Delta_{[\ell+1-N/p']},\ < h_\pi, q >= 0.$$

Recall that the operator

$$\Delta\ : W^{1,p}_\ell(\mathbb{R}^N) \to W^{-1,p}_\ell \perp P^\Delta_{[\ell+1-\frac{N}{p'}]} \quad \text{if } \ell \geq 1$$

and

$$\Delta\ : W^{1,p}_0(\mathbb{R}^N)/P_{[1-\frac{N}{p}]} \to W^{-1,p}_0(\mathbb{R}^N) \perp P_{[1-\frac{N}{p'}]} \quad \text{if } \ell = 0$$

are isomorphisms. Then, there exists $\widetilde{v} \in W^{1,p}_\ell(\mathbb{R}^N)$ such that

$$-\Delta\widetilde{v} = h_\pi.$$

Now, remark that the function $w = \frac{1}{2} \sqcap \widetilde{v} \in W^{1,p}_\ell(\mathbb{R}^N_+)$ and

$$-\Delta w = h \quad \text{in} \quad \mathbb{R}^N_+ \quad \text{and} \quad w = 0 \quad \text{on} \quad \Gamma,$$

i.e. w is solution of our problem. □

REMARK 1.22.

(i) The kernel $A^\Delta_{[-l+1-N/p]}$ is reduced to $\{0\}$ if $\ell \geq 0$ and to $P_{[1-N/p]}$ if $\ell = 0$.

(ii) With similar arguments, we can show an analogous result if $\ell < 0$:

$$f \in W^{-1,p}_\ell(\mathbb{R}^N_+), \quad g \in W^{\frac{1}{p'},p}_\ell(\Gamma),$$

then

$$u \in W^{1,p}_\ell(\mathbb{R}^N_+)/A^\Delta_{[-\ell+1-N/p]}.$$

6. The biharmonic problem in $\mathbb{R}^N_+$

We are interested here by the biharmonic problem (see [6, 7, 8, 10, 15]):

$$(B)\ \begin{cases} \Delta^2 u = f & \text{in } \mathbb{R}^N_+, \\ u = g_0 & \text{on } \Gamma, \\ \partial_N u = g_1 & \text{on } \Gamma. \end{cases}$$

As for Problem (P_D), we have the following theorem.

THEOREM 1.23 (Generalized solutions). *Let $\ell \in \mathbb{Z}$ be such that $\frac{N}{p'} \notin \{1, \dots, \ell\}$ and $\frac{N}{p} \notin \{1, \dots, -\ell\}$. There exists $C > 0$ such that for all $f \in W^{-2,\,p}_{\ell}(\mathbb{R}^N_+)$, $g_0 \in W^{2-1/p,\,p}_{\ell}(\Gamma)$ and $g_1 \in W^{1-1/p,\,p}_{\ell}(\Gamma)$ satisfying the compatibility condition*

$$\forall \varphi \in \mathcal{B}_{[2+\ell-N/p']} : \quad \langle f, \varphi \rangle_{W^{-2,\,p}_{\ell}(\mathbb{R}^N_+) \times \mathring{W}^{2,\,p'}_{-\ell}(\mathbb{R}^N_+)} + \langle g_1, \Delta\varphi \rangle_{\Gamma} - \langle g_0, \partial_N \Delta \varphi \rangle_{\Gamma} = 0,$$

there exists a unique $u \in W^{2,\,p}_{\ell}(\mathbb{R}^N_+)/\mathcal{B}_{[2-\ell-N/p]}$ solution to (B), with the estimate

$$\inf_{q \in \mathcal{B}_{[2-\ell-N/p]}} \|u + q\|_{W^{2,\,p}_{\ell}(\mathbb{R}^N_+)} \leqslant C(\|f\|_{W^{-2,\,p}_{\ell}} + \|g_0\|_{W^{2-1/p,\,p}_{\ell}} + \|g_1\|_{W^{1-1/p,\,p}_{\ell}}).$$

PROOF. First, we recall the reflection principle for Δ^2. If $u \in \mathcal{D}'(\mathbb{R}^N_+)$ is such that $\Delta^2 u = 0$, then $u \in \mathcal{C}^\infty(\mathbb{R}^N_+)$.

Let u be a biharmonic function in $\mathbb{R}^N_+$ with $u = \partial_N u = 0$ on Γ. Then, by the Schwarz reflection principle, u can be extended uniquely by

$$\tilde{u}(x', x_N) \ = \ \begin{cases} u(x', x_N) & \text{if } x_N \geqslant 0, \\ (-u - 2x_N \partial_N u - x_N^2 \Delta u)(x', -x_N) & \text{if } x_N < 0. \end{cases}$$

to a biharmonic function on $\mathbb{R}^N$ (see [15]). Now, if $u \in W^{2,\,p}_{\ell}(\mathbb{R}^N_+)$, we show that $\tilde{u} \in \mathcal{S}'(\mathbb{R}^N)$ and then $u \in \mathcal{P}^{\Delta^2}$. Consequently,

$$\mathrm{Ker}(\Delta^2, \gamma_0, \gamma_1) = \mathcal{B}_{[2-\ell-N/p]} = \left\{ u \in \mathcal{P}^{\Delta^2}_{[2-\ell-N/p]} \ ; \ u = \partial_N u = 0 \text{ on } \Gamma \right\}$$

with

$$(\Delta^2, \gamma_0, \gamma_1) : W^{2,\,p}_{\ell}(\mathbb{R}^N_+) \ \longrightarrow \ W^{-2,\,p}_{\ell}(\mathbb{R}^N_+) \times W^{2-1/p,\,p}_{\ell}(\Gamma) \times W^{1-1/p,\,p}_{\ell}(\Gamma).$$

To prove the existence of solutions, we first consider the homogeneous problem corresponding to $f = 0$:

$$(P^0)\ \begin{cases} \Delta^2 u = 0 & \text{in } \mathbb{R}^N_+, \\ u = g_0 & \text{on } \Gamma, \\ \partial_N u = g_1 & \text{on } \Gamma. \end{cases}$$

We solve successively (see Amrouche-Nečasová (2001) and Amrouche (2002))

$$(R^0)\ \begin{cases} \Delta\vartheta = 0 & \text{in } \mathbb{R}^N_+, \\ \vartheta = g_0 & \text{on } \Gamma, \end{cases} \quad \textit{and} \quad (S^0)\ \begin{cases} \Delta\zeta = 0 & \text{in } \mathbb{R}^N_+, \\ \partial_N \zeta = g_1 & \text{on } \Gamma. \end{cases}$$

Then

$$u = x_N \, \partial_N(\zeta - \vartheta) + \vartheta \in W^{1,\,p}_{\ell-1}(\mathbb{R}^N_+)$$

satisfies (P^0) and we can show that $u \in W^{2,\,p}_{\ell}(\mathbb{R}^N_+)$.

For the complete problem, we use lifting of the boundary conditions:

$$\exists u_g \in W^{2,\,p}_{\ell}(\mathbb{R}^N_+), \quad (u_g,\, \partial_N u_g) = (g_0,\, g_1) \ \text{ on } \Gamma.$$

If we put $h = f - \Delta^2 u_g \in W^{-2,\,p}_{\ell}(\mathbb{R}^N_+)$ and $v = u - u_g$, the problem (P) is equivalent to the following problem:

$$(P^\star) \ \begin{cases} \Delta^2 v = h & \text{in } \mathbb{R}^N_+, \\ v = 0 & \text{on } \Gamma, \\ \partial_N v = 0 & \text{on } \Gamma, \end{cases}$$

with $h \perp \mathcal{B}_{[2+\ell-N/p']}$. To solve the problem $(P^\star)$, we will prove that

$$\Delta^2 : \mathring{W}^{2,\,p}_{\ell}(\mathbb{R}^N_+)/\mathcal{B}_{[2-\ell-N/p]}) \ \longrightarrow W^{-2,\,p}_{\ell}(\mathbb{R}^N_+) \perp \mathcal{B}_{[2+\ell-N/p']}$$

is an isomorphism.

`Step 1.` Case $2 + \ell - N/p' < 0$. Let $h \perp \mathcal{B}_{[2+\ell-N/p']}$. Then $h = \operatorname{div}\operatorname{div}\mathbb{H}$ with $\mathbb{H} \in W^{0,\,p}_{\ell}(\mathbb{R}^N_+)^{N^2}$. We extend $\mathbb{H}$ by zero and set $\tilde{h} = \operatorname{div}\operatorname{div}\tilde{\mathbb{H}}$. We know that there exists $\tilde{z} \in W^{2,\,p}_{\ell}(\mathbb{R}^N)$ satisfying $\Delta^2 \tilde{z} = \tilde{h}$ in $\mathbb{R}^N$. Setting $z = \tilde{z}|_{\mathbb{R}^N_+} \in W^{2,\,p}_{\ell}(\mathbb{R}^N_+)$, there exists $w \in W^{2,\,p}_{\ell}(\mathbb{R}^N_+)$ satisfying

$$\Delta^2 w = 0 \text{ in } \mathbb{R}^N_+, \ w = z \text{ and } \partial_N w = \partial_N z \text{ on } \Gamma.$$

Then $v = z - w$ answers to $(P^\star)$.

`Step 2.` Case $2 - \ell - N/p < 0$. We deduce the result by duality from the previous case, because Δ^2 is selfadjoint.

`Step 3.` We finish by the case $2 + \ell - N/p' \geqslant 0$ and $2 - \ell - N/p \geqslant 0$ which implies that $\ell \in \{-1, 0, 1\}$. □

In the next theorem, we obtain strong solutions which are regular when the data are also regular.

THEOREM 1.24. *Let $\ell \in \mathbb{Z}$, $m \in \mathbb{N}$ be such that*

$$\frac{N}{p'} \notin \{1, \ldots, \ell + \min\{m,\, 2\}\} \quad \text{and} \quad \frac{N}{p} \notin \{1, \ldots, -\ell - m\}.$$

There exists $C > 0$ such that for all $f \in W^{m-2,\,p}_{m+\ell}(\mathbb{R}^N_+)$, $g_0 \in W^{m+2-1/p,\,p}_{m+\ell}(\Gamma)$ and $g_1 \in W^{m+1-1/p,\,p}_{m+\ell}(\Gamma)$ satisfying the compatibility condition

$$\forall \varphi \in \mathcal{B}_{[2+\ell-N/p']} : \quad \langle f, \varphi \rangle_{W^{-2,\,p}_{\ell}(\mathbb{R}^N_+) \times \mathring{W}^{2,\,p'}_{-\ell}(\mathbb{R}^N_+)} + \langle g_1, \Delta\varphi \rangle_\Gamma - \langle g_0, \partial_N \Delta \varphi \rangle_\Gamma = 0,$$

there exists a unique $u \in W^{m+2,\,p}_{m+\ell}(\mathbb{R}^N_+)/\mathcal{B}_{[2-\ell-N/p]}$, solution of (P), with the estimate

$$\inf_{q \in \mathcal{B}_{[2-\ell-N/p]}} \|u + q\|_{W^{m+2,\,p}_{m+\ell}(\mathbb{R}^N_+)} \leqslant C(\|f\|_{W^{m-2,\,p}_{m+\ell}(\mathbb{R}^N_+)} + \|g_0\|_{W^{m+2-1/p,\,p}_{m+\ell}(\Gamma)} + \|g_1\|_{W^{m+1-1/p,\,p}_{m+\ell}(\Gamma)}).$$

We can now give a panorama of basic cases:

- For $\ell = 0$

$$\begin{array}{lrcll}
\Delta^2: & \overset{\circ}{W}{}^{2,\,p}_{0}(\mathbb{R}^N_+) & \overset{iso}{\longrightarrow} & W^{-2,\,p}_{0}(\mathbb{R}^N_+). & \\
\Delta^2: & \overset{\star}{W}{}^{3,\,p}_{1}(\mathbb{R}^N_+) & \overset{iso}{\longrightarrow} & W^{-1,\,p}_{1}(\mathbb{R}^N_+), & \text{if } N/p' \neq 1. \\
\Delta^2: & \overset{\star}{W}{}^{4,\,p}_{2}(\mathbb{R}^N_+) & \overset{iso}{\longrightarrow} & W^{0,\,p}_{2}(\mathbb{R}^N_+), & \text{if } N/p' \notin \{1,2\}.
\end{array}$$

- For $\ell = -1$

$$\begin{array}{lrcll}
\Delta^2: & \overset{\circ}{W}{}^{2,\,p}_{-1}(\mathbb{R}^N_+)/\mathcal{B}_{[3-N/p]} & \overset{iso}{\longrightarrow} & W^{-2,\,p}_{-1}(\mathbb{R}^N_+), & \text{if } N/p \neq 1. \\
\Delta^2: & \overset{\star}{W}{}^{3,\,p}_{0}(\mathbb{R}^N_+)/\mathcal{B}_{[3-N/p]} & \overset{iso}{\longrightarrow} & W^{-1,\,p}_{0}(\mathbb{R}^N_+). & \\
\Delta^2: & \overset{\star}{W}{}^{4,\,p}_{1}(\mathbb{R}^N_+)/\mathcal{B}_{[3-N/p]} & \overset{iso}{\longrightarrow} & W^{0,\,p}_{1}(\mathbb{R}^N_+), & \text{if } N/p' \neq 1.
\end{array}$$

- For $\ell = -2$

$$\begin{array}{lrcll}
\Delta^2: & \overset{\circ}{W}{}^{2,\,p}_{-2}(\mathbb{R}^N_+)/\mathcal{B}_{[4-N/p]} & \overset{iso}{\longrightarrow} & W^{-2,\,p}_{-2}(\mathbb{R}^N_+), & \text{if } N/p \notin \{1,2\}. \\
\Delta^2: & \overset{\star}{W}{}^{3,\,p}_{-1}(\mathbb{R}^N_+)/\mathcal{B}_{[4-N/p]} & \overset{iso}{\longrightarrow} & W^{-1,\,p}_{-1}(\mathbb{R}^N_+), & \text{if } N/p \neq 1. \\
\Delta^2: & \overset{\star}{W}{}^{4,\,p}_{0}(\mathbb{R}^N_+)/\mathcal{B}_{[4-N/p]} & \overset{iso}{\longrightarrow} & L^p(\mathbb{R}^N_+). &
\end{array}$$

The space $\overset{\star}{W}{}^{m,\,p}_{\alpha}(\mathbb{R}^N_+)$ denotes the subspace of functions u in $W^{m,\,p}_{\alpha}(\mathbb{R}^N_+)$ satisfying $u = 0$ and $\frac{\partial u}{\partial x_N} = 0$ on Γ.

We will now consider the case of very weak solutions of the homogeneous biharmonic problem:

$$(P^0) \quad \Delta^2 u = 0 \ \text{ in } \mathbb{R}^N_+, \quad u = g_0, \ \partial_N u = g_1 \ \text{ on } \Gamma.$$

THEOREM 1.25. *Let $\ell \in \mathbb{Z}$ be such that*

$$\frac{N}{p'} \notin \{1, \dots, \ell - 2\} \ \text{ and } \ \frac{N}{p} \notin \{1, \dots, -\ell + 2\}. \tag{1.3}$$

There exists $C > 0$ such that for any $g_0 \in W^{-1/p,\,p}_{\ell-2}(\Gamma)$ and $g_1 \in W^{-1-1/p,\,p}_{\ell-2}(\Gamma)$ satisfying the compatibility condition

$$\forall \varphi \in \mathcal{B}_{[2+\ell-N/p']}: \quad \langle g_1, \Delta\varphi\rangle_\Gamma - \langle g_0, \partial_N \Delta\varphi\rangle_\Gamma = 0,$$

there exists a unique $u \in W^{0,\,p}_{\ell-2}(\mathbb{R}^N_+)/\mathcal{B}_{[2-\ell-N/p]}$, solution of (P^0), with the estimate

$$\inf_{q \in \mathcal{B}_{[2-\ell-N/p]}} \|u + q\|_{W^{0,\,p}_{\ell-2}(\mathbb{R}^N_+)} \ \leqslant \ C \left(\|g_0\|_{W^{-1/p,\,p}_{\ell-2}(\Gamma)} + \|g_1\|_{W^{-1-1/p,\,p}_{\ell-2}(\Gamma)} \right).$$

PROOF. The existence of solution $u \in W^{0,\,p}_{\ell-2}(\mathbb{R}^N_+)$ of (P^0) is obtained thanks to a duality argument. Let us introduce the following space:

$$Y^p_{\ell,1}(\mathbb{R}^N_+) = \left\{ v \in W^{0,\,p}_{\ell-2}(\mathbb{R}^N_+); \ \Delta^2 v \in W^{0,\,p}_{\ell+2,1}(\mathbb{R}^N_+) \right\}.$$

Under hypothesis (1.3), we show that the space $\mathcal{D}\big(\overline{\mathbb{R}^N_+}\big)$ is dense in $Y^p_{\ell,1}(\mathbb{R}^N_+)$. Next, we deduce that

$$(\gamma_0, \gamma_1) : Y^p_{\ell,1}(\mathbb{R}^N_+) \longrightarrow W^{-1/p,\,p}_{\ell-2}(\Gamma) \times W^{-1-1/p,\,p}_{\ell-2}(\Gamma)$$

is continuous and we have the following Green formula. For any $v \in Y^p_{\ell,1}(\mathbb{R}^N_+)$ and any $\varphi \in \overset{\star}{W}{}^{4,\,p'}_{-\ell+2}(\mathbb{R}^N_+)$,

$$\begin{aligned}&\left\langle \Delta^2 v, \varphi \right\rangle_{W^{0,\,p}_{\ell+2,1}(\mathbb{R}^N_+)\times W^{0,\,p'}_{-\ell-2,-1}(\mathbb{R}^N_+)} - \left\langle v, \Delta^2 \varphi \right\rangle_{W^{0,\,p}_{\ell-2}(\mathbb{R}^N_+)\times W^{0,\,p'}_{-\ell+2}(\mathbb{R}^N_+)}\\ &= \langle v, \partial_N \Delta\varphi\rangle_{W^{-1/p,\,p}_{\ell-2}(\Gamma)\times W^{1/p,\,p'}_{-\ell+2}(\Gamma)} - \langle \partial_N v, \Delta\varphi\rangle_{W^{-1-1/p,\,p}_{\ell-2}(\Gamma)\times W^{1+1/p,\,p'}_{-\ell+2}(\Gamma)}\,.\end{aligned}$$

Finally, Problem (P^0) is equivalent to the variational formulation: Find $u \in Y^p_{\ell,\,1}(\mathbb{R}^N_+)$ such that for any $v \in \overset{\star}{W}{}^{4,\,p'}_{-\ell+2}(\mathbb{R}^N_+)$,

$$\left\langle u, \Delta^2 v \right\rangle_{W^{0,\,p}_{\ell-2}(\mathbb{R}^N_+)\times W^{0,\,p'}_{-\ell+2}(\mathbb{R}^N_+)} \;=\; \langle g_1, \Delta v\rangle_\Gamma - \langle g_0, \partial_N \Delta v\rangle_\Gamma\,.$$

To solve this last problem, we note that for any $f \in W^{0,\,p'}_{-\ell+2}(\mathbb{R}^N_+) \perp \mathcal{B}_{[2-\ell-N/p]}$, the problem

$$\Delta^2 v = f \text{ in } \mathbb{R}^N_+, \quad v = \partial_N v = 0 \text{ on } \Gamma,$$

admits a unique solution $v \in W^{4,\,p'}_{-\ell+2}(\mathbb{R}^N_+)/\mathcal{B}_{[2+\ell-N/p']}$. Because $T : f \longmapsto \langle g_1, \Delta v\rangle_\Gamma - \langle g_0, \partial_N \Delta v\rangle_\Gamma$ is a linear continuous mapping, there exists a unique $u \in W^{0,\,p}_{\ell-2}(\mathbb{R}^N_+)/\mathcal{B}_{[2-\ell-N/p]}$ such that $Tf = \langle u, f\rangle_{W^{0,\,p}_{\ell-2}(\mathbb{R}^N_+)\times W^{0,\,p'}_{-\ell+2}(\mathbb{R}^N_+)}$. □

7. Stokes problem with Dirichlet or Navier boundary conditions

We will now study Stokes system with Dirichlet or Navier boundary conditions (see [4, 8, 10]). We begin with the case of Dirichlet boundary conditions:

$$(S_D)\quad \left\{\begin{array}{rll} -\Delta \boldsymbol{u} + \nabla\pi &= \boldsymbol{f} & \text{in } \mathbb{R}^N_+,\\ \operatorname{div} \boldsymbol{u} &= h & \text{in } \mathbb{R}^N_+,\\ \boldsymbol{u} &= \boldsymbol{g} & \text{on } \Gamma.\end{array}\right.$$

The following theorem gives the existence and uniqueness of generalized solutions with weight 0.

THEOREM 1.26. *For any $\boldsymbol{f} \in \boldsymbol{W}^{-1,\,p}_0(\mathbb{R}^N_+)$, $h \in L^p(\mathbb{R}^N_+)$ and $\boldsymbol{g} \in \boldsymbol{W}^{1-1/p,\,p}_0(\Gamma)$, there exists a unique $(\boldsymbol{u}, \pi) \in \boldsymbol{W}^{1,\,p}_0(\mathbb{R}^N_+) \times L^p(\mathbb{R}^N_+)$ solution to (S_D), with the estimate*

$$\|\boldsymbol{u}\|_{\boldsymbol{W}^{1,\,p}_0(\mathbb{R}^N_+)} + \|\pi\|_{L^p(\mathbb{R}^N_+)} \leqslant C\,(\|\boldsymbol{f}\|_{\boldsymbol{W}^{-1,\,p}_0(\mathbb{R}^N_+)} + \|h\|_{L^p(\mathbb{R}^N_+)} + \|\boldsymbol{g}\|_{\boldsymbol{W}^{1-1/p,\,p}_0(\Gamma)}),$$

where $C > 0$ is a constant depending only on p and N.

PROOF. **Sketch of the proof**. We start by the homogeneous problem:

$$(S^0_D)\qquad -\Delta \boldsymbol{u} + \nabla\pi = \boldsymbol{0} \;\text{ and }\; \operatorname{div} \boldsymbol{u} = 0 \;\text{ in } \mathbb{R}^N_+, \quad \boldsymbol{u} = \boldsymbol{g} \;\text{ on } \Gamma.$$

`Step 1.` We induce the following problems:

$$(P)\qquad \Delta^2 u_N = 0 \;\text{ in } \mathbb{R}^N_+, \quad u_N = g_N \;\text{ and }\; \partial_N u_N = -\operatorname{div}' \boldsymbol{g}' \;\text{ on } \Gamma,$$

$$(Q)\qquad \Delta\pi = 0 \;\text{ in } \mathbb{R}^N_+, \quad \partial_N \pi = \Delta u_N \;\text{ on } \Gamma,$$

$$(R)\qquad \Delta \boldsymbol{u}' = \nabla'\pi \;\text{ in } \mathbb{R}^N_+, \quad \boldsymbol{u}' = \boldsymbol{g}' \;\text{ on } \Gamma.$$

`Step 2.` Because of the regularity of the data, there exist a unique very weak solution $u_N \in W^{1,\,p}_0(\mathbb{R}^N_+)$, a unique very weak solution $\pi \in L^p(\mathbb{R}^+_N)$, and a unique generalized solution $\boldsymbol{u}' \in \boldsymbol{W}^{1,\,p}_0(\mathbb{R}^N_+)$ to Problem (P), (Q) and (R) respectively.

`Step 3.` From (P), (Q), (R), we deduce that $\Delta(\Delta u_N - \partial_N \pi) = 0$ in $\mathbb{R}^N_+$ and $\Delta u_N - \partial_N \pi = 0$ on Γ. Because $\Delta u_N - \partial_N \pi \in W_0^{-1,\,p}(\mathbb{R}^N_+)$, then $-\Delta u_N + \partial_N \pi = 0$ in $\mathbb{R}^N_+$. For the condition of the divergence, we have $\operatorname{div}' \boldsymbol{u}' = \operatorname{div}' \boldsymbol{g}'$ on Γ, where $\operatorname{div}' \boldsymbol{u}' = \sum_{j=1}^{N-1} \frac{\partial u_j}{\partial x_j}$. Since $\Delta \operatorname{div} \boldsymbol{u} = 0$ in $\mathbb{R}^N_+$ and $\operatorname{div} \boldsymbol{u} \in L^p(\mathbb{R}^N_+)$, we deduce that $\operatorname{div} \boldsymbol{u} = 0$ in $\mathbb{R}^N_+$.

`Step 4.` For the uniqueness, we use the uniqueness result of the Laplace's equation with Dirichlet or Neumann boundary condition.

To study the nonhomogeneous problem (S_D), we extend to the whole space the data $\boldsymbol{f}$ and h and the problem is again reduced to the homogeneous case. □

We can also prove existence of strong solutions and some regularity results. More precisely, we have:

THEOREM 1.27. *Assume that* $\frac{N}{p'} \neq 1$. *For any*

$$\boldsymbol{f} \in \boldsymbol{W}_1^{0,\,p}(\mathbb{R}^N_+),\ h \in W_1^{1,\,p}(\mathbb{R}^N_+) \quad \text{and} \quad \boldsymbol{g} \in \boldsymbol{W}_1^{2-1/p,\,p}(\Gamma),$$

there exists a unique $(\boldsymbol{u},\, \pi) \in \boldsymbol{W}_1^{2,\,p}(\mathbb{R}^N_+) \times W_1^{1,\,p}(\mathbb{R}^N_+)$, *solution to* (S_D), *with the estimate*

$$\|\boldsymbol{u}\|_{\boldsymbol{W}_1^{2,\,p}(\mathbb{R}^N_+)} + \|\pi\|_{W_1^{1,\,p}(\mathbb{R}^N_+)} \leqslant C(\|\boldsymbol{f}\|_{\boldsymbol{W}_1^{0,\,p}(\mathbb{R}^N_+)} + \|h\|_{W_1^{1,\,p}(\mathbb{R}^N_+)} + \|\boldsymbol{g}\|_{\boldsymbol{W}_1^{2-1/p,\,p}(\Gamma)}),$$

where $C > 0$ *is a constant depending only on* p *and* N.

COROLLARY 1.28. *Let* $m \in \mathbb{N}$ *and assume that* $\frac{N}{p'} \neq 1$ *if* $m \geqslant 1$. *For any*

$$\boldsymbol{f} \in \boldsymbol{W}_m^{m-1,\,p}(\mathbb{R}^N_+),\ h \in W_m^{m,\,p}(\mathbb{R}^N_+)) \quad \text{and} \quad \boldsymbol{g} \in \boldsymbol{W}_m^{m+1-1/p,\,p}(\Gamma),$$

there exists a unique $(\boldsymbol{u},\, \pi) \in \boldsymbol{W}_m^{m+1,\,p}(\mathbb{R}^N_+) \times W_m^{m,\,p}(\mathbb{R}^N_+)$, *solution to* (S_D), *satisfying the estimate*

$$\|\boldsymbol{u}\|_{\boldsymbol{W}_m^{m+1,\,p}} + \|\pi\|_{W_m^{m,\,p}} \leqslant C(\|\boldsymbol{f}\|_{\boldsymbol{W}_m^{m-1,\,p}(\mathbb{R}^N_+)} + \|h\|_{W_m^{m,\,p}(\mathbb{R}^N_+)} + \|\boldsymbol{g}\|_{\boldsymbol{W}_m^{m+1-1/p,\,p}(\Gamma)}),$$

where $C > 0$ *is a constant depending only on* p, m *and* N.

Using duality argument and the existence and uniqueness results of strong solutions, we prove the following theorem concerning the very weak solutions:

THEOREM 1.29. *Assume that* $\frac{N}{p} \neq 1$. *For all* $\boldsymbol{g} \in \boldsymbol{W}_{-1}^{-1/p,\,p}(\Gamma)$, *there exists a unique* $(\boldsymbol{u},\, \pi) \in \boldsymbol{W}_{-1}^{0,\,p}(\mathbb{R}^N_+) \times W_{-1}^{-1,\,p}(\mathbb{R}^N_+)$, *solution to* (S_D^0), *with the estimate*

$$\|\boldsymbol{u}\|_{\boldsymbol{W}_{-1}^{0,\,p}(\mathbb{R}^N_+)} + \|\pi\|_{W_{-1}^{-1,\,p}(\mathbb{R}^N_+)} \leqslant C\|\boldsymbol{g}\|_{\boldsymbol{W}_{-1}^{-1/p,\,p}(\Gamma)},$$

where $C > 0$ *is a constant depending only on* p *and* N.

For the other behaviour at infinity with various weights, we have:

THEOREM 1.30 (Generalized solutions). *Let* $\ell \in \mathbb{Z}$ *be such that*

$$\frac{N}{p'} \notin \{1, \dots, \ell\} \quad \text{and} \quad \frac{N}{p} \notin \{1, \dots, -\ell\}. \tag{1.4}$$

For any

$$\boldsymbol{f} \in \boldsymbol{W}_\ell^{-1,\,p}(\mathbb{R}^N_+),\ h \in W_\ell^{0,\,p}(\mathbb{R}^N_+) \quad \text{and} \quad \boldsymbol{g} \in \boldsymbol{W}_\ell^{1-1/p,\,p}(\Gamma),$$

satisfying the compatibility condition

$$\forall \boldsymbol{\varphi} \in \boldsymbol{\mathcal{A}}^{\Delta}_{[1+\ell-N/p']}, \qquad \langle \boldsymbol{f} - \nabla h,\, \boldsymbol{\varphi}\rangle_{\boldsymbol{W}^{-1,\,p}_{\ell}(\mathbb{R}^N_+)\times \mathring{\boldsymbol{W}}^{1,\,p'}_{-\ell}(\mathbb{R}^N_+)} +$$
$$+ \left\langle \operatorname{div} \boldsymbol{f},\, \Pi_D \operatorname{div}' \boldsymbol{\varphi}' + \Pi_N \partial_N \varphi_N \right\rangle_{W^{-2,\,p}_{\ell}(\mathbb{R}^N_+)\times \mathring{\boldsymbol{W}}^{2,\,p'}_{-\ell}(\mathbb{R}^N_+)} + \langle \boldsymbol{g},\, \partial_N \boldsymbol{\varphi}\rangle_{\Gamma} = 0,$$

there exists a unique $(\boldsymbol{u},\, \pi) \in \boldsymbol{W}^{1,\,p}_{\ell}(\mathbb{R}^N_+) \times W^{0,\,p}_{\ell}(\mathbb{R}^N_+)/\mathcal{S}^D_{[1-\ell-N/p]}$ *solution to* (S_D), *with the estimate*

$$\inf_{(\boldsymbol{\lambda},\,\mu)\in\mathcal{S}^D_{[1-\ell-N/p]}} \left(\|\boldsymbol{u} + \boldsymbol{\lambda}\|_{\boldsymbol{W}^{1,\,p}_{\ell}(\mathbb{R}^N_+)} + \|\pi + \mu\|_{W^{0,\,p}_{\ell}(\mathbb{R}^N_+)} \right)$$
$$\leqslant C(\ \|\boldsymbol{f}\|_{\boldsymbol{W}^{-1,\,p}_{0}(\mathbb{R}^N_+)} + \|h\|_{W^{0,\,p}_{\ell}(\mathbb{R}^N_+)} + \|\boldsymbol{g}\|_{\boldsymbol{W}^{1-1/p,\,p}_{\ell}(\Gamma)}),$$

where $C > 0$ *is a constant depending only on* p, ℓ *and* N.

The operators Π_D and Π_N are defined as follows:

$$\forall r \in \mathscr{A}^{\Delta}_k, \quad \Pi_D r = \frac{1}{2}\int_0^{x_N} t\, r(x', t)\, dt,$$
$$\forall s \in \mathscr{N}^{\Delta}_k, \quad \Pi_N s = \frac{1}{2}\, x_N \int_0^{x_N} s(x', t)\, dt,$$

and under hypothesis (1.4) we have the following characterization:

$$\mathscr{B}_{[2-l-N/p]} = \Pi_D \mathscr{A}^{\Delta}_{[-l-N/p]} \oplus \Pi_N \mathscr{N}^{\Delta}_{[-l-N/p]}. \tag{1.5}$$

Moreover, a direct calculation with these operators yields the following formulas:

$$\forall r \in \mathscr{A}^{\Delta}_k, \quad \begin{cases} \Delta \Pi_D r = r & \text{in } \mathbb{R}^N_+, \\ \partial_N \Pi_D r = \dfrac{1}{2} x_N r & \text{in } \mathbb{R}^N_+, \\ \Pi_D r = \partial_N \Pi_D r = 0 & \text{on } \Gamma, \end{cases} \tag{1.6}$$

and

$$\forall s \in \mathscr{N}^{\Delta}_k, \quad \begin{cases} \Delta \Pi_N s = s & \text{in } \mathbb{R}^N_+, \\ \partial_N \Pi_N s = \dfrac{1}{2}\left(x_N s + \displaystyle\int_0^{x_N} s(x', t)\, dt \right) & \text{in } \mathbb{R}^N_+, \\ \Pi_N s = \partial_N \Pi_N s = 0 & \text{on } \Gamma. \end{cases} \tag{1.7}$$

In the following lemma, we give the reflection principle for the Stokes system:

LEMMA 1.31. *Let* $\ell \in \mathbb{Z}$ *satisfy*

$$N/p' \notin \{1, \dots, \ell - 1\} \quad \text{and} \quad N/p \notin \{1, \dots, -\ell + 1\}, \tag{1.8}$$

and $(\boldsymbol{u},\, \pi) \in \boldsymbol{W}^{0,\,p}_{\ell-1}(\mathbb{R}^N_+) \times W^{-1,\,p}_{\ell-1}(\mathbb{R}^N_+)$ *satisfy*

$$-\Delta \boldsymbol{u} + \nabla \pi = \boldsymbol{0} \ \text{ and } \ \operatorname{div} \boldsymbol{u} = 0 \ \text{ in } \mathbb{R}^N_+, \quad \boldsymbol{u} = \boldsymbol{0} \ \text{ on } \Gamma.$$

There exists a unique extension $(\tilde{\boldsymbol{u}},\, \tilde{\pi}) \in \boldsymbol{\mathcal{D}}'(\mathbb{R}^N) \times \mathcal{D}'(\mathbb{R}^N)$ *of* $(\boldsymbol{u},\, \pi)$ *satisfying*

$$-\Delta \tilde{\boldsymbol{u}} + \nabla \tilde{\pi} = \boldsymbol{0} \ \text{ and } \ \operatorname{div} \tilde{\boldsymbol{u}} = 0 \quad \text{in } \mathbb{R}^N, \tag{1.9}$$

which is given for all $(\boldsymbol{\varphi},\,\psi)\in\boldsymbol{\mathcal{D}}(\mathbb{R}^N)\times\mathcal{D}(\mathbb{R}^N)$ *by*

$$\begin{aligned}\langle\tilde{\boldsymbol{u}},\,\boldsymbol{\varphi}\rangle=&\int_{\mathbb{R}^N_+}[\boldsymbol{u}\cdot(\boldsymbol{\varphi}-\boldsymbol{\varphi}^*)-2\,u_N\,\varphi_N^*+2\,u_N\,x_N\,(\operatorname{div}\boldsymbol{\varphi})^*]\,\mathrm{d}x\\&+\big\langle\pi,\,2\,x_N\,\varphi_N^*-x_N^2\,(\operatorname{div}\boldsymbol{\varphi})^*\big\rangle_{W^{-1,\,p}_{\ell-1}(\mathbb{R}^N_+)\times\mathring{W}^{1,\,p'}_{-\ell+1}(\mathbb{R}^N_+)}\end{aligned}\tag{1.10}$$

and

$$\begin{aligned}\langle\tilde{\pi},\,\psi\rangle=&\big\langle\pi,\,\psi-\psi^*-2\,x_N\,\partial_N\psi^*\big\rangle_{W^{-1,\,p}_{\ell-1}(\mathbb{R}^N_+)\times\mathring{W}^{1,\,p'}_{-\ell+1}(\mathbb{R}^N_+)}\\&+4\int_{\mathbb{R}^N_+}u_N\,\partial_N\psi^*\,\mathrm{d}x,\end{aligned}\tag{1.11}$$

where $\varphi^*(x)=\varphi(x',\,-x_N)$. *Moreover, we have* $(\tilde{\boldsymbol{u}},\,\tilde{\pi})\in\boldsymbol{W}^{-2,\,p}_{\ell-3}(\mathbb{R}^N)\times W^{-2,\,p}_{\ell-2}(\mathbb{R}^N)$ *with the estimate*

$$\|(\tilde{\boldsymbol{u}},\,\tilde{\pi})\|_{\boldsymbol{W}^{-2,\,p}_{\ell-3}(\mathbb{R}^N)\times W^{-2,\,p}_{\ell-2}(\mathbb{R}^N)}\;\leqslant\;C\,\|(\boldsymbol{u},\,\pi)\|_{\boldsymbol{W}^{0,\,p}_{\ell-1}(\mathbb{R}^N_+)\times W^{-1,\,p}_{\ell-1}(\mathbb{R}^N_+)}.\tag{1.12}$$

Let us consider now the case of Navier boundary conditions:

$$(S_N)\quad\left\{\begin{array}{rcll}-\Delta\boldsymbol{u}+\nabla\pi=\boldsymbol{f}&\text{and}&\operatorname{div}\boldsymbol{u}=h&\text{in }\mathbb{R}^N_+,\\u_N=g_N&\text{and}&\partial_N\boldsymbol{u}'=\boldsymbol{g}'&\text{on }\Gamma.\end{array}\right.$$

Theorem 1.32 (Generalized solutions). *Assume that* $\frac{N}{p'}\neq1$. *For any* $\boldsymbol{f}\in\boldsymbol{W}^{0,\,p}_1(\mathbb{R}^N_+)$, $h\in W^{1,\,p}_1(\mathbb{R}^N_+)$, $g_N\in W^{1-1/p,\,p}_0(\Gamma)$ *and* $\boldsymbol{g}'\in\boldsymbol{W}^{-1/p,\,p}_0(\Gamma)$, *satisfying the compatibility condition:*

$$\forall i\in\{1,\dots,N-1\},\;\int_{\mathbb{R}^N_+}f_i\,\mathrm{d}x=\langle g_i,\,1\rangle_{W^{-1/p,\,p}_0(\Gamma)\times W^{1/p,\,p'}_0(\Gamma)},\quad\text{if } N<p',\tag{1.13}$$

there exists a unique solution $(\boldsymbol{u},\,\pi)\in\boldsymbol{W}^{1,\,p}_0(\mathbb{R}^N_+)\times L^p(\mathbb{R}^N_+)/{\mathcal{P}_{[1-N/p]}}^{N-1}\times\{0\}^2$ *to* (S_N), *with the estimate*

$$\begin{aligned}&\inf_{\boldsymbol{\chi}\in\mathbb{R}^{N-1}\times\{0\}}\|\boldsymbol{u}+\boldsymbol{\chi}\|_{\boldsymbol{W}^{1,\,p}_0(\mathbb{R}^N_+)}+\|\pi\|_{L^p(\mathbb{R}^N_+)}\;\leqslant\\&\qquad C(\|\boldsymbol{f}\|_{\boldsymbol{W}^{0,\,p}_1(\mathbb{R}^N_+)}+\|h\|_{W^{1,\,p}_1(\mathbb{R}^N_+)}+\|g_N\|_{W^{1-1/p,\,p}_0(\Gamma)}+\|\boldsymbol{g}'\|_{\boldsymbol{W}^{-1/p,\,p}_0(\Gamma)}),\end{aligned}$$

where $C>0$ *is a constant depending only on* p *and* N.

Concerning the existence and uniqueness of very weak solutions, we consider here only the homogeneous case:

$$(S^0_N)\quad\left\{\begin{array}{rcll}-\Delta\boldsymbol{u}+\nabla\pi=\boldsymbol{0}&\text{and}&\operatorname{div}\boldsymbol{u}=0&\text{in }\mathbb{R}^N_+,\\u_N=g_N&\text{and}&\partial_N\boldsymbol{u}'=\boldsymbol{g}'&\text{on }\Gamma.\end{array}\right.$$

Theorem 1.33. *Assume that* $\frac{N}{p}\neq1$. *For any* $\boldsymbol{g}'\in\boldsymbol{W}^{-1-1/p,\,p}_{-1}(\Gamma)$ *such that* $\boldsymbol{g}'\perp\mathbb{R}^{N-1}$ *if* $N\leqslant p'$ *and* $g_N\in W^{-1/p,\,p}_{-1}(\Gamma)$, *there exists a unique* $(\boldsymbol{u},\,\pi)\in\boldsymbol{W}^{0,\,p}_{-1}(\mathbb{R}^N_+)\times W^{-1,\,p}_{-1}(\mathbb{R}^N_+)/(\mathcal{P}_{[1-N/p]})^{N-1}\times\{0\}^2$ *solution to* (S^0_N), *with the estimate*

$$\inf_{\boldsymbol{\chi}\in\mathbb{R}^{N-1}\times\{0\}}\|\boldsymbol{u}+\boldsymbol{\chi}\|_{\boldsymbol{W}^{0,\,p}_{-1}}+\|\pi\|_{W^{-1,\,p}_{-1}}\;\leqslant C(\|g_N\|_{W^{-1/p,\,p}_{-1}(\Gamma)}+\|\boldsymbol{g}'\|_{\boldsymbol{W}^{-1-1/p,\,p}_{-1}(\Gamma)}),$$

where $C>0$ *is a constant depending only on* p *and* N.

REMARK 1.34. *A generalized Stokes system was studied by [9]*

$$(S_N^e)\ \begin{cases} -\nu\,\Delta \boldsymbol{u} - \mu\,\nabla \operatorname{div} \boldsymbol{u} + \nabla \pi = \boldsymbol{f} & \text{and} \quad \lambda\,\pi + \operatorname{div} \boldsymbol{u} = h & \text{in } \mathbb{R}^N_+, \\ u_N = g_N & \text{and} \quad \partial_N \boldsymbol{u}' = \boldsymbol{g}' & \text{on } \Gamma, \end{cases}$$

where the constants ν, μ and λ satisfy $\nu > 0$, $\lambda \geqslant 0$ and $\mu + \nu > 0$.

8. Elliptic systems with data in $L^1(\mathbb{R}^N_+)$

The purpose of this section is to give some estimates for div-curl-grad operators and elliptic problems with L^1-data in the half-space. We know that if $f \in L^N(\mathbb{R}^N_+)$, there exists

$$\boldsymbol{u} \in \mathring{\mathbf{W}}{}^{1,N}_0(\mathbb{R}^N_+) \quad \text{such that } \operatorname{div} \boldsymbol{u} = f$$

But does

$$\boldsymbol{u} \in \mathring{\mathbf{W}}{}^{1,N}_0(\mathbb{R}^N_+) \cap \mathbf{L}^\infty(\mathbb{R}^N_+)$$

hold?

THEOREM 1.35. *Let $f \in L^N(\mathbb{R}^N_+)$. Then there exists $\boldsymbol{u} \in \mathring{\mathbf{W}}{}^{1,N}_0(\mathbb{R}^N_+) \cap \mathbf{L}^\infty(\mathbb{R}^N_+)$ such that $\operatorname{div} \boldsymbol{u} = f$ with the following estimate (see [11, 12])*

$$\|\boldsymbol{u}\|_{\mathbf{L}^\infty(\mathbb{R}^N_+)} + \|\boldsymbol{u}\|_{\mathbf{W}^{1,N}_0(\mathbb{R}^N_+)} \leqslant C\,\|f\|_{L^N(\mathbb{R}^N_+)}. \tag{1.14}$$

Thanks to the above theorem, we have the following estimate.

COROLLARY 1.36. *There exists $C > 0$ such that for all $u \in L^{N/(N-1)}(\mathbb{R}^N_+)$, we have the following estimate*

$$\|u\|_{L^{N/(N-1)}(\mathbb{R}^N_+)} \leqslant C \inf_{\boldsymbol{f}+\boldsymbol{g}=\nabla u} \left(\|\boldsymbol{f}\|_{\mathbf{L}^1(\mathbb{R}^N_+)} + \|\boldsymbol{g}\|_{\mathbf{W}^{-1,N/(N-1)}_0(\mathbb{R}^N_+)} \right) \tag{1.15}$$

with $\boldsymbol{f} \in \mathbf{L}^1(\mathbb{R}^N_+)$ and $\boldsymbol{g} \in \mathbf{W}^{-1,N/(N-1)}_0(\mathbb{R}^N_+)$.

THEOREM 1.37. *Let $\boldsymbol{\varphi} \in \mathring{\mathbf{W}}{}^{1,N}_0(\mathbb{R}^N_+)$. Then there exist $\boldsymbol{\psi} \in \mathring{\mathbf{W}}{}^{1,N}_0(\mathbb{R}^N_+) \cap \mathbf{L}^\infty(\mathbb{R}^N_+)$ and $\eta \in \mathring{W}{}^{2,N}_0(\mathbb{R}^N_+)$ such that*

$$\boldsymbol{\varphi} = \boldsymbol{\psi} + \nabla \eta.$$

Moreover, we have the following estimate

$$\|\boldsymbol{\psi}\|_{\mathbf{W}^{1,N}_0(\mathbb{R}^N_+)} + \|\boldsymbol{\psi}\|_{\mathbf{L}^\infty(\mathbb{R}^N_+)} + \|\eta\|_{W^{2,N}_0(\mathbb{R}^N_+)} \leqslant C\|\boldsymbol{\varphi}\|_{\mathbf{W}^{1,N}_0(\mathbb{R}^N_+)}. \tag{1.16}$$

We will study now some elliptic problems with data given in $L^1(\mathbb{R}^N_+)$. First, we set

$$\mathbf{X}(\mathbb{R}^N_+) = \{\boldsymbol{f} \in \mathbf{L}^1(\mathbb{R}^N_+),\ \operatorname{div} \boldsymbol{f} \in W_0^{-2,N/(N-1)}(\mathbb{R}^N_+)\},$$

which is Banach space endowed with the following norm

$$\|\boldsymbol{f}\|_{\mathbf{X}} = \|\boldsymbol{f}\|_{\mathbf{L}^1} + \|\operatorname{div} \boldsymbol{f}\|_{W_0^{-2,N/(N-1)}}.$$

THEOREM 1.38. *Let $\boldsymbol{f} \in \mathbf{X}(\mathbb{R}^N_+)$. Then for any $\boldsymbol{\varphi} \in \mathcal{D}(\mathbb{R}^N_+)$,*

$$|<\boldsymbol{f}, \boldsymbol{\varphi}>| \leqslant C\,\|\boldsymbol{f}\|_{\mathbf{X}} \|\nabla \boldsymbol{\varphi}\|_{\mathbf{L}^N}.$$

By density of $\mathcal{D}(\mathbb{R}^N_+))$ in $\mathring{\mathbf{W}}{}^{1,N}_0(\mathbb{R}^N_+)$, we have $\boldsymbol{f} \in \mathbf{W}^{-1,N/(N-1)}_0(\mathbb{R}^N_+)$ and the following estimate holds

$$\|\boldsymbol{f}\|_{\mathbf{W}^{-1,N/(N-1)}_0(\mathbb{R}^N_+)} \leqslant C\,\|\boldsymbol{f}\|_{\mathbf{X}(\mathbb{R}^N_+)}.$$

COROLLARY 1.39. *Let* $\boldsymbol{f} \in \mathbf{X}(\mathbb{R}^N_+)$. *Then the following problem*

$$-\Delta \boldsymbol{u} = \boldsymbol{f} \ \text{ in } \mathbb{R}^N_+ \ \text{ and } \ \boldsymbol{u} = \mathbf{0} \ \text{ on } \Gamma = \mathbb{R}^{N-1}, \tag{1.17}$$

has a unique solution $\boldsymbol{u} \in \mathbf{W}_0^{1,N/(N-1)}(\mathbb{R}^N_+)$ *and we have the following estimate*

$$|| \boldsymbol{u} ||_{\mathbf{W}_0^{1,N/(N-1)}(\mathbb{R}^N_+)} \leqslant C || \boldsymbol{f} ||_{\mathbf{X}(\mathbb{R}^N_+)}.$$

THEOREM 1.40.

(i) Let $\boldsymbol{f} \in \mathbf{L}^3(\mathbb{R}^3_+)$ *such that* $\operatorname{div} \boldsymbol{f} = 0$ *in* $\mathbb{R}^3_+$ *and* $f_3 = 0$ *on* Γ. *Then there exists* $\boldsymbol{\psi} \in \mathring{\mathbf{W}}{}_0^{1,3}(\mathbb{R}^3_+) \cap \mathbf{L}^\infty(\mathbb{R}^3_+)$ *such that* $\boldsymbol{f} = \mathbf{curl}\, \boldsymbol{\psi}$ *and we have the following estimate*

$$|| \boldsymbol{\psi} ||_{\mathbf{W}_0^{1,3}(\mathbb{R}^3_+)} + || \boldsymbol{\psi} ||_{\mathbf{L}^\infty(\mathbb{R}^3_+)} \leqslant C || \boldsymbol{f} ||_{\mathbf{L}^3(\mathbb{R}^3_+)}.$$

(ii) Let $\boldsymbol{f} \in \mathbf{L}^3(\mathbb{R}^3_+)$. *Then there exist* $\boldsymbol{\varphi} \in \mathring{\mathbf{W}}{}_0^{1,3}(\mathbb{R}^3_+) \cap \mathbf{L}^\infty(\mathbb{R}^3_+)$ *and* $\pi \in W_0^{1,3}(\mathbb{R}^3_+)$ *unique up to an additive constant and satisfying*

$$\boldsymbol{f} = \mathbf{curl}\, \boldsymbol{\varphi} + \nabla \pi.$$

Moreover, we have the following estimate

$$|| \boldsymbol{\varphi} ||_{\mathbf{W}_0^{1,3}(\mathbb{R}^3_+)} + || \boldsymbol{\varphi} ||_{\mathbf{L}^\infty(\mathbb{R}^3_+)} + ||\nabla \pi||_{\mathbf{L}^3(\mathbb{R}^3_+)} \leqslant C || \boldsymbol{f} ||_{\mathbf{L}^3(\mathbb{R}^3_+)}.$$

COROLLARY 1.41. *Let* $\boldsymbol{f} \in \mathbf{L}^1(\mathbb{R}^3_+)$ *such that* $\operatorname{div} \boldsymbol{f} = 0$. *For all* $\boldsymbol{\varphi} \in \mathring{\mathbf{W}}{}_0^{1,3}(\mathbb{R}^3_+)$, *we have the following estimate*

$$| < \boldsymbol{f}, \boldsymbol{\varphi} > | \leqslant C || \boldsymbol{f} ||_{\mathbf{L}^1} ||\mathbf{curl}\, \boldsymbol{\varphi}||_{\mathbf{L}^3}.$$

THEOREM 1.42. *Let* $\boldsymbol{f} \in \mathbf{L}^1(\mathbb{R}^N_+) + \mathbf{W}_0^{-1,N/(N-1)}(\mathbb{R}^N_+)$ *satisfy the following compatibility condition*

$$\forall \boldsymbol{v} \in \mathbf{V}_0^{1,N}(\mathbb{R}^N_+) \cap \mathbf{L}^\infty(\mathbb{R}^N_+), \quad < \boldsymbol{f}, \boldsymbol{v} > = 0, \tag{1.18}$$

with

$$\mathbf{V}_0^{1,N}(\mathbb{R}^N_+) = \{ \boldsymbol{v} \in \mathring{\mathbf{W}}{}_0^{1,N}(\mathbb{R}^3_+), \ \operatorname{div} \boldsymbol{v} = 0 \text{ in } \mathbb{R}^N_+ \}.$$

Then there exists a unique $\pi \in L^{N/(N-1)}(\mathbb{R}^N_+)$ *such that* $\boldsymbol{f} = \nabla \pi$.

PROPOSITION 1.43. *Let* $\boldsymbol{f} \in \mathbf{L}^1(\mathbb{R}^3_+)$ *such that* $\operatorname{div} \boldsymbol{f} = 0$ *in* $\mathbb{R}^3_+$. *Then there exists a unique* $\boldsymbol{\varphi} \in \mathbf{L}^{3/2}(\mathbb{R}^3_+)$ *such that* $\mathbf{curl}\, \boldsymbol{\varphi} = \boldsymbol{f}$, $\operatorname{div} \boldsymbol{\varphi} = 0$ *in* $\mathbb{R}^3_+$ *and* $\varphi_3 = 0$ *on* Γ *satisfying the following estimate*

$$|| \boldsymbol{\varphi} ||_{\mathbf{L}^{3/2}(\mathbb{R}^3_+)} \leqslant C || \boldsymbol{f} ||_{\mathbf{L}^1(\mathbb{R}^3_+)}.$$

We set

$$\mathbf{H}_p(\mathbb{R}^N_+) = \{ \boldsymbol{v} \in \mathbf{L}^p(\mathbb{R}^N_+), \ \operatorname{div} \boldsymbol{v} = 0 \text{ in } \mathbb{R}^N_+, \ v_N = 0 \text{ on } \Gamma \}.$$

PROPOSITION 1.44. *Let* $\boldsymbol{f} \in \mathbf{L}^1(\mathbb{R}^3_+) + \mathbf{W}_0^{-1,3/2}(\mathbb{R}^3_+)$ *such that* $\operatorname{div} \boldsymbol{f} = 0$. *Then there exists a unique* $\boldsymbol{\varphi} \in \mathbf{L}^{3/2}(\mathbb{R}^3_+)$ *such that* $\mathbf{curl}\, \boldsymbol{\varphi} = \boldsymbol{f}$ *and* $\operatorname{div} \boldsymbol{\varphi} = 0$ *in* $\mathbb{R}^3_+$ *satisfying the following estimate*

$$|| \boldsymbol{\varphi} ||_{\mathbf{L}^{3/2}(\mathbb{R}^3_+)} \leqslant C || \boldsymbol{f} ||_{\mathbf{L}^1(\mathbb{R}^3_+) + \mathbf{W}_0^{-1,3/2}(\mathbb{R}^3_+)}.$$

THEOREM 1.45. *Let $\boldsymbol{f} \in \mathbf{X}(\mathbb{R}^3_+)$. Then there exists a unique $\boldsymbol{\varphi} \in \mathbf{L}^{3/2}(\mathbb{R}^3_+)$ such that* $\operatorname{div} \boldsymbol{\varphi} = 0$ *with* $\varphi_3 = 0$ *on* Γ *and a unique* $p \in L^{3/2}(\mathbb{R}^3_+)$ *satisfying*

$$\boldsymbol{f} = \mathbf{curl}\, \boldsymbol{\varphi} + \nabla p$$

and the following estimate holds

$$|| \boldsymbol{\varphi} ||_{\mathbf{L}^{3/2}(\mathbb{R}^3_+)} + || p ||_{L^{3/2}(\mathbb{R}^3_+)} \leqslant C || \boldsymbol{f} ||_{\mathbf{X}(\mathbb{R}^3_+)}.$$

We can finally solve the following elliptic systems.

THEOREM 1.46.

(i) Let $\boldsymbol{g}' \in \mathbf{L}^1(\Gamma)$ and $g_N \in W_0^{-1+\frac{1}{N},\frac{N}{N-1}}(\Gamma)$ satisfy the compatibility conditions $\int_\Gamma \boldsymbol{g}' = \mathbf{0}$ and $< g_N, 1 >= 0$. If $\operatorname{div}' \boldsymbol{g}' \in W_0^{-2+\frac{1}{N},\frac{N}{N-1}}(\Gamma)$, then the system

$$-\Delta \boldsymbol{u} = \mathbf{0} \ \text{ in } \mathbb{R}^N_+ \quad \text{and} \quad \boldsymbol{u} = \boldsymbol{g} \text{ on } \Gamma$$

has a unique very weak solution $\boldsymbol{u} \in \mathbf{L}^{N/(N-1)}(\mathbb{R}^N_+)$.

(ii) Let $\boldsymbol{f} \in \mathbf{L}^1(\mathbb{R}^N_+)$, $\boldsymbol{g}' \in \mathbf{L}^1(\Gamma)$ and $g_N \in W_0^{-1+\frac{1}{N},\frac{N}{N-1}}(\Gamma)$ satisfy the compatibility condition $\int_{\mathbb{R}^N_+} \boldsymbol{f}' + \int_\Gamma \boldsymbol{g}' = \mathbf{0}$ and $\int_{\mathbb{R}^N_+} f_N + < g_N, 1 >= 0$. If

$$[\boldsymbol{f}, \boldsymbol{g}'] = \sup_{\xi \in W_0^{2,N}(\mathbb{R}^N_+),\ \xi \neq 0} \frac{| \int_{\mathbb{R}^N_+} \boldsymbol{f} \cdot \nabla \xi + \int_\Gamma \boldsymbol{g}' \cdot \nabla' \xi |}{|| \xi ||_{W_0^{2,N}(\mathbb{R}^N_+)}} < \infty,$$

then the system

$$-\Delta \boldsymbol{u} = \boldsymbol{f} \ \text{ in } \mathbb{R}^N_+ \quad \text{and} \quad \frac{\partial \boldsymbol{u}}{\partial x_N} = \boldsymbol{g} \text{ on } \Gamma$$

has a unique solution $\boldsymbol{u} \in \mathbf{W}_0^{1,N/(N-1)}(\mathbb{R}^N_+)$.

(iii) Let $\boldsymbol{f} \in \mathbf{L}^1(\mathbb{R}^N_+)$ such that $\operatorname{div} \boldsymbol{f} \in [W_0^{2,N}(\mathbb{R}^N_+) \cap \overset{\circ}{W}{}_{-1}^{1,N}(\mathbb{R}^N_+)]'$ and $\int_{\mathbb{R}^N_+} f_N = 0$. Then $\boldsymbol{f} \in \mathbf{W}_0^{-1,N/(N-1)}(\mathbb{R}^N_+)$ and the system

$$-\Delta \boldsymbol{u} = \boldsymbol{f} \ \text{ in } \mathbb{R}^N_+; \quad \boldsymbol{u}' = \mathbf{0} \quad \text{and} \quad \frac{\partial u_N}{\partial x_N} = 0 \ \text{ on } \Gamma$$

has a unique solution $\boldsymbol{u} \in \mathbf{W}_0^{1,N/(N-1)}(\mathbb{R}^N_+)$.

(iv) Let $\boldsymbol{f} \in \mathbf{L}^1(\mathbb{R}^N_+)$ such that $\int_{\mathbb{R}^N_+} \boldsymbol{f}\,' = \mathbf{0}$. If

$$[\boldsymbol{f}] = \sup_{\xi \in \mathcal{D}(\overline{\mathbb{R}^N_+}),\ \frac{\partial \xi}{\partial x_N} = 0 \ \text{on } \Gamma} \frac{| \int_{\mathbb{R}^N_+} \boldsymbol{f} \cdot \nabla \xi |}{|| \xi ||_{W_0^{2,N}(\mathbb{R}^N_+)}} < \infty$$

holds, then the system

$$-\Delta \boldsymbol{u} = \boldsymbol{f} \ \text{ in } \mathbb{R}^N_+, \quad u_N = 0 \quad \text{and} \quad \frac{\partial \boldsymbol{u}'}{\partial x_N} = \mathbf{0} \quad \text{on } \Gamma$$

has a unique solution $\boldsymbol{u} \in \mathbf{W}_0^{1,N/(N-1)}(\mathbb{R}^N_+)$.

Bibliography

[1] C. Amrouche, Traces in the half-space for weighted Sobolev spaces $W^{m,p}_{\alpha}(\mathbb{R}^n_+)$. In preparation.

[2] C. Amrouche, V. Girault, J. Giroire. Weighted Sobolev spaces for Laplace's equation in $\mathbb{R}^n$, *J. Math. Pures Appl.* **73-6** (1994), 579–606.

[3] C. Amrouche, S. Nečasová, Laplace equation in the half-space with a nonhomogeneous Dirichlet boundary condition, *Mathematica Bohemica* **126-2** (2001), 265–274.

[4] C. Amrouche, S. Nečasová, Y. Raudin, Very weak, generalized and strong solutions to the Stokes system in the half space, *J. Differential Equations* **244** (2008), 887–915.

[5] C. Amrouche, H. H. Nguyen, New estimates for the div-curl-grad operators and elliptic problems with L^1-data in the whole space and in the half-space, *J. Differential Equations* **250-7** (2011), 3150–3195.

[6] C. Amrouche, Y. Raudin, Nonhomogeneous biharmonic problem in the half-space, L^p theory and generalized solutions, *J. Differential Equations* **236** (2007), 57– 81.

[7] C. Amrouche, Y. Raudin, Singular boundary conditions and regularity for the biharmonic problem in the half-space, *Comm. Pure Appl. Anal.* **6-4** (2007), 957–982.

[8] C. Amrouche, Y. Raudin, Reflection principles and kernels in $\mathbb{R}^n_+$ for the biharmonic and Stokes operators. Solutions in a large class of weighted Sobolev spaces, *Adv. Differential Equations* **15** (2010), no. 3-4, 201–230.

[9] H. Beirao da Veiga, Regularity of solutions to a non-homogeneous boundary value problem for general Stokes systems in $\mathbb{R}^n_+$, *Math. Ann.* **331-1** (2005), 203–217.

[10] T. Z. Boulmezaoud, On the Stokes system and the biharmonic equation in the half-space: an approach via weighted Sobolev spaces, *Math. Meth. Appl. Sci.* **25** (2002) 373–398.

[11] J. Bourgain, H. Brézis, On the equation $\operatorname{div} Y = f$ and application to control of phases, *Journal of the American Mathematical Society* **16-2** (2002), 393–426.

[12] J. Bourgain, H. Brézis, New estimates for elliptic equations and Hodge type systems, *Journal of the European Mathematical Society* **9-2** (2007), 277–315.

[13] H. Brézis, J. V. Schaftingen, Boundary extimates for elliptic systems with $\mathbf{L}^1$-data, *Calculus of Variations and Partial Differential Equations* **30-3** (2007), 369–388.

[14] L. Cattabriga, Su un problema al contorno relativo al sistema di equazioni di Stokes, *Rend. Sem. Mat. Padova* **31** (1961), 308–340.

[15] R. J. Duffin, Continuation of biharmonic functions by reflection, *Duke Math. J.* **22** (1955), 313–324.

[16] R. Farwig, A Note on the Reflection Principle for the Biharmonic Equation and the Stokes system, *Acta Appl. Math.* **25** (1994), 41–51.

[17] R. Farwig, and H. Sohr, On the Stokes and Navier–Stokes system for domains with noncompact boundary in L^q-spaces, *Math. Nachr.* **170** (1994), 53–77.

[18] V. Maz'ya, T. Shaposhnikova, On the Bourgain, Brezis, and Mironescu theorem concerning limiting embeddings of fractional Sobolev spaces, *J. Funct. Anal.* **195-2** (2002), 230–238.

[19] V. G. Maz'ya, B. A. Plamenevskiĭ, L. I. Stupyalis, The three-dimentional problem of steady-state motion of a fluid with a free surface, *Amer. Math. Soc. Transl.* **123** (1984), 171–268.

[20] N. Tanaka, On the boundary value problem for the stationary Stokes system in the half-space, *J. Differential Equations* **115** (1995), 70–74.

[21] S. Ukai, A solution formula for the Stokes equation in R^n_+, *Comm. Pure Appl. Math.* **40** (1987), no. 5, 611–621.

Part 2

Generalised trigonometric functions, compact operators and the p-Laplacian

David E. Edmunds

2000 *Mathematics Subject Classification.* 35P30, 46E30, 47A75, 47B06, 47B40

Key words and phrases. embeddings, s-numbers, compact linear operators, p-Laplacian, generalised trigonometric functions, Schauder basis

Abstract. A survey is given of some recent developments involving the generalised trigonometric functions and the representation of compact linear operators acting between Banach spaces. There are applications to the Dirichlet problem for the $p-$Laplacian.

Contents

CHAPTER 1

Generalised trigonometric functions, compact operators and the $p-$Laplacian

1. Introduction

The lectures on which these notes are based had two main components: the theory of generalised trigonometric functions and the representation in series form of the action of a compact linear operator acting between Banach spaces. Part of the motivation for the study of the first topic arises from the Dirichlet problem for the one-dimensional $p-$Laplacian $(1 < p < \infty)$ on the unit interval $(0, 1)$: this asks for the existence of u and λ such that

$$-\Delta_p u := -\left(|u'|^{p-2}\, u'\right)' = \lambda\, |u|^{p-2}\, u \text{ on } (0,1), u(0) = u(1) = 0. \qquad (1.1)$$

The case $p = 2$ is a familiar question for the Laplace operator, with eigenvalues $(n\pi)^2$ and corresponding eigenvectors $\sin(n\pi t)$ $(n \in \mathbb{N})$. For a general $p \in (1, \infty)$ it turns out (see, for example, [4]) that the problem has eigenvalues

$$\lambda_n = (p-1)(n\pi_p)^p, \text{ where } \pi_p = \frac{2\pi}{p \sin(\pi/p)},$$

and associated eigenfunctions $\sin_p(n\pi_p t)$ $(n \in \mathbb{N})$. Here the function $\sin_p$ is a generalisation of the classical sine function that has properties in common with (as well as differences from) it. However, the interest in such $p-$trigonometric functions is not solely dependent upon the $p-$Laplacian, and we endeavour to make the case that there are now so many remarkable formulae and identities involving them that analysts would do well to be acquainted with them. For additional details of the topic we refer to [3], [8] and the references given in these works.

The object of the second component is to obtain a Banach space analogue of the Schmidt representation of compact linear operators acting between Hilbert spaces. This is based on the recent work [6] in which a representation in series form is obtained of the action of a compact linear map $T : X \to Y$ when the only conditions imposed are that X, Y are reflexive Banach spaces with strictly convex duals. The proof proceeds by means of a sequential procedure based on the familiar process used when X and Y are Hilbert spaces, and results in the construction of a decreasing sequence of subspaces X_n of X with intersection contained in the kernel of T : for each n, λ_n is the norm of the restriction of T to X_n, attained at a point $x_n \in X_n$ with unit norm. The λ_n and x_n correspond to an 'eigenvalue' and 'eigenvector' respectively of a nonlinear operator equation involving a duality map that becomes the identity map in the Hilbert space case. Application of these abstract results to the case in which T is the embedding of a Sobolev space in

a Lebesgue space gives the existence of a countable family of 'eigenvalues' of the Dirichlet problem for the $p-$Laplacian. The relationship between these eigenvalues and eigenvectors and their classical counterparts obtained via the $p-$trigonometric functions mentioned above is discussed.

2. The $p-$trigonometric functions

Throughout we shall assume that $p \in (1, \infty)$. Define $F_p : [0,1] \to \mathbb{R}$ by

$$F_p(x) = \int_0^x (1-t^p)^{-1/p} dt, \ x \in [0,1].$$

Note that $F_2 = \sin^{-1}$. Since F_p is strictly increasing it has an inverse, denoted by $\sin_p$ to emphasise its connection with the usual sine function, and defined on the interval $[0, \pi_p/2]$, where

$$\begin{aligned}\pi_p/2 = \sin_p^{-1}(1) &= \int_0^1 (1-t^p)^{-1/p} dt = p^{-1} \int_0^1 (1-s)^{-1/p} s^{-1/p'} ds \\ &= p^{-1}\mathbf{B}(1/p', 1/p),\end{aligned}$$

where $\mathbf{B}$ is the usual beta function and $p' = p/(p-1)$. Use of the Euler reflection formula for the Gamma function shows that

$$\pi_p = \frac{2\pi}{p \sin(\pi/p)}.$$

Clearly $\pi_2 = \pi$ and

$$p\pi_p = 2\Gamma(1/p')\Gamma(1/p) = p'\pi_{p'}.$$

In addition, π_p decreases as p increases, and

$$\lim_{p\to 1} \pi_p = \infty, \ \lim_{p\to\infty} \pi_p = 2, \ \lim_{p\to 1}(p-1)\pi_p = \lim_{p\to 1} \pi_{p'} = 2.$$

The function $\sin_p$ is strictly increasing on $[0, \pi_p/2]$, $\sin_p(0) = 0$ and $\sin_p(\pi_p/2) = 1$. It may be extended to $[0, \pi_p]$ by defining $\sin_p x = \sin_p(\pi_p - x)$ for $x \in [\pi_p/2, \pi_p]$; further extension to $[-\pi_p, \pi_p]$ is made by oddness, and finally $\sin_p$ is extended to the whole of $\mathbb{R}$ by $2\pi_p-$periodicity. This extension belongs to $C^1(\mathbb{R})$.

Now define $\cos_p : \mathbb{R} \to \mathbb{R}$ by

$$\cos_p x = \frac{d}{dx} \sin_p x, \ x \in \mathbb{R}.$$

Plainly $\cos_p$ is even, $2\pi_p-$periodic and odd about $\pi_p/2$. If $x \in [0, \pi_p/2]$ and we put $y = \sin_p x$, then

$$\cos_p x = (1-y^p)^{1/p} = (1-(\sin_p x)^p)^{1/p}. \tag{1.2}$$

Hence $\cos_p$ is strictly decreasing on $[0, \pi_p/2]$, $\cos_p(0) = 1$ and $\cos_p(\pi_p/2) = 0$; moreover,

$$|\sin_p x|^p + |\cos_p x|^p = 1. \tag{1.3}$$

This is clear from (1.2) if $x \in [0, \pi_p/2]$ and follows for all $x \in \mathbb{R}$ by symmetry and periodicity. Analogues of the other trigonometric functions may be given in the natural way: thus $\tan_p$ is defined by

$$\tan_p x = \frac{\sin_p x}{\cos_p x}$$

whenever $\cos_p x \neq 0$; that is, for all $x \in \mathbb{R}$ except for the points $(k+1/2)\pi_p$ $(k \in \mathbb{Z})$. Evidently $\tan_p$ is odd and π_p-periodic, while $\tan_p(0) = 0$. Note that when $p \neq 2$ the extended $\sin_p$ function does not have the smoothness properties of its classical counterpart. In particular, it is not in $C^\infty(\mathbb{R})$: for example, its second derivative at x is $-h(\sin_p x)$, where

$$h(y) = (1 - y^p)^{\frac{2}{p}-1} y^{p-1},$$

and so is not continuous at $\pi_p/2$ if $2 < p < \infty$. Nevertheless, $\sin_p$ is of class C^∞ on $[0, \pi_p/2)$.

To illustrate the behaviour of the $p-$trigonometric functions when differentiated, we give the following identities, which are immediate consequences of the definitions and (1.3):

$$\frac{d}{dx}\cos_p x = -\sin_p^{p-1} x \cos_p^{2-p} x, \ \frac{d}{dx}\tan_p x = 1 + \tan_p^p x,$$
$$\frac{d}{dx}\cos_p^{p-1} x = -(p-1)\sin_p^{p-1} x, \ \frac{d}{dx}\sin_p^{p-1} x = (p-1)\sin_p^{p-2} x \cos_p x.$$

Here $x \in (0, \pi_p/2)$. The classical Jordan inequality has a $p-$analogue:

$$\frac{2}{\pi_p} \leq \frac{\sin_p x}{x} < 1 \text{ for all } x \in (0, \pi_p/2]. \tag{1.4}$$

To establish this, use a change of variable to obtain

$$\sin_p^{-1} x = x \int_0^1 (1 - x^p s^p)^{-1/p} ds,$$

from which we have

$$x = (\sin_p x) \int_0^1 (1 - (\sin_p x)^p s^p)^{-1/p} ds.$$

Since

$$1 \leq \int_0^1 (1 - (\sin_p x)^p s^p)^{-1/p} ds \leq \frac{\pi_p}{2},$$

the result follows.

Connections with functions occurring in classical analysis are important. From

$$\sin_p^{-1} x = \frac{x}{p} \int_0^1 t^{-1/p'} (1 - x^p t)^{-1/p} dt,$$

we have the representations, valid for $0 \leq x < 1$,

$$\sin_p^{-1} x = xF(1/p, 1/p; 1 + 1/p; x^p) = x(1 - x^p)^{1/p'} F(1, 1; 1 + 1/p; x^p),$$

where F is the hypergeometric function (see [1], Theorems 2.2.1 and 2.2.5). Since

$$F(a, b; c; x) = \sum_{n=0}^{\infty} \frac{\Gamma(a+n)\Gamma(b+n)\Gamma(c)}{\Gamma(a)\Gamma(b)\Gamma(c+n)} \frac{x^n}{n!},$$

we obtain the power series expansion of $\sin_p^{-1} x$ as

$$\sin_p^{-1} x = x \sum_{n=0}^{\infty} \frac{\Gamma(n + 1/p)}{(np+1)\Gamma(1/p)} \frac{x^{np}}{n!} \ (0 \leq x < 1). \tag{1.5}$$

From this an expansion of $\sin_p x$ may be obtained, the first three terms being

$$\sin_p x = x - \frac{1}{p(p+1)}x^{p+1} - \frac{(p^2-2p-1)}{2p^2(p+1)(2p+1)}x^{2p+1} + ... \ (0 \le x < \pi_p/2).$$

The later terms have very complicated coefficients, with no obvious regular pattern.

Turning next to integration, it is elementary to show that for all $k, l \ge 0$,

$$\int_0^{\pi_p/2} \sin_p^k x \cos_p^l x dx = \frac{1}{p} B\left(\frac{k+1}{p}, 1 + \frac{l-1}{p}\right). \tag{1.6}$$

At a slightly more sophisticated level we have, as a consequence of (1.5),

$$x = \sin_p x \sum_{n=0}^{\infty} \frac{\Gamma(n+1/p)}{(np+1)\Gamma(1/p)} \frac{(\sin_p x)^{np}}{n!}, \ 0 \le x < \frac{\pi_p}{2},$$

and so, with the aid of (1.6),

$$\int_0^{\pi_p/2} \frac{x}{\sin_p x} dx = \frac{\pi_p}{2} \sum_{n=0}^{\infty} \left(\frac{\Gamma(n+1/p)}{n!\Gamma(1/p)}\right)^2 \frac{1}{np+1}.$$

Since it is known that (see [10], 1.7.4)

$$\int_0^{\pi/2} \frac{x}{\sin x} dx = 2G,$$

where G is the Catalan constant defined by

$$G = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)^2},$$

this gives a representation of G in the form

$$G = \frac{\pi}{4} \sum_{n=0}^{\infty} \left(\frac{(2n)!}{(n!)^2 2^{2n}}\right)^2 \frac{1}{2n+1}.$$

Another entertaining result is that

$$\int_0^1 \frac{\sin_p^{-1} x}{x} dx = -\frac{\pi_p}{2p}\left(\frac{\Gamma'(1/p)}{\Gamma(1/p)} + \gamma\right),$$

where γ is Euler's constant.

Now we turn to the basis properties of the $\sin_p$ functions, beginning with the Fourier sine coefficients of $\sin_p(n\pi_p t)$. Given $n \in \mathbb{N}$ we write

$$f_{n,p}(t) = \sin_p(n\pi_p t), \ e_n = f_{n,2},$$

so that $e_n(t) = \sin(n\pi t)$. Each $f_{n,p}$ belongs to $C^1([0,1])$ and so is continuous with bounded variation on $[0,1]$: thus it has a Fourier sine expansion:

$$f_{n,p}(t) = \sum_{k=1}^{\infty} \widehat{f_{n,p}}(k) \sin(k\pi t), \ \widehat{f_{n,p}}(k) = 2\int_0^1 f_{n,p}(t) \sin(k\pi t) dt.$$

The symmetry of $f_{1,p}$ about $t = 1/2$ implies that $\widehat{f_{1,p}}(k) = 0$ when k is even and that

$$\widehat{f_{n,p}}(k) = \begin{cases} \widehat{f_{1,p}}(m), & \text{if } mn = k \text{ for some odd } m, \\ 0, & \text{otherwise.} \end{cases}$$

For economy of expression put $\tau_m(p) = \widehat{f_{1,p}}(m)$. Since all the Fourier sine coefficients of the $f_{n,p}$ may be expressed in terms of the $\tau_m(p)$, we concentrate on the behaviour of these terms. For even m, $\tau_m(p) = 0$. When m is odd, say $m = 2k+1$, integration by parts plus change of variable gives

$$\begin{aligned}\tau_{2k+1}(p) &= 4\int_0^{1/2} \sin_p(\pi_p t)\sin((2k+1)\pi t)dt \\ &= \frac{4\pi_p}{(2k+1)^2\pi^2}\int_0^1 \sin\left(\frac{(2k+1)\pi}{\pi_p}\cos_p^{-1} x\right)dx. \end{aligned} \tag{1.7}$$

From (1.7) it is immediate that

$$|\tau_{2k+1}(p)| \le \frac{4\pi_p}{(2k+1)^2\pi^2} \ (k \in \mathbb{N}). \tag{1.8}$$

With slightly more effort and using an estimate of van der Corput type to deal with the oscillatory integral, it can be shown that if $1 < p < 2$, then the $\tau_{2k+1}(p)$ decay faster as k increases:

$$|\tau_{2k+1}(p)| \le \frac{16\pi_p^2}{m_p(2k+1)^3\pi^3} \ (k \in \mathbb{N}), \tag{1.9}$$

where $m_p = \left\{(2-p)^{-(2-p)}(p-1)^{-(p-1)}\right\}^{1/p}$.

It is well known that $(\exp(in\pi x))_{n\in\mathbb{N}}$ is a (Schauder) basis in $L_q(-1,1)$ for every $q \in (1,\infty)$: see, for example, [9], 12.10.1. Given any element of $L_q(0,1)$, its odd extension to $L_q(-1,1)$ has a unique representation in terms of the functions $\sin(n\pi x)$, which means that $(\sin(n\pi x))$ is a basis of $L_q(0,1)$. Here we discuss a recent paper [2], the object of which was to show that if p is not too close to 1, then the functions $\sin_p(n\pi_p x)$ have the same basis property. The analysis presented here is based on [3] and [2].

Given any function f on $[0,1)$, extend it to a function $\widetilde{f}$ on $[0,\infty)$ by setting $\widetilde{f}(t) = -\widetilde{f}(2k-t)$ for $t \in [k,k+1)$, $k \in \mathbb{N}$; define $M_m : L_q(0,1) \to L_q(0,1)$ by $M_m g(t) = \widetilde{g}(mt)$ ($m \in \mathbb{N}$, $1 < q < \infty$) and note that $M_m e_n = e_{mn}$. In [2] it is shown that M_m is a linear isometry and that the map T defined by $Tg(t) = \sum_{m=1}^{\infty}\tau_m(p)M_m g(t)$ is a bounded linear map of $L_q(0,1)$ to itself with the property that for all $n \in \mathbb{N}$, $Te_n = f_{n,p}$. If it can be shown that T is a homeomorphism, then it will follow from standard results (see [16], p. 75) that the $f_{n,p}$ form a basis in $L_q(0,1)$ for every $q \in (1,\infty)$. A sufficient condition for T to be a homeomorphism is that it is not too far from the identity map. More precisely, since M_1 is the identity map id and each M_m is a linear isometry,

$$\|T - \tau_1(p)\mathrm{id}\| \le \sum_{k=1}^{\infty}|\tau_{2k+1}(p)|,$$

from which it follows from the classical Neumann theorem that T is invertible if

$$\sum_{k=1}^{\infty}|\tau_{2k+1}(p)| < |\tau_1(p)|. \tag{1.10}$$

It is in trying to satisfy this inequality that restrictions on p appear. In view of (1.8) we have

$$\sum_{k=1}^{\infty} |\tau_{2k+1}(p)| \leq \frac{4\pi_p}{\pi^2}\left(\frac{\pi^2}{8} - 1\right), \tag{1.11}$$

and it remains to estimate $|\tau_1(p)|$ from below. By the $p-$Jordan inequality (1.4),

$$\sin_p(\pi_p t) > 2t \text{ if } 0 < t < 1/2,$$

and hence

$$\tau_1(p) = 4\int_0^{1/2} f_{1,p}(t)\sin(\pi t)dt > 4\int_0^{1/2} 2t\sin(\pi t)dt = 8/\pi^2. \tag{1.12}$$

Together with (1.10), (1.11) and the fact that π_p decreases as p increases, this shows that if $2 \leq p < \infty$, the map $T : L_q(0,1) \to L_q(0,1)$ is a homeomorphism for every $q \in (1,\infty)$. When $p < 2$ the argument is more delicate, a key step being the next result.

PROPOSITION 1.1. *Suppose that* $1 < p < q < \infty$. *Then the function* f *defined by*

$$f(x) = \frac{\sin_q^{-1} x}{\sin_p^{-1} x}$$

is strictly decreasing on $(0,1)$.

PROOF. Let

$$g(t) = \frac{(1-t^q)^{1/q}}{(1-t^p)^{1/p}} \ (0 < t < 1)$$

and observe that for all $t \in (0,1)$, $g'(t) > 0$. Now put

$$G(t) = \sin_p^{-1}(t) - g(t)\sin_q^{-1}(t).$$

Since

$$G'(t) = -\left(\sin_q^{-1}(t)\right) g'(t) < 0 \text{ in } (0,1),$$

it follows that $G(t) < 0$ in $(0,1)$, so that

$$f'(t) = \frac{G(t)}{\left(\sin_q^{-1}(t)\right)^2 (1-t^q)^{1/q}} < 0 \text{ in } (0,1).$$

□

From this it is easy to see that if $1 < p \leq q < \infty$, then

$$\sin_p(\pi_p x) \geq \sin_q(\pi_q x) \text{ when } x \in [0, 1/2].$$

This implies that if $1 < p < 2$, then

$$\tau_1(p) = 4\int_0^{1/2} \sin_p(\pi_p t)\sin(\pi t)dt > 4\int_0^{1/2} \sin^2(\pi t)dt = 1.$$

With (1.10) and (1.11) this shows that if $p_0 < p < \infty$, where p_0 is defined by the equation

$$\pi_{p_0} = \frac{2\pi^2}{\pi^2 - 8}, \tag{1.13}$$

then T is a homeomorphism for every $q \in (1,\infty)$.

We now need the standard result concerning preservation of bases mentioned earlier.

LEMMA 1.2. *Let (x_n) be a basis of a Banach space X, let T be a linear homeomorphism of X onto itself and for each $n \in \mathbb{N}$ put $y_n = Tx_n$. Then (y_n) is a basis of X.*

PROOF. Let $x \in X$ and let the biorthogonal functionals associated to the given basis be denoted by x_k^*. Then $x = Ty$ for some unique $y \in X$. Hence

$$\left\| x - \sum_{k=1}^{n} \langle y, x_k^* \rangle \, y_k \right\| = \left\| T\left(y - \sum_{k=1}^{n} \langle y, x_k^* \rangle \, x_k \right) \right\| \leq \|T\| \left\| y - \sum_{k=1}^{n} \langle y, x_k^* \rangle \, x_k \right\| \to 0$$

as $n \to \infty$. Thus each $x \in X$ is representable in the form $x = \sum_{k=1}^{\infty} \langle y, x_k^* \rangle \, y_k$. To show that this representation is unique, suppose that $x = \sum_{k=1}^{\infty} a_k y_k$. Then

$$\begin{aligned} \left\| \sum_{k=1}^{n} (a_k - \langle y, x_k^* \rangle) \, x_k \right\| &= \left\| T^{-1} \left(\sum_{k=1}^{n} (a_k - \langle y, x_k^* \rangle) \, y_k \right) \right\| \\ &\leq \left\| T^{-1} \right\| \left\| \sum_{k=1}^{n} (a_k - \langle y, x_k^* \rangle) \, y_k \right\| \to 0 \end{aligned}$$

as $n \to \infty$. Since (x_n) is a basis, $a_k = \langle y, x_k^* \rangle$ for all k. □

The basis properties of the $\sin_p$ functions now follow immediately.

THEOREM 1.3. *The functions $\sin_p(n\pi_p x)$ form a basis in $L_q(0,1)$ for every $q \in (1, \infty)$ if $p_0 < p < \infty$, where p_0 is defined by equation (1.13),*

$$\pi_{p_0} = \frac{2\pi^2}{\pi^2 - 8}.$$

Numerical solution of (1.13) shows that p_0 is approximately equal to 1.05. Some improvement of this result can be obtained by use of the faster decay properties of the $\tau_k(p)$ given in (1.9), but the improvement is very modest. Indeed, it can be shown (see [3]) that Theorem 1.3 is close to the limit of what can be established by the method represented by satisfaction of (1.10): this proof technique cannot work for $p \leq 1.0439898$. Whether or not the basis property holds for all $p \in (1, \infty)$ does not appear to be known. For basis properties of functions closely related to $\sin_p$ see [7].

3. Representation of compact linear operators

3.1. Abstract theory. The spectral theory of compact linear operators acting between Hilbert spaces is a most attractive part of functional analysis. Thus if H is a Hilbert space and $A : H \to H$ is compact and self-adjoint, then A has the representation

$$Ax = \sum_n \lambda_n (x, \phi_n) \phi_n \ (x \in H),$$

where the λ_n are eigenvalues of A, each repeated according to multiplicity and ordered by decreasing modulus, while the ϕ_n are orthonormal eigenvectors of A corresponding to the eigenvalues λ_n; here the inner product in H is denoted by

$(\cdot,\cdot)$. The number of eigenvalues is finite if and only if the rank of A is finite; if there are infinitely many eigenvalues they can accumulate only at 0. If the kernel of A is trivial, then the ϕ_n form a complete orthonormal set in H. More generally, if H_1, H_2 are Hilbert spaces and $B : H_1 \to H_2$ is compact and linear, then B has the Schmidt representation

$$Bx = \sum_n \mu_n (x, \phi_n)_1 \psi_n \ (x \in H_1),$$

where $(\cdot,\cdot)_1$ denotes the inner product in H_1 and this time the ϕ_n are orthonormal eigenvectors of the positive square root $|B|$ of B^*B with corresponding eigenvalues μ_n (the μ_n are the singular values of B) and $\psi_n = \mu_n^{-1} B\phi_n$ ($\mu_n \neq 0$). Details of these assertions are given in [5], Chapter II, for example. Here we describe recent work aimed at the extension of these Hilbert space results to the case of a compact linear map acting between reflexive Banach spaces with strictly convex duals.

First we recall some concepts from the geometry of Banach spaces: to avoid complications we shall always assume that the spaces involved are infinite-dimensional, real and reflexive. The modulus of convexity of such a space X (with norm $\|\cdot\|_X$) is the map $\delta_X : [0,2] \to [0,1]$ defined by

$$\delta_X(\varepsilon) = \inf\left\{1 - \frac{\|x+y\|_X}{2} : x, y \in X; \|x\|_X = \|y\|_X = 1, \|x-y\|_X = \varepsilon\right\},$$

and X is said to be uniformly convex if for all $\varepsilon \in (0,2]$, $\delta_X(\varepsilon) > 0$. The modulus of smoothness of X is the function $\rho_X : (0,\infty) \to [0,\infty)$ given by

$$\rho_X(\tau) = \sup\left\{\frac{\|x+y\|_X + \|x-y\|_X}{2} - 1 : \|x\|_X = 1, \|y\|_X = \tau\right\},$$

and X is called uniformly smooth if $\lim_{\tau\to 0} \rho_X(\tau)/\tau = 0$. Note that in the definition of $\delta_X(\varepsilon)$ the same quantity results if the infimum is taken over all $x, y \in X$ with $\|x\|_X, \|y\|_X \leq 1$ and $\|x-y\|_X \geq \varepsilon$; moreover, in the definition of $\rho_X(\tau)$ the supremum may be taken over all $x, y \in X$ with $\|x\|_X \leq 1$ and $\|y\|_X \leq \tau$ without affecting the outcome. Every Hilbert space H is both uniformly convex and uniformly smooth, for from the parallelogram identity it follows that

$$\delta_H(\varepsilon) = 1 - \left(1 - \varepsilon^2/4\right)^{1/2} \text{ and } \rho_H(\tau) = (1+\tau^2)^{1/2} - 1.$$

Similarly, from the Clarkson inequalities (see [15], pp. 225-227) it can be shown that every L_p space with $1 < p < \infty$ is both uniformly convex and uniformly smooth. In fact, Hilbert spaces are the 'most' uniformly convex and the 'most' uniformly smooth spaces: for every X,

$$\delta_X(\varepsilon) \leq \delta_H(\varepsilon) \text{ and } \rho_X(\tau) \geq \rho_H(\tau).$$

These two notions are dual in the sense that X is uniformly convex if and only if X^* is uniformly smooth. We refer to [18], 1e for proofs of these claims and for further information.

Two useful properties of any uniformly convex space X are worth mentioning:

(i) If (x_k) is a sequence in X that converges weakly to $x \in X$, with $\|x_k\|_X \to \|x\|_X$, then $\|x_k - x\|_X \to 0$.

(ii) Let K be a closed, convex, non-empty subset of X. Then given any $x \in X$, there is a unique point $P_K x \in K$ at which the distance of x from K is attained.

The map $P_K : X \to K$ thus defined is called the projection of X onto K and is continuous.

Proofs of these assertions may be found in [8], for example.

A weaker condition than that of uniform convexity is strict convexity: X is strictly convex if the statement $\|x\|_X = \|y\|_X = \|(x+y)/2\|_X = 1$ implies that $x = y$. In other words, the unit sphere of X contains no line segments. This is linked to the concept of smoothness: X is called smooth if, for every x belonging to the unit sphere of X, there is a unique $x^* \in X^*$ such that $\|x^*\|_{X^*} = \langle x, x^* \rangle_X = 1$. Here $\langle \cdot, \cdot \rangle_X$ corresponds to the duality pairing between X and X^*. It turns out that X is smooth if and only if the norm $\|\cdot\|_X$ is Gâteaux differentiable on $X \backslash \{0\}$; given $x \in X \backslash \{0\}$, the unique $x^* \in X^*$ such that for all $y \in X$,

$$\langle y, x^* \rangle_X = \lim_{t \to 0} \left(\|x + ty\|_X - \|x\|_X \right) / t$$

is called the Gâteaux derivative of $\|\cdot\|_X$ at x and is denoted by $\widetilde{J}_X(x)$ or grad $\|x\|_X$. Note that X is uniformly smooth if and only if this limit is uniform on the set $\{(x, y) : \|x\|_X = \|y\|_X = 1\}$. The linkage just mentioned is that X is smooth if and only if X^* is strictly convex. When X is smooth, the Gâteaux derivative $\widetilde{J}_X$ of $\|\cdot\|_X$, has the property that for each $x \in X \backslash \{0\}$, $\widetilde{J}_X(x)$ is the unique element of X^* such that

$$\left\| \widetilde{J}_X(x) \right\|_{X^*} = 1 \text{ and } \left\langle x, \widetilde{J}_X(x) \right\rangle_X = \|x\|_X .$$

We define

$$(x, h)_X = \|x\|_X \langle h, \operatorname{grad} \|x\|_X \rangle_X \text{ for } x, h \in X, x \neq 0,$$

and set $(0, h)_X = 0$ for all $h \in X$: $(x, h)_X$ is called the semi-inner product of x and h. It depends linearly on h and $(x, x)_X = \|x\|^2$, while in general, $(x, h)_X \neq (h, x)_X$.

The property of being strictly convex can be transmitted. In fact, if M is a closed linear subspace of X, then if X is strictly convex so are M and X/M; while if X^* is strictly convex, so are $(X/M)^*$ and $M^0 := \{x^* \in X^* : \langle x, x^* \rangle_X = 0$ for all $x \in M\}$, the polar of M. For future use we remark here that if N is a linear subspace of X^*, then we shall write ${}^0N$ for $\{x \in X : \langle x, x^* \rangle_X = 0$ for all $x^* \in N\}$.

Related to $\widetilde{J}_X$ are the duality maps. To explain this, let μ be a gauge function; that is, a continuous, strictly increasing function $\mu : [0, \infty) \to [0, \infty)$ such that $\mu(0) = 0$, $\lim_{t \to \infty} \mu(t) = \infty$. Then the duality map with gauge function μ is the map $J_X : X \to X^*$ defined by

$$J_X(x) = \mu\left(\|x\|_X\right) \widetilde{J}_X(x) \text{ if } x \neq 0, \; J_X(0) = 0.$$

It has the properties that for all $x \in X$,

$$\langle x, J_X(x) \rangle_X = \|J_X(x)\|_{X^*} \|x\|_X , \;\; \|J_X(x)\|_{X^*} = \mu\left(\|x\|_X\right).$$

Again we refer to [8] for further details and proofs of the statements made above.

From now on we shall assume that X and Y are real, reflexive spaces with strictly convex duals X^* and Y^*. Under these conditions, given any gauge function μ, the corresponding duality map $J_X : X \to X^*$ is strictly monotone (that is, $\langle x - y, J_X x - J_X y \rangle_X > 0$ if $x, y \in X, x \neq y$) and weakly continuous (that is, if $x_k \to x$ in X, then $J_X x_k$ converges in the weak* sense to $J_X x$). If, in addition, X is strictly convex, then J_X is surjective and its inverse $(J_X)^{-1} : X^* \to X$ is a duality map of X^* onto X with gauge function μ (see [19]). Henceforth the only

duality maps J_X on X that we shall consider correspond to gauge functions μ_X that are normalised by $\mu_X(1) = 1$. Moreover, when X is a Hilbert space, $\widetilde{J}_X(x) = x/\|x\|_X$ $(x \neq 0)$ and we shall identify J_X (with gauge function $\mu(t) = t$) with the identity map of X to itself.

After these preliminaries we can begin the promised analysis of compact linear maps $T : X \to Y,\ T \neq 0$. The first step is the following familiar result.

PROPOSITION 1.4. *There exists $x_1 \in X$, with $\|x_1\|_X = 1$, such that $\|T\| = \|Tx_1\|_Y$. Also, $x = x_1$ satisfies*

$$T^* \widetilde{J}_Y T x = \nu \widetilde{J}_X x, \tag{1.14}$$

with $\nu = \|T\|$; in terms of duality maps this equation has the form

$$T^* J_Y T x = \nu_1 J_X x,\ \nu_1 = \|T\|\, \mu_Y(\|T\|).$$

If $x \in X\backslash\{0\}$ satisfies (1.14) for some ν, then $0 \leq \nu \leq \|T\|$ and $\|Tx\|_Y = \nu \|x\|_X$.

PROOF. First let (w_n) be a sequence in the unit sphere of X such that $\|Tw_n\|_Y \to \|T\|$. As X is reflexive we may suppose that (w_n) converges weakly, to w, say. By the compactness of T, $Tw_n \to Tw$. Thus $\|Tw\|_Y = \|T\|$, so that $\|w\|_X = 1$. The existence of x_1 follows. To see that it satisfies (1.14), note that

$$\|T\| = \|Tx_1\|_Y = \max_{x \in X\backslash\{0\}} \frac{\|Tx\|_Y}{\|x\|_X},$$

and that consequently, for all $x \in X$,

$$\frac{d}{dt}\left(\frac{\|Tx_1 + tTx\|_Y}{\|x_1 + tx\|_X}\right)\Big|_{t=0} = 0.$$

In terms of duality pairings this gives

$$\left\langle Tx, \widetilde{J}_Y T x_1 \right\rangle_Y = \|Tx_1\|_Y \left\langle x, \widetilde{J}_X x_1 \right\rangle_X,$$

which amounts to

$$T^* \widetilde{J}_Y T x_1 = \lambda \widetilde{J}_X x_1$$

with $\lambda = \|T\|$. For the remaining part, suppose that $x \in X\backslash\{0\}$ satisfies (1.14) for some ν. Then

$$\|Tx\|_Y = \left\langle Tx, \widetilde{J}_Y T x \right\rangle_Y = \left\langle x, T^* \widetilde{J}_Y T x \right\rangle_X = \nu \left\langle x, \widetilde{J}_X x \right\rangle_X = \nu \|x\|_X.$$

Hence $0 \leq \nu \leq \|T\|$. □

Equation (1.14) can be thought of as the Euler equation for maximising $\|Tx\|_Y$ subject to the condition $\|x\|_X = 1$.

To proceed further a usable adaptation of the procedure followed in the Hilbert space case is desirable. We recall how this goes for a compact self-adjoint map S of a Hilbert space H into itself. First it is shown that S has an eigenvalue λ_1 with corresponding eigenvector e_1; then, denoting the orthogonal complement of the span of e_1 by H_2, this argument is applied to the restriction S_2 of S to H_2, giving another pair λ_2, e_2; and so on. Such an adaptation was provided in [6] and is as follows. With sp W denoting the span of W we set $X_1 = X$, $M_1 = \text{sp}\,\{J_X x_1\}$, $X_2 = {}^0M_1$, $N_1 = \text{sp}\,\{J_Y T x_1\}$, $Y_2 = {}^0N_1$ and $\lambda_1 = \|T\|$. Since X_2 and Y_2 are closed subspaces of reflexive spaces they are reflexive. As $X_2^* = ({}^0M_1)^*$ is isometrically

isomorphic to X_1^*/M_1, it follows that X_2^* is strictly convex; the same argument applies to Y_2^*. By Proposition 1.4,

$$\langle Tx, J_Y Tx_1\rangle_Y = \nu_1 \langle x, J_X x_1\rangle_X \text{ for all } x \in X;$$

hence T maps X_2 to Y_2. The restriction T_2 of T to X_2 is compact, and if it is not the zero map, we may apply Proposition 1.4 again and conclude that there exists $x_2 \in X_2\backslash\{0\}$ such that, with obvious notation,

$$\langle T_2 x, J_{Y_2} Tx_2\rangle_{Y_2} = \nu_2 \langle x, J_{X_2} x_2\rangle_{X_2} \text{ for all } x \in X_2,$$

where $\nu_2 = \lambda_2 \mu_Y(\lambda_2)$, $\lambda_2 = \|Tx_2\|_Y = \|T_2\|$. Plainly $\lambda_2 \le \lambda_1$ and $\nu_2 \le \nu_1$. In this way we obtain elements $x_1, x_2, ..., x_n$ of X, each with unit norm, subspaces $M_1, ..., M_n$ of X^* and $N_1, ..., N_n$ of Y^*, where

$$M_k = \text{ sp } \{J_X x_1, ..., J_X x_k\} \text{ and } N_k = \text{ sp } \{J_Y Tx_1, ..., J_Y Tx_k\},$$

and decreasing families $X_1, ..., X_n$ and $Y_1, ..., Y_n$ of subspaces of X and Y respectively given by

$$X_k = {}^0M_{k-1},\ Y_k = {}^0N_{k-1}\ (k = 2, ..., n).$$

For each $k \in \{1, ..., n\}$, T maps X_k into Y_k, $x_k \in X_k$ and, with T_k standing for the restriction of T to X_k, $\lambda_k = \|T_k\| = \|Tx_k\|_Y$, $\nu_k = \lambda_k \mu_Y(\lambda_k)$, we have

$$\langle T_k x, J_{Y_k} T_k x_k\rangle_{Y_k} = \nu_k \langle x, J_{X_k} x_k\rangle_{X_k} \text{ for all } x \in X_k, \tag{1.15}$$

and so

$$T_k^* J_{Y_k} T_k x_k = \nu_k J_{X_k} x_k.$$

It turns out that (1.15) is equivalent to

$$\langle Tx, J_Y Tx_k\rangle_Y = \nu_k \langle x, J_X x_k\rangle_X \text{ for all } x \in X_k.$$

Since $Tx_k \in Y_k = {}^0N_{k-1}$, we see that

$$\langle Tx_k, J_Y Tx_l\rangle_Y = 0 \text{ if } l < k. \tag{1.16}$$

The process stops with λ_n, x_n and X_{n+1} if and only if the restriction of T to X_{n+1} is the zero operator. In that case, the range of T is the linear space spanned by $Tx_1, ..., Tx_n$.

Some fundamental properties of the quantities arising in the above scheme are given next.

PROPOSITION 1.5. *(i) The sequence (x_k) is semi-orthogonal in the sense that*

$$(x_l, x_k)_X = 0 \text{ if } l < k.$$

(ii) If rank $T = \infty$, the sequence (λ_k) is infinite and converges to zero. Moreover,

$$\ker T \supset \cap_{k=1}^{\infty} X_k := X_\infty.$$

PROOF. (i) This is an immediate consequence of (1.16).

(ii) From (1.16),

$$\left\langle Tx_n, \widetilde{J}_Y Tx_m\right\rangle_Y = 0 \text{ if } m < n.$$

Thus if $m < n$,

$$\begin{aligned}\limsup_{k\to\infty}\lambda_k &\leq \|Tx_m\|_Y = \left\langle Tx_m, \widetilde{J}_Y Tx_m\right\rangle_Y = \left\langle Tx_m - Tx_n, \widetilde{J}_Y Tx_m\right\rangle_Y \\ &\leq \|Tx_m - Tx_n\|_Y \left\|\widetilde{J}_Y Tx_m\right\|_{Y^*} = \|Tx_m - Tx_n\|_Y .\end{aligned}$$

As $\{x_n\}$ is bounded and T is compact, some subsequence of $\{Tx_n\}$ must converge, The proof is complete. □

Now we introduce the family of maps

$$S_k : X \to \mathcal{M}'_{k-1} := \text{ sp } \{x_1, ..., x_{k-1}\},\ k \geq 2,$$

determined by the condition that $x - S_k x \in X_k$ for all $x \in X$. By induction it follows that S_k is uniquely given by

$$S_k x = \sum_{j=1}^{k-1} \xi_j(x) x_j,$$

where

$$\begin{aligned}\xi_j(x) &= \left\langle x - \sum_{i=1}^{j-1} \xi_i(x) x_i, J_X x_j \right\rangle_X \quad \text{if } j > 1, \\ \xi_1(x) &= \langle x, J_X x_1\rangle_X .\end{aligned}$$

Thus each S_k is linear; moreover, in view of the uniqueness, it follows that $S_k^2 = S_k$ and S_k is a linear projection of X onto $\mathcal{M}'_{k-1}$. It can be shown that for each $k \geq 2$, the spaces X and X^* have the direct sum decompositions

$$X = X_k \oplus \mathcal{M}'_{k-1},\ X^* = M_{k-1} \oplus \left(\mathcal{M}'_{k-1}\right)^0 .$$

The map S_k^* is a linear projection of X^* onto M_{k-1}.

Now let P_k, P_∞ be the projections of X onto the subspaces X_k, X_∞ respectively.

PROPOSITION 1.6. *Let X be uniformly convex. Then for all $x \in X$, $P_k x \to P_\infty x$ as $k \to \infty$.*

PROOF. Since $\|x - P_k x\|_X = \|[x]\|_{X/X_k} \leq \|x\|_X$, we see that $\|P_k x\|_X \leq 2\|x\|_X$. Thus $\{P_k x\}$ has a subsequence that converges weakly, to $y \in X_\infty$, say. We claim that $y = P_\infty x$. If this were not so, then

$$\|x - y\|_X > \|x - P_\infty x\|_X = \|[x]\|_{X/X_\infty} .$$

Hence

$$\begin{aligned}\|x - P_k x\|_X &\geq \left\langle x - P_k x, \widetilde{J_X}(x - y)\right\rangle_X \to \left\langle x - y, \widetilde{J_X}(x - y)\right\rangle_X \\ &= \|x - y\|_X > \|[x]\|_{X/X_\infty} .\end{aligned}$$

It follows that for some $k \in \mathbb{N}$, $\|x - P_k x\|_X > \|[x]\|_{X/X_\infty}$, which means that $\|[x]\|_{X/X_k} > \|[x]\|_{X/X_\infty}$ and contradicts the fact that $X_\infty \subset X_k$. Thus every weakly convergent subsequence of $\{P_k x\}$ has weak limit $P_\infty x$, from which it follows from a standard contradiction argument that the whole sequence $\{P_k x\}$ converges weakly to $P_\infty x$. However, it is not difficult to see that $\|x - P_k x\|_X \to \|x - P_\infty x\|_X$ .The result now follows from the uniform convexity of X. □

THEOREM 1.7. *Let X be uniformly convex. Then for all $x \in X$,*

$$x = \lim_{k\to\infty} (id - P_k)\, S_k x + P_\infty x, \text{ where } id : X \to X \text{ is the identity.} \tag{1.17}$$

If ker $(T) = \{0\}$, then

$$x = \lim_{k\to\infty} (id - P_k)\, S_k x,\ Tx = \lim_{k\to\infty} (T - TP_k)\, S_k x.$$

If, in addition, $\lim_{k\to\infty} S_k x$ exists, then

$$x = \textstyle\sum_{j=1}^{\infty} \xi_j(x) x_j \text{ and } Tx = \sum_{j=1}^{\infty} \lambda_j \xi_j(x) y_j,\ y_j = Tx_j / \left\|Tx_j\right\|_Y .$$

PROOF. For any closed linear subspace L of X and any $u \in X$, $P_L u$ is the unique element $w \in L$ for which $\left\|u - w\right\|_X$ is minimal. Hence if $u - v \in L$ we have $P_L(u - v) = u - v$, and as

$$\left\|u - (u - v + P_L v)\right\|_X = \left\|v - P_L v\right\|_X ,$$

we also have $P_L u = u - v + P_L v$, so that

$$P_L u - P_L v = u - v = P_L(u - v).$$

Put $s_k = s_k(x) := S_k x$. Then

$$(s_k - P_k s_k) - (x - P_k x) = s_k - x - P_k(s_k - x) = 0.$$

Now (1.17) follows with the aid of Proposition 1.6. The rest is clear; in particular, if ker $(T) = \{0\}$ and $\lim_{k\to\infty} S_k x$ exists, then $\lim_{k\to\infty} P_k s_k$ exists and is 0, since $X_\infty = \{0\}$. □

COROLLARY 1.8. *If X is a Hilbert space, then for all $x \in X$,*

$$x = \textstyle\sum_{j=1}^{\infty} (x, x_j)_X x_j + P_\infty x,$$

so that if $\ker T = \{0\}$,

$$x = \textstyle\sum_{j=1}^{\infty} (x, x_j)_X x_j \text{ and } Tx = \sum_{j=1}^{\infty} \lambda_j (x, x_j)_X y_j.$$

PROOF. In this case, $S_k x \in \mathcal{M}'_{k-1} = X_k^{\perp}$, and so $P_k S_k x = 0$. □

Results corresponding to those in the theorem can be obtained if the hypothesis of uniform convexity is made about Y rather than X. Under this assumption, put $Y_\infty = \cap_{k=1}^{\infty} Y_k$ and let Q_k, Q_∞ be the projections of Y onto the subspaces Y_k, Y_∞ respectively. Then it turns out that for all $x \in X$,

$$Tx = \lim_{k\to\infty} (I - Q_k)\, TS_k x + Q_\infty Tx,$$

where

$$TS_k x = \textstyle\sum_{j=1}^{k-1} \xi_j(x) Tx_j = \sum_{j=1}^{k-1} \lambda_j \xi_j(x) y_j.$$

If Y is a Hilbert space, then for all $x \in X$,

$$Tx = \textstyle\sum_{j=1}^{\infty} \lambda_j \xi_j(x) y_j + Q_\infty Tx.$$

Finally, we observe that when both X and Y are Hilbert spaces, the λ_n are eigenvalues of the positive square root of T^*T. The privileged rôle of eigenvalues is absent in the general situations we have been discussing, in which each λ_n is simply the norm of the restriction of T to the subspace X_n. Note also that there is a relation between the λ_n and the Gelfand numbers of T. We recall that for

each $n \in \mathbb{N}$, the n^{th} Gelfand number of any bounded linear map $S : X \to Y$ is the quantity

$$c_n(S) = \inf \|S \restriction_V\|,$$

where the infimum is taken over all those linear subspaces V of X with codimension less than n. These numbers form a sequence that converges to zero if and only if S is compact; they form one of several sequences (the so-called $s-$numbers-see [20]) that are used to give an idea of 'how compact' a map is. Since

$$\text{codim } X_k = \dim \text{ sp } \{J_X x_1, ..., J_X x_{k-1}\},$$

it follows immediately that

$$c_n(T) \leq \lambda_n \ (n \in \mathbb{N}).$$

3.2. Application: the $p-$Laplacian. Let Ω be a bounded open subset of $\mathbb{R}^n$, let $p \in (1, \infty)$, $D_j = \partial/\partial x_j$ and write

$$W_p^1(\Omega) = \{u \in L_p(\Omega) : D_j u \in L_p(\Omega) \text{ for } j = 1, ..., n\};$$

this is the familiar first-order Sobolev space and is endowed with the norm

$$\left(\int_\Omega \left\{ |u|^p + \sum_{j=1}^n |D_j u|^p \right\} dx \right)^{1/p}.$$

Put $X = \overset{0}{W}{}_p^1(\Omega)$, the closure in $W_p^1(\Omega)$ of the C^∞ functions with compact support in Ω. The norm of $u \in X$ is defined to be

$$\|u\|_X = \left(\int_\Omega \sum_{j=1}^n |D_j u|^p \, dx \right)^{1/p}.$$

The Friedrichs inequality shows that this norm is equivalent to that inherited by X from $W_p^1(\Omega)$. Let $Y = L_p(\Omega)$ and $T = \text{id}: X \to Y$. Both X and Y are reflexive and strictly convex, with strictly convex duals. Direct verification shows that

$$\widetilde{J}_Y u = \|u\|_p^{-(p-1)} |u|^{p-2} u,$$

while

$$\widetilde{J}_X u = - \|u\|_X^{-(p-1)} \Delta_p u \text{ in the sense of distributions,}$$

where

$$\Delta_p u = \sum_{j=1}^n D_j \left(|D_j u|^{p-2} D_j u \right),$$

corresponding to a version of the $p-$Laplacian. This follows since

$$\begin{aligned} \left\langle u, - \|u\|_X^{-(p-1)} \Delta_p u \right\rangle_X &= - \|u\|_X^{-(p-1)} \langle u, \Delta_p u \rangle_X \\ &= \|u\|_X^{-(p-1)} \int_\Omega \sum_{j=1}^n (D_j u) |D_j u|^{p-2} D_j u dx \\ &= \|u\|_X . \end{aligned}$$

With $\mu_X(t) = \mu_Y(t) = t^{p-1}$, the associated duality maps J_X, J_Y are given by

$$J_X(u) = -\Delta_p u, \ J_Y(u) = |u|^{p-2} u.$$

The basic eigenvalue equation (1.14) then gives the existence of $u_1 \in X_1 = X$ such that

$$\left\langle v, \widetilde{J}_Y u_1 \right\rangle_Y = \lambda_1 \left\langle v, \widetilde{J}_X u_1 \right\rangle_X \text{ for all } v \in X.$$

Since $\|u_1\|_Y = \|\mathrm{id}\| = \lambda_1$, this implies that

$$\int_\Omega v \left|u_1\right|^{p-2} u_1 dx = \lambda_1^p \int_\Omega \sum_{j=1}^n (D_j v) \left|D_j u_1\right|^{p-2} D_j u_1 dx,$$

so that u_1 is a weak solution of the Dirichlet eigenvalue problem

$$-\Delta_p u_1 = \lambda_1^{-p} \left|u_1\right|^{p-2} u_1 \text{ in } \Omega,\ u_1 = 0 \text{ on } \partial\Omega.$$

As in the general theory of 3.1, denote the restriction of id to X_k by id_k. Since rank $id_k = \infty$ for each k, our general procedure ensures that for each $k \in \mathbb{N}$, there are an "eigenvector" u_k and a corresponding "eigenvalue" λ_k^{-p} that satisfy

$$-\Delta_p u_k = \lambda_k^{-p} \left|u_k\right|^{p-2} u_k \text{ in } \Omega,\ u_k = 0 \text{ on } \partial\Omega, \tag{1.18}$$

in the sense that for all $v \in X_k$,

$$\int_\Omega v \left|u_k\right|^{p-2} u_k dx = \lambda_k^p \int_\Omega \sum_{j=1}^n (D_j v) \left|D_j u_k\right|^{p-2} D_j u_k dx.$$

The function u_k is called a $k-$weak solution of (1.18): note that when $k = 1$ all functions in $X_1 = X$ are allowed as test functions, so that u_1 is a weak solution of the Dirichlet problem in the conventional sense. However, when $k > 1$ the only test functions allowed are the elements of X_k, which is a proper subset of X, and so the u_k need not be weak solutions in the classical sense. Information about the growth of the λ_k^{-p} as $k \to \infty$ can be obtained without difficulty from $s-$number estimates. For from the definition of Gelfand numbers together with [17], Theorem 3.c.5 and Remark 3.c.7 (1), we see that

$$\lambda_k \geq c_k(\mathrm{id}_k) \geq c_k(\mathrm{id}) \geq c k^{-1/n},$$

where c is a positive constant independent of k. Thus the eigenvalues λ_k^{-p} of (1.18) are $O(k^{p/n})$. This upper estimate of the growth of the eigenvalues is exactly that obtained in [11], [12], [13] (together with lower bounds of the same order) for the Lyusternik-Schnirel'mann eigenvalues, which correspond to classical weak solutions. We recall that the m^{th} such eigenvalue, denoted here by $\hat{\lambda}_m$, is given by

$$\hat{\lambda}_m = \inf_{K \in \mathcal{A}_m} \sup_{u \in K} \|u\|_p^{-p},$$

where $\mathcal{A}_m$ is the family of all compact, symmetric subsets K of $\{u \in \overset{0}{W}{}_p^1(\Omega) : \|\nabla u\|_p = 1\}$ with genus $\gamma(K) \geq m$, the genus being defined as

$$\gamma(K) = \inf\{k \in \mathbb{N} : \text{there is a continuous odd map } h : K \to \mathbb{R}^k \backslash \{0\}\}.$$

The corresponding quantities λ_m^{-p} obtained by the method presented here are expressible as

$$\lambda_m^{-p} = \inf_{u \in X_m \backslash \{0\}} \frac{\|\nabla u\|_p^p}{\|u\|_p^p}.$$

The relationship between these two sets of eigenvalues remains unclear, and the question as to whether there are eigenvalues not found by either method remains unanswered.

As remarked in Section 2, the classical solutions of the Dirichlet problem (1.1) for the $p-$Laplacian on the unit interval $(0,1)$ are exactly the functions $\sin_p(n\pi_p t)$ $(n \in \mathbb{N})$. However, the '$k-$weak' solutions of this problem that are generated by our procedure appear to be different from these $p-$trigonometric functions, for numerical computations indicate that the $\sin_p(n\pi_p t)$ functions do not have the semi-orthogonality properties possessed by these weak solutions.

Higher-order problems may be tackled in the same way. Thus let Ω and p be as above, but now take $Y = L_p(\Omega)$ and let $X = \overset{0}{W}{}^2_p(\Omega)$, the completion of $C_0^\infty(\Omega)$ with respect to the norm

$$\|u\|_X = \|\Delta u \mid L_p(\Omega)\|.$$

This norm is equivalent to the more usual norm

$$\left(\sum_{|\alpha|\leq 2} \|D^\alpha u \mid L_p(\Omega)\|^p\right)^{1/p}$$

(standard notation being employed) in view of [14], remarks following Corollary 7.11, together with Corollary 9.10. As before it can be checked that X and Y have the properties required for application of the abstract theory, and that

$$\widetilde{J}_X u = \|u\|_X^{-(p-1)} \, \Delta\left(|\Delta u|^{p-2}\,\Delta u\right).$$

The operator $\Delta\left(|\Delta u|^{p-2}\,\Delta u\right)$ is often referred to as the $p-$biharmonic operator. Thus $T := \mathrm{id} : X \to Y$ is compact and we obtain the existence of a countable family of 'weak' eigenvectors and eigenvalues of the Dirichlet problem

$$\Delta\left(|\Delta u|^{p-2}\,\Delta u\right) = \lambda\,|u|^{p-2}\,u \text{ in } \Omega,\ u = |\nabla u| = 0 \text{ on } \partial\Omega,$$

with the k^{th} eigenvalue being $O\left(k^{2p/n}\right)$.

Bibliography

[1] G. E. Andrews, R. Askey and R. Roy, *Special functions,* Cambridge Univ. Press, Cambridge, 1999.

[2] P. Binding, L. Boulton, J. Čepička, P. Drábek and P. Girg, *Basis properties of the p−Laplacian*, Proc. American Math. Soc. **134** (2006), 3487-3494.

[3] P. J. Bushell and D. E. Edmunds, *Remarks on generalised trigonometric functions,* Rocky Mountain J. Math., to appear.

[4] P. Drábek and R. Manásevich, *On the closed solution to some nonhomogeneous eigenvalue problems with p−Laplacian,* J. Diff. Integral Equations **12** (1999), 773-788.

[5] D. E. Edmunds and W. D. Evans, *Spectral theory and differential operators,* Oxford University Press, Oxford, 1987.

[6] D. E. Edmunds, W. D. Evans and D. J. Harris, *Representations of compact linear operators in Banach spaces and nonlinear eigenvalue problems,* J. London Math. Soc. **78** (2008), 65-84.

[7] D. E. Edmunds, P. Gurka and J. Lang, *Properties of generalized trigonometric functions*, J. Approx. Theory **164** (2012), 47-56.

[8] D. E. Edmunds and J. Lang, *Eigenvalues, embeddings and generalised trigonometric functions*, Lecture Notes in Mathematics 2016, Springer, Heidelberg-Dordrecht-London-New York, 2011.

[9] R. E. Edwards, *Fourier series, a modern introduction,* Vol. 2 (2nd ed.), Springer-Verlag, 1982.

[10] S. R. Finch, *Mathematical constants,* Cambridge University Press, Cambridge, 2003.

[11] L. Friedlander, *Asymptotic behavior of eigenvalues of the p−Laplacian,* Comm. Partial Diff. Equations **14** (1989), 1059-1069.

[12] J. P. García-Azorero and I. Peral Alonso, *Existence and nonuniqueness for the p−Laplacian: nonlinear eigenvalues*, Comm. Partial Diff. Equations **12** (1987), no. 12, 1389-1430.

[13] J. P. García-Azorero and I. Peral Alonso, *Comportement asymptotique des valeurs propres du p−laplacien*, C. R. Acad. Sci. Paris Sér. I Math. **307** (1988), no. 2, 75-78.

[14] D. Gilbarg and N. S. Trudinger, *Elliptic partial differential equations of second order*, 2nd ed., Revised 3rd printing, Springer-Verlag, Berlin-Heidelberg-New York, 1998.

[15] E. Hewitt and K. Stromberg, *Real and abstract analysis,* Springer-Verlag, New York, 1965.

[16] J. R. Higgins, *Completeness and the basis property of sets of special functions,* Cambridge Tracts in Mathematics, vol. 72, Cambridge Univ. Press, Cambridge, 1977.

[17] H. König, *Eigenvalue distribution of compact operators,* Birkhäuser, Basel, 1986.

[18] J. Lindenstrauss and L. Tzafriri, *Classical Banach spaces II,* Springer, Berlin, 1979.

[19] J.-L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Paris, 1969.

[20] A. Pietsch, *History of Banach spaces and linear operators,* Birkhäuser, Boston-Basel-Berlin, 2007.

Part 3

The control of PDEs: some basic concepts, recent results and open problems

Enrique Fernández-Cara

2000 *Mathematics Subject Classification.* Primary: 49J20, 93B05, 93C20, 49N90, Secondary: 70Q05, 76B75.

Key words and phrases. control theory, partial differential equations, optimal control, controllability, observability, equations from fluid mechanics

Abstract. These Notes deal with the control of systems governed by some PDEs. I will mainly consider time-dependent problems. The aim is to present some fundamental results, some applications and some open problems related to the optimal control and the controllability properties of these systems.

In Chapter 1, I will review part of the existing theory for the optimal control of partial differential systems. This is a very broad subject and there have been so many contributions in this field over the last years that we will have to limit considerably the scope. In fact, I will only analyze a few questions concerning some very particular PDEs. We shall focus on the Laplace, the stationary Navier-Stokes and the heat equations. Of course, the existing theory allows to handle much more complex situations.

Chapter 2 is devoted to the controllability of some systems governed by linear time-dependent PDEs. I will consider the heat and the wave equations. I will try to explain which is the meaning of controllability and which kind of controllability properties can be expected to be satisfied by each of these PDEs. The main related results, together with the main ideas in their proofs, will be recalled.

Finally, Chapter 3 is devoted to present some controllability results for other time-dependent, mainly nonlinear, parabolic systems of PDEs. First, we will revisit the heat equation and some extensions. Then, some controllability results will be presented for systems governed by stochastic PDEs. Finally, I will consider several nonlinear systems from fluid mechanics: Burgers, Navier–Stokes, Boussinesq, micropolar, etc.

Along these Notes, a set of questions (some of them easy, some of them more intrincate or even difficult) will be stated. Also, several open problems will be mentioned. I hope that all this will help to understand the underlying basic concepts and results and to motivate research on the subject.

Acknowledgement. I want to thank very deeply the Nečas Center and very particularly Prof. Josef Málek for their kind invitation to come to Prague and participate and collaborate in the programmed activities. I also want to thank Prof. Petr Kaplický for his invitation to write these Notes. This work has been partially supported by Grant MTM2010-15992, DGES-MICINN (Spain).

Contents

CHAPTER 1

Optimal control of systems governed by PDEs

In this Lecture, I will review part of the existing theory for the optimal control of partial differential systems. This is a very broad subject and there have been so many contributions in this field over the last years that we will have to limit considerably the scope. In fact, I will only analyze a few questions concerning some very particular PDEs. We shall focus on the Laplace, the stationary Navier-Stokes and the heat equations. Of course, the existing theory allows to handle much more complex situations. The optimal control of (elliptic, parabolic and hyperbolic) partial differential systems was addressed in [76]. Many other details can be found for instance in [32, 59, 72, 75] and the references therein. Along the text, several questions have been stated. They are of different nature and level of difficulty and it is highly recommended to the interested reader to try to answer them.

1. Some examples

It will be assumed that $\Omega \subset \mathbb{R}^N$ is a bounded, regular and connected open set, with boundary $\Gamma = \partial\Omega$.

The first example concerns the optimal control of a *capacitor*.

Let $\omega \subset\subset \Omega$ be a non-empty open set. For each $u \in L^2(\omega)$, we consider the state system

$$\begin{cases} -\Delta y = 1_\omega u & \text{in} \quad \Omega, \\ y = 0 & \text{on} \quad \Gamma, \end{cases} \tag{1.1}$$

where 1_ω is the characteristic function of ω.

The solution $y = y(x)$ to (1.1) can be interpreted as the *electric potential* of a capacitor to which a *density of charge* $1_\omega v$ is applied; $E = -\nabla y$ is the associated electric field.

In practice, it may be important to know how to choose v in a subset $\mathcal{U}_{\rm ad} \subset L^2(\omega)$ in order to obtain a potential y as close as possible to a prescribed function y_d without too much effort. For instance, $\mathcal{U}_{\rm ad}$ can be a ball in $L^2(\omega)$. It can also be a set of the form

$$\mathcal{U}_{\rm ad} = \{\, u \in L^2(\omega) : \underline{u} \leq u(x) \leq \overline{u} \ \text{ a.e.}\,\}, \tag{1.2}$$

where $\underline{u}, \overline{u} \in \mathbb{R}$.

Thus, let us fix $y_d \in L^2(\Omega)$ and let us introduce the *cost functional* J, with

$$J(u) = \frac{a}{2}\int_\Omega |y - y_d|^2\, dx + \frac{b}{2}\int_\omega |u|^2\, dx \tag{1.3}$$

where $a, b > 0$. The optimal control problem we want to solve is then:

PROBLEM P1: *To find* $\hat{u} \in \mathcal{U}_{\rm ad}$ *such that* $J(\hat{u}) \leq J(u)$ *for all* $u \in \mathcal{U}_{\rm ad}$ *, where* J *is given by* (1.3).

We will see below that this problem can be solved. We will also see the way the solution (the optimal control) can be characterized by an appropriate *optimality system*. Additionally, we will present some generalizations and variants.

In our second problem, the control is performed through the coefficients of the system.

Assume that Ω is composed of two *dielectric materials* whose properties and prices are different. We want to build a *nonhomogeneous plate* with these two materials in such an optimal way. Here, the word optimal means that, under an applied density of charge (fixed and known), the associated potential is as close as possible to a prescribed state y_d .

Let α and β be the *permeability coefficients* of the first and the second material, respectively. We assume that $0 < \alpha < \beta$. Let $\{G_1, G_2\}$ be a *partition* of Ω (G_1 and G_2 are measurable sets) and set

$$a(x) = \begin{cases} \alpha & \text{if } x \in G_1\,, \\ \beta & \text{if } x \in G_2\,. \end{cases} \tag{1.4}$$

Then the *electrostatic potential* $y = y(x)$ corresponding to this distribution of the materials is the solution of the system

$$\begin{cases} -\nabla \cdot (a(x)\nabla y) = f(x) & \text{in} \quad \Omega, \\ y = 0 & \text{on} \quad \Gamma, \end{cases} \tag{1.5}$$

where $f \in H^{-1}(\Omega)$ (for instance) is given. In this example, the coefficient $a = a(x)$ is the control and y is the state.

Let us put

$$j(a) = \frac{1}{2}\int_\Omega |y - y_d|^2\, dx \quad \forall a \in \mathcal{A}_{\rm ad}\,, \tag{1.6}$$

where $y_d \in L^2(\Omega)$ and, by definition, we have

$$\mathcal{A}_{\rm ad} = \{\, a \in L^\infty(\Omega) :\ a(x) = \alpha \text{ or } a(x) = \beta \text{ a.e.}\,\} \tag{1.7}$$

The second problem we want to consider in this Section is then:

PROBLEM P2: *To find* $\hat{a} \in \mathcal{A}_{\rm ad}$ *such that* $j(\hat{a}) \leq j(a)$ *for all* $a \in \mathcal{A}_{\rm ad}$ *, where* j *is given by* (1.6).

It is well known that, in general, this problem has no solution and a "generalized" or "relaxed" version has to be introduced in order to describe the limitting behavior of the minimizing sequences. This is in fact typical in control problems where the control enters in the system through its coefficients and, specially, in the principal part of the operator. Phenomena of this kind have led to a very rich development of the theory. We will see later what can be done and which is the physical interpretation of the "generalized" or "relaxed solution".

The third example is an *optimal design* problem.

We will assume that Ω is filled with a viscous incompressible fluid and we will try to find the optimal shape of a body travelling at constant velocity in Ω. Thus, assume that $B \subset \Omega$ is a non-empty closed subset whose shape is in principle unknown. We will assume that B is the closure of a connected open set and ∂B is

piecewise Lipschitz-continuous. Let us choose a reference system fixed with respect to B. We will consider the following Navier-Stokes system in $\Omega \setminus B$:

$$\begin{cases} -\nu\Delta y + (y\cdot\nabla)y + \nabla\pi = 0, \quad \nabla\cdot y = 0 & \text{in} \quad \Omega\setminus B, \\ y = y_\infty & \text{on} \quad \Gamma, \\ y = 0 & \text{on} \quad \partial B. \end{cases} \tag{1.8}$$

Here, (y,π) is the state (the velocity field and the pressure of the fluid). The positive coefficient ν is the viscosity of the fluid. We have assumed that the velocity of the fluid particles on the *exterior* boundary Γ, that is, far from the body, is y_∞ (a constant vector). We have also imposed the usual *no-slip condition* on ∂B. These boundary conditions in mean that the body travels with velocity $-y_\infty$ and the fluid particles on ∂B adhere to the body.

For each B in a family $\mathcal{B}_{\rm ad}$ of *admissible bodies*, the state system (1.8) possesses at least one *weak solution* (y,π), with $y\in H^1(\Omega;\mathbb{R}^2)$ and $\pi\in L^2(\Omega)$. Now, we can associate to each solution the quantity

$$T(B,y) = 2\nu\int_\Omega |Dy|^2\,dx, \tag{1.9}$$

where

$$Dy = \frac{1}{2}(\nabla y + \nabla y^t)$$

is the symmetric part of the gradient ∇y. It can be seen that $T(B,y)$ is in fact the *hydrodynamical drag* of the fluid, that is

$$T(B,y) = -y_\infty\cdot\int_{\partial B}\big(-\pi\, I + \nu D(y)\big)\cdot n\, ds$$

(the projection in the direction of the velocity of the body of the force exerted by the fluid particles).

Our third problem is the following:

> PROBLEM P3: *To find* $\hat B\in\mathcal{B}_{\rm ad}$ *such that the corresponding system* (1.8) *possesses a solution* $(\hat y,\hat\pi)$ *satisfying* $T(\hat B,\hat y)\le T(B,y)$ *whenever* (y,π) *is a solution to* (1.8) *and* $B\in\mathcal{B}_{\rm ad}$.

We will see below that, unless the family $\mathcal{B}_{\rm ad}$ satisfies particular and in some sense artificial conditions, it is not possible to prove an existence result for Problem P3.

Besides existence, another interesting question is to analyze the way $T(B,y)$ depends on B. In fact, we will show that, at least when y_∞ is small, the mapping $B\mapsto T(B,y)$ is well-defined and in some sense of class C^∞. We will also indicate how to compute its "derivative".

We will now consider an optimal control problem for a parabolic system with origin in biomedical science. As shown below, the control is oriented to the determination of cancer therapy strategies.

The state system is nonlinear and reads:

$$\begin{cases} c_t - \nabla\cdot(D(x)\nabla c) = f(c) - F(c,\beta) & \text{in} \quad Q = \Omega\times(0,T), \\ \beta_t - \mu\Delta\beta = -h(\beta) - H(c,\beta) + v1_\omega & \text{in} \quad Q = \Omega\times(0,T), \\ c = 0 & \text{on} \quad \Sigma = \partial\Omega\times(0,T), \\ \beta = 0 & \text{on} \quad \Sigma = \partial\Omega\times(0,T), \\ c(x,0) = c_0(x) & \text{in} \quad \Omega, \\ \beta(x,0) = \beta_0(x) & \text{in} \quad \Omega. \end{cases} \tag{1.10}$$

We assume that Ω is an organ, where we find a population of cancer cells with density $c = c(x,t)$ and a distribution of inhibitors (or antibodies), of density $\beta = \beta(x,t)$. The antibodies are generated through a therapy process, determined by the control v and localized in a small open set $\omega \subset \Omega$. This can be used to model the evolution of a glioblastoma, i.e. a brain tumor, under radiotherapy, see [103, 104].

The functions f and h define the proliferation and death rates of c and β, respectively. On the other hand, F and H determine the way c and β interact. In the simplest cases we just take

$$f(c) = \rho c, \quad h(\beta) = -m\beta, \quad F(c,\beta) = Rc\beta, \quad H(c,\beta) = Mc\beta, \tag{1.11}$$

for some positive constants ρ, m, R and M.

For a large family of functions f, h, F and H, for any $v \in L^2(\omega\times(0,T))$ there exists at least one solution (c,β) to (1.10).

Obviously, in order to make the problem realistic, we have to impose constraints on v. Thus, we will assume that $v \in \mathcal{V}_{\rm ad}$, where $\mathcal{V}_{\rm ad}$ is a bounded, closed and convex set of $L^2(\omega\times(0,T))$. A natural choice is the following:

$$\mathcal{V}_{\rm ad} = \{\, v \in L^2(\omega\times(0,T)) : 0 \le v \le A, \ \int_0^T v\,dt \le B, \ \ v = 0 \ \text{ for } \ t \notin \mathcal{I}\,\},$$

where $\mathcal{I}$ is a (small) closed set of times where the therapy is applied.

There are different possible choices for the cost function. A reasonable (but maybe nor the best) choice is the following:

$$K(c,\beta,v) = \frac{a}{2}\int_\Omega |c(x,T)|^2\,dx + \frac{b}{2}\int_{\omega\times(0,T)} |v|^2\,dx\,dt. \tag{1.12}$$

The fourth considered problem is then:

> PROBLEM P4: *To find $\hat v \in \mathcal{V}_{\rm ad}$ such that the corresponding system* (1.10) *possesses a solution $(\hat c,\hat\beta)$ satisfying $K(\hat c,\hat\beta,\hat v) \le K(c,\beta,v)$ whenever (c,β) is a solution to* (1.10) *and $v \in \mathcal{V}_{\rm ad}$.*

Under very general conditions, we will give below an existence result for Problem P4. We will also find the *optimality system* for this problem.

2. Existence, uniqueness and optimality results

Our first result is the following:

THEOREM 1.1. *Assume that $\mathcal{U}_{\rm ad}$ is a non-empty closed convex set of $L^2(\omega)$. Then, Problem* P1 *possesses exactly one solution.*

PROOF. For the proof we only have to check that $u \mapsto J(u)$ is a strictly convex, coercive and weakly lower semicontinuous function on $L^2(\omega)$.

But this is very easy to verify. In fact, $u \mapsto J(u)$ can be written in the form

$$J(u) = \frac{1}{2}\, a_0(u,u) + a_1(u) + a_2 \quad \forall u \in \mathcal{U}_{\rm ad}\,, \tag{1.13}$$

where $a_0(\cdot\,,\cdot)$ is a continuous and coercive bilinear form on $L^2(\omega)$, $a_1(\cdot)$ is a continuous linear form on $L^2(\omega)$ and $a_2 \in \mathbb{R}$.

The forms $a_0(\cdot\,,\cdot)$ and $a_1(\cdot)$ are given as follows:

$$a_0(u,v) = a\int_\Omega yz\,dx + b\int_\omega uv\,dx$$

and

$$a_1(u) = -a\int_\Omega y_d y\,dx,$$

where y (resp. z) is the solution to (1.1) (resp. (1.1) with u replaced by v). On the other hand,

$$a_3 = \frac{a}{2}\int_\Omega |y_d|^2\,dx.$$

Hence, the usual arguments of the *direct method of the Calculus of Variations* lead to the existence and uniqueness of solution, as assserted. □

QUESTION 1. *What can be said if, in* (1.3)*, we assume that* $b = 0$*? Which interpretation can be given to the corresponding optimal control problem?*

We will now be concerned with the computation of $J'(u)$ and the obtention of an optimality system. Our result is the following:

THEOREM 1.2. *Assume that* $\mathcal{U}_{\rm ad} \subset L^2(\omega)$ *is a non-empty closed convex set and let* $\hat{u}$ *be the solution to Problem* P1*. Then there exists* $\hat{y}$ *and* $\hat{p}$ *such that the following optimality system is satisfied:*

$$\left\{\begin{array}{lll} -\Delta\hat{y} = \hat{u}1_\omega & in & \Omega,\\ \hat{y} = 0 & on & \Gamma, \end{array}\right. \tag{1.14}$$

$$\left\{\begin{array}{lll} -\Delta\hat{p} = \hat{y} - y_d & in & \Omega,\\ \hat{p} = 0 & on & \Gamma, \end{array}\right. \tag{1.15}$$

$$\int_\omega (a\hat{p} + b\hat{u})(u - \hat{u})\,dx \geq 0 \quad \forall u \in \mathcal{U}_{\rm ad}\,. \tag{1.16}$$

PROOF. For the proof, we argue as follows. Since $\hat{u}$ is the solution to Problem P1, we must have

$$\langle J'(\hat{u}), u - \hat{u}\rangle \geq 0 \quad \forall u \in \mathcal{U}_{\rm ad}\,, \quad \hat{u} \in \mathcal{U}_{\rm ad}\,. \tag{1.17}$$

Here, $\langle\cdot\,,\cdot\rangle$ denotes the scalar product in $L^2(\omega)$. Taking into account (1.13), this can be written as follows:

$$a_0(\hat{u}, u - \hat{u}) + a_1(u - \hat{u}) \geq 0$$

that is to say,

$$a\int_\Omega (\hat{y} - y_d)(y - \hat{y})\,dx + b\int_\omega \hat{u}\,(u - \hat{u})\,dx \geq 0 \tag{1.18}$$

for all $u \in \mathcal{U}_{\rm ad}$. Of course, in (1.18) y is the solution to (1.1) and $\hat{y}$ is the solution to (1.1) with u replaced by $\hat{u}$.

Let $\hat{p}$ be the solution to (1.15), the *adjoint system*. It is then clear that

$$\int_\Omega (\hat{y} - y_d)(y - \hat{y})\, dx = \int_\Omega \nabla \hat{p} \cdot \nabla (y - \hat{y})\, dx = \int_\omega \hat{p}\, (u - \hat{u})\, dx.$$

Consequently, (1.18) is equivalent to (1.16). This proves that the optimality system (1.14) – (1.16) must hold. □

REMARK 1.3. In this particular case, we also have the reciprocal or theorem 1.2: If $\hat{u} \in \mathcal{U}_{\rm ad}$ and there exist $\hat{y}$ and $\hat{p}$ such that (1.14) – (1.16) holds, then $\hat{u}$ is the unique solution to Problem P1. □

It is usual to say that $\hat{p}$ is the *adjoint state* associate to the optimal control $\hat{u}$. In fact, in view of the previous argument, for each $u \in \mathcal{U}_{\rm ad}$, we have

$$\langle J'(u), v\rangle = \int_\omega (ap + bu)\, v\, dx \quad \forall v \in \mathcal{U}_{\rm ad}\,, \tag{1.19}$$

where p is the adjoint state associate to u, i.e. the solution to

$$\left\{ \begin{array}{lll} -\Delta p = y - y_d & \text{in} & \Omega, \\ p = 0 & \text{on} & \Gamma. \end{array} \right. \tag{1.20}$$

This provides a very useful technique to compute the derivative $J'(u)$ for a given u. From the practical viewpoint this is very important, since a method to compute $J'(u)$ permits the use of *descent methods* in order to determine the optimal control $\hat{u}$.

QUESTION 2. *The optimality system in theorem 1.2 suggests the following iterative method for the computation of $\hat{u}$:*

$$\left\{ \begin{array}{lll} -\Delta y^n = u^{n-1} 1_\omega & in & \Omega, \\ y^n = 0 & on & \Gamma, \end{array} \right. \tag{1.21}$$

$$\left\{ \begin{array}{lll} -\Delta p^n = y^n - y_d & in & \Omega, \\ p^n = 0 & on & \Gamma, \end{array} \right. \tag{1.22}$$

$$\int_\omega (ap^n + bu^n)(u - u^n)\, dx \geq 0 \quad \forall u \in \mathcal{U}_{\rm ad}. \tag{1.23}$$

What can be said on the convergence of these iterates?

QUESTION 3. *In view of* (1.19) – (1.20)*, how can we apply (for instance) the fixed-step gradient method to produce a sequence $\{u^n\}$ of controls converging to the optimal control $\hat{u}$? What about the optimal-step gradient method? What about the fixed-step and optimal-step conjugate gradient methods?*

The previous ideas can be generalized in several directions. We will present a generalization involving nonlinear elliptic state systems and nonquadratic cost functionals.

Thus, let us introduce the system

$$\left\{ \begin{array}{lll} Ay + f(y) = 1_\omega u & \text{in} & \Omega, \\ y = 0 & \text{on} & \Gamma, \end{array} \right. \tag{1.24}$$

where A is a linear second order partial differential operator given by

$$Ay = -\sum_{i,j=1}^{2} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial y}{\partial x_j}\right) + \sum_{j=1}^{2} b_j(x)\frac{\partial y}{\partial x_j} + c(x)y \tag{1.25}$$

and $f : \mathbb{R} \mapsto \mathbb{R}$ is (for instance) a nondecreasing C^1 function satisfying

$$|f(s)| \leq C(1 + |s|) \quad \forall s \in \mathbb{R}. \tag{1.26}$$

We will assume that the coefficients a_{ij} , b_i and c satisfy:

$$\begin{aligned} &a_{ij}\,, b_i\,, c \in L^\infty(\Omega), \quad c \geq 0, \\ &\sum_{i,j=1}^{2} a_{ij}(x)\xi_i\xi_j \geq \alpha|\xi|^2 \quad \forall \xi \in \mathbb{R}^2 \quad \text{a.e. in } \Omega, \quad \alpha > 0. \end{aligned} \tag{1.27}$$

For each $u \in L^2(\omega)$, the corresponding system (1.24) possesses exactly one solution $y \in H_0^1(\Omega)$. Let $\mathcal{U}_{\rm ad} \subset L^2(\omega)$ be a family of admissible controls. We will now set

$$J(u) = \int_\Omega F(x, y(x), u(x))\, dx \quad \forall u \in \mathcal{U}_{\rm ad}\,, \tag{1.28}$$

where $F = F(x, s, v)$ is assumed to be a *Carathéodory function*, defined for $(x, s, v) \in \Omega \times \mathbb{R} \times \mathbb{R}$. We consider the following generalization of Problem P1:

PROBLEM P1′: *To find $\hat{u} \in \mathcal{U}_{\rm ad}$ such that $J(\hat{u}) \leq J(u)$ for all $u \in \mathcal{U}_{\rm ad}$, where J is given by* (1.24),(1.28).

Among all possible results that can be established in this context, let us indicate the following, that has been taken from [16]:

THEOREM 1.4. *Assume that $\mathcal{U}_{\rm ad}$ is a closed convex subset of $L^2(\omega)$. Also, assume that F is of the form*

$$F(x, s, v) = F_0(x, s) + F_1(x, v)\, 1_\omega(x),$$

where F_0 and F_1 are Carathéodory functions satisfying:

$$\begin{cases} |F_0(x, s)| \leq C(1 + |s|^2) \quad \forall (x, s) \in \Omega \times \mathbb{R}, \\ a|v|^2 \leq F_1(x, v) \leq C(1 + |v|^2) \quad \forall (x, v) \in \omega \times \mathbb{R}, \quad a > 0, \\ F_1(x, \cdot) \text{ is convex for each } x \in \omega. \end{cases} \tag{1.29}$$

Then Problem P1′ *possesses at least one solution $\hat{u}$.*

The proof relies on arguments similar to those above but technically more involved. It will not be given here; see [16] for the details.

QUESTION 4. *What can be said if, in* (1.29)*, we have $a = 0$?*

Notice that, in the previous result, the convexity hypothesis on $F_1(x, \cdot)$ is essential. Indeed, let us consider the particular case in which the state system is

$$\begin{cases} -\Delta y = u & \text{in} \quad \Omega, \\ y = 0 & \text{on} \quad \Gamma, \end{cases} \tag{1.30}$$

the set $\mathcal{U}_{\rm ad}$ is

$$\mathcal{U}_{\rm ad} = \{\, u \in L^2(\Omega) : |u| \leq 1 \quad \text{a.e. in } \Omega \,\} \tag{1.31}$$

and the cost functional is given by

$$J(u) = \int_\Omega \left(|u|^2 - 1\right)^2 \, dx + \frac{1}{2}\int_\Omega |y|^2 \, dx \quad \forall u \in \mathcal{U}_{\rm ad}\,. \tag{1.32}$$

Then, it can be shown that

$$\inf_{u\in\mathcal{U}_{\rm ad}} J(u) = 0$$

and however

$$J(u) > 0 \quad \forall u \in \mathcal{U}_{\rm ad}\,,$$

whence the optimal control problem associate to (1.30), (1.31) and (1.32) has no solution.

To end this Subsection, let us recall a result concerning the optimality system for Problem P1′. We will need the adjoint operator A^*, which is given as follows:

$$A^*p = -\sum_{i,j=1}^{2} \frac{\partial}{\partial x_j}\left(a_{ij}(x)\frac{\partial p}{\partial x_i} + b_j(x)p\right) + c(x)p. \tag{1.33}$$

Then, one has:

THEOREM 1.5. *Assume that F is as above, that F_0 and F_1 possess bounded partial derivatives and, also, that* (1.29) *is satisfied. Let $\hat{u}$ be a solution to Problem* P1′. *Then there exist $\hat{y}$ and $\hat{p}$ such that the following optimality system is satisfied:*

$$\begin{cases} A\hat{y} + f(\hat{y}) = \hat{u}1_\omega & \text{in} \quad \Omega, \\ \hat{y} = 0 & \text{on} \quad \Gamma, \end{cases} \tag{1.34}$$

$$\begin{cases} A^*\hat{p} + f'(\hat{y})\hat{p} = \dfrac{\partial F_0}{\partial s}(x,\hat{y}) & \text{in} \quad \Omega, \\ \hat{p} = 0 & \text{on} \quad \Gamma, \end{cases} \tag{1.35}$$

$$\int_\omega \left(\hat{p} + \frac{\partial F_1}{\partial v}(x,\hat{u})\right)(u-\hat{u})\,dx \geq 0 \quad \forall u \in \mathcal{U}_{\rm ad}\,. \tag{1.36}$$

As before, the method of proof of this result provides an expression for the derivative $J'(u)$ of J at each u. More precisely, one finds that

$$\langle J'(u), v\rangle = \int_\omega \left(p + \frac{\partial F_1}{\partial v}(x,u)\right) v\,dx \quad \forall v \in \mathcal{U}_{\rm ad}\,, \tag{1.37}$$

where p is the adjoint state associate to u, i.e. the solution to

$$\begin{cases} A^*p + f'(y)p = \dfrac{\partial F_0}{\partial s}(x,y) & \text{in} \quad \Omega, \\ p = 0 & \text{on} \quad \Gamma \end{cases} \tag{1.38}$$

and y is the state, i.e. the solution to (1.24).

For other similar results, see for instance [14] and [17].

QUESTION 5. *Is there a way to use the optimality system in theorem 1.5 to prove a uniqueness result?*

QUESTION 6. *The optimality system in theorem 1.5 also suggests a "natural" iterative method for the computation of $\hat{u}$. Which one? What can be said on the convergence of the iterates?*

QUESTION 7. *In view of* (1.37) – (1.38)*, how can we apply gradient and conjugate gradient method to produce a sequence of controls that converge to an optimal control?*

3. Control on the coefficients, nonexistence and relaxation

In this Section we assume for simplicity that $N = 2$ and we consider Problem P2.

We will try to show the complexity of the problems in which the control is applied through coefficients in the principal part of the operator. We will first see that, in general, there exists no solution to this problem.

The following notation is needed. For given α and β with $\alpha, \beta > 0$, let us denote by $\mathcal{A}(\alpha, \beta)$ the family of 2×2 matrices A with components $A_{ij} \in L^\infty(\Omega)$ such that

$$A(x)\xi \cdot \xi \geq \alpha|\xi|^2, \quad (A(x))^{-1}\xi \cdot \xi \geq \frac{1}{\beta}|\xi|^2 \quad \forall \xi \in \mathbb{R}^2, \quad x \text{ a.e. in } \Omega. \tag{1.39}$$

It will be useful to recall the concept of *H-convergence*, which was introduced by F. Murat in 1978 (see [84],[85] and [88]):

DEFINITION 1.6. *Assume that $A^n \in \mathcal{A}(\alpha, \beta)$ for each $n \geq 1$ and that $A^0 \in \mathcal{A}(\alpha, \beta)$. It will be said that A^n H-converges to A^0 in Ω if, for any non-empty open set $\mathcal{O} \subset \Omega$ and any $g \in H^{-1}(\mathcal{O})$, the solution y^n of the elliptic problem*

$$\begin{cases} -\nabla \cdot (A^n(x)\nabla y) = g & \text{in} \quad \mathcal{O}, \\ y = 0 & \text{on} \quad \partial\mathcal{O}, \end{cases} \tag{1.40}$$

satisfies

$$y^n \to y^0 \quad \text{weakly in } H_0^1(\mathcal{O})$$

and

$$A^n \nabla y^n \to A^0 \nabla y^0 \quad \text{weakly in } L^2(\mathcal{O}),$$

where y^0 is the unique solution of the problem

$$\begin{cases} -\nabla \cdot (A^0(x)\nabla y) = g & \text{in} \quad \mathcal{O}, \\ y = 0 & \text{on} \quad \partial\mathcal{O}. \end{cases} \tag{1.41}$$

It can be seen that the family $\mathcal{A}(\alpha, \beta)$ is *closed* for the H-convergence. The following is also true:

THEOREM 1.7. *The family $\mathcal{A}(\alpha, \beta)$ is compact for the H-convergence. In other words, any sequence in $\mathcal{A}(\alpha, \beta)$ possesses subsequences that H-converge in $\mathcal{A}(\alpha, \beta)$.*

A key point is that we can have all the A^n of the form

$$A^n = a^n I \quad \forall n \geq 1,$$

while the H-limit A^0 can have extra-diagonal terms. In fact, explicit examples can be constructed and, in particular, we can find $A^0 \in \mathcal{A}(\alpha, \beta)$ and $f^0 \in H^{-1}(\Omega)$ with the following two properties:

(a) A^0 is the H-limit of a sequence of the form $a^n I$, with $a^n(x) = \alpha$ or $a^n(x) = \beta$ a.e.

(b) Let y^0 be the solution to (1.41) with g replaced by f^0. Then there is no function a with $a(x) = \alpha$ or $a(x) = \beta$ a.e. such that y^0 solves (1.5) with f replaced by f^0.

We are now ready to prove that Problem P2 has no solution in general. Let us take $f = f^0$ and $y_d = y^0$, where y^0 is the solution of (1.41) with g replaced by f^0. In view of the properties of A^0, it is clear that

$$\inf_{a \in \mathcal{A}_{\rm ad}} j(a) = 0$$

(recall that $\mathcal{A}_{\rm ad}$ is given by (1.7)). However, in view of the properties of f^0, we also have

$$j(a) > 0 \quad \forall a \in \mathcal{A}_{\rm ad}\,.$$

As a consequence, we must modify the definition of *optimal material*. Note that minimizing sequences do exist and that, in fact, they "describe" the optimal behavior. Consequently, it seems natural to adopt a new formulation in which the limits of minimizing sequences are distinguished material configurations. A satisfactory strategy consists of introducing a *relaxed problem*.

Relaxation is a useful tool in Optimization. Roughly speaking, to *relax* an extremal problem, say (P), is to introduce a second one, denoted by (Q), satisfying the following three conditions:

(a) (Q) possesses at least one solution.
(b) Any solution to (Q) can be written as the limit (in some sense) of a minimizing sequence for (P).
(c) Conversely, any minimizing sequence for (P) contains a subsequence that converges (in the same sense) to a solution of (Q).

For an overview on the role of the notion of relaxation in control problems, see [67] and [93]. We will only present here an intuitive and very simple argument which leads to a relaxed problem for P2.

The main point is to determine the "closure" in $\mathcal{A}(\alpha, \beta)$ of the family formed by the matrices of the form aI, with $a \in \mathcal{A}_{\rm ad}$. The answer is given by the following result:

THEOREM 1.8. *Let $\tilde{\mathcal{A}}_{\rm ad}$ be the family of all $A \in \mathcal{A}(\alpha, \beta)$ with the following two properties:*

(a) $A(x)$ is symmetric for x a.e. in Ω.
(b) For almost all x, the eigenvalues $\lambda_1(x)$ and $\lambda_2(x)$ of the matrix $A(x)$ satisfy:

$$\alpha \le \lambda_1(x) \le \lambda_2(x) \le \beta, \quad \frac{\alpha\beta}{\alpha + \beta - \lambda_2(x)} \le \lambda_1(x). \tag{1.42}$$

Then, if A is given in $\mathcal{A}(\alpha, \beta)$, one has $A \in \tilde{\mathcal{A}}_{\rm ad}$ if and only if A can be written as the H-limit of a sequence of matrices of the form $a^n I$, with $a^n \in \mathcal{A}_{\rm ad}$ for all n.

This is proved in [106] (see also [88]). At this respect, it is worth mentioning that, in a similar N-dimensional situation with $N \ge 3$, the determination of the set of H-limits of the matrices of the form aI with $a \in \mathcal{A}_{\rm ad}$ is an open problem.

The previous result permits to introduce a new control problem which is nothing but the relaxation of Problem P2.

Namely, for each $A \in \tilde{\mathcal{A}}_{\rm ad}$, let us consider the (relaxed) state system

$$\begin{cases} -\nabla \cdot (A(x)\nabla Y) = f(x) & \text{in} \quad \Omega, \\ Y = 0 & \text{on} \quad \Gamma \end{cases} \tag{1.43}$$

and let us set

$$k(A) = \frac{1}{2}\int_\Omega |Y - y_d|^2\, dx \quad \forall A \in \tilde{\mathcal{A}}_{\rm ad}\,. \tag{1.44}$$

The relaxed problem is then:

> PROBLEM P2′: *To find $\hat{A} \in \tilde{\mathcal{A}}_{\rm ad}$ such that $k(\hat{A}) \le k(A)$ for all $A \in \tilde{\mathcal{A}}_{\rm ad}$, where $\tilde{j}$ is given by* (1.44).

Indeed, the following can be proved:

THEOREM 1.9. *Assume that $f \in H^{-1}(\Omega)$ and $y_d \in L^2(\Omega)$ are given. Then, there exists at least one solution $\hat{A}$ to Problem* P2′. *This can be written as the H-limit of a minimizing sequence for Problem* P2. *Furthermore, any minimizing sequence for Problem* P2 *contains a subsequence that H-converges to a solution of Problem* P2′.

The proof of this result is not difficult, taking into account the definition of H-convergence and the fact that $\tilde{\mathcal{A}}_{\rm ad}$ is the H-closure of $\mathcal{A}_{\rm ad}$.

From a physical viewpoint, we see that the "generalized" solution to the original problem is a *composite material*. In general, it is anisotropic, i.e. $\hat{A}_{ij}(x)$ may be $\neq 0$ for $i \neq j$.

QUESTION 8. *Is it possible to deduce an optimality system for the solutions to Problem* P2′*? Which one? Does this optimality system lead to convergent iterates?*

QUESTION 9. *Is it possible to compute $k'(A)$ easily and use this computation to apply gradient and/or conjugate gradient methods in the context of Problem* P2′*?*

The reader is referred to [82] and the references therein for more details on the control of coefficients, the generation of composite materials and other related topics.

4. Optimal design and domain variations

We will now consider Problem P3.

This is an *optimal design* problem. The feature is that, now, the control is a geometric datum in (1.8) (the set B). Accordingly, we have to minimize a function over a set $\mathcal{B}_{\rm ad}$ where there is no vector structure at our disposal. It is thus reasonable to expect a higher level of difficulty than for other optimal control problems.

As mentioned above, the existence of a solution to Problem P3 is not clear at all. To simplify our arguments, let us introduce two non-empty open sets D_0 and D_1, with

$$D_0 \subset\subset D_1 \subset\subset \Omega$$

and let us first assume that $\mathcal{B}_{\rm ad}$ is the family of the non-empty closed sets B with piecewise Lipschitz-continuous boundary that satisfy

$$\overline{D_0} \subset B \subset \overline{D_1}\,. \tag{1.45}$$

Also, assume that $|y_\infty|$ is small enough (depending on ν and Ω). Then, for each $B \in \mathcal{B}_{\rm ad}$, the state system (1.8) possesses exactly one solution (y, π) (the pressure π is unique up to an additive constant). Consequently, we can assign to B a *drag* $D(B) = T(B, y)$, given by (1.9).

In other words, in this case the function $B \mapsto D(B)$ is well-defined and Problem P3 reads:

To find $\hat{B} \in \mathcal{B}_{\rm ad}$ *such that* $D(\hat{B}) \leq D(B)$ *for all* $B \in \mathcal{B}_{\rm ad}$.

Let $\{B^n\}$ be a minimizing sequence. For each $n \geq 1$, let us denote by y^n the velocity field associated to B^n by (1.8). Then, it is clear that y^n is uniformly bounded in the H^1-norm. More precisely, the extensions-by-zero of y^n to the whole domain Ω, that we denote by $\tilde{y}^n$, are uniformly bounded in $H^1(\Omega; \mathbb{R}^2)$. We can thus assume that $\tilde{y}^n$ converges weakly in $H^1(\Omega; \mathbb{R}^2)$, strongly in $L^2(\Omega; \mathbb{R}^2)$ and a.e. to a function $\tilde{y}^0$. This is a consequence of the compactness of the embedding $H^1(\Omega) \hookrightarrow L^2(\Omega)$; see for instance [1].

On the other hand, since $\{B^n\}$ is a sequence of closed sets of Ω, we can also assume that B^n converges in the sense of the *Haussdorf distance* d_H to a closed set B^0. This is a consequence of the fact that the family of closed subsets of Ω is compact for d_H; see [25].

At this respect, recall that, when B and B' closed sets in $\mathbb{R}^2$, the Haussdorf distance $d_H(B, B')$ is given by

$$d_H(B, B') = \max\{\rho(B, B'), \rho(B', B)\},$$

where

$$\rho(B, B') = \sup_{x \in B} d(x, B') \quad \text{and} \quad d(x, B') = \inf_{x' \in B'} |x - x'| \quad \text{for all } B \text{ and } B'$$

and a similar definition holds for $\rho(B', B)$.

The set B^0 satisfies (1.45). However, the uniform bound in the H^1 norm does not give enough regularity for B^0 and it is not clear whether the restriction of $\tilde{y}^0$ to the limit set $\Omega \setminus B^0$ is, together with some π^0, the solution of (1.8) with B replaced by B^0.

We can overcome this difficulty by introducing a more restrictive family $\mathcal{B}_{\rm ad}$.

For instance, let us now assume that $\mathcal{B}_{\rm ad}$ is the family of the non-empty closed sets B satisfying (1.45) whose boundaries are *uniformly Lipschitz-continuous* with constant $L > 0$. By this we mean that the boundary ∂B of any $B \in \mathcal{B}_{\rm ad}$ can be written in the form

$$\partial B = \{\, x(\theta) : \theta \in [0, 1] \,\}, \tag{1.46}$$

where the function $\theta \mapsto x(\theta)$ satisfies $x(0) = x(1)$ and is Lipschitz-continuous on $[0, 1]$ with Lipschitz constant L. Obviously, $\mathcal{B}_{\rm ad}$ is non-empty if L is large enough.

It is clear that we can argue as before and find a limit set B^0 and a vector field $\tilde{y}^0$, defined in Ω. In this particular case, the set B^0 belongs to $\mathcal{B}_{\rm ad}$, that is, its boundary is also of the form (1.46), see [21]. In view of this regularity property for B^0, it can also be proved that the restriction of $\tilde{y}^0$ to $\Omega \setminus B^0$ is, together with an appropriate π^0, the solution of (1.8) with $B = B^0$.

QUESTION 10. *How can this be proved?*

Unfortunately, this new definition of the admissible set $\mathcal{B}_{\rm ad}$ can be too restrictive.

Actually, this is a common fact for optimal design problems: either we choose the apparently natural definition of $\mathcal{B}_{\rm ad}$ (and then existence is not known) or we make it more restrictive (and then the problem can become unrealistic). For more details on these and other similar results, see [60, 95, 96].

We will now study the behavior of the function $B \mapsto D(B)$. Let $\hat{B}$ be a reference shape for the body (arbitrary in $\mathcal{B}_{\rm ad}$ but fixed). The body variations are described by a field $u = u(x)$ and we search for a formula of the kind

$$D(\hat{B} + u) = D(\hat{B}) + D'(\hat{B}; u) + o(u), \tag{1.47}$$

where the modified fluid domain is

$$(\Omega \setminus \hat{B}) + u = \Omega \setminus (\hat{B} + u) = \{\, x \in \mathbb{R}^2 : x = (I + u)(\xi),\ \xi \in \Omega \setminus \hat{B} \,\}$$

and

$$o(u)\|u\|^{-1}_{W^{1,\infty}} \to 0 \ \text{ as } \ \|u\|_{W^{1,\infty}} \to 0.$$

We are thus led to an analysis of the differentiability of the function $u \mapsto D(\hat{B}+u)$.

A lot of work has been made for the definition and computation of the variations with respect to a domain of functionals defined through the solutions to boundary value problems. The reader is referred to [102] and the references therein.

We will recall briefly a variant of a general method introduced by F. Murat and J. Simon in [86] and [87][1]. This is taken from [10]. Notice that some formal computations of the derivative were previously carried out by O. Pironneau in [94] (see also [96]), using "normal" variations.

We will choose fields $u \in W^{1,\infty}(\mathbb{R}^2;\mathbb{R}^2)$ such that $u = 0$ on Γ. This includes many interesting situations in which $\partial(\Omega \setminus (\hat{B} + u))$ possesses "corner" points. Furthermore, the equality $u = 0$ on Γ expresses the fact that the outer boundary limiting the fluid is fixed.

We will also assume that $\|u\|_{W^{1,\infty}} \le \eta$, with η being small enough to ensure that the boundary of $\Omega \setminus (\hat{B} + u)$ is Lipschitz-continuous and also that $\hat{B} + u$ is included in a fixed open set D_2 satisfying

$$\hat{B} \subset\subset D_2 \subset\subset \Omega$$

(such a constant $\eta > 0$ exists, see [10] for a proof).

For the sequel, we introduce

$$\mathcal{W} = \{\, u \in W^{1,\infty}(\mathbb{R}^2;\mathbb{R}^2) :\ \|u\|_{W^{1,\infty}} \le \eta, \quad u = 0 \text{ on } \partial\Omega \,\}.$$

Now, we choose g satisfying

$$\nabla \cdot g = 0, \quad g = y_\infty \text{ in a neighborhood of } \partial\Omega, \quad g = 0 \text{ in a neighborhood of } D_2$$

(such a function g always exists; see for instance [51]). If $u \in \mathcal{W}$, one has $g = 0$ in a neighborhood of $\partial\hat{B} + u$. After normalization of the pressure, the Navier-Stokes

[1]The general method in [86] and [87] cannot be directly applied to the Stokes and Navier-Stokes cases. This is due to the incompressibility condition.

problem in $\Omega \setminus (\hat{B}+u)$ can be written as follows:

$$\begin{cases} -\nu\Delta y(u) + (y(u)\cdot\nabla)y(u) + \pi(u) = 0, \quad \nabla\cdot y(u) = 0 \quad \text{in } \Omega\setminus(\hat{B}+u), \\ y(u) - g \in H_0^1(\Omega\setminus(\hat{B}+u);\mathbb{R}^2), \\ \pi(u) \in L^2(\Omega\setminus(\hat{B}+u)), \quad \displaystyle\int_{\Omega\setminus\hat{B}} \pi(u)\circ(I+u)\,dx = 0. \end{cases} \tag{1.48}$$

The drag associated to $\hat{B}+u$ can be defined and is given by

$$D(\hat{B}+u) = T(\hat{B}+u, y(u)) = 2\nu\int_{\Omega\setminus(\hat{B}+u)} |Dy(u)|^2\,dx, \tag{1.49}$$

where $Dy(u) = \frac{1}{2}(\nabla y(u) + \nabla y(u)^t)$.

Under these conditions, it is proved in [10] that the equality (1.47) is satisfied, with the first order term $D'(\hat{B};u)$ given by

$$D'(\hat{B};u) = 4\nu\int_{\Omega\setminus\hat{B}} Dy\cdot\left(D\dot{y}(u) - E(u,y) + \frac{1}{2}(\nabla\cdot u)Dy\right)\,dx.$$

Here, we have introduced the following notation:

(a) $(\dot{y}(u),\dot{\pi}(u))$ is the unique solution to the linear problem

$$\begin{cases} -\nu\Delta\dot{y}(u) + (y\cdot\nabla)\dot{y}(u) + (\dot{y}(u)\cdot\nabla)y + \dot{\pi}(u) = G(u,y,\pi),\ \nabla\cdot\dot{y}(u) = 0 \text{ in } \Omega\setminus\hat{B}, \\ \dot{y}(u) \in H_0^1(\Omega\setminus\hat{B};\mathbb{R}^2), \\ \dot{\pi}(u) \in L^2(\Omega\setminus\hat{B}), \quad \displaystyle\int_{\Omega\setminus\hat{B}} \dot{\pi}(u)\,dx = 0, \end{cases}$$

where

$$G(u,y,\pi) = -\nu\Delta((u\cdot\nabla)y) + \big(((u\cdot\nabla)y)\cdot\nabla\big)y + (y\cdot\nabla)((u\cdot\nabla)y) + \nabla(u\cdot\nabla\pi).$$

(b) $E(u,y)$ is the 2×2 tensor whose (i,j)-th component is given by

$$E_{ij}(u,y) = \frac{1}{2}\sum_k\left(\frac{\partial u_k}{\partial x_i}\frac{\partial y_j}{\partial x_k} + \frac{\partial u_k}{\partial x_j}\frac{\partial y_i}{\partial x_k}\right).$$

(c) $y = y(0)$ and $\pi = \pi(0)$, i.e. (y,π) is the solution to (1.48) for $u = 0$.

It can also be proved that, if B and Ω are $W^{2,\infty}$ domains and $u \in W^{2,\infty}(\mathbb{R}^2;\mathbb{R}^2)$, then $y \in H^2(\Omega;\mathbb{R}^2)$, $\pi \in H^1(\Omega)$ and

$$D'(\hat{B};u) = \int_{\partial\hat{B}}\left(\frac{\partial w}{\partial n} - \frac{\partial y}{\partial n}\right)\cdot\frac{\partial y}{\partial n}\,(u\cdot n)\,d\sigma, \tag{1.50}$$

with (w,q) being the unique solution to the "adjoint" problem

$$\begin{cases} -\nu\Delta w_i + \sum_j \partial_i y_j\, w_j - \sum_j y_j\partial_j w_i + \partial_i q = -2\nu\,\Delta y_i\ (1\le i\le 2),\ \nabla\cdot w = 0, \\ w \in H_0^1(\Omega\setminus\hat{B};\mathbb{R}^2)\cap H^2(\Omega\setminus\hat{B};\mathbb{R}^2), \\ q \in H^1(\Omega\setminus\hat{B}),\ \displaystyle\int_{\Omega\setminus\hat{B}} q\,dx = 0, \end{cases} \tag{1.51}$$

Notice that, in order to compute the derivative of the drag in several directions u, it is interesting to use the identity (1.50). Indeed, it suffices to solve (1.8)

and (1.51) only once. Then, to determine $D'(\hat{B};u)$ for a given u, we will only have to compute one integral on $\partial\hat{B}$.

QUESTION 11. *Assume that $\mathcal{B}_{\rm ad}$ is the family of the non-empty closed sets B satisfying* (1.45) *whose boundaries are uniformly Lipschitz-continuous with Lipschitz constant L. How can* (1.50) *be used to produce a sequence $\{B^n\}$ "converging" to a solution to Problem* P3*?*

To end this Section, let us state another result from [10]:

THEOREM 1.10. *There exists $\alpha>0$ such that, if $|y_\infty|\le\alpha\nu$, then $u\mapsto D(\hat{B}+u)$ is a C^∞ mapping in the set $\mathcal{W}$.*

One can also obtain expressions for the derivatives of higher orders. This must be made with caution; indeed, $D''(\hat{B};\cdot,\cdot)$ (i.e. the second derivative at 0 of $u\mapsto D(\hat{B}+u)$) does not coincide with $(D'(\hat{B};\cdot)';\cdot)$ (i.e. the derivative at 0 of the mapping $u\mapsto D'(\hat{B}+u;\cdot)$), see [101].

5. Optimal control for a system modelling tumor growth

This Section deals with Problem P4. For simplicity, we will assume that the functions f, h, F and H are given by (1.11), where ρ, m, R and M are positive constants. We will also assume that the initial data in (1.10) satisfy:

$$c_0\,,\beta_0\in L^\infty(\Omega)\cap H_0^1(\Omega),\quad c_0\,,\beta_0\ge 0.$$

For each $v\in L^2(\omega\times(0,T))$ with $v\ge 0$, there exists at least one solution (c,β) to (1.10), with

$$c\in L^\infty(Q),\quad c_t\,,\frac{\partial c}{\partial x_i}\,,\frac{\partial^2 c}{\partial x_i\partial x_j}\in L^2(Q)$$

and the same properties for β.

QUESTION 12. *Why is this true? What about uniqueness?*

Then the following results can be proved:

THEOREM 1.11. *Assume that $\mathcal{V}_{\rm ad}$ is a non-empty closed convex set of $L^2(\omega)$ and all $v\in\mathcal{V}_{\rm ad}$ satisfy $v\ge 0$. Then Problem* P4 *possesses at least one solution.*

THEOREM 1.12. *Let the assumptions of theorem 1.11 be satisfied and let $\hat{u}$ be a solution to Problem* P4*. Then there exists $(\hat{c},\hat{\beta})$ and $(\hat{p},\hat{\eta})$ such that*

$$\begin{cases}
\hat{c}_t-\nabla\cdot(D(x)\nabla\hat{c})=\rho\hat{c}-R\hat{c}\hat{\beta} & \text{in} \quad Q=\Omega\times(0,T),\\
\hat{\beta}_t-\mu\Delta\hat{\beta}=-m\hat{\beta}-M\hat{c}\hat{\beta}+v1_\omega & \text{in} \quad Q=\Omega\times(0,T),\\
\hat{c}=0 & \text{on} \quad \Sigma=\partial\Omega\times(0,T),\\
\hat{\beta}=0 & \text{on} \quad \Sigma=\partial\Omega\times(0,T),\\
\hat{c}(x,0)=c_0(x) & \text{in} \quad \Omega,\\
\hat{\beta}(x,0)=\beta_0(x) & \text{in} \quad \Omega,
\end{cases}\tag{1.52}$$

$$\begin{cases} -\hat{p}_t - \nabla\cdot(D(x)\nabla\hat{p}) = \rho\hat{p} - R\hat{\beta}\hat{p} - M\hat{\beta}\hat{\eta} & \textit{in} \quad Q=\Omega\times(0,T), \\ -\hat{\eta}_t - \mu\Delta\hat{\eta} = -m\hat{\eta} - R\hat{c}\hat{p} - M\hat{c}\hat{\eta} & \textit{in} \quad Q=\Omega\times(0,T), \\ \hat{p}=0 & \textit{on} \quad \Sigma=\partial\Omega\times(0,T), \\ \hat{\eta}=0 & \textit{on} \quad \Sigma=\partial\Omega\times(0,T), \\ \hat{p}(x,T)=\hat{c}(x,T) & \textit{in} \quad \Omega, \\ \hat{\eta}(x,T)=0 & \textit{in} \quad \Omega, \end{cases} \tag{1.53}$$

$$\iint_{\omega\times(0,T)} (a\hat{p}+b\hat{u})(u-\hat{u})\,dx\,dt \geq 0 \quad \forall u\in\mathcal{V}_{\rm ad}. \tag{1.54}$$

For the proofs, the arguments are not too different from those in Section 2.

Again, it is common to say that $(\hat{p},\hat{\eta})$ is the *adjoint state* associate to the optimal control $\hat{u}$. Also,

$$\langle J'(u), v\rangle = \iint_{\omega\times(0,T)} (ap+bu)\,v\,dx\,dt \quad \forall v\in\mathcal{V}_{\rm ad}\,, \tag{1.55}$$

where (p,η) is the adjoint state associate to u, i.e. the solution to

$$\begin{cases} -p_t - \nabla\cdot(D(x)\nabla p) = \rho p - R\beta p - M\beta\eta & \text{in} \quad Q=\Omega\times(0,T), \\ -\eta_t - \mu\Delta\eta = -m\eta - Rcp - Mc\eta & \text{in} \quad Q=\Omega\times(0,T), \\ p=0 & \text{on} \quad \Sigma=\partial\Omega\times(0,T), \\ \eta=0 & \text{on} \quad \Sigma=\partial\Omega\times(0,T), \\ p(x,T)=c(x,T) & \text{in} \quad \Omega, \\ \eta(x,T)=0 & \text{in} \quad \Omega. \end{cases}$$

Once more, this provides very useful techniques to compute, for any control u, the associate $J'(u)$.

Question 13. *Can the optimality system in theorem 1.12 be used to prove a uniqueness result for Problem* P4*?*

Question 14. *Again, a "natural" iterative method for the computation of $\hat{u}$ is suggested by the optimality system in theorem 1.12. Which is this method? What can be said on the convergence of the iterates?*

Question 15. *How can we apply gradient and conjugate gradient method to produce a sequence of controls that converge to an optimal control in he context of Problem* P4*?*

This optimal control problem has been solved numerically in [28]; more results will be given in a forthcoming paper.

CHAPTER 2

Controllability of the linear heat and wave PDEs

This Lecture is devoted to the controllability of some systems governed by linear time-dependent PDEs. I will consider the heat and the wave equations. I will try to explain which is the meaning of controllability and which kind of controllability properties can be expected to be satisfied by each of these PDEs. The main related results, together with the main ideas in their proofs, will be recalled.

1. Introduction

Let us first make some very general considerations on the following abstract problem:

$$\begin{cases} y_t - Ay = Bv, \quad t \in (0,T), \\ y(0) = y^0, \end{cases} \tag{2.1}$$

where A and B are linear operators, $v = v(t)$ is the control and $y = y(t)$ is the state.

For fixed $T > 0$, we choose y^0 and y^1 in the space of states (the space where y "lives") and we try to answer the following question:

QUESTION 1. *Can one find a control v such that the solution y associated to v and y^0 takes the value y^1 at $t = T$?*

This is an *exact controllability* problem. The control requirement $y(T) = y^1$ can be relaxed in various ways, leading to other notions of controllability.

Of course, the solvability of problems of this kind depends very much on the nature of the system under consideration; in particular, the following features may play a crucial role: time reversibility, regularity of the state, structure of the set of admissible controls, etc.

The controllability of partial differential equations has been the object of intensive research since more than 30 years. However, the subject is older than that. In 1978, D.L. Russell [99] made a rather complete survey of the most relevant results that were available in the literature at that time. In that paper, the author described a number of different tools that were developed to address controllability problems, often inspired and related to other subjects concerning partial differential equations: multipliers, moment problems, nonharmonic Fourier series, etc. More recently, J.-L. Lions introduced the so called *Hilbert Uniqueness Method* (H.U.M.; see [77, 78]). That was the starting point of a fruitful period for this subject.

It would be impossible to present here all the important results that have been proved in this area. I will thus only consider some model examples where the most interesting difficulties are found.

Several important related topics, like numerical computation and simulation in controllability problems, stabilizability, connections with finite dimensional controllability theory, etc. have been left out. However, some useful references for these issues have been included; see [23, 24, 57, 58, 59, 111].

2. Basic results for the linear heat equation

Let $\Omega \subset \mathbb{R}^N$ be a bounded domain ($N \geq 1$), with boundary Γ of class C^2. Let ω be an open and non-empty subset of Ω. Let $T > 0$ and consider the linear controlled heat equation in the cylinder $Q = \Omega \times (0, T)$:

$$\begin{cases} y_t - \Delta y = v 1_\omega & \text{in} \quad Q, \\ y = 0 & \text{on} \quad \Sigma, \\ y(x, 0) = y^0(x) & \text{in} \quad \Omega. \end{cases} \tag{2.2}$$

In (2.2), $\Sigma = \Gamma \times (0, T)$ is the lateral boundary of Q, 1_ω is the characteristic function of the set ω, $y = y(x, t)$ is the state and $v = v(x, t)$ is the control. Since v is multiplied by 1_ω, the action of the control is limitted to $\omega \times (0, T)$.

We assume that $y^0 \in L^2(\Omega)$ and $v \in L^2(\omega \times (0, T))$, so that (2.2) admits a unique solution

$$y \in C^0([0, T]; L^2(\Omega)) \cap L^2(0, T; H_0^1(\Omega)).$$

We will set $R(T; y^0) = \{ y(\cdot, T) : v \in L^2(\omega \times (0, T)) \}$. Then:

(a) It is said that system (2.2) is *approximately controllable* (at time T) if $R(T; y^0)$ is dense in $L^2(\Omega)$ for all $y^0 \in L^2(\Omega)$.
(b) It is said that (2.2) is *exactly controllable* if $R(T; y^0) = L^2(\Omega)$ for all $y^0 \in L^2(\Omega)$.
(c) Finally, it is said that (2.2) is *null controllable* if $0 \in R(T; y^0)$ for all $y^0 \in L^2(\Omega)$.

It will be seen below that approximate and null controllability hold for every non-empty open set $\omega \subset \Omega$ and every $T > 0$.

On the other hand, it is clear that exact controllability cannot hold, except possibly in the case in which $\omega = \Omega$. Indeed, due to the regularizing effect of the heat equation, the solutions of (2.2) at time $t = T$ are smooth in $\Omega \setminus \overline{\omega}$. Therefore, if $\omega \neq \Omega$, $R(T; y^0)$ is strictly contained in $L^2(\Omega)$ for all $y^0 \in L^2(\Omega)$.

Our first remark is that null controllability implies that the whole range of the semigroup generated by the heat equation is reachable too. Let us make this statement more precise.

Let us denote by $S(t)$ the semigroup generated by the heat equation (2.2) without control, i.e. with $v = 0$. Then, if null controllability holds, it follows that for any $y^0 \in L^2(\Omega)$ and any $y^1 \in S(T)(L^2(\Omega))$ there exists $v \in L^2(\omega \times (0, T))$ such that the solution of (2.2) satisfies $y(x, T) \equiv y^1(x)$. In other words,

$$S(T)(L^2(\Omega)) \subset R(T; y^0) \quad \forall y^0 \in L^2(\Omega).$$

Question 2. *Why is this true?*

The space $S(T)(L^2(\Omega))$ is dense in $L^2(\Omega)$. Therefore, null controllability implies approximate controllability. Observe however that the reachable states we obtain by this argument are smooth, due to the regularizing effect of the heat equation.

Notice that proving that null controllability implies approximate controllability requires the use of the density of $S(T)(L^2(\Omega))$ in $L^2(\Omega)$. In the case of the linear heat equation this is easy to check developing solutions in Fourier series. However, if the equation contains time or space-time dependent coefficients, this is true but not so immediate. In those cases, the density of the range of the "semigroup", can be reduced by duality to a backward uniqueness property, in the spirit of J.-L. Lions and B. Malgrange [81].

Our first main result is the following:

THEOREM 2.1. *System* (2.2) *is approximately controllable for any non-empty open set* $\omega \subset \Omega$ *and any* $T > 0$.

PROOF. This is an easy consequence of Hahn-Banach theorem. For completeness, we will reproduce the argument here.

Let us fix ω and $T > 0$. Then, it is clear that (2.2) is approximately controllable if and only if $R(T;0)$ is dense in $L^2(\Omega)$. But this is true if and only if any φ^0 in the orthogonal complement $R(T;0)^\perp$ is necessarily zero.

Let $\varphi^0 \in L^2(\Omega)$ be given and assume that it belongs to $R(T;0)^\perp$. Let us introduce the following backwards in time system:

$$\begin{cases} -\varphi_t - \Delta\varphi = 0 & \text{in} \quad Q, \\ \varphi = 0 & \text{on} \quad \Sigma, \\ \varphi(x,T) = \varphi^0(x) & \text{in} \quad \Omega. \end{cases} \tag{2.3}$$

Then, if $v \in L^2(\omega \times (0,T))$ is given and y is the solution to (2.2) with $y^0 = 0$, we have

$$\iint_{\omega\times(0,T)} \varphi v \, dx\, dt = \int_\Omega \varphi^0(x) y(x,T)\, dx = 0.$$

Consequently, approximate controllability holds if and only if the following uniqueness property is true:

> *If* φ *solves* (2.3) *and* $\varphi = 0$ *in* $\omega \times (0,T)$, *then necessarily* $\varphi \equiv 0$, *i.e.* $\varphi^0 = 0$.

But this is a well known uniqueness property for the heat equation, a consequence of the fact that the solutions to (2.3) are analytic in space.

This proves that approximate controllability holds for (2.2). □

Following the variational approach in [80], we can also determine the way the "good" control can be constructed. First of all, observe that it is sufficient to consider the particular case $y^0 = 0$. Then, let us fix $y^1 \in L^2(\Omega)$ and $\varepsilon > 0$ and let us introduce the following functional on $L^2(\Omega)$:

$$J_\varepsilon(\varphi^0) = \frac{1}{2}\iint_{\omega\times(0,T)} |\varphi|^2\, dx\, dt + \varepsilon\|\varphi^0\|_{L^2} - \int_\Omega \varphi^0 y^1\, dx, \tag{2.4}$$

where for each φ^0 we have denoted by φ the solution to the corresponding problem (2.3).

The functional J_ε is continuous and strictly convex in $L^2(\Omega)$. On the other hand, in view of the unique continuation property above, it can be proved that

$$\liminf_{\|\varphi^0\|_{L^2}\to\infty} \frac{J_\varepsilon(\varphi^0)}{\|\varphi^0\|_{L^2}} \geq \varepsilon. \tag{2.5}$$

Hence, J_ε admits a unique minimizer $\hat{\varphi}^0$ in $L^2(\Omega)$. The control $u = \hat{\varphi}|_{\omega\times(0,T)}$, where $\hat{\varphi}$ solves (2.3) with $\hat{\varphi}^0$ as final data is such that the solution of (2.2) (with $y^0 = 0$) satisfies

$$\|y(\cdot,T) - y^1\|_{L^2} \leq \varepsilon. \tag{2.6}$$

Question 3. *Why is* (2.5) *true? How can we prove* (2.6) *for this control?*

With a slight change in the definition of J_ε, we are also able to build *bang-bang* controls. Indeed, it suffices to consider the new functional

$$\tilde{J}_\varepsilon(\varphi^0) = \frac{1}{2}\left(\iint_{\omega\times(0,T)} |\varphi|\,dx\,dt\right)^2 + \varepsilon\|\varphi^0\|_{L^2} - \int_\Omega \varphi^0 y^1\,dx. \tag{2.7}$$

Then $\tilde{J}_\varepsilon$ is continuous and convex in $L^2(\Omega)$ and satisfies the coercivity property (2.5) too.

Let $\hat{\varphi}^0$ be a minimizer of $\tilde{J}_\varepsilon$ in $L^2(\Omega)$ and let $\hat{\varphi}$ be the corresponding solution of (2.3). Let us set

$$u = \left(\iint_{\omega\times(0,T)} |\hat{\varphi}|\,dx\,dt\right) \operatorname{sgn}(\hat{\varphi})|_{\omega\times(0,T)}, \tag{2.8}$$

where sgn is the multivalued sign function: $\operatorname{sgn}(s) = 1$ if $s > 0$, $\operatorname{sgn}(0) = [-1,1]$ and $\operatorname{sgn}(s) = -1$ when $s < 0$. Again, the control u given by (2.8) is such that the solution to (2.2) with zero initial data satisfies (2.6).

Due to the regularizing effect of the heat equation, the zero set of nontrivial solutions of (2.3) is of zero $(n+1)-$dimensional Lebesgue measure. Thus, the control u in (2.8) belongs to $L^\infty(Q)$ and is of *bang-bang* form, i.e. $u = \pm\lambda$ a.e. in $\omega\times(0,T)$, where

$$\lambda = \iint_{\omega\times(0,T)} |\hat{\varphi}|\,dx\,dt.$$

In fact, it can be proved that u minimizes the L^∞-norm in the set of all controls such that (2.6) is satisfied (we refer to [31] for a proof of this assertion).

Following [110], we can improve the previous argument and show that, for any ω, any $T > 0$ and any finite-dimensional subspace $E \subset L^2(\Omega)$, (2.2) is E-approximate controllable. This means that, for arbitrary $y^0, y^1 \in L^2(\Omega)$ and any $\varepsilon > 0$, there exists a control $v \in L^2(\omega\times(0,T))$ such that the corresponding solution to (2.2) satisfies:

$$\|y(\cdot,T) - y^1\|_{L^2} \leq \varepsilon, \quad \pi_E(y(\cdot,T)) = \pi_E(y^1). \tag{2.9}$$

Here, $\pi_E : L^2(\Omega) \mapsto E$ stands for the usual orthogonal projector on E.

Indeed, it suffices to modify J_ε (or $\tilde{J}_\varepsilon$) and use instead the functional J^E_ε, where

$$J^E_\varepsilon(\varphi^0) = \frac{1}{2}\iint_{\omega\times(0,T)} |\varphi|^2\,dx\,dt + \varepsilon\|(I-\pi_E)\varphi^0\|_{L^2} - \int_\Omega \varphi^0 y^1\,dx. \tag{2.10}$$

As before, J^E_ε is continuous, strictly convex and coercive in $L^2(\Omega)$. Once again, let us denote by $\hat{\varphi}^0$ its unique minimizer and let us set $u = \hat{\varphi}|_{\omega\times(0,T)}$. Then the associate state satisfies (2.9).

Question 4. *Which is in this case the argument leading to* (2.9)*? Is the hypothesis "E is finite-dimensional" essential?*

Let us now analyze the null controllability of (2.2).

The null controllability property for system (2.2), together with a L^2- estimate of the control, is equivalent to the following observability inequality for the adjoint system (2.3):

$$\|\varphi(\cdot,0)\|^2_{L^2} \leq C \iint_{\omega\times(0,T)} |\varphi|^2\, dx\, dt \quad \forall \varphi^0 \in L^2(\Omega). \tag{2.11}$$

QUESTION 5. *Which is the proof of this assertion?*

Due to the regularizing effect of the heat equation, the norm in the left hand side of (2.11) is very weak. However, the irreversibility of the system makes (2.11) difficult to prove. For instance, multiplier methods do not apply in this context.

Thus, we see that the approximate (resp. null) controllability of (2.2) is related to the unique continuation property (resp. the observability) of (2.3).

Historically, it seems that the first null controllability results established for the heat equation involved boundary controls. They were given in [99] in the one-dimensional case, using moment problems and classical results on the linear independence in $L^2(0,T)$ of families of real exponentials. Later, in [100], a deep general result was proved. Roughly speaking, the following was shown:

> *If the wave equation is controllable for some $T > 0$ with controls supported in ω, then the heat equation* (2.2) *is null controllable for every $T > 0$ with controls supported in ω.*

In view of the controllability results in Section 3, according to this principle, it follows that the heat equation (2.2) is null controllable for all $T > 0$ provided ω satisfies a specific geometric control condition. However, this geometric condition does not seem to be natural in the context of the heat equation and, therefore, this result is not completely satisfactory.

More recently, the following was shown by G. Lebeau and L. Robbiano [70]:

THEOREM 2.2. *System* (2.2) *is null controllable for any non-empty open set $\omega \subset \Omega$ and any $T > 0$.*

SKETCH OF THE PROOF. A slightly simplified proof of this result was given in [74]. The main ingredient is an observability estimate for the eigenfunctions of the Dirichlet-Laplace operator:

$$\begin{cases} -\Delta w_j = \lambda_j w_j & \text{in} \quad \Omega, \\ w_j = 0 & \text{on} \quad \partial\Omega. \end{cases} \tag{2.12}$$

Recall that the eigenvalues $\{\lambda_j\}$ form a nondecreasing sequence of positive numbers such that $\lambda_j \to \infty$ as $j \to \infty$ and the associated eigenfunctions $\{w_j\}$ form an orthonormal basis in $L^2(\Omega)$.

The following holds:

> *Let $\Omega \subset \mathbb{R}^N$ be a bounded smooth domain. For any open set $\omega \subset \Omega$, there exist positive constants $C_1, C_2 > 0$ such that*
>
> $$\sum_{\lambda_j \leq \mu} |a_j|^2 \leq C_1 e^{C_2\sqrt{\mu}} \int_\omega \Big| \sum_{\lambda_j \leq \mu} a_j w_j(x) \Big|^2 dx \tag{2.13}$$
>
> *whenever $\{a_j\} \in \ell^2$ and $\mu > 0$.*

This result was implicitly used in [70] and is proved in [74]. A consequence is that the observability inequality (2.11) holds for the solutions to (2.3) with initial data in

$$E_\mu = \text{span}\{\, \varphi_j : \lambda_j \leq \mu \,\},$$

the constant being of the order of $\exp\left(C\sqrt{\mu}\right)$.

This shows that the projection on E_μ of the solution of (2.3) can be controlled to zero with a control of size $\exp\left(C\sqrt{\mu}\right)$. Thus, when controlling the frequencies $\lambda_j \leq \mu$, one increases the L^2-norm of the high frequencies $\lambda_j > \mu$ by a multiplicative factor of the order of $\exp\left(C\sqrt{\mu}\right)$.

However, it was observed in [70] that any solution of the heat equation (2.2) with $v = 0$ such that the projection on E_μ of $y(\cdot, 0)$ vanishes decays in $L^2(\Omega)$ at a rate of the order of $\exp(-\mu t)$.

Consequently, if we divide the time interval $[0, T]$ in two parts $[0, T/2]$ and $[T/2, T]$, we control to zero the frequencies $\lambda_j \leq \mu$ in the interval $[0, T/2]$ and then allow the equation to evolve without control in the interval $[T/2, T]$, it follows that, at time $t = T$, the projection of the solution y over E_μ vanishes and the norm of the high frequencies does not exceed the norm of the initial data.

This argument allows to control to zero the projection over E_μ for any $\mu > 0$, but not the whole solution. To do that, an iterative argument is needed. Thus, we decompose the interval $[0, T)$ in disjoint subintervals of the form $[T_j, T_{j+1})$ for $j \in \mathbb{N}$, with a suitable choice of the sequence $\{T_j\}$. In each interval $[T_j, T_{j+1}]$, we control to zero the frequencies $\lambda_k \leq 2j$. By letting $j \to \infty$, we obtain a control $v \in L^2(\omega \times (0, T))$ such that the solution of (2.2) satisfies

$$y(x, T) \equiv 0. \tag{2.14}$$

□

Once it is known that (2.2) is null controllable, one can obtain the control with minimal L^2-norm satisfying (2.14). It suffices to minimize the functional

$$J(\varphi^0) = \frac{1}{2} \iint_{\omega \times (0,T)} |\varphi|^2 \, dx \, dt + \int_\Omega \varphi(x, 0) y^0(x) \, dx$$

over the Hilbert space

$$H = \{\, \varphi^0 : \text{ the solution } \varphi \text{ of (2.3) satisfies } \iint_{\omega \times (0,T)} |\varphi|^2 \, dx \, dt < \infty \,\}.$$

As a consequence of this theorem, we also have the null boundary controllability of the heat equation, with controls in an arbitrarily small open subset of the boundary. See [70] for more details.

Question 6. *Why does theorem 2.2 imply null boundary controllability?*

The previous controllability results also hold for linear parabolic equations with lower order terms depending on time and space.

For instance, the following system can be considered:

$$\begin{cases} y_t - \Delta y + a(x,t) y = v 1_\omega & \text{in} \quad Q, \\ y = 0 & \text{on} \quad \Sigma, \\ y(x, 0) = y^0(x) & \text{in} \quad \Omega. \end{cases} \tag{2.15}$$

Here, we assume that $a \in L^\infty(Q)$. In this case, the adjoint system is

$$\begin{cases} -\varphi_t - \Delta\varphi + a(x,t)\varphi = 0 & \text{in} \quad Q, \\ \varphi = 0 & \text{on} \quad \Sigma, \\ \varphi(x,T) = \varphi^0(x) & \text{in} \quad \Omega. \end{cases} \tag{2.16}$$

Again, the null controllability of (2.15), together with a L^2- estimate of the control, is equivalent to an observability inequality. Hence, in order to obtain a null controllability result for (2.15), what we have to do is to prove the estimate (2.11) for the solutions to (2.16).

The controllability properties of systems of this kind have been analyzed by several authors. Among them, let us mention the work of A.V. Fursikov and O.Yu. Imanuvilov (for instance, see [19, 45, 46, 47, 48, 64]; more complicate linear heat equations involving first-order terms of the form $B(x,t)\cdot\nabla y$ have recently been considered in [66]). Their approach to the controllability problem is different and more general than the previous one and relies on appropriate (global) *Carleman inequalities.*

A general global Carleman inequality is an estimate of the form

$$\iint_{\Omega\times(0,T)} \rho^{-2}\,|\varphi|^2\,dx\,dt \le C \iint_{\omega\times(0,T)} \rho^{-2}\,|\varphi|^2\,dx\,dt, \tag{2.17}$$

where $\rho = \rho(x,t)$ is continuous, strictly positive and bounded from below. For an appropriate ρ that depends on Ω, ω, T and $\|a\|_{L^\infty(Q)}$, it is possible to deduce (2.17) and, consequently, also estimates of the form

$$\iint_{\Omega\times(T/4,3T/4)} |\varphi|^2\,dx\,dt \le C \iint_{\omega\times(0,T)} |\varphi|^2\,dx\,dt. \tag{2.18}$$

This, together with the properties of the solutions of (2.16), leads to (2.11) and, therefore, implies the null controllability property for (2.15); see also [26, 43, 66] for some improved estimates.

QUESTION 7. *How can* (2.11) *be proved from* (2.18)*?*

Thus, at present we can affirm that, as in the case of the classical heat equation, (2.15) is both approximately and null controllable for any ω and any $T > 0$. Once more, null controllability implies approximate controllability for (2.15); this has been shown in [43].

An interesting question analyzed in [43] deals with explicit estimates of the *cost* in $L^2(Q)$ of the approximate, E-approximate (E is a finite-dimensional space) and null controllability of (2.15).

For instance, let us recall the results concerning the costs of approximate and null controllability. In the remainder of this Section, it will be assumed that C is a generic positive constant that only depends on Ω and ω.

Let us consider the linear state equation (2.15), where $a \in L^\infty(Q)$. For each $y^0 \in L^2(\Omega)$, $y^1 \in L^2(\Omega)$ and $\varepsilon > 0$, let us introduce the corresponding *set of admissible controls*

$$\mathcal{U}_{\rm ad}(y^0, y^1; \varepsilon) := \{\, v \in L^2(Q) :\ \text{the solution of (2.15) satisfies (2.6)}\,\} \tag{2.19}$$

and the following quantity, which measures the *cost of approximate controllability* or, more precisely, the cost of achieving (2.6):

$$\mathcal{C}(y^0, y^1; \varepsilon) := \inf_{v \in \mathcal{U}_{ad}(y^0, y^1; \varepsilon)} \|v\|_{L^2(Q)} \,. \tag{2.20}$$

Then, the question is: can we obtain "explicit" upper bounds for $\mathcal{C}(y^0, y^1; \varepsilon)$?

Taking into account that system (2.15) is linear, one can assume, without loss of generality, that $y^0 = 0$. Indeed,

$$\mathcal{C}(y^0, y^1; \varepsilon) = \mathcal{C}(0, z^1; \varepsilon) \,, \tag{2.21}$$

where $z^1 = y^1 - z(\cdot, T)$ and z is the solution of (2.15) with $v \equiv 0$.

Let us denote by $\|\cdot\|_\infty$ the usual norm in $L^\infty(Q)$. Then the following is satisfied:

THEOREM 2.3. *For any $y^1 \in H^2(\Omega) \cap H_0^1(\Omega)$, $\varepsilon > 0$, $T > 0$ and $a \in L^\infty(Q)$, one has:*

$$\mathcal{C}(0, y^1; \varepsilon) \leq \exp\left[C\left[1 + \frac{1}{T} + T\|a\|_\infty + \|a\|_\infty^{2/3} + \frac{\|a\|_\infty \|y^1\|_{L^2} + \|\Delta y^1\|_{L^2}}{\varepsilon}\right]\right] \|y^1\|_{L^2} \,. \tag{2.22}$$

Notice that (2.22) is only of interest when

$$\frac{\|\Delta y^1\|_{L^2}}{\lambda_1} > \varepsilon \,,$$

with λ_1 being the first eigenvalue of the Dirichlet Laplacian $-\Delta$. Otherwise, we would have $\|y^1\|_{L^2} \leq \varepsilon$ and then, taking $v \equiv 0$ in (2.15) for $y^0 = 0$, we would trivially obtain $y \equiv 0$ and

$$\|y(\cdot, T) - y^1\|_{L^2} \leq \varepsilon \,.$$

In other words,

$$\mathcal{C}(0, y^1; \varepsilon) = 0 \ \text{ if } \ \frac{\|\Delta y^1\|_{L^2}}{\lambda_1} \leq \varepsilon \,.$$

Furthermore, if instead of assuming $y^1 \in D(-\Delta) = H_0^1(\Omega) \cap H^2(\Omega)$ we assume that $y^1 \in D((-\Delta)^{\gamma/2})$ with $0 < \gamma \leq 2$, other estimates similar to (2.22) can be established. See [43] for the details.

For the proof of (2.22), we first have to obtain sharp bounds on the cost of controlling to zero. Recall that (2.16) is the adjoint system of (2.15). Then we have the following explicit observability estimate:

LEMMA 2.4. *For any solution of* (2.16) *and for any $a \in L^\infty(Q)$, one has*

$$\|\varphi(\cdot, 0)\|_{L^2}^2 \leq \exp\left(C\left(1 + \frac{1}{T} + T\|a\|_\infty + \|a\|_\infty^{2/3}\right)\right) \iint_{\omega \times (0,T)} |\varphi|^2 \, dx \, dt \,. \tag{2.23}$$

The proof of (2.23) relies on global Carleman inequalities as in [47], but paying special attention to the constants arising in the integrations by parts. Once (2.23) is known, (2.22) can be proved easily.

QUESTION 8. *How can* (2.22) *be proved from* (2.23)*?*

As we have already seen, (2.23) implies the null controllability of (2.15). But it also provides an estimate for the associated cost $\mathcal{C}(y^0, 0)$. More precisely, one has:

THEOREM 2.5. *For each $y^0 \in L^2(\Omega)$, the set $\mathcal{U}_{\rm ad}(y^0,0)$ is non-empty. Moreover, the associated cost $\mathcal{C}(y^0,0)$ satisifes:*

$$\mathcal{C}(y^0,0) \leq \exp\left(C\left(1+\frac{1}{T}+T\|a\|_\infty+\|a\|_\infty^{2/3}\right)\right)\|y^0\|_{L^2}\,. \tag{2.24}$$

QUESTION 9. *How can* (2.24) *be proved from* (2.23)*?*

In the particular case in which $a \equiv$ Const, (2.22) can be improved. More precisely, we can obtain a bound of the cost of approximate controllability of the order of $\exp(1/\sqrt{\varepsilon})$. Furthermore, it can be proved that this estimate is optimal in an appropriate sense; see [43] for the details.

REMARK 2.6. We can be more explicit on the way the constants C in (2.22) and (2.24) depend on Ω and ω: there exist "universal" constants $C_0 > 0$ and $m \geq 1$ such that C can be taken of the form

$$C = \exp\left(C_0\|\psi\|_{C^2}^m\right),$$

where $\psi \in C^2(\overline{\Omega})$ is any function satisfying $\psi > 0$ in Ω, $\psi = 0$ on $\partial\Omega$ and $\nabla\psi \neq 0$ in $\overline{\Omega}\setminus\omega$. All this is a consequence of the particular form that must have ρ in order to ensure (2.17). □

The results of this Section can be extended to more general equations of the form

$$\begin{cases} y_t - \Delta y + \nabla\cdot(yB(x,t)) + a(x,t)y = v1_\omega & \text{in} \quad Q,\\ y = 0 & \text{on} \quad \Sigma,\\ y(x,0) = y^0(x) & \text{in} \quad \Omega, \end{cases} \tag{2.25}$$

where $a \in L^\infty(Q)$ and $B \in L^\infty(Q;\mathbb{R}^N)$.

To do that, it is sufficient to obtain suitable observability estimates for the solutions of adjoint systems of the form

$$\begin{cases} -\varphi_t - \Delta\varphi - B(x,t)\cdot\nabla\varphi + a(x,t)\varphi = 0 & \text{in} \quad Q,\\ \varphi = 0 & \text{on} \quad \Sigma,\\ \varphi(x,T) = \varphi^0(x) & \text{in} \quad \Omega. \end{cases} \tag{2.26}$$

More precisely, we can deduce that

$$\|\varphi(\cdot,0)\|_{L^2}^2 \leq \exp\left(C\left(1+\frac{1}{T}+T\|a\|_\infty+\|a\|_\infty^{2/3}+T^2\|B\|_\infty^2\right)\right)\iint_{\omega\times(0,T)}|\varphi|^2\,dx\,dt \tag{2.27}$$

for any solution of (2.26) and for all $a \in L^\infty(Q)$, $B \in L^\infty(Q;\mathbb{R}^N)$. Then, arguments similar to those above lead to an estimate of the cost of approximate controllability.

The situation is more complicate when the state equation is of the form

$$\begin{cases} y_t - \Delta y + B(x,t)\cdot\nabla y + a(x,t)y = 0 & \text{in} \quad Q\\ y = 0 & \text{on} \quad \Sigma\\ y(x,0) = y^0(x) & \text{in} \quad \Omega\,. \end{cases} \tag{2.28}$$

Indeed, if B is only assumed to be in $L^\infty(Q;\mathbb{R}^N)$, the adjoint systems take the form

$$\begin{cases} -\varphi_t - \Delta\varphi - \nabla\cdot(\varphi B(x,t)) + a(x,t)\varphi = 0 & \text{in} \quad Q\\ \varphi = 0 & \text{on} \quad \Sigma,\\ \varphi(x,T) = \varphi^0(x) & \text{in} \quad \Omega \end{cases} \tag{2.29}$$

and, therefore, the usual Carleman inequalities do not suffice. These questions have been considered and solved in [26], using some ideas from [66]. We omit the details.

To end this Section, let us make some comments on the convergence rate of algorithms devised to construct "good" controls.

It is rather natural to build approximate controls by *penalizing* a suitable optimal control problem. This has been done systematically, for instance, in the works by R. Glowinski [55] and R. Glowinski et al. [56]. This method has also been used to prove the approximate controllability for some linear and semilinear heat equations in [79] and [33], respectively.

Let us briefly describe the procedure in the case of the linear heat equation. First of all, without loss of generality, we set $y^0 = 0$. Given $y^1 \in L^2(\Omega)$, we introduce the functional F_k, with

$$F_k(v) = \frac{1}{2}\iint_{\omega\times(0,T)} |v|^2\,dx\,dt + \frac{k}{2}\|y(\cdot,T) - y^1\|^2_{L^2} \quad \forall v \in L^2(\omega\times(0,T)), \quad (2.30)$$

where y is the solution of (2.2) with $y^0 = 0$.

It was proved in [79] that F_k has a unique minimizer $v_k \in L^2(\omega\times(0,T))$ for all $k > 0$ and that the associated states y_k satisfy

$$y_k(\cdot,T) \to y^1 \text{ in } L^2(\Omega) \text{ as } k \to \infty. \quad (2.31)$$

In view of (2.31), in order to compute a control satisfying (2.6), it suffices to take $v = v_k$ for a sufficiently large $k = k(\varepsilon)$.

Using the results above, it is easy to get explicit estimates of the rate of convergence in (2.31) (we refer to [43] for the details of the proof):

THEOREM 2.7. *Under the previous conditions, there exists $C > 0$ such that*

$$\|y_k(\cdot,T) - y^1\| \le \frac{C}{\log k} \quad (2.32)$$

and

$$\|v_k\|_{L^2(Q)} \le \frac{C\sqrt{k}}{\log k} \quad (2.33)$$

as $k \to \infty$.

QUESTION 10. *How can* (2.32) *and* (2.33) *be proved?*

Notice that (2.32) provides logarithmic (and therefore very slow) convergence rates. This fact agrees with the extremely high cost (exponentially depending on $1/\varepsilon$) of approximate controllability.

The methods of this Section can also be applied to obtain estimates on the cost of controllability when the control acts on a non-empty open subset of $\partial\Omega$.

3. Basic results for the linear wave equation

Let us now consider the linear controlled wave equation

$$\begin{cases} y_{tt} - \Delta y = v1_\omega & \text{in} \quad Q, \\ y = 0 & \text{on} \quad \Sigma, \\ y(x,0) = y^0(x), \quad y_t(x,0) = y^1(x) & \text{in} \quad \Omega. \end{cases} \quad (2.34)$$

In (2.34), we have used the same notation as in Section 2. Again, $y = y(x,t)$ is the state and $v = v(x,t)$ is the control. For any $(y^0, y^1) \in H_0^1(\Omega) \times L^2(\Omega)$ and any $v \in L^2(\omega \times (0,T))$, (2.34) possesses exactly one solution $y \in C^0([0,T]; H_0^1(\Omega)) \cap C^1([0,T]; L^2(\Omega))$.

Roughly speaking, the *controllability* problem for (2.34) consists on *describing the set of reachable final states*

$$R(T; y^0, y^1) := \{\, (y(\cdot,T), y_t(\cdot,T)) : v \in L^2(\omega \times (0,T)) \,\}.$$

As in the case of the heat equation, we may distinguish several degrees of controllability:

(a) It is said that (2.34) is *approximately controllable* at time T if $R(T; y^0, y^1)$ is dense in $H_0^1(\Omega) \times L^2(\Omega)$ for every $(y^0, y^1) \in H_0^1(\Omega) \times L^2(\Omega)$.
(b) It is said that (2.34) is *exactly controllable* at time T if $R(T; y^0, y^1) = H_0^1(\Omega) \times L^2(\Omega)$ for every $(y^0, y^1) \in H_0^1(\Omega) \times L^2(\Omega)$.
(c) Finally, it is said that (2.34) is *null controllable* at time T if $(0,0) \in R(T; (y^0, y^1))$ for every $(y^0, y^1) \in H_0^1(\Omega) \times L^2(\Omega)$.

The previous controllability properties can also be formulated in other function spaces in which the wave equation is well posed.

Since we are now dealing with solutions to the wave equation, for any of these properties to hold, the control time T has to be sufficiently large, due to the finite speed of propagation. On the other hand, since (2.34) is linear and reversible in time, null and exact controllability are equivalent notions. As we have seen, the situation is completely different in the case of the heat equation.

QUESTION 11. *Why do we need large T for any kind of controllability of the wave equation? Why are null controllability and exact controllability equivalent properties?*

Clearly, every exactly controllable system is approximately controllable too. However, (2.34) may be approximately but not exactly controllable.

Let us now briefly discuss the *approximate controllability problem* for the wave equation.

Again, it is easy to see that approximate controllability is equivalent to a specific *unique continuation property*. More precisely, let us introduce the *adjoint system*

$$\begin{cases} \varphi_{tt} - \Delta\varphi = 0 & \text{in} \quad Q, \\ \varphi = 0 & \text{on} \quad \Sigma, \\ \varphi(x,T) = \varphi^0(x), \quad \varphi_t(x,T) = \varphi^1(x) & \text{in} \quad \Omega. \end{cases} \tag{2.35}$$

Then, (2.34) is approximately controllable with controls that depend continously on the data if and only if the following unique continuation property is fulfilled:

If φ solves (2.35) and $\varphi = 0$ in $\omega \times (0,T)$, then necessarily $\varphi \equiv 0$, i.e. $(\varphi^0, \varphi^1) = (0,0)$.

In fact, that the previous uniqueness property implies approximate controllability can be checked at least in two ways:

(a) Applying the Hahn-Banach theorem; see [78].
(b) Using the variational approach developed in [80].

Both approaches have been considered in the context of the heat equation. They will not be revisited here, for reasons of space.

QUESTION 12. *Which are the detailed arguments?*

In view of a well known consequence of *Holmgren's uniqueness theorem*, it can be easily seen that, for any non-empty open set $\omega \subset \Omega$, the previous unique continuation property holds if T is large enough (depending on Ω and ω). We refer to Chapter 1 in [78] and [18] for a discussion on this problem.

Therefore, the following result holds:

THEOREM 2.8. *Let $\omega \subset \Omega$ be a non-empty open set. There exists $T_1 > 0$, only depending on Ω and ω, such that, for any $T > T_1$, the linear system* (2.34) *is approximately controllable at time T.*

When approximate controllability holds, the following (apparently stronger) property is also satisfied:

> *Let E be a finite dimensional subspace of $H_0^1(\Omega) \times L^2(\Omega)$ and let us denote by $\pi_E : H_0^1(\Omega) \times L^2(\Omega) \mapsto E$ the corresponding orthogonal projector. Then, for any $(y^0, y^1), (z^0, z^1) \in H_0^1(\Omega) \times L^2(\Omega)$ and any $\varepsilon > 0$, there exists $v \in L^2(\omega \times (0,T))$ such that the solution of* (2.34) *satisfies*

$$\|(y(\cdot,T) - z^0, y_t(\cdot,T) - z^1)\|_{H_0^1 \times L^2} \leq \varepsilon,\ \pi_E(y(\cdot,T), y_t(\cdot,T)) = \pi_E(z^0, z^1). \quad (2.36)$$

In other words, if $T > 0$ is large enough to ensure approximate controllability, for any finite dimensional subspace $E \subset H_0^1(\Omega) \times L^2(\Omega)$ we also have E-approximate controllability.

QUESTION 13. *Why does approximate controllability imply E-approximate controllability for any finite-dimensional space $E \subset H_0^1(\Omega) \times L^2(\Omega)$?*

The previous results hold for wave equations with analytic coefficients too. However, the problem is not completely solved in the frame of the wave equation with lower order potentials $a \in L^\infty(Q)$ of the form

$$y_{tt} - \Delta y + a(x,t)y = v1_\omega \text{ in } Q.$$

We refer to [3, 98, 105] for some deep results in this direction.

Let us now consider the *exact controllability problem.*

It was shown by J.-L. Lions in [78] using the so called H.U.M. that exact controllability holds (with controls $v \in L^2(\omega \times (0,T))$) if and only if

$$\|(\varphi(\cdot,0), \varphi_t(\cdot,0))\|^2_{L^2 \times H^{-1}} \leq C \iint_{\omega \times (0,T)} |\varphi|^2 \, dx\, dt \quad (2.37)$$

for any solution φ to the adjoint system (2.35).

This is an observability inequality, playing in this context the role played by (2.11) in Section 2. It provides an estimate of the *total energy* of the solution (2.35) by means of a measurement in the control region $\omega \times (0,T)$.

Notice that the energy

$$E(t) = \|(\varphi(\cdot,t), \varphi_t(\cdot,t))\|^2_{L^2 \times H^{-1}}$$

of any solution to (2.35) is conserved. Thus, (2.37) is equivalent to the so called *inverse inequality*

$$\|(\varphi^0,\varphi^1)\|^2_{L^2\times H^{-1}} \le C\iint_{\omega\times(0,T)} |\varphi|^2\,dx\,dt. \tag{2.38}$$

QUESTION 14. *Why is* (2.37) *equivalent to the exact controlability of* (2.34)*?*

When (2.37) holds, one can minimize the functional W, with

$$W(\varphi^0,\varphi^1) = \frac{1}{2}\iint_{\omega\times(0,T)} |\varphi|^2\,dx\,dt + \langle(\varphi(\cdot,0),\varphi_t(\cdot,0)),(y^1,-y^0)\rangle, \tag{2.39}$$

in the space $L^2(\Omega)\times H^{-1}(\Omega)$. Indeed, the following result is easy to prove:

LEMMA 2.9. *Assume that* (2.37) *holds and* $(y^0,y^1)\in H_0^1(\Omega)\times L^2(\Omega)$ *is given. Then* W *possesses a unique minimizer* $(\hat\varphi^0,\hat\varphi^1)$ *in* $L^2(\Omega)\times H^{-1}(\Omega)$. *The control* $v=\hat\varphi 1_\omega$ *, where* $\hat\varphi$ *is the solution to* (2.35) *corresponding to the final data* $(\hat\varphi^0,\hat\varphi^1)$*, is such that the associated state satisfies*

$$y(x,T)\equiv y_t(x,T)\equiv 0. \tag{2.40}$$

QUESTION 15. *How can we prove lemma 2.9?*

As a consequence, the exact controllability problem is reduced to the analysis of the inequality (2.38). Let us now indicate what is known about this inequality:

- Using multipliers techniques in the spirit of C. Morawetz, L.F. Ho proved in [63] that, for any subset of Γ of the form

  $$\Gamma(x^0) = \{\, x\in\Gamma : (x-x^0)\cdot n(x) > 0\,\}$$

  with $x^0\in\mathbb{R}^N$ ($n(x)$ is the outward unit normal to Ω at $x\in\Gamma$) and any sufficiently large T, the following boundary observability inequality holds:

  $$\|(\varphi(\cdot,0),\varphi_t(\cdot,0))\|^2_{H_0^1\times L^2} \le C\iint_{\Gamma(x^0)\times(0,T)} \left|\frac{\partial\varphi}{\partial n}\right|^2 d\Gamma\,dt \tag{2.41}$$

  for every couple $(\varphi^0,\varphi^1)\in H_0^1(\Omega)\times L^2(\Omega)$.

  This is the observability inequality that is required to solve a boundary controllability problem similar to the one we are considering here.

  Later, (2.41) was proved in [77, 78] for any

  $$T > T(x^0) = 2\|x-x^0\|_{L^\infty}. \tag{2.42}$$

  In fact, this is the optimal observability time that one may obtain by means of multipliers.

  Proceeding as in Vol. 1 of [78], one can easily prove that (2.41) implies (2.37) when ω is a neighborhood of $\Gamma(x^0)$ in Ω and $T > T(x^0)$. Consequently, the following result holds:

  THEOREM 2.10. *Assume that* $x^0\in\mathbb{R}^N$*,* ω *is a neighborhood of* $\Gamma(x^0)$ *in* Ω *and* (2.42) *is satisfied. Then* (2.34) *is exactly controllable at time* T*.*

  More recently, A. Osses has introduced in [89] a new multiplier which is essentially a rotation of the one in [78]. In this way, he proved that the class of subsets of the boundary for which observability holds is considerably larger.

- C. Bardos, G. Lebeau and J. Rauch [9] proved that, in the class of C^∞ domains, the observability inequality (2.37) holds if and only if the couple (ω, T) satisfies the following *geometric control condition* in Ω:

  Every ray of geometric optics that begins to propagate in Ω at time $t = 0$ and is reflected on its boundary Γ enters ω at a time $t < T$.

  This result was proved with *microlocal analysis techniques*. Recently, the microlocal approach has been greatly simplified by N. Burq [15] by using the microlocal defect measures introduced by P. Gerard [50]. In [15], the geometric control condition was shown to be sufficient for exact controllability for domains Ω of class C^3 and equations with C^2 coefficients.

  Therefore, one has:

  THEOREM 2.11. *Let Ω be of class C^3, let $\omega \subset \Omega$ be a non-empty open set and let us assume that the couple (ω, T) satisfies the previous geometric condition. Then* (2.34) *is exactly controllable at time T.*

- Let us finally indicate that other methods have also been developed to address controllability problems for wave equations: Moment problems, the use of fundamental solutions, controllability via stabilization, Carleman estimates, etc. We will not present them here; for more details, we refer to the survey paper by D.L. Russell [99] and also to the works of J.-P. Puel [97] and X. Zhang [107].

As in the case of the heat equation, it is also natural to study the cost of the approximate controllability of the wave equation or, in other words, the minimal size of a control needed to reach the ε-neighborhood of a final state. The same can be said in the context of null controllability. These questions were considered by G. Lebeau in [69], with techniques which are not the same we used in Section 2.

CHAPTER 3

Controllability results for other time-dependent PDEs

This Lecture is devoted to present some controllability results for several time-dependent, mainly nonlinear, parabolic systems of PDEs. First, we will revisit the heat equation and some extensions. Then, some controllability results will be presented for systems governed by stochastic PDEs. Finally, we will consider several nonlinear systems from fluid mechanics: Burgers, Navier-Stokes, Boussinesq, micropolar, etc. Along this Lecture, several open questions will be stated.

1. Introduction. Recalling general ideas

Let us first recall some general ideas, many of them already mentioned in the previous Lecture.

Suppose that we are considering an abstract *state equation* of the form

$$\begin{cases} y_t - A(y) = Bv, \quad t \in (0,T), \\ y(0) = y^0, \end{cases} \tag{3.1}$$

which governs the behavior of a physical system. It is assumed that

- $y : [0,T] \mapsto H$ is the *state*, i. e. the variable that serves to identify the physical properties of the system,
- $v : [0,T] \mapsto U$ is the *control*, i.e. the variable we can choose (for simplicity, we assume that U and H are Hilbert spaces),
- $A : D(A) \subset H \mapsto H$ is a (generally nonlinear) operator with $A(0) = 0$, $B \in \mathcal{L}(U;H)$ and $y^0 \in H$.

Suppose that (3.1) is well-posed in the sense that, for each $y^0 \in H$ and each $v \in L^2(0,T;U)$, it possesses exactly one solution. Then the *null controllability* problem for (3.1) can be stated as follows:

> *For each $y^0 \in H$, find $v \in L^2(0,T;U)$ such that the corresponding solution of* (3.1) *satisfies $y(T) = 0$.*

More generally, the *exact controllability to the trajectories* problem for (3.1) is the following:

> *For each free trajectory $\overline{y} : [0,T] \mapsto H$ and each $y^0 \in H$, find $v \in L^2(0,T;U)$ such that the corresponding solution of* (3.1) *satisfies $y(T) = \overline{y}(T)$.*

Here, by a *free* or *uncontrolled* trajectory we mean any (sufficiently regular) function $\overline{y} : [0,T] \mapsto H$ satisfying $\overline{y}(t) \in D(A)$ for all t and

$$\overline{y}_t - A(\overline{y}) = 0, \quad t \in (0,T).$$

Notice that the exact controllability to the trajectories is a very useful property from the viewpoint of applications: if we can find a control such that $y(T) = \overline{y}(T)$, then after time T we can switch off the control and let the system follow the "ideal" trajectory $\overline{y}$.

For each system of the form (3.1), these problems lead to several interesting questions. Among them, let us indicate the following:

- First, are there controls v such that $y(T) = 0$ and/or $y(T) = \overline{y}(T)$?
- Then, if this is the case, which is the *cost* we have to pay to drive y to zero and/or $\overline{y}(T)$? In other words, which is the minimal norm of a control $v \in L^2(0,T;U)$ satisfying these properties?
- How can these controls be computed?

As indicated in Lecture 2, the controllability of differential systems is a very relevant area of research and has been the subject of a lot of work the last years. In particular, in the context of PDEs, the null controllability problem was first analyzed in [64, 70, 77, 78, 99, 100]. For semilinear systems of this kind, the first contributions have been given in [30, 47, 68, 109].

In this Lecture, I will consider several linear and nonlinear parabolic PDEs. First, we will recall the results satisfied by the classical heat equation in a bounded N-dimensional domain, complemented with appropriate initial and boundary-value conditions. Secondly, we will deal with similar stochastic PDEs. We will then consider the viscous Burgers equation. We will see that, for this PDE, the null controllability problem (with distributed and locally supported control) is well understood.[1] We will also consider the Navier-Stokes and Boussinesq equations and some other systems from mechanics.

2. The heat equation. Observability and Carleman estimates

Let us consider the following control system for the heat equation:

$$\begin{cases} y_t - \Delta y = v1_\omega, & (x,t) \in \Omega \times (0,T), \\ y(x,t) = 0, & (x,t) \in \partial\Omega \times (0,T), \\ y(x,0) = y^0(x), & x \in \Omega. \end{cases} \tag{3.2}$$

Here, we conserve the notation of Lecture 2. In particular, $\Omega \subset \mathbb{R}^N$ is a nonempty regular and bounded domain, $\omega \subset\subset \Omega$ is a (small) nonempty open subset (1_ω is the characteristic function of ω) and $y^0 \in L^2(\Omega)$.

It is well known that, for every $y^0 \in L^2(\Omega)$ and every $v \in L^2(\omega \times (0,T))$, there exists a unique solution y to (3.2), with $y \in L^2(0,T;H_0^1(\Omega)) \cap C^0([0,T];L^2(\Omega))$.

In view of the results in Lecture 2, (3.2) is approximately, E-approximately and null controllable.

Also, if we introduce for each $\varphi^0 \in L^2(\Omega)$ the adjoint system

$$\begin{cases} -\varphi_t - \Delta\varphi = 0, & (x,t) \in \Omega \times (0,T), \\ \varphi(x,t) = 0, & (x,t) \in \partial\Omega \times (0,T), \\ \varphi(x,T) = \varphi^0(x), & x \in \Omega, \end{cases} \tag{3.3}$$

[1] More precisely, let us denote by $T^*(r)$ the minimal time needed to drive any initial state with L^2 norm $\leq r$ to zero. Then we will show that $T^*(r) > 0$, with explicit sharp estimates from above and from below.

we know that the null controllability of (3.2) is equivalent to the *observability* of (3.3), that is, to the following estimate:

$$\|\varphi(\cdot,0)\|_{L^2}^2 \leq C \iint_{\omega\times(0,T)} |\varphi|^2 \, dx\, dt \quad \forall \varphi^0 \in L^2(\Omega) \tag{3.4}$$

(where C only depends on Ω, ω and T).

We have already seen that the estimates (3.4) are implied by the so called global Carleman inequalities. These have been introduced in the context of the controllability of PDEs by Fursikov and Imanuvilov; see [47, 64]. When they are applied to the solutions to the adjoint system (3.3), they take the form

$$\iint_{\Omega\times(0,T)} \rho^{-2} \, |\varphi|^2 \, dx\, dt \leq K \iint_{\omega\times(0,T)} \rho^{-2} \, |\varphi|^2 \, dx\, dt \quad \forall \varphi^0 \in L^2(\Omega), \tag{3.5}$$

where $\rho = \rho(x,t)$ is an appropriate weight depending on Ω, ω and T and the constant K only depends on Ω and ω.[2] Combining (3.5) and the dissipativity of the backwards heat equation (3.3), it is not difficult to deduce (3.4) for some C only depending on Ω, ω and T.

Since (3.2) is linear, null controllability is equivalent in this case to *exact controllability to the trajectories.* This means that, for any uncontrolled solution $\overline{y}$ and any $y^0 \in L^2(\Omega)$, there exists $v \in L^2(\omega \times (0,T))$ such that the associated state y satisfies

$$y(x,T) = \overline{y}(x,T) \ \text{ in } \ \Omega.$$

Remark 3.1. Notice that the null controllability of (3.2) holds for *any* ω and T. This is a consequence of the fact that, in a parabolic equation, the transmission of information is instantaneous. Recall that this was not the case for the wave equation. Again, this is not the case for the transport equation. Thus, let us consider the control system

$$\begin{cases} y_t + y_x = v 1_\omega, & (x,t) \in (0,L)\times(0,T), \\ y(0,t) = 0, & t \in (0,T), \\ y(x,0) = y^0(x), & x \in (0,L), \end{cases} \tag{3.6}$$

with $\omega = (a,b) \subset\subset (0,L)$. Then, if $0 < T < a$, null controllability does not hold, since the solution always satisfies

$$y(x,T) = y^0(x-T) \quad \forall x \in (T,a),$$

independently of the choice of v; see [23] for more details and similar results concerning other control systems for the wave, Schrödinger and Korteweg-De Vries equations. □

There are many generalizations and variants of the previous argument that provide the null controllability of other similar linear (parabolic) state equations:

- Time-space dependent (and sufficiently regular) coefficients can appear in the equation, other boundary conditions can be used, boundary control (instead of distributed control) can be imposed, etc.; see [47]; see also [36] for a review of related results.

[2] In order to prove (3.5), we have to use a weight ρ that blows up as $t \to 0$ and also as $t \to T$, for instance exponentially.

- The null controllability of Stokes-like systems of the form

$$y_t - \Delta y + (a \cdot \nabla)y + (y \cdot \nabla)b + \nabla p = v1_\omega, \quad \nabla \cdot y = 0, \tag{3.7}$$

  where a and b are regular enough, can also be analyzed with these techniques. See for instance [39]; see also [29] for other controllability properties. We will come back in Section 5 to systems of this kind.
- Other linear parabolic (non-scalar) systems can also be considered, etc.

However, there are several interesting problems related to the controllability of linear parabolic systems that remain open. Let us mention some of them.

First, let us consider the controlled system

$$\begin{cases} y_t - \nabla \cdot (a(x)\nabla y) = v1_\omega, & (x,t) \in \Omega \times (0,T), \\ y(x,t) = 0, & (x,t) \in \partial\Omega \times (0,T), \\ y(x,0) = y^0(x), & x \in \Omega, \end{cases} \tag{3.8}$$

where y^0 and v are as before and the coefficient a is assumed to satisfy

$$a \in L^\infty(\Omega), \quad 0 < a_0 \leq a(x) \leq a_1 < +\infty \quad \text{a.e.} \tag{3.9}$$

It is natural to consider the null controllability problem for (3.8). Of course, this is equivalent to the observability of the associated adjoint system

$$\begin{cases} -\varphi_t - \nabla \cdot (a(x)\nabla\varphi) = 0, & (x,t) \in \Omega \times (0,T), \\ \varphi(x,t) = 0, & (x,t) \in \partial\Omega \times (0,T), \\ y\varphi(x,T) = \varphi^1(x), & x \in \Omega, \end{cases} \tag{3.10}$$

that is to say, to the fact that an inequality like (3.4) holds for the solutions to (3.10).

To our knowledge, it is at present unknown whether (3.8) is null controllable. In fact, it is also unknown whether approximate controllability holds.

Recently, some partial results have been obtained in this context.

Thus, when $N = 1$, the null controllability of (3.8) has been established in [2] for general a satisfying (3.9). The techniques in the proof rely on the theory of quasi-conformal complex mappings and can be applied only to the one-dimensional case, with a independent of t. Furthermore, they only serve to apply directly the Lebeau-Robbiano method (recall the proof of theorem 2.2 in Lecture 2), that is, they do not lead to a Carleman estimate of the form (3.5).

When $N \geq 2$, it is known that (3.8) is null controllable under the following assumption

$$\exists \textit{ smooth open set } \Omega_0 \subset\subset \Omega \textit{ such that } a \textit{ is } C^1 \textit{ in } \overline{\Omega_0} \textit{ and } \overline{\Omega \setminus \Omega_0}. \tag{3.11}$$

This has been proved in [73]. A slight improvement has been performed in [13], where Ω_0 is allowed to touch the boundary of Ω. Again, the proofs use that a is independent of t in an essential way and do not clarify whether (3.5) holds.

In fact, it is an open question whether a Carleman estimate like (3.5) holds for the solutions to (3.10) even if $N = 1$ or (3.11) holds.

In order to have (3.5), we apparently need more regularity for a; see [12] for a proof when $N = 1$, a satisfies (3.9) and

$$a \in BV(\Omega); \tag{3.12}$$

see also [27] for a proof when $N \geq 2$, a is piecewise C^1 and satisfies (3.9) and some additional "sign" conditions.

At present, the following questions are open:

- Is (3.8) is null controllable when $N \geq 2$ and a satisfies (3.9) and (3.12)? Is (3.5) satisfied in this case?
- Is (3.5) satisfied when $N = 1$ and a only satisfies (3.9)?

QUESTION 1. *Assume that $N = 1$ and a is piecewise constant and satisfies* (3.9). *Is* (3.8) *approximately controllable?*

A similar question can be asked when $N \geq 2$. Which is the rigorous question and which is the answer?

Let us now consider the non-scalar system

$$\begin{cases} y_t - D\Delta y = My + Bv1_\omega, & (x,t) \in \Omega \times (0,T), \\ y(x,t) = 0, & (x,t) \in \partial\Omega \times (0,T), \\ y(x,0) = y^0(x), & x \in \Omega, \end{cases} \tag{3.13}$$

where $y = (y_1, \dots, y_n)$ is the state, $v = (v_1, \dots, v_m)$ is the control and D, M and B are constant matrices, with $D, M \in \mathcal{L}(\mathbb{R}^n; \mathbb{R}^n)$ and $B \in \mathcal{L}(\mathbb{R}^m; \mathbb{R}^n)$. It is assumed that $n \geq 2$ and D is definite positive, that is,

$$D\xi \cdot \xi \geq d_0|\xi|^2 \quad \forall \xi \in \mathbb{R}^n, \quad d_0 > 0. \tag{3.14}$$

When D is diagonal (or similar to a diagonal matrix), the null controllability problem for (3.13) is well understood. In view of the results in [4], (3.13) is null controllable if and only if

$$\text{rank } [(-\lambda_i D + M); B] = n \quad \forall i \geq 1, \tag{3.15}$$

where the λ_i are the eigenvalues of the Dirichlet-Laplace operator and, for any matrix $H \in \mathcal{L}(\mathbb{R}^n; \mathbb{R}^n)$, $[H; B]$ stands for the $n \times nm$ matrix

$$[H; B] := [B|HB| \cdots |H^{n-1}B].$$

Therefore, it is natural to search for (algebraic) conditions on D, M and B that ensure the null controllability of (3.13) in the general case. But, to our knowledge, this is unknown.

The results in [4] have been extended recently to the case of any D having no eigenvalue of geometric multiplicity > 4; see [35].

QUESTION 2. *Under which conditions the system*

$$\begin{cases} y_t - D\Delta y = M(x,t)y + Bv1_\omega, & (x,t) \in \Omega \times (0,T), \\ y(x,t) = 0, & (x,t) \in \partial\Omega \times (0,T), \\ y(x,0) = y^0(x), & x \in \Omega, \end{cases} \tag{3.16}$$

where D is a diagonal matrix satisfying (3.14), *$M \in L^\infty(Q; \mathcal{L}(\mathbb{R}^n; \mathbb{R}^n))$ and $B \in \mathcal{L}(\mathbb{R}^m; \mathbb{R}^n)$, is null controllable?*

REMARK 3.2. As we have said, global Carleman estimates are the main tool we can use to establish the observability property (3.4). These open questions can be viewed, at least in part, as a confirmation of the limitations of Carleman estimates: first, they need sufficiently regular coefficients; then, they are actually well-suited only for *scalar* equations. □

3. Some remarks on the controllability of stochastic PDEs

In this Section, we deal briefly with a system governed by a linear stochastic partial differential equation:

$$\begin{cases} y_t - \Delta y = v1_\omega + B(t)\,\dot{w}_t & \text{in} \quad Q, \\ y = 0 & \text{on} \quad \Sigma, \\ y(x,0) = y^0(x) & \text{in} \quad \Omega. \end{cases} \tag{3.17}$$

Here, v is again the control and $\dot{w}_t$ is a *Gaussian random field* (white noise in time). For instance, it can be regarded as the distributional time derivative of a Wiener process w_t . The equations are required to be satisfied $P-$a.e., i.e. P-almost surely, in a given probability space $\{\Lambda, \mathcal{F}, P\}$.

In the sequel, we are going to see that, for general y^0, y^1 and $B = B(t)$, one can obtain final states $y(\cdot, T)$ arbitrarily close to y^1 in quadratic mean by choosing v appropriately (an approximate controllability result). We will also see that, if B is not random and in some sense small, then one can also choose v such that $y(\cdot, T) = 0$ (amnull controllability result).

3.1. Some basic results from probability calculus. In order to present the results without too much ambiguity, we will first recall some basic definitions and results.

Thus, assume that a *complete probability space* $\{\Lambda, \mathcal{F}, P\}$ is given. If X is a Banach space and $f \in L^1(\Lambda, \mathcal{F}; X)$, we will denote by Ef the expectation of f:

$$Ef = \int_\Lambda f(\lambda)\, dP(\lambda).$$

Assume that a separable Hilbert space K and a *Wiener process* w_t on $\{\Lambda, \mathcal{F}, P\}$ with values in K are given. This means that

$$w_t = \sum_{k=1}^{\infty} \beta_t^k e_k \quad \forall t \geq 0,$$

where $\{e_k\}$ is an orthonormal basis in K and the β_t^k are mutually independent *real Wiener processes* satisfying

$$E|\beta_t^k|^2 = \mu_k^2 t, \qquad \sum_{k=1}^{\infty} \mu_k^2 < +\infty. \tag{3.18}$$

A normalized real Wiener process β_t is a measurable function $(\lambda, t) \mapsto \beta_t(\lambda)$ which is defined $P-$ a.s. in Λ for all $t \in \mathbb{R}_+$ and satisfies the following:

(a) $\beta_0 = 0$,

(b) For each t, β_t is *normally distributed,* with mean 0 and variance t, i.e.

$$E\beta_t = 0, \quad E|\beta_t|^2 = t.$$

(c) $E(\beta_t\, \beta_s) = \sqrt{t}\sqrt{s}$ for all $t, s \geq 0$.

For other equivalent definitions and basic properties of real Wiener processes, see [6]. Recall that, in particular, the real processes β_t^k and the K-valued process w_t have Hölder-continuous sample paths $t \mapsto \beta_t^k(\lambda)$ and $t \mapsto w_t(\lambda)$.

In the sequel, we put

$$\mathcal{F}_t := \sigma(\, w_s,\ 0 \le s \le t\,)$$

($\mathcal{F}_t$ is the σ-algebra spanned by w_s for $0 \le s \le t$, completed with the negligeable sets in $\mathcal{F}$). Obviously, $\{\mathcal{F}_t\}$ is an increasing family of sub σ-algebras of $\mathcal{F}$ and, among other things, one has:

$$\mathcal{F}_t = \sigma\left(\bigcup_{s<t} \mathcal{F}_s\right) \quad \forall t > 0. \tag{3.19}$$

Let H be a Hilbert space. For any $f \in L^1(\Lambda, \mathcal{F}; H)$, we denote by $E[f|\mathcal{F}_t]$ the *conditional expectation* of f with respect to $\mathcal{F}_t$, i.e. the unique element in $L^1(\Lambda, \mathcal{F}_t; H)$ such that

$$\int_A E[f|\mathcal{F}_t]\, dP = \int_A f\, dP \quad \forall A \in \mathcal{F}_t\,.$$

The existence and uniqueness of $E[f|\mathcal{F}_t]$ is implied by the celebrated Radon-Nykodim theorem. For the main properties of the conditional expectation, see for instance [91]. In particular, recall that, if $f \in L^2(\Lambda, \mathcal{F}; H)$, then $E[f|\mathcal{F}_t] \in L^2(\Lambda, \mathcal{F}_t; H)$ and icoincides with the orthogonal projection of f in $L^2(\Lambda, \mathcal{F}_t; H)$.

Let X be a Banach space. We denote by $I^2(0,T;X)$ the space formed by all stochastic processes $\Phi \in L^2(\Lambda \times (0,T), dP \otimes dt; X)$ which are $\mathcal{F}_t$-adapted a.e. in $(0,T)$, i.e. such that

$$\lambda \mapsto \Phi(\lambda, t) \text{ is } \mathcal{F}_t\text{-measurable for almost all } t \in (0,T)$$

In the case $X = \mathcal{L}(K;H)$, measurability will be understood in the *strong* sense, i.e. the measurability of $\lambda \mapsto \Phi(\lambda,t)w$ for each $w \in K$. Then, $I^2(0,T;X)$ is a closed subspace of $L^2(\Lambda \times (0,T), dP \otimes dt; X)$.

Recall that, for any $b \in I^2(0,T;\mathbb{R})$, any real-valued Wiener process β_t and any fixed $t \in [0,T]$, we can introduce a random variable $I_t(f) : \Lambda \mapsto \mathbb{R}$ known as the Ito stochastic integral in $[0,t]$:

$$I_t(f) = \int_0^t f(s)\, d\beta_s\,.$$

The stochastic process $(\lambda, t) \mapsto I_t(f)(\lambda)$ again belongs to $I^2(0,T;\mathbb{R})$ and, among other properties, satisfies the following:

$$E\int_0^t f(s)\, d\beta_s = 0$$

and

$$E\left|\int_0^t f(s)\, d\beta_s\right|^2 = \int_0^t E|f(s)|^2\, ds$$

for all $t \in [0,T]$.

Now, assume that a stochastic process B is given, with

$$B \in I^2(0,T;\mathcal{L}(K;H)) \tag{3.20}$$

(H is a Hilbert space). Then the stochastic integral of B with respect to w_t is defined by the formula

$$\int_0^t B(s)\,dw_s = \sum_{k=1}^{\infty} \int_0^t B(s)e_k\,d\beta_s^k \qquad \forall t \in [0,T].$$

Here, the convergence of the series is understood in the sense of $L^2(\Lambda, \mathcal{F}_t; H)$. The stochastic integrals in the right hand side are defined by the equalities

$$(\int_0^t B(s)e_k\,d\beta_s^k, h) = \int_0^t (B(s)e_k, h)\,d\beta_s^k \qquad \forall h \in H,$$

where the latter are usual *Ito stochastic integrals* with respect to the real-valued processes β_t^k; see [6] for more details.

3.2. The controllability results. In the remainder of this Section, H and V will denote the Hilbert spaces $L^2(\Omega)$ and $H_0^1(\Omega)$, respectively.

Assume we are given an arbitrary but fixed initial state

$$y^0 \in H, \tag{3.21}$$

a Wiener process w_t with values in the separable Hilbert space K and a stochastic process $B \in I^2(0,T;\mathcal{L}(K;H))$. Let $A = -\Delta$ be the usual Laplace-Dirichlet operator in Ω, with domain $D(A) = H_0^1(\Omega) \cap H^2(\Omega)$. For each control $v \in I^2(0,T;H)$, there exists exactly one solution y to the state system

$$\begin{cases} y \in I^2(0,T;V) \cap L^2(\Lambda; C^0([0,T];H)), \\ y(\cdot,t) = y^0 + \displaystyle\int_0^t \{-Ay(\cdot,s) + 1_\omega v(\cdot,s)\}\,ds + \int_0^t B(s)\,dw_s \quad \forall t \in [0,T]. \end{cases} \tag{3.22}$$

In (0.1), the equalities have to be understood $P-$a.s. in V'.

Notice that we choose $\mathcal{F}_t$-adapted controls to govern the state system. This is a natural assumption from the stochastic viewpoint since, once w_t is given, only $\mathcal{F}_t$-adapted processes can be regarded as *statistically observable.*

Let $S(t)$ be the semigroup generated in H by A. Then, in accordance with the results in [24, 90], one has:

$$\begin{cases} y(\cdot,t) = S(t)y^0 + \displaystyle\int_0^t S(t-s)(1_\omega v(\cdot,s))\,ds + \int_0^t S(t-s)B(s)\,dw_s \\ \forall t \in [0,T] \end{cases} \tag{3.23}$$

Our first result deals with approximate controllability:

THEOREM 3.3. *The linear manifold* $Y_T = \{\, y(\cdot,T) : v \in I^2(0,T;H) \,\}$ *is dense in* $L^2(\Lambda, \mathcal{F}_T; H)$. *In other words: for any* $y^1 \in L^2(\Lambda, \mathcal{F}_T; H)$ *and any* $\varepsilon > 0$, *there exists a control* $v \in I^2(0,T;H)$ *such that the associated solution to* (0.1) *satisfies:*

$$E\|y(\cdot,T) - y^1\|_{L^2}^2 \le \varepsilon.$$

Accordingly, it is said that (0.1) *is approximately controllable in quadratic mean.*

PROOF. We will argue as in the deterministic case. In view of (3.23), it will suffice to check that, if $f \in L^2(\Lambda, \mathcal{F}_T; H)$ and

$$E(\int_0^T S(T-s)(1_\omega v(\cdot,s))\,ds,\ f)_{L^2} = 0 \quad \forall v \in I^2(0,T;H), \tag{3.24}$$

then necessarily $f = 0$.

Let f be a function in $L^2(\Lambda, \mathcal{F}_T; H)$ satisfying (3.24) and assume that $\phi \in I^2(0,T;H)$ is defined pathwise by

$$\begin{cases} -\phi_t + A\phi = 0 & \text{in} \quad Q, \\ \phi = 0 & \text{on} \quad \Sigma, \\ \phi(x,T) = f(x) & \text{in} \quad \Omega, \end{cases}$$

i.e. $\phi(\cdot,t) = S(T-t)f$ for all t. It will be sufficient to prove that

$$E[\phi(\cdot,t)|\mathcal{F}_t] = 0 \quad \forall t \in (0,T). \tag{3.25}$$

Indeed, this and the continuity property (3.19) of the family $\{\mathcal{F}_t\}$ clearly imply that

$$f = E[\phi(\cdot,T)|\mathcal{F}_T] = 0.$$

□

QUESTION 3. *Why do* (3.25) *and* (3.19) *imply that* $f = 0$*?*

We know that

$$E\int_0^T (v(\cdot,s),\ 1_\omega\phi(\cdot,s))_{L^2}\,ds = 0 \qquad \forall v \in I^2(0,T;H).$$

Thus, $1_\omega E[\phi(\cdot,t)|\mathcal{F}_t]$ is a stochastic process in $I^2(0,T;H)$ such that

$$E\int_0^T (v(\cdot,s),\ 1_\omega E[\phi(\cdot,s)|\mathcal{F}_s])\,ds = \int_0^T E(v(\cdot,s),\ 1_\omega\phi(\cdot,s))\,ds = 0$$

for all $v \in I^2(0,T;H)$ and, consequently,

$$1_\omega E[\phi(\cdot,t)|\mathcal{F}_t] = 0. \tag{3.26}$$

For each $t \in (0,T)$, $E[\phi(\cdot,t)|\mathcal{F}_t] = S(T-t)E[f|\mathcal{F}_t]$ is real analytic in the variable $x \in \Omega$. Hence, one must necessarily have $E[\phi(\cdot,t)|\mathcal{F}_t] = 0$ for all $t \in (0,T)$ and the result is proved. □

A consequence of this theorem is that, for any $y^1 \in L^2(\Lambda, \mathcal{F}_T; H)$, $\varepsilon > 0$ and $\delta > 0$, a control v can be found such that

$$P\{\,\|y(\cdot,T) - y^1\|_{L^2} < \varepsilon\,\} \geq 1 - \delta.$$

However, the existence of a control $v \in I^2(0,T;H)$ such that $P\{\,\|y(\cdot,T) - y^1\|_{L^2} < \varepsilon\,\} = 1$ is an interesting open question.

The approximate controllability in quadratic mean remains true for systems governed by more general linear equations. More precisely, the following result is proved in [42]:

THEOREM 3.4. *Assume that, in* (0.1)*, A is an operator of the form*

$$Ay = -\sum_{i,j=1}^{N} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial y}{\partial x_j}\right) + \sum_{j=1}^{N} b_j \frac{\partial y}{\partial x_j} + cy, \tag{3.27}$$

where the coefficients satisfy

$$a_{ij} \in C^1(\overline{\Omega}), \quad b_j, c \in L^\infty(\Omega)$$

and the usual ellipticity condition

$$\sum_{i,j=1}^{N} a_{ij}(x)\xi_j\xi_i \geq \alpha|\xi|^2 \quad \forall \lambda \in \mathbb{R}^N, \quad \forall x \in \Omega, \quad \alpha > 0.$$

Then the corresponding $Y_T = \{\, y(\cdot, T) : v \in I^2(0,T;H)\,\}$ is dense in $L^2(\Lambda, \mathcal{F}_T; H)$.

We will now recall a null controllability result for (0.1) from [42]. Again, this is the analog of a deterministic result.

THEOREM 3.5. *Let us set $\gamma(t) := t(T-t)$. Assume that B is not random, $B \in C^1([0,T]; \mathcal{L}(K;H))$ and, also, that the support of $B(t)w$ does not intersect ω for any t and $w \in K$. Then there exists a positive function $\beta = \beta(x)$ such that, if*

$$\iint_Q t\left(\gamma(t)^{-1}\|B\|^2_{\mathcal{L}(K;H)} + \gamma(t)^3\|B_t\|^2_{\mathcal{L}(K;H)}\right) e^{2\beta(x)/\gamma(t)}\,dx\,dt < +\infty, \tag{3.28}$$

for each $y^0 \in H$ there exists $v \in I^2(0,T;H)$ satisfying $y(x,T) \equiv 0$, i.e. (0.1) *is null controllable.*

As in the deterministic case, the proof relies on an observability estimate for the solution of the adjoint system.

The situation is more complicate in the case of a *multiplicative noise*, that is, for systems of the form

$$\begin{cases} y \in I^2(0,T;V) \cap L^2(\Omega; C^0([0,T],H)), \\ y(\cdot,t) = y^0 + \displaystyle\int_0^t \{-Ay(\cdot,s) + 1_\omega v(\cdot,s)\}\,ds + \int_0^t By(\cdot,s)\,dw_s \quad \forall t \in [0,T]. \end{cases} \tag{3.29}$$

Here, B is given by $(By)(x) = b(x)y(x)$ for some $b \in W^{1,\infty}(\Omega)$ and (for simplicity) w_t is a real Wiener process.

From the theory of stochastic partial differential equations, it follows in particular that, for each $v \in I^2(0,T;H)$, there exists exactly one solution y to (3.29), see [90].

The approximate controllability in quadratic mean of (3.29) is equivalent to the unique continuation property for the following backward (adjoint) stochastic system:

$$\begin{cases} p \in I^2(0,T;V) \cap L^2(\Omega; C^0([0,T];H)), \quad q \in I^2(0,T;H), \\ p(\cdot,t) = f + \displaystyle\int_t^T \{A^*p(\cdot,s) + Bq(\cdot,s)\}\,ds - \int_t^T q(\cdot,s)\,dw_s \quad \forall t \in [0,T]. \end{cases} \tag{3.30}$$

In [8], a global Carleman estimate has been established for this system when $A = -\Delta$ and $b \in C^2(\overline{\Omega})$. Of course, this implies unique continuation for (3.30)

and, consequently, approximate controllability in quadratic mean for (3.29) in this particular case.

On the other hand, an appropriate unique continuation property for (3.30) has been proved in [34] in a more general case. As a consequence, one has approximate controllability in quadratic mean for (3.29). In fact, when b is a constant, and Γ is of class C^∞, we can also prove approximate controllability in all spaces $L^r(\Lambda, \mathcal{F}_T; L^q(\mathcal{O}))$ with $1 \leq r, q < +\infty$.

The previous analysis can also be made for stochastic Stokes systems; see [41].

For more results concerning the approximate and null controllability of stochastic PDEs, see the recent paper [108].

4. Positive and negative results for the Burgers equation

In this Section, we will be concerned with the null controllability of the following system for the viscous Burgers equation:

$$\begin{cases} y_t - y_{xx} + yy_x = v1_\omega\,, & (x,t) \in (0,1)\times(0,T), \\ y(0,t) = y(1,t) = 0, & t \in (0,T), \\ y(x,0) = y^0(x), & x \in (0,1). \end{cases} \tag{3.31}$$

Recall that some controllability properties of (3.31) have been studied in [47] (see Chapter 1, theorems 6.3 and 6.4). There, it is shown that, in general, a stationary solution of (3.31) with large L^2-norm cannot be reached (not even approximately) at any time T. In other words, with the help of one control, the solutions of the Burgers equation cannot go anywhere at any time.

For each $y^0 \in L^2(0,1)$, let us introduce

$$T(y^0) = \inf\{\, T > 0 : (3.31) \text{ is null controllable at time } T\,\}.$$

Then, for each $r > 0$, let us define the quantity

$$T^*(r) = \sup\{\, T(y^0) : \|y^0\|_{L^2} \leq r\,\}.$$

Our main purpose is to show that $T^*(r) > 0$, with explicit sharp estimates from above and from below. In particular, this will imply that (global) null controllability at any positive time does not hold for (3.31).

More precisely, let us set $\phi(r) = (\log \frac{1}{r})^{-1}$. We have the following result from [37]:

THEOREM 3.6. *One has*

$$C_0\phi(r) \leq T^*(r) \leq C_1\phi(r) \quad as \quad r \to 0, \tag{3.32}$$

for some positive constants C_0 and C_1 not depending of r.

REMARK 3.7. The same estimates hold when the control v acts on system (3.31) through the boundary *only* at $x = 1$ (or only at $x = 0$). Indeed, it is easy to transform the boundary controlled system

$$\begin{cases} y_t - y_{xx} + yy_x = 0, & (x,t) \in (0,1)\times(0,T), \\ y(0,t) = 0, \quad y(1,t) = w(t), & t \in (0,T), \\ y(x,0) = y^0(x), & x \in (0,1) \end{cases} \tag{3.33}$$

into a system of the kind (3.31). The boundary controllability of the Burgers equation with *two* controls (at $x = 0$ and $x = 1$) has been analyzed in [54]. There, it is shown that even in this more favorable situation null controllability does not hold for small time. It is also proved in that paper that exact controllability does not hold for large time.[3] □

REMARK 3.8. It is proved in [20] that the Burgers equation is *globally* null controllable when we act on the system through two boundary controls and an additional right hand side only depending on t. In other words, for any $y^0 \in L^2(0,1)$, there exist w_1 , w_2 and h in $L^2(0,T)$ such that the soution to

$$\begin{cases} y_t - y_{xx} + yy_x = h(t), & (x,t) \in (0,1)\times(0,T), \\ y(0,t) = w_1(t), \quad y(1,t) = w_2(t), & t \in (0,T), \\ y(x,0) = y^0(x), & x \in (0,1) \end{cases} \tag{3.34}$$

satisfies

$$y(x,T) = 0 \ \text{ in } \ (0,1).$$

However, it is unknown whether this global property is conserved when one of the boundary controls w_1 or w_2 is eliminated. □

The proof of the estimate from above in (3.32) can be obtained by solving the null controllability problem for (3.31) via a (more or less) standard fixed point argument, using global Carleman inequalities to estimate the control and energy inequalities to estimate the state and being very careful with the role of T in these inequalities.

The proof of the estimate from below is inspired by the arguments in [5] and is implied by the following property: there exist positive constants C_0 and C_0' such that, for any sufficiently small $r > 0$, we can find initial data y^0 and associated states y satisfying $\|y^0\|_{L^2} \leq r$ and

$$|y(x,t)| \geq C_0' r \ \text{ for some } x \in (0,1) \text{ and any } t \text{ satisfying } 0 < t < C_0\phi(r).$$

For more details, see [37].

5. The Navier-Stokes and Boussinesq systems

There is a lot of more realistic nonlinear equations and systems from mechanics that can also be considered in this context. First, we have the well known Navier-Stokes equations:

$$\begin{cases} y_t + (y\cdot\nabla)y - \Delta y + \nabla p = v1_\omega, \quad \nabla\cdot y = 0, & (x,t) \in Q, \\ y = 0, & (x,t) \in \Sigma, \\ y(x,0) = y^0(x), & x \in \Omega. \end{cases} \tag{3.35}$$

Here and below, $N = 2$ or $N = 3$ and (again) $\omega \subset \Omega$ is a nonempty open set.

In (3.35), (y,p) is the state (the velocity field and the pressure distribution) and v is the control (a field of external forces applied to the fluid particles located at ω). To our knowledge, the best results concerning the controllability of this system

[3] Let us remark that the results in [54] do not allow to estimate $T(r)$; in fact, the proofs are based in contradiction arguments.

have been given in [39] and [40].[4] Essentially, these results establish the local exact controllability of the solutions of (3.35) to bounded uncontrolled trajectories.

In order to be more specific, let us recall the definition of some usual spaces in the context of Navier-Stokes equations:

$$V := \left\{ y \in H_0^1(\Omega)^N : \nabla \cdot y = 0 \text{ in } \Omega \right\}$$

and

$$H := \{ y \in L^2(\Omega)^N : \nabla \cdot y = 0 \text{ in } \Omega,\ y \cdot n = 0 \text{ on } \partial\Omega \}.$$

Of course, it will be said that (3.35) is *exactly controllable to the trajectories* if, for any trajectory $(\overline{y}, \overline{p})$, i.e. any solution of the uncontrolled Navier-Stokes system

$$\begin{cases} \overline{y}_t + (\overline{y} \cdot \nabla)\overline{y} - \Delta \overline{y} + \nabla \overline{p} = 0, \quad \nabla \cdot \overline{y} = 0, & (x,t) \in Q, \\ \overline{y} = 0, & (x,t) \in \Sigma \end{cases} \tag{3.36}$$

and any $y^0 \in H$, there exist controls $v \in L^2(\omega \times (0,T))^N$ and associated solutions (y,p) such that

$$y(x,T) = \overline{y}(x,T) \text{ in } \Omega. \tag{3.37}$$

At present, we do not know any global result concerning exact controllability to the trajectories for (3.35). However, the following local result holds:

THEOREM 3.9. *Let $(\overline{y}, \overline{p})$ be a strong solution of* (3.36)*, with*

$$\overline{y} \in L^\infty(Q)^N, \quad \overline{y}(\cdot\,, 0) \in V. \tag{3.38}$$

Then, there exists $\delta > 0$ such that, for any $y^0 \in H \cap L^{2N-2}(\Omega)^N$ satisfying $\|y^0 - \overline{y}^0\|_{L^{2N-2}} \leq \delta$, we can find a control $v \in L^2(\omega \times (0,T))^N$ and an associated solution (y,p) to (3.35) *such that* (3.37) *holds.*

In other words, the local exact controllability to the trajectories holds for (3.35) in the space $X = L^{2N-2}(\Omega)^N \cap H$ Similar questions were addressed (and solved) in [46] and [45]. The fact that we consider here Dirichlet boundary conditions and locally supported distributed control increases a lot the mathematical difficulty of the control problem.

REMARK 3.10. It is clear that we cannot expect exact controllability for the Navier-Stokes equations with an arbitrary target function, because of the dissipative and irreversible properties of the system. On the other hand, approximate controllability is still an open question for this system. Some results in this direction have been obtained in [22] for different boundary conditions (Navier slip boundary conditions) and in [29] with a different nonlinearity. However, the notion of approximate controllability does not appear to be optimal from a practical viewpoint. Indeed, even if we could reach an arbitrary neighborhood of a given target y^1 at time T by the action of a control, the question of what to do afterwards to stay in the same neighbourhood would remain open. □

The proof of theorem 3.9 can be obtained as an application of *Liusternik's inverse mapping theorem* in an appropriate framework.

[4] The main ideas come from [49, 65]; some similar results have been given more recently in [52].

A key point in the proof is a related null controllability result for the linearized Navier-Stokes system at $(\overline{y}, \overline{p})$, that is to say:

$$\begin{cases} y_t + (\overline{y} \cdot \nabla) y + (y \cdot \nabla)\overline{y} - \Delta y + \nabla p = v 1_\omega, & (x,t) \in Q, \\ \nabla \cdot y = 0, & (x,t) \in Q, \\ y = 0, & (x,t) \in \Sigma, \\ y(x,0) = y^0(x), & x \in \Omega. \end{cases} \tag{3.39}$$

This is implied by a global Carleman inequality of the kind (3.5) that can be established for the solutions to the adjoint of (3.39), which is the following:

$$\begin{cases} -\varphi_t - (\nabla\varphi + \nabla\varphi^t)\,\overline{y} - \Delta\varphi + \nabla\pi = g, & (x,t) \in Q, \\ \nabla \cdot \varphi = 0, & (x,t) \in Q, \\ \varphi = 0, & (x,t) \in \Sigma, \\ \varphi(x,T) = \varphi^0(x), & x \in \Omega. \end{cases} \tag{3.40}$$

The details can be found in [39].

Similar results have been given in [53] for the Boussinesq equations

$$\begin{cases} y_t + (y \cdot \nabla) y - \Delta y + \nabla p = v 1_\omega + \theta\, e_N, \quad \nabla \cdot y = 0 & (x,t) \in Q, \\ \theta_t + y \cdot \nabla\theta - \Delta\theta = h 1_\omega, & (x,t) \in Q, \\ y = 0, \quad \theta = 0, & (x,t) \in \Sigma, \\ y(x,0) = y^0(x), \quad \theta(x,0) = \theta^0(x), & x \in \Omega. \end{cases} \tag{3.41}$$

Here, the state is the triplet (y, p, θ) (θ is interpreted as a temperature distribution) and the control is (v, h) (as before, v is a field of external forces; h is an external heat source).

QUESTION 4. *Can we deduce from theorem 3.9 a null controllability result for* (3.35) *for large* T*? What about* (3.41)*?*

QUESTION 5. *Does local null controllability imply local exact controllability to the trajectories in the context of* (3.35)*? What about* (3.41)*?*

An interesting question concerning both (3.35) and (3.41) is whether we can still get local exact controllability to the trajectories with a reduced number of scalar controls. This is partially answered in [40], where the following results are proved:

THEOREM 3.11. *Assume that the following property is satisfied:*

$$\exists x^0 \in \partial\Omega,\ \exists \varepsilon > 0 \ \text{ such that } \ \overline{\omega} \cap \partial\Omega \supset B(x^0; \varepsilon) \cap \partial\Omega. \tag{3.42}$$

Here, $B(x^0; \varepsilon)$ *is the ball centered at* x^0 *of radius* ε*. Then, for any* $T > 0$*,* (3.35) *is locally exactly controllable at time* T *to the trajectories satisfying* (3.38) *with controls* $v \in L^2(\omega \times (0,T))^N$ *having one component identically zero.*

THEOREM 3.12. *Assume that* ω *satisfies* (3.42) *with* $n_k(x^0) \neq 0$ *for some* $k < N$*. Then, for any* $T > 0$*,* (3.41) *is locally exactly controllable at time* T *to the trajectories* $(\overline{y}, \overline{p}, \overline{\theta})$ *satisfying* (3.38) *and*

$$\overline{\theta} \in L^\infty(Q), \quad \overline{\theta}(\cdot\,, 0) \in H_0^1(\Omega), \tag{3.43}$$

with controls $v \in L^2(\omega \times (0,T))^N$ *and* $h \in L^2(\omega \times (0,T))$ *such that* $v_k \equiv v_N \equiv 0$. *In particular, if* $N = 2$, *we have local exact controllability to these trajectories with controls* $v \equiv 0$ *and* $h \in L^2(\omega \times (0,T))$.

The proofs of theorems 3.11 and 3.12 are similar to the proof of theorem 3.9. We have again to rewrite the controllability property as a nonlinear equation in a Hilbert space. Then, we have to check that the hypotheses of Liusternik's theorem are fulfilled.

Again, a crucial point is to prove the null controllability of certain linearized systems, this time with *modified* controls. For instance, when dealing with (3.35), the task is reduced to prove that, for some appropriate weights ρ, ρ_0 and some $K > 0$, the solutions to (3.40) satisfy the following Carleman-like estimates:

$$\iint_{\Omega\times(0,T)} \rho^{-2}|\varphi|^2\,dx\,dt \le K \iint_{\omega\times(0,T)} \rho_0^{-2}(|\varphi_1|^2 + |\varphi_2|^2)\,dx\,dt \quad \forall \varphi^0 \in H. \tag{3.44}$$

This inequality can be proved using the assumption (3.42) and the incompressibility identity $\nabla \cdot \varphi = 0$; see [40].

6. Some other nonlinear systems from mechanics

The previous arguments can be applied to other similar partial differential systems arising in mechanics. For instance, this is done in [38] in the context of micropolar fluids.

To fix ideas, let us assume that $N = 3$. The behavior of a micropolar three-dimensional fluid is governed by a system which has the form

$$\begin{cases} y_t - \Delta y + (y\cdot\nabla)y + \nabla p = \nabla \times w + v1_\omega, \quad \nabla \cdot y = 0, & (x,t) \in Q, \\ w_t + (y\cdot\nabla)w - \Delta w - \nabla(\nabla\cdot w) = \nabla \times y + u1_\omega, & (x,t) \in Q, \\ y = 0, \quad w = 0 & (x,t) \in \Sigma, \\ y(x,0) = y^0(x), \quad w(x,0) = w^0(x) & x \in \Omega. \end{cases} \tag{3.45}$$

Here, the state is (y, p, w) and the control is (v, u). As usual, y and p stand for the velocity field and pressure and w is the microscopic velocity of rotation of the fluid particles.

The following result holds:

THEOREM 3.13. *Let* $(\overline{y}, \overline{p}, \overline{w})$ *be such that*

$$\overline{y},\, \overline{w} \in L^\infty(Q) \cap L^2(0,T; H^2(\Omega)), \quad \overline{y}_t,\, \overline{w}_t \in L^2(Q) \tag{3.46}$$

and

$$\begin{cases} \overline{y}_t - \Delta\overline{y} + (\overline{y}\cdot\nabla)\overline{y} + \nabla\overline{p} = \nabla\times\overline{w}, \quad \nabla\cdot\overline{y} = 0, & (x,t) \in Q, \\ \overline{w}_t + (\overline{y}\cdot\nabla)\overline{w} - \Delta\overline{w} - \nabla(\nabla\cdot\overline{w}) = \nabla\times\overline{y}, & (x,t) \in Q, \\ \overline{y} = 0, \quad \overline{w} = 0 & (x,t) \in \Sigma. \end{cases} \tag{3.47}$$

Then, for each $T > 0$, (3.45) *is locally exactly controllable to* $(\overline{y}, \overline{p}, \overline{w})$ *at time* T. *In other words, there exists* $\delta > 0$ *such that, for any initial data* $(y^0, w^0) \in (H^2(\Omega) \cap V) \times H_0^1(\Omega)$ *satisfying*

$$\|(y^0, w^0) - (\overline{y}(\cdot\,,0), \overline{w}(\cdot\,,0))\|_{H^2\times H_0^1} \le \delta, \tag{3.48}$$

there exist L^2 controls u and v and associated solutions (y, p, w) satisfying

$$y(x,T) = \overline{y}(x,T), \quad w(x,T) = \overline{w}(x,T) \quad in \quad \Omega. \tag{3.49}$$

Notice that this case involves a nontrivial difficulty. Indeed, w is a non-scalar variable and the equations satisfied by its components w_i are coupled through the second-order terms $\partial_i(\nabla \cdot w)$. This is a serious inconvenient. An appropriate strategy has to be applied in order to deduce the required Carleman estimates.

Let us also mention [7, 61, 62], where the controllability of the MHD and other related equations has been analyzed.

For all these systems, the proof of the controllability can be achieved arguing as in the first part of the proof of theorem 3.9. This is the general structure of the argument:

- First, rewrite the original controllability problem as a nonlinear equation in a space of admissible "state-control" variables.
- Then, prove an appropriate global Carleman inequality and a regularity result and deduce that the linearized equation possesses at least one solution. This provides a controllability result for a related linear problem.
- Finally, check that the hypotheses of a suitable implicit function theorem are satisfied and deduce a local result.

REMARK 3.14. Recall that an alternative strategy was introduced in [109] in the context of the semilinear wave equation:

- First, consider a linearized similar problem and rewrite the original controllability problem in terms of a fixed point equation.
- Then, prove a global Carleman inequality and deduce an observability estimate for the adjoint system and a controllability result for the linearized problem.
- Finally, prove appropriate estimates for the control and the state (this usually needs some kind of *smallness* of the data), prove an appropriate compactness property of the state and deduce that there exists at least one fixed point.

This method has been used in [30] and [44] in the context of semilinear heat equations and in [52] to prove a result similar to theorem 3.9. □

REMARK 3.15. Observe that all these results are positive, in the sense that they provide local controllability properties. At present, no negative result is known to hold for these nonlinear systems (except for the already considered one-dimensional Burgers equation). □

To end this Section, let us mention two systems from fluid mechanics, apparently not much more complex than (3.35), for which local controllability to the trajectories is an open question.

The first system is the following:

$$\begin{cases} y_t + (y \cdot \nabla)y - \nabla \cdot (\nu(|Dy|)Dy) + \nabla p = v1_\omega\,, & (x,t) \in Q, \\ \nabla \cdot y = 0, & (x,t) \in Q, \\ y = 0, & (x,t) \in \Sigma, \\ y(x,0) = y^0(x), & x \in \Omega. \end{cases} \tag{3.50}$$

Here, $Dy = \frac{1}{2}(\nabla y + \nabla y^t)$ and $\nu : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is a regular function (for example, we can take $\nu(s) \equiv a + bs^{r-1}$ for some $a, b > 0$ and some $r > 1$). This models the behavior of a *quasi-Newtonian* fluid; for a mathematical analysis, see [11, 83].

In view of the new nonlinear diffusion term $\nabla \cdot (\nu(|Dy|)Dy)$, the control properties of (3.51) are much more difficult to analyze than for (3.35). In particular, it is unknown whether the local approximate and the local null controllability properties hold for (3.50).

For the second system, we suppose that $N = 2$. It reads:

$$\begin{cases} \theta_t + (y \cdot \nabla)\theta - \Delta\theta = v1_\omega\,, & (x,t) \in Q, \\ y = \nabla \times \left((-\Delta)^{-a}\theta\right), & (x,t) \in Q, \\ \theta = 0, & (x,t) \in \Sigma, \\ \theta(x,0) = \theta^0(x), & x \in \Omega, \end{cases} \tag{3.51}$$

where $a \in [1/2, 1]$. We are now modelling the behavior of a *quasi-geostrophic* fluid. The state variables θ and y may be viewed as a *generalized vorticity* and velocity field, respectively (notice that, for $a = 1$, we find again the Navier-Stokes system written in terms of y and $\nabla \times y$; see for instance [92]).

It is possible to prove a local null controllability result for (3.51). However, to our knowledge, the local approximate controllability and the local exact controllability to the trajectories are open problems.

Bibliography

[1] R.A. Adams, *Sobolev Spaces,* Academic Press, New York 1975.

[2] G. Alessandrini and L. Escauriaza, *Null-controllability of one-dimensional parabolic equations,* ESAIM Control Optim. Calc. Var. 14, no. 2, 2007, 284–293.

[3] S. Alinhac, *Non unicité du problème de Cauchy*, Annals of Mathematics, **117** (1983), 77–108.

[4] F. Ammar-Khodja, A, Benabdallah, C. Dupaix and M. González-Burgos, *A Kalman rank condition for the localized distributed controllability of class of linear parabolic systems,* J. Evol. Equ. **9** (2009), no. 2, 267–291.

[5] S. Anita and D. Tataru, *Null controllability for the dissipative semilinear heat equation,* Appl. Math. Optim. **46**, 2002, 97–105.

[6] L. Arnold, *Stochastic differential equations: theory and applications,* Wiley-Interscience [John Wiley & Sons], New York-London-Sydney, 1974.

[7] V. Barbu V, T. Havarneanu, C. Popa and S.S. Sritharan SS, *Exact controllability for the magnetohydrodynamic equations,* Comm. Pure Appl. Math. **56**, 2003, 732–783.

[8] V. Barbu, A. Rascanu and G. Tesitore, *Carleman estimates and controllability of linear stochastic heat equations,* Appl. Math. Optim. **47** (2003), no. 2, 97–120.

[9] C. Bardos, G. Lebeau and J. Rauch, Sharp sufficient conditions for the observation, control and stabilization of waves from the boundary, SIAM J. Cont. Optim., **30** (1992), 1024–1065.

[10] J.A. Bello, E. Fernández-Cara, J. Lemoine and J. Simon, *The differentiability of the drag with respect to the variations of a Lipschitz domain in a Navier-Stokes flow,* SIAM J. Control Optimiz., Vol. 35, No. 2, pp. 626–640, 1997.

[11] H. Bellout, F. Bloom and J. Nečas, *Young measure-valued solutions for non-Newtonian incompressible fluids,* Comm. PDE **19**, no. 11–12, 1994, 1763–1803.

[12] A. Benabdallah, Y. Dermenjian, Yves and J. Le Rousseau, *Carleman estimates for the one-dimensional heat equation with a discontinuous coefficient and applications to controllability and an inverse problem,* J. Math. Anal. Appl. **336** (2007), no. 2, 865–887.

[13] A. Benabdallah, Y. Dermenjian, Yves and J. Le Rousseau, *to appear.*

[14] J.F. Bonnans and E. Casas, *Optimal control of semilinear multistate systems with state constraints,* SIAM J. Control Optimiz. **27** (1989), No. 2, 446–455.

[15] N. Burq, Contrôle de l'équation des ondes dans des ouverts peu réguliers, Asymptotic Analysis, **14** (1997), 157–191.

[16] E. Casas, *Boundary control of semilinear elliptic equations with pointwise state constraints,* SIAM J. Control Optimiz. **31** (1993), No. 4, 996–1006.

[17] E. Casas and L.A. Fernández, *Optimal control of semilinear elliptic equations with pointwise constraints on the gradient of the state,* Appl. Math. Optim. **27** (1993), No. 1, 35–56.

[18] T. Cazenave, *On the propagation of confined waves along the geodesics*, J. Math. Anal. Appl., **146** (1990), 591–603.

[19] D. Chae, O.Yu. Imanuvilov and S.M. Kim, *Exact controllability for semilinear parabolic equations with Neumann boundary conditions*, J. Dynamical and Control Systems **2** (1996), 449–483.

[20] M. Chapouly, *Global controllability of nonviscous and viscous Burgers-type equations,* SIAM J. Control Optim. **48** (2009), no. 3, 1567–1599.

[21] D. Chenais, *On the existence of a solution in a domain identification problem,* J. Math. Anal. Appl. **52** (2) (1975), 430–445.

[22] J.-M. Coron, *On the controllability of the 2-D incompressible Navier-Stokes equations with the Navier slip boundary conditions,* ESAIM Control Optim. Calc. Var. **1**, (1995/96), 35–75.

[23] J.-M. Coron, *Control and nonlinearity, Control and nonlinearity,* Mathematical Surveys and Monographs, 136. American Mathematical Society, Providence, RI, 2007.

[24] G. DaPrato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Cambridge University Press, Cambridge 1992.

[25] C. Dellacherie, *Analytical Sets, Capacities and Haussdorff Measures,* Lectures Notes in Mathematics, No. 180, Springer-Verlag, Berlin 1972.

[26] A. Doubova, E. Fernández-Cara, M. González-Burgos and E. Zuazua, *On the controllability of parabolic systems with a nonlinear term involving the state and the gradient,* SIAM J. Control Optim. **41** (2002), no. 3, 798–819.

[27] A. Doubova, A. Osses and J.-P. Puel, *Exact controllability to trajectories for semilinear heat equations with discontinuous diffusion coefficients,* ESAIM Control Optim. Calc. Var. **8** (2002), 621–661.

[28] R. Echevarría, A. Doubova, E. Fernández-Cara and I. Gayte, *Control de EDPs orientado a la terapia de un tumor cerebral,* Communication to CEDYA 2007, Sevilla, 2007.

[29] C. Fabre, *Uniqueness results for Stokes equations and their consequences in linear and nonlinear control problems,* ESAIM Control Optim. Calc. Var. **1**, (1995/96), 267–302.

[30] C. Fabre, J.-P. Puel and E. Zuazua, *Approximate controllability of the semilinear heat equation,* Proc. Royal Soc. Edinburgh A **125** (1995), 31–61.

[31] C. Fabre, J.P. Puel and E. Zuazua, Contrôlabilité approchée de l'équation de la chaleur linéaire avec des contrôles de norme L^∞ minimale, C.R. Acad. Sci. Paris, **316** (1993), 679–684.

[32] H.O. Fattorini, *Infinite Dimensional Optimization and Control Theory,* Encyclopedia of Mathematics and its Applications **62**, Cambridge University Press, 1999.

[33] L.A. Fernández and E. Zuazua, *Approximate controllability for the semilinear heat equation involving gradient terms,* J. Optim. Theory Appl. **101** (1999), no. 2, 307–328.

[34] E. Fernández-Cara, M.J. Garrido-Atienza and J. Real, *On the approximate controllability of a stochastic parabolic equation with a multiplicative noise,*

C.R. Acad. Sci. Paris **328** (1999), 675–680.

[35] E. Fernández-Cara, M. González-Burgos and L. De Teresa, *Null-exact controllability of a semilinear cascade system of parabolic-hyperbolic equations,* Comm. Pure Applied Anal. **5** (2008), no. 3, 639–658.

[36] E. Fernández-Cara and S.Guerrero, *Global Carleman inequalities for parabolic systems and applications to controllability,* SIAM J. Control Optim. **45** (2006), no. 4, 1399–1446.

[37] E. Fernández-Cara and S. Guerrero, *Null controllability of the Burgers system with distributed controls,* Systems Control Lett. **56** (2007), no. 5, 366–372.

[38] E. Fernández-Cara and S. Guerrero, *Local exact controllability of micropolar fluids,* J. Math. Fluid. Mech. **8** (2006), 1–35.

[39] E. Fernández-Cara, S. Guerrero, O.Yu. Imanuvilov and J.-P. Puel, *Local exact controllability to the trajectories of the Navier-Stokes equations,* J. Math. Pures Appl. **83** (2004), no. 12, 1501–1542.

[40] E. Fernández-Cara, S. Guerrero, O.Yu. Imanuvilov and J.-P. Puel, *Some controllability results for the N-dimensional Navier-Stokes and Boussinesq systems with $N-1$ scalar controls,* SIAM J. Control and Optim. **45** (2006), no. 1, 146–173.

[41] E. Fernández-Cara. J.D. Martín and J. Real, *On the approximate controllability of stochastic Stokes systems,* Stoch. Anal. Appl. **17** (1999), no. 4, 563–577.

[42] E. Fernández-Cara and J. Real, *Remarks on the controllability of some stochastic partial differential equations,* in "Control and Estimation of Distributed Parameter Systems", p. 141–151, International Series of Numerical Mathematics, Birkhäuser Verlag, Basel 1998.

[43] E. Fernández-Cara and E. Zuazua, *The cost of approximate controllability for heat equations: The linear case*, Advances Diff. Eqs. **5** (2000), no. (4–6), 465–514.

[44] E. Fernández-Cara and E. Zuazua, *Null and approximate controllability for weakly blowing up semilinear heat equations,* Ann. Inst. H. Poincaré Anal. Non Linéaire **17** (2000), no. 5, 583–616.

[45] A.V. Fursikov, *Exact boundary zero-controllability of three-dimensional Navier-Stokes equations,* J. Dynam. Control Systems **1** (1995), no. 3, 325–350.

[46] A.V. Fursikov and O.Yu. Imanuvilov, *On exact boundary zero-controllability of two-dimensional Navier-Stokes equations,* Acta Applicandae Mathematicae **37** (1994), 67–76.

[47] A.V. Fursikov and O.Yu. Imanuvilov, *Controllability of evolution equations*, Lecture Notes Series # 34, Research Institute of Mathematics, Global Analysis Research Center, Seoul National University, 1996.

[48] A.V. Fursikov and O.Yu. Imanuvilov, Local exact controllability of the Boussinesq equation, SIAM J. Cont. Opt. **36** (1998), no. 2, 391–421.

[49] A.V. Fursikov and O. Yu. Imanuvilov, *Exact controllability of the Navier-Stokes and Boussinesq equations (Russian),* Uspekhi Mat. Nauk **54** (1999), no. 3(327), 93–146; translation in Russian Math. Surveys **54** (1999), no. 3, 565–618.

[50] P. Gérard, *Microlocal defect measures*, Comm. P.D.E. **16** (1991), 1761–1794.

[51] V. Girault and Ph.-A. Raviart, *Finite element methods for Navier-Stokes equations. Theory and algorithms,* Springer-Verlag, Berlin, 1986.

[52] M. González-Burgos, S. Guerrero and J.-P. Puel, *Local exact controllability to the trajectories of the Boussinesq system via a fictitious control on the divergence equation,* Commun. Pure Appl. Anal. **8** (2009), no. 1, 311–333.

[53] S. Guerrero, *Local exact controllability to the trajectories of the Boussinesq system,* Annales IHP, Anal. non linéaire **23** (2006), 29–61.

[54] S. Guerrero and O. Yu. Imanuvilov, *Remarks on global controllability for the Burgers equation with two control forces,* Ann. Inst. H. Poincaré Anal. Non Linéaire **24** (2007), no. 6, 897–906.

[55] R. Glowinski, *Ensuring well-posedness by analogy; Stokes problem and boundary control for the wave equation*, J. Compt. Physics **103** (1992), no. 2, 189–221.

[56] R. Glowinski, C.H. Li and J.-L. Lions, *A numerical approach to the exact boundary controllability of the wave equation (I). Dirichlet controls: Description of the numerical methods*, Japan J. Appl. Math. **7** (1990), 1–76.

[57] R. Glowinski and J.-L. Lions, *Exact and approximate controllability for distributed parameter systems*, Acta Numerica 1994, p. 269–378.

[58] R. Glowinski, J.-L. Lions and J. He, *Exact and approximate controllability for distributed parameter systems. A numerical approach,* Encyclopedia of Mathematics and its Applications, 117. Cambridge University Press, Cambridge, 2008.

[59] M.D. Gunzburger, *Perspectives in flow control and optimization,* Advances in design and control, 5. SIAM, Philadelphia, 2003.

[60] J. Haslinger and P. Neittaanmäki, *Finite Element Approximation for Optimal Shape Design,* John Wiley and Sons, Chichester, 1988.

[61] T. Havarneanu, C. Popa and S.S. Sritharan, *Exact internal controllability for the magnetohydrodynamic equations in multi-connected domains,* Adv. Differ. Equations **11** (2006), no. 8, 893–929.

[62] T. Havarneanu, C. Popa and S.S. Sritharan, *Exact internal controllability for the two-dimensional magnetohydrodynamic equations,* SIAM J. Control Optim. **46** (2007), no. 5, 1802–1830.

[63] L.F. Ho, Observabilité frontière de l'équation des ondes, C.R. Acad. Sci. Paris **302** (1986), 443–446.

[64] O.Yu. Imanuvilov, *Boundary controllability of parabolic equations*, Russian Acad. Sci. Sb. Math. **186** (1995), 109–132 (in Russian).

[65] O. Yu. Imanuvilov, *Remarks on exact controllability for the Navier-Stokes equations,* ESAIM Control Optim. Calc. Var. **6** (2001), 39–72.

[66] O.Yu. Imanuvilov and M. Yamamoto, *Carleman inequalities for parabolic equations in Sobolev spaces of negative order and exact controllability for semilinear parabolic equations,* Publ. Res. Inst. Math. Sci. **39** (2003), no. 2, 227–274.

[67] R. Kohn and G. Strang, *Optimal design and relaxation of variational problems, I, II and III,* Comm. Pure Appl. Math. **34** (1986), 113–137, 139–182 and 353–377.

[68] I. Lasiecka and R. Triggiani, *Exact controllability of semilinear abstract systems with application to waves and plates boundary control problems,* Appl. Math. Optim. **23** (1991), no. 2, 109–154.

[69] G. Lebeau, *Contrôle analytique I: estimations a priori,* Duke Math. J. **68** (1992), no. 1, 1–30.

[70] G. Lebeau and L. Robbiano, Contrôle exact de l'équation de la chaleur, Comm. P.D.E. **20** (1995), 335–356.

[71] E.B. Lee and L. Markus, *Foundations of Optimal Control Theory,* The SIAM Series in Applied Mathematics, John Wiley & Sons, New York 1967.

[72] S. Lenhart and J.T. Workman *Optimal control applied to biological models,* Chapman & Hall / CRC, 2007.

[73] J. Le Rousseau and L. Robbiano, *Carleman estimate for elliptic operators with coefficients with jumps at an interface in arbitrary dimension and application to the null controllability of linear parabolic equations,* Arch. Ration. Mech. Anal. **195** (2010), no. 3, 953–990.

[74] G. Leugering and E. Zuazua, *On exact controllability of generic trees,* in "Contrôle de Systèmes Gouvernés par des Equations aux Dérivées Partielles", ESAIM Proceedings, Vol. 8, F. Conrad and M. Tucsnak eds., pp. 95–105.

[75] X. Li and J. Yong, *Optimal Control Theory for Infinite Dimensional Systems,* Birkhäuser, Boston, 1995.

[76] J.-L. Lions, *Contrôle Optimale des Systèmes Gouvernés par des Equations aux Dérivées Partielles,* Dunod, Gauthiers-Villars, Paris, 1968.

[77] J.-L. Lions, Exact controllability, stabilizability and perturbations for distributed systems, SIAM Review **30** (1988), 1–68.

[78] J.-L. Lions, *Contrôlabilité Exacte, Stabilisation et Perturbations de Systèmes Distribués, Tomes 1 & 2.* Masson, RMA **8** & **9**, Paris 1988.

[79] J.-L. Lions, Remarques sur la contrôlabilité approchée, in *Jornadas Hispano-Francesas sobre Control de Sistemas Distribuidos*, University of Málaga, Spain, 1991, pp. 77–87.

[80] J.-L. Lions, Remarks on approximate controllability, J. Anal. Math. **59** (1992), 103–116.

[81] J.-L. Lions and B. Malgrange, Sur l'unicité rétrograde dans les problèmes mixtes paraboliques, Math. Scan. **8** (1960), 277–286.

[82] K.A. Lurie, *Applied Optimal Control Theory of Distributed Systems,* Plenum Press, New York, 1993.

[83] J. Málek, J. Nečas, M. Rokyta, M. Ružička M, *Weak and measure-valued solutions to evolutionary PDEs,* Applied Mathematics and Mathematical Computation, 13, Chapman & Hall, London, 1996.

[84] F. Murat, *H-convergence,* Lectures at the University of Alger, Alger, 1978.

[85] F. Murat, *H-convergence,* in "Topics in the Mathematical Modelling of Composite Materials", A. Cherkaev and R. Kohn eds., Birkhäuser, Boston, 1997.

[86] F. Murat and J. Simon, *Quelques résultats sur le contrôle par un domaine géométrique,* Rapport du L.A. 189 No. 74003, Université Paris VI, 1974.

[87] F. Murat and J. Simon, *Sur le contrôle par un domaine géométrique,* Rapport du L.A. 189 No. 76015, Université Paris VI, 1976.

[88] F. Murat and L. Tartar, *On the control of coefficients in partial differential equations,* in "Topics in the Mathematical Modelling of Composite Materials", A. Cherkaev and R. Kohn eds., Birkhäuser, Boston 1997.

[89] A. Osses, *A rotated multiplier applied to the controllability of waves, elasticity, and tangential Stokes control,* SIAM J. Control Optim. **40** (2001), no. 3, 777–800

[90] E. Pardoux, *PhD Thesis,* Université Paris XI (France), 1975.

[91] E. Pardoux, *Intégrales stochastiques hilbertiennes,* Cahiers de Mathématiques de la Décision No. 7617, Université Paris IX,1976.

[92] J. Pedlosky, *Geophysical fluid dynamics, Second edition,* Springer, New York, 1987.

[93] P. Pedregal, *Optimization, relaxation and Young measures,* Bull. A.M.S. **36** (1999), no. 1, 27–58.

[94] O. Pironneau, *On optimum design in fluid mechanics,* J. Fluid Mech. **64** (1974), 97–110.

[95] O. Pironneau, *PhD Thesis,* Université Pierre & Marie Curie (Paris VI; France), 1976.

[96] O. Pironneau, *Optimal Shape Design for Elliptic Systems,* Springer-Verlag, New-York, 1984.

[97] J.-P. Puel, *Global Carleman inequalities for the wave equations and applications to controllability and inverse problems,* in "Control of Solids and Structures: Mathematical Modelling and Engineering Applications", Udine (Italy), June 2004.

[98] L. Robbiano and C. Zuily, *Uniqueness in the Cauchy problem for operators with partially holomorphic coefficients*, Invent. Math. **131** (1998), 493–539.

[99] D.L. Russell, *Controllability and stabilizability theory for linear partial differential equations. Recent progress and open questions*, SIAM Review **20** (1978), 639–739.

[100] D.L. Russell, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Studies in Appl. Math., **52** (1973), 189–221.

[101] J. Simon, *Second variation in domain optimization problems,* in "Control and Estimation of Distributed Parameter Systems", F. Kappel, K. Kunish and W. Schappacher eds., Int. Series of Numer. Math., No. 91, p. 361–378, Birkhauser, 1989.

[102] J. Sokolowski and J.-P. Zolesio, *Introduction to Shape Optimization,* Springer-Verlag, Berlin, 1992.

[103] K. R. Swanson, E. C. Alvord Jr. and J. D. Murray, *A quantitative model for differential motility of gliomas in grey and white matter,* Cell Prolif. **33** (2000), 317–329.

[104] K. R. Swanson, E. C. Alvord Jr. and J. D. Murray, *Quantifying the efficacy of chemotherapy of brain tumors (gliomas) with homogeneous and heterogeneous drug delivery,* Acta Biotheoretica **50** (2002), no. 4, 223–237.

[105] D. Tataru, *Unique continuation for solutions to PDE's: between Hörmander's theorem and Holmgren's theorem*, Comm. PDE **20** (1995), no. 6–7, 855–884.

[106] L. Tartar, *Problèmes de contrôle des coefficients dans des équations aux dérivées partielles,* in "Control theory, numerical methods and computer systems modelling", A. Bensoussan and J.L. Lions eds., Lecture Notes in Economics and Mathematical Systems No. 107, Springer-Verlag, Berlin, 1975.

[107] X. Zhang, *Explicit observability estimate for the wave equation with potential and its application,* R. Soc. Lond. Proc. Ser. A, Math. Phys. Eng. Sci. **456** (2000), no. 1997, 1101–1115.

[108] X. Zhang, *A unified controllability/observability theory for some stochastic and deterministic partial dfferential equations,* Proc. ICM, Hyderabad, India, 2010.

[109] E. Zuazua E, *Exact boundary controllability for the semilinear wave equation,* in "Nonlinear Partial Differential Equations and their Applications", Vol. X (p. 357–391), H. Brezis and J.L. Lions eds., Pitman, New York, 1991.

[110] E. Zuazua, Finite dimensional null controllability for the semilinear heat equation, J. Math. Pures Appl. **76** (1997), 570–594.

[111] E. Zuazua *Controllability and observability of partial differential equations: some results and open problems,* Handbook of differential equations: evolutionary equations, Vol. III, 527–621, Handb. Differ. Equ., Elsevier/North-Holland, Amsterdam, 2007.

Part 4

Fluid-structure interaction for shear-dependent non-Newtonian fluids

Anna Hundertmark-Zaušková,
Mária Lukáčová-Medviďová, Gabriela Rusnáková

2000 *Mathematics Subject Classification.* Primary 65M60, 35Q30, 74F10, 65N30, 76D05, 35M99

Key words and phrases. non-Newtonian fluids, fluid-structure interaction, shear-thinning flow, hemodynamical wall parameters, stenosis, kinematical splitting, numerical stability, weak solution

Abstract. We present our recent results on mathematical modelling and numerical simulation of non-Newtonian flows in compliant two-dimensional domains having applications in hemodynamics. Two models of the shear-thinning non-Newtonian fluids, the power law Carreau model and the logarithmic Yeleswarapu model, will be considered. For the structural model the generalized string equation for radially symmetric tubes will be generalized to stenosed vessels and vessel bifurcations.

The arbitrary Lagrangian-Eulerian approach is used in order to take into account moving computational domains. To represent the fluid-structure interaction we use two different methods: the global iterative approach and the kinematical splitting. We will show that the latter method is more efficient and stable without any additional subiterations. The analytical result for the existence of a weak solution for the shear-thickening power-law fluid is based on the global iteration with respect to the domain deformation, energy estimates, compactness arguments using the semi-continuity in time and the theory of monotone operators. The numerical part of paper contains several experiments for the Carreau and the Yeleswarapu model, comparisons of the non-Newtonian and Newtonian models and the results for hemodynamical wall parameters; the wall shear stress and the oscillatory shear index. Numerical experiments confirm higher order accuracy and the reliability of new fluid-structure interaction methods.

Acknowledgement. This research has been supported by the European Union's 6th Framework Programme DEASE under the No. MEST-CT-2005-021122, the German Science Foundation (DFG) project ZA 613/1-1 as well as by the Nečas Centrum for Mathematical Modelling LC06052 (financed by MSMT). The third author has been partially supported by the Center for Computational Sciences in Mainz and the German Academic Exchange Service DAAD. The authors gratefully acknowledge these supports.

Contents

CHAPTER 1

Fluid-structure interaction methods

1. Introduction

In the recent years there is a growing interest in the use of mathematical models and numerical methods arising from other fields of computational fluid dynamics in hemodynamics, see, e.g., [6, 8, 18, 19, 21, 27, 31, 36, 37, 39, 40, 41] just to mention some of them.

Many numerical methods used for blood flow simulations are based on the Newtonian model using the Navier-Stokes equations. This is efficient and useful, especially if the flow in large arteries is modeled. However, in small vessels or dealing with patients with a cardiovascular disease more complex models for blood rheology should be considered [31]. In capillaries blood is even not a homogenized continuum and more precise models, for example mixture theories need to be used. But even in the intermediate-size vessels the non-Newtonian behavior of blood has been demonstrated, see, e.g., [2], [43] and the references therein. In fact, blood is a complex mixture showing several non-Newtonian properties, such as the shear-thinning, viscoelasticity [48], [49] the yield stress or the stress relaxation [43].

The aim of this overview paper is to report on our recent results on mathematical and numerical modelling of shear-dependent flow in moving vessels. The application to hemodynamics will be pointed out. We will address the significance of non-Newtonian models for reliable hemodynamical modelling. In particular, we will show that the rheological properties of fluid have an influence on the wall deformation as well as on the hemodynamical wall parameters, such as the wall shear stress and oscillatory shear index. Consequently these models yield a more reliable prediction of critical vessel areas, see also our previous results [24, 28, 29].

The paper is organized as follows. In Section 2 we recall the conservation laws for shear-dependent fluids and present typical models for non-constant blood viscosity. The generalized string model for the vessel deformation [40] is generalized to the case of reference radius, which is dependent on longitudinal variable. The derivation of this model for radially symmetric domains follows in Section 3.

Section 4 is devoted to two strategies to model the coupling between a fluid and a structure. *The global iterative method* with respect to the domain, presented in Section 4.1, provides besides the numerical scheme also a strategy to prove the existence of a weak solution. Mathematical analysis of the well-posedness of a coupled fluid-structure model arising from the blood flow in a compliant vessel is of great interest. In the literature there are already several results for the Newtonian fluid flow in time-dependent domains, see, e.g., [4, 5, 6, 7, 9, 10, 11, 12, 13, 20, 22, 23, 33, 47, 51] and others. The well-posedness of non-Newtonian fluids has been studied only in the fixed domains, see, e.g., [17, 34, 35, 50]. In these works the

technique of monotone operators and the Lipschitz- or L^∞-truncation techniques are applied in order to control the additional nonlinearities in the diffusion terms arising from the non-Newtonian viscosity. In this overview paper we also present our recent result on the existence of a weak solution for the shear-thickenning fluid in compliant vessels, cf. [25]. The proof is based on the global iterative method with respect to the domain deformation [13, 51], theory of monotone operators as well as the techniques for moving domains developed in [7, Chambolle, Desjarden, Esteban, Grandmont].

The second fluid-structure interaction approach, that will be presented in Section 4.2, is the loosely-coupled fluid-structure interaction algorithm based on the *kinematical splitting* [21]. This is a novel way how to avoid instabilities due to the added mass effect and the additional stabilization through subiterations. Subsection 4.2.1 is devoted to stability analysis of the kinematical splitting method. Further details can be found in our recent paper [28].

Results of numerical experiments are described in Section 5. We apply both fluid-structure interaction methods and compare domain deformations as well the hemodynamical wall indices measuring the danger of atherosclerotic plaque caused by temporal oscillation or low values of the wall shear stress. We use two types of data: the model data proposed by Sequeira and Nadau [31] and the physiological data from the iliac artery and the carotid bifurcation measurements. In the hemodynamical wall parameters the effects due to the fluid-structure interaction as well as the blood rheology have been observed. Finally the experimental order of convergence for a rigid as well as a moving domain for both fluid-structure interaction methods will be investigated. In the case of kinematical splitting method second order convergence will be confirmed.

2. Mathematical model for shear-dependent fluids

Flow of incompressible fluid is governed by the momentum and the continuity equation

$$\begin{aligned} \rho\frac{\partial \boldsymbol{u}}{\partial t} + \rho\left(\boldsymbol{u}\cdot\nabla\right)\boldsymbol{u} - \operatorname{div}\left[2\mu\mathbf{D}(\boldsymbol{u})\right] + \nabla p &= \boldsymbol{f} \\ \operatorname{div}\boldsymbol{u} &= 0. \end{aligned} \tag{2.1}$$

Here ρ denotes the constant density of fluid, $\boldsymbol{u} = (u_1, u_2)$ the velocity vector, p the pressure, $\mathbf{D}(\boldsymbol{u}) = \frac{1}{2}(\nabla\boldsymbol{u} + \nabla\boldsymbol{u}^T)$ the deformation tensor and μ the viscosity of the fluid. In the literature various non-Newtonian models for the blood flow can be found. Here we will consider shear-dependent fluids, in particular the Carreau model and the Yeleswarapu-viscosity model [48], see also Fig. 1. For the Carreau model the viscosity function depends on the shear rate $|\mathbf{D}(\boldsymbol{u})| = \sqrt{\mathbf{D}:\mathbf{D}} = \sqrt{tr(\mathbf{D}^2)}$ in the following way

$$\mu = \mu(\mathbf{D}(\boldsymbol{u})) = \mu_\infty + (\mu_0 - \mu_\infty)(1 + |\gamma\mathbf{D}(\boldsymbol{u})|^2)^q, \quad q = \frac{p-2}{2} \leq 0, \tag{2.2}$$

where $q, \mu_0, \mu_\infty, \gamma$ are rheological parameters. According to [48] the physiological values for blood are $\mu_0 = 0.56P$, $\mu_\infty = 0.0345P$, $\gamma = 3.313$, $q = -0.322$. Note that in the case $q = 0$ the model reduces to the linear Newtonian model used in

the Navier-Stokes equations. The Yeleswarapu viscosity model reads

$$\mu = \mu(\mathbf{D}(\boldsymbol{u})) = \mu_\infty + (\mu_0 - \mu_\infty)\frac{\ln(1+\gamma|\mathbf{D}(\boldsymbol{u})|)+1}{(1+\gamma|\mathbf{D}(\boldsymbol{u})|)}. \tag{2.3}$$

The physiological measurements give $\mu_0 = 0.736P,\ \mu_\infty = 0.05P,\ \gamma = 14.81$ [48]. Time-dependent computational domain

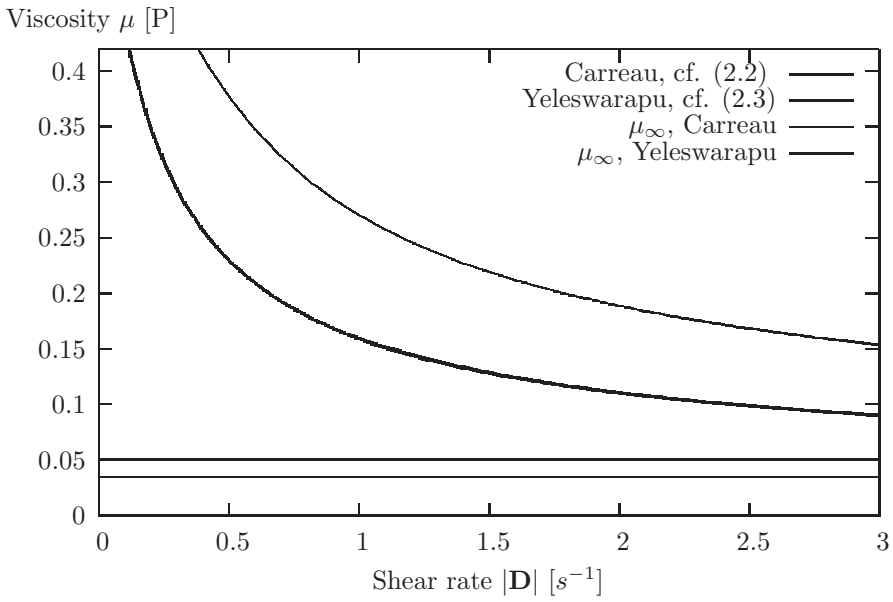


FIGURE 1. Shear-thinning viscosity (2.2), (2.3) for physiological blood parameters.

$$\Omega(\eta(t)) \equiv \{(x_1,x_2):\ -L < x_1 < L,\ 0 < x_2 < R_0(x_1) + \eta(x_1,t)\},\ 0 \le t \le T$$

is given by a reference radius function $R_0(x_1)$ and an unknown free boundary function $\eta(x_1,t)$ describing the domain deformation. For simplicity we will also use a shorter notation $\Omega_t := \Omega(\eta(t))$. We restrict ourselves to two-dimensional domains.

In order to capture movement of a deformable computational domain and preserve the rigidness of inflow and outflow parts, the conservation laws are rewritten using the so-called ALE (**Arbitrary Lagrangian-Eulerian**) mapping $\mathcal{A}_t$, see Fig. 3. It is a continuous bijective mapping from the reference configuration Ω_{ref}, e.g. at time $t = 0$, onto the current one $\Omega_t = \Omega(\eta(t)),\ \mathcal{A}_t : \Omega_{ref} \to \Omega_t$. Introducing the

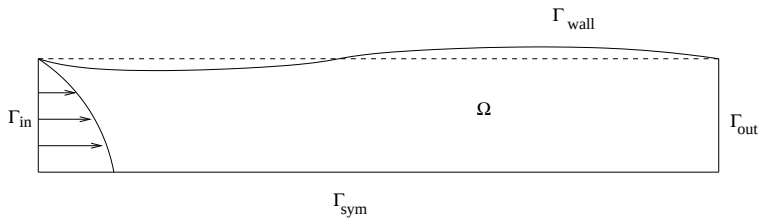


FIGURE 2. Computational domain geometry.

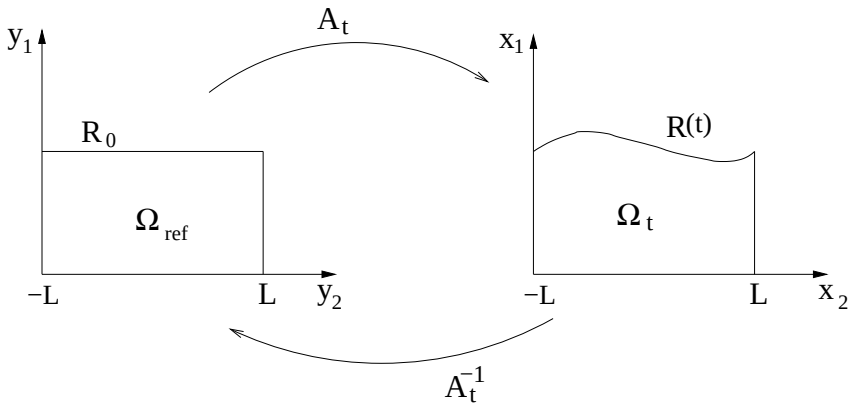


FIGURE 3. ALE-mapping $\mathcal{A}_t$ for a domain with moving boundary.

so-called ALE-derivative

$$\frac{\mathcal{D}^{\mathcal{A}}\boldsymbol{u}(\boldsymbol{x},t)}{\mathcal{D}t} := \frac{\partial \boldsymbol{u}(\boldsymbol{Y},t)}{\partial t}\Big|_{\boldsymbol{Y}=\mathcal{A}^{-1}(\boldsymbol{x})} = \frac{\partial \boldsymbol{u}(\boldsymbol{x},t)}{\partial t}\Big|_{\boldsymbol{x}} + \boldsymbol{w}(\boldsymbol{x},t)\cdot\bigtriangledown\boldsymbol{u}(\boldsymbol{x},t), \quad \boldsymbol{x}\in\Omega_t,\ \boldsymbol{Y}\in\Omega_{ref} \tag{2.4}$$

and the domain velocity $\boldsymbol{w}(\boldsymbol{x},t) := \frac{\partial \mathcal{A}(\boldsymbol{Y})}{\partial t}\Big|_{\boldsymbol{Y}=\mathcal{A}^{-1}(\boldsymbol{x})} = \frac{\partial \boldsymbol{x}}{\partial t}$ for $\boldsymbol{x}\in\Omega_t,\ \boldsymbol{Y}\in\Omega_{ref}$ we rewrite the governing equations (2.1) into a formulation that takes explicitly into account time-dependent behaviour of the domain, i.e.

$$\rho\Big[\frac{\mathcal{D}^{\mathcal{A}}\boldsymbol{u}}{\mathcal{D}t} + ((\boldsymbol{u}-\boldsymbol{w})\cdot\bigtriangledown)\boldsymbol{u}\Big] - \operatorname{div}\Big[2\mu(\mathbf{D}(\boldsymbol{u}))\ \mathbf{D}(\boldsymbol{u})\Big] + \bigtriangledown p = \boldsymbol{f} \quad \operatorname{div}\boldsymbol{u} = 0 \quad \text{on} \quad \Omega(\eta(t)). \tag{2.5}$$

Equation (2.5) is equipped with the initial and boundary conditions

$$\boldsymbol{u} = \boldsymbol{u}_0 \quad \text{with} \quad \operatorname{div}\boldsymbol{u}_0 = 0 \qquad \text{on} \quad \Omega_0, \tag{2.6}$$

$$\left(\mathbf{T}(\boldsymbol{u},p)-\frac{1}{2}|\boldsymbol{u}|^2\mathbf{I}\right)\cdot\boldsymbol{n}=-P_{in}\,\mathbf{I}\cdot\boldsymbol{n},\qquad \text{on}\quad \Gamma_{in},\ t\in(0,T),\tag{2.7}$$

$$\left(\mathbf{T}(\boldsymbol{u},p)-\frac{1}{2}|\boldsymbol{u}|^2\mathbf{I}\right)\cdot\boldsymbol{n}=-P_{out}\,\mathbf{I}\cdot\boldsymbol{n},\qquad \text{on}\quad \Gamma_{out},\ t\in(0,T),\tag{2.8}$$

$$\frac{\partial u_1}{\partial x_2}=0,\quad u_2=0,\qquad \text{on}\quad \Gamma_{sym},\ t\in(0,T).\tag{2.9}$$

Conditions (2.7) and (2.8) are called the kinematical pressure conditions. The fluid velocity is coupled with the velocity of wall deformation by the so-called kinematical coupling condition

$$\boldsymbol{u}=\boldsymbol{w}:=\left(0,\frac{\partial\eta}{\partial t}\right)^T\quad \text{on}\quad \Gamma_{wall},\ t\in(0,T).\tag{2.10}$$

3. Generalized string model for the wall deformation

In order to model biological structure several models have been proposed in literature. For example, to model flow in a collapsible tubes a two-dimensional thin shell model can be used, see results of Wall et al. [16]. Recently Čanič et al. [6] developed a new one-dimensional model for arterial walls, the linearly viscoelastic cylindrical Koiter shell model, that is closed and rigorously derived by energy estimates, asymptotic analysis and homogenization techniques. The viscous fluid dissipation imparts long-term viscoelastic memory effects represented by higher order derivatives.

In the present work we will consider the *generalized string model* for vessel wall deformation. The original generalized string model, see [40], was valid only for radially symmetric domains with a constant reference radius R_0. In order to model stenotic occlusions we will extend this model and assume that the reference radius at rest R_0 depends on the longitudinal variable.

Let us consider a three-dimensional radially symmetric domain. We assume that the deformations are only in the radial direction and set $x_1 = z$ - longitudinal direction and $x_2 = r$ - radial direction. The radial wall displacement, constant with respect to the angle θ, is defined as

$$\eta(z,t)=R(z,t)-R_0(z),\quad z\in(-L,L),\ t\in(0,T),$$

where $R(z,t)$ is the actual radius and $R_0(z)$ is the reference radius at rest. Since the actual radius of the compliant tube is given by $R(z,t) = R_0(z) + \eta(z,t)$, the reference radius R_0 and the actual radius R coincides for fixed solid domains and are dependent only on the spatial variable z. The assumption of radially symmetric geometry and radial displacement allow us to approximate the length of arc in the reference configuration by $dc_0 \approx R_0 d\theta$ and the length of the deformed arc as $dc \approx Rd\theta$, see Fig. 4 and also [40]. Further, we assume that the gradient of displacement $\partial_z\eta$ is small, which implies the linear constitutive law (linear elasticity) of the vessel wall. The wall thickness is assumed to be small and constant. Moreover we approximate the infinitesimal surface S of Γ_{wall} in the following way $S \approx dc\, dl$.

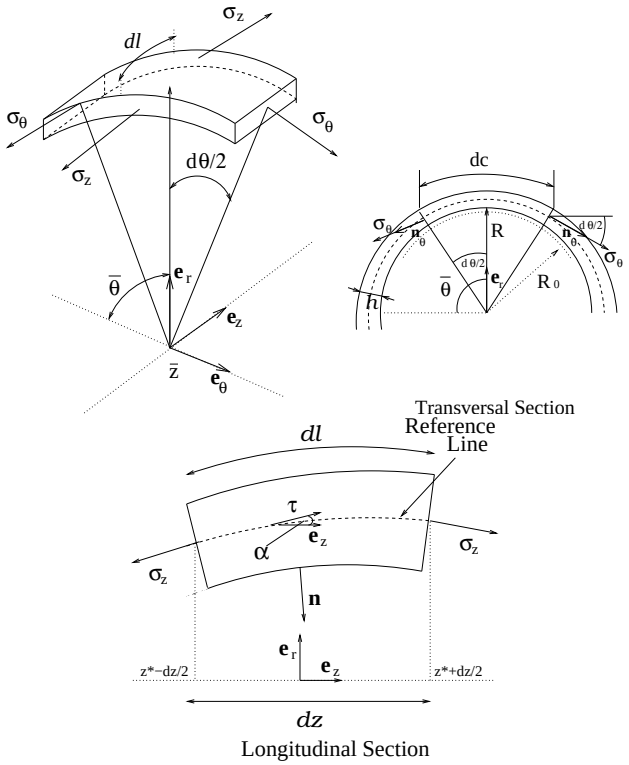


FIGURE 4. Small portion of vessel wall with physical characteristics [40].

The linear momentum law: $Force = Mass \times Acceleration$ is applied in the radial direction to obtain the equation for η.

$$Mass = \rho_w \hbar \, dc \, dl, \qquad Acceleration = \frac{\partial^2 R(z,t)}{\partial t^2} = \frac{\partial^2 \eta(z,t)}{\partial t^2}, \tag{3.1}$$

where ρ_w is the density of the wall and $\hbar$ its thickness.

Now we evaluate forces acting on the vessel wall. The tissue surrounding the vessel wall interacts with the vessel wall by exerting a constant pressure P_w. The resulting tissue force is $\boldsymbol{f}_{tissue} = -P_w \boldsymbol{n} \, dc \, dl \approx -P_w \boldsymbol{n} R \, d\theta \, dl$.

The forces from the fluid on Γ_{wall} are represented by the normal component of the Cauchy stress tensor $\boldsymbol{f}_{fluid} = -\mathbf{T}\boldsymbol{n} \, dc \, dl$, $\mathbf{T} = -p\mathbf{I} + 2\mu(\mathbf{D}(\boldsymbol{u}))\mathbf{D}(\boldsymbol{u})$. By summing the tissue and fluid forces we get the resulting external force acting on the vessel wall along the radial direction ($\boldsymbol{f}_{ext} = \boldsymbol{f}_{tissue} + \boldsymbol{f}_{fluid}$):

$$f_{ext}\Big|_{\Gamma_{wall}} = \boldsymbol{f}_{ext}\Big|_{\Gamma_{wall}} \cdot \boldsymbol{e}_r = (-\mathbf{T} - P_w\mathbf{I})\boldsymbol{n}\Big|_{\Gamma_{wall}} \cdot \boldsymbol{e}_r \, dc \, dl,$$

where $\boldsymbol{e}_r$ is the unit vector in the radial direction and $\boldsymbol{n} = \frac{1}{\sqrt{1+(\partial_z R)^2}}(-\partial_z R, 1)$ the unit outward normal to the boundary Γ_{wall}. We transform this force from the

Eulerian to the Lagrangian coordinates, see, e.g., [24] for more details.

$$f_{ext}\Big|_{\Gamma^0_{wall}} = -(\tilde{\mathbf{\Pi}} + \tilde{P}_w\mathbf{I})\tilde{\boldsymbol{n}}\Big|_{\Gamma^0_{wall}} \cdot \boldsymbol{e}_r \frac{R\sqrt{1+(\partial_z R)^2}}{R_0\sqrt{1+(\partial_z R_0)^2}} dc_0\, dl_0,$$

here $\boldsymbol{n}(x) = \tilde{\boldsymbol{n}}(\tilde{x})$, $x = (z, R(z,t)) \in \Gamma_{wall}, \tilde{x} = (z, R_0(z)) \in \Gamma^0_{wall}$. The term $\frac{R\sqrt{1+(\partial_z R)^2}}{R_0\sqrt{1+(\partial_z R_0)^2}}$ arrives from the transformation to the Lagranian coordinates, in particular we have the transformation of the curve $\Gamma_{wall}(t) = \{(z, R(z,t)), z \in (-L, L)\}$ to the curve $\Gamma^0_{wall} = \{(z, R_0(z)), z \in (-L, L)\}$.

The internal forces acting on the vessel portion are due to the circumferential stress $\boldsymbol{\sigma}_\theta$ (constant with respect to the angle) and the longitudinal stress $\boldsymbol{\sigma}_z$. Both stresses are directed along the normal to the surface to which they act. Let us denote $\sigma_\theta = \boldsymbol{\sigma}_\theta \cdot \boldsymbol{n}$. Further the longitudinal stress $\boldsymbol{\sigma}_z$ is parallel to tangent, i.e. $\boldsymbol{\sigma}_z = \pm\sigma_z\boldsymbol{\tau}$. The sign is positive if the versus of the normal to the surface, on which $\boldsymbol{\sigma}_z$ is acting, is the same as those chosen for $\boldsymbol{\tau}$.

We have $f_{int} = (\boldsymbol{f}_\theta + \boldsymbol{f}_z) \cdot \boldsymbol{e}_r$ and

$$\begin{aligned}
\boldsymbol{f}_\theta \cdot \boldsymbol{e}_r &= \Big[\boldsymbol{\sigma}_\theta\Big(\bar{\theta} + \frac{d\theta}{2}\Big) + \boldsymbol{\sigma}_\theta\Big(\bar{\theta} - \frac{d\theta}{2}\Big)\Big] \cdot \boldsymbol{e}_r \hbar\, dl = 2|\sigma_\theta| cos\Big(\frac{\pi}{2} + \frac{d\theta}{2}\Big)\hbar\, dl \\
&= -2|\sigma_\theta| \sin\left(\frac{d\theta}{2}\right)\hbar\, dl \approx -|\sigma_\theta|\hbar\, d\theta\, dl = -E\frac{\eta}{R_0}\hbar\, d\theta\, dl, \\
\boldsymbol{f}_z \cdot \boldsymbol{e}_r &= \Big[\boldsymbol{\sigma}_z\Big(z^* + \frac{dz}{2}\Big) + \boldsymbol{\sigma}_z\Big(z^* - \frac{dz}{2}\Big)\Big] \cdot \boldsymbol{e}_r \hbar\, dc \\
&= \frac{\boldsymbol{\tau}(z^* + \frac{dz}{2}) - \boldsymbol{\tau}(z^* - \frac{dz}{2})}{dz} \cdot \boldsymbol{e}_r \hbar|\sigma_z|\, dz\, dc \\
&\approx |\sigma_z|\Big[\frac{d\boldsymbol{\tau}}{dz}(z^*)\Big] \cdot \boldsymbol{e}_r \hbar\, dz\, dc \\
&\approx \left(\frac{\partial^2\eta}{\partial z^2} + \frac{\partial^2 R_0}{\partial z^2}\right)\left[1 + \left(\frac{\partial R_0}{\partial z}\right)^2\right]^{-1} \boldsymbol{n} \cdot \boldsymbol{e}_r|\sigma_z|\hbar dz\, dc.
\end{aligned}$$

Here we have used the following properties. According to the linear elasticity assumption the stress tensor $\boldsymbol{\sigma}_\theta$ is proportional to the relative circumferential prolongation, i.e.

$$\boldsymbol{\sigma}_\theta = E\frac{2\pi(R - R_0)}{2\pi R_0} = E\frac{\eta}{R_0}, \quad E \text{ is Young's modulus of elasticity.}$$

To evaluate the longitudinal force we have used the following result, that is a generalization of an analogous lemma from [40].

LEMMA 3.1. *If $\frac{\partial\eta}{\partial z}$ is small then*

$$\frac{d\boldsymbol{\tau}}{dz}(z^*) \approx \left(\frac{\partial^2\eta}{\partial z^2} + \frac{\partial^2 R_0}{\partial z^2}\right)\left[1 + \left(\frac{\partial R_0}{\partial z}\right)^2\right]^{-1} \boldsymbol{n}.$$

PROOF. Let a parametric curve $\mathbf{c}$ be defined at each t on the plane (z, r) by

$$\mathbf{c} : \mathbb{R} \to \mathbb{R}^2,\ z \to (c_1(z), c_2(z)) = (z, R(z,t)) = (z, R_0(z,t) + \eta(z,t)),$$

and $\boldsymbol{\tau}$, $\boldsymbol{n}$, κ denote the tangent, the normal and the curvature of $\mathbf{c}$, respectively. Then according to the Serret-Frenet formula [40] we have

$$\frac{d\tau}{dz}(z) = \left|\frac{d\mathbf{c}}{dz}(z)\right| \kappa(z)\tilde{\boldsymbol{n}}(z).$$

Here $\tilde{\boldsymbol{n}} = \pm\boldsymbol{n}$ is the normal oriented towards the center of curvature. Furthermore since we assume $\frac{\partial\eta}{\partial z}$ to be small, we have

$$\begin{aligned}
\left|\frac{d\mathbf{c}}{dz}(z)\right| &= \left[1+\left(\frac{\partial R}{\partial z}\right)^2\right]^{1/2} \approx \left[1+\left(\frac{\partial R_0}{\partial z}\right)^2\right]^{1/2} \quad \text{and} \\
\kappa &= \left|\frac{dc_1}{dz}\frac{d^2c_2}{dz^2} - \frac{dc_2}{dz}\frac{d^2c_1}{dz^2}\right| \left|\frac{d\mathbf{c}}{dz}\right|^{-3} = \left|\frac{\partial^2 R}{\partial z^2}\right| \left[1+\left(\frac{\partial R}{\partial z}\right)^2\right]^{-\frac{3}{2}} \\
&\approx \left|\frac{\partial^2 R_0 + \partial^2\eta}{\partial z^2}\right| \left[1+\left(\frac{\partial R_0}{\partial z}\right)^2\right]^{-\frac{3}{2}}.
\end{aligned}$$

Since the sign of $\frac{\partial^2 R}{\partial z^2}$ determines the convexity of curve, $\tilde{\boldsymbol{n}} = sign\left(\frac{\partial^2 R}{\partial z^2}\right)\boldsymbol{n}$, we obtain the desired result. □

Now we use the assumption of the incompressibility of material; the volume of the infinitesimal portion remains constant under the deformation: $\hbar dc dl = \hbar dc_0 dl_0$. Using this assumption the internal forces can be expressed as

$$f_{int} \approx \left\{-E\frac{\eta}{RR_0} + \left(\frac{\partial^2\eta}{\partial z^2} + \frac{\partial^2 R_0}{\partial z^2}\right)\left[1+\left(\frac{\partial R_0}{\partial z}\right)^2\right]^{-1} \boldsymbol{n}\cdot\boldsymbol{e}_r|\sigma_z|\frac{dz}{dl}\right\}\hbar\, dc_0\, dl_0.$$

Moreover, we use the fact that $\boldsymbol{n}\cdot\boldsymbol{e}_r = 1/\sqrt{1+(\partial_z R)^2} \approx 1/\sqrt{1+(\partial_z R_0)^2}$, and

$$\frac{dz}{dl} \approx \cos(\measuredangle(\boldsymbol{e}_z, \boldsymbol{\tau})) = \boldsymbol{e}_z\cdot\boldsymbol{\tau} \approx 1/\sqrt{1+(\partial_z R_0)^2},$$

compare Fig. 4.

Summing up all contributions of balancing forces acting on the infinitesimal portion of Γ_{wall} we obtain from the linear momentum law (3.1) using the transformation to Γ^0_{wall}

$$\left\{\rho_w\hbar\frac{\partial^2\eta}{\partial t^2} - |\sigma_z|\frac{\left(\frac{\partial^2\eta}{\partial z^2}+\frac{\partial^2 R_0}{\partial z^2}\right)}{\left[1+\left(\frac{\partial R_0}{\partial z}\right)^2\right]^2}\hbar + E\hbar\frac{\eta}{R_0 R} + (\tilde{\mathbf{T}} + \tilde{P_w}\mathbf{I})\tilde{\boldsymbol{n}}\cdot\boldsymbol{e}_r\frac{R\sqrt{1+\left(\frac{\partial(R_0+\eta)}{\partial z}\right)^2}}{R_0\sqrt{1+(\frac{\partial R_0}{\partial z})^2}}\right\} R_0 d\theta\, dl_0 = o(d\theta\, dl_0).$$

Thus by dividing the above equation by $\rho_w\hbar\, R_0\, d\theta\, dl_0$ and passing to the limit for $d\theta \to 0$, $dl_0 \to 0$ we obtain the so called *vibrating string model*. Adding a damping term $-c\partial^3_{tzz}\eta$ (or $-c\partial^5_{tzzzz}\eta$) $c > 0$ at the left hand side we get the *generalized string model* for radially symmetric domains with non-constant reference radius $R_0(z)$

$$\left[\frac{\partial^2\eta}{\partial t^2} - \frac{|\sigma_z|}{\rho_w}\frac{\left(\frac{\partial^2\eta}{\partial z^2}+\frac{\partial^2 R_0}{\partial z^2}\right)}{\left[1+(\partial_z R_0)^2\right]^2} + \frac{E\eta}{\rho_w R_0(R_0+\eta)} - c\frac{\partial^3\eta}{\partial t\partial^2 z}\right](z,t)$$

$$= \left[-(\tilde{\mathbf{T}}+\tilde{P_w}\mathbf{I})\tilde{\boldsymbol{n}}\right](z,R_0(z))\cdot\boldsymbol{e}_r\frac{(R_0+\eta)(z,t)}{R_0(z)\rho_w\hbar}\frac{\sqrt{1+(\partial_z R_0+\partial_z\eta)^2}}{\sqrt{1+(\partial_z R_0)^2}}. \quad (3.2)$$

The generalized string model for structure (3.2) is completed with the initial and boundary conditions

$$\eta = 0, \quad \frac{\partial\eta}{\partial t} = \boldsymbol{u}_0|_{\Gamma^0_{wall}}\cdot\boldsymbol{e}_r \qquad \text{on} \quad \Gamma^0_{wall}, \quad (3.3)$$

$$\eta(-L,t)=\eta_1, \quad \eta(L,t)=\eta_2, \qquad \text{for} \quad t\in(0,T). \quad (3.4)$$

Let us point out that the coupling of fluid and structure is realized by the kinematical and dynamical coupling conditions. The dynamical coupling is represented by the continuity of stresses, i.e. the fluid forces acting on the structure are due to fluid stress tensor at the right hand side of the structure equation (3.2). The kinematical coupling represents the continuity of velocities at the moving boundaries, which is the condition (2.10).

4. Fluid-structure interaction methods

In what follows we describe two numerical schemes for coupling the fluid and the structure. The first approach, called *the global iterative method*, is based on the global iterations with respect to the domain geometry. This method belongs to the strong coupling-type methods. In the second approach, *the kinematical splitting*, the structure equation (3.2) is splitted into two parts, which are solved consequently. Using this splitting, no additional iterations between the fluid and the structure are necessary. The second method belongs to the class of weakly coupled methods.

4.1. Strong coupling: global iterative method. Assume that the domain deformation $\eta = \eta^{(k)}$ is a given function, take $\eta^{(0)} = \eta(\cdot,0)$. The vector $(\boldsymbol{u}^{(k+1)}, p^{(k+1)}, \eta^{(k+1)})$ is obtained as a solution of (2.1), (3.2) for all $x\in\Omega(\eta^{(k)})$, $x_1\in(-L,L)$ and all $t\in(0,T)$. Instead of condition (2.10) we use

$$u_2(x_1,x_2,t) = \frac{\partial\eta^{(k)}}{\partial t}(x_1,t) = w_2(x_1,x_2,t), \;\; u_1(x_1,x_2,t)=0 \qquad \text{on } \Gamma^{(k)}_{wall}(t), \;\; (4.1)$$

where $\Gamma^{(k)}_{wall}(t) = \{(x_1,x_2);\; x_2 = R_0(x_1)+\eta^{(k)}(x_1,t),\; x_1\in(-L,L)\}$, $t\in(0,T)$ and $\boldsymbol{w}$ is the velocity of mesh movement related to smoothing the grid after moving its boundary (we allow just movement in the x_2 direction, x_1 direction is neglected), see also [51].

Further we linearize the equation (3.2) replacing the non-linear term on its left hand side by $E\eta/(\rho_w(R_0+\eta^{(k)})R_0)$. In order to decouple (2.1) or (2.5) and (3.2) we evaluate the forcing term at the right hand side of (3.2) at the old time step t^{n-1}, see also Fig 6. Convergence of this global method was verified experimentally. Our extensive numerical experiments show fast convergence of domain

deformation, two iteration of domain deformation differ about $10^{-4}cm$ (for e.g., $R_0 = 1cm$) pointwisely after few, about 5 iterations. As an example we have depicted in Figure 3 a deformed vessel wall after 1, 2, 3 and 9 global iterations at the same time $T = 0.36s$. It illustrates that the vessel wall converges to one curve and does not change significantly already after second iteration, see Fig. 5. Theoretical proof of the convergence $\eta^{(k)} \to \eta$ can be obtained by means of the Schauder fixed point theorem, cf. [25] and the following subsection.

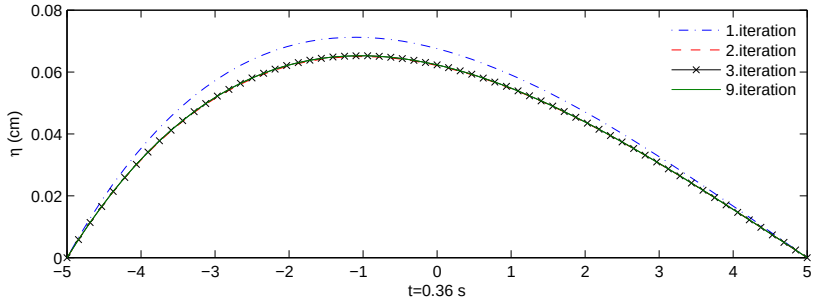


FIGURE 5. Several iterations of the wall deformation η at time $t = 0.36s$, after a few iterations curves coincide. Computed for the Carreau model with $Re = 40$, cf. (1.3).

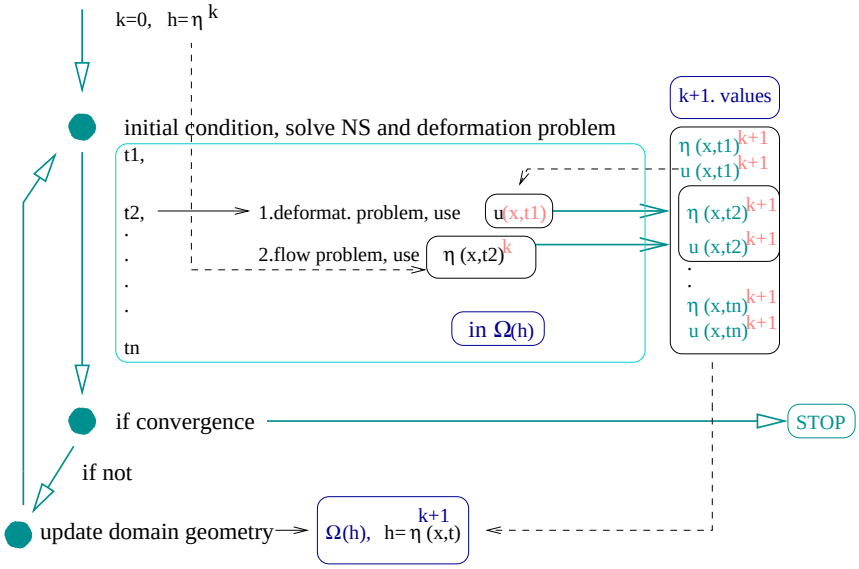


FIGURE 6. The sketch of the global iterative method.

4.1.1. *Existence of a weak solution to the coupled problem.* In the recent years the well-posedness of fluid-structure interaction is being extensively studied. In particular, the well-possedness of the mathematical model describing the Newtonian fluid flow in compliant vessels has been studied in [6, 7, 13, 23, 25, 47], see also [4, 5] for related results. In [47] local existence in time of strong solutions is shown, provided the initial data are sufficiently small. Cheng, Shkoller and Coutand [9, 11] studied coupled problem consisting of viscous incompressible fluid and elastic solid shell. Mathematically, the shell encloses the fluid and creates a time-dependent boundary of viscous fluid. In [9] Cheng and Shkoller proved local (in time) existence and uniqueness of regular solutions. For three-dimensional problems in addition some smallness of shell thickness has to be assumed. The difficulty of the coupled model lies in a parabolic-hyperbolic coupling of viscous fluid and hyperbolic structure. A new idea presented in their recent works [10, 11, 12] is based on introducing a functional framework that scales in a hyperbolic fashion and is thus driven by the elastic structure. The problem has been reformulated in the Lagrangian coordinates.

In contrast to these local existence and uniqueness results for regular solutions we can find already several results on global existence of weak solutions. Recently, Padula et al. [23] showed the global existence in time of weak solutions when initial data are sufficiently close to equilibrium. If no restriction on initial data is assumed, then weak solutions exist as long as the elastic wall does not touch the rigid bottom [23]. In [22] uniqueness and continuous dependence on initial data for weak solutions has been studied.

Similarly, in [7] Chambole et al. proved the global existence of weak solutions until a contact of the viscoelastic and the rigid boundary, see also [20] for the existence result of stationary solution and the elastic Saint-Venant-Kirchhoff material and [44] for related results.

In [25] we have proved the existence of a weak solution of fully coupled fluid-structure interaction problem between the non-Newtonian shear-thickening fluid and linear viscoelastic structure. In order to obtain enough regular η we need to regularize the structure equation (3.2) with η_{txxxx} instead of η_{txx}.

Now, assuming that η is enough regular (see below) and taking into account the results from [7] we can define functional spaces that give sense to the trace of velocity from $W^{1,p}(\Omega(\eta(t)))$ and thus define the weak solution of the problem. We assume that $R_0 \in C_0^2(0, L)$. Note that p is the exponent of the polynomial viscosity function, see, e.g., (2.2). In [25] more general non-Newtonian models with a polynomially growing potential of the stress tensor have been analyzed as well.

Definition 4.1 (Weak formulation). We say that $(\boldsymbol{u}, \eta)$ is a weak solution of (2.1), (3.2), (2.10) with the initial and boundary conditions (2.6), (2.7), (2.8), (2.9), (3.3), (3.4) on $[0, T)$ if the following conditions hold

- $\boldsymbol{u} \in L^p(0, T; W^{1,p}(\Omega(\eta(t)))) \cap L^\infty(0, T; L^2(\Omega(\eta(t))))$
- $\eta \in W^{1,\infty}(0, T; L^2(-L, L)) \cap H^1(0, T; H_0^2(-L, L))$
- $\operatorname{div} \boldsymbol{u} = 0$ a.e. on $\Omega(\eta(t))$
- $\boldsymbol{u} = (0, \eta_t)$ for a.e. $\boldsymbol{x} \in \Gamma_w(t),\ t \in (0, T)$,

$$\int_0^T \int_{\Omega(\eta(t))} \left\{ -\rho \boldsymbol{u} \cdot \frac{\partial \boldsymbol{\varphi}}{\partial t} + 2\mu(|(\boldsymbol{u})|)\mathbf{D}(\boldsymbol{u})\mathbf{D}(\boldsymbol{\varphi}) + \rho \sum_{i,j=1}^{2} u_i \frac{\partial u_j}{\partial x_i} \boldsymbol{\varphi}_j \right\} d\boldsymbol{x}\, dt$$
$$+ \int_0^T \int_0^{R_0(L)} \left(P_{out} - \frac{\rho}{2} |u_1|^2 \right) \varphi_1(L, x_2, t)\, dx_2\, dt \tag{4.2}$$
$$- \int_0^T \int_0^{R_0(0)} \left(P_{in} - \frac{\rho}{2} |u_1|^2 \right) \varphi_1(-L, x_2, t)\, dx_2\, dt$$
$$+ \int_0^T \int_{-L}^{L} P_w \varphi_2(x_1, R_0(x_1) + \eta(x_1, t), t) - a \frac{\partial^2 R_0}{\partial x_1^2} \xi\, dx_1\, dt$$
$$+ \int_0^T \int_{-L}^{L} -\frac{\partial \eta}{\partial t} \frac{\partial \xi}{\partial t} + c\, \frac{\partial^3 \eta}{\partial x_1^2 \partial t} \frac{\partial^2 \xi}{\partial x_1^2} + a \frac{\partial \eta}{\partial x_1} \frac{\partial \xi}{\partial x_1} + b\eta\, \xi\, dx_1\, dt \;=\; 0$$

for every test functions

$$\begin{aligned} &\boldsymbol{\varphi}(x_1, x_2, t) \in H^1(0, T; W^{1,p}(\Omega(\eta(t)))) \text{ such that} \\ &\operatorname{div} \boldsymbol{\varphi} = 0 \text{ a.e on } \Omega(\eta(t)), \\ &\varphi_2|_{\Gamma_w(t)} \in H^1(0, T; H_0^2(\Gamma_w(t))) \quad \text{and} \\ &\xi(x_1, t) = \tilde{E} \rho\, \boldsymbol{\varphi}_2(x_1, R_0(x_1) + \eta(x_1, t), t), \end{aligned}$$

where $\tilde{E}$ is a given constant depending on the structural material properties.

THEOREM 4.2 (Existence of a weak solution [25]). *Let $p \geq 2$. Assume that the boundary data fulfill $P_{in} \in L^{p'}(0, T; L^2(0, R_0(0)))$, $P_{out} \in L^{p'}(0, T; L^2(0, R_0(L)))$, $P_w \in L^{p'}(0, T; L^2(-L, L))$, $\frac{1}{p} + \frac{1}{p'} = 1$. Furthermore, assume that the viscous stress tensor τ has a potential $\mathcal{U} \in C^2(\mathbb{R}^{2\times 2})$ satisfying the following conditions*

$$\frac{\partial \mathcal{U}(\eta)}{\partial \eta_{ij}} = \tau_{ij}(\eta) \tag{4.3}$$
$$\mathcal{U}(\mathbf{0}) = \frac{\partial \mathcal{U}(\mathbf{0})}{\partial \eta_{ij}} = 0 \tag{4.4}$$
$$\frac{\partial^2 \mathcal{U}(\eta)}{\partial \eta_{mn} \partial \eta_{rs}} \xi_{mn} \xi_{rs} \geq C_1 (1 + |\eta|)^{p-2} |\xi|^2 \tag{4.5}$$
$$\left| \frac{\partial^2 \mathcal{U}(\eta)}{\partial \eta_{ij} \partial \eta_{kl}} \right| \leq C_2 (1 + |\eta|)^{p-2}. \tag{4.6}$$

Then there exists a weak solution $(\boldsymbol{u}, \eta)$ of the problem (2.1), (3.2), (2.10) with the initial and boundary conditions (2.6), (2.7), (2.8), (2.9), (3.3), (3.4) such that

(i) $\boldsymbol{u} \in L^p(0, T; W^{1,p}(\Omega(\eta(t)))) \cap L^\infty(0, T; L^2(\Omega(\eta(t))))$,
$\eta \in W^{1,\infty}(0, T; L^2(-L, L)) \cap H^1(0, T; H_0^2(-L, L))$,
(ii) $\boldsymbol{u} = (0, \eta_t)$ *for a.e.* $\boldsymbol{x} \in \Gamma_w(t),\ t \in (0, T)$,
(iii) $\boldsymbol{u}$ *satisfies the condition* $div\, \boldsymbol{u} = 0$ *a.e on* $\Omega(\eta(t))$ *and (4.2) holds.*

The proof of existence is realized in several steps:

a) Approximation of the solenoidal spaces on a moving domain by the artificial compressibility approach: ε - approximation

$$\varepsilon\left(\frac{\partial p_\varepsilon}{\partial t}-\Delta p_\varepsilon\right)+\operatorname{div}\boldsymbol{v}_\varepsilon=0\quad\text{in }\Omega(\eta^{(k)}),\tag{4.7}$$
$$\frac{\partial p_\varepsilon}{\partial \boldsymbol{n}}=0\ \text{ on }\ \partial\Omega(\eta^{(k)}),\quad \varepsilon>0.$$

b) Splitting the boundary conditions (2.10), (3.2) by introducing the semipervious boundary: κ - approximation.

$$\left[\mu(|e(\boldsymbol{v})|)\left\{-\left(\frac{\partial v_2}{\partial x_1}+\frac{\partial v_1}{\partial x_2}\right)\frac{\partial h}{\partial x_1}+2\frac{\partial v_2}{\partial x_2}\right\}-p+P_w\right](\bar{x},t)\tag{4.8}$$
$$-\frac{\rho}{2}v_2\Big(v_2(\bar{x},t)-\frac{\partial \eta^{(k)}}{\partial t}(x_1,t)\Big)=\rho\kappa\Big(\frac{\partial\eta}{\partial t}(x_1,t)-v_2(\bar{x},t)\Big)$$

and

$$-\tilde{E}\left[\frac{\partial^2\eta}{\partial t^2}-a\frac{\partial^2\eta}{\partial x_1^2}+b\eta+c\frac{\partial^5\eta}{\partial t\partial x_1^4}-a\frac{\partial^2 R_0}{\partial x_1^2}\right](x_1,t)\tag{4.9}$$
$$=\kappa\Big(\frac{\partial\eta}{\partial t}(x_1,t)-v_2(\bar{x},t)\Big)\qquad \bar{x}=(x_1,\eta^{(k)}(x_1,t)),\ x_1\in(-L,L)$$

with $\kappa\gg 1$. For finite κ the boundary Γ_w is partly permeable, but letting $\kappa\to\infty$ it becomes impervious. In fact, we can prove the existence of solution if $\kappa\to\infty$ and thus we get the original boundary condition.

c) Transformation of the weak formulation on a time dependent domain $\Omega(\eta(t))$ to a fixed reference domain $D=(-L,L)\times(0,1)$ using a given domain deformation $\eta=\eta^{(k)}$: k - approximation.

The (κ,ε)-approximated problem is defined on a moving domain depending on function $h=R_0+\eta^{(k)}$. We will transform it to a fixed rectangular domain and set

$$\begin{aligned}\boldsymbol{v}(y_1,y_2,t)&\overset{\text{def}}{=}\boldsymbol{u}(y_1,h(y_1,t)y_2,t)\\ q(y_1,y_2,t)&\overset{\text{def}}{=}\rho^{-1}p(y_1,h(y_1,t)y_2,t)\\ \sigma(y_1,t)&\overset{\text{def}}{=}\frac{\partial\eta}{\partial t}(y_1,t)\end{aligned}\tag{4.10}$$

for $y\in D=\{(y_1,y_2);\ -L<y_1<L,\ 0<y_2<1\}$, $0<t<T$.

d) Limiting process for $\varepsilon\to 0$, $\kappa\to\infty$ and $k\to\infty$, respectively.

We firstly show the existence of weak solutions of stationary problems obtained by time discretization. Furthermore, we derive suitable a priori estimates for piecewise approximations in time. By using the theory of monotone operators, the Minty-Browder theorem and the compactness arguments due to the Lions-Aubin lemma, we can show the convergence of time approximations to its weak unsteady solution. Thus we obtain the existence of a weak solution to the (κ,ε,k) - approximate problem. The next step are the limiting processes for κ and ε. First of all we show the limiting process in $\varepsilon\to 0$ since necessary a priori estimates obtained by means of the energy method are independent on ε. In order to realize the limiting

process in κ; $\kappa \to \infty$, we however need new a priori estimates and show the semi-continuity in time. Thus, letting $\varepsilon \to 0$ and $\kappa \to \infty$ we obtain the existence of weak solution to the $k-$approximate problem depending only on the approximation of domain deformation $h(x_1,t) = R_0(x_1) + \eta^{(k)}(x_1,t)$. The final limiting process with respect to the domain deformation, i.e. for $k \to \infty$ will be realized by the Schauder fixed point arguments for a regularized problem and consequently by passing to the limit with the regularizing parameter. This will yield the existence of at least one weak solution of the fully coupled unsteady fluid-structure interaction between the non-Newtonian shear-dependent fluid and the viscoelastic string.

The existence result from [25] is the generalization of the results of Filo and Zaušková [13] where the Newtonian fluids were considered. In [13] the generalized string equation with a third order regularizing term was considered, but the final limiting step for $k \to \infty$ was open.

4.2. Weak coupling: kinematical splitting algorithm. First, let us rewrite the generalized string model (3.2) in the following way

$$\frac{\partial^2 \eta}{\partial t^2} - a\frac{\partial^2 \eta}{\partial x_1^2} + b\eta - c\frac{\partial^3 \eta}{\partial t \partial x_1^2} = -\frac{(\mathbf{T} + P_w\ \mathbf{I}) \cdot \boldsymbol{n} \cdot \boldsymbol{e}_r}{\rho_w \hbar} + a\frac{\partial^2 R_0}{\partial x_1^2} \quad \text{on} \quad \Gamma_{wall}(t) \tag{4.11}$$

or

$$\frac{\partial^2 \eta}{\partial t^2} - a\frac{\partial^2 \eta}{\partial x_1^2} + b\eta - c\frac{\partial^3 \eta}{\partial t \partial x_1^2} = -\frac{(\tilde{\mathbf{T}} + \tilde{P}_w\ \mathbf{I}) \cdot \tilde{\boldsymbol{n}} \cdot \boldsymbol{e}_r}{\rho_w \hbar} \frac{R}{R_0} \frac{\sqrt{1 + (\partial_{x_1} R)^2}}{\sqrt{1 + (\partial_{x_1} R_0)^2}} + a\frac{\partial^2 R_0}{\partial x_1^2} \quad \text{on} \quad \Gamma^0_{wall}\ . \tag{4.12}$$

Here the parameters are defined as follows

$$a = \frac{|\sigma_{x_1}|}{\rho_w}\left[1 + \left(\frac{\partial R_0}{\partial x_1}\right)^2\right]^{-2}, \quad b = \frac{E}{\rho_w R_0 (R_0 + \eta)}, \quad c = \frac{\gamma}{\rho_w \hbar}. \tag{4.13}$$

Recall that E is the Young modulus, $\hbar$ the thickness of the vessel wall, ρ_w its density, γ is a positive viscoelastic constant and $|\sigma_{x_1}|$ magnitude of the stress tensor component in the longitudinal direction, cf. also Subsection 1.3 for typical physiological values. The kinematical splitting algorithm is based on the kinematical coupling condition

$$\boldsymbol{u} = \boldsymbol{w} := \left(0, \frac{\partial \eta}{\partial t}\right) \quad \text{on} \quad \Gamma^0_{wall} \tag{4.14}$$

and special splitting of the structure equation into the hyperbolic and parabolic part. We define the operator A that includes the fluid solver for (2.5) and the viscoelastic part of structure equation

$$\textbf{A operator (hydrodynamic)} \begin{cases} \text{fluid solver } (\boldsymbol{u}, p), \\ \xi := u_2|_{\Gamma_{wall}}, \\ \dfrac{\partial \xi}{\partial t} = c\dfrac{\partial^2 \xi}{\partial {x_1}^2} + H(\boldsymbol{u}, p) \end{cases} \tag{4.15}$$

and the operator B for purely elastic load of the structure

$$\textbf{B operator (elastic)} \begin{cases} \dfrac{\partial \eta}{\partial t} = \xi, \\ \dfrac{\partial \xi}{\partial t} = a\dfrac{\partial^2 \eta}{\partial {x_1}^2} - b\eta + H(R_0), \end{cases} \tag{4.16}$$

where

$$H(\boldsymbol{u}, p) := -\frac{(\tilde{\mathbf{T}} + \tilde{P}_w \mathbf{I}) \cdot \tilde{\boldsymbol{n}} \cdot \boldsymbol{e}_r}{\rho_w \hbar} \frac{(R_0 + \eta)}{R_0} \frac{\sqrt{1 + (\partial_{x_1} R)^2}}{\sqrt{1 + (\partial_{x_1} R_0)^2}}, \quad H(R_0) := a\frac{\partial^2 R_0}{\partial x_1^2} \ . \tag{4.17}$$

Here we note that the coupling condition allowed us to rewrite the hydrodynamic part of structure equation in the terms of wall velocity ξ. Time discretization of our problem is done in the following way: from the fluid equation we compute new velocities $\boldsymbol{u}^{n+1}$ and pressures p^{n+1} for $x^n \in \Omega^n$ (i.e. Ω_t for $t = t^n$). Note that $\tilde{\boldsymbol{u}}^{n+1} = \boldsymbol{u}^{n+1} \circ \mathcal{A}_{t^{n+1}} \circ \mathcal{A}_{t^n}^{-1}$ and $\tilde{p}^{n+1} = p^{n+1} \circ \mathcal{A}_{t^{n+1}} \circ \mathcal{A}_{t^n}^{-1}$, where $\mathcal{A}_{t^n}$ is the ALE-mapping from a reference domain Ω_{ref} onto Ω^n. Then we continue with computing of the wall velocity $\xi^{n+\frac{1}{2}}$ from the hydrodynamic part of structure equation (4.15). Further on we proceed with the operator B and compute new wall displacement η^{n+1} and new wall velocity ξ^{n+1}. Finally, knowing η^{n+1} the geometry is updated from Ω^n to Ω^{n+1} and new values of fluid velocity $\boldsymbol{u}^{n+1}$ and pressure p^{n+1} are transformed onto Ω^{n+1}. In order to update the domain Ω^n we need to define the grid velocity $\boldsymbol{w}$. First, we set $\boldsymbol{w}|_{\Gamma_{wall}} = \xi^{n+1}$. In order to prescribe the grid velocity also inside Ω we can solve an auxiliary problem, cf., e.g., [15] or interpolate $\boldsymbol{w}$. Consequently, we get $\boldsymbol{w}^{n+1} = \partial \boldsymbol{x}/\partial t$, $\boldsymbol{x} \in \Omega^{n+1}$.

4.2.1. *Stability analysis.* In what follows we will briefly describe stability analysis of the semi-discrete scheme for the kinematical coupling approach. More details on the derivation can be found in [28]. Now, let us consider the weak formulation of the fluid equation and set for the test function $\boldsymbol{u}$. Integrating over Ω^n and approximating the time derivative by the backward Euler method the operator A yields the following equation for new intermediate velocities $\tilde{\boldsymbol{u}}^{n+1}$, $\xi^{n+\frac{1}{2}}$

$$\begin{aligned} &\int_{\Omega^n} \tilde{\boldsymbol{u}}^{n+1} \cdot \frac{\tilde{\boldsymbol{u}}^{n+1} - \boldsymbol{u}^n}{\Delta t} \, \mathrm{d}\omega + \frac{2}{\rho} \int_{\Omega^n} \mu(|\mathbf{D}(\tilde{\boldsymbol{u}}^{n+1})|) \, \mathbf{D}(\tilde{\boldsymbol{u}}^{n+1}) : \mathbf{D}(\tilde{\boldsymbol{u}}^{n+1}) \, \mathrm{d}\omega \\ &+ \frac{1}{2} \int_{\Omega^n} |\tilde{\boldsymbol{u}}^{n+1}|^2 \mathrm{div} \, \boldsymbol{w}^n \, \mathrm{d}\omega \quad = \quad -\rho_w \hbar \int_{\Gamma^0_{wall}} \left[\frac{\xi^{n+\frac{1}{2}} - \xi^n}{\Delta t} \right] \xi^{n+\frac{1}{2}} \, \mathrm{d}l_0 \\ &- \rho_w \hbar c \int_{\Gamma^0_{wall}} \left[\frac{\partial \xi^{n+\frac{1}{2}}}{\partial x_1} \right]^2 \mathrm{d}l_0 - \int_{\Gamma^n_{wall}} \frac{\tilde{P}_w(t^{n+1}) \, \tilde{u}_2^{n+1}}{\sqrt{1 + (\partial_{x_1} R_0)^2}} \, \mathrm{d}l \ + \int_{\Omega^n} \tilde{\boldsymbol{u}}^{n+1} \cdot \boldsymbol{f}^{n+1} \, \mathrm{d}\omega \\ &+ \int_0^{R_0} P_{in}(t^{n+1}) \tilde{u}_1^{n+1}|_{x_1=0} \, \mathrm{d}x_2 - \int_0^{R_0} P_{out}(t^{n+1}) \tilde{u}_1^{n+1}|_{x_1=L} \, \mathrm{d}x_2. \end{aligned} \tag{4.18}$$

Moreover, we have div $\tilde{\boldsymbol{u}}^{n+1} = 0$ in Ω^n. The operator B is discretized in time via the Crank-Nicolson scheme, i.e.

$$\frac{\eta^{n+1} - \eta^n}{\Delta t} = \frac{1}{2}\big(\xi^{n+1} + \xi^{n+\frac{1}{2}}\big) \ , \tag{4.19}$$

$$\frac{\xi^{n+1} - \xi^{n+\frac{1}{2}}}{\Delta t} = \frac{a}{2}\big(\eta^{n+1}_{x_1x_1} + \eta^{n}_{x_1x_1}\big) - \frac{b}{2}\big(\eta^{n+1} + \eta^{n}\big) + H(R_0). \tag{4.20}$$

The discrete scheme (4.19)-(4.20) is also reported in literature as the Newmark scheme.

First we look for an energy estimate of the semi-discrete weak formulation of the momentum equation (4.18). In order to control the energy of the operator A we apply the Young, the trace and the Korn inequality for the individual terms from (4.18). After some manipulations, cf. [28], we obtain

$$\begin{aligned}
&||\tilde{\boldsymbol{u}}^{n+1}||^2_{L^2(\Omega^n)} + C^*\Delta t||\tilde{\boldsymbol{u}}^{n+1}||^p_{W^{1,p}(\Omega^n)} \\
&+\rho_w\hbar\Big[||\xi^{n+\frac{1}{2}}||^2_{L^2(\Gamma^0_{wall})} - ||\xi^{n}||^2_{L^2(\Gamma^0_{wall})} + 2\Delta t c\ ||\xi^{n+\frac{1}{2}}_{x_1}||^2_{L^2(\Gamma^0_{wall})}\Big] \\
&\leq ||\boldsymbol{u}^n||^2_{L^2(\Omega^n)} + \alpha^n\Delta t||\tilde{\boldsymbol{u}}^{n+1}||^2_{L^2(\Omega^n)} + \frac{\Delta t}{2\varepsilon}\mathrm{RHS}^{\mathrm{n+1}} + 2C^*\kappa\ \Delta t \ ,
\end{aligned} \tag{4.21}$$

where $\kappa = 0$ for $p \geq 2$ and $\kappa = 1$ for $1 \leq p < 2$, $\alpha^n := ||\mathrm{div}\ \boldsymbol{w}^n||_{L^\infty(\Omega^n)}$,

$$\begin{aligned}
\mathrm{RHS}^{\mathrm{n+1}} \ := \ & ||P_{in}(t^{n+1})||^{p'}_{L^{p'}(\Gamma_{in})} + ||P_{out}(t^{n+1})||^{p'}_{L^{p'}(\Gamma_{out})} + ||\tilde{P}_w(t^{n+1})||^{p'}_{L^{p'}(\Gamma^n_{wall})} \\
& +||\boldsymbol{f}^{n+1}||^{p'}_{L^{p'}(\Omega^{n+1})}
\end{aligned}$$

and C^*, C^{tr}, ε are positive constants. The dual argument $p' \geq 1$ satisfies $1/p + 1/p' = 1$.

In order to rewrite the term containing the norm $||\tilde{\boldsymbol{u}}^{n+1}||_{L^2(\Omega^n)}$ by means of $||\tilde{\boldsymbol{u}}^{n+1}||_{L^2(\Omega^{n+1})}$ and $||\boldsymbol{u}^n||_{L^2(\Omega^n)}$ we use the so-called **Geometric Conservation Law** (GCL), cf. [15, 26, 36]. It requires that a numerical scheme should reproduce a constant solution, i.e.

$$\int_{\Omega^{n+1}} \mathrm{d}\omega^{n+1} - \int_{\Omega^n} \mathrm{d}\omega^n = \int_{t^n}^{t^{n+1}} \int_{\Omega_t} \mathrm{div}\ \boldsymbol{w}\ \mathrm{d}\omega\ \mathrm{d}t \ . \tag{4.22}$$

Applying (4.22) to the function $|\tilde{\boldsymbol{u}}^{n+1}|^2$ we obtain

$$||\boldsymbol{u}^{n+1}||^2_{L^2(\Omega^{n+1})} - ||\tilde{\boldsymbol{u}}^{n+1}||^2_{L^2(\Omega^n)} = \int\limits_{t^n}^{t^{n+1}} \int\limits_{\Omega_t} |\tilde{\boldsymbol{u}}|^2\ \mathrm{div}\ \boldsymbol{w}\ \mathrm{d}\omega\ \mathrm{d}t \ . \tag{4.23}$$

Taking into account the ALE-mapping, we have for $t \in (t^n, t^{n+1})$

$$\boldsymbol{x} = \mathcal{A}_{t^n,t^{n+1}}(\boldsymbol{x}^n), \quad \mathrm{d}\omega^n = |J^{-1}_{A_{t^n,t^{n+1}}}|\ \mathrm{d}\omega \ ,$$

where $\mathcal{A}_{t^n,t^{n+1}} := \mathcal{A}_{t^{n+1}} \circ \mathcal{A}^{-1}_{t^n}$ denotes the ALE-mapping between two time levels, J_A is the determinant of the Jacobian matrix of the ALE mapping. The right hand

side of (4.23) can be further estimated in the following way

$$\int_{t^n}^{t^{n+1}} \int_{\Omega_t} |\tilde{\boldsymbol{u}}|^2 \text{ div } \boldsymbol{w} \text{ d}\omega \text{ d}t \leq \beta^n \Delta t ||\boldsymbol{u}^n||^2_{L^2(\Omega^n)} \ , \tag{4.24}$$

where $\beta^n := \sup_{t\in(t^n,t^{n+1})} \Big\{ ||\text{div } \boldsymbol{w} \cdot |J^{-1}_{A_{t^n,t^{n+1}}}| \ ||_{L^\infty(\Omega^n)} \Big\}$. Inserting (4.24) to (4.23) we obtain the desired estimate

$$||\tilde{\boldsymbol{u}}^{n+1}||^2_{L^2(\Omega^n)} \geq ||\boldsymbol{u}^{n+1}||^2_{L^2(\Omega^{n+1})} - \beta^n \Delta t ||\boldsymbol{u}^n||^2_{L^2(\Omega^n)} \ . \tag{4.25}$$

Moreover, we also obtain from (4.23)

$$\alpha^n \Delta t ||\tilde{\boldsymbol{u}}^{n+1}||^2_{L^2(\Omega^n)} \leq \alpha^n \Delta t ||\boldsymbol{u}^{n+1}||^2_{L^2(\Omega^{n+1})} + \alpha^n \beta^n (\Delta t)^2 ||\boldsymbol{u}^n||^2_{L^2(\Omega^n)}. \tag{4.26}$$

Using the inequalities (4.25)-(4.26) and summing up (4.21) for the first $n+1$ time steps we obtain the following estimate for the operator A

$$\begin{aligned} &||\boldsymbol{u}^{n+1}||^2_{L^2(\Omega^{n+1})} + C^* \Delta t \ \sum_{i=0}^{n} ||\tilde{\boldsymbol{u}}^{i+1}||^p_{W^{1,p}(\Omega^i)} \\ &\qquad + \rho_w \hbar \sum_{i=0}^{n} \left[||\xi^{i+\frac{1}{2}}||^2_{L^2(\Gamma^0_{wall})} - ||\xi^i||^2_{L^2(\Gamma^0_{wall})} + 2\Delta t c ||\xi^{i+\frac{1}{2}}_{x_1}||^2_{L^2(\Gamma^0_{wall})} \right] \\ &\leq \left[1 + \Delta t \beta^0 + (\Delta t)^2 \alpha^0 \beta^0 \right] ||\boldsymbol{u}^0||^2_{L^2(\Omega^0)} + \Delta t \sum_{i=1}^{n+1} \left[\beta^i (1 + \alpha^i \Delta t) + \alpha^{i-1} \right] ||\boldsymbol{u}^i||^2_{L^2(\Omega^i)} \\ &\qquad + \frac{\Delta t}{2\varepsilon} \ \sum_{i=1}^{n+1} \text{RHS}^{\text{i}} + 2C^* \kappa T. \end{aligned} \tag{4.27}$$

In order to estimate of the operator B we firstly multiply the equation (4.19) by $b(\eta^{n+1} + \eta^n)$ and the equation (4.20) by $(\xi^{n+1} + \xi^{n+\frac{1}{2}})$, secondly sum up the multiplied equations and then integrate them over Γ^0_{wall}. Finally, after some manipulation [28], we obtain

$$\begin{aligned} &a||\eta^{n+1}_{x_1}||^2_{L^2(\Gamma^0_{wall})} + \frac{b}{2} \ ||\eta^{n+1}||^2_{L^2(\Gamma^0_{wall})} + ||\xi^{n+1}||^2_{L^2(\Gamma^0_{wall})} \\ &\leq a||\eta^0_{x_1}||^2_{L^2(\Gamma^0_{wall})} + \frac{3b}{2} \ ||\eta^0||^2_{L^2(\Gamma^0_{wall})} + ||\xi^0||^2_{L^2(\Gamma^0_{wall})} \\ &\quad + \sum_{i=0}^{n} \Big(||\xi^{i+\frac{1}{2}}||^2_{L^2(\Gamma^0_{wall})} - ||\xi^i||^2_{L^2(\Gamma^0_{wall})} \Big) + \frac{aL|\Gamma^0_{wall}|}{\delta}. \end{aligned} \tag{4.28}$$

Here $L := \left\| \dfrac{\partial^2 R_0}{\partial x_1^2} \right\|^2_{L^\infty(\Gamma^0_{wall})}$ and δ is a small positive number.

Combining the estimates for the operator A, cf. (4.27), with the operator B, cf. (4.28), we obtain

$$E^{n+1} + \Delta t \ \sum_{i=1}^{n+1} G^i \leq E^0 + Q^0 + \Delta t \ \sum_{i=1}^{n+1} P^i + \Delta t \sum_{i=1}^{n+1} \left[\beta^i (1 + \alpha^i \Delta t) + \alpha^{i-1} \right] E^i,$$

where

$$E^i := ||\boldsymbol{u}^i||^2_{L^2(\Omega^i)} + \rho_s\hbar\left[||\eta^i_{x_1}||^2_{L^2(\Gamma^0_{wall})} + \frac{b}{2}\,||\eta^i||^2_{L^2(\Gamma^0_{wall})} + ||\xi^i||^2_{L^2(\Gamma^0_{wall})}\right],$$

$$G^i := C^*||\tilde{\boldsymbol{u}}^i||^p_{W^{1,p}(\Omega^{i-1})} + 2\rho_w\hbar\; c||\xi^{i-\frac{1}{2}}_{x_1}||^2_{L^2(\Gamma^0_{wall})},$$

$$Q^0 := \left[\Delta t\beta^0 + (\Delta t)^2\alpha^0\beta^0\right]||\boldsymbol{u}^0||^2_{L^2(\Omega^0)} + \rho_w\hbar b\;||\eta^0||^2_{L^2(\Gamma^0_{wall})} + \frac{aL|\Gamma^0_{wall}|}{\delta} + 2C^*\kappa\; T,$$

$$P^i := \frac{1}{2\varepsilon}\;\mathrm{RHS}^{\mathrm{i}},$$

and $i = 0, \ldots, n+1$. Finally, using the discrete Gronwall lemma, cf. [42], we obtain

$$E^{n+1} + \Delta t\sum_{i=1}^{n+1} G^i \le \left[E^0 + Q^0 + \Delta t\sum_{i=1}^{n+1} P^i\right]\exp\left\{\sum_{i=1}^{n+1}\frac{(\beta^i(1+\alpha^i\Delta t)+\alpha^{i-1})\Delta t}{1-(\beta^i(1+\alpha^i\Delta t)+\alpha^{i-1})\Delta t}\right\} \tag{4.29}$$

with the following condition on the time step

$$\Delta t \le \frac{1}{\beta^i(1+\alpha^i\Delta t)+\alpha^{i-1}} \quad \text{for} \quad i = 0, \ldots, n+1. \tag{4.30}$$

We would like to point out that assuming a smooth grid movement the coefficients α^i and β^i are sufficiently small and thus the condition (4.30) is not very restrictive. Indeed, our estimate is more general than those obtained by Formaggia et al. [15]. The estimate (4.29) states that the kinetic and the dissipative energy $E^{n+1} + \Delta t\sum_{i=1}^{n+1} G^i$ is bounded with the initial and boundary data as well as a small constant arising from the smooth mesh movement.

Remark 4.3. *Applying the midpoint rule for approximation of the convective ALE-term we can derive a corresponding energy estimate of the semi-discrete scheme without any dependence on the domain velocity $\boldsymbol{w}$. Here, we use the fact that in two-dimensional case the integrand on the left hand side of the geometric conservation law (4.23) can be exactly computed using the midpoint integration rule, cf. [15, 26, 36], i.e.*

$$\int_{t^n}^{t^{n+1}}\int_{\Omega_t} |\tilde{\boldsymbol{u}}^{n+1}|^2\; div\;\boldsymbol{w}\;\mathrm{d}\omega\;\mathrm{d}t\; = \Delta t\int_{\Omega^{n+1/2}} |\hat{\boldsymbol{u}}^{n+1}|^2\; div\;\boldsymbol{w}^{n+1/2}\;\mathrm{d}\omega. \tag{4.31}$$

Here $\hat{\boldsymbol{u}}^{n+1} = \boldsymbol{u}^{n+1}\circ A_{t^{n+1}}\circ A^{-1}_{t^{n+1/2}}$ is defined on $\Omega^{n+1/2}$. As a consequence, the ALE-term will exactly balance out the integral on right hand side of (4.31) and the total energy at the new time step t^{n+1} will be bounded only with the initial energy and the boundary data

$$E^{n+1} + \Delta t\sum_{i=1}^{n+1} G^i \le\; E^0 + \rho_w\hbar b\;||\eta^0||^2_{L^2(\Gamma^0_{wall})} + \frac{aL|\Gamma^0_{wall}|}{\delta} + 2C^*\kappa T + \Delta t\sum_{i=1}^{n+1} P^i\;. \tag{4.32}$$

For more details on the derivation of energy estimates (4.29) and (4.32) the reader is referred to [28].

CHAPTER 2

Numerical study

1. Numerical study

1.1. Hemodynamical indices. Several hemodynamical indices have been proposed in literature in order to measure the risk zones in a blood vessel. They have been introduced to describe the mechanisms correlated to intimal thickening of vessel wall. Many observations show that one reason is the blood flow oscillations during the diastolic phase of every single heart beat. To identify the occlusion risk zones the *Oscillatory Shear Index* is usually studied in literature, see [41]

$$OSI := \frac{1}{2}\left(1 - \frac{\int_0^T \tau_w \, dt}{\int_0^T |\tau_w| \, dt}\right), \tag{1.1}$$

where $(0, T)$ is the time interval of a single heart beat ($T \approx 1sec$) and τ_w is the *Wall Shear Stress* (WSS) defined as

$$WSS := \tau_w = -\mathbf{T}\boldsymbol{n} \cdot \boldsymbol{\tau}. \tag{1.2}$$

Here $\boldsymbol{n}$ and $\boldsymbol{\tau}$ are the unit outward normal and the unit tangential vector on the arterial wall $\Gamma_{wall}(t)$, respectively. OSI index measures the temporal oscillations of the shear stress pointwisely without taking into account the shear stress behavior in an immediate neighborhood of a specific point.

It is known that the typical range of WSS in a normal artery is $[1.0, 7.0]$ Pa and in the venous system it is $[0.1, 0.6]$ Pa, see [30]. The regions of artery that are athero-prone, i.e. stimulates an atherogenic phenotype, are in the range of ± 0.4 Pa. On the other hand, the WSS greater than 1.5 Pa induces an anti-proliferative and anti-thrombotic phenotype and therefore is found to be athero-protective. In the range of $[7, 10]$ Pa high-shear thrombosis is likely to be found.

Since the viscosity of the non-Newtonian fluid is a function of shear rate, see Fig. 1, for comparison with Newtonian flow we introduce the Reynolds number for non-Newtonian models using averaged viscosity

$$Re = \frac{\rho V l}{\frac{1}{2}(\mu_0 + \mu_\infty)}, \tag{1.3}$$

where ρ is the fluid density, V is the characteristic velocity (e.g. maximal inflow velocity), l is the characteristic length (we take the diameter of a vessel). In order to take into account also the effects of asymptotical viscosity values, we define $Re_0 = \rho V l/\mu_0$, $Re_\infty = \rho V l/\mu_\infty$ and introduce them in the Table 1 below as well.

1.2. Computational geometry and parameter settings. We will present our numerical experiments for two test geometries. In the first one, Fig. 2 or Fig. 1,

the two dimensional symmetric vessel with a smooth stenosed region is considered. Due to the symmetry we can restrict our computational domain to the upper half of the vessel. Let $\Gamma_{in} = \{(-L, x_2);\, x_2 \in (0, R(-L,t))\}$, $\Gamma_{out} = \{(L, x_2);\, x_2 \in (0, R(L,t))\}$, $\Gamma_{sym} = \{(x_1, 0);\, x_1 \in (-L, L)\}$ denote the inflow, outflow and symmetry boundary, respectively. The impermeable moving wall $\Gamma_{wall}(t)$ is modeled as a smooth stenosed constriction given as, see [31],

$$R_0(x_1) = \begin{cases} R_0(-L)\left[1 - \frac{g}{2}\left(1 + \cos\left(\frac{5\pi x_1}{2L}\right)\right)\right] & \text{if } x_1 \in [-0.4L; 0.4L] \\ R_0(-L) & \text{if } x_1 \in [-L; -0.4L) \cup (0.4L; L]. \end{cases}$$

We took $L = 5$ cm, $g = 0.3$ with $R(-L,t) = R(L,t) = 1$ cm for experiments with model data. These values give a stenosis with 30% area reduction which corresponds to a relatively mild occlusion, leading to local small increment of the Reynolds number. When considering physiological pulses prescribed by the iliac flow rate (Fig. 3, left) and the physiological viscosities (Tab. 1), the radius $R(-L,t) = R(L,t) = 0.6$ cm and the length $L = 3$ cm were chosen. This radius represents the physiological radius of an iliac artery, i.e. a daughter artery of the abdominal aorta bifurcation, cf. [46].

The bifurcation geometry shown in Fig. 2 represents the second test domain. This is a more complex geometry with asymmetric daughter vessels and the so-called sinus bulb area. Indeed, it is a simplified example of a realistic carotid artery bifurcation, see [38]. The radii of the mother vessel (i.e. common carotid artery), daughter vessels (i.e. external and internal carotid artery) and the maximal radius of the sinus bulb area are: $r_0 = 0.31$ cm, $r_1 = 0.22$ cm, $r_2 = 0.18$ cm and $r_S = 0.33$ cm. The branching angles for the bifurcation in Fig. 2 are $\gamma_1 = \gamma_2 = 25°$. We

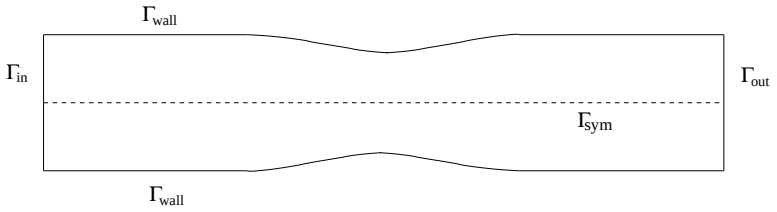


FIGURE 1. Stenotic reference geometry.

note that since the generalized string model has been derived for radially symmetric domains we need to preserve the radial symmetry of the geometry also after the bifurcation divider. For this purpose we rotate the original coordinate system with respect to the bifurcation angle γ_1 of the daughter vessel. In our simulations for simplicity we assume that only one part of the boundary Γ_{wall} (this corresponds to the boundary Γ^m_{wall} in Fig. 2) is allowed to move. This is motivated by the fact that atherogenesis occurs preferably at the outer wall of daughter vessel, in particular in the carotid sinus, see [30]. Therefore this is the area of a special interest. Note that we use two different reference frames. One corresponds to the mother vessel and in the second reference frame the x_1-axis coincides with the axis of symmetry of the daughter vessel.

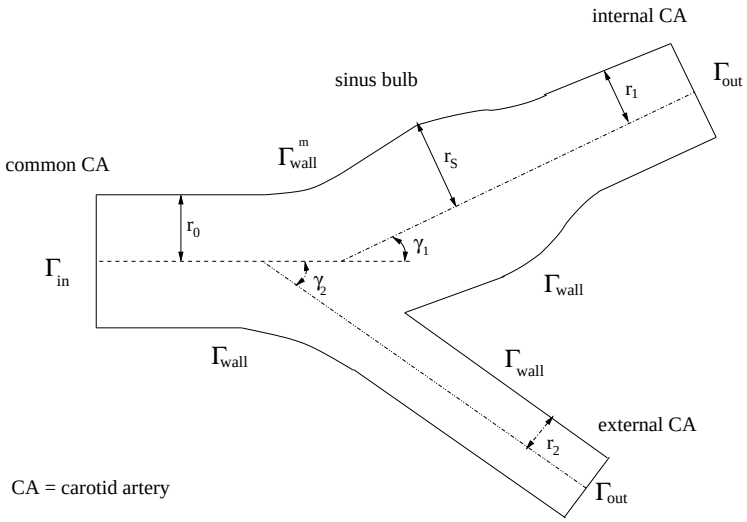


FIGURE 2. Bifurcation reference geometry, see [38].

1.2.1. *Boundary conditions.* For Γ_{sym} the symmetry the boundary conditions $\partial_{x_2} u_1 = 0,\ u_2 = 0$ is prescribed, for Γ_{out} the Neumann type boundary condition $-\mathbf{T}\boldsymbol{n} = P_{out}\mathbf{I}\boldsymbol{n}$ is used. We prescribe the pulsatile parabolic velocity profile on the inflow boundary

$$u_1(-L, x_2) = V_{max}\frac{(R^2 - x_2^2)}{R^2} f(t), \qquad u_2(-L, x_2) = 0, \tag{1.4}$$

where $R = R(-L,t) = R_0(-L) + \eta(-L,t)$ and $V_{max} = u_1(-L,0)$ is the maximal velocity at the inflow. For temporal function modeling pulses of heart $f(t)$ we have used two variants: $f(t) = sin^2(\pi t/\omega)$ with the period $\omega = 1s$, and $f(t)$ coming from physiological pulses of heart and iliac artery flow rate $Q(t)$, depicted in Fig. 3.

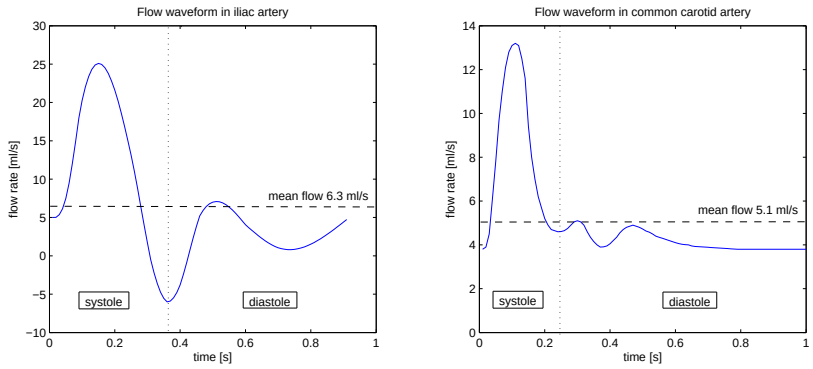


FIGURE 3. Flow rate $Q(t)$ in the iliac artery (left) and in a common carotid artery (right), see [38, 46].

Indeed, the flow rate is obtained as an integral over the inflow surface S_{in}, in our case

$$Q(t) = \int_{S_{in}} u_1 dS_{in} = 2\pi \int_0^R x_2 u_1(-L, x_2) dx_2, \quad R = R(-L, t).$$

Taking into account inflow velocity (1.4) we obtain that $Q(t) = \frac{1}{2}\pi V_{max} R^2 f(t)$. Consequently we get the relation for temporal function $f(t)$, which we use in (1.4),

$$f(t) = \frac{2Q(t)}{\pi V_{max} R^2(-L, t)}. \tag{1.5}$$

Note that the mean inflow velocity and the maximal inflow velocity are defined by $\bar{U} = Q(t)/(\pi R^2)$ and $V_{max} = 2Q(t)/(\pi R^2)$, respectively.

1.2.2. *Parameter settings.* In the first part we have chosen in analogy to Nadau and Sequeira [31], $Re_0 = 30$ or $Re_0 = 60$ and $\mu_\infty = \frac{1}{2}\mu_0$ for the Carreau model (2.2) as well as for the Yeleswarapu model (2.3), cf. Section 5.1. For these (artificial) viscosities we will compare the hemodynamical wall parameters and study the experimental convergence of our methods. Moreover we test the stability and robustness of the method for physiological viscosity parameters [48]. The viscosity parameters for experiments with model and for physiological data are collected in Table 1 below. In order to model pulsatile flow in a vessel we use sinus pulses introduced above. In the second part we test the stability and robustness of the

TABLE 1. Parameters for numerical experiments.

$Re = 34$	$Re = 80$	$Re = 40$	$Re = 80$
Carreau model $q = 0, -0.322, -0.2,\ \gamma = 1$		**Yeleswarapu model** $\gamma = 14.81$	
$\mu_\infty = 1.26P$ $\mu_0 = 2.53P$ $V_{max} = 38\,cm/s$ $Re_0 = 30,\ Re_\infty = 60$	$\mu_\infty = 0.63P$ $\mu_0 = 1.26P$ $V_{max} = 38\,cm/s$ $Re_0 = 60,\ Re_\infty = 121$	$\mu_\infty = 1.26P$ $\mu_0 = 2.53P$ $V_{max} = 38\,cm/s$ $Re_0 = 30,\ Re_\infty = 60$	$\mu_\infty = 0.63P$ $\mu_0 = 1.26P$ $V_{max} = 38\,cm/s$ $Re_0 = 60,\ Re_\infty = 121$
physiological parameters $q = -0.322,\ \gamma = 3.313$		physiological parameters	
$\mu_\infty = 0.0345P,\ \mu_0 = 0.56P$ $V_{max} = 17cm/s$ $Re = 114,\ Re_\infty = 986$		$\mu_\infty = 0.05P,\ \mu_0 = 0.736P$ $V_{max} = 22.3cm/s$ $Re = 113,\ Re_\infty = 892$	

method for physiological viscosities [48] in some relevant physiological situations, e.g. in the iliac artery or in the common carotid bifurcation artery.

We note than in the human circulatory system, the Reynolds number varies significantly. Over one cycle it reaches the values from 10^{-3} up to 6000. A typical critical number for a normal artery is around 2300, for bifurcation it is around 600. However, the recirculation zones start to be created already at the Reynolds number around 170. This explains the fact that small recirculation zones appear even in healthy bifurcations. The part of a bifurcation that is the most sensitive to the local change of flow is the so-called sinus bulb area. This is a part of a daughter vessel, where an atherosclerosis is usually formed, see Fig. 2. Indeed, our analysis of the local hemodynamical parameters confirms this fact.

In the following table we give the overview of the Reynolds numbers Re_0 and Re_∞ defined in Section 5.1 for physiological viscosities. The characteristic velocity V is taken to be the mean inflow velocity $\bar{U}$. In the Tab. 2 the Reynolds numbers for physiological pulses corresponding to the iliac artery flow rate (Fig. 3, left) and common carotid artery (Fig. 3, right) are computed. We denote by Q_{mean}, Q_{max} and Q_{min} the mean, the maximal and the minimal flow rate, respectively. The Newtonian viscosity corresponds here to μ_∞ in the Carreau model.

TABLE 2. Reynolds numbers for physiological data and pulses.

Iliac artery $R(0,t) = 0.6$ cm	**Newtonian model** $\mu = 0.0345P$	**Carreau model**	**Yeleswarapu model**
$Q_{mean}(t) = 6.3$ ml.s^{-1} $\bar{U} = 5.6$ cm.s^{-1}	$\mathrm{Re} \approx 195$	$\mathrm{Re}_0 \approx 12$ $\mathrm{Re}_\infty \approx 195$	$\mathrm{Re}_0 \approx 9$ $\mathrm{Re}_\infty \approx 134$
$Q_{max}(t) = 25.1$ ml.s^{-1} $\bar{U} = 22.2$ cm.s^{-1}	$\mathrm{Re} \approx 772$	$\mathrm{Re}_0 \approx 48$ $\mathrm{Re}_\infty \approx 772$	$\mathrm{Re}_0 \approx 36$ $\mathrm{Re}_\infty \approx 533$
$Q_{min}(t) = -6.0$ ml.s^{-1} $\bar{U} = -5.3$ cm.s^{-1}	$\mathrm{Re} \approx 185$	$\mathrm{Re}_0 \approx 14$ $\mathrm{Re}_\infty \approx 185$	$\mathrm{Re}_0 \approx 10$ $\mathrm{Re}_\infty \approx 114$
Carotid artery $R(0,t) = 0.31$ cm	**Newtonian model** $\mu = 0.0345P$	**Carreau model**	**Yeleswarapu model**
$Q_{mean}(t) = 5.1$ ml.s^{-1} $\bar{U} = 16.9$ cm.s^{-1}	$\mathrm{Re} \approx 304$	$\mathrm{Re}_0 \approx 19$ $\mathrm{Re}_\infty \approx 304$	$\mathrm{Re}_0 \approx 14$ $\mathrm{Re}_\infty \approx 210$
$Q_{max}(t) = 13.2$ ml.s^{-1} $\bar{U} = 43.7$ cm.s^{-1}	$\mathrm{Re} \approx 785$	$\mathrm{Re}_0 \approx 48$ $\mathrm{Re}_\infty \approx 785$	$\mathrm{Re}_0 \approx 37$ $\mathrm{Re}_\infty \approx 542$
$Q_{min}(t) = 3.9$ ml.s^{-1} $\bar{U} = 12.9$ cm.s^{-1}	$\mathrm{Re} \approx 232$	$\mathrm{Re}_0 \approx 14$ $\mathrm{Re}_\infty \approx 232$	$\mathrm{Re}_0 \approx 11$ $\mathrm{Re}_\infty \approx 160$

1.3. Discretization methods. For the numerical approximation of (2.1), (3.2) and (2.10) we have used as a basis the UG software package [1] and extended it for the shear-dependent fluids as well as by adding the solver for the wall deformation equation (3.2). In the UG package the problem class library for the Navier-Stokes equations in moving domain is based on the ALE formulation, see [3]. The Euler implicit method, the Crank-Nicolson method or the second order backward differentiation formula can be applied for time-discretization. The spatial discretization of the fluid equations (2.1), or (2.5), is realized by the finite volume method with the pseudo-compressibility stabilization. This stabilization results in the elliptic equation for the pressure. The non-linear convective term is linearized by the Newton or fixed point method, see e.g., [32]. In what follows we explain the treatment of the non-linear viscous term, which we have implemented within the UG software package.

1.3.1. *Linearization of the viscous term.* According to Taylor's expansion we have

$$\begin{aligned} \mu(\mathbf{D}(\boldsymbol{u}))\mathbf{D}(\boldsymbol{u}) &= \mu(\mathbf{D}(\boldsymbol{u}^{old}))\mathbf{D}(\boldsymbol{u}^{old}) \\ &+ \frac{d\,[\mu(\mathbf{D}(\boldsymbol{u}))\mathbf{D}(\boldsymbol{u})]}{d(\nabla \boldsymbol{u})}(\boldsymbol{u}^{old})(\nabla \boldsymbol{u} - \nabla \boldsymbol{u}^{old}) + \mathcal{O}((\nabla \boldsymbol{u} - \nabla \boldsymbol{u}^{old})^2), \end{aligned} \tag{1.6}$$

where

$$\frac{d\,[\mu(\mathbf{D}(\boldsymbol{u}))\mathbf{D}(\boldsymbol{u})]}{d(\nabla \boldsymbol{u})}(\boldsymbol{u}^{old}) = \mu(\mathbf{D}(\boldsymbol{u}^{old}))\frac{1}{2}(I + I^T) + \frac{d\,\mu(\mathbf{D}(\boldsymbol{u}))}{d\nabla \boldsymbol{u}}(\boldsymbol{u}^{old})\mathbf{D}(\boldsymbol{u}^{old})$$

and $(.)^{old}$ denotes the previous iteration. Plugging the above expression for $\frac{d[\mu(\mathbf{D}(\boldsymbol{u}))\mathbf{D}(\boldsymbol{u})]}{d(\nabla \boldsymbol{u})}$ into (1.6) and neglecting the higher order term $\mathcal{O}((\nabla \boldsymbol{u}-\nabla \boldsymbol{u}^{old})^2)$ we obtain the Newton type iteration $\mu(\mathbf{D}(\boldsymbol{u}))\mathbf{D}(\boldsymbol{u}) \approx \mu(\mathbf{D}(\boldsymbol{u}^{old}))\mathbf{D}(\boldsymbol{u})+(\nabla \boldsymbol{u}-\nabla \boldsymbol{u}^{old})$ $\frac{d\,\mu(\mathbf{D}(\boldsymbol{u}))}{d\nabla \boldsymbol{u}}(\boldsymbol{u}^{old})\mathbf{D}(\boldsymbol{u}^{old})$. By neglecting the term $\mathcal{O}(|\nabla \boldsymbol{u} - \nabla \boldsymbol{u}^{old}|)$ we get the fixed point approximation

$$\mu(\mathbf{D}(\boldsymbol{u}))\mathbf{D}(\boldsymbol{u}) \approx \mu(\mathbf{D}(\boldsymbol{u}^{old}))\mathbf{D}(\boldsymbol{u}). \tag{1.7}$$

We iterate with respect to $\boldsymbol{u}; \boldsymbol{u}_\ell \equiv \boldsymbol{u}^{old}$, see (1.8). The fixed point iteration can be also understood as the Newton iteration with an incomplete Jacobian matrix, since the second part of the Jacobian matrix $\frac{d\,\mu(\mathbf{D}(\boldsymbol{u}))}{d(\nabla \boldsymbol{u})}(\boldsymbol{u}^{old})\mathbf{D}(\boldsymbol{u}^{old})$ is neglected.

Now, we present the finite volume method used in the UG package with the fixed point linearization for the viscous and convective terms and the Euler implicit time discretization

$$\begin{aligned}
&\int\limits_{\Omega_i} \begin{pmatrix} (\boldsymbol{u}_{\ell+1}^{n+1} - \boldsymbol{u}^n) \\ 0 \end{pmatrix} \, d\omega + \Delta t \int\limits_{\Omega_i} \begin{pmatrix} (\mathrm{div}\boldsymbol{w}^n)\boldsymbol{u}_{\ell+1}^{n+1} \\ 0 \end{pmatrix} \, d\omega \qquad (1.8)\\
+\Delta t &\int\limits_{\partial\Omega_i} \begin{pmatrix} [(\boldsymbol{u}_\ell^{n+1} - \boldsymbol{w}^n)\cdot \boldsymbol{n}]\boldsymbol{u}_{\ell+1}^{n+1} + [(\boldsymbol{u}_{\ell+1}^{n+1} - \boldsymbol{u}_\ell^{n+1})\cdot \boldsymbol{n}]\boldsymbol{u}_\ell^{n+1} \\ 0 \end{pmatrix} \, dS \\
+\Delta t &\int\limits_{\partial\Omega_i} \begin{pmatrix} -(1/\rho)\mu(\mathbf{D}(\boldsymbol{u}_\ell^{n+1}))(\bigtriangledown \boldsymbol{u}_{\ell+1}^{n+1}\cdot \boldsymbol{n}) + (1/\rho)p_{\ell+1}^{n+1}\ (\mathbf{I}\cdot \boldsymbol{n}) \\ \boldsymbol{u}_{\ell+1}^{n+1}\cdot \boldsymbol{n} - h^2 \bigtriangledown (p_{\ell+1}^{n+1} - p_\ell^{n+1})\cdot \boldsymbol{n} \end{pmatrix} \, dS = 0.
\end{aligned}$$

In the case of the global iterative method Ω_i denotes the i-th control volume at time t^{n+1} given from the previous iteration, i.e. $\Omega_i = \Omega_i^{(k-1)}(t^{n+1})$. The grid velocity $\boldsymbol{w}^n = \boldsymbol{w}(t^n)$ is obtained using the backward difference of the grid position $\boldsymbol{w}^n = \frac{\boldsymbol{x}^{n+1,\,(k-1)} - \boldsymbol{x}^n}{\Delta t}$. In the case of the kinematical splitting we have $\Omega_i = \Omega_i(t^n)$ and $\boldsymbol{w}^n = \frac{\boldsymbol{x}^n - \boldsymbol{x}^{n-1}}{\Delta t}$.

1.3.2. *Discretization of structure equation.* In order to approximate the structure equation we apply the finite difference method. For the global iterative method we will rewrite the second order equation (3.2) as a system of two first order equations. Set $\xi = \partial_t \eta$. Time discretization is realized by the following scheme

$$\begin{aligned}
\frac{\xi^{n+1} - \xi^n}{\Delta t} - a\alpha \frac{\partial^2 \eta^{n+1}}{\partial x_1^2} \quad &+ \quad b\alpha\eta^{n+1} - c\alpha \frac{\partial^2 \xi^{n+1}}{\partial x_1^2} \qquad (1.9)\\
&= \quad H + a(1-\alpha)\frac{\partial^2 \eta^n}{\partial x_1^2} - b(1-\alpha)\eta^n + c(1-\alpha)\frac{\partial^2 \xi^n}{\partial x_1^2} \\
\frac{\eta^{n+1} - \eta^n}{\Delta t} \quad &= \quad \alpha\xi^{n+1} + (1-\alpha)\xi^n,
\end{aligned}$$

where

$$a = \frac{|\sigma_{x_1}|}{\rho_w}\left[1 + \left(\frac{\partial R_0}{\partial x_1}\right)^2\right]^{-2}, \quad b = \frac{E}{\rho_w R_0(R_0+\eta)} + \frac{(\tilde{\mathbf{T}} + \tilde{P_w}\mathbf{I})\cdot \tilde{\boldsymbol{n}}\cdot \boldsymbol{e}_r}{R_0\rho_w\hbar},$$

$$c = \frac{\gamma}{\rho_w\hbar}, \quad H = H(\boldsymbol{u}, p) + H(R_0) = -\frac{(\tilde{\mathbf{T}} + \tilde{P_w}\mathbf{I})\cdot \tilde{\boldsymbol{n}}\cdot \boldsymbol{e}_r}{\rho_w\hbar} + a\frac{\partial^2 R_0}{\partial x_1^2}\ .$$

We note that in contrary to the definitions given by (4.13) and (4.17), we included a part of the right hand side term having the factor η/R_0 to the coefficient b. Moreover, due to the linear elasticity assumption we assumed small deformation gradient $\partial\eta/\partial x_1$, which yields $\frac{\sqrt{1+\left(\partial_{x_1}(R_0+\eta)\right)^2}}{\sqrt{1+\left(\partial_{x_1}R_0\right)^2}} \approx 1$.

Constants appearing in the coefficients a, b, c have typically following values, see [14]: the Young modulus is $E = 0.75 \times 10^5 dyn.cm^{-2}$, the wall thickness $\hbar = 0.1\ cm$, the density of the vessel wall tissue $\rho_w = 1.1\ g.cm^{-3}$, the viscoelasticity constant $\gamma = 2 \times 10^4\ P.s.cm^{-1}$, $|\sigma_z| = G\kappa$, where $\kappa = 1$ is the Timoshenko shear correction factor and G is the shear modulus, $G = E/2(1+\sigma)$, where $\sigma = 1/2$ for incompressible materials.

If $\alpha = 0$ we have an explicit scheme in time, for $\alpha = 1$ we obtain an implicit scheme. The parameter $\alpha = \frac{1}{2}$ yields the Newmark scheme, which is proven to be unconditionally stable at least in the case of homogeneous Dirichlet boundary conditions, see [36].

In the case of kinematical splitting algorithm, the structure equation (4.12) is discretized using the splitting approach (4.15)-(4.16). The operator A consists of (1.8) and (1.10), where

$$\frac{\xi^{n+1/2} - \xi^n}{\Delta t} = c\alpha\ \xi^{n+1/2}_{x_1x_1} + c(1-\alpha)\ \xi^n_{x_1x_1} + H(p^{n+1}, \boldsymbol{u}^{n+1}). \tag{1.10}$$

The parameter α is chosen to be either 0.5 or 1. A new solution obtained from (1.8), (1.10) is the velocity $\tilde{\boldsymbol{u}}^{n+1}$ and the pressure $\tilde{p}^{n+1}$ on Ω^n as well as the wall velocity function $\xi^{n+1/2}$ on Γ^n_{wall}. The second step is the operator B, that combines the purely elastic part of structure equation and the kinematical coupling condition (4.14). This can be discretized in an explicit or implicit way. An explicit scheme reads as follows

$$\begin{aligned}
\frac{\eta^{n+1} - \eta^n}{\Delta t} &= \alpha_1\ \xi^{n+1/2} + (1-\alpha_1)\ \xi^n,\\
\frac{\xi^{n+1} - \xi^{n+1/2}}{\Delta t} &= a\alpha_2\ \eta^{n+1}_{x_1x_1} + a(1-\alpha_2)\ \eta^n_{x_1x_1} - b\alpha_2\ \eta^{n+1} - b(1-\alpha_2)\ \eta^n + H(R_0)
\end{aligned} \tag{1.11}$$

for $\alpha_1 = 0.5,\ \alpha_2 \in \{0.5; 1\}$. An implicit scheme has the following form

$$\begin{aligned}
\frac{\eta^{n+1} - \eta^n}{\Delta t} &= \alpha_1\ \xi^{n+1} + (1-\alpha_1)\ \xi^{n+1/2}\\
\frac{\xi^{n+1} - \xi^{n+1/2}}{\Delta t} &= a\alpha_2\ \eta^{n+1}_{x_1x_1} + a(1-\alpha_2)\ \eta^n_{x_1x_1} - b\alpha_2\ \eta^{n+1} - b(1-\alpha_2)\ \eta^n + H(R_0)
\end{aligned} \tag{1.12}$$

for $\alpha_1 = 0.5,\ \alpha_2 = 0.5$. We note that once new values for the wall displacement η^{n+1} and the velocity ξ^{n+1} are known, we update the fluid velocity on the moving boundary to $\boldsymbol{u}^{n+1}$ and update the mesh. In our experiments we have used both, the explicit (1.11) as well as the implicit method (1.12). The implicit coupling was typically more stable. In order to combine the operators A and B we may use the first order **Marchuk-Yanenko operator splitting** or the second order **Strang splitting** scheme. The Marchuk-Yanenko scheme

$$\mathcal{U}^{n+1} = B_{\triangle t} A_{\triangle t}\ \mathcal{U}^n,$$

where $\mathcal{U}^n$ is the approximate solution of the coupled fluid-structure interaction problem at the time level t^n. The second order Strang splitting yields

$$\mathcal{U}^{n+1} = B_{\triangle t/2} A_{\triangle t} B_{\triangle t/2}\, \mathcal{U}^n.$$

1.4. Numerical experiments. We start with the comparison of our two fluid-structure interaction schemes: the global iterative method and the kinematical splitting. In Fig. 4 we can see the domain deformation at two different time steps, that was obtained using the global iterative method, cf. (1.9) with $\alpha = \frac{1}{2}$ (Newmark scheme) and the explicit kinematical splitting (1.10), (1.11) with $\alpha = \alpha_1 = \alpha_2 = \frac{1}{2}$. We can see that both methods yield analogous results.

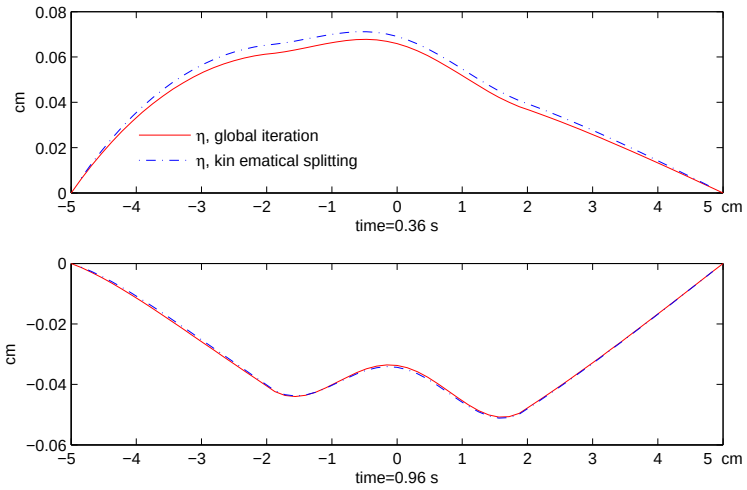


FIGURE 4. Comparison of the wall deformation η in a stenosed vessel for the global iterative method and the explicit kinematical splitting, Re=40.

Our next aim is to investigate differences in the behavior of Newtonian and non-Newtonian fluids in moving domains for different viscosity data and Reynolds numbers. Experiments presented in Section 1.4.1 as well as in the first part of Section 1.4.2 were obtained using the global iterative method. In Section 1.4.2 experiments using the kinematical splitting method for the iliac artery and the carotid bifurcation will be presented.

1.4.1. *Numerical experiments for model data.* In this subsection we present experiments with the viscosity parameters introduced in the first three lines of Table 1 for model data. The Reynolds number for non-Newtonian fluids varies between two values Re_0 and Re_∞. In order to compare similar flow regimes, the same Reynolds number $Re = 40$ for the Newtonian as well as the non-Newtonian fluids was used. The corresponding Newtonian viscosity was chosen such that it coincides with the averaged non-Newtonian viscosity $\frac{1}{2}(\mu_0 + \mu_\infty)$, see (1.3).

We use the Dirichlet inflow boundary condition (1.4), which models pulsatile parabolic velocity profile at the inflow. Here we took $f(t) = sin^2(\pi t/\omega)$ with $\omega = 1s$. We have chosen two non-Newtonian models for the blood flow often used in the literature, the Carreau and the Yeleswarapu model. Further, we study the influence of non-Newtonian rheology and of fluid-structure interaction on some hemodynamical wall parameters such as the wall shear stress WSS and the oscillatory shear index OSI. In what follows we plot the results comparing several aspects of Newtonian and non-Newtonian flow in the straight channel and in the channel with a stenotic occlusion.

Fig. 5 describes time evolution of the wall deformation function η at two time instances $t = 0.36s$ and $t = 0.96s$ for the straight and stenotic compliant vessel and for different non-Newtonian viscosities. Clearly, we can see effects due to the presence of stenosis in Fig. 5. The differences in wall deformation for non-Newtonian and Newtonian fluids are not significant.

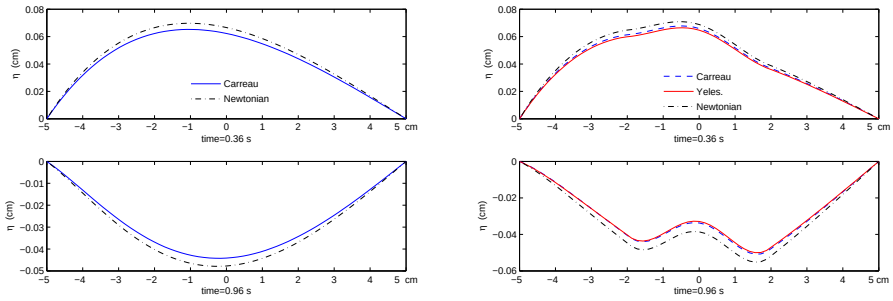


FIGURE 5. Deformation of the compliant wall η, left: a straight vessel, right: a stenosed vessel, $Re = 40$.

Fig. 6, 7 describe the wall shear stress distribution WSS along the moving or fixed (solid) wall in the straight and in stenotic vessel, respectively. We compare the WSS for the Newtonian and non-Newtonian fluids. Analogously as before we see that the WSS depends considerably on the geometry. In Fig. 7 peaks in the WSS due to the stenosis can be identified clearly for both Newtonian and non-Newtonian models. Fluid rheology is even more significant for WSS measurements; see different behaviour of WSS at $t = 0.36s$ in Fig. 6 and Fig. 7. Moreover, we can conclude that the WSS at $t = 0.36s$ is in general lower in a compliant vessel than in a solid one, see Fig. 6 for the straight and Fig. 7 for the stenotic vessel.

Another important hemodynamical wall parameter is the oscillatory shear index OSI. Fig. 8 describes the behavior of the OSI in the straight and stenotic vessel (both solid and compliant case). We can see new effects due to the presence of stenosis in the OSI. Moreover the peaks in the OSI are more dominant for the non-Newtonian models in comparison to the Newtonian flow. High OSI values indicate the areas with the large stenotic plug danger. Fig. 8 indicates, that such areas appear at the end of stenotic reduction. Numerical simulation with solid

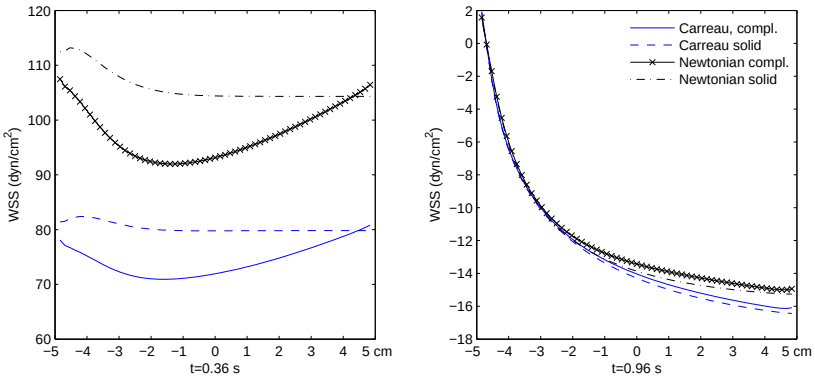


FIGURE 6. WSS along the straight vessel with solid as well as compliant walls, the Newtonian and the Carreau model, $Re = 40$, left: $t = 0.36s$, right: $t = 0.96s$.

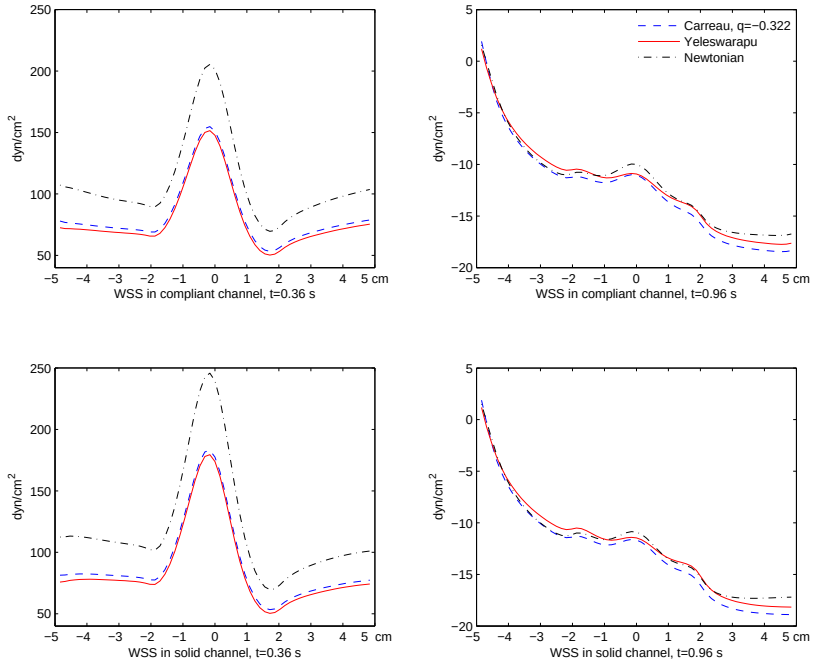


FIGURE 7. WSS along the vessel wall in a stenosed compliant (top) and solid (bottom) vessel at two time instances, $Re = 40$.

vessel walls indicates even higher oscillation of the wall shear stress. Thus, simulations without fluid-structure interaction would indicate more critical shear stress situation in vessels as they are actually present in elastic moving vessels.

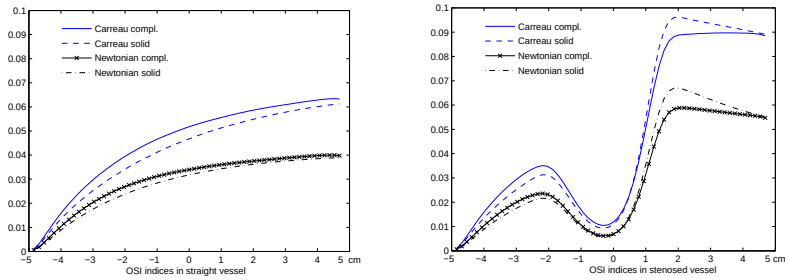


FIGURE 8. OSI indices along the compliant and the solid vessel wall, the straight vessel (left) and the stenosed vessel (right), $Re = 40$.

We conclude this subsection with a statement, that the fluid rheology and domain geometry may have a considerable influence on the hemodynamical wall parameters WSS and OSI. The fluid-structure interaction aspect plays definitely significant role in the prediction of hemodynamical indices and should be involved in reliable computer simulations.

1.4.2. *Numerical experiments for physiological data.* Several results comparing the behavior of both non-Newtonian models, the Carreau and the Yeleswarapu model with corresponding physiological viscosities from Table 1, cf. lines 4 and 5, are presented below. We consider here pulsatile velocity profile at the inflow as in Section 1.4.1 and zero Dirichlet boundary conditions for η.

Fig. 9 describes the velocity field, streamlines and the pressure distribution at two time instances. We can clearly notice reversal flow areas due to pulsatile behavior of blood flow. At time $t = 0.96$ s, where the inflow velocity is decreasing, we can observe vortices in the streamlines.

In what follows we compare measurements for increasing Reynolds numbers, namely for $Re = 114$ and $Re = 182$. For comparison with the Newtonian case, we are using two values of the corresponding Newtonian viscosity. One Newtonian viscosity, similarly as in the previous Section 1.4.1, is obtained from the averaged non-Newtonian viscosity $\mu = \frac{1}{2}(\mu_0 + \mu_\infty)$. The second one is the physiological viscosity $\mu = 0.0345\ P$.

Our numerical experiments confirm, that the differences between Newtonian (both physiological and averaged viscosity) and non-Newtonian fluids in the wall deformation, the wall shear stress WSS as well as in the oscillatory shear index OSI are more dominant with increasing Reynolds numbers, see Figs. 10, 11, 12. We can also observe that the amplitude of wall displacement and wall shear stress is smaller for the physiological Newtonian velocity and larger for the averaged Newtonian velocity. This means that concerning physiological Newtonian viscosity (or averaged Newtonian viscosity) the hemodynamical parameters predict more (or

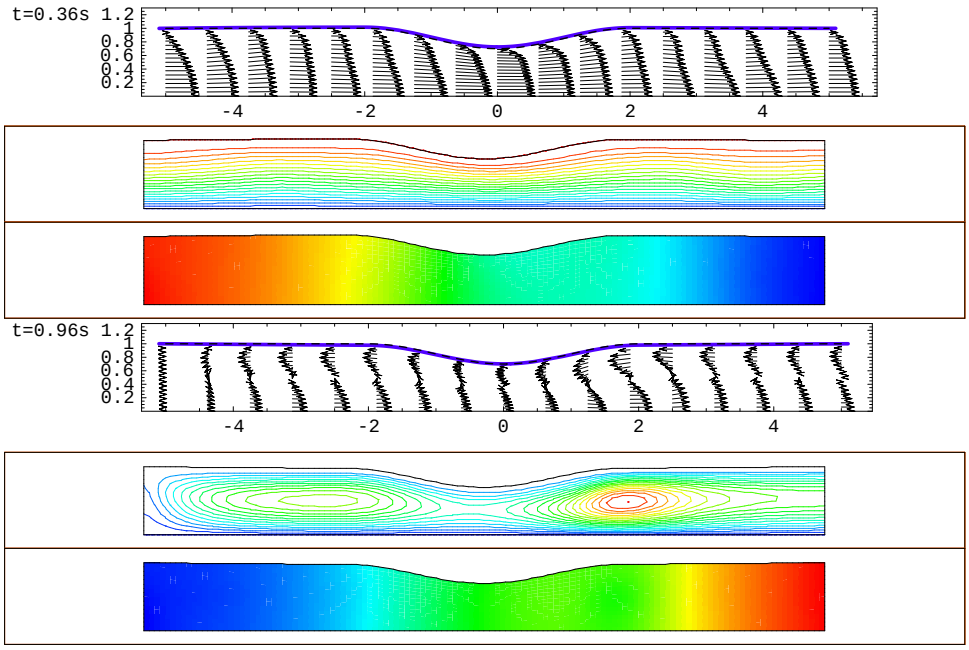


FIGURE 9. Numerical experiment using physiological parameters: the Carreau model, $t = 0.36s$ (top) and $t = 0.96s$ (bottom), from above: velocity field, streamlines and pressure distribution.

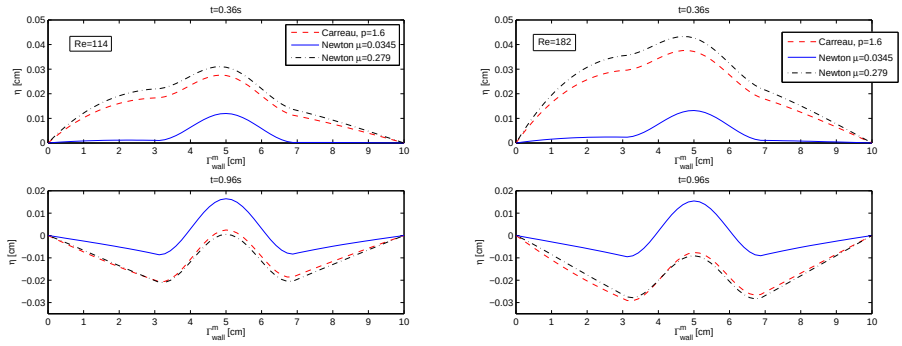


FIGURE 10. Comparison of wall deformations for different Reynolds numbers, left: Re=114, right: Re=182, at two time instances $t = 0.36s$, $t = 0.96s$.

less) critical situation in the case of Newtonian flow than by the non-Newtonian flow.

Concerning the oscillatory shear index OSI, see Fig. 12, the higher Reynolds number corresponds to the higher values of OSI. It shows that increasing the

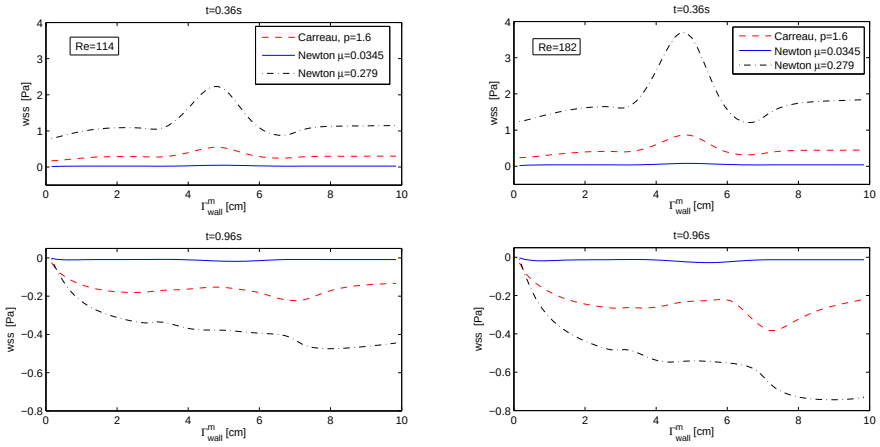


FIGURE 11. Comparison of wall shear stresses WSS for different Reynolds numbers, left: Re=114, right: Re=182, at two time instances $t = 0.36s$, $t = 0.96s$.

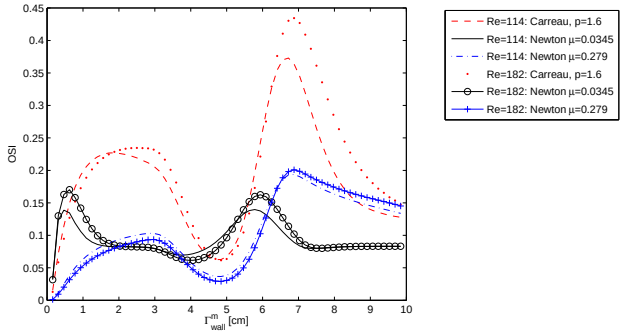


FIGURE 12. OSI indices for different Reynolds numbers using physiological viscosities.

Reynolds number, the amplitude of WSS increases and the period of reversed flow prolongates.

1.4.3. *Iliac artery and carotid bifurcation.* In this part we present experiments for physiological situations, including a simplified but realistic geometry (Figs. 1,2), physiological flow rate (Fig. 3) as well as the physiological viscosities (Newtonian viscosity $\mu = 0.0345$ P and non-Newtonian viscosities from Tab. 1).

Figs. 13, 14 (left) demonstrate dependence of the wall displacement on the reference geometry. From the evolution of the wall movement for several time instances corresponding to systolic maximum, systolic minimum, diastolic minimum and the final phase of the physiological flow for the Carreau viscosity we conclude that the presence of stenosis as well as bifurcation divider influences the compliance

of the vessel wall. Comparing the Newtonian as well as non-Newtonian rheology (Figs. 13, 14, right) we see that the difference between them is not significant.

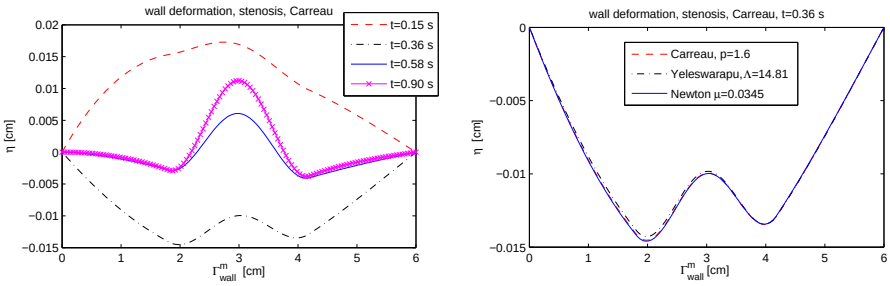


FIGURE 13. The evolution of η along the line $x_2 = R_0$ for stenosed vessel. Left: comparison at several time instances, right: comparison of different constitutive models at $t = 0.36$ s.

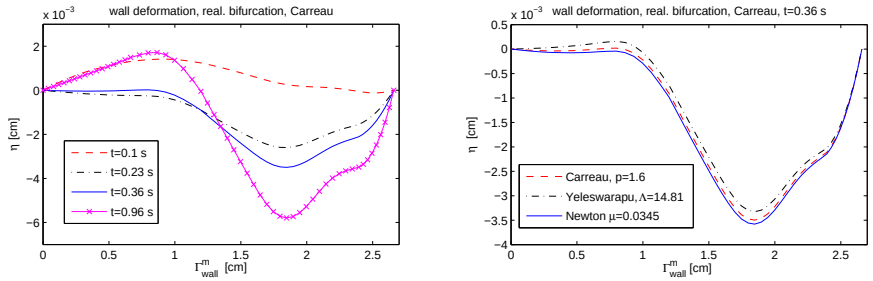


FIGURE 14. The evolution of η along the moving boundary Γ^m_{wall} for bifurcation geometry. Left: comparison at different time instances, right: comparison of constitutive models at $t = 0.36$ s.

The WSS distribution along the moving boundary for both types of geometry is presented in Figs. 15, 16. We see that the peak of the WSS corresponds to the stenosed area (see Fig. 15) and to the bifurcation divider (see Fig. 16). In the case of stenosed vessel we observe one reversed vortex at $t = 0.36$ s and two vortices at $t = 0.58$ s. Moreover, at time instances $t = 0.58$ s and $t = 0.90$ s, the WSS belongs to the athero-prone range. Looking at the bifurcation geometry (Fig. 16), a reversed flow at all time instances around the sinus bulb appears. Moreover, approaching the bifurcation divider lower values of the WSS can be found. In both cases, stenotic iliac and bifurcation carotid artery, we observe that the WSS corresponding to the non-Newtonian model gives higher extremal values.

In Fig. 17 the OSI index along the moving boundary for stenotic (left) and bifurcation geometry (right) is presented. Due to the high shear flow in a stenosed region, direction-varying WSS, in particular before and after stenosis, can be found.

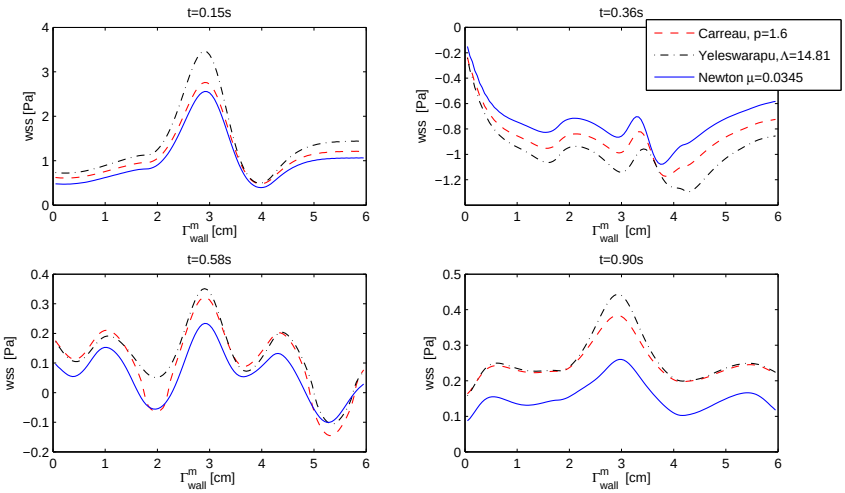


FIGURE 15. WSS along Γ_{wall} for the stenotic vessel geometry at several time instances.

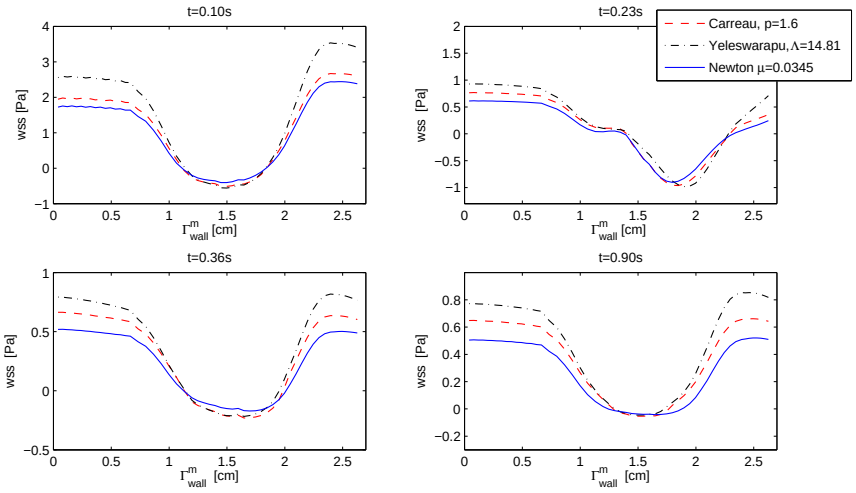


FIGURE 16. WSS along Γ^m_{wall} for the bifurcation geometry at several time instances.

In the case of carotid bifurcation, the OSI peak corresponds to the sinus bulb area. These measurements agree with the observations from the clinical praxis, e.g. [30].

Finally, Figs. 18, 19 present several snapshots for the stenotic and bifurcation geometry depicting the velocity streamlines, velocity vector field, pressure isolines and horizontal velocity isolines. In Fig. 18a we observe the development of recirculation

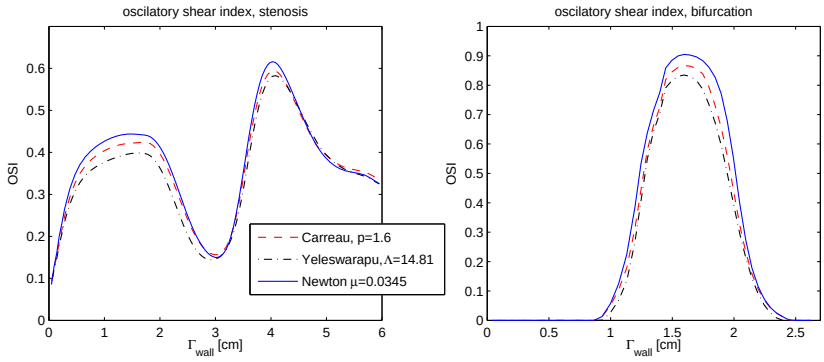
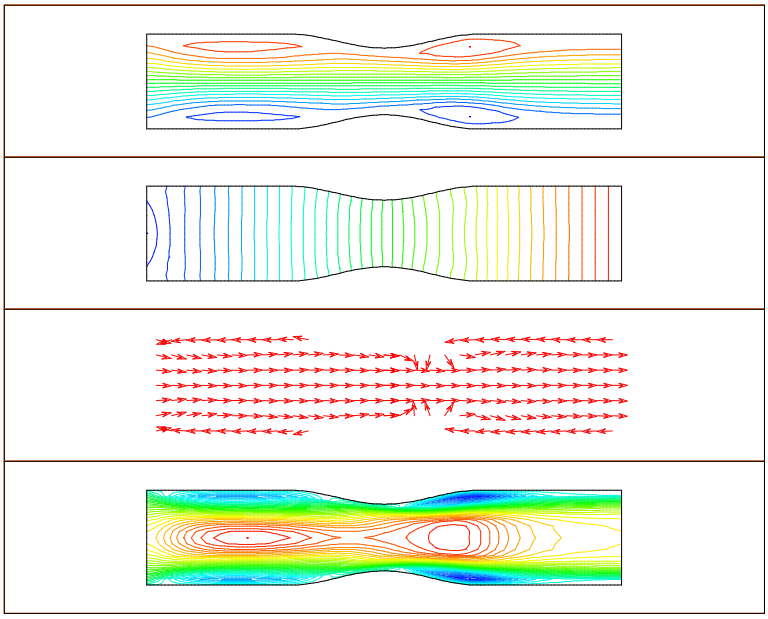


FIGURE 17. OSI along Γ_{wall}. Left: stenosed geometry, right: bifurcation geometry.

zones around the stenosed area. At the systolic minimum (Fig. 18b) we can see that the negative flow with pressures from $[-51.6, 0]$ Pa and horizontal velocities from $[-21.5, 7.8]$ cm.s^{-1} develops. In the case of bifurcation geometry, we observe the development of reversed flow in particular in the sinus bulb area (Fig. 19). We note that due to the bifurcation geometry the axial velocity profiles in daughter vessels are asymmetric. For more details on physiological experiments, including the evolution of WSS for chosen points on moving boundary, see [28].

a) $t = 0.27$ s



b) $t = 0.36$ s

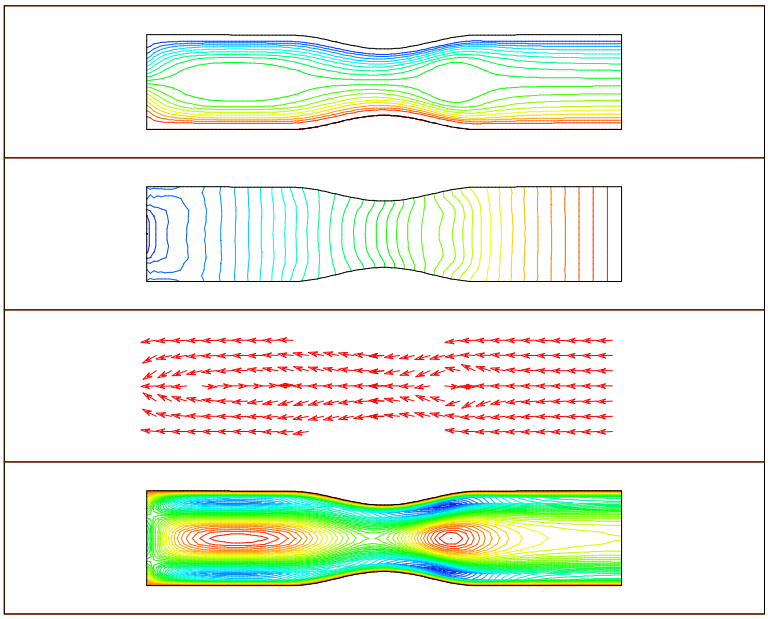
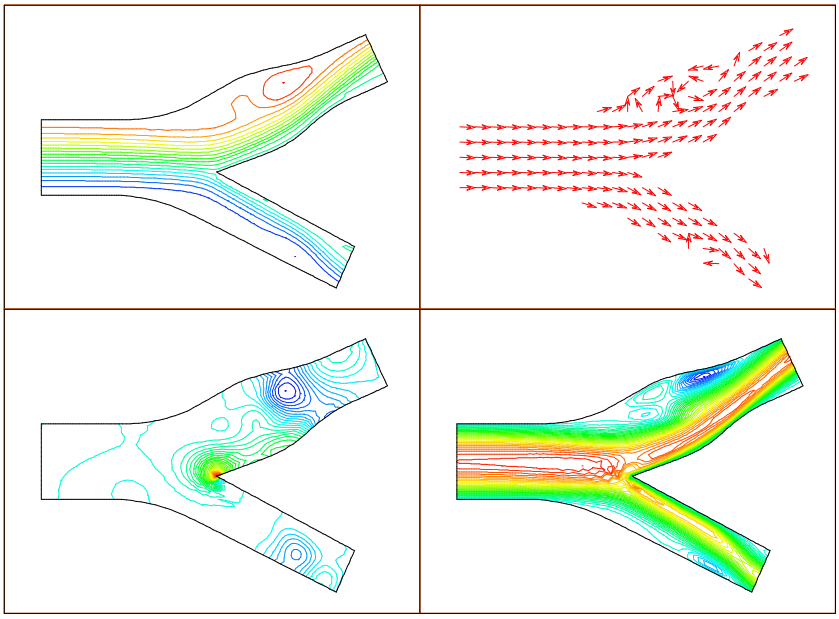


FIGURE 18. Velocity streamlines, pressure isolines, velocity vector field and horizontal velocity isolines for the stenosed vessel at two time instances.

a) $t = 0.23$ s



b) $t = 0.36$ s

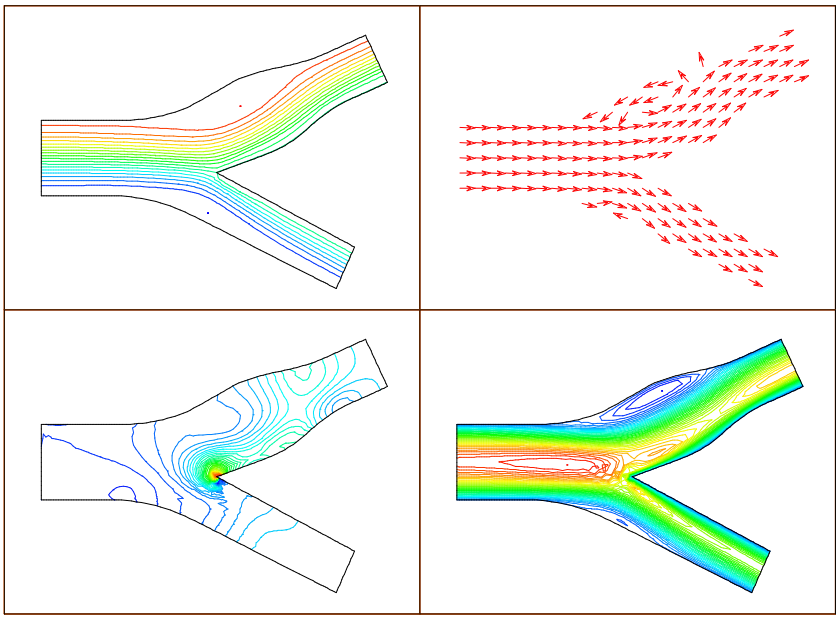


FIGURE 19. Velocity streamlines, velocity vector field, pressure isolines and horizontal velocity isolines for the bifurcation geometry at two time instances.

1.5. Convergence study. In our numerical experiments we have used piecewise linear approximations for fluid velocities and for pressure and backward Euler method for time discretization. In the case of global iterative method the structure equation is approximated by the second order Newmark method. For the kinematical splitting method the second order approximation in space and time (Crank-Nicolson method) was applied. Both the first order Marchuk-Yanenko splitting as well as the second order Strang splitting schemes have been tested.

In order to study accuracy of the coupled fluid-structure interaction problem the so-called **experimental order of convergence** (EOC) is computed. The EOC in space is defined in the following way

$$EOC(\boldsymbol{u}) = \log_2 \frac{\|\boldsymbol{u}_{h,\Delta t} - \boldsymbol{u}_{\frac{h}{2},\Delta t}\|_{L^p} / \|\Omega_h\|_{L^p}}{\|\boldsymbol{u}_{\frac{h}{2},\Delta t} - \boldsymbol{u}_{\frac{h}{4},\Delta t}\|_{L^p} / \|\Omega_{\frac{h}{2}}\|_{L^p}} \,. \tag{1.13}$$

To evaluate the EOC in space, the computational domain $\Omega(\eta)$ is consecutively divided into 16×2 elements (1. refinement), 32×4 elements (2. refinement), 64×8 elements (3. refinement), 128×16 elements (4. refinement), where the element size $h = (\Delta x, \Delta y)$ is halved. The space errors and the EOC were computed at $T = 0.8s$. The constant time step was chosen to be enough small and set to $\Delta t = 0.002s$.

We consider here the Carreau model for non-Newtonian fluid as well as the Newtonian fluid. The index p denotes the corresponding exponent in the power-law model used for the non-Newtonian Carreau viscosity function, see (2.2). In our case we took $q = -0.2$, which yields $p = 1.6$ as well as $q = 0$ ($p = 2$) in the Newtonian case. Note that due to the regularity results presented in Subsection 4.1.1 we measure the errors in the L^p, or $W^{1,p}$, norm for velocity for the non-Newtonian fluid and in the L^2, or $W^{1,2}$, norm for the Newtonian fluid. Due to the artificial compressibility regularization of the continuity equation we can measure pressure in the L^2 norm.

Let us firstly present the convergence results in space in term of the EOC values for velocity and pressure in a rigid domain, Tab. 3. For each quantity a following notation for the normalized L^p-error was used

$$Err(\boldsymbol{u}) = \frac{\|\boldsymbol{u}_{h,\Delta t} - \boldsymbol{u}_{h/2,\Delta t}\|_{L^p}}{\|\Omega_h\|_{L^p}}.$$

Table 3 presents convergence results in a fixed domain with the symmetry boundary

TABLE 3. Convergence rates in space, rigid domain.

	Newtonian fluid				Carreau fluid, $q = -0.2$			
# of refin.	$Err(\boldsymbol{u})$	EOC	$Err(p)$	EOC	$Err(\boldsymbol{u})$	EOC	$Err(p)$	EOC
	L^2 norm				L^p norm		L^2 norm	
2/1	1.0783		3.5199		0.9083		3.7209	
3/2	0.2758	1.967	0.6870	2.357	0.2494	1.865	0.7073	2.395
4/3	0.084	1.715	0.3204	1.101	0.1092	1.192	0.1577	2.165

conditions at the central line, see Fig. 20. We can notice slightly worse than second order convergence rate in velocity for the Newtonian case. Moreover, in the non-Newtonian case the convergence in velocity is reduced to 1. This effect can be explained by the influence of symmetry boundary conditions coupled with

the Neumann boundary conditions. On the other hand this boundary conditions improve convergence of pressure in the non-Newtonian case to the second order.

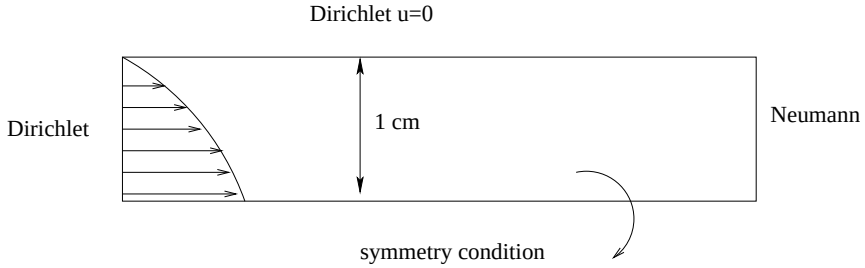


FIGURE 20. Boundary conditions in measurements of EOC .

In the next experiment we compare the order of convergence of the kinematical splitting algorithm, see Tab. 5 and the global iterative method, see Tab. 4. We can

TABLE 4. Convergence rates in space: global iterative method, the Carreau model.

# of refin.	$Err\ (\boldsymbol{u})$	EOC	$Err\ (\eta)$	EOC	$Err\ (p)$	EOC
	L^p-norm		L^2-norm			
2/1	0.9512		2.81 e-3		3.3925	
3/2	0.2563	1.89	8.88 e-4	1.69	0.7113	2.25
4/3	0.1074	1.26	1.85 e-4	2.23	0.1577	2.17

TABLE 5. Convergence rates in space: kinematical splitting (Marchuk-Yanenko), the Carreau model.

# of refin	$Err\ (\boldsymbol{u})$	EOC	$Err(\nabla \boldsymbol{u})$	EOC	$Err\ (\eta)$	EOC	$Err\ (p)$	EOC
	L^p-norm				L^2-norm			
2/1	0.8971		0.9682		2.62 e-4		3.1338	
3/2	0.2466	1.86	0.1408	2.78	1.84 e-5	0.51	0.7026	2.16
4/3	0.1051	1.23	0.0435	1.69	0.38 e-5	2.26	0.1461	2.27

clearly see the similar convergence rates in velocities, pressures and displacements that are higher than first order for velocities and the second order for pressures and wall displacements. We also observe decreasing convergence rate for velocity for finer meshes, which was also observed in the rigid domain, Table 3, and may be caused by the boundary conditions. Let us point out that the kinematical splitting approach yields 10 times smaller relative errors in the wall displacement than the strong coupling scheme. This weak coupling method is also more efficient as the global iterations with respect to the domain geometry are not needed anymore.

In the third series of experiments we compare the kinematical splitting method for both, the Newtonian (Tab. 6) and the non-Newtonian Carreau (Tab. 5) fluid. For the Newtonian fluid we observe again higher order convergence rate in velocity.

Considering the non-Newtonian rheology the convergence rate in pressure is of the second order. It is typically better than in the Newtonian case. Since the convergence of η depends on the convergence rates of $\nabla \boldsymbol{u}$ and p, we see in the case of non-Newtonian rheology improvement of the convergence order for the wall displacement, too.

TABLE 6. Convergence rates in space for Marchuk-Yanenko splitting, Newtonian viscosity.

# of refin.	*Err* ($\boldsymbol{u}$)	*EOC*	*Err*($\nabla \boldsymbol{u}$)	*EOC*	*Err* (η)	*EOC*	*Err* (p)	*EOC*
	L^2-norm							
2/1	1.0566		1.2723		1.56e-4		3.0076	
3/2	0.2780	1.93	0.2171	2.55	1.94e-4	-0.32	0.7081	2.09
4/3	0.0872	1.67	0.0483	2.17	1.12e-4	0.79	0.0313	1.18

Now we compare the convergence results in space for four possible schemes of the kinematical splitting method, i.e. explicit Marchuk-Yanenko splitting (Tab. 7), implicit Marchuk-Yanenko splitting (Tab. 8), explicit Strang splitting (Tab. 9) and implicit Strang splitting (Tab. 10). As it is expected there is no significant

TABLE 7. Convergence rates in space; explicit Marchuk-Yanenko splitting, the Carreau viscosity.

# of refin	*Err* ($\boldsymbol{u}$)	*EOC*	*Err*($\nabla \boldsymbol{u}$)	*EOC*	*Err* (η)	*EOC*	*Err* (p)	*EOC*
	L^p-norm				L^2-norm			
2/1	0.8667		0.9292		1.01 e-4		3.0097	
3/2	0.2338	1.89	0.1283	2.86	0.66 e-5	0.61	0.6451	2.22
4/3	0.1032	1.18	0.0437	1.55	0.25 e-5	1.37	0.1791	1.85

TABLE 8. Convergence rates in space; implicit Marchuk-Yanenko splitting, the Carreau viscosity.

# of refin	*Err* ($\boldsymbol{u}$)	*EOC*	*Err*($\nabla \boldsymbol{u}$)	*EOC*	*Err* (η)	*EOC*	*Err* (p)	*EOC*
	L^p-norm				L^2-norm			
2/1	1.0866		0.9470		1.04e-4		3.4332	
3/2	0.3300	1.72	0.1626	2.54	4.26e-5	1.29	0.8190	2.07
4/3	0.1034	1.67	0.0442	1.88	2.51e-5	0.76	0.1960	2.06

improvement in the convergence rate in space due to the Strang splitting, however the implicit scheme improves the convergence rate.

The last series of experiments focuses on the evaluation of the EOC in time, that is given as follows

$$EOC(\boldsymbol{u}) = \log_2 \frac{\left(\sum_{j=1}^{N} ||\boldsymbol{u}^j_{h,\Delta t} - \boldsymbol{u}^j_{h,\Delta t/2}||^p_{L^p} / |\Omega^j_{h,\Delta t}|^p \right)^{1/p}}{\left(1/2 \sum_{j=1}^{2N} ||\boldsymbol{u}^j_{h,\Delta t/2} - \boldsymbol{u}^j_{h,\Delta t/4}||^p_{L^p} / |\Omega^j_{h,\Delta t/2}|^p \right)^{1/p}}. \tag{1.14}$$

TABLE 9. Convergence rates in space; explicit Strang splitting, the Carreau viscosity.

# of refin	*Err* ($\boldsymbol{u}$)	*EOC*	*Err*($\nabla\boldsymbol{u}$)	*EOC*	*Err* (η)	*EOC*	*Err* (p)	*EOC*
	L^p-norm				L^2-norm			
2/1	0.8015		0.8274		9.14e-4		3.0562	
3/2	0.2144	1.90	0.1170	2.82	9.77e-5	3.23	0.5403	2.50
4/3	0.1000	1.10	0.0447	1.39	8.34e-5	0.22	0.1580	1.77

TABLE 10. Convergence rates in space; implicit Strang splitting, the Carreau viscosity.

# of refin	*Err* ($\boldsymbol{u}$)	*EOC*	*Err*($\nabla\boldsymbol{u}$)	*EOC*	*Err* (η)	*EOC*	*Err* (p)	*EOC*
	L^p-norm				L^2-norm			
2/1	0.8646		0.9260		1.23e-4		3.0057	
3/2	0.2332	1.89	0.1280	2.86	6.11e-5	1.01	0.6420	2.23
4/3	0.1032	1.18	0.0444	1.53	2.77e-5	1.14	0.1933	1.73

Here $\boldsymbol{u}^j_{h,\Delta t}$ is the velocity at the time instance $j\Delta t$. For the EOC in time, going from one time refinement to the other, the time step is halved. The time interval was $t \in [0.2; 0.8]$ and the initial time step was taken $\Delta t = 0.025$ s. In order to evaluate the EOC in time we compute also the normalized relative $L^p(0, T; L^p(\Omega))$ error in time. This is defined by

$$Err(\boldsymbol{u}) = \frac{1}{T}\left(\sum_{j=1}^{N} \Delta t \left(\frac{||\boldsymbol{u}^j_{h,\Delta t} - \boldsymbol{u}^j_{h,\Delta t/2}||_{L^p}}{|\Omega^j_{h,\Delta t}|}\right)^p\right)^{1/p}, \quad T = N\Delta t.$$

Similarly as before, we compare the explicit and implicit Marchuk-Yanenko kinematical splitting scheme (Tabs. 11, 12) and the explicit and implicit Strang splitting scheme (Tabs. 13, 14). For the explicit kinematical splitting scheme the EOC is around first order. The second order explicit Strang splitting improves the convergence orders in comparison to the Marchuk-Yanenko splitting. The implicit kinematical splitting schemes yield better convergence than the explicit kinematical splitting schemes. Finally, Tab. 14 shows convergence rates for the implicit Strang splitting that are significantly improved. We can conclude that the Strang splitting strategy gives better convergence results for both, the explicit and the implicit schemes.

TABLE 11. Convergence rates in time; explicit Marchuk-Yanenko splitting, the Carreau model.

# of refin ($\triangle t$)	*Err* ($\boldsymbol{u}$)	*EOC*	*Err*($\nabla\boldsymbol{u}$)	*EOC*	*Err* (η)	*EOC*	*Err* (p)	*EOC*
	$L^p(L^p)$-norm				$L^2(L^2)$-norm			
2/1	0.0246		0.0159		0.0088		0.2905	
3/2	0.0132	0.89	0.0088	0.86	0.0060	0.56	0.1422	1.03
4/3	0.0070	0.92	0.0046	0.93	0.0041	0.53	0.0697	1.03
5/4	0.0042	0.74	0.0030	0.61	0.0016	1.40	0.0336	1.05

TABLE 12. Convergence rates in time; implicit Marchuk-Yanenko scheme the Carreau model.

# of refin ($\triangle t$)	$Err\ (\boldsymbol{u})$	EOC	$Err(\nabla \boldsymbol{u})$	EOC	$Err\ (\eta)$	EOC	$Err\ (p)$	EOC
	$L^p(L^p)$-norm				$L^2(L^2)$-norm			
2/1	0.1491		0.1633		0.1640		0.5616	
3/2	0.1532	-0.03	0.1600	0.03	0.2706	-0.72	0.4332	0.37
4/3	0.0705	1.12	0.0747	1.10	0.2000	0.44	0.2286	0.92
5/4	0.0218	1.69	0.0234	1.67	0.0915	1.13	0.0683	1.74

TABLE 13. Convergence rates in time; explicit Strang splitting, the Carreau model.

# of refin ($\triangle t$)	$Err\ (\boldsymbol{u})$	EOC	$Err(\nabla \boldsymbol{u})$	EOC	$Err\ (\eta)$	EOC	$Err\ (p)$	EOC
	$L^p(L^p)$-norm				$L^2(L^2)$-norm			
2/1	0.0564		0.0252		0.0583		0.3363	
3/2	0.0195	1.53	0.0081	1.65	0.0234	1.32	0.1539	1.13
4/3	0.0077	1.34	0.0024	1.75	0.0089	1.40	0.0712	1.11
5/4	0.0044	0.83	0.0013	0.90	0.0054	0.72	0.0315	1.18

TABLE 14. Convergence rates in time; implicit Strang splitting, the Carreau model.

# of refin ($\triangle t$)	$Err\ (\boldsymbol{u})$	EOC	$Err(\nabla \boldsymbol{u})$	EOC	$Err\ (\eta)$	EOC	$Err\ (p)$	EOC
	$L^p(L^p)$-norm				$L^2(L^2)$-norm			
2/1	0.1826		0.1936		0.2211		0.5969	
3/2	0.0578	1.66	0.0609	1.67	0.1140	0.96	0.2411	1.31
4/3	0.0241	1.26	0.0243	1.32	0.0441	1.37	0.0896	1.43
5/4	0.0088	1.44	0.0078	1.64	0.0173	1.35	0.0297	1.60

2. Concluding remarks

In this overview paper we have summarized our recent results on mathematical modelling and numerical simulation of fluid-structure interaction of a shear-dependent non-Newtonian fluid and a viscoelastic membrane. We have presented mathematical models for both the shear-dependent fluids and the generalized string equation. Further, we have derived two fluid-structure interaction methods, the global iterative method belonging to the class of strongly coupled schemes, and the kinematical splitting which is a weakly coupled method. The global iterative method has been also used in order to prove existence of a weak solution to fully coupled interactions between the shear-dependent fluids and the viscoelastic structure. As far as we know the result presented in [25] is the first contribution to the well-posedness of fluid-structure interaction for non-Newtonian fluids.

The kinematical splitting yields a numerical scheme that is more efficient than the global iterative method while having typically smaller global errors. We have analysed stability of the semi-discrete kinematical splitting method and shown that depending on the choice of discretization for the ALE convective term we may obtain a semi-discrete scheme that does or does not depend on a time step. Indeed, using the implicit Euler time discretization we obtain the stability condition for time

step depending on a mesh velocity. If the midpoint rule is used in order to discretize the ALE convective term the semi-discrete kinematical splitting is unconditionally stable, cf. also [28].

An application which is of particular interest is blood flow in elastic vessels. We have simulated blood flow in a stenotic vessel and a carotid bifurcation and analyzed some hemodynamical control quantities. We have modeled blood as a shear-thinning non-Newtonian fluid and chosen two well-known models, the Carreau model and the Yeleswarapu model. Comparisons with the Newtonian model are presented as well.

Further, we have investigated the wall deformation and the hemodynamical wall parameters, the wall shear stress WSS and the oscillatory shear index OSI, for a straight and stenotic vessel as well as for a carotid bifurcation. Numerical simulations demonstrate a significant influence of the non-Newtonian rheology for hemodynamical wall parameters. According to some authors [31] negative values of WSS indicates occurance of recirculation zones and reversal flows around stenosis, which seems to be better predicted by the non-Newtonian models. Further, the domain geometry has also a considerable influence on the wall deformation as well as on the WSS and OSI. The maximum values of OSI are larger for the Newtonian fluid. Such high OSI values at the end of stenotic occlusion indicate a large oscillatory nature of the wall shear stress and could yield further to additional stenotic plug.

Comparisons of WSS and OSI for a solid and compliant vessel showed significantly higher oscillations of the wall shear stress for fixed solid vessels. This leads to the conclusion that the fluid-structure interaction aspect is important for hemodynamical modelling and should be involved in reliable computational models.

It would be interesting to extend mathematical models and consider the generalized Oldroyd-B model that includes viscoelastic properties of blood as well. More realistic models for vessel walls, see, e.g., [6], allowing deformation in both directions would be more appropriate in order to consider more complex vessel geometries. An important point of numerical simulation is a correct outflow boundary condition, reflecting the influence of the rest of circulatory system. According to [45] this can be realized by the so-called impedance condition arising from coupling the fluid equations with some less dimensional model (1D or 0D lumped model).

Bibliography

[1] Bastian P., Johannsen K., Reichenberger V.: *UG tutorial*, 2001.

[2] Bodnár T., Sequeira A.: Numerical study of the significance of the non-Newtonian nature of blood in steady flow through a stenosed vessel, In: Rannacher R. et.al. (Eds.), *Advances in Mathematical Fluid Mechanics*, 83–104, Springer Verlag, 2010.

[3] Broser Ph. J.: *Simulation von Strömungen in Blutgefässen*, Master Thesis, Ruprechts-Karl University, Heidelberg, 2001.

[4] Bucur D., Fereisl E., Nečasová Š.: Influence of wall roughness on the slip behavior of viscous fluids, *Proc. R. Soc. Edinb., Sect. A, Math.* 138(5) (2008), 957-973.

[5] Bucur D., Feireisl E., Nečasová Š., Wolf J.: On the asymptotic limit of the Navier-Stokes system on domains with rough boundaries, J. of Differ. Equations 244 (2008), 2890–2908.

[6] Čanić S., Tambača J., Guidoboni G., Mikelič A., Hartley C.J., Rosenstrauch D.: Modeling viscoelastic behavior of arterial walls and their intercation with pulsatile blood flow, *SIAM J. Appl. Math.* 67(1) (2006), 164–193.

[7] Chambolle A., Desjardin B., Esteban M.J., Grandmont C.: Existence of weak solutions for unsteady fluid-plate interaction problem, *J. Math. Fluid Mech.* 7 (3) (2005), 368–404.

[8] Černý J.: Numerical modelling of non-Newtonian flows with application in hemodynamics, *Report TU Hamburg-Harburg*, 2004.

[9] Cheng A., Shkoller S.: The interaction of the 3D Navier-Stokes equations with a moving nonlinear Koiter elastic shell, *SIAM J. Math. Anal.* 42(3) (2010), 1094-1155.

[10] Cheng A., Coutand D., Shkoller S.: Navier-Stokes equations interacting with a nonlinear elastic biofluid shell, *SIAM J. Math. Anal.* 39(3) (2007), 742-800.

[11] Coutand D., Shkoller S.: Motion of an elastic solid inside an incompressible viscous fluid, *Arch. Ration. Mech. Anal.* 176(1) (2005), 25-102.

[12] Coutand D., Shkoller S.: Interaction between quasilinear elasticity and the Navier-Stokes equations, *Arch. Ration. Mech. Anal.* 179 (2006), 303-352.

[13] Filo J., Zaušková A.: 2D Navier-Stokes equations in a time dependent domain with Neumann type boundary conditions, *J. Math. Fluid Mech.* 10 (2008), 1–46.

[14] Formaggia L., Gerbeau J.F., Nobile F., Quarteroni A.: On the coupling of 3D and 1D Navier-Stokes equations for flow problems in compliant vessels, *Comput. Meth. Appl. Mech. Eng.* 191 (6-7) (2001), 561–582.

[15] Formaggia L., Nobile F.: A stability analysis for the arbitrary Lagrangian Eulerian formularion with finite elements, *East-West J. Numer. Math.* 7 (2) (1999), 103–131.

[16] Förster C., Scheven von M., Wall W.A., Ramm E.: Coupling of incompressible flows and thin-walled structures, In: Papadrakakis M. et. al. (Eds.) *Coupled Problems 2007 International Conference on Computational Methods for Coupled Problems in Science and Engineering*, 541–544, Greece, 2007.

[17] Frehse J., Málek J., Steinhauer M.: On analysis of steady flow of fluid with shear dependent viscosity based on the Lipschitz truncation method, *SIAM J. Math. Anal.* 35(5) (2003), 1064–1083.

[18] Galdi G.P., Rannacher R., Robertson A.M., Turek S.: *Hemodynamical Flows. Modeling, Analysis, Simulation.*, Oberwolfach Seminars Vol. 37, Birkhäuser Verlag, 2008.

[19] Gijsen F.J.H., van de Vosse F.N., Janssen J.D.: Influence of the non-Newtonian properties of blood flow on the flow in large arteries: steady flow in a carotid bifurcation model, *J. Biomech.* 32 (1999), 601–608.

[20] Grandmont C.: Existence of weak solutions for the unsteady interaction of a viscous fluid with an elastic plate, *SIAM J. Math. Anal.* 40(2) (2008), 716-737.

[21] Guidoboni G., Glowinski R., Cavallini N., Čanić S.: Stable-loosely coupled-type algorithm for fluid-structure interaction on blood flow, *J. Comput. Phys.* 228 (18) (2009) 6916–6937.

[22] Guidoboni G., Guidorzi M., Padula M.: Continuous dependence on initial data in fluid-structure motion, *J. Math. Fluid Mech.* (2010) published online: http://dx.doi.org/10.1007/s00021-010-0031-0.

[23] Guidorzi M., Padula M., Plotnikov P.I.: Hopf solutions to a fluid-elastic interaction model, *Math. Models Meth. Appl. Sci.* 18 (2) (2008), 215–270.

[24] Hundertmark-Zaušková A., Lukáčová-Medviďová M.: Numerical study of shear-dependent non-Newtonian fluids in compliant vessels, *Comput. Math. Appl.* 60 (2010), 572–590.

[25] Hundertmark-Zaušková A., Lukáčová-Medviďová M., Nečasová Š.: On the existence of a weak solution to the coupled fluid-structure interaction problem for the non-Newtonian shear-dependent fluid, *Preprint University of Mainz*, 2011.

[26] Lesoinne M., Farhat Ch.: Geometric conservation laws for flow problems with moving boundaries and deformable meshes and their impact on aeroelastic computations, *Comput. Meth. Appl. Mech. Eng.* 134 (1996), 71–90.

[27] Lukáčová-Medviďová M., Černý J.: Numerical modelling of non-Newtonian shear-thinning viscoelastic fluids with application in hemodynamics, In: Sequeira A. et.al. (Eds.) *Proceedings of the Second International Symposium Modelling of Physiological Flows*, Portugal, 2005.

[28] Lukáčová-Medviďová M., Rusnáková G.: Kinematical splitting algorithm for fluid-structure interaction in hemodynamics, *Preprint University of Mainz & University of Košice*, 2011.

[29] Lukáčová-Medviďová M., Zaušková A.: Numerical modelling of shear-thinning non-Newtonian flows in compliant vessels, *Int. J. Numer. Methods Fluids*, 56 (2008), 1409–1415.

[30] Malek A.M., Alper A.L., Izumo S.: Hemodynamic shear stress and its role in atherosclerosis, *J. Am. Med. Assoc.* 282 (1999), 2035–2042.

[31] Nadau L., Sequeira A.: Numerical simulations of shear dependent viscoelastic flows with a combined finite element - finite volume method, *Comput. Math. Appl.* 53 (2007), 547–568.

[32] Nägele S.: *Mehrgitterverfahren für incompressiblen Navier-Stokes Gleichungen im laminaren und turbulenten Regime unter Berücksichtigung verschiedener Stabilisierungsmethoden*, PhD Thesis, Ruprechts-Karl University, Heidelberg, 2003.

[33] Neustupa J.: Existence of weak solution to the Navier-Stokes equation in a general time-varying domain by Rothe method, *Math. Methods Appl. Sci.* 32(6) (2009), 653–683.

[34] Málek J., Nečas J., Růžička M.: On the weak solutions to a class of non-Newtonian incompressible fluids in bounded three-dimensional domans: the case $p \geq 2$, *Adv. Differ. Equat.* 6(3) (2001), 257–302.

[35] Málek J., Nečas J., Rokyta M., Růžička M.: *Weak and Measure-Valued Solutions to Evolutionary PDEs*, Chapman and Hall, London, 1996.

[36] Nobile F.: *Numerical Approximation of Fluid-Structure Interaction Problems with Application to Haemodynamics*, PhD Thesis, EPFL Lausanne, 2001.

[37] Perktold K., Rappitsch G.: Mathematical modelling of local arterial flow and vessel mechanics, In: Crolet J.-M. et.al. (Eds.) *Computational Methods for Fluid-Structure Interaction*, 230–245, Longman, Harlow, 1994.

[38] Perktold K., Resch M., Reinfried O.P.: Three-dimensional analysis of pulsative flow and wall shear stress in the carotid artery bifurcation, *J. Biomech.* 24(6) (1991), 409–420.

[39] Quarteroni A.: Mathematical and numerical simulation of the cardiovascular system, In: *Proceedings of the International Congress of Mathematicians, Vol. III*, 839–849, Higher Ed. Press, Beijing, 2002.

[40] Quarteroni A., Formaggia L.: Mathematical modelling and numerical simulation of the cardiovascular system, In: Ciarlet P.G. et. al. (Eds.) *Handbook of Numerical Analysis*, Elsevier, Amsterdam, 2002.

[41] Quarteroni A., Gianluigi R.: Optimal control and shape optimization of aorto-coronaric bypass anastomoses, M^3AS, *Math. Models Methods Appl. Sci.* 13 (12) (2003), 1801–1823.

[42] Quarteroni A., Valli A.: *Numerical Approximation of Partial Differential Equations*, Springer Verlag, 2008.

[43] Rajagopal K., Lawson J.: Regulation of hemostatic system function by biochemical and mechanical factors, Modeling of Biological Materials, In: Bellomo N. (Ed.) *Modeling and Simulation in Science, Engineering and Technology*, 179–201, Birkhäuser, Boston, 2007.

[44] Surulescu C.: On the stationary interaction of a Navier-Stokes fluid with an elastic tube wall, *Appl. Anal.* 86(2) (2007), 149–165.

[45] Taylor Ch.A., Vignon-Clementel I.E., Figueroa C.A., Jansenca K.E.: Outflow boundary conditions for three-dimensional finite element modeling of blood flow and pressure in arteries, *Comput. Meth. Appl. Mech. Eng.* 195 (2006), 3776–3796.

[46] Taylor Ch.A.., Hughes J.R., Zarins C.K.: Effect of exercise on hemodynamic conditions in the abdominal aorta, *J. Vasc. Surg.*, 29 (6) (1999), 1077–1089.

[47] Beirão da Veiga H.: On the existence of strong solution to a coupled fluid-structure evolution problem, *J. Math. Fluid Mech.* 6 (1) (2001), 21–52.

[48] Yeleswarapu K.K.: *Evaluation of Continuum Models for Chatacterizing the Constitutive Behavior of Blood*, PhD Thesis, University of Pittsburg, 1996.

[49] Yeleswarapu K.K., Kameneva M.V., Rajagopal K.R., Antaki J.F.: The flow of blood in tubes: theory and experiments, *Mech. Res.Commun.* 25 (3) (1998), 257–262.

[50] Wolf J.: Existence of weak solution to the equations of non stationary motion of non-Newtonian fluids with shear rate dependent viscosity, *J. Math. Fluid Mech.* 9 (2007) 104–138.

[51] Zaušková A.: *2D Navier-Stokes Equations In a Time Dependent Domain*, PhD Thesis, Comenius University, Bratislava, 2007.

Part 5

Analysis in Orlicz spaces

Ron Kernan

2000 *Mathematics Subject Classification.* Primary 46E35, 35J65

Key words and phrases. Orlicz spaces, gauge norms, duality, decreasing rearrangement, rearrangement invariance, elliptic partial differential equations

Abstract. The purpose of these notes is to present in a direct manner the theory of a class of function spaces introduced and investigated by Wladyslav Orlicz in the 1930's. The theory is then applied to certain nonlinear elliptic boundary value problems.

Acknowledgement. The author is grateful to the Nečas Center—in particular to Josef Málek and Luboš Pick—for the opportunity to give the lectures. He would like to thank Martin Franců, Colin Phipps and Susanna Spektor for their help in preparing these notes.

Contents

CHAPTER 1

Analysis in Orlicz spaces

1. Introduction

The purpose of these notes is to present in a direct manner the theory of a class of function spaces introduced and investigated by Wladyslaw Orlicz in the 1930's. Their goal is to facilitate the application of these spaces to problems in partial differential equations.

Orlicz had in mind the application of his spaces to the convergence of orthogonal series. Thus, it seems appropriate to begin by describing how they arise in the study of classical Fourier series.

First, we recall that for measurable $f : \Omega \to \mathbb{R}$, $\Omega \subset \mathbb{R}^n$, and $1 \leq p < \infty$,

$$\|f\|_{p,\Omega} := \left[\int_\Omega |f(x)|^p \, dx\right]^{\frac{1}{p}},$$

while

$$\|f\|_{\infty,\Omega} := \operatorname{ess\,sup}_{x\in\Omega} |f(x)| .$$

It is well-known that

$$L_p(\Omega) := \left\{f : \Omega \to \mathbb{R} : \|f\|_{p,\Omega} < \infty\right\}$$

is a Banach space under the norm $\| \quad \|_{p,\Omega}$.

The connection of the L_p spaces to Fourier series is as follows. Given $f \in L_1(I)$, $I = [-\pi, \pi]$, its Fourier series is

$$\frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos kx + b_k \sin kx,$$

in which

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(y) \cos ky \, dy, \; k = 0, 1, \ldots$$

and

$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(y) \sin ky \, dy, \; k = 1, 2, \ldots .$$

Now, the pointwise convergence of the K^{th} partial sum of the Fourier series of f, namely

$$(S_K f)(x) := \frac{a_0}{2} + \sum_{k=1}^{K} a_k \cos kx + b_k \sin kx,$$

to $f(x)$, $\frac{f(x-o)+f(x+o)}{2}$ or anything else was found to be difficult to prove. To overcome this problem other kinds of convergence were considered. Thus, J. P. Gram in 1883 showed that for suitable f

$$\lim_{K\to\infty}\int_{-\pi}^{\pi}|f(x)-(S_Kf)(x)|^2\,dx=0;$$

that is, S_Kf converges to f in $L_2(I)$.

This was later refined in 1907 by F. Riesz and E. Fischer to the assertion that, in our notation,

$$\lim_{K\to\infty}\|f-S_Kf\|_{2,I}=0$$

if and only if $f\in L_2(I)$. Then, in 1927, M. Riesz proved that, for $1<p<\infty$,

$$\lim_{K\to\infty}\|f-S_Kf\|_{p,I}=0 \tag{1.1}$$

if and only if $f\in L_p(I)$. However, there are functions $f\in L_1(I)$ and $g\in L_\infty(I)$ such that

$$\lim_{K\to\infty}\|f-S_Kf\|_{1,I}\neq 0$$

and

$$\lim_{K\to\infty}\|g-S_Kg\|_{\infty,I}\neq 0.$$

One may ask if there is a condition stronger than

$$\int_{-\pi}^{\pi}|f(x)|\,dx<\infty$$

that guarantees (1.1) when $p=1$. A. Zygmund, in 1928, showed that such a condition is

$$\int_{-\pi}^{\pi}|f(x)|\log_+|f(x)|\,dx<\infty.$$

As will be seen later, this is the condition for f to belong to the Orlicz space $L\log L(I)$.

When $p=\infty$, we don't ask if there is a condition stronger than

$$\|f\|_{\infty,I}=\operatorname{ess\,sup}_{x\in\Omega}|f(x)|<\infty$$

that guarantees (1.1), but rather if there is a norm smaller than $\|\quad\|_{\infty,I}$ in which we have convergence as in (1.1) for all $g\in L_\infty(I)$. There is indeed such a norm, namely the one for the Orlicz space $L_{exp}(I)$, to which g belongs if and only if

$$\int_{-\pi}^{\pi}\exp(k\,|g(x)|)dx<\infty \tag{1.2}$$

for some $k>0$. Observe that every function in $L_\infty(I)$ is in $L_{\exp}(I)$, as is the unbounded function $g(x):=\log\frac{\pi}{|x|}$, since any $k<1$ works in (1.2).

We conclude this introductory section by considering the more recent application of Orlicz spaces to Sobolev-type imbeddings. In them Ω will be a bounded *domain* in $\mathbb{R}^n$, a domain being an open connected set. The simplest case of the classical Sobolev imbedding inequality asserts that

$$\left[\int_\Omega|u(x)|^q\,dx\right]^{\frac{1}{q}}\leq C\left[\int_\Omega|(\nabla u)(x)|^p\,dx\right]^{\frac{1}{p}} \tag{1.3}$$

for all $u \in C_0^\infty(\Omega)$. Here,

$$\nabla := \left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots, \frac{\partial}{\partial x_n}\right)$$

is the usual gradient operator,

$$|(\nabla u)(x)| := \sqrt{\left(\frac{\partial u}{\partial x_1}\right)^2 + \left(\frac{\partial u}{\partial x_2}\right)^2 + \cdots + \left(\frac{\partial u}{\partial x_n}\right)^2}$$

and

$$\frac{1}{q} = \frac{1}{p} - \frac{1}{n},$$

where $1 < p < n$. Gagliardo and Nirenberg, independently, showed the result holds for $p = 1$.

When $p = n$, $\frac{1}{q} = 0$. One might expect that

$$\operatorname{ess\,sup}_{x\in\Omega} |u(x)| \le C \left[\int_\Omega |(\nabla u)(x)|^n \, dx\right]^{\frac{1}{n}}.$$

This is not the case. A norm smaller than that of $L_\infty(\Omega)$ is needed.

In the early 1960's, Yudovich, Pohozhaev and Trudinger each showed, Yudovich being the first, that, when $n = 2$, the norm of $L_\infty(\Omega)$ should be replaced by that of the Orlicz space $L_{\exp(t^2)}(\Omega)$. Strichartz later showed that for $n > 2$ the correct norm is that of $L_{\exp(t^{\frac{n}{n-1}})}(\Omega)$. Of course, $f \in L_{\exp(t^{\frac{n}{n-1}})}(\Omega)$ if and only if

$$\int_\Omega \exp\left(\left[k\,|f(x)|\right]^{\frac{n}{n-1}}\right) dx < \infty$$

for some $k > 0$.

This so-called limiting Sobolev inequality was used by the above-mentioned authors to prove the existence of eigenvalues for the Laplace operator.

2. The Orlicz class $L_\Phi(\Omega)$

As we saw in the examples in the previous section, membership in an Orlicz space is equivalent to a condition of the form

$$\int_\Omega \Phi\left(k\,|f(x)|\right) dx < \infty \tag{2.1}$$

for some $k > 0$. Often, (2.1) holds for all $k > 0$ if and only if it holds for one such k.

In the L_p case, $\Phi(t) = \Phi_p(t) = t^p$. Since we'll be generalizing L_p spaces we begin by asking what special properties Φ_p has when $1 < p < \infty$. The key property, it turns out, is

$$\Phi_p(t) = \int_0^t p s^{p-1} ds,$$

where $\phi_p(s) := p s^{p-1}$ increases strictly from $\phi_p(0) = 0$ to ∞ as $s \to \infty$.

DEFINITION 2.1. The function $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$, $\mathbb{R}_+ := (0, \infty)$, is said to be a **Young function** provided

$$\Phi(t) := \int_0^t \phi(s) ds,$$

in which $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is increasing and, say, left continuous, with $\phi(0+) = 0$ and $\lim_{t\to\infty} \phi(t) = \infty$.

We observe that Φ is convex, that is, if $\alpha, \beta \geq 0$,$\alpha + \beta = 1$, then,

$$\Phi(\alpha t_1 + \beta t_2) \leq \alpha\Phi(t_1) + \beta\Phi(t_2).$$

Indeed, the point $(\alpha t_1 + \beta t_2, \alpha\Phi(t_1) + \beta\Phi(t_2))$ is on the line

$$y = \Phi(t_1) + \frac{\Phi(t_2) - \Phi(t_1)}{t_2 - t_1}(t - t_1) = \Phi(t_1) + \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} \phi(s)ds(t - t_1)$$

joining $(t_1, \Phi(t_1))$ and $(t_2, \Phi(t_2))$ and that line lies above the graph of

$$y = \Phi(t) = \Phi(t_1) + \frac{1}{t - t_1}\int_{t_1}^{t} \phi(s)ds(t - t_1),$$

since the average value of ϕ on $[t_1, t]$, namely

$$\frac{1}{t - t_1}\int_{t_1}^{t} \phi(s)ds,$$

incerases on $[t_1, t_2]$.

EXAMPLE 2.2.

(1) The Orlicz norm of $L_{\log}L$ is given in terms of

$$\Phi(t) = \int_0^t \log(s + 1)ds = (t + 1)log(t + 1) - t,\ t \in \mathbb{R}_+.$$

(2) The Orlicz norm of $L_{\exp}$ is defined through

$$\Phi(t) = \int_0^t (e^s - 1)ds = e^t - t - 1,\ t \in \mathbb{R}_+.$$

DEFINITION 2.3. Let Φ be a Young function. Then, the corresponding **gauge norm** at a measurable function $f : \Omega \to \mathbb{R}$, $\Omega \subset \mathbb{R}^n$, is

$$\|f\|_{\Phi,\Omega} := \inf\left\{\lambda > 0 : \int_\Omega \Phi\left(\frac{|f(x)|}{\lambda}\right) dx \leq 1\right\}.$$

EXAMPLE 2.4.

(1) Consider $\Phi = \Phi_p$, $1 < p < \infty$. Then,

$$\int_\Omega \Phi_p\left(\frac{|f(x)|}{\lambda}\right) dx = \int_\Omega \left(\frac{|f(x)|}{\lambda}\right)^p dx \leq 1$$

if and only if

$$\int_\Omega |f(x)|^p\, dx \leq \lambda^p$$

or

$$\left[\int_\Omega |f(x)|^p\, dx\right]^{\frac{1}{p}} \leq \lambda.$$

So,

$$\|f\|_{\Phi_p,\Omega} = \left[\int_\Omega |f(x)|^p\, dx\right]^{\frac{1}{p}} = \|f\|_{p,\Omega}.$$

(2) Let Ω and E be measurable subsets of $\mathbb{R}^n$, with $E \subset \Omega$, $|E| < \infty$. Denote by χ_E the characteristic function of E.

Then, for any Young function Φ,

$$\|\chi_E\|_{\Phi,\Omega} = \frac{1}{\Phi^{-1}(|E|^{-1})}.$$

Indeed,

$$1 \geq \int_\Omega \Phi\left(\frac{\chi_E(x)}{\lambda}\right) dx = \int_E \Phi\left(\frac{1}{\lambda}\right) dx = \Phi\left(\frac{1}{\lambda}\right)|E|$$

amounts to

$$\lambda \geq \frac{1}{\Phi^{-1}\left(|E|^{-1}\right)},$$

whence

$$\|\chi_E\|_{\Phi,\Omega} = \frac{1}{\Phi^{-1}\left(|E|^{-1}\right)},$$

as claimed.

REMARK 2.5. *Suppose $\|f\|_{\Phi,\Omega}$ is finite and positive. Then, there exists a strictly decreasing sequence $\{\lambda_n\}$ in $\mathbb{R}_+$ such that $\lambda_n \downarrow \|f\|_{\Phi,\Omega}$. This means that $\frac{|f(x)|}{\lambda_n} \uparrow \frac{|f(x)|}{\|f\|_{\Phi,\Omega}}$, so, by the Monotone Convergence Theorem,*

$$\int_\Omega \Phi\left(\frac{|f(x)|}{\|f\|_{\Phi,\Omega}}\right) dx = \lim_{n\to\infty} \int_\Omega \Phi\left(\frac{|f(x)|}{\lambda_n}\right) dx \leq 1.$$

Moreover, if $\int_\Omega \Phi\left(\frac{|f(x)|}{\lambda}\right) dx < \infty$ for $\lambda \in \mathbb{R}_+$, as is the case for simple functions f, then, by the same theorem,

$$\int_\Omega \Phi\left(\frac{|f(x)|}{\|f\|_{\Phi,\Omega}}\right) dx = 1.$$

Finally, $\|f\|_{\Phi,\Omega} = 0$ implies $\int_\Omega \Phi\left(\frac{|f(x)|}{\lambda}\right) dx \leq 1$ for all $\lambda > 0$, which implies $f = 0$ a. e.; otherwise there would exist $k > 0$ and $E \subset \Omega$, $|E| > 0$, with $|f(x)| \geq k$ on E or

$$\Phi\left(\frac{k}{\lambda}\right)|E| = \int_\Omega \Phi\left(\frac{k\chi_E}{\lambda}\right) dx \leq \int_\Omega \Phi\left(\frac{|f(x)|}{\lambda}\right) dx \leq 1,$$

that is,

$$|E| \leq \frac{1}{\Phi\left(\frac{k}{\lambda}\right)},\ \lambda \in \mathbb{R}_+,$$

or $|E| = 0$, a contradiction.

DEFINITION 2.6. Let $\Omega \subset \mathbb{R}^n$ be measurable and suppose Φ is a Young function. The **Orlicz class** $L_\Phi(\Omega)$ is then defined as the set

$$\left\{f : \Omega \to \mathbb{R} : \|f\|_{\Phi,\Omega} < \infty\right\}.$$

PROPOSITION 2.7. *Let $\Omega \subset \mathbb{R}^n$ be measurable and suppose Φ is a Young function. Then,*

(1) $f \in L_\Phi(\Omega)$, $c \in \mathbb{R}$ *implies* $cf \in L_\Phi(\Omega)$, *with*

$$\|cf\|_{\Phi,\Omega} = |c|\,\|f\|_{\Phi,\Omega}\,;$$

(2) $f, g \in L_\Phi(\Omega)$ *implies* $f+g \in L_\Phi(\Omega)$, *with*

$$\|f+g\|_{\Phi,\Omega} \le \|f\|_{\Phi,\Omega} + \|f\|_{\Phi,\Omega}\,.$$

PROOF. Only the second item needs proving. One may assume $\|f\|_{\Phi,\Omega} > 0$, $\|g\|_{\Phi,\Omega} > 0$. Set $\gamma := \|f\|_{\Phi,\Omega} + \|g\|_{\Phi,\Omega}$, $\alpha := \frac{\|f\|_{\Phi,\Omega}}{\gamma}$ and $\beta := \frac{\|g\|_{\Phi,\Omega}}{\gamma}$, so that $\alpha, \beta > 0$, $\alpha + \beta = 1$. One has

$$\begin{aligned}
\int_\Omega \Phi\left(\frac{|f+g|}{\gamma}\right) dx &\le \int_\Omega \Phi\left(\frac{|f|+|g|}{\gamma}\right) dx \\
&= \int_\Omega \Phi\left(\frac{\alpha|f|}{\|f\|_{\Phi,\Omega}} + \frac{\beta|g|}{\|g\|_{\Phi,\Omega}}\right) dx \\
&\le \int_\Omega \left[\alpha\Phi\left(\frac{|f|}{\|f\|_{\Phi,\Omega}}\right) + \beta\Phi\left(\frac{|g|}{\|g\|_{\Phi,\Omega}}\right)\right] dx \\
&= \alpha\int_\Omega \Phi\left(\frac{|f|}{\|f\|_{\Phi,\Omega}}\right) dx + \beta\int_\Omega \Phi\left(\frac{|g|}{\|g\|_{\Phi,\Omega}}\right) dx \\
&\le \alpha+\beta \\
&= 1.
\end{aligned}$$

We conclude that

$$\|f+g\|_{\Phi,\Omega} \le \gamma = \|f\|_{\Phi,\Omega} + \|g\|_{\Phi,\Omega}\,.$$

□

Proposition 2.7 shows that $L_\Phi(\Omega)$ is a normed linear space under the functional $f \to \|f\|_{\Phi,\Omega}$. We next establish further poperties of this functional that will enable us to prove, in the next section, that $L_\Phi(\Omega)$ is complete and thus a Banach space.

PROPOSITION 2.8. *Let* $\Omega \subset \mathbb{R}^n$ *be measurable and suppose* Φ *is a Young function. Then,* $0 \le f_k \uparrow f$ *a. e. on* Ω *implies*

$$\|f_k\|_{\Phi,\Omega} \uparrow \|f\|_{\Phi,\Omega}\,.$$

PROOF. We first show $\|f_k\|_{\Phi,\Omega} \le \|f_{k+1}\|_{\Phi,\Omega}$, $k = 1, 2, \ldots$. Indeed,

$$\int_\Omega \Phi\left(\frac{f_k}{\|f_{k+1}\|_{\Phi,\Omega}}\right) dx \le \int_\Omega \Phi\left(\frac{f_{k+1}}{\|f_{k+1}\|_{\Phi,\Omega}}\right) dx \le 1,$$

so

$$\|f_k\|_{\Phi,\Omega} \le \|f_{k+1}\|_{\Phi,\Omega}$$

Next, let $\alpha_k = \|f_k\|_{\Phi,\Omega}$ and put $\alpha = \sup_k \alpha_k$. Since $f \ge f_k$ we have $\|f\|_{\Phi,\Omega} \ge \|f_k\|_{\Phi,\Omega}$ for each k, whence $\|f\|_{\Phi,\Omega} \ge \alpha$. We need only show equality holds. This is clear if $\alpha = 0$ or $\alpha = \infty$, so we assume $0 < \alpha_k \le \alpha < \infty$. In that case,

$$\int_\Omega \Phi\left(\frac{f_k}{\alpha}\right) dx \le \int_\Omega \Phi\left(\frac{f_k}{\alpha_k}\right) dx \le 1$$

and the Monotone Convergence Theorem shows

$$\int_\Omega \Phi\left(\frac{f}{\alpha}\right) dx \leq 1$$

and hence

$$\|f\|_{\Phi,\Omega} \leq \alpha.$$

□

The property of $\| \quad \|_{\Phi,\Omega}$ embodied in Proposition 2.8 is called the Fatou Property.

PROPOSITION 2.9. *Let $\Omega \subset \mathbb{R}^n$ be measurable and suppose Φ is a Young function. Then, given measurable $E \subset \Omega$, with $0 < |E| < \infty$, there exists a positive constant $K = K_E$, independent of f, such that*

$$\int_E |f| dx \leq K \|f\|_{\Phi,\Omega}.$$

PROOF. It suffices to consider f with $0 < \|f\|_{\Phi,\Omega} < \infty$. Jensen's inequality, which asserts that

$$\Phi\left(\frac{1}{|E|}\int_E |g| dx\right) \leq \frac{1}{|E|}\int_E \Phi(|g|)\, dx,$$

is a simple consequence of the convexity of Φ. In particular, given $k = \|f\|^{-1}_{\Phi,\Omega}$, one has

$$\begin{aligned}\Phi\left(\frac{1}{|E|}\int_E k|f| dx\right) &\leq \frac{1}{|E|}\int_E \Phi(k|f|) dx \\ &= \frac{1}{|E|}\int_E \Phi\left(\frac{|f|}{\|f\|_{\Phi,\Omega}}\right) dx \\ &\leq \frac{1}{|E|},\end{aligned}$$

or

$$\frac{1}{|E|}\int_E k|f| dx \leq \Phi^{-1}\left(\frac{1}{|E|}\right);$$

that is,

$$\int_E |f| dx \leq K \|f\|_{\Phi,\Omega},$$

where

$$K = |E|\Phi^{-1}\left(\frac{1}{|E|}\right).$$

□

3. The completeness of $L_\Phi(\Omega)$

We have so far shown that an Orlicz class $L_\Phi(\Omega)$ is a normed linear space. It is the purpose of this section to prove

THEOREM 3.1. *Let Ω be a measurable subset of $\mathbb{R}^n$ and suppose Φ is a Young function. Then, $L_\Phi(\Omega)$ is a Banach space in the sense that every Cauchy sequence in it is convergent.*

PROOF. We first show that if $g_n \in L_\Phi(\Omega)$, $n = 1, 2, \dots$ and $\sum_{n=1}^\infty \|g_n\|_{\Phi,\Omega} < \infty$, then the series $\sum_{n=1}^\infty g_n$ converges to a function $g \in L_\Phi(\Omega)$ and $\|g_n\|_{\Phi,\Omega} \leq \sum_{n=1}^\infty \|g_n\|_{\Phi,\Omega}$.

To this end, let

$$t := \sum_{n=1}^\infty |g_n| \quad \textit{and} \quad t_N := \sum_{n=1}^N |g_n|, \quad N = 1, 2, \dots$$

so

$$0 \leq t_N \uparrow t.$$

Since

$$\|t_N\|_{\Phi,\Omega} \leq \sum_{n=1}^N \|g_n\|_{\Phi,\Omega} \leq \sum_{n=1}^\infty \|g_n\|_{\Phi,\Omega},$$

it follows from the Fatou Property that

$$\|t\|_{\Phi,\Omega} \leq \sum_{n=1}^\infty \|g_n\|_{\Phi,\Omega} < \infty,$$

that is, $t \in L_\Phi(\Omega)$. In particular, $t = \sum_{n=1}^\infty |g_n| < \infty$ *a.e.* by the local imbedding of $L_\Phi(\Omega)$ in $L_1(\Omega)$, whence $\sum_{n=1}^\infty g_n$ converges *a.e.* Thus, if

$$g := \sum_{n=1}^\infty g_n \quad \textit{and} \quad S_N := \sum_{n=1}^N g_n, \quad N = 1, 2, \dots$$

one has $S_N \longrightarrow g$ *a.e.* Hence, for any $M \in \mathbb{Z}_+$,

$$S_N - S_M \longrightarrow g - S_M \quad \textit{a.e.}, \text{ as } N \longrightarrow \infty.$$

Again,

$$\inf_{n \geq N} |S_n - S_M| \leq |S_N - S_M| \leq \sum_{N=M}^\infty |g_N|$$

and

$$\inf_{n \geq N} |S_n - S_M| \uparrow |g - S_M| \quad \textit{a.e.}, \text{ as } N \longrightarrow \infty,$$

so,

$$\begin{aligned} \|g - S_M\|_{\Phi,\Omega} &= \lim_{N \longrightarrow \infty} \| \inf_{n \leq N} |S_n - S_M| \|_{\Phi,\Omega}, \quad \textit{by the Fatou Property,} \\ &\leq \| \sum_{N=M}^\infty |g_N| \|_{\Phi,\Omega}, \quad \textit{by the monotonicity of } \|\|_{\Phi,\Omega}, \\ &\leq \sum_{N=M}^\infty \|g_N\|_{\Phi,\Omega} < \infty, \quad \textit{by the triangle inequality.} \end{aligned}$$

Thus, $g \in L_\Phi(\Omega)$ and $S_M \longrightarrow g$ in $L_\Phi(\Omega)$. It is now a simple exercise to show

$$\|g_n\|_{\Phi,\Omega} \leq \sum_{n=1}^{\infty} \|g_n\|_{\Phi,\Omega}.$$

Suppose, next, $\{f_n\}$ is a Cauchy sequence in $L_\Phi(\Omega$, namely, given $\varepsilon > 0$ there is an $N \in \mathbb{Z}_+$ such that

$$\|f_m - f_n\|_{\Phi,\Omega} < \varepsilon,$$

whenever $m, n \geq N$. To show there exists an $f \in L_\Phi(\Omega)$ with $f_n \longrightarrow f$ with respect to the norm of $L_\Phi(\Omega)$, it suffices to do so for a subsequence, $\{f_{n_k}\}$, of $\{f_n\}$. But, $\{f_n\}$ a Cauchy sequence means there exists a subsequence $\{f_{n_k}\}$ of $\{f_n\}$ for which

$$\|f_{n_{k+1}} - f_{n_k}\|_{\Phi,\Omega} < 2^{-k}, \quad k = 1, 2, \ldots$$

Set

$$g_1 = f_{n_1}, \quad \textit{and} \quad g_k = f_{n_k} - f_{n_{k-1}}, \quad k = 2, \ldots$$

Then,

$$\begin{aligned}\sum_{k=1}^{\infty} \|g_k\|_{\Phi,\Omega} &= \|g_1\|_{\Phi,\Omega} + \sum_{k=2}^{\infty} \|g_k\|_{\Phi,\Omega} \\ &= \|f_{n_1}\|_{\Phi,\Omega} + \sum_{k=2}^{\infty} \|f_{n_k} - f_{n_{k-1}}\|_{\Phi,\Omega} \\ &\leq \|f_{n_1}\|_{\Phi,\Omega} + \sum_{k=2}^{\infty} 2^{-k+1} < \infty.\end{aligned}$$

Hence, $S_K := \sum_{k=1}^{K} g_k$ converges to a function f in $L_\Phi(\Omega)$. Since $f_{N_K} = S_K$ this says $f_{N_K} \longrightarrow f$ in $L_\Phi(\Omega)$, as required. □

4. Duality

Orlicz spaces are examples of Banach function spaces. A general theory of the latter was developed by Luxemburg in [L]; see also [BS]. The principal result of the theory is a duality theorem involving the so-called Köthe dual, $\| \quad \|'$, of a Banach function norm $\| \quad \|$. In the typical case of a gauge norm, $\| \quad \|_{\Phi,\Omega}$, this dual is defined by

$$\|g\|'_{\Phi,\Omega} := \sup\{\int_\Omega |fg| : \|f\|_{\Phi,\Omega} \leq 1\};$$

here, of course, f and g are measurable functions on $\Omega \subset \mathbb{R}^n$.

Our aim in this section is to show $\| \quad \|'_{\Phi,\Omega}$ is equivalent to a gauge norm, $\| \quad \|_{\Psi,\Omega}$, whose Young function Ψ is intimately connected to Φ.

Definition 4.1. Let $\Phi(t) = \int_0^t \varphi(s)ds$ be a Young function. If φ is strictly increasing and continuous, set $\psi(y) := \varphi^{-1}(y)$, $y \in \mathbb{R}_+$. Otherwise, take $\psi(s) := \inf\{y : \varphi(y) \geq s\}$, the so-called left continuous inverse of φ.

One readily shows ψ is increasing and left continuous, with $\psi(0+) = 0$ and $\lim_{s\longrightarrow\infty} \psi(s) = \infty$, so that $\Psi(t) := \int_0^t \psi(s)ds$ is a Young function, referred to as the Young function complementary to Φ. It is easily seen, that, in turn, Φ is the

Young function complementary to Ψ. In view of all this, we speak of Φ and Ψ as complementary Young functions.

EXAMPLE 4.2.

(1) The gauge norm of L_p has

$$\begin{aligned}\Phi(t) &= \frac{t^p}{p}, \quad 1 < p < \infty\\ \phi(s) &= s^{p-1}\\ \psi(s) &= s^{q-1}, \quad q = \frac{p}{p-1},\\ \Psi(t) &= \frac{t^q}{q}.\end{aligned}$$

(2) For the gauge norm of $L \log L$

$$\begin{aligned}\Phi(t) &= (t+1)\log(t+1) - t\\ \phi(s) &= \log(s+1)\\ \psi(s) &= e^s - 1\\ \Psi(t) &= e^t - t - 1.\end{aligned}$$

The following inequality of W. H. Young generalizes the classical AM-GM inequality.

THEOREM 4.3. *Let Φ and Ψ be complementary Young functions. Then,*

$$st \le \Phi(s) + \Psi(t), \quad s, t \in \mathbb{R}_+,$$

with equality if and only if $t = \varphi(s)$ or $s = \psi(t)$.

PROOF. One can do no better than reproduce the diagram on p.5 of [KR],with its notation altered to agree with ours, see figure 1.

□

EXAMPLE 4.4. *In case $\Phi(t) = \dfrac{t^p}{p}$ and hence $\Psi(t) = \dfrac{t^q}{q}$, $q = \dfrac{p}{p-1}$, Young's inequality reads*

$$st \le \frac{t^p}{p} + \frac{t^q}{q},$$

with equality if and only if $t = s^{p-1}$.

On the basis of Young's inequality one can get a Young' function, χ, say, such that $\| \quad \|'_{\Phi,\Omega} = \| \quad \|_{\chi,\Omega}$, namely,

$$\chi(t) := \sup_{s \in \mathbb{R}_+} [ts - \Phi(s)].$$

However, the function χ is somehow less appealing than ψ, so we continue with the latter. A substitute for Hölder's inequality involving Ψ is given in

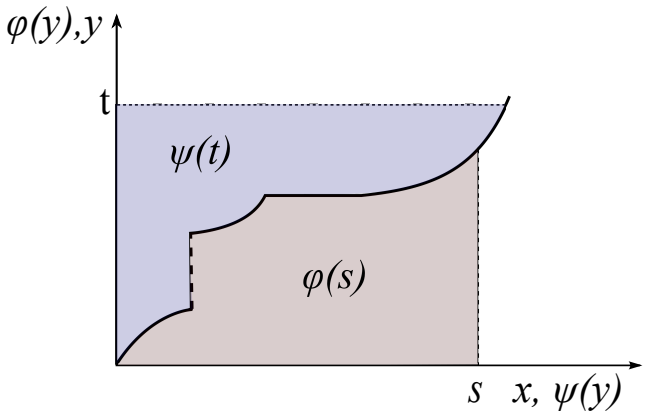


FIGURE 1. Diagram from [KR] with altered notation.

THEOREM 4.5. *Let Φ and Ψ be complementary Young functions and suppose $f \in L_\Phi(\Omega)$, $g \in L_\Psi(\Omega)$ for some domain $\Omega \subset \mathbb{R}^n$. Then, $fg \in L_1(\Omega)$ and*

$$\int_\Omega |fg| \leq 2\|f\|_{\Phi,\Omega}\|g\|_{\Psi,\Omega},$$

from which we conclude that

$$\|g\|'_{\Phi,\Omega} \leq 2\|g\|_{\Psi,\Omega}.$$

PROOF. Let $\lambda, \mu \in \mathbb{R}_+$ be such that

$$\int_\Omega \Phi\left(\frac{|f|}{\lambda}\right), \quad \int_\Omega \Psi\left(\frac{|g|}{\mu}\right) \leq 1.$$

From Theorem 4.3 we get

$$\frac{|f|}{\lambda}\frac{|g|}{\mu} \leq \Phi\left(\frac{|f|}{\lambda}\right) + \Psi\left(\frac{|g|}{\mu}\right),$$

so

$$\int_\Omega \frac{|fg|}{\lambda\mu} \leq \int_\Omega \Phi\left(\frac{|f|}{\lambda}\right) + \int_\Omega \Psi\left(\frac{|g|}{\mu}\right) \leq 2,$$

that is,

$$\int_\Omega |fg| \leq 2\lambda\mu.$$

Taking the minimum over λ and then over μ yields

$$\int_\Omega |fg| \leq 2\|f\|_{\Phi,\Omega}\|g\|_{\Psi,\Omega},$$

as asserted. □

THEOREM 4.6. *Suppose $g : \Omega \longrightarrow \mathbb{R}$, $\Omega \subset \mathbb{R}^n$, is such that $\|g\|'_{\Phi,\Omega} < \infty$. Then, $g \in L_\Psi(\Omega)$ and*

$$\|g\|_{\Psi,\Omega} \leq \|g\|'_{\Phi,\Omega}.$$

PROOF. It suffices to show

$$\int_\Omega \Psi\left(\frac{g}{\|g\|'_{\Phi,\Omega}}\right) \leq 1.$$

We may suppose that g is a nonnegative function and, in view of Fatou's lemma, a simple function. The case $\|g\|'_{\Phi,\Omega} = 0$ being trivial, we also assume $\|g\|'_{\Psi,\Omega} > 0$. Now, for $f := \psi\left(\frac{g}{\|g\|'_{\Phi,\Omega}}\right)$, one has equality in Young's inequality, namely,

$$\Phi(f) + \Psi\left(\frac{g}{\|g\|'_{\Phi,\Omega}}\right) = \frac{fg}{\|g\|'_{\Phi,\Omega}}.$$

Moreover, f, like g, is a nonnegative function, so

$$\int_\Omega \frac{fg}{\|g\|'_{\Phi,\Omega}} \leq \int_\Omega \Phi(f) + 1.$$

Indeed, f is a simple nonnegative function with $\|f\|_{\Phi,\Omega} > 0$, which means

$$\int_\Omega \frac{f}{\|f\|_{\Phi,\Omega}} g \leq \|g\|'_{\Phi,\Omega}$$

or

$$\int_\Omega \frac{fg}{\|g\|'_{\Phi,\Omega}} \leq \|f\|_{\Phi,\Omega}.$$

But,

$$\|f\|_{\Phi,\Omega} \leq \max\left[1, \int_\Omega \Phi(f)\right],$$

since, if $\|f\|_{\Phi,\Omega} > 1$, Remark 2.5 guarantees,

$$1 \leq \int_\Omega \Phi\left(\frac{f}{\|f\|_{\Phi,\Omega}}\right) \leq \frac{1}{\|f\|_{\Phi,\Omega}} \int_\Omega \Phi(f),$$

or

$$\|f\|_{\Phi,\Omega} \leq \int_\Omega \Phi(f).$$

Altogether, then,

$$\int_\Omega \Phi(f) + \int_\Omega \Psi\left(\frac{g}{\|g\|'_{\Phi,\Omega}}\right) \leq \int_\Omega \Phi(f) + 1$$

and, hence, canceling the finite positive number $\int_\Omega \Phi(f)$, we get

$$\int_\Omega \Psi\left(\frac{g}{\|g\|'_{\Phi,\Omega}}\right) \leq 1.$$

□

REMARK 4.7. *Theorems 4.5 and 4.6, combined, yield the equivalence*

$$\|g\|_{\Psi,\Omega} \leq \|g\|'_{\Phi,\Omega} \leq 2\|g\|_{\Psi,\Omega}.$$

5. The rearrangement invariance of $L_\Phi(\Omega)$

The gauge norm of a function depends only on the distribution of its values and not where they are taken on. To quantify this fact we associate to a given measurable function f on $\Omega \subset \mathbb{R}^n$ its distribution function

$$\mu_f(\lambda) := |\{x \in \Omega : |f(x)| > \lambda\}|, \quad \lambda \in \mathbb{R}_+.$$

Our claim is that if $g : \Omega \longrightarrow \mathbb{R}$ is measurable and $\mu_g = \mu_f$, then, for any Young function Φ,

$$\|g\|_{\Phi,\Omega} = \|f\|_{\Phi,\Omega}.$$

To establish it we work with the generalized right continuous inverse, f^*, of μ_f, namely

$$f^*(t) := \inf\{t : \mu_f(\lambda) \leq t\}, \quad t \in \mathbb{R}_+.$$

The function f^* thus defined on $I_\Omega := (0, |\Omega|)$ is called the decreasing rearrangement of f. We observe that, when μ_f is strictly decreasing and continuous, $f^* = \mu_f^{-1}$. More important, for our purposes, $\mu_f = \mu_g$ if and only if $f^* = g^*$. The above claim follows from

THEOREM 5.1. *Suppose $f : \Omega \longrightarrow \mathbb{R}$, $\Omega \subset \mathbb{R}^n$ is measurable. Then,*

$$\|f\|_{\Phi,\Omega} = \|f^*\|_{\Phi,I_\Omega}.$$

PROOF. The result is clear when f is a simple function, in which case f^* is a step function on I_Ω. Again, to every nonnegative measurable function f there corresponds a sequence, $\{S_k\}$, of nonnegative simple functions such that $S_k \uparrow f$. The proof can then be completed by invoking the Fatou property of $\| \quad \|_{\Phi,\Omega}$ and $\| \quad \|_{\Phi,I_\Omega}$. □

In the next section we will indicate the role the decreasing rearrangement plays in the theory of Sobolev imbeddings and through them in PDE. For the present, we indicate its use in computing gauge norms.

To begin, we describe how to approximate the rearrangement of a function defined on a polyhedral domain, P, in $\mathbb{R}^n$. For simplicity we restrict attention to the case $n = 2$.

One decomposes P into a finite number of small triangles, $T_i, i \in I$, so that any two of the triangles intersect at most along a common side or at a common vertex. On a given triangle, T, one interpolates f by a linear function ℓ_T, equal to f at the

vertices of T. The continuous linear spline, s, equal to ℓ_{T_i} on each $T_i, i \in I$, is then rearranged to give s^* as an approximation to f^*. One has

$$\mu_s(\lambda) = \sum_{i \in I} \mu_{\ell_{T_i}}(x),$$

where

$$\mu_{\ell_{T_i}}(\lambda) = \begin{cases} \left(1 - \dfrac{(\lambda - z_1)^2}{(z_2 - z_1)(z_3 - z_1)}\right)|T_i|, & z_1 < \lambda < z_2, \\ \dfrac{(z_3 - \lambda)^2}{(z_3 - z_1)(z_3 - z_2)}|T_i|, & z_2 < \lambda < z_3, \\ 0, & \textit{otherwise}; \end{cases}$$

here, z_1, z_2 and z_3 are the values of f, in increasing order, at the vertices of T_i. See [FKPS].

Suppose, next, f is a function on a polygonal domain, P, such that $0 < \int_P \Phi\left(\frac{|f|}{\lambda}\right) < \infty$ for all $\lambda \in \mathbb{R}_+$, in which case $\|f\|_{\Phi,P}$ will be the unique λ satisfying the equation

$$\int_P \Phi\left(\frac{|f|}{\lambda}\right) = 1.$$

Indeed,

$$\int_P \Phi\left(\frac{|f|}{\lambda}\right) = \int_0^{|P|} \Phi\left(\frac{f^*}{\lambda}\right), \quad \lambda \in \mathbb{R}_+,$$

so, $\|f\|_{\Phi,P}$ will be the unique λ for which

$$\int_0^{|P|} \Phi\left(\frac{f^*}{\lambda}\right) = 1.$$

Consider, then, the equation

$$0 = F(\lambda) := \int_0^{|P|} \Phi\left(\frac{f^*}{\lambda}\right) - 1. \tag{5.1}$$

One has F decreasing convexly, since

$$F'(\lambda) = -\int_0^{|P|} \varphi\left(\frac{f^*}{\lambda}\right)\lambda^{-2}$$

increases from $-\infty$ at 0. Thus, Newton's method gives rise to an iteration sequence converging to the root of (5.1), if we start with a λ_0 such that $F(\lambda_0) > 0$. Such a λ_0 is the largest number of the form 2^{-k}, $k \in \mathbb{Z}_+$, with $F(2^{-k}) > 0$. It will be convenient to use, all along, the fact that

$$\int_0^{|K|} \Phi\left(\frac{f^*}{\lambda}\right) = \lambda^{-1} \int_0^{M_f} \varphi\left(\frac{s}{\lambda}\right)\mu_f(s)ds$$

and

$$\int_0^{|K|} \varphi\left(\frac{f^*}{\lambda}\right) = \lambda^{-1} \int_0^{M_f} \varphi'\left(\frac{s}{\lambda}\right)\mu_f(s)ds,$$

so the Newton iteration equation becomes

$$\lambda_{n+1} = \lambda_n + \frac{\lambda_n^{-1}\int_0^{M_f}\varphi\left(s/\lambda_n\right)\mu_f(s)ds - 1}{\lambda_n^{-3}\int_0^{M_f}\varphi'\left(s/\lambda_n\right)\mu_f(s)ds}, \quad M_f := \|f\|_{\infty,P}. \tag{5.2}$$

It is important to compute the ratio as it appears here rather than to attempt to simplify it.

EXAMPLE 5.2. *For simplicity, we consider a function whose distribution function can be calculated directly. Thus, let* $g(t) := e^{-(1-t)^{-1}}\chi_{(0,1)}(t)$ *and at* $X = (x, y)$ *in* $Q := [0, 1] \times [0, 1]$ *take*

$$f(X) := 9g(6|X - X_1|) + \frac{64}{3}g(8|X - X_2|) + \frac{125}{3}g(10|X - X_3|),$$

$$X_1 = \left(\frac{1}{3}, \frac{1}{3}\right), \quad X_2 = \left(\frac{2}{3}, \frac{1}{3}\right), \quad \textit{and} \quad X_3 = \left(\frac{1}{3}, \frac{2}{3}\right).$$

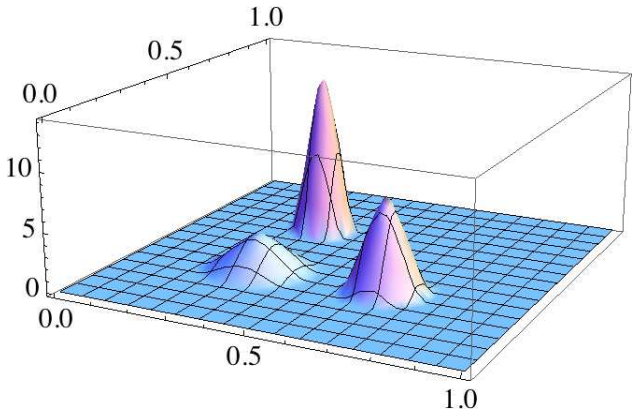


FIGURE 2. Graph of $f(X)$.

Then, with $h(t) := \pi\left(1 - \frac{1}{\log(\frac{1}{t})}\right)^2$, $0 < t < 1$, *one has*

$$\mu_f(\lambda) = \begin{cases} \frac{1}{36}h\left(\frac{\lambda}{9}\right) + \frac{1}{64}h\left(\frac{3\lambda}{64}\right) + \frac{1}{100}h\left(\frac{3\lambda}{125}\right), & 0 < \lambda \le 9e^{-1}, \\ \frac{1}{64}h\left(\frac{3\lambda}{64}\right) + \frac{1}{100}h\left(\frac{3\lambda}{125}\right), & 9e^{-1} < \lambda \le \frac{64}{3}e^{-1} \\ \frac{1}{100}h\left(\frac{3\lambda}{125}\right), & \frac{64}{3}e^{-1} < \lambda \le \frac{125}{3}e^{-1}, \\ 0, & \lambda > \frac{125}{3}e^{-1}. \end{cases}$$

Using (5.2) we obtain

$$\|f\|_{L\log L(Q)} = .7606 \quad \textit{and} \quad \|f\|_{L\exp(Q)} = 2.3498$$

where, respectively, their Young functions

$$\Phi_1(t) = \frac{(t+1)\log(t+1) - t}{\log 4 - 1}$$

and

$$\Phi_2(t) = \frac{e^t - t - 1}{e - 2}$$

are so chosen that

$$\|\chi_Q\|_{L\log L(Q)} = \|\chi_Q\|_{L\exp(Q)} = 1.$$

We observe that

$$\int_Q f = \int_Q e^{-(1-|x|)^{-1}} dx = \int_0^1 e^{-(1-r)^{-1}} r dr$$
$$= \int_0^\infty e^{-y}[y^{-2} - y^{-3}]dy = \frac{1}{2}\int_1^\infty e^{-y}\frac{dy}{y} \approx 0.1097.$$

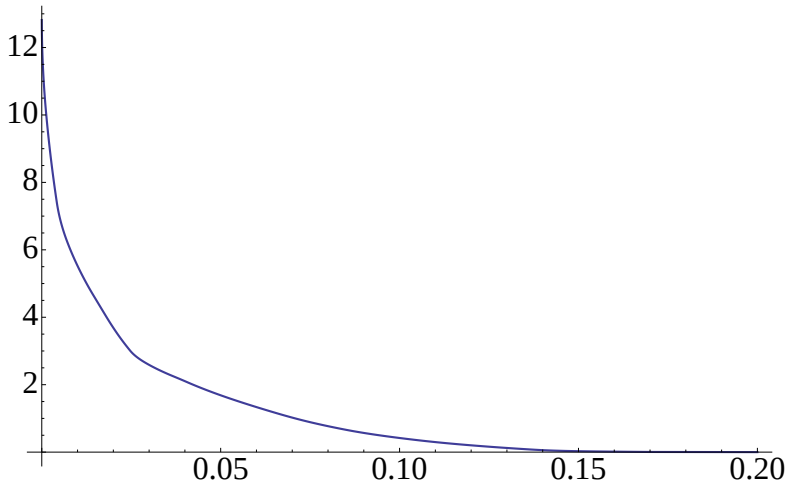


FIGURE 3. Graph of $f^*(t)$.

6. The role of Orlicz spaces in the theory of elliptic PDE

This final section looks at the role of gauge norms in the theory of elliptic PDE. As mentioned in Section 1 they first occurred in endpoint Sobolev imbedding inequalities, specifically in the inequality

$$\|u\|_{\Phi_1,\Omega} \le C\left[\int_\Omega |(\nabla u)(x)|^{\frac{1}{n}} dx\right]^{\frac{1}{n}}, \quad u \in C_0^\infty(\Omega), \tag{6.1}$$

with

$$\Phi_1(t) := \int_0^t (\exp(s^{\frac{n}{n-1}} - 1))ds, \quad t \in \mathbb{R}_+.$$

One can improve on (6.1) in two ways. Thus, in [BW], it is shown $\| \quad \|_{\Phi_1,\Omega}$ can be replaced by the larger (Lorentz) norm

$$\left[\int_0^{|\Omega|} u^*(t)^n \log^{-n}\left(\frac{|\Omega|}{t}\right)\frac{dt}{t}\right]^{\frac{1}{n}}.$$

Recently, an improvement on the Lebesgue norm in (6.1) was obtained, in [KP2], namely, the (smaller) functional – equivalent to a norm –

$$\rho(f) := \left\| t^{-\frac{1}{n}} \int_t^{|\Omega|} f^*(s) s^{-\frac{1}{n}} ds \right\|_{\Phi_2, I_\Omega},$$

defined in terms of the Young function

$$\Phi_2(t) := \int_0^t s^{n-1} \log^{-n}(s+10n) ds \approx t^n \log^{-n}(t+10n), \quad t \in \mathbb{R}_+.$$

Finally, we focus on an elliptic nonlinear boundary value problem considered in detail by J. P. Gossez in [G]. Its very statement involves a Young function. For concreteness, we state it in the context of $\mathbb{R}^2$.

Let $\Phi(t) = \int_o^t \varphi(s) ds, \quad t \in \mathbb{R}_+$, be a Young function and suppose $f : \Omega \longrightarrow \mathbb{R}$, where Ω is a bounded domain in $\mathbb{R}^2$. Find a function $u = u(x, y)$ on Ω satisfying

$$\begin{gathered} -\frac{\partial}{\partial x}\left(\varphi\left(\frac{\partial u}{\partial x}\right)\right) - \frac{\partial}{\partial y}\left(\varphi\left(\frac{\partial u}{\partial y}\right)\right) = f \quad in \quad \Omega \\ u = 0 \quad on \quad \partial\Omega; \end{gathered} \tag{6.2}$$

here, $\varphi(-t) = -\varphi(t)$, $t \in \mathbb{R}_+$.

To further simplify things we assume that Φ satisfies the $\triangle_2$ condition, namely $\Phi(2t) \le c\Phi(t)$ for all t greater than some $t_0 \in \mathbb{R}_+$.

Gossez sought a solution of (6.2) in the Orlicz–Sobolev space $W_0^1 L_\Phi(\Omega)$, which we now define. To do this we require the norm of another Orlicz–Sobolev space, that of

$$W^1 L_\Phi(\Omega) := \left\{ u \in L_\Phi(\Omega) : \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y} \quad belong\ to \quad L_\Phi(\Omega) \right\},$$

whose norm is given through the equation

$$\|u\|_{1,\Phi,\Omega}^2 := \|u\|_{\Phi,\Omega}^2 + \left\|\frac{\partial u}{\partial x}\right\|_{\Phi,\Omega}^2 + \left\|\frac{\partial u}{\partial y}\right\|_{\Phi,\Omega}^2; \tag{6.3}$$

the partial derivatives appearing in (6.3) are the usual weak derivatives. The space $W_0^1 L_\Phi(\Omega)$ is now defined to be the closure of C_0^∞ in $W^1 L_\Phi(\Omega)$.

To motivate the condition to be put on the function f in (6.2) we first consider the so-called Nemickii operator

$$T : u(x) \longrightarrow \varphi(u(x)).$$

Specifically, we study

$$D(T) := \{u \in L_\Phi(\Omega) : \varphi(u(x)) \in L_\Psi(\Omega)\},$$

in which Ψ is the Young function complementary to Φ. To this end, we define

$$E_\Phi(\Omega) := \{f : \Omega \longrightarrow \mathbb{R} : \int_\Omega \Phi(k|f|) < \infty \quad \textit{for all} \quad k > 0\}$$

and

$$\mathcal{L}_\Phi(\Omega) := \{f : \Omega \longrightarrow \mathbb{R} : \int_\Omega \Phi(|f|) < \infty\}.$$

One has

LEMMA 6.1.

$$E_\Phi \subset D(T) \subset \mathcal{L}_\Phi$$

In particular, T is a bounded operator from $L_\Phi(\Omega)$ into $L_\Psi(\Omega)$ if and only if Φ satisfies the $\triangle_2$ condition.

PROOF. If $u \in D(T)$, then $|u| \in L_\Phi(\Omega)$ and $\varphi(|u|) \in L_\Psi(\Omega)$, so $\infty > \int_\Omega |u|\varphi(|u|) \geq \int_\Omega \Phi(|u|)$; that is, $u \in \mathcal{L}_\Phi(\Omega)$.

Again, we claim

$$\Psi(\varphi(t)) \leq \Phi(2t), \quad t \in \mathbb{R}_+, \tag{6.4}$$

assuming which, $u \in E_\Phi(\Omega)$ implies $2u \in \mathcal{L}_\Phi(\Omega)$ and, therefore,

$$\infty > \int_\Omega \Phi(2|u|) \geq \int_\Omega \Psi(\varphi(|u|)),$$

or

$$\varphi(u) = \operatorname{sgn}(u)\varphi(|u|) \in L_\Psi(\Omega). \tag{6.5}$$

To see (6.4), observe that

$$\begin{aligned}\Phi(2t) = \int_0^{2t} \varphi(s)ds &\geq \int_t^{2t} \varphi(s)ds \\ &\geq t\varphi(t) \\ &= \Phi(t) + \Psi(\varphi(t)), \quad \textit{by Theorem 4.3}, \\ &\geqslant \Psi(\varphi(t)).\end{aligned}$$

As for the second assertion, we first show that T bounded implies

$$E_\Phi(\Omega) = L_\Phi(\Omega) = \mathcal{L}_\Phi(\Omega)$$

and hence that Φ satisfies Δ_2 condition. Indeed, suppose $u \in L_\Phi(\Omega)$ and, for $n \in \mathbb{Z}_+$, set

$$u_n(x) = (\min[u_+(x), n] - \min[u_-(x), n])\chi_{(-n,n)}(x).$$

Then, $u_n \in E_\Phi(\Omega) \subset D(T)$ and

$$\sup_n \|u_n\|_{\Phi,\Omega} \leqslant \|u\|_{\Phi,\Omega} < \infty.$$

Since T is bounded

$$\sup_n \|Tu_n\|_{\Psi,\Omega} < \infty.$$

Hölder's inequality then yields

$$\begin{aligned}\int_\Omega \Phi(|u_n|) &\le \int_\Omega |u_n|\varphi(|u_n|)\\ &\le 2\|u_n\|_{\Phi,\Omega}\|\varphi(|u_n|)\|_{\Psi,\Omega}\\ &\le K<\infty, \quad n=1,2,\dots,\end{aligned}$$

whence, by Fatou's lemma,

$$\int_\Omega \Phi(|u|)\le K<\infty;$$

that is, $u\in\mathcal{L}_\Phi(\Omega)$, so (6.5) holds.

Assume, next, Φ satisfies the $\triangle_2$ condition and let u vary in a bounded set in $D(T)$; in particular, there exists $k>1$ with

$$\int_\Omega \Phi\left(\frac{2|u|}{k}\right)\le 1$$

for all u in the set.

Let $t_0>0$ and $C>1$ be such that

$$\Phi(kt)\le C\Phi(t), \quad \text{for } t\ge t_0>0,$$

Then

$$\begin{aligned}\int_\Omega \Phi\left(2|u|\right) &= \int_{\frac{2|u|}{k}\ge t_0}\Phi\left(2|u|\right)+\int_{\frac{2|u|}{k}<t_0}\Phi\left(2|u|\right)\\ &\le C\int_{\frac{2|u|}{k}\ge t_0}\Phi\left(\frac{2|u|}{k}\right)+\Phi(kt_0)|\Omega|\le C+\Phi(kt_0)|\Omega|,\end{aligned}$$

for all u in the set. Further, (6.4) yields

$$\int_\Omega \Psi(\varphi(|u|))\le\int_\Omega\Phi(2|u|)\le C+\Phi(kt_0)|\Omega|$$

where, for all u in the set,

$$\|\varphi(u)\|_{\Psi,\Omega}\le\max\left[1,\int_\Omega\Psi(\varphi(|u|))\right]\le C+\Phi(kt_0)|\Omega|<\infty.$$

□

REMARK 6.2. *Suppose the Young function Φ satisfies the $\triangle_2$ condition and consider $u\in W_0^1L_\Phi(\Omega)$. Such a u being in $L_\Phi(\Omega)$, one has, in view of Lemma 6.1, that $\varphi(u)\in L_\Psi(\Omega)$. But, then $\dfrac{\partial\varphi(u)}{\partial x}$, say, defines a bounded linear functional on $W_0^1L_\Phi(\Omega)$, since, for $v\in W_0^1L_\Phi(\Omega)$,*

$$\begin{aligned}\left|\int_\Omega v\frac{\partial\varphi(u)}{\partial x}\right| &= \left|\int_\Omega\varphi(u)\frac{\partial v}{\partial x}\right|\\ &\leqslant 2\|\varphi(u)\|_{\Psi,\Omega}\left\|\frac{\partial v}{\partial x}\right\|_{\Phi,\Omega}\\ &\leqslant 2\|\varphi(u)\|_{\Psi,\Omega}\|v\|_{W_0^1L_\Phi(\Omega)}.\end{aligned}$$

This tell us it is reasonable to take f in (6.2) to be a bounded linear functional on $W_0^1 L_\Phi(\Omega)$.

It only remains to clarify the sense in which

$$u = 0 \quad on \quad \partial\Omega$$

when $u \in W^1 L_\Phi(\Omega)$.

We assume Φ satisfies the $\triangle_2$ condition and Ω is a Lipschitz domain, so, in particular, it has the segment property, whereby there exists a locally finite open covering, $\{O_i\}$, of $\partial\Omega$ and corresponding vectors, $\{y_i\}$, such that for $x \in \partial\Omega$ and $t \in (0,1)$, one has $x + ty_i \in \Omega$.

As proved in section 2.2 of [G], $C^\infty(\bar{\Omega})$ is dense in $W^1 L_\Phi(\Omega)$. Again, in section 2.3 it is shown that for the restriction mapping

$$\hat{\gamma} : C^\infty(\bar{\Omega}) \longrightarrow C(\partial\Omega)$$

defined by

$$\hat{\gamma}(u) := u|_{\partial\Omega},$$

there holds

$$\|\hat{\gamma}(u)\|_{\Phi,\partial\Omega} \leq K\|u\|_{W^1 L_\Phi(\Omega)}, \tag{6.6}$$

when $u \in C^\infty(\bar{\Omega})$. The above-mentioned density of $C^\infty(\bar{\Omega})$ in $W^1 L_\Phi(\Omega)$ then ensures that $\hat{\gamma}$ can be extended to all of $W^1 L_\Phi(\Omega)$, this extension, denoted by γ, being called the trace operator, with, of course,

$$\|\gamma(u)\|_{\Phi,\partial\Omega} \leq K\|u\|_{W^1 L_\Phi(\Omega)}, \quad u \in W^1 L_\Phi(\Omega).$$

Lastly, the Proposition of section 2.3 tells us that the kernel of γ is in $W_0^1 L_\Phi(\Omega)$, so, if $u \in C^\infty(\bar{\Omega})$,

$$u|_{\partial\Omega} \equiv 0.$$

We are now able to properly state

THEOREM 6.3. *Let Ω be a bounded Lipschitz domain in $\mathbb{R}^2$. Suppose $\varphi : \mathbb{R} \longrightarrow \mathbb{R}$ is continuous and strictly increasing, with $\varphi(0) = 0$ and $\lim_{t\longrightarrow\infty} \varphi(t) = \infty$. Assume,*

$$\Phi(t) := \int_0^t \varphi(s)ds, \quad t \in \mathbb{R}_+$$

satisfies the $\triangle_2$ condition.

Then, to any bounded linear functional, f, on $W^1 L_\Phi(\Omega)$ there corresponds a unique $u \in W_0^1 L_\Phi(\Omega)$ such that $\varphi\left(\frac{\partial u}{\partial x}\right)$ and $\varphi\left(\frac{\partial u}{\partial y}\right)$ belong to the Köthe dual, $L_\Psi(\Omega)$, of $L_\Phi(\Omega)$ and, as distributions,

$$-\frac{\partial}{\partial x}\varphi\left(\frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial y}\varphi\left(\frac{\partial u}{\partial y}\right) = f.$$

EXAMPLE 6.4. *The Young function $\Phi(t) = \int_0^t \log(1+s)ds$ for the Orlicz space $L\log L$ satisfies the $\triangle_2$ condition, so Theorem 6.3 guarantees the unique solution, u, to (6.3) in any Lipschitz domain Ω, whenever T is a bounded linear functional on $W_0^1 L\log L(\Omega)$. It is a conseuence of the imbedding theory in*

[KP1] that, in fact, u belongs to an Orlicz space smaller than $L\log L(\Omega)$*, namely to* $L^{n'}(\log L)^{n'}(\Omega)$,$n' = \dfrac{n}{n-1}$*, defined through the Young function*

$$\Phi(t) = \int_0^t s^{n'-1}\log^{n'}(1+s)ds, \quad t \in \mathbb{R}_+.$$

Bibliography

[BS] C. Bennett and R. Sharpley, *Interpolation of Operators*, Pure and Applied Mathematics Vol. 129, Academic Press, Boston 1988.

[BW] H. Brézis and S. Wainger, *A note on limiting cases of Sobolev embeddings and convolution inequalities*, Comm. Partial Differential Equations **5** (1980), 773–789.

[FKPS] M. Francu, R. Kerman, C. Phipps and A. Sayfy, *Finite element approximations to rearrangements*, Preprint.

[G] J. P. Gossez, *Orlicz–Sobolev spaces and nonlinear elliptic boundary value problems*, NAFSA Spring School, Horni Bradlo, 1978, http://project.dml.cz.

[KP1] R. Kerman and L. Pick, *Optimal Sobolev imbeddings*, Forum Math. **18** (2006), no. 4, 535–570.

[KP2] R. Kerman and L. Pick, *Explicit formulas for optimal rearrangement-invariant norms in Sobolev imbedding inequalities*, Preprint.

[KR] M. A. Krasnoselskii and Ya. B. Rutickii, *Convex functions and Orlicz spaces*, GITTL, Moscow, 1958, english transl./ Noordhoff, Groningen, 1961.

[L] W. A. J. Luxemburg, *Banach function spaces*, Ph.D. Thesis, Delft Institute of technology, Assen (Netherlands), 1955.

Part 6

Topics in the Q-tensor theory of liquid crystals

Arghir Zarnescu

2000 *Mathematics Subject Classification.* Primary 82D30 (35A15 35B25 76A15 76M30)

Key words and phrases. complex non-Newtonian fluids, liquid crystals, defects, biaxial, uniaxial, Q-tensor, Navier–Stokes, harmonic maps

Abstract. These lecture notes are intended to be a brief introduction into the mathematical problems of the Q-tensor theory of liquid crystals. Emphasis is placed on those aspects in which tensorial features are significant and on their physical relevance.

Acknowledgement. The author gratefully acknowledges the hospitality of Nečas Center in September 2010.

Contents

CHAPTER 1

Mathematical modelling

1. Some history and the main physical aspects

The story of liquid crystals began in 1888 in the beautiful city of Prague with a discovery of the Austrian botanical physiologist Friedrich Reinitzer. During his studies at Charles University, Reinitzer noticed that an organic substance related to cholesterol "melted" into a cloudy liquid at approximately 145 °C and became a clear liquid at approximately 178 °C. The cloudy liquid was what is now known as a *cholesteric liquid crystals*.

Reinitzer asked for help from a physicist, Otto Lehmann, who was working at that time in Aachen. They exchanged letters and samples and Lehmann became very interested in the newly discovered material particularly because its specificity that was revealed by a novel (at that time) technique that Lehmann was very interested in, namely polarized microscopy. Under the polarized microscope, the apparently liquid substance revealed optical properties more specific to crystals than to a liquid. It was Lehmann who coined the "liquid crystal" name and who fought a bitter battle for the recognition of the new material. However it took until the early 1920s for liquid crystals to became recognized as a new, overall, phase of matter, with properties between liquids and solids.

Liquid crystals are "crystals that flow". Although it sounds contradictory, the name captures the specificity of liquid crystals as phase of matter that is intermediate between crystalline solids and isotropic fluids, by sharing properties with both phases. The material flows like a nearly incompressible viscous fluid and still retains several features, especially optical, characteristic of crystals. Nowadays the material is mathematically modeled as a complex non-Newtonian fluid.

Liquid crystal materials are called *mesogenic* because of their capacity to generate intermediate phases of matter, that is *mesophases*. There are essentially two ways of obtaining a liquid crystalline phase in a mesogenic substance:

- by changing the temperature, in which case we obtain *thermotropic liquid crystals* or
- by changing the concentration of molecules in a solvent in order to obtain *lyotropic liquid crystals*.

It was G. Friedl who in 1922 in [Fri22] proposed the modern classification of the main types of liquid crystals, separating them into *nematic, cholesteric* and *smectic*. In the rest of these notes we shall be concerned only with the simplest type: the nematics (more precisely thermotropic nematics).

Nematic liquid crystals consist of molecules that have rod-like shape, whose thickness is much smaller than their length. Although the centres of mass of the

molecules are isotropically distributed, the direction of the molecules is, on average over small regions in space, constant! Moreover, the rod-like molecules are locally invariant with respect to reflection in the plane perpendicular to the preferred direction. This is commonly referred to as the 'head-to-tail' symmetry [Gen74].

The name *"nematic"* comes from the Greek $\nu\hat{\eta}\mu\alpha$ which means "thread". We will see in the next subsection a justification for this name.

1.1. Birefringence and "defect" textures. Liquid crystals are an anisotropic, inhomogeneous material that is optically birefringent. That is, it has (locally) two indices of refraction. Light polarized parallel to the director has a different index of refraction (in particular it travels at a different velocity) than light polarized perpendicular to the director. However not every direction is a direction of double refraction. Directions of single refraction are called *"optic axis directions"*. Thus light transmitted along a path parallel to the optic axis has only a single refractive index. Depending on the properties of the liquid crystal material at any given point there can be either one or two directions of single refraction. Points where there is a single optic axis are called *"unixial"* while the other ones are called *"biaxial"*. There can also be (rare) points where the light has the same index of refraction in any direction, and these points are called *"isotropic"*. This optic terminology is relevant to the Q-tensor one in the following section.

The most common way of probing the nature of a liquid crystal material is by sending monochromatic (single wave-length) linearly polarized light through a thin layer of liquid crystal material placed between two crossed polarizers. Because of the birefringent nature of the sample, the incoming linearly polarized light becomes elliptically polarized. When this ray reaches the second polarizer, there is now (generically) a component that can pass through, and the region appears bright through the second polarizer. However this is not always the case. If the transmission axis of the first polarizer is parallel to either of the two special directions (called the ordinary or extraordinary directions), the light is not broken up into components, and no change in the polarization state occurs. In this case, there is no transmitted component and the region appears dark through the second polarizer. This property is responsible for the basic mechanism of a liquid crystal display. In the simplest such display one places a source of light of a certain frequency under a thin layer of liquid crystals . In the absence of an exterior field one would only see a dark state on the other side. However, switching on an electric field the state of the molecules will be distorted so that the light will pass and one will observe a bright state.

Up to this point, we have considered using only monochromatic light (single wave-length). However if the liquid crystal material is illuminated by white light (all light frequencies) the wavelengths will experience some retardation and emerge from the full-wave plate in a variety of polarization states, that appear as various colors.

Among the patterns observed in experiments, there exist at least two types that are of a particular significance. A first type refers to systems of dark, flexible filaments. It is these thread-like filaments that are responsible for the name of "nematics". As De Gennes notes in [Gen74], Ch.4 some appear to be floating

entirely the fluid, while others have both ends attached to the walls, or are entirely attached to the walls. De Gennes called these structures *"black filaments"* or *"structures à fils"*. F.C. Frank in [Fra58] referred to them as *"disclinations"* and they are commonly referred to as *"line defects"*.

On the other hand it is possible to see only systems of singular points connected by black stripes. These points are called *"structures a noyaux"* by De Gennes in [Gen74] and they (together with the patterns around them) are referred to as *"schlieren"* textures in the German literature. They are commonly referred to as *"point defects"*.

One can naturally ask that if one has 1D defects (line defects) as well as 0D defects (point defects) why not 2D defects? In [Gen74], Ch.4 De Gennes provides a crude explanation for the lack of defects, arguing that they are "energetically expensive" hence they cannot appear in energy minimizers. In this explanation De Gennes uses the Oseen–Frank director description of nematics (see the next section). However, when using his own Q-tensor description of the nematics he argues against the existence of wall defects based on what are essentially symmetry considerations [Gen72].

Defect patterns are an essential characteristic of nematic liquid crystals and thus they are a benchmark test for determining the relevance of a nematic liquid crystal model. A good model should be able to describe these patterns and predict their appearance. In Section 5 we will see to what extent the available theories are capable of describing these patterns.

2. The probability distribution function and the Q-tensor

The main physical characteristics of nematics (locally preferred directions and 'head-to-tail' symmetry) can be modelled by assigning to each material point x, in the region Ω occupied by the liquid crystal, a probability measure $\mu(x,\cdot) : \mathcal{L}(\mathbb{S}^2) \to [0,1]$ for describing the orientation of the molecules, where $\mathcal{L}(\mathbb{S}^2)$ denotes the family of Lebesgue measurable sets on the unit sphere. Thus $\mu(x,A)$ gives the probability that the molecules with centre of mass in a very small neighbourhood of the point $x \in \Omega$ are pointing in a direction contained in $A \subset \mathbb{S}^2$.

Nematic liquid crystals are locally invariant with respect to reflection in the plane perpendicular to the preferred direction. This is commonly referred to as the 'head-to-tail' symmetry, see [Gen74]. This means that $\mu(x,A) = \mu(x,-A)$, for all $x \in \Omega$, $A \subset \mathcal{L}(\mathbb{S}^2)$. Note that because of this symmetry the first moment of the probability measure vanishes:

$$\langle p \rangle \stackrel{\text{def}}{=} \int_{\mathbb{S}^2} p \, d\mu(p) = \frac{1}{2}\left[\int_{\mathbb{S}^2} p \, d\mu(p) + \int_{\mathbb{S}^2} -p \, d\mu(-p)\right] = 0.$$

The first nontrivial information on μ comes from the tensor of second moments:

$$M_{ij} \stackrel{\text{def}}{=} \int_{\mathbb{S}^2} p_i p_j \, d\mu(p),\ i,j = 1,2,3$$

We have $M = M^T$ and $\operatorname{tr} M = \int_{\mathbb{S}^2} d\mu(p) = 1$. Let e be a unit vector. Then

$$e \cdot Me = \int_{\mathbb{S}^2} (e \cdot p)^2 \, d\mu(p) = \langle \cos^2(\theta) \rangle$$

where θ is the angle between p and e.

If the orientation of the molecules is equally distributed in all directions we say that the distribution is *isotropic* and then $\mu = \mu_0$ where $d\mu_0(p) = \frac{1}{4\pi} dA$. The corresponding second moment tensor is

$$M_0 \stackrel{\text{def}}{=} \frac{1}{4\pi} \int_{\mathbb{S}^2} p \otimes p \, dA = \frac{1}{3} Id$$

(since $\int_{\mathbb{S}^2} p_1 p_2 \, d\mu(p) = 0$, $\int_{\mathbb{S}^2} p_1^2 \, d\mu(p) = \int_{\mathbb{S}^2} p_2^2 \, d\mu(p) = \int_{\mathbb{S}^2} p_3^2 \, d\mu(p)$ and $\operatorname{tr} M_0 = 1$).

The de Gennes order-parameter tensor Q is then defined as

$$Q \stackrel{\text{def}}{=} M - M_0 = \int_{\mathbb{S}^2} \left(p \otimes p - \frac{1}{3} Id \right) d\mu(p) \tag{2.1}$$

and measures the deviation of the second moment tensor from its isotropic value.

Since Q is symmetric and $\operatorname{tr} Q = 0$, Q has, by the spectral theorem, the representation:

$$Q = \lambda_1 \hat{e}_1 \otimes \hat{e}_1 + \lambda_2 \hat{e}_2 \otimes \hat{e}_2 - (\lambda_1 + \lambda_2) \hat{e}_3 \otimes \hat{e}_3 \tag{2.2}$$

where $\hat{e}_1$, $\hat{e}_2$, $\hat{e}_3$ is an orthonormal basis of eigenvectors of Q with the corresponding eigenvalues λ_1, λ_2 and $\lambda_3 = -(\lambda_1 + \lambda_2)$.

When two of the eigenvalues are equal (and non-zero) the order parameter Q is called *uniaxial*, otherwise being *biaxial*. It is an easy exercise, using the spectral representation and the fact that $Id = \sum_{i=1}^{3} e_i \otimes e_i$ to show that any uniaxial Q-tensor can be represented in the form

$$Q = s \left(n \otimes n - \frac{1}{3} Id \right)$$

with $n \in \mathbb{S}^2$ and for some scalar $s \in \mathbb{R}$.

The $Q = 0$ tensor is called the *isotropic* state. This kind of classification of nematics comes from optics, and we saw in the previous section the optical relevance of the terms "biaxial" and "uniaxial". From now on we will refer to symmetric, traceless, d x d-matrices ($d = 2$ or $d = 3$) as *Q-tensors*. More about the physical interpretation and relevance of the Q-tensors will be mentioned in Section 3.1 (see for further details [MN04], [Maj10]).

Equilibrium configurations of liquid crystals are obtained, for instance, as energy minimizers, subject to suitable boundary conditions. The simplest commonly used energy functional is

$$\mathcal{F}_{LG}[Q] = \int_\Omega \left[\frac{L}{2} \sum_{i,j,k=1}^{3} Q_{ij,k} Q_{ij,k} - \frac{a}{2} \operatorname{tr} Q^2 - \frac{b}{3} \operatorname{tr} Q^3 + \frac{c}{4} \left(\operatorname{tr} Q^2 \right)^2 \right] dx \tag{2.3}$$

where a, b, c are temperature and material dependent constants and $L > 0$ is the elastic constant. In the physically significant limit $L \to 0$ (and for appropriate boundary conditions, see for details Chapter 2) we have that the energy minimizers are suitably approximated by minimizers of the corresponding *'Oseen–Frank energy functional'*

$$\mathcal{F}_{OF}[Q] = \int_\Omega \sum_{i,j,k=1}^{3} Q_{ij,k} Q_{ij,k} \, dx$$

in the restricted class of $Q \in W^{1,2}($ where $W^{1,2}(\Omega; S_0)$ is the Sobolev space of square-integrable Q-tensors with square-integrable first derivatives [Eva98]), with Q uniaxial a.e., so that

$$Q = s\left(n \otimes n - \frac{1}{3}Id\right) \tag{2.4}$$

with $s \in \mathbb{R}$ (an explicit *fixed* constant depending on a, b and c but *independent* of $x \in \Omega$) and $n(x) \in \mathbb{S}^2$ a.e. $x \in \Omega$.

This convergence, as $L \to 0$, was studied initially in [MZ10] and further refined in [NZ]. If we restrict ourselves to working with tensors Q that admit a representation as in (2.4), for a fixed s independent of $x \in \Omega$ we refer to this as *the constrained De Gennes theory*.

The reason for referring to (2) as an "Oseen–Frank functional" has to do with the relations that the functional has with the celebrated *Oseen–Frank theory of nematics*. The theory was proposed in the 1950s, in [Fra58]. The Oseen–Frank theory does not use matrices as in (2.4) but rather unit vectors, hence the material is modeled through a vector-field $n : \Omega \subset \mathbb{R}^d \to \mathbb{S}^{d-1}$, $d = 2, 3$. The theory has the deficit of ignoring the physical head-to-tail symmetry of the material but has the advantage of being simpler as it is more convenient to manipulate vector-valued functions, rather than matrix-valued ones. In particular the theory uses mathematical objects with just two degrees of freedom, namely vectors $n \in \mathbb{S}^2$, unlike the theory of De Gennes that uses objects with five degrees of freedom (Q-tensors).

If one formally replaces the formula (2.4) into (2.3) (with s a fixed constant, independent of $x \in \Omega$) one obtains:

$$\mathcal{F}_{LG}[Q] = \int_\Omega \frac{L}{2} \sum_{i,j,k=1}^{3} Q_{ij,k}Q_{ij,k}\, dx + C = \mathcal{F}_{OF}[n] + C = 2s^2 \int_\Omega \sum_{i,j,k=1}^{3} n_{i,k}n_{i,k}\, dx + C$$

(where C is an explicitly computable constant) so *apparently* nothing is lost, at energy level, when passing from the matrix representation to the vector representation form of the Oseen–Frank theory. However, in order to derive the equality above we implicitly assumed that the vector-valued functions n are as smooth as the matrix-valued functions Q. This might not always be the case and indeed it turns out that there can be significant differences between the classical Oseen–Frank theory and the constrained De Gennes theory. These differences are analyzed in [BZ].

An important deficit of the Oseen–Frank theory has to do with its incapacity of predicting higher dimensional defects, as we will discuss in Section 5. The theory can predict only point-like defects, but not the more complicated line or wall defects.

Returning to the Q-tensors and taking into account how they are obtained, namely the definition (2.1) of the tensor Q and assuming that Q is uniaxial so it has a representation as in (2.4) we have that

$$Qn \cdot n = \frac{2}{3}s = \langle (p \cdot n)^2 - \frac{1}{3}\rangle = \langle \cos^2\theta - \frac{1}{3}\rangle$$

where θ is the angle between p and n. We have $s = \frac{3}{2}\langle \cos^2\theta - \frac{1}{3}\rangle$ and so we must necessarily have $-\frac{1}{2} \leq s \leq 1$ with $s = 1$ when we have perfect ordering parallel to n, $s = -\frac{1}{2}$ when all molecules are perpendicular to n and $s = 0$ iff $Q = 0$ (which

occurs if μ is isotropic). Thus s is a measure of order and is called the *scalar order parameter associated to the tensor* Q. In the physical literature it is often assumed that s is constant almost everywhere. For experimental determinations of s see for instance [CEG$^+$06]. Various experimental methods for the determination of s are also presented in the older paper of A. Rapini, [Rap73], pp. $362 - 364$.

Motivated by this representation, and in an attempt to overcome the limitations of the Oseen–Frank theory, J.L. Ericksen proposed in the 1990s, in [Eri90] a more elaborate theory than that of Oseen–Frank. His theory describes nematics through pairs $(s, n) : \Omega \to [-\frac{1}{2}, 1] \times \mathbb{S}^2$ so that n describes the orientation of the director and s is the degree of order. If $s = 0$ then n needs not be defined. The nematics in Ericksen's theory are thus described as objects with three degrees of freedom, intermediary between the theory of Oseen–Frank that uses only two and that of De Gennes that uses five degrees of freedom. Ericksen theory proposes an alternative definition of defects, and manages to predict the more complicated line or wall defects, but however it does not capture properly the head-to-tail symmetry of the molecules.

We have thus three major theories attempting to describe nematics: Oseen–Frank, Ericksen and Landau-de Gennes. Each theory uses different mathematical objects to describe the same physical reality and, as we will see in Section 5, each proposes different explanations of the defect patterns.

2.1. Q-tensors: beyond liquid crystals. The type of local orientational order that exists in nematics is by no means an isolated phenomena in nature. Liquid crystal ordering is found in various different systems, including polymers, amphiphile solutions, elastomers, carbon nanotubes and biological systems (for example, the liquid crystal states of DNA). Thus the Q-tensor description should be suitable for the treatment of such complex systems. More on the nematic defects as they appear in carbon nanotubes can be found in [ZK08].

3. Some simple properties of Q-tensors

3.1. Physical Q-tensors and eigenvalues constraints. Let us recall, (2.1), that a Q-tensor was defined as a normalized second-order moment of a probability measure μ. One can easily see that there can be different probability measures μ_0 and μ_1 such that the corresponding Q-tensors are the same. For instance we can take μ_0 to be the uniform probability measure as before and $\mu_1 = \frac{1}{6}\sum_{i=}^{3}[\delta_{e_i} + \delta_{-e_i}]$ where δ_{e_i} denotes the Dirac delta measure concentrated at e_i, with e_1, e_2, e_3 an orthonormal basis in $\mathbb{R}^3$. Then the corresponding Q-tensors are the same, namely the isotropic Q-tensors.

On the other hand the representation formula (2.1) imposes some restrictions on the eigenvalues of Q. Recalling the spectral representation of Q,(2.2), and the probabilistic definition of Q, (2.1), we have:

$$\lambda_i = (Q\hat{e}_i) \cdot \hat{e}_i = \int_{\mathbb{S}^2} (p \cdot \hat{e}_i)^2 d\mu(p) - \frac{1}{3}$$

Since μ is a probability measure we have $0 \leq \int_{\mathbb{S}^2} (p \cdot \hat{e}_i)^2 \, d\mu(p) \leq 1$ and thus the last relation implies a constraint on the eigenvalues $\lambda_i, i = 1, 2, 3$ of Q, namely:

$$-\frac{1}{3} \leq \lambda_i \leq \frac{2}{3}, \ i = 1, 2, 3 \tag{3.1}$$

Further discussions on the physical relevance and constraints on the Q-tensors are available in [Maj10]. A way of enforcing, at a variational level, the eigenvalues of Q-tensors to stay within the physical region delimitated in (3.1) is by using a potential that blows-up logarithmically when the eigenvalues approach the limit values [BM10].

Let us denote by $|Q|$ the Frobenius norm of a Q-tensor, namely:

$$|Q| \stackrel{\text{def}}{=} \sqrt{\text{tr}(QQ^t)} = \sqrt{\text{tr}(Q^2)} = \sqrt{\sum_{i,j=1}^{3} Q_{ij} Q_{ij}} = \sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}$$

For the physical Q-tensors, the restrictions (3.1) impose a bound on the size of $|Q|$. On the other hand a bound on the size of $|Q|$ imposes restrictions only on the size of $\lambda_1^2 + \lambda_2^2 + \lambda_3^2$ and thus it might not be enough to obtain the more complicated constraints (3.1).

3.2. Characteristic equation and uniaxial Q-tensors. Using the characteristic equation and the Cayley-Hamilton theorem one easily obtains:

PROPOSITION 3.1. *For any Q-tensor we have:*

(i)

$$Q^3 - \frac{1}{2}|Q|Q - \det Q \cdot Id = 0,$$

(ii)

$$\text{tr}(Q^3) = 3 \det(Q),$$

(iii)

$$\text{tr}(Q^4) = \frac{1}{2}|Q|^4.$$

Let us recall that a Q-tensor is uniaxial if it has two equal eigenvalues. We can easily characterize such Q-tensors:

PROPOSITION 3.2. *A Q-tensor is uniaxial if and only if*

$$|Q|^6 = 54(\det Q)^2 = 6 \, \text{tr}(Q^3) \tag{3.2}$$

PROOF. The characteristic equation of Q is:

$$\det(Q - \lambda Id) = -\lambda^3 + \underbrace{\text{tr}(Q)}_{=0}\lambda^2 - \lambda \, \text{tr}(\text{cof}\, Q) + \det Q.$$

But $2 \, \text{tr}(\text{cof}\, Q) = 2(\lambda_2\lambda_3 + \lambda_3\lambda_1 + \lambda_1\lambda_2) = \underbrace{(\lambda_1 + \lambda_2 + \lambda_3)}_{=\text{tr}(Q)=0}^2 - (\lambda_1^2 + \lambda_2^2 + \lambda_3^2) = -|Q|^2$.

Hence the characteristic equation becomes:

$$\lambda^3 - \frac{1}{2}|Q|^2\lambda - \det Q = 0$$

and the condition that $\lambda^3 - p\lambda + q = 0$ has two equal eigenvalues becomes: $p \geq 0$ and $4p^3 = 27q^2$, which gives the first equality in the conclusion. The second equality is easily obtained by observing that we just need to verify it on diagonal matrices as for any Q there exists a $R \in O(3)$ such that RQR^t is diagonal and replacing Q by RQR^t does not change $\det Q$ nor $\mathrm{tr}(Q^3)$. □

REMARK 3.3. *An alternative proof of the previous proposition, without using facts about third-order equations, is provided in Lemma 0.5*

Let us recall that a uniaxial Q-tensor can be written as $Q = s\left(n \otimes n - \frac{1}{3}Id\right)$, where $s \in \mathbb{R}$ and $n \in \mathbb{S}^2$. If we fix the s to be a given non-zero constant, say s_+, we obtain the class of constrained Q-tensors:

$$\mathcal{S}_* \stackrel{\text{def}}{=} \{s_+ \left(n \otimes n - \frac{1}{3}Id\right), n \in \mathbb{S}^2\} \tag{3.3}$$

One can easily show (see also [NZ]):

LEMMA 3.4. *A Q-tensor belongs to $\mathcal{S}_*$ if and only if one of the following holds: (i) $\mathrm{tr}(Q^2) = \frac{2s_+^2}{3}$ and $\mathrm{tr}(Q^3) = \frac{2s_+^3}{9}$. (ii) The minimal polynomial of Q is $\lambda^2 - \frac{s_+}{3}\lambda = \frac{2}{9}s_+^2$.*

4. A Q-tensor model: the Beris-Edwards system

The most complete descriptions of liquid crystals regard them as a complex non-Newtonian fluid. There exist various specific models that all use the Q-tensor description and a comparative discussion of the main models is available for instance in [SMV04].

In these notes we use a model proposed by Beris and Edwards [BE94], that one can find in the physics literature for instance in [DOY], [TDY03] (however in the last two references the fluid is assumed compressible (for computational purposes only) and with variable density). An important feature of this model is that if one assumes smooth solutions and one formally takes $Q(x) = s_+(n(x) \otimes n(x) - \frac{1}{3}Id)$, with s_+ a constant (depending on the parameters of the system, see for instance [MZ10]) and $n : \mathbb{R}^d \to \mathbb{S}^{d-1}$ smooth, then the equations reduce (see [DOY]) to the generally accepted equations of Ericksen, Leslie and Parodi [Les68]. The system is related structurally to other models of complex fluids coupling a transport equation with a forced Navier-Stokes system [CFTZ07], [CM01], [LM00], [LZZ08],[Lin91], [Mas08], [Sch09]. In our case the Navier-Stokes equations are coupled with a parabolic type system, but we also have two more derivatives (than in the previously mentioned models) in the forcing term of the Navier-Stokes equations. The Ericksen-Leslie-Parodi system describing nematic liquid crystals, whose structure is closer to our system (but that has one less derivative in the forcing term of the Navier-Stokes equations) was studied in [LL95], [LL00], [LLW10].

In the following we use a partial Einstein summation convention, that is we assume summation over repeated indices. We consider the equations as described in [DOY], [TDY03] but assume that the fluid has constant density in time and is incompressible.

We denote

$$S(\nabla u, Q) \stackrel{def}{=} (\xi D+\Omega)(Q+\frac{1}{d}Id)+(Q+\frac{1}{d}Id)(\xi D-\Omega)-2\xi(Q+\frac{1}{d}Id)\,\mathrm{tr}(Q\nabla u) \quad (4.1)$$

where $D \stackrel{def}{=} \frac{1}{2}(\nabla u + (\nabla u)^t)$ and $\Omega \stackrel{def}{=} \frac{1}{2}(\nabla u - (\nabla u)^t)$ are the symmetric part and the antisymmetric part, respectively, of the velocity gradient tensor ∇u. The constant d is the dimension of the space and Q is a function on $\mathbb{R}^d$ that is Q-tensor valued. The term $S(\nabla u, Q)$ appears in the equation of motion of the order-parameter, Q, and describes how the flow gradient rotates and stretches the order-parameter. The constant ξ depends on the molecular details of a given liquid crystal and measures the ratio between the tumbling and the aligning effect that a shear flow would exert over the liquid crystal directors.

We also denote:

$$H \stackrel{def}{=} -aQ + b[Q^2 - \frac{\mathrm{tr}(Q^2)}{d}Id] - cQ\,\mathrm{tr}(Q^2) + L\Delta Q \quad (4.2)$$

where $L > 0$.

With the notations above we have the coupled system:

$$\begin{cases} (\partial_t + u\cdot\nabla)Q - S(\nabla u, Q) = \Gamma H \\ (\partial_t + u_\beta\partial_\beta)u_\alpha = \nu\partial_\beta^2 u_\alpha + \partial_\alpha p + \partial_\beta\tau_{\alpha\beta} + \partial_\beta\sigma_{\alpha\beta} \\ \partial_\gamma u_\gamma = 0 \end{cases} \quad (4.3)$$

where $\Gamma > 0$, $\nu > 0$ and we have the symmetric part of the additional stress tensor:

$$\tau_{\alpha\beta} \stackrel{\mathrm{def}}{=} -\xi\left(Q_{\alpha\gamma} + \frac{\delta_{\alpha\gamma}}{d}\right)H_{\gamma\beta} - \xi H_{\alpha\gamma}\left(Q_{\gamma\beta} + \frac{\delta_{\gamma\beta}}{d}\right) + 2\xi\left(Q_{\alpha\beta} + \frac{\delta_{\alpha\beta}}{d}\right)Q_{\gamma\delta}H_{\gamma\delta} - L\partial_\beta Q_{\gamma\delta}\partial_\alpha Q_{\gamma\delta}$$

and an antisymmetric part:

$$\sigma_{\alpha\beta} \stackrel{\mathrm{def}}{=} Q_{\alpha\gamma}H_{\gamma\beta} - H_{\alpha\gamma}Q_{\gamma\beta}$$

In the rest of the notes we restrict ourselves to the case $\xi = 0$. This means that the molecules are such that they only tumble in a shear flow, but are not aligned by such a flow. In this case the system (4.3) reduces to:

$$\begin{cases} (\partial_t + u_\gamma\cdot\partial_\gamma)Q_{\alpha\beta} - \Omega_{\alpha\gamma}Q_{\gamma\beta} + Q_{\alpha\gamma}\Omega_{\gamma\beta} \\ \qquad = \Gamma\left(L\Delta Q_{\alpha\beta} - aQ_{\alpha\beta} + b[Q_{\alpha\gamma}Q_{\gamma\beta} - \frac{\delta_{\alpha\beta}}{d}\,\mathrm{tr}(Q^2)] - cQ_{\alpha\beta}\,\mathrm{tr}(Q^2)\right) \\ (\partial_t + u_\beta\partial_\beta)u_\alpha = \nu\Delta u_\alpha + \partial_\alpha p - L\partial_\beta\left(\partial_\alpha Q_{\zeta\delta}\partial_\beta Q_{\zeta\delta}\right) \\ \qquad\qquad + L\partial_\beta\left(Q_{\alpha\gamma}\Delta Q_{\gamma\beta} - \Delta Q_{\alpha\gamma}Q_{\gamma\beta}\right) \\ \partial_\gamma u_\gamma = 0 \end{cases} \quad (4.4)$$

in $\mathbb{R}^d$, $d = 2, 3$ where we use the summation convention over repeated indices and note that we have that the matrices $-aQ + b[Q^2 - \frac{\mathrm{tr}(Q^2)}{d}Id] - cQ\,\mathrm{tr}(Q^2)$ and Q commute, hence $QH - HQ = L(Q\Delta Q - \Delta QQ)$.

We also need to assume from now on that

$$c > 0 \quad (4.5)$$

This assumption is necessary from a modelling point of view (see [Maj10],[MZ10]) so that the energy $\mathcal{F}$ is bounded from below, and is also necessary for having global solutions (see Proposition 1.2 and its proof).

5. Defect patterns and their diverse descriptions

An important test of any theory attempting to describe nematics concerns the predictive capacities of the theory with regard to defect patterns. The simplest theory, the Oseen–Frank theory uses unit-vectors to describe nematics, that is the nematics are modelled as vectors fields $n : \Omega \subset \mathbb{R}^d \to \mathbb{S}^{d-1}$. In this theory defects are simply discontinuities of vectors and this is one reason why his theory has been extremely successful with mathematicians interested in regularity problems.

In Oseen–Frank theory equilibrium configurations are obtained as energy minimizers of the energy functional that we saw in a previous section, namely:

$$\mathcal{F}_{OF} = \int_\Omega |\nabla n(x)|^2 \, dx \tag{5.1}$$

where the minimization is done the space

$$\begin{aligned} &W^{1,2}_\varphi(\Omega, \mathbb{S}^{d-1}) \\ &\stackrel{\text{def}}{=} \{n : \Omega \to \mathbb{R}^d;\, n(x) \in \mathbb{S}^{d-1} \text{ a.e. } x \in \Omega, \text{ and } n(x) = \varphi(x), \text{ for a.e. } x \in \partial\Omega\} \end{aligned} \tag{5.2}$$

with $\varphi : \partial\Omega \to \mathbb{S}^{d-1}$ an imposed boundary condition.

The minimizers satisfy the Euler–Lagrange system of equations:

$$\Delta n_i = -n_i |\nabla n|^2, i = 1, \ldots, d \tag{5.3}$$

$$|n(x)| = 1 \, a.e. \, x \in \Omega \tag{5.4}$$

The above equations are known as the harmonic map equations and they have been widely studied, particularly from the point of view of regularity and partial regularity, a recent monograph in the area being [LW08].

One of the important limitations of the Oseen–Frank theory is that it only recognizes points defects, as defects with finite energy. While line and wall defects might be interpreted in this framework, they always have infinite energy.

In the more complex Ericksen theory the equilibrium configurations are obtained as minimizers of an energy functional that in its simplest form is:

$$\mathcal{F}_E(s, n) = \int_\Omega k|\nabla s(x)|^2 + s(x)|\nabla n(x)|^2 + w_0(s(x)) \, dx \tag{5.5}$$

where $k > 0$ is a constant and the minimization is in the space

$$\mathcal{A} \stackrel{\text{def}}{=} \{(n, s) : \Omega \to \mathbb{S}^{d-1} \times \mathbb{R}, (s \cdot n, s) \in H^1, \text{ and } s = s_0, \, sn = s_0 n_0 \text{ on } \partial\Omega\}$$

where $(s_0, n_0) : \partial\Omega \to [-\frac{1}{2}, 1] \times \mathbb{S}^{d-1}$ are imposed boundary conditions.

The Ericksen model was studied in several papers, mainly related to Fanghua Lin, see for instance [LP94]. In Ericksen theory the defects are regarded as the zero set of s, that is the points where the degree of order is zero, i.e. there is no preferred direction of the molecules. This proposes a compelling explanation for the defects and the energy minimizers in this theory are found indeed to have both

line and wall defects, of finite energy, unlike in the Oseen–Frank model. However, the actual experimental techniques do not allow to probe close enough to the core of a defect to experimentally check the relevance of this proposed explanation for defects.

On the other hand in the Q-tensor theory of De Gennes defects seem to have a very different nature. The problem of defects in the framework of De Gennes theory has not attracted much attention in the physics literature, perhaps because the Oseen–Frank theory, despite its limitations and thanks to its relative simplicity, has proved to be very attractive for describing defects, particularly from a topological point of view. Indeed, in fact one can talk about line and wall defects even in the framework of the Oseen–Frank theory provided that one is willing to allow for regions of infinite energy. The overall perspective in the physics literature is that if one is willing to provide a substitute for these infinite energy regions, that are the core of the defects, then the Oseen–Frank theory works reasonably well outside the cores. Indeed, it is in these defect regions of infinite energy where the theory of De Gennes proves its utility by providing finite energy, together with a possibly different structure for the core of the defects. Thus physicists often adopt an ambiguous position by using the simpler Oseen–Frank theory outside defects and substituting the more complex De Gennes energy near the core of the defects. The simple variational model of De Gennes allows, in some sense, for a justification of this perspective, see next chapter and [MZ10].

However if one attempts a definition of defects intrinsic to the Q-tensor theory, it seems that defects have quite a different nature, in this theory, when compared to the other two theories. In his relatively unknown paper [Gen72] in the 1970, De Gennes proposes to describe defects essentially as discontinuities of eigenvectors. This definition is in some sense misleading and it should be interpreted in the sense that one has a point of discontinuity x_0 for the eigenvectors of the matrix $Q(x)$ if it is not possible to find some continuous functions on a neighbourhood of x_0, say $e_i(x), i = 1, 2, 3$ so that $Q(x)e_i(x) = \lambda_i(x)e_i(x), i = 1, 2, 3$ (for some $\lambda_i(x)$) and $(e_i(x), e_j(x)) = \delta_{ij}, i, j = 1, 2, 3)$. An example is provided by the function $Q : [-1, 1]^3 \to S_0$ (where S_0 denotes the set of traceless and symmetric 3 x 3 matrices),

$$Q(x, y, z) = \begin{pmatrix} 1+x & y & 0 \\ y & 1-x & 0 \\ 0 & 0 & -2 \end{pmatrix}$$

For $x = 0$ and $(y, z) \in [-1, 1]^2$ the basis of eigenvectors is made of $\begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{pmatrix}$, $\begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$. For $y = 0$ and $(x, y) \in [-1, 1]$ the basis of eigenvectors is made of $\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, $\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$. Thus the line $\{0\} \times \{0\} \times [-1, 1]$ is a

line of discontinuity of eigenvectors. Numerical computations, [SKH95] show that a situation as described above actually appears in the equilibrium solutions.

From a mathematical point of view an interesting (and specific) feature of matrix-valued functions is that even if they are real analytic it is still possible for them to have discontinuities of eigenvectors in the sense mentioned above. These discontinuities are related to the number of distinct eigenvalues. Thus, if one is in a biaxial region, or in a uniaxial region (i.e. the number of distinct eigenvalues is constant) then one can choose the eigenvectors to be as smooth as Q(see for instance [Kat76],[Nom73]). Thus the discontinuities of eigenvectors can only occur at the biaxial-uniaxial, uniaxial-isotropic or biaxial-isotropic interfaces. However the presence of such an interface is just a necessary criterion for the existence of a discontinuity of eigenvectors. A simple criterion for checking the presence of such an interface can be provided by the " biaxiality coefficient" $\beta(Q) = 1 - 6\frac{\left(\operatorname{tr} Q^3\right)^2}{(\operatorname{tr} Q^2)^3}$ that is used in the physics literature to provide a measure of "how far a matrix is from being uniaxial", and this is motivated by the fact that $\beta(Q) \in [0, 1]$ with $\beta(Q) = 0$ if and only if $Q = 0$ or Q is uniaxial, see Appendix A.

No necessary and sufficient analytic criterion seems to be known and devising this can be an interesting problem. Having a sufficiently simple analytic criterion for detecting discontinuities of eigenvectors would be of paramount importance in attempting to study the evolution in time of these defects.

Particularly relevant to this problem could be the examples of eigenvector discontinuities that were constructed in [Lin96].

CHAPTER 2

Qualitative features of a stationary problem: Oseen–Frank limit

Liquid crystals are a complex non-Newtonian fluid and the equations that are needed to describe it are quite complicated (as we saw in the previous sections). In this chapter we restrict ourselves to physical situations in which the effect of the flow is negligible and aim to obtain qualitative features of equilibrium solutions, ignoring the effect of the flow. This is done by considering global energy minimizers of a suitable energy functional.

The Landau–De Gennes energy functional, $\mathcal{F}_{LG}$, is a nonlinear integral functional of Q and its spatial derivatives. Physical symmetries impose certain restrictions on the form of the energy functional, as noted in [Bal06]. In these notes we work with the simplest form of $\mathcal{F}_{LG}$ (see also [MN04]) with Dirichlet boundary conditions, Q_b (refer to (1.4) next section, for a precise description), on three-dimensional domains $\Omega \subset \mathbb{R}^3$ that are simply-connected. We take $\mathcal{F}_{LG}$ to be [PWS75]

$$\mathcal{F}_{LG}[Q] = \int_\Omega \frac{L}{2}|\nabla Q|^2(x) + f_B(Q(x))\,dx \tag{0.1}$$

where f_B is the bulk energy density that accounts for bulk effects,

$$|\nabla Q|^2 = \sum_{i,j,k=1}^{3} Q_{ij,k}Q_{ij,k}$$

is the elastic energy density that penalizes spatial inhomogeneities and $L > 0$ is a material-dependent elastic constant. We take f_B to be a quartic polynomial in the Q-tensor components, since this is the simplest form of f_B that allows for multiple local minima and a first-order nematic-isotropic phase transition [Gen74, Vir94]. This form of f_B has been widely-used in the literature and is defined as follows

$$f_B(Q) = \frac{\alpha(T-T^*)}{2}\operatorname{tr}\left(Q^2\right) - \frac{\bar{b}}{3}\operatorname{tr}\left(Q^3\right) + \frac{\bar{c}}{4}\left(\operatorname{tr}Q^2\right)^2$$

where $\alpha, \bar{b}, \bar{c} \in \mathbb{R}$ are material-dependent positive constants, T is the absolute temperature and T^* is a characteristic liquid crystal temperature. We work in the low-temperature regime $T < T^*$ for which $\alpha(T-T^*) < 0$. Keeping this in mind, we recast the bulk energy density as follows:

$$f_B(Q) = -\frac{a^2}{2}\operatorname{tr}\left(Q^2\right) - \frac{b^2}{3}\operatorname{tr}\left(Q^3\right) + \frac{c^2}{4}\left(\operatorname{tr}Q^2\right)^2, \tag{0.2}$$

where $a^2, b^2, c^2 \in \mathbb{R}^+$ are material-dependent and temperature-dependent non-zero constants. The equilibrium configurations (the physically observable configurations) then correspond to minimizers of $\mathcal{F}_{LG}$, subject to the imposed boundary conditions.

In Section 2, we study the limit of vanishing elastic constant $L \to 0$ for global minimizers, $Q^{(L)}$, of $\mathcal{F}_{LG}$. This study is in the spirit of the asymptotics for minimizers of Ginzburg-Landau functionals for superconductors [BBH93]. The limit $L \to 0$ is a physically relevant limit since the elastic constant L is typically very small after a suitable non-dimensionalization of the system. For $MBBA$ for instance (see [PWS75]) the elastic constant L, is of order $10^{-11}\,\mathrm{J\,m^{-1}}$, and the material coefficients A, B, C are of order $k = 10^4\,\mathrm{J\,m^{-3}}$. Since the typical size d of the liquid crystalline system is of order $10^{-6}\,\mathrm{m}$, after non-dimensionalization we obtain that on a domain of size 1 we have $\tilde{L} = \frac{L}{kd^2} \sim 10^{-3}$, $\tilde{A}, \tilde{B}, \tilde{C} \sim 1$.

We define a *limiting harmonic map* $Q^{(0)}$ as follows

$$Q^{(0)} = s_+ \left(n^{(0)} \otimes n^{(0)} - \frac{1}{3} Id \right)$$

where s_+ is defined in (1.3), $n^{(0)}$ is a minimizer of the Oseen–Frank energy, $\mathcal{F}_{OF}$, defined as:

$$\mathcal{F}_{OF}[n] = \int_\Omega n_{i,k}(x) n_{i,k}(x)\ dx,$$

subject to the fixed boundary condition $n = n_b \in C^\infty(\partial\Omega, \mathbb{S}^2)$ and Q_b and n_b are related as in (1.4). Then we can relatively easy show that there exists a sequence of global minimizers $\left\{Q^{(L_k)}\right\}$ such that $Q^{(L_k)} \overset{L_k \to 0^+}{\longrightarrow} Q^{(0)}$ strongly in the Sobolev space $W^{1,2}$. The sequence $\left\{Q^{(L_k)}\right\}$ converges uniformly to $Q^{(0)}$ as $L_k \to 0$, in the interior of Ω, away from the (possible) singularities of $Q^{(0)}$.

These results show that the predictions of the Oseen–Frank theory (described by the limiting map $Q^{(0)}$) and the Landau–De Gennes theory agree away from the singularities of $Q^{(0)}$. The presentation in this chapter follows [MZ10].

1. Preliminaries

We take our domain, $\Omega \subset \mathbb{R}^3$, to be bounded and simply-connected with smooth boundary, $\partial\Omega$. Let $S_0 \subset \mathbb{M}^{3\times 3}$ denote the space of Q-tensors, i.e.

$$S_0 \overset{def}{=} \left\{ Q \in \mathbb{M}^{3\times 3}; Q_{ij} = Q_{ji},\ Q_{ii} = 0 \right\}$$

where we have used the Einstein summation convention; the Einstein convention will be assumed in the rest of the paper. The corresponding matrix norm is defined to be

$$|Q| \overset{def}{=} \sqrt{\operatorname{tr} Q^2} = \sqrt{Q_{ij} Q_{ij}}.$$

As stated in the introduction, we take the bulk energy density term to be

$$f_B(Q) = -\frac{a^2}{2} \operatorname{tr}\left(Q^2\right) - \frac{b^2}{3} \operatorname{tr}\left(Q^3\right) + \frac{c^2}{4} \left(\operatorname{tr}(Q^2)\right)^2$$

where $a^2, b^2, c^2 \in \mathbb{R}^+$ are material-dependent and temperature-dependent non-zero constants. One can readily verify that f_B is bounded from below, and we define a

non-negative bulk energy density, $\tilde{f}_B$, that differs from f_B by an additive constant as follows:

$$\tilde{f}_B(Q) = f_B(Q) - \min_{Q \in S_0} f_B(Q). \tag{1.1}$$

It is clear that $\tilde{f}_B(Q) \geq 0$ for all $Q \in S_0$ and the set of minimizers of $\tilde{f}_B$ coincides with the set of minimizers for f_B. In Proposition 0.1, we show that the function $\tilde{f}_B$ attains its minimum on the set of uniaxial Q-tensors with constant order parameter s_+ as shown below

$$\begin{aligned} &\tilde{f}_B(Q) = 0 \Leftrightarrow Q \in Q_{min} \text{ where} \\ &Q_{\min} = \left\{ Q \in S_0, Q = s_+ \left(n \otimes n - \frac{1}{3} Id \right), n \in \mathbb{S}^2 \right\} \end{aligned} \tag{1.2}$$

with

$$s_+ = \frac{b^2 + \sqrt{b^4 + 24a^2c^2}}{4c^2}. \tag{1.3}$$

We work with Dirichlet boundary conditions, referred to as *strong anchoring* in the liquid crystal literature [Gen74]. The boundary condition $Q_b(x) \in Q_{\min}$ is smooth and is given by

$$Q_b = s_+ \left(n_b \otimes n_b - \frac{1}{3} Id \right), \; n_b \in C^\infty \left(\partial\Omega; S^2 \right). \tag{1.4}$$

We define our admissible space to be

$$\mathcal{A}_Q = \left\{ Q \in W^{1,2}(\Omega; S_0); Q = Q_b \text{ on } \partial\Omega, \text{ with } Q_b \text{ as in (1.4)} \right\}, \tag{1.5}$$

The corresponding $W^{1,2}$-norm is given by $\|Q\|_{W^{1,2}(\Omega)} = \left(\int_\Omega |Q|^2 + |\nabla Q|^2 dx \right)^{\frac{1}{2}}$. In addition to the $W^{1,2}$-norm, we also use the L^∞-norm in this paper, defined to be $\|Q\|_{L^\infty(\Omega)} = \text{ess sup}_{x \in \Omega} |Q(x)|$.

We study global minimizers of a modified Landau–De Gennes energy functional, $\tilde{F}_{LG}$, in the admissible space $\mathcal{A}_Q$. The functional $\tilde{F}_{LG}$ differs from $\mathcal{F}_{LG}$ in (0.1) by an additive constant and is defined to be

$$\tilde{F}_{LG}[Q] = \int_\Omega \frac{L}{2} Q_{ij,k}(x) Q_{ij,k}(x) + \tilde{f}_B(Q(x)) \, dx. \tag{1.6}$$

For a fixed $L > 0$, let $Q^{(L)}$ denote a global minimizer of $\tilde{F}_{LG}$ in the admissible class, $\mathcal{A}_Q$. The existence of $Q^{(L)}$ is immediate from the direct methods in the calculus of variations [Eva98]. The bulk energy density, $\tilde{f}_B$, is bounded from below, the energy density is convex in ∇Q and therefore, $\tilde{F}_{LG}$ is weakly sequentially lower semi-continuous. Moreover, it is clear that $\tilde{F}_{LG}$ and $\mathcal{F}_{LG}$ have the same set of global minimizers for a fixed set of material-dependent and temperature-dependent constants $\left\{ a^2, b^2, c^2, L \right\}$.

The global minimizer $Q^{(L)}$ is a weak solution of the corresponding Euler-Lagrange equations [BM10]

$$L\Delta Q_{ij} = -a^2 Q_{ij} - b^2 \left(Q_{ik} Q_{kj} - \frac{\delta_{ij}}{3} \operatorname{tr}(Q^2) \right) + c^2 Q_{ij} \operatorname{tr}(Q^2) \quad i,j = 1,2,3. \tag{1.7}$$

where the term $b^2 \frac{\delta_{ij}}{3} \operatorname{tr}(Q^2)$ is a Lagrange multiplier that enforces the tracelessness constraint. It follows from standard arguments in elliptic regularity that $Q^{(L)}$ is

actually a classical solution of (1.7) and $Q^{(L)}$ is smooth and real analytic (see also [MZ10]).

Finally, we introduce a *"limiting uniaxial harmonic map"* $Q^{(0)} : \Omega \to Q_{min}$; $Q^{(0)}$ is defined to be a global minimizer (not necessarily unique) of $\tilde{F}_{LG}$ in the restricted class, $\mathcal{A}_Q \cap \{Q : \Omega \to S_0, Q(x) \in Q_{min} \text{ a.e. } x \in \Omega\}$. Then $Q^{(0)}$ is necessarily of the form

$$Q^{(0)} = s_+ \left(n^{(0)} \otimes n^{(0)} - \frac{1}{3} Id \right), \tag{1.8}$$

where $n^{(0)}$ is a global minimizer of $\mathcal{F}_{OF}$ (see [BZ]),

$$\int_\Omega |\nabla n^0(x)|^2 \, dx = \min_{n \in \mathcal{A}_n} \int_\Omega |\nabla n(x)|^2 \, dx \tag{1.9}$$

in the admissible class $\mathcal{A}_n = \left\{ n \in W^{1,2}\left(\Omega; S^2\right) ; \ n = n_b \text{ on } \partial\Omega \right\}$ and n_b and Q_b are related as in (1.4). This "limiting harmonic" map $Q^{(0)}$ is therefore obtained from an energy minimizer, n^0, (not necessarily unique) within the Oseen–Frank theory for uniaxial nematic liquid crystals with constant order parameter (for more results about the relation between $n^{(0)}$ and $Q^{(0)}$ see [BZ]). It follows from standard results in harmonic maps [Vir94] that $Q^{(0)}$ has at most a finite number of isolated point singularities (points where $n^{(0)}$ has singularities). In the following sections we will elaborate on the relation between $Q^{(L)}$ and $Q^{(0)}$.

2. The limiting harmonic map and the uniform convergence

We first obtain apriori L^∞ bounds, independent of L.

PROPOSITION 2.1. *Let $\Omega \subset \mathbb{R}^3$ be a bounded and simply-connected open set with smooth boundary. Let $Q^{(L)}$ be a global minimizer of $\tilde{F}_{LG}$, (1.6), in the space (1.5). Then*

$$\|Q^{(L)}\|_{L^\infty(\Omega)} \leq \sqrt{\frac{2}{3}} s_+ \tag{2.1}$$

where s_+ is defined in (1.3).

PROOF. The proof proceeds by contradiction. In the following we drop the superscript L for convenience. We assume that there exists a point $x^* \in \bar{\Omega}$ where $|Q|$ attains its maximum and $|Q(x^*)| > \sqrt{\frac{2}{3}} s_+$. On $\partial\Omega$, $|Q| = \sqrt{\frac{2}{3}} s_+$ by our choice of the boundary condition Q_b (note that if $Q \in Q_{min}$ then $|Q| = \sqrt{\frac{2}{3}} s_+$). If Q is a global minimizer of $\tilde{F}_{LG}$ then Q is a classical solution (see Section 3.2 for regularity) of the Euler-Lagrange equations

$$L \Delta Q_{ij} = -a^2 Q_{ij} - b^2 \left(Q_{ip} Q_{pj} - \frac{1}{3} \operatorname{tr} Q^2 \delta_{ij} \right) + c^2 \left(\operatorname{tr} Q^2 \right) Q_{ij}. \tag{2.2}$$

Since the function $|Q|^2 : \bar{\Omega} \to \mathbb{R}$ must attain its maximum at $x^* \in \Omega$, we necessarily have that

$$\Delta \left(\frac{1}{2} |Q|^2 \right) (x^*) \leq 0 \tag{2.3}$$

We multiply both sides of (2.2) by Q_{ij} and obtain

$$L\,\Delta\left(\frac{1}{2}|Q|^2\right) = -a^2\,\mathrm{tr}\,Q^2 - b^2\,\mathrm{tr}\,Q^3 + c^2\left(\mathrm{tr}\,Q^2\right)^2 + L|\nabla Q|^2. \tag{2.4}$$

We note that

$$-a^2\,\mathrm{tr}(Q^2) - b^2\,\mathrm{tr}(Q^3) + c^2\left(\mathrm{tr}(Q^2)\right)^2 \geq f(|Q|) \tag{2.5}$$

where

$$f(|Q|) = -a^2|Q|^2 - \frac{b^2}{\sqrt{6}}|Q|^3 + c^2|Q|^4, \tag{2.6}$$

since $\mathrm{tr}(Q^3) \leq \frac{|Q|^3}{\sqrt{6}}$ (see Lemma 0.5 in the Appendix). One can readily verify that

$$f(|Q|) > 0 \quad \text{for } |Q| > \sqrt{\frac{2}{3}}s_+ \tag{2.7}$$

which together with (2.4) and (2.5) imply that

$$\Delta\left(\frac{1}{2}|Q|^2\right)(x) > 0 \tag{2.8}$$

for all interior points $x \in \Omega$, where $|Q(x)| > \sqrt{\frac{2}{3}}s_+$. This contradicts (2.3) and thus gives the conclusion. □

In what follows, let $e_L(Q(x))$ denote the energy density $e_L(Q(x)) \stackrel{def}{=} \frac{1}{2}|\nabla Q|^2 + \frac{\tilde{f}_B(Q(x))}{L}$. We consider the normalized energy on balls $B(x,r) \subset \Omega = \{y \in \Omega;\ |x-y| \leq r\}$

$$\mathcal{F}(Q,x,r) \stackrel{def}{=} \frac{1}{r}\int_{B(x,r)} e_L(Q(x))\,dx = \frac{1}{r}\int_{B(x,r)} \frac{1}{2}\,|\nabla Q|^2 + \frac{\tilde{f}_B(Q)}{L}\,dx. \tag{2.9}$$

We have:

LEMMA 2.2. *(Monotonicity lemma) Let $Q^{(L)}$ be a global minimizer of $\tilde{F}_{LG}$, (1.6), in the space (1.5). Then*

$$\mathcal{F}(Q^{(L)},x,r) \leq \mathcal{F}(Q^{(L)},x,R), \forall x \in \Omega, r \leq R, \text{ so that } B(x,R) \subset \Omega \tag{2.10}$$

PROOF. The proof follows a standard pattern (see for instance [LR99]) and is a consequence of the Pohozaev identity. We assume, without loss of generality, that $x = 0$ and $0 < R < d(0,\partial\Omega)$, where d denotes the Euclidean distance. Since $Q^{(L)}$ is a global energy minimizer, it is a classical solution (see Section 3.2 for regularity) of the system (1.7):

$$\Delta Q_{ij} = \frac{1}{L}\left[\frac{\partial \tilde{f}_B(Q)}{\partial Q_{ij}} + b^2\frac{\delta_{ij}}{3}\,\mathrm{tr}(Q^2)\right] \tag{2.11}$$

In (2.11) and in what follows, we drop the superscript L for convenience.

We multiply (2.11) by $x_k \cdot \partial_k Q_{ij}$, sum over repeated indices and integrate over $B(0,R)$ to obtain the following

$$
\begin{aligned}
0 &= \int_{B(0,R)} Q_{ij,ll}(x)\cdot x_k\cdot \partial_k Q_{ij}(x) - \frac{1}{L}\frac{\partial \tilde{f}_B(Q(x))}{\partial Q_{ij}}\cdot x_k\cdot \partial_k Q_{ij}(x) \\
&\qquad -\frac{1}{L}\int_{B(0,R)} b^2\frac{\delta_{ij}}{3}\operatorname{tr}(Q^2(x))\cdot x_k\cdot \partial_k Q_{ij}(x)\,dx \\
&= \underbrace{\int_{B(0,R)} Q_{ij,ll}(x)\cdot x_k\cdot \partial_k Q_{ij}(x)\,dx}_{I} - \underbrace{\int_{B(0,R)} \frac{1}{L}\frac{\partial \tilde{f}_B(Q(x))}{\partial Q_{ij}}\cdot x_k\cdot \partial_k Q_{ij}\,dx}_{II} \qquad (2.12)
\end{aligned}
$$

where we have used the tracelessness condition $Q_{ii}=0$.

Integrating by parts, we have that:

$$
\begin{aligned}
I &= \int_{B(0,R)} Q_{ij,ll}(x)x_k\partial_k Q_{ij}(x)\,dx \\
&= -\int_{B(0,R)} Q_{ij,l}(\delta_{lk}Q_{ij,k}(x)+x_kQ_{ij,kl}(x))dx + \int_{\partial B(0,R)} Q_{ij,l}x_kQ_{ij,k}\frac{x_l}{R}\,d\sigma \\
&= -\int_{B(0,R)} Q_{ij,l}(x)Q_{ij,l}(x)\,dx + 3\int_{B(0,R)}\frac{1}{2}Q_{ij,l}(x)Q_{ij,l}(x)\,dx \\
&\quad -\int_{\partial B(0,R)}\frac{Q_{ij,l}(x)Q_{ij,l}(x)}{2}\frac{x_k\cdot x_k}{R}\,d\sigma + \int_{\partial B(0,R)}\frac{(Q_{ij,k}(x)\cdot x_k)^2}{R}\,d\sigma \qquad (2.13)
\end{aligned}
$$

$$
\begin{aligned}
II &= \int_{B(0,R)}\frac{1}{L}\frac{\partial \tilde{f}_B(Q(x))}{\partial Q_{ij}}\cdot x_k\cdot\partial_k Q_{ij}(x)\,dx = \frac{1}{L}\int_{B(0,R)}\partial_k\tilde{f}_B(Q(x))\cdot x_k\,dx \\
&= -\frac{3}{L}\int_{B(0,R)}\tilde{f}_B(Q(x))\,dx + \frac{1}{L}\int_{\partial B(0,R)}\tilde{f}_B(Q(x))\cdot\frac{x_k\cdot x_k}{R}\,d\sigma \qquad (2.14)
\end{aligned}
$$

Hence (2.12) becomes:

$$
\begin{aligned}
&-\int_{B(0,R)}\frac{Q_{ij,l}(x)Q_{ij,l}(x)}{2}+\frac{\tilde{f}_B(Q(x))}{L}\,dx \\
&\qquad + R\int_{\partial B(0,R)}\frac{Q_{ij,k}(x)Q_{ij,k}(x)}{2}+\frac{\tilde{f}_B(Q(x))}{L}\,d\sigma \\
&\qquad = \frac{1}{R}\int_{\partial B(0,R)}(Q_{ij,k}(x)\cdot x_k)^2\,d\sigma + 2\int_{B(0,R)}\frac{\tilde{f}_B(Q(x))}{L}\,dx \qquad (2.15)
\end{aligned}
$$

We have

$$
\begin{aligned}
&\frac{\partial}{\partial R}\left(\frac{1}{R}\int_{B(0,R)}\frac{Q_{ij,l}(x)Q_{ij,l}(x)}{2}+\frac{\tilde{f}_B(Q(x))}{L}\,dx\right) = \\
&\qquad -\frac{1}{R^2}\int_{B(0,R)}\frac{Q_{ij,l}(x)\cdot Q_{ij,l}(x)}{2}+\frac{\tilde{f}_B(Q(x))}{L}\,dx \\
&\qquad +\frac{1}{R}\int_{\partial B(0,R)}\frac{Q_{ij,l}(x)\cdot Q_{ij,l}(x)}{2}+\frac{\tilde{f}_B(Q(x))}{L}\,d\sigma. \qquad (2.16)
\end{aligned}
$$

The right-hand side of (2.16) is positive from (2.15) and hence the conclusion. □

LEMMA 2.3. *($W^{1,2}$-convergence to harmonic maps) Let $\Omega \subset \mathbb{R}^3$ be a simply-connected bounded open set with smooth boundary. Let $Q^{(L)}$ be a global minimizer of $\tilde{F}_{LG}$, (1.6), in the space (1.5). Then there exists a sequence $L_k \to 0$ so that $Q^{(L_k)} \to Q^{(0)}$ strongly in $W^{1,2}(\Omega; S_0)$, where $Q^{(0)}$ is a limiting harmonic map defined in (1.8).*

PROOF. Our proof follows closely, up to a point, the ideas of Proposition 1 in [BBH93]. Firstly, we note that the limiting harmonic map $Q^{(0)}$ belongs to our admissible space $\mathcal{A}_Q$ and since $Q^{(0)}(x) \in Q_{\min}$, a.e. $x \in \Omega$ (see Section 1) we have that $\tilde{f}_B\left(Q^{(0)}(x)\right) = 0$ a.e. $x \in \Omega$. Therefore

$$\int_\Omega \frac{1}{2} Q^{(L)}_{ij,k}(x) Q^{(L)}_{ij,k}(x)\, dx \leq \int_\Omega \frac{1}{2} Q^{(L)}_{ij,k}(x) Q^{(L)}_{ij,k}(x) + \frac{1}{L} \tilde{f}_B(Q^{(L)}(x))\, dx \leq \int_\Omega \frac{1}{2} Q^{(0)}_{ij,k}(x) Q^{(0)}_{ij,k}(x)\, dx. \tag{2.17}$$

The $Q^{(L)}$'s are subject to the same boundary condition, Q_b, for all L. Therefore (2.17) shows that the $W^{1,2}$-norms of the $Q^{(L)}$'s are bounded uniformly in L. Hence there exists a weakly-convergent subsequence $Q^{(L_k)}$ such that $Q^{(L_k)} \rightharpoonup Q^{(1)}$ in $W^{1,2}$, for some $Q^{(1)} \in \mathcal{A}_Q$ as $L_k \to 0$. Using the lower semicontinuity of the $W^{1,2}$-norm with respect to the weak convergence, we have that

$$\int_\Omega |\nabla Q^{(1)}(x)|^2\, dx \leq \int_\Omega |\nabla Q^{(0)}(x)|^2\, dx. \tag{2.18}$$

Relation (2.17) shows that $\int_\Omega \tilde{f}_B(Q^{(L_k)}(x))\, dx \leq L_k \int_\Omega Q^{(0)}_{ij,k}(x) Q^{(0)}_{ij,k}(x)\, dx$ and hence $\int_\Omega \tilde{f}_B(Q^{(L_k)}(x))\, dx \to 0$ as $L_k \to 0$. Taking into account that $\tilde{f}_B(Q) \geq 0, \forall Q \in S_0$ we have that, on a subsequence L_{k_j}, $\tilde{f}_B(Q^{(L_{k_j})}(x)) \to 0$ for almost all $x \in \Omega$. From Proposition 3.1, we know that $\tilde{f}_B(Q) = 0$ if and only if $Q \in Q_{\min}$ i.e. if $Q = s_+\left(n \otimes n - \frac{1}{3} Id\right)$ for $n \in S^2$. On the other hand, the sequence $Q^{(L_k)}$ converges weakly in $W^{1,2}$ and, on a subsequence, strongly in L^2 to $Q^{(1)}$. Therefore, the weak limit $Q^{(1)}$ is of the form

$$Q^{(1)}(x) = s_+\left(n^{(1)}(x) \otimes n^{(1)}(x) - \frac{1}{3} Id\right),\ n^{(1)}(x) \in \mathbb{S}^2,\ a.e.\, x \in \Omega \tag{2.19}$$

It was proved in [BZ] that if $Q^{(1)} \in W^{1,2}$ and the domain Ω is simply-connected, we can assume, without loss of generality, that $n^{(1)} \in W^{1,2}(\Omega, \mathbb{S}^2)$ and its trace is n_b. Then (2.19) implies $|\nabla Q^{(1)}(x)|^2 = 2 s_+^2 |\nabla n^{(1)}(x)|^2$ for a.e. $x \in \Omega$. Also, recalling the definition of $Q^{(0)}$ from Section 1 we have $|\nabla Q^{(0)}(x)|^2 = 2 s_+^2 |\nabla n^{(0)}(x)|^2$ for a.e. $x \in \Omega$.

Combining (2.18) with (1.9) and the observations in the previous paragraph, we obtain $\int_\Omega |\nabla n^{(1)}(x)|^2\, dx = \int_\Omega |\nabla n^{(0)}(x)|^2\, dx$ and $\int_\Omega |\nabla Q^{(1)}(x)|^2\, dx = \int_\Omega |\nabla Q^{(0)}(x)|^2\, dx$. Then:

$$\int_\Omega |\nabla Q^{(0)}(x)|^2\,dx \le \liminf_{L_{k_j}\to 0}\int_\Omega |\nabla Q^{(L_{k_j})}(x)|^2\,dx \le \limsup_{L_{k_j}\to 0}\int_\Omega |\nabla Q^{(L_{k_j})}(x)|^2\,dx \le \int_\Omega |\nabla Q^{(0)}(x)|^2\,dx,$$

which demonstrates that $\lim_{L_{k_j}\to 0}\|\nabla Q^{(L_{k_j})}\|_{L^2} = \|\nabla Q^{(0)}\|_{L^2}$. This together with the weak convergence $Q^{(L_{k_j})} \rightharpoonup Q^{(0)}$ suffices to show the strong convergence $Q^{(L_{k_j})} \to Q^{(0)}$ in $W^{1,2}$. □

The following has an elementary proof, that will be omitted:

LEMMA 2.4. *The function $\tilde f_B : S_0 \to \mathbb{R}_+$ is locally Lipschitz.*

We can now prove the uniform convergence of the bulk energy density in the interior, away from the singularities of the limiting harmonic map $Q^{(0)}$.

PROPOSITION 2.5. *Let $\Omega \subset \mathbb{R}^3$ be a simply-connected bounded open set with smooth boundary. Let $Q^{(L)}$ be a global minimizer of $\tilde F_{LG}$, (1.6), in the space (1.5). Assume that we have a sequence $Q^{(L_k)}$ with $L_k \to 0$ as $k \to \infty$, such that $Q^{(L_k)} \to Q^{(0)}$ as $k \to \infty$.*

For any compact $K \subset \Omega$ such that $Q^{(0)}$ has no singularity in K we have

$$\lim_{L_k\to 0} \tilde f_B(Q^{(L_k)}(x)) = 0 \quad x \in K \tag{2.20}$$

and the limit is uniform on K.

PROOF. *Lemma* 2.3 shows that the strong limit $Q^{(0)}$ is a limiting harmonic map, as defined in Section 1, $Q^{(0)} = s_+(n^{(0)}(x)\otimes n^{(0)}(x) - \frac{1}{3}Id)$ where $n^{(0)} \in W^{1,2}(\Omega, \mathbb{S}^2)$ is a global energy minimizer of the harmonic map problem, subject to the boundary condition $n = n_b$ on $\partial\Omega$.

Let $\alpha_{L_k} = \tilde f_B(Q^{(L_k)}(x_0))$, for $x_0 \in K$ an arbitrary point. *Proposition* 2.1 and *Lemma* 2.4 imply that there exists a constant β (independent of x_0) so that

$$|\tilde f_B(Q^{(L)}(x)) - \tilde f_B(Q^{(L)}(y))| \le \beta |Q^{(L)}(x) - Q^{(L)}(y)| \tag{2.21}$$

for any $x, y \in \Omega, L > 0$.

We then have

$$\begin{aligned}\alpha_{L_k} &\le \tilde f_B(Q^{(L_k)}(x)) + \beta|Q^{(L_k)}(x) - Q^{(L_k)}(x_0)|\\ &\le \tilde f_B(Q^{(L_k)}(x)) + \beta\|\nabla Q^{(L_k)}\|_{L^\infty(K')}|x - x_0|\\ &\le \tilde f_B(Q^{(L_k)}(x)) + \frac{\tilde C\beta}{\sqrt{L_k}}|x - x_0|, \forall x \in K'\end{aligned} \tag{2.22}$$

where $K' \subset \Omega$ is a compact neighborhood of K to be precisely defined later. In the last relation above we use *Lemma A.*1 from [BBH93] and the apriori bound given by *Proposition* 2.1. For reader's convenience we recall that *Lemma A.*1 in [BBH93] states that if u is a scalar-valued function such that $-\Delta u = f$ on $\Omega \subset \mathbb{R}^n$ then $|\nabla u(x)|^2 \le C\left(\|f\|_{L^\infty(\Omega)}\|u\|_{L^\infty(\Omega)} + \frac{1}{\mathrm{dist}^2(x,\partial\Omega)}\|u\|^2_{L^\infty(\Omega)}\right)$ where C is a constant

that depends on n only. In our case the constant $\tilde{C}$ depends on the dimension, $n = 3$, on a^2, b^2, c^2 and on the distance $\sup_{y\in K'} d(y, \partial\Omega)$ only.

From (2.22) we have that

$$\alpha_{L_k} - \frac{\tilde{C}\beta\rho_k}{\sqrt{L_k}} \leq \tilde{f}_B(Q^{(L_k)}(x)), \forall x \in K', |x - x_0| < \rho_k \tag{2.23}$$

We argue similarly as in [BBH93] and divide by L_k and integrate over $B_{\rho_k}(x_0)$ to obtain:

$$\frac{\rho_k^3}{L_k}(\alpha_{L_k} - \frac{\tilde{C}\beta\rho_k}{\sqrt{L_k}}) \leq \int_{B_{\rho_k}(x_0)} \frac{\tilde{f}_B(Q^{L_k}(x))}{L_k}\, dx \tag{2.24}$$

Take an arbitrary $\varepsilon > 0$. Recall that K is a compact set that does not contain singularities of $Q^{(0)}$. Then there exists a larger compact set K', so that $K \subset K'$, that does not contain singularities either, and a constant $C_{K'}$ such that $|\nabla Q^{(0)}(x)|^2 < C_{K'}, \forall x \in K'$. For R_0 small enough, with $R_0 < \operatorname{dist}(K, \partial\Omega)$ and such that $B(x_0, R_0) \subset K', \forall x_0 \in K$ we have

$$\frac{1}{R_0}\int_{B_{R_0}(x_0)} \frac{|\nabla Q^{(0)}(x)|^2}{2}\, dx \leq \frac{4\pi}{6} C_{K'} R_0^2 \leq \frac{\varepsilon}{3}, \forall x_0 \in K \tag{2.25}$$

We fix an R_0 as before. As $Q^{(L_k)} \to Q^{(0)}$ in $W^{1,2}$, we have that there exists an $\bar{L}_0 > 0$ so that:

$$\frac{1}{R_0}\int_{B_{R_0}(x_0)} \frac{|\nabla Q^{(L_k)}(x)|^2}{2}\, dx < \frac{1}{R_0}\int_{B_{R_0}(x_0)} \frac{|\nabla Q^{(0)}(x)|^2}{2}\, dx + \frac{\varepsilon}{3}, \text{ for } L_k < \bar{L}_0, \quad \forall x_0 \in K$$

The arguments in [BBH93] fail to work in our case as we have a three dimensional domain, unlike in the quoted paper, where the domain is two dimensional. In our case, using the monotonicity formula from *Lemma* 2 and taking $\rho_k < R_0$ we obtain:

$$\int_{B_{\rho_k}(x_0)} \frac{\tilde{f}_B(Q^{(L_k)}(x))}{L_k}\, dx \leq \frac{\rho_k}{R_0}\int_{B_{R_0}(x_0)} \frac{|\nabla Q^{(L_k)}(x)|^2}{2} + \frac{\tilde{f}_B(Q^{(L_k)}(x))}{L_k}\, dx \leq \rho_k\left(\frac{2\varepsilon}{3} + \frac{\varepsilon}{3}\right) \tag{2.26}$$

for $L_k < \min\left\{\bar{L}_1, \bar{L}_0\right\}$ with $\bar{L}_1$ small so that $\frac{1}{R_0}\int_{B_{R_0}(x_0)} \frac{\tilde{f}_B(Q^{(L_k)}(x))}{L_k}\, dx < \frac{\varepsilon}{3}$ (note that there exists such an $\bar{L}_1$ as the proof of *Lemma* 3 shows that

$$\int_\Omega \frac{\tilde{f}_B(Q^{(L_k)}(x))}{L_k}\, dx = o(1)$$

as $L_k \to 0$).

We take $\rho_k = \frac{\alpha_{L_k}\sqrt{L_k}}{2\tilde{C}\beta}$. Then, from (2.24) and (2.26) we obtain

$$\alpha_{L_k}^3 < 8(\tilde{C}\beta)^2\varepsilon$$

for $L_k < \min\{\bar{L}_0, \bar{L}_1\}$. As $\varepsilon > 0$ is arbitrary and the estimate on $\alpha_{L_k} = \tilde{f}_B(Q^{(L_k)}(x_0))$ with $x_0 \in K$ is obtained in a manner independent of x_0, we have the claimed result. □

We also need the following

LEMMA 2.6. *There exists $\varepsilon_0 > 0$ so that:*

$$\frac{1}{\tilde{C}}\tilde{f}_B(Q) \leq \sum_{i,j=1}^{3}\left(\frac{\partial \tilde{f}_B(Q)}{\partial Q_{ij}} + b^2\frac{\delta_{ij}}{3}\operatorname{tr}(Q^2)\right)^2 \leq \tilde{C}\tilde{f}_B(Q)$$

$$\forall Q \in S_0 \text{ such that } |Q - s_+(n\otimes n - \frac{1}{3}Id)| \leq \varepsilon_0, \text{ for some } n \in \mathbb{S}^2 \tag{2.27}$$

where $s_+ = \frac{b^2+\sqrt{b^4+24a^2c^2}}{4c^2}$ and the constant $\tilde{C}$ is independent of Q, but depends on a^2, b^2, c^2.

PROOF. Recall from Proposition 3.1, [BM10] that $\tilde{f}_B(Q) \geq 0$ and $\tilde{f}_B(Q) = 0 \leftrightarrow Q = s_+(n\otimes n - \frac{1}{3}Id)$ with $s_+ = \frac{b^2+\sqrt{b^4+24a^2c^2}}{4c^2}$ and $n \in \mathbb{S}^2$.

Let the eigenvalues of Q be $x, y, -x-y$. We define $F(x,y) \overset{def}{=} -a^2(x^2+y^2+xy) + b^2xy(x+y) + c^2(x^2+y^2+xy)^2$ and $D \overset{def}{=} \min_{(x,y)\in\mathbb{R}^2} F(x,y)$. Then $\tilde{F}(x,y) \overset{def}{=} F(x,y) - D = \tilde{f}_B(Q)$.

Then $\tilde{F} = 0$ only at three pairs (x,y) namely $(-\frac{s_+}{3}, -\frac{s_+}{3}), (-\frac{s_+}{3}, 2\frac{s_+}{3})$ and $(2\frac{s_+}{3}, -\frac{s_+}{3})$.

On the other hand we have

$$\sum_{i,j=1}^{3}\left(\frac{\partial \tilde{f}_B}{\partial Q_{ij}} + \frac{b^2\delta_{ij}}{3}\operatorname{tr}(Q^2)\right)^2 = a^4\operatorname{tr}(Q^2) + (\frac{b^4}{6} - 2a^2c^2)(\operatorname{tr}(Q^2))^2$$
$$+c^4(\operatorname{tr}(Q^2))^3 + 2a^2b^2\operatorname{tr}(Q^3) - 2b^2c^2\operatorname{tr}(Q^2)\operatorname{tr}(Q^3) \tag{2.28}$$

(where we used the identity $\operatorname{tr}(Q^4) = \frac{\left[\operatorname{tr}(Q^2)\right]^2}{2}$, valid for a traceless symmetric 3×3 matrix).

If we denote $h(Q) = \sum_{i,j=1}^{3}\left(\frac{\partial \tilde{f}_B(Q)}{\partial Q_{ij}} + b^2\frac{\delta_{ij}}{3}\operatorname{tr}(Q^2)\right)^2$ we have $h(Q) = H(x,y)$ where $H : \mathbb{R}^2 \to \mathbb{R}$ is given by

$$H(x,y) \overset{def}{=} 2a^4(x^2+y^2+xy) + 4(\frac{b^4}{6} - 2a^2c^2)(x^2+y^2+xy)^2 + 8c^4(x^2+y^2+xy)^3$$
$$+12b^2c^2xy(x+y)(x^2+y^2+xy) - 6a^2b^2xy(x+y)$$

We claim that there exist $\varepsilon_1, \varepsilon_2, \varepsilon_3 > 0$ so that

$$\frac{1}{\tilde{C}}\tilde{F}(x,y) \leq H(x,y) \leq \tilde{C}\tilde{F}(x,y),$$
$$\forall (x,y) \in B_{\varepsilon_1}(-\frac{s_+}{3}, -\frac{s_+}{3}),\ B_{\varepsilon_2}(-\frac{s_+}{3}, 2\frac{s_+}{3}),\ B_{\varepsilon_3}(2\frac{s_+}{3}, -\frac{s_+}{3}) \tag{2.29}$$

which gives the conclusion.

We prove the inequality (2.29) only for $(x,y) \in B_{\varepsilon_1}(-\frac{s_+}{3}, -\frac{s_+}{3})$; the other two cases can be dealt with similarly.

Careful computations show:

$$H(-\frac{s_+}{3},-\frac{s_+}{3})=\frac{\partial H}{\partial x}(-\frac{s_+}{3},-\frac{s_+}{3})=\frac{\partial H}{\partial y}(-\frac{s_+}{3},-\frac{s_+}{3})=0$$

$$\frac{\partial^2 H}{\partial y^2}(-\frac{s_+}{3},-\frac{s_+}{3})=\frac{\partial^2 H}{\partial x^2}(-\frac{s_+}{3},-\frac{s_+}{3})=4(b^4+6a^2c^2)\frac{b^4+12a^2c^2+b^2\sqrt{b^4+24a^2c^2}}{24c^4}$$

$$\frac{\partial^2 H}{\partial x\partial y}(-\frac{s_+}{3},-\frac{s_+}{3})=-2(b^4-12a^2c^2)\frac{b^4+12a^2c^2+b^2\sqrt{b^4+24a^2c^2}}{24c^4}$$

$$\tilde{F}(-\frac{s_+}{3},-\frac{s_+}{3})=\frac{\partial \tilde{F}}{\partial x}(-\frac{s_+}{3},-\frac{s_+}{3})=\frac{\partial \tilde{F}}{\partial y}(-\frac{s_+}{3},-\frac{s_+}{3})=0$$

$$\frac{\partial^2 \tilde{F}}{\partial y^2}(-\frac{s_+}{3},-\frac{s_+}{3})=\frac{\partial^2 \tilde{F}}{\partial x^2}(-\frac{s_+}{3},-\frac{s_+}{3})=\frac{1}{4c^2}(b^4+12a^2c^2+b^2\sqrt{b^4+24a^2c^2})$$

$$\frac{\partial^2 \tilde{F}}{\partial x\partial y}(-\frac{s_+}{3},-\frac{s_+}{3})=3a^2$$

Let $(x_0,y_0)=(-\frac{s_+}{3},-\frac{s_+}{3})$. We have

$$\frac{H(x,y)}{\tilde{F}(x,y)}=\frac{H_1(x,y)+R_H(x,y)}{\tilde{F}_1(x,y)+R_{\tilde{F}}(x,y)}$$

where $H_1(x,y)=(x-x_0)^2\frac{\partial^2 H}{\partial x^2}(x_0,y_0)+2(x-x_0)(y-y_0)\frac{\partial^2 H}{\partial x\partial y}(x_0,y_0)+(y-y_0)^2\frac{\partial^2 H}{\partial y^2}(x_0,y_0)$ and $\tilde{F}_1(x,y)=(x-x_0)^2\frac{\partial^2 \tilde{F}}{\partial x^2}(x_0,y_0)+2(x-x_0)(y-y_0)\frac{\partial^2 \tilde{F}}{\partial x\partial y}(x_0,y_0)+(y-y_0)^2\frac{\partial^2 \tilde{F}}{\partial y^2}(x_0,y_0)$ with R_H, $R_{\tilde{F}}$ the remainders in the Taylor expansions around (x_0,y_0).

From the definition of the Taylor expansions, and the continuity of the eigenvalues as functions of matrices [Nom73], we have that there exists $\varepsilon_0,\varepsilon_1>0$ so that on $B_{\varepsilon_1}(x_0,y_0)$ we have

$$|R_H(x,y)|\leq\frac{1}{2}H_1(x,y) \text{ and } |R_{\tilde{F}}(x,y)|\leq\frac{1}{2}\tilde{F}_1(x,y),\forall(x,y)\in B_{\varepsilon_1}(x_0,y_0) \quad (2.30)$$

On the other hand we have

$$\tilde{F}_1(x,y)\frac{3}{\tilde{C}}\leq H_1(x,y)\leq\frac{\tilde{C}}{3}\tilde{F}_1(x,y)\,\forall(x,y)\in\mathbb{R}^2 \quad (2.31)$$

where $\tilde{C}>3$ is a constant depending only on a^2,b^2 and c^2 hence, combining (2.30) and (2.31), we get:

$$\tilde{F}(x,y)\frac{1}{\tilde{C}}\leq H(x,y)\leq\tilde{C}\tilde{F}(x,y),\forall(x,y)\in B_{\varepsilon_1}(-\frac{s_+}{3},-\frac{s_+}{3}) \quad (2.32)$$

which yields claim (2.29) for $(x,y)\in B_{\varepsilon_1}(-\frac{s_+}{3},-\frac{s_+}{3})$. The other two cases can be analyzed analogously. □

We continue by proving a Bochner-type inequality that is crucial for the derivation of uniform (in L) Lipschitz bounds, away from the singularities of the limiting harmonic map. This type of inequalities were first used (to the best of our knowledge) in the context of harmonic maps (see [Sch84] and the references there) and later adapted to other, more complicated contexts (see for instance [CL95]). The main difficulty in the proof of Proposition 2.9 (to follow) is the derivation of the next lemma.

LEMMA 2.7. *There exists $\varepsilon_0 > 0$ and a constant $C > 0$, independent of L, so that for $Q^{(L)}$ a global minimizer of $\tilde{F}_{LG}$, (1.6), in the space (1.5), we have*

$$-\Delta e_L(Q^{(L)})(x) \leq C e_L^2(Q^{(L)}(x)) \tag{2.33}$$

provided there exists a ball $B_{\rho(x)}(x)$ for some $\rho(x) > 0$ such that

$$|Q^{(L)}(y) - s_+ \left(m(y) \otimes m(y) - \frac{1}{3} Id \right)| < \varepsilon_0$$

with $m(y) \in \mathbb{S}^2$ for all $y \in B_{\rho(x)}(x)$.

PROOF. In the following we drop the superscript L for convenience. We have:

$$\begin{aligned} -\Delta \left(\frac{Q_{ij,k} Q_{ij,k}}{2} \right) &= -\Delta Q_{ij,k} Q_{ij,k} - Q_{ij,kl} Q_{ij,kl} \leq \\ &\leq -\partial_k \left[\frac{1}{L} \frac{\partial \tilde{f}_B}{\partial Q_{ij}}(Q(x)) + \frac{b^2 \delta_{ij}}{3L} \operatorname{tr}(Q^2) \right] Q_{ij,k} \\ &= -\partial_k \left[\frac{1}{L} \frac{\partial \tilde{f}_B}{\partial Q_{ij}}(Q(x)) \right] Q_{ij,k} \end{aligned} \tag{2.34}$$

On the other hand:

$$\begin{aligned} -\Delta \left[\frac{1}{L} \tilde{f}_B(Q(x)) \right] &= -\partial_k \left(\frac{1}{L} \frac{\partial \tilde{f}_B}{\partial Q_{ij}}(Q(x)) \partial_k Q_{ij} \right) \\ &= -\partial_k \left[\left[\frac{1}{L} \frac{\partial \tilde{f}_B}{\partial Q_{ij}}(Q(x)) + \frac{b^2 \delta_{ij}}{3L} \operatorname{tr}(Q^2) \right] \partial_k Q_{ij} \right] \\ &= - \underbrace{\left(\frac{1}{L} \frac{\partial \tilde{f}_B}{\partial Q_{ij}}(Q(x)) + \frac{b^2 \delta_{ij}}{3L} \operatorname{tr}(Q^2) \right)}_{= \Delta Q_{ij}} \times \Delta Q_{ij} \end{aligned}$$

$$-\partial_k \left(\frac{1}{L} \frac{\partial \tilde{f}_B}{\partial Q_{ij}}(Q(x)) \right) Q_{ij,k} \leq -\partial_k \left(\frac{1}{L} \frac{\partial \tilde{f}_B}{\partial Q_{ij}}(Q(x)) \right) Q_{ij,k} \tag{2.35}$$

We take $\varepsilon_1 > 0$ a small number, to be made precise later. For any such ε_1 we can pick $\varepsilon_0 > 0$ small enough so that if the eigenvalues of $Q(x)$ are $(\lambda_1, \lambda_2, -\lambda_1 - \lambda_2)$ then one of the three numbers $(\lambda_1 + \frac{s_+}{3})^2 + (\lambda_2 + \frac{s_+}{3})^2 + (\lambda_1 + \lambda_2 + 2\frac{s_+}{3})^2$, $(\lambda_1 + \frac{s_+}{3})^2 + (\lambda_2 - 2\frac{s_+}{3})^2 + (\lambda_1 + \lambda_2 - \frac{s_+}{3})^2$, $(\lambda_1 - 2\frac{s_+}{3})^2 + (\lambda_2 + \frac{s_+}{3})^2 + (\lambda_1 + \lambda_2 - \frac{s_+}{3})^2$ is less than or equal to ε_1 (this can be done because the eigenvalues are continuous functions of matrices, [Kat76], and the matrix $s_+(n \otimes n - \frac{1}{3} Id)$ has eigenvalues $-\frac{s_+}{3}, -\frac{s_+}{3}$ and $2\frac{s_+}{3}$). Note moreover that we need to choose ε_0 to be smaller than

the choice (of ε_0) in *Lemma* 2.6 as we will need to use that lemma in the remainder of this proof.

For the matrix $Q(x)$, let us denote its eigenvectors by $n_1(x), n_2(x), n_3(x)$ and let $\lambda_1(x), \lambda_2(x)$, $\lambda_3(x) = -\lambda_1(x) - \lambda_2(x)$ denote the corresponding eigenvalues. From the preceeding discussion, we can, without loss of generality, assume that

$$(\lambda_1 + \frac{s_+}{3})^2 + (\lambda_2 + \frac{s_+}{3})^2 + (\lambda_1 + \lambda_2 + 2\frac{s_+}{3})^2 < \varepsilon_1 \tag{2.36}$$

We define the matrix

$$Q^x \stackrel{def}{=} -\frac{s_+}{3} n_1(x) \otimes n_1(x) - \frac{s_+}{3} n_2(x) \otimes n_2(x) + \frac{2s_+}{3} n_3(x) \otimes n_3(x)$$

(Note that $Q^x = s_+(n_3(x) \otimes n_3(x) - \frac{1}{3}Id)$).

Taking into account (2.36) and the fact that $Q(x)$ and Q^x have the same eigenvectors, we have :

$$\mathrm{tr}(Q(x) - Q^x)^2 = (\lambda_1 + \frac{s_+}{3})^2 + (\lambda_2 + \frac{s_+}{3})^2 + (\lambda_1 + \lambda_2 + 2\frac{s_+}{3})^2 < \varepsilon_1 \tag{2.37}$$

Using the of Taylor expansion of $\frac{1}{2}\frac{\partial^2 \tilde{f}_B}{\partial Q_{ij}\partial Q_{mn}}(Q(x))$ around Q^x we obtain:

$$\frac{1}{2}\frac{\partial^2 \tilde{f}_B}{\partial Q_{ij}\partial Q_{mn}}(Q(x)) = \frac{1}{2}\frac{\partial^2 \tilde{f}_B}{\partial Q_{ij}\partial Q_{mn}}(Q^x) + \frac{1}{2}\frac{\partial^3 \tilde{f}_B}{\partial Q_{ij}\partial Q_{mn}\partial Q_{pq}}(Q^x)(Q_{pq}(x) - Q^x_{pq}) + \mathcal{R}^{ijmn}(Q^x, Q(x)) \tag{2.38}$$

where $\mathcal{R}^{ijmn}(Q^x, Q(x))$ is the remainder.

From (2.38) we have:

$$-\partial_k\left(\frac{1}{L}\frac{\partial \tilde{f}_B}{\partial Q_{ij}}(Q(x))\right) Q_{ij,k} = -\frac{1}{L}\frac{\partial^2 \tilde{f}_B}{\partial Q_{ij}\partial Q_{mn}} Q_{mn,k} Q_{ij,k} =$$

$$= \underbrace{-\frac{1}{L}\frac{\partial^2 \tilde{f}_B}{\partial Q_{ij}\partial Q_{mn}}(Q^x) Q_{mn,k} Q_{ij,k}}_{\leq 0}$$

$$-\frac{1}{L}\frac{\partial^3 \tilde{f}_B}{\partial Q_{ij}\partial Q_{mn}\partial Q_{pq}}(Q^x)(Q_{pq}(x) - Q^x_{pq}) Q_{ij,k} Q_{mn,k} -$$

$$-\frac{1}{L}\mathcal{R}^{ijmn}(Q(x), Q^x) Q_{ij,k} Q_{mn,k} \leq$$

$$\leq \frac{C_0\delta}{L^2}\Sigma^3_{i,j,m,n=1}\left(\frac{\partial^3 \tilde{f}_B}{\partial Q_{ij}\partial Q_{mn}\partial Q_{pq}}(Q^x)\right)^2 (Q_{pq}(x) - Q^x_{pq})^2$$

$$+\frac{C_0\delta}{L^2}\sum_{i,j,m,n=1}^{3}\left(\mathcal{R}^{ijmn}\right)^2(Q(x), Q^x) + \frac{1}{\delta}|\nabla Q|^4 \leq$$

$$\leq \frac{\delta}{L^2}\sum_{i,j,m,n=1}^{3}\left[\bar{C}_0\left(\frac{\partial^3 f}{\partial Q_{ij}\partial Q_{mn}\partial Q_{pq}}(Q^x)\right)^2+C_1\right](Q_{pq}(x)-Q^x_{pq})^2+\frac{1}{\delta}|\nabla Q|^4$$
$$\leq \frac{C_2\delta}{L^2}\operatorname{tr}(Q(x)-Q^x)^2+\frac{1}{\delta}|\nabla Q|^4 \tag{2.39}$$

where $0<\delta<1$ and $C_0,\bar{C}_0,C_1,C_2$ are independent of L and x. For the first term in the second line above we use the fact that the Hessian matrix of a function $\tilde{f}_B$ is non-negative definite at a global minimum (which holds true in our case as well, as one can easily check, even though we have $\tilde{f}_B$ restricted to the linear space S_0).

Let us recall (from the proof of the previous lemma) the definitions of F and $\tilde{F}$. Then, for a matrix $Q\in S_0$ with eigenvalues $(\lambda_1,\lambda_2,-\lambda_1-\lambda_2)$ we have

$$\tilde{f}_B(Q)=\tilde{F}(\lambda_1,\lambda_2) \tag{2.40}$$

We claim that for $\varepsilon_1>0$ small enough there exists C_2 independent of L,λ_1,λ_2 so that

$$C_2\left((\lambda_1+\frac{s_+}{3})^2+(\lambda_2+\frac{s_+}{3})^2+(\lambda_1+\lambda_2+2\frac{s_+}{3})^2\right)\leq\tilde{F}(\lambda_1,\lambda_2)$$
$$\text{for all } (\lambda_1,\lambda_2) \text{ so that} (\lambda_1+\frac{s_+}{3})^2+(\lambda_2+\frac{s_+}{3})^2+(\lambda_1+\lambda_2+2\frac{s_+}{3})^2<\varepsilon_1. \tag{2.41}$$

Careful computations show:

$$\tilde{F}(-\frac{s_+}{3},-\frac{s_+}{3})=\frac{\partial\tilde{F}}{\partial\lambda_1}(-\frac{s_+}{3},-\frac{s_+}{3})=\frac{\partial\tilde{F}}{\partial\lambda_2}(-\frac{s_+}{3},-\frac{s_+}{3})=0$$

$$\frac{\partial^2\tilde{F}}{\partial\lambda_2^2}(-\frac{s_+}{3},-\frac{s_+}{3})=\frac{\partial^2\tilde{F}}{\partial\lambda_1^2}(-\frac{s_+}{3},-\frac{s_+}{3})=\frac{1}{4c^2}(b^4+12a^2c^2+b^2\sqrt{b^4+24a^2c^2})$$

$$\frac{\partial^2\tilde{F}}{\partial\lambda_1\partial\lambda_2}(-\frac{s_+}{3},-\frac{s_+}{3})=3a^2$$

Using a Taylor expansion around $(\lambda_1,\lambda_2)=(-\frac{s_+}{3},-\frac{s_+}{3})$ we have

$$\tilde{F}(\lambda_1,\lambda_2)=\frac{1}{8c^2}\left(b^4+12a^2c^2+b^2\sqrt{b^4+24a^2c^2}\right)[(\lambda_1+\frac{s_+}{3})^2+(\lambda_2+\frac{s_+}{3})^2]+$$
$$+3a^2(\lambda_1+\frac{s_+}{3})(\lambda_2+\frac{s_+}{3})+R(\lambda_1,\lambda_2)\geq$$
$$\geq\frac{1}{2}\left\{\frac{1}{8c^2}\left(b^4+12a^2c^2+b^2\sqrt{b^4+24a^2c^2}\right)[(\lambda_1+\frac{s_+}{3})^2+(\lambda_2+\frac{s_+}{3})^2]\right\}$$
$$+\frac{3a^2}{2}(\lambda_1+\frac{s_+}{3})(\lambda_2+\frac{s_+}{3}) \tag{2.42}$$

where $R(\lambda_1,\lambda_2)$ is the remainder in the Taylor expansion, and the inequality holds *provided* that the remainder R is small enough. We choose $\varepsilon_1>0$ to be small enough so that if $(\lambda_1+\frac{s_+}{3})^2+(\lambda_2+\frac{s_+}{3})^2+(\lambda_1+\lambda_2+2\frac{s_+}{3})^2<\varepsilon_1$ then R is small enough and the inequality above holds.

As the quadratic form $\frac{1}{16c^2}\left(b^4+12a^2c^2+b^2\sqrt{b^4+24a^2c^2}\right)[(\lambda_1+\frac{s_+}{3})^2+(\lambda_2+\frac{s_+}{3})^2]+\frac{3}{2}a^2(\lambda_1+\frac{s_+}{3})(\lambda_2+\frac{s_+}{3})$ is positive definite, there exists a $C_2>0$, depending only on a^2,b^2 and c^2 such that

$$\frac{1}{2}\left\{\frac{1}{8c^2}\left(b^4+12a^2c^2+b^2\sqrt{b^4+24a^2c^2}\right)[(\lambda_1+\frac{s_+}{3})^2+(\lambda_2+\frac{s_+}{3})^2]\right\}$$
$$+\frac{3a^2}{2}(\lambda_1+\frac{s_+}{3})(\lambda_2+\frac{s_+}{3})$$
$$\geq C_2\left((\lambda_1+\frac{s_+}{3})^2+(\lambda_2+\frac{s_+}{3})^2+(\lambda_1+\lambda_2+\frac{2s_+}{3})^2\right)\ \forall(\lambda_1,\lambda_2)\in\mathbb{R}^2$$

Combining this last inequality with (2.42) we obtain the claim (2.41).

The relation (2.41) together with (2.40) and (2.37) show that $\mathrm{tr}(Q(x)-Q^x)^2\leq C_3\tilde{f}_B(Q(x))$ for some C_3 independent of L and x, which combined with (2.39) shows

$$-\partial_k\left(\frac{1}{L}\frac{\partial\tilde{f}_B}{\partial Q_{ij}}(Q(x)\right)Q_{ij,k}\leq\frac{\delta C_4}{L^2}\tilde{f}_B(Q(x))+\frac{1}{\delta}|\nabla Q(x)|^4$$

with C_4 a constant independent of L and x and any $\delta>0$. This last inequality together with (2.34) and (2.35) show:

$$-\Delta e_L+\frac{1}{L^2}\sum_{i,j=1}^{3}\left(\frac{\partial\tilde{f}_B}{\partial Q_{ij}}+\frac{b^2\delta_{ij}}{3}\,\mathrm{tr}(Q^2)\right)^2\leq\frac{\delta C_4}{L^2}\tilde{f}_B(Q)+\frac{1}{\delta}|\nabla Q|^4$$

Taking into account *Lemma* 2.6 and choosing δ small enough (depending only on C_4 and the constant $\tilde{C}$ from *Lemma* 2.6) we can absorb the term $\frac{\delta C_4}{L^2}\tilde{f}_B(Q)$ on the right hand side into the left hand side and obtain

$$-\Delta e_L\leq\frac{1}{\delta}|\nabla Q|^4,$$

giving the desired conclusion. □

LEMMA 2.8. *Let $\Omega\subset\mathbb{R}^3$ be a simply-connected bounded open set with smooth boundary. Let $Q^{(L_k)}\in W^{1,2}(\Omega,S_0)$ be a sequence of global minimizers for the energy $\tilde{F}_{LG}$, (1.6), in the admissible space (1.5). Assume that as $L_k\to 0$ we have $Q^{(L_k)}\to Q^{(0)}$ in $W^{1,2}(\Omega,S_0)$.*

Let $K\subset\Omega$ be a compact set which contains no singularity of $Q^{(0)}$. There exists $C_1>0, C_2>0, \bar{L}_0>0$(all constants independent of L_k) so that if for $a\in K, 0<r<d(a,\partial K)$ we have

$$\frac{1}{r}\int_{B_r(a)}e_{L_k}(Q^{(L_k)}(x))\,dx\leq C_1$$

then

$$r^2\sup_{B_{\frac{r}{2}}(a)}e_{L_k}(Q^{(L_k)})\leq C_2.$$

for all $L_k<\bar{L}_0$.

PROOF. Taking into account our assumptions on the sequence $\left(Q^{(L_k)}\right)_{k\in\mathbb{N}}$, *Proposition* 2.5 shows that for any given $\tilde{\varepsilon}_0$ smaller than ε_0 in *Lemma* 2.6 and also smaller than the ε_0 in *Lemma* 2.7, we have that there exists a $\bar{L}_0$ so that for $L_k < \bar{L}_0$ we have

$$\|Q^{(L_k)}(x) - s_+\left(n(x)\otimes n(x) - \frac{1}{3}Id\right)\| \leq \tilde{\varepsilon}_0,\ \forall x\in K, \text{ for some } n(x)\in\mathbb{S}^2 \quad (2.43)$$

We continue reasoning similarly as in [Sch84]. We fix an arbitrary $L_k < \bar{L}_0$ and an $a\in K$ and take a $r>0$ so that $0<r<d(a,\partial K)$. We let $r_1>0$ and $x_1\in B_{r_1}(a)$ be such that

$$\max_{0\leq s\leq\frac{2}{3}r}(\frac{2}{3}r-s)^2\max_{B_s(a)}e_{L_k}(Q^{(L_k)}) = (\frac{2}{3}r-r_1)^2\max_{B_{r_1}(a)}e_{L_k}(Q^{(L_k)})$$
$$= (\frac{2}{3}r-r_1)^2 e_{L_k}(Q^{(L_k)}(x_1)) \quad (2.44)$$

Define $e_1^{(L_k)} \stackrel{def}{=} \max_{B_{r_1}(a)} e_{L_k}(Q^{(L_k)})$. Then:

$$\max_{B_{\frac{2/3\cdot r-r_1}{2}}(x_1)} e_{L_k}(Q^{(L_k)}) \leq \max_{B_{\frac{2/3\cdot r+r_1}{2}}(a)} e_{L_k}(Q^{(L_k)})$$
$$\leq \frac{(2/3\cdot r-r_1)^2\max_{B_{r_1}(a)}e_{L_k}(Q^{(L_k)})}{(2/3\cdot r-(2/3\cdot r+r_1)/2)^2} = 4\max_{B_{r_1}(a)}e_{L_k}(Q^{(L_k)}) = 4e_1^{(L_k)} \quad (2.45)$$

where for the first inequality we use the fact that $B_{\frac{(2/3r-r_1)}{2}}(x_1)\subset B_{\frac{(2/3\cdot r+r_1)}{2}}(a)$ and for the second inequality, we use the definition of r_1.

Let $r_2 = \frac{(2/3\cdot r-r_1)\sqrt{e_1^{(L_k)}}}{2}$ and define $R^{(L_k)}(x) = Q^{(L_k)}\left(x_1 + \frac{x}{\sqrt{e_1^{(L_k)}}}\right)$. We let $\bar{L}_k = e_1^{(L_k)}L_k$ and then

$$e_{\bar{L}_k}(R^{(L_k)}) = \frac{1}{2}|\nabla R^{(L_k)}|^2 + \frac{\tilde{f}_B(R^{(L_k)})}{\bar{L}_k}$$
$$= \frac{1}{2}\frac{|\nabla Q^{(L_k)}|^2}{e_1^{(L_k)}} + \frac{\tilde{f}_B(Q^{(L_k)})}{e_1^{(L_k)}L_k} = \frac{1}{e_1^{(L_k)}}e_{L_k}(Q^{(L_k)})$$

Equation (2.45) then implies

$$\max_{x\in B_{r_2}(0)} e_{\bar{L}_k}(R^{(L_k)}) = \max_{x\in B_{\frac{(2/3r-r_1)}{2}}(x_1)}\frac{e_{L_k}(Q^{(L_k)}(x))}{e_1^{L_k}} \leq 4$$

where the equality above follows from the definitions of r_2 and $R^{(L_k)}$ and the inequality above follows from equation (2.45). Thus, we have

$$\max_{B_{r_2(0)}} e_{\bar{L}_k}(R^{(L_k)}) \leq 4,\ e_{\bar{L}_k}(R^{(L_k)})(0) = 1 \quad (2.46)$$

where $R^{(L_k)}$ satisfies the following system of elliptic PDEs

$$\bar{L}_k R^{(L_k)}_{ij,mm} = -a^2 R^{(L_k)}_{ij} - b^2\left(R^{(L_k)}_{in}R^{(L_k)}_{nj} - \frac{\delta_{ij}}{3}\operatorname{tr}((R^{(L_k)})^2)\right) + c^2 R^{(L_k)}_{ij}\operatorname{tr}((R^{(L_k)})^2). \tag{2.47}$$

We now claim that

$$r_2 \leq 1. \tag{2.48}$$

It is clear that $r_2 \leq 1$ implies the conclusion. Let us assume for contradiction that $r_2 > 1$. Then we claim that there exists a constant $C > 0$, independent of L_k, so that

$$1 \leq C\int_{B_1} e_{\bar{L}_k}(R^{(L_k)})(x)\,dx \tag{2.49}$$

The matrix $R^{(L_k)}$ satisfies the system (2.47) (which is the rescaled version of (1.7)); using relation (2.43) and the definition of $R^{(L_k)}$ as well as the fact that $r_2 > 1$, we can apply *Lemma* 2.7 to $e_{\bar{L}_k}(R^{(L_k)})$ and obtain

$$-\Delta e_{\bar{L}_k}(R^{(L_k)}(x)) \leq C e^2_{\bar{L}_k}(R^{(L_k)}(x)) \overset{(2.46)}{\leq} 4C e_{\bar{L}_k}(R^{(L_k)}(x)), \forall x \in B_1(0)$$

Combining (2.46) and the Harnack inequality (see for instance [Tay11], Ch.14, Thm. 9.3) along with the above relation we obtain (2.49).

We have

$$\begin{aligned}\int_{B_1} e_{\bar{L}_k}(R^{(L_k)}(x))\,dx &\leq \frac{1}{r_2}\int_{B_{r_2}(0)} \frac{|\nabla R^{(L_k)}(x)|^2}{2} + \frac{\tilde{f}_B(R^{(L_k)}(x))}{L_k e_1^{(L_k)}}\,dx = \\ = \frac{2}{2/3\cdot r - r_1}\int_{B_{(2/3\cdot r - r_1)/2}(x_1)} e_{L_k}(Q^{(L_k)}(x))\,dx &\leq \frac{3}{r}\int_{B_{r/3}(x_1)} e_{L_k}(Q^{(L_k)}(x))\,dx \\ &\leq \frac{3}{r}\int_{B_r(a)} e_{L_k}(Q^{(L_k)}(x))\,dx \leq 3C_1 \end{aligned} \tag{2.50}$$

where for the first inequality we use the monotonicity inequality (Lemma 2.2) and the assumption that $r_2 \geq 1$ (note that the equation satisfied by $R^{(L_k)}$, equation (2.47) is the same as the equation satisfied by $Q^{(L_k)}$, up to a different elastic constant, hence the use of Lemma 2.2 here is justified). For the equality in relation (2.50) we use the change of variables $y = x_1 + \frac{x}{\sqrt{e_1^{L_k}}}$ and use the relation: $e_{\bar{L}_k}(R^{(L_k)}) = \frac{1}{e_1^{(L_k)}} e_{L_k}(Q^{(L_k)})$. For the second inequality in (2.50) we use the monotonicity inequality and the fact that $\frac{2/3r - r_1}{2} \leq \frac{r}{3}$. For the third inequality in (2.50) we use the fact that $B_{r/3}(x_1) \subset B_r(a)$ since $|x_1 - a| < r_1 < \frac{2}{3}r$. The last step in (2.50) follows from the hypothesis of the Lemma.

Choosing C_1 small enough we reach a contradiction with (2.49) which in turn implies that $r_2 \leq 1$ and hence the conclusion. □

We can now prove the uniform convergence of $Q^{(L_k)}$ away from singularities of the limiting harmonic map $Q^{(0)}$:

PROPOSITION 2.9. *Let $\Omega \subset \mathbb{R}^3$ be a simply-connected bounded open set with smooth boundary. Let $Q^{(L_k)} \in W^{1,2}(\Omega, S_0)$ be a sequence of global minimizers for*

the energy $\tilde{F}_{LG}$, (1.6), in the admissible space (1.5). Assume that as $L_k \to 0$ we have $Q^{(L_k)} \to Q^{(0)}$ in $W^{1,2}(\Omega, S_0)$.

Let $K \subset \Omega$ be a compact set which contains no singularity of $Q^{(0)}$. Then

$$\lim_{k\to\infty} Q^{(L_k)}(x) = Q^{(0)}(x), \text{ uniformly for } x \in K. \tag{2.51}$$

PROOF. From the hypothesis and *Proposition* 2.5 we have that $\tilde{f}_B(Q^{(L_k)}) \to 0$ uniformly in K. Thus for any $\varepsilon_0 > 0$ there exists a $\bar{L}_0 > 0$ such that for $L_k < \bar{L}_0$ we have that $|Q^{(L_k)}(x) - s_+\left(n(x)\otimes n(x) - \frac{1}{3}Id\right)| < \varepsilon_0$ for all $x \in K$ (and for each $x \in K$, we have $n(x) \in \mathbb{S}^2$). Thus we can apply Lemmas 2.6, 2.7 and 2.8.

Recall that we have $Q^{(L_k)} \to Q^{(0)}$ in $W^{1,2}$. In order to show the uniform convergence it suffices to show that we have uniform (independent of L_k) Lipschitz bounds on $Q^{(L_k)}$ for $x \in K$. We reason similarly to the proof in Proposition 2.5 (see also [CL95]). We first claim that there exists an $\varepsilon_1 > 0$ so that

$$\begin{aligned} &\forall \varepsilon \in (0, \varepsilon_1), \text{ there exists } r_0(\varepsilon) \text{ depending only on } \varepsilon, \Omega, K, \\ &\qquad\qquad \text{and boundary data } Q_b \text{ so that} \\ &\frac{1}{r_0}\int_{K\cap B_{r_0}(x)} \frac{1}{2}|\nabla Q^{(L_k)}(x)|^2 + \frac{\tilde{f}_B(Q^{(L_k)}(x))}{L_k}\,dx \le \varepsilon\,, \forall x \in K, \\ &\qquad\qquad \text{provided that } L_k < L_*(\varepsilon, r_0(\varepsilon)) \end{aligned} \tag{2.52}$$

In order to prove the claim let us first recall that $Q^{(0)}$ has no singularities on the compact set K. Thus there exists a larger compact set K' with $K \subset K'$ and a constant $C > 0$ so that $|\nabla Q^{(0)}(x)| \le C, \forall x \in K'$. We choose $\varepsilon_1 > 0$ so that $B(x, \varepsilon_1) \cap K \subset K'$ hence for an arbitrary $\varepsilon \in (0, \varepsilon_1)$ there exists $r_0(\varepsilon) > 0$ so that

$$\frac{1}{r_0}\int_{K\cap B_{r_0}(x)} \frac{1}{2}|\nabla Q^{(0)}(x)|^2\,dx < \frac{\varepsilon}{3}$$

provided that $x \in K$ and $r_0(\varepsilon)$ is chosen small enough. We also have, from the $W^{1,2}(\Omega, S_0)$ convergence of $Q^{(L_k)}$ to $Q^{(0)}$, that there exists $\bar{L}(\varepsilon)$ so that

$$\frac{1}{r_0}\int_{K\cap B_{r_0}(x)} \frac{1}{2}|\nabla Q^{(L_k)}(x)|^2\,dx \le \frac{1}{r_0}\int_{K\cap B_{r_0}(x)} \frac{1}{2}|\nabla Q^{(0)}(x)|^2\,dx + \frac{\varepsilon}{3}, \forall L_k < \bar{L}(\varepsilon)$$

Recall from the proof of *Lemma* 2.3 that $\lim_{L_k\to 0}\int_\Omega \frac{\tilde{f}_B(Q^{(L_k)}(x))}{L_k}\,dx = 0$. Hence there exists $\tilde{L}(\varepsilon)$ so that $\frac{1}{r_0}\int_{K\cap B_{r_0}(x)} \frac{\tilde{f}_B(Q^{(L_k)}(x))}{L_k}\,dx < \frac{\varepsilon}{3}, \forall L < \tilde{L}(\varepsilon)$. Letting $L_*(\varepsilon, r_0(\varepsilon)) = \min\{\bar{L}, \tilde{L}\}$ and combining the two relations above we obtain the claim (2.52).

Choosing $\varepsilon > 0$ smaller than the constant C_1 from *Lemma* 2.8, we apply *Lemma* 2.8 to conclude that $|\nabla Q^{(L_k)}|$ can be bounded, independently of L_k, on the set K. The uniform convergence result now follows. □

3. Biaxiality and uniaxiality

3.1. The bulk energy density. Our first proposition concerns the stationary points of the bulk energy density.

PROPOSITION 3.1. *[BM10] Consider the bulk energy density $\tilde{f}_B(Q)$ given by*

$$\tilde{f}_B(Q) = -\frac{a^2}{2}\operatorname{tr} Q^2 - \frac{b^2}{3}\operatorname{tr} Q^3 + \frac{c^2}{4}\left(\operatorname{tr} Q^2\right)^2 + \frac{a^2}{3}s_+^2 + \frac{2b^2}{27}s_+^3 - \frac{c^2}{9}s_+^4. \tag{3.1}$$

Then $\tilde{f}_B(Q)$ attains its minimum for uniaxial Q-tensors of the form

$$Q = s_+\left(n \otimes n - \frac{1}{3}\right), \tag{3.2}$$

where

$$s_+ = \frac{b^2 + \sqrt{b^4 + 24a^2c^2}}{4c^2} \tag{3.3}$$

and $n : \Omega \to S^2$ is a unit eigenvector of Q.

PROOF. Proposition 3.1 was proved in [BM10] and we reproduce the proof in the *Appendix* for completeness. □

In the following proposition, we estimate $\tilde{f}_B(Q)$ in terms of $|Q|$ and the biaxiality parameter $\beta(Q)$.

PROPOSITION 3.2. *Let $Q \in S_0$. Then the bulk energy density $\tilde{f}_B(Q)$ is bounded from below by*

$$\tilde{f}_B(Q) \geq \Gamma(a^2, b^2, c^2)\left(|Q| - \sqrt{\frac{2}{3}}s_+\right)^2 + \frac{b^2}{6\sqrt{6}}\beta(Q)|Q|^3 \tag{3.4}$$

where

$$\Gamma(a^2, b^2, c^2) = \frac{b^2}{216c^2}\left\{3\sqrt{b^4 + 24a^2c^2} - b^2\right\} \tag{3.5}$$

and s_+ has been defined in (3.3).

PROOF. From Lemma 0.5, we have the inequality,

$$\operatorname{tr} Q^3 = |Q|^3\sqrt{\left(\frac{1-\beta}{6}\right)} \leq \frac{|Q|^3}{\sqrt{6}}\left(1 - \frac{\beta}{2}\right) \quad for\ Q \in S_0.$$

From the definition of $\tilde{f}_B(Q)$ and s_+ in (3.1) and (3.3), we can obtain a lower bound for $\tilde{f}_B(Q)$ in terms of $|Q|$ and $\beta(Q)$ as follows i.e.

$$\begin{aligned}\tilde{f}_B(Q) = -\frac{a^2}{2}|Q|^2 - \frac{b^2}{3\sqrt{6}}|Q|^3\sqrt{1-\beta} + \frac{c^2}{4}|Q|^4 + \frac{a^2}{2}\left(\sqrt{\frac{2}{3}}s_+\right)^2 \\ + \frac{b^2}{3}\frac{2s_+^3}{9} - \frac{c^2}{4}\left(\sqrt{\frac{2}{3}}s_+\right)^4 \end{aligned} \tag{3.6}$$

$$\geq \left[-\frac{a^2}{2}|Q|^2 - \frac{b^2}{3\sqrt{6}}|Q|^3 + \frac{c^2}{4}|Q|^4 + \frac{a^2}{3}s_+^2 + \frac{2b^2}{27}s_+^3 - \frac{c^2}{9}s_+^4\right] + \frac{b^2}{6\sqrt{6}}\beta(Q)|Q|^3. \tag{3.7}$$

The bracketed term in (3.7) can be further simplified by carrying out a series of calculations. Let $\delta = |Q| - \sqrt{\frac{2}{3}}s_+$ where $\frac{2}{3}c^2 s_+^2 = \frac{b^2}{\sqrt{6}}\sqrt{\frac{2}{3}}s_+ + a^2$ by the definition of s_+. Then

$$\left[-\frac{a^2}{2}|Q|^2 - \frac{b^2}{3\sqrt{6}}|Q|^3 + \frac{c^2}{4}|Q|^4 + \frac{a^2}{3}s_+^2 + \frac{2b^2}{27}s_+^3 - \frac{c^2}{9}s_+^4\right] = \tag{3.8}$$

$$= \delta\left[-a^2\sqrt{\frac{2}{3}}s_+ - \frac{\sqrt{2}b^2}{3\sqrt{3}}s_+^2 + \frac{2\sqrt{2}}{3\sqrt{3}}c^2 s_+^3\right] + \delta^2\left[-\frac{a^2}{2} - \frac{b^2}{3}s_+ + s_+^2 c^2\right] +$$

$$+\delta^3\left[c^2\sqrt{\frac{2}{3}}s_+ - \frac{b^2}{3\sqrt{6}}\right] + \frac{c^2}{4}\delta^4. \tag{3.9}$$

The coefficient of δ vanishes from the definition of s_+. Therefore, we have that

$$\left[-\frac{a^2}{2}|Q|^2 - \frac{b^2}{3\sqrt{6}}|Q|^3 + \frac{c^2}{4}|Q|^4 + \frac{a^2}{3}s_+^2 + \frac{2b^2}{27}s_+^3 - \frac{c^2}{9}s_+^4\right] = \tag{3.10}$$

$$= \delta^2\left(\alpha + \beta\delta + \gamma\delta^2\right) \text{ where}$$

$$\alpha = \left[-\frac{a^2}{2} - \frac{b^2}{3}s_+ + s_+^2 c^2\right] = \frac{a^2}{2} + \frac{c^2 s_+^2}{3}$$

$$\beta = \left[c^2\sqrt{\frac{2}{3}}s_+ - \frac{b^2}{3\sqrt{6}}\right] \geq \frac{2b^2}{3\sqrt{6}} > 0 \text{ and}$$

$$\gamma = \frac{c^2}{4}. \tag{3.11}$$

It is readily verified that the function

$$f(\delta) = \left(\alpha + \beta\delta + \gamma\delta^2\right)$$

attains its minimum value at $\delta_{\min} = -\frac{\beta}{2\gamma}$ and the minimum value is

$$f(\delta_{\min}) = \frac{b^2}{216c^2}\left\{3\sqrt{b^4 + 24a^2c^2} - b^2\right\}. \tag{3.12}$$

Substituting (3.12) into (3.10), we have that

$$\left[-\frac{a^2}{2}|Q|^2 - \frac{b^2}{3\sqrt{6}}|Q|^3 + \frac{c^2}{4}|Q|^4 + \frac{a^2}{3}s_+^2 + \frac{2b^2}{27}s_+^3 - \frac{c^2}{9}s_+^4\right] \geq \Gamma(a^2, b^2, c^2)\delta^2 \tag{3.13}$$

where Γ has been defined in (3.5) and on combining (3.13) with (3.7), the lower bound (3.4) follows. □

We can obtain explicit lower bounds for $\tilde{f}_B$ in terms of the order parameters s and r in Proposition 0.1, as shown below.

PROPOSITION 3.3. *Let $Q \in S_0$ be represented as in Proposition 0.1*

$$Q = s\left(n \otimes n - \frac{1}{3}Id\right) + r\left(m \otimes m - \frac{1}{3}Id\right)$$

with either $0 \leq r \leq \frac{s}{2}$ or $\frac{s}{2} \leq r \leq 0$. Case (i) Non-negative order parameters, $0 \leq r \leq \frac{s}{2}$ with $0 \leq s \leq s_+$, where s_+ is defined in (3.3). Then the bulk energy

density, $\tilde{f}_B(Q)$, is bounded from below by

$$\tilde{f}_B(Q) \geq (s_+ - s)^2 \gamma(a^2, b^2, c^2) + \frac{r(s-r)}{9}\left(3a^2 + b^2 s - 2c^2 s^2\right) + \frac{5b^2}{27} r^2 s \geq 0 \qquad 0 \leq s \leq s_+ \quad (3.14)$$

where $\gamma\left(a^2, b^2, c^2\right)$ is an explicitly computable positive constant.

Case (ii) Non-negative order parameters, $0 \leq r \leq \frac{s}{2}$ and $s \geq s_+$. Then

$$\tilde{f}_B(Q) \geq \frac{b^2}{216c^2}\left\{3(b^4 + 24a^2c^2)^{1/2} - b^2\right\} \min\left\{\frac{2}{3}(s - s_+)^2,\ \frac{1}{6}\left(\sqrt{3}s - 2s_+\right)^2\right\} + \tau b^2 s_+^3 \left(\frac{r^2(s-r)^2}{s^4}\right) \quad (3.15)$$

where τ is an explicitly computable positive constant, independent of a^2, b^2, c^2.

Case (iii) If $\frac{s}{2} \leq r \leq 0$, then

$$\tilde{f}_B(Q) = \tilde{f}_B(-Q) + \frac{2b^2}{27}\left(2|s|^3 + 2|r|^3 - 3s^2|r| - 3|s|r^2\right), \quad (3.16)$$

where $-Q \in S_0$ has positive order parameters $0 \leq -r \leq -\frac{s}{2}$ and $\tilde{f}_B(-Q)$ can be estimated using (3.14) and (3.15). In particular,

$$\tilde{f}_B(Q) \geq -\frac{a^4}{4c^2} - \frac{s_+^3}{3}\left(\frac{b^2}{9} - \frac{c^2}{3}s_+\right) > 0 \quad (3.17)$$

for Q-tensors with $\frac{s}{2} \leq r \leq 0$.

Proof. From Proposition 0.1, it suffices to consider the two cases $0 \leq r \leq \frac{s}{2}$ and $\frac{s}{2} \leq r \leq 0$.

Case (i): We can explicitly express the bulk energy density, $\tilde{f}_B(Q)$, in terms of s and r as follows –

$$\tilde{f}_B(Q) = -\frac{a^2}{3}\left(s^2 + r^2 - sr\right) - \frac{b^2}{27}\left(2s^3 + 2r^3 - 3s^2 r - 3sr^2\right) + \frac{c^2}{9}\left(s^4 + r^4 + 3s^2r^2 - 2sr^3 - 2s^3 r\right) + \frac{a^2}{3}s_+^2 + \frac{2b^2}{27}s_+^3 - \frac{c^2}{9}s_+^4, \quad (3.18)$$

where we have expressed $\operatorname{tr} Q^2$ and $\operatorname{tr} Q^3$ in terms of s and r

$$\operatorname{tr} Q^2 = \frac{2}{3}\left(s^2 + r^2 - sr\right)$$

and

$$\operatorname{tr} Q^3 = \frac{1}{9}\left(2s^3 + 2r^3 - 3s^2 r - 3sr^2\right).$$

The function $\tilde{f}_B(Q)$ consists of two components – $\tilde{f}_B(Q) = F(s) + G(s, r)$, where

$$F(s) = -\frac{a^2}{3}\left(s^2 - s_+^2\right) - \frac{2b^2}{27}\left(s^3 - s_+^3\right) + \frac{c^2}{9}\left(s^4 - s_+^4\right)$$
$$G(s, r) = \frac{a^2}{3}\left(sr - r^2\right) + \frac{b^2}{27}\left(3s^2 r + 3sr^2 - 2r^3\right) + \frac{c^2}{9}\left(-2s^3 r + 3s^2 r^2 - 2sr^3 + r^4\right). \quad (3.19)$$

Recalling that $2c^2s_+^2 = b^2s_+ + 3a^2$ (from the definition of s_+ in (3.3)), the function $F(s)$ can be expressed in terms of $\delta = s_+ - s \geq 0$ as follows –

$$F(s) = \frac{\delta}{27}\left(18a^2s_+ + 6b^2s_+^2 - 12c^2s_+^3\right) + \delta^2\left(\frac{3b^2}{27}s_+ + \frac{18a^2}{27} + \delta\left(\frac{2b^2}{27} - \frac{4c^2}{9}s_+ + \frac{c^2}{9}\delta\right)\right). \quad (3.20)$$

The coefficient of δ vanishes by virtue of the definition of s_+ in (3.3). We note that the function

$$\bar{G}(\delta) \stackrel{def}{=} \delta\left(\frac{2b^2}{27} - \frac{4c^2}{9}s_+ + \frac{c^2}{9}\delta\right) \quad (3.21)$$

attains a minimum for

$$\delta_{\min} = 2s_+ - \frac{b^2}{3c^2} > s_+ \quad (3.22)$$

and, since $\bar{G}$ is non-increasing on $[\delta, s_+]$:

$$\bar{G}(\delta) \geq \bar{G}(s_+) = \frac{1}{27}\left(2b^2s_+ - 9c^2s_+^2\right). \quad (3.23)$$

We substitute (3.23) into (3.20) to obtain the following lower bound for $F(s)$ -

$$F(s) \geq \frac{c^2s_+^2 + 3a^2}{27}(s_+ - s)^2. \quad (3.24)$$

We can analyze the function $G(s,r)$, in (3.19), in an analogous manner. Let $\gamma = \frac{r}{s} \in \left[0, \frac{1}{2}\right]$. Then

$$G(s,r) = \gamma s^2\left[\frac{a^2}{3} + \frac{3b^2}{27}s - \frac{2c^2}{9}s^2\right] + \gamma^2 s^2\left[-\frac{a^2}{3} + \frac{3b^2}{27}s + \frac{3c^2}{9}s^2\right] + \gamma^3 s^3\left[-\frac{2b^2}{27} - \frac{2c^2 s}{9} + \gamma\frac{c^2 s}{9}\right]. \quad (3.25)$$

The coefficient of γ is non-negative for all $s \leq s_+$. Using the inequality $\gamma \leq \frac{1}{2}$, one readily obtains the following lower bound for $G(s,r)$ -

$$G(s,r) \geq \gamma s^2\left[\frac{a^2}{3} + \frac{3b^2}{27}s - \frac{2c^2}{9}s^2\right] + \gamma^2 s^2\left[-\frac{a^2}{3} + \frac{2b^2}{27}s + \frac{2c^2}{9}s^2\right] \geq \geq \frac{r(s-r)}{9}\left(3a^2 + b^2s - 2c^2s^2\right) + \frac{5b^2}{27}r^2 s. \quad (3.26)$$

Combining (3.24) and (3.26), the lower bound for $0 \leq s \leq s_+$ in (3.14) follows.

Case (ii) The case $s \geq s_+$ can be dealt with similarly. For any $Q \in S_0$ with $0 \leq r \leq \frac{s}{2}$, we have that

$$\frac{s}{\sqrt{2}} \leq |Q| = \sqrt{\frac{2}{3}}\sqrt{(s^2 + r^2 - sr)} \leq \sqrt{\frac{2}{3}}s. \quad (3.27)$$

For $s \geq s_+$, $|Q|^3 \geq \frac{s_+^3}{2\sqrt{2}}$ and

$$\beta(Q) \geq \eta\left(\frac{r^2(s-r)^2}{s^4}\right) \quad (3.28)$$

where $\beta(Q)$ is the biaxiality parameter defined in (0.9) and η is a positive constant independent of a^2, b^2 or c^2 or L. Combining (3.27), (3.28) and (3.4), we readily obtain the lower bound

$$\begin{aligned}\tilde{f}_B(Q) \geq \frac{b^2}{216c^2}\left\{3(b^4+24a^2c^2)^{1/2}-b^2\right\}\left(|Q|-\sqrt{\frac{2}{3}}s_+\right)^2+\frac{b^2}{6\sqrt{6}}\beta(Q)|Q|^3 \geq \\ \geq \frac{b^2}{216c^2}\left\{3(b^4+24a^2c^2)^{1/2}-b^2\right\}\min\left\{\frac{2}{3}\left(s-s_+\right)^2,\ \frac{1}{6}\left(\sqrt{3}s-2s_+\right)^2\right\} \\ +\tau b^2 s_+^3\left(\frac{r^2(s-r)^2}{s^4}\right)\end{aligned} \tag{3.29}$$

where τ is an explicitly computable positive constant.

Case (iii) Finally, we consider $Q \in S_0$ with negative order parameters $\frac{s}{2} \leq r \leq 0$. In this case, one can directly check that

$$\operatorname{tr} Q^3 = \frac{1}{9}\left(2s^3+2r^3-3s^2r-3sr^2\right) \leq 0$$

and therefore,

$$\begin{aligned}\tilde{f}_B(Q) &= -\frac{a^2}{2}|Q|^2-\frac{b^2}{3}\operatorname{tr} Q^3+\frac{c^2}{4}|Q|^4+\frac{a^2}{3}s_+^2+\frac{2b^2}{27}s_+^3-\frac{c^2}{9}s_+^4 \\ &= -\frac{a^2}{2}|Q|^2-\frac{b^2}{3}\operatorname{tr}(-Q)^3+\frac{c^2}{4}|Q|^4+\frac{a^2}{3}s_+^2+\frac{2b^2}{27}s_+^3-\frac{c^2}{9}s_+^4+\frac{2b^2}{3}\left|\operatorname{tr} Q^3\right|,\end{aligned} \tag{3.30}$$

since $\frac{b^2}{3}\operatorname{tr}(-Q)^3 = -\frac{b^2}{3}\operatorname{tr} Q^3$ and $-\frac{b^2}{3}\operatorname{tr} Q^3 = \frac{b^2}{3}\left|\operatorname{tr} Q^3\right|$. The equality (3.16) follows from (3.30) upon expressing $\operatorname{tr} Q^3$ in terms of s and r.

For (3.17), it suffices to note that for $s, r \leq 0$, $\operatorname{tr} Q^3 \leq 0$ and therefore,

$$\begin{aligned}\tilde{f}_B(Q) &\geq -\frac{a^2}{2}|Q|^2+\frac{c^2}{4}|Q|^4+\frac{a^2}{3}s_+^2+\frac{2b^2}{27}s_+^3-\frac{c^2}{9}s_+^4 \\ &= -\frac{a^2}{3}\left(s^2+r^2-sr\right)+\frac{c^2}{9}\left(s^2+r^2-sr\right)^2-\frac{s_+^3}{3}\left(\frac{b^2}{9}-\frac{c^2}{3}s_+\right).\end{aligned} \tag{3.31}$$

A straightforward computation shows that the function

$$-\frac{a^2}{3}\left(s^2+r^2-sr\right)+\frac{c^2}{9}\left(s^2+r^2-sr\right)^2 \geq -\frac{a^4}{4c^2}$$

and

$$\frac{s_+^3}{3}\left(\frac{b^2}{9}-\frac{c^2}{3}s_+\right) < -\frac{a^4}{4c^2}.$$

The inequality (3.17) now follows. □

REMARK 3.4. *One can readily obtain lower bounds for $\tilde{f}_B(Q)$ in terms of the order parameters $(S,\ R)$ in Proposition 0.3, following the methods outlined in Proposition 3.3. The details are omitted here for brevity.*

REMARK 3.5. *Relation (3.17) shows that if* $f_B(Q^{(L_k)}(x)) \to 0$ *as* $L_k \to 0$ *then* $Q^{(L_k)}(x)$ *cannot have an* (s,r) *representation with* $\frac{s}{2} < r < 0$*, if* L_k *is sufficiently small.*

In view of Propositions 2.1 and 3.2, we can make qualitative predictions about the size of regions where a global Landau–De Gennes minimizer Q^* can have $|Q^*| << \sqrt{\frac{2}{3}}s_+$ and the size of regions where Q^* can be strongly biaxial.

PROPOSITION 3.6. *Let* Q^* *be a global minimizer of* $\tilde{F}_{LG}$*, (1.6), in the space (1.5). Let* $\Omega^* = \left\{x \in \Omega;\ |Q^*(x)| \leq \frac{1}{2}\sqrt{\frac{2}{3}}s_+\right\}$*. Then*

$$|\Omega^*| \leq \alpha \frac{L}{\Gamma(a^2,b^2,c^2)} \int_\Omega |\nabla n^{(0)}(x)|^2\ dx, \tag{3.32}$$

where $n^{(0)}$ *is defined in (1.9) and* α *is an explicitly computable positive constant independent of* a^2, b^2, c^2 *or* L*. The constant* $\Gamma(a^2,b^2,c^2)$ *is defined in (3.5).*

PROOF. From Proposition 3.2, we have that

$$\tilde{f}_B(Q^*(x)) \geq \frac{1}{\alpha}\Gamma s_+^2, \quad x \in \Omega^* \tag{3.33}$$

for some explicitly computable positive constant α, since $|Q^*| \leq \frac{1}{2}\sqrt{\frac{2}{3}}s_+ = \frac{1}{\sqrt{6}}s_+$ on Ω^*. On the other hand, recalling the definition of $Q^{(0)}$ in (1.8) and since Q^* is a global minimizer of $\tilde{F}_{LG}[Q]$, we have that

$$\int_{\Omega^*} \tilde{f}_B(Q^*(x))\,dx \leq \tilde{F}_{LG}[Q^{(0)}] = \int_\Omega \frac{L}{2}|\nabla Q^{(0)}|^2 + \tilde{f}_B(Q^{(0)})\,dx = Ls_+^2 \int_\Omega |\nabla n^{(0)}|^2\,dx, \tag{3.34}$$

since $\tilde{f}_B(Q^{(0)}) = 0$ everywhere in Ω. Substituting (3.33) into (3.34), we obtain

$$\frac{1}{\alpha}\Gamma(a^2,b^2,c^2)s_+^2|\Omega^*| \leq Ls_+^2 \int_\Omega |\nabla n^{(0)}|^2\ dx, \tag{3.35}$$

from which the inequality (3.32) follows. □

PROPOSITION 3.7. *Let* Q^* *be a global minimizer of* $\tilde{F}_{LG}$*, (1.6), in the space (1.5). Let*

$$\Omega^\lambda = \left\{x \in \Omega;\ |Q^*(x)| \geq \frac{1}{2}\sqrt{\frac{2}{3}}s_+,\ \beta(Q(x)) > \lambda\right\}$$

for some positive constant λ*. Then,*

$$|\Omega^\lambda| \leq \alpha \frac{L}{\lambda s_+ b^2} \int_\Omega |\nabla n^{(0)}|^2\,dx \tag{3.36}$$

where $n^{(0)}$ *is defined in (1.9) and* α *is an explicitly computable positive constant independent of* a^2, b^2, c^2 *or* L*.*

PROOF. From Proposition 3.2, we have that

$$\tilde{f}_B(Q^*(x)) \geq \frac{b^2}{6\sqrt{6}}\beta(Q^*(x))|Q^*(x)|^3 \geq \frac{1}{\alpha}b^2\lambda s_+^3 \quad x \in \Omega^\lambda \tag{3.37}$$

for some explicitly computable positive constant α, since $|Q^*| \geq \frac{1}{2}\sqrt{\frac{2}{3}}s_+ = \frac{1}{\sqrt{6}}s_+$ on Ω^λ. On the other hand, recalling the definition of $Q^{(0)}$, (1.8), and since Q^* is a global minimizer of $\tilde{F}_{LG}[Q]$, we have that

$$\int_{\Omega^\lambda} \tilde{f}_B(Q^*(x))\, dx \leq \int_\Omega \frac{L}{2}|\nabla Q^{(0)}|^2 + \tilde{f}_B(Q^{(0)})\, dx = Ls_+^2 \int_\Omega |\nabla n^{(0)}|^2\, dx \quad (3.38)$$

since $\tilde{f}_B(Q^{(0)}) = 0$ everywhere in Ω. Substituting (3.37) into (3.38), we obtain

$$\frac{1}{\alpha} b^2 \lambda s_+^3 |\Omega^\lambda| \leq Ls_+^2 \int_\Omega |\nabla n^{(0)}|^2\, dx, \quad (3.39)$$

from which the inequality (3.36) follows. □

Proposition 3.6 is relevant to the size of defect cores in global energy minimizers whereas Proposition 3.7 is relevant to the equilibrium behaviour far away from the defect cores.

3.2. Analyticity and uniaxiality. We define a new biaxiality parameter $\tilde{\beta}(Q)$ as follows:

$$\tilde{\beta}(Q) \stackrel{def}{=} (\mathrm{tr}(Q^2))^3 - 6(\mathrm{tr}(Q^3))^2.$$

Then $\tilde{\beta}(Q) \geq 0$ with $\tilde{\beta}(Q) = 0$ if and only if Q is uniaxial i.e. $Q = s\left(n \otimes n - \frac{1}{3}Id\right)$ for some $s \in \mathbb{R}\setminus\{0\}, n \in \mathbb{S}^2$ or $Q = 0$. The function $\tilde{\beta}(Q)$ is a real analytic function of Q and this is particularly important given that global energy minimizers of $\tilde{F}_{LG}$ in the admissible space $\mathcal{A}_Q$ are real analytic:

PROPOSITION 3.8. *Let Ω be a simply-connected bounded open set with smooth boundary. Let $Q^{(L)}$ be a global minimizer of $\tilde{F}_{LG}$, (1.6), in the space (1.5). This global minimizer $Q^{(L)}$ is a solution of (1.7) and is real analytic in Ω.*

PROOF. We drop the superscript L from $Q^{(L)}$ for convenience. As $-\frac{a^2}{2}\,\mathrm{tr}\left(Q^2\right) - \frac{b^2}{3}\,\mathrm{tr}\left(Q^3\right) + \frac{c^2}{4}\left(\mathrm{tr}\,Q^2\right)^2$ is bounded from below (see also the *Appendix*) we have that there exists an H^1-global energy minimizer. This is a weak solution of the Euler-Lagrange system:

$$L\Delta Q_{ij} = -a^2 Q_{ij} - b^2\left(Q_{ik}Q_{kj} - \frac{\delta_{ij}}{3}\,\mathrm{tr}(Q^2)\right) + c^2 Q_{ij}\,\mathrm{tr}(Q^2)$$

For Q an H^1 solution of the equation one uses $H^1 \hookrightarrow L^6$ (in $\mathbb{R}^3$) and Hölder's inequality to obtain that the right hand side of each equation is in L^2. Elliptic regularity gives that $Q \in H^2 \hookrightarrow W^{1,6} \hookrightarrow L^\infty$ hence the right hand side of each equation is in H^1. Elliptic regularity gives $Q \in H^3$ and one can continue bootstrapping to obtain the C^∞ regularity.

We obtain the analyticity by observing that the nonlinearity is polynomial and then standard results (see for instance [Fri58], p.45) imply that that the solutions are real analytic in Ω. □

PROPOSITION 3.9. *Let Q be a real analytic function $Q : \Omega \subset \mathbb{R}^3 \to S_0$. Then the set where Q is uniaxial or isotropic is either Ω itself or has zero Lebesgue measure.*

PROOF. If there is no $x \in \Omega$ such that $\tilde{\beta}(Q(x)) \neq 0$ then Q is uniaxial or isotropic everywhere. If there exists a $P \in \Omega$ such that $\tilde{\beta}(Q(P)) \neq 0$ then let us consider the lines passing through P. The restriction of Q to any such line is real analytic and then so is $\tilde{\beta}(Q)$. Thus $\tilde{\beta}(Q)$ has at most countably many zeroes on such a line. We claim that this implies that the set of zeroes of $\tilde{\beta}(Q)$ in Ω is of measure zero.

We assume, without loss of generality, that $P = 0$. We denote $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$ and decompose $\Omega = \cup_{n\in\mathbb{N}^*} \left(\overline{B_{\frac{1}{n}} \setminus B_{\frac{1}{n+1}}} \cap \Omega\right) \cup \left(\cup_{n\in\mathbb{N}^*} \left(\overline{B_{n+1} \setminus B_n} \cap \Omega\right)\right) \cup \{0\}$. We claim that for any $n \in \mathbb{N}^*$ the set $\left(\tilde{\beta}(Q)\right)^{-1}(0) \cap \left(\Omega \cap \overline{B_{\frac{1}{n}} \setminus B_{\frac{1}{n+1}}}\right)$ is a set of measure zero. This implies that $\tilde{\beta}(Q)^{-1}(0) \cap \Omega$, which is a countable union of sets as before, is also a set of measure zero.

We consider the bi-Lipschitz functions

$$f_n : [\frac{1}{n+1}, \frac{1}{n}] \times [0, \pi] \times [0, 2\pi) \to \overline{B_{\frac{1}{n}} \setminus B_{\frac{1}{n+1}}}, \forall n \in \mathbb{N}$$

that realize the change of coordinates from polar to usual cartesian coordinates.

We have that $f_n^{-1}\left(\tilde{\beta}(Q)^{-1}(0) \cap \Omega \cap \overline{B_{\frac{1}{n}} \setminus B_{\frac{1}{n+1}}}\right) \subset [\frac{1}{n+1}, \frac{1}{n}] \times [0, \pi] \times [0, 2\pi)$. We recall that the Lebesgue measure μ on the 3-dimensional product space $[\frac{1}{n+1}, \frac{1}{n}] \times [0, \pi] \times [0, 2\pi)$ is the completion of the product measure $\mu_1 \times \mu_2$ where μ_1 is the 1-dimensional Lebesgue measure on $[\frac{1}{n+1}, \frac{1}{n}]$ and μ_2 is the 2-dimensional Lebesgue measure on $[0, \pi] \times [0, 2\pi)$. Then for any set $E \subset [\frac{1}{n+1}, \frac{1}{n}] \times [0, \pi] \times [0, 2\pi)$ we have

$$(\mu_1 \times \mu_2)(E) = \int_{[0,\pi]\times[0,2\pi)} \mu_1(E^y)\mu_2(dy)$$

where $E^y = \{x \in [\frac{1}{n+1}, \frac{1}{n}], (x, y_1, y_2) \in E\} \subset [\frac{1}{n+1}, \frac{1}{n}]$. In our case, letting

$$E \stackrel{def}{=} f_n^{-1}\left(\tilde{\beta}(Q)^{-1}(0) \cap \Omega \cap \overline{B_{\frac{1}{n}} \setminus B_{\frac{1}{n+1}}}\right)$$

we have that E^y is made of finitely many points for almost all $y \in [0, \pi] \times [0, 2\pi)$(this is a consequence of the first paragraph in this proof. Indeed, on the segment through P that is in $\Omega \cap \overline{B_{\frac{1}{n}} \setminus B_{\frac{1}{n+1}}}$ and has direction given in polar coordinates by $y \in [0, \pi] \times [0, 2\pi)$ there are only finitely many points that are isotropic or uniaxial. The set E^y is just the set of the distances to P of the uniaxial or isotropic points that are on such a segment). Thus $\mu_1(E^y) = 0, \mu_2 - a.e.\, y$ hence $\mu_1 \times \mu_2(E) = 0$ thus $\mu(E) = 0$.

As bi-Lipschitz functions carry sets of measure zero into sets of measure zero we have that $\tilde{\beta}(Q)^{-1}(0) \cap \Omega \cap \overline{B_{\frac{1}{n}} \setminus B_{\frac{1}{n+1}}}$ is a set of measure zero for all $n \in \mathbb{N}$. A similar argument holds for the sets $\tilde{\beta}^{-1}(0) \cap \Omega \cap \overline{B_{n+1} \setminus B_n}$ for all $n \in \mathbb{N}$. On the other hand $\tilde{\beta}(Q)^{-1}(0) \cap \Omega$ is a countable union of sets as before, hence it has measure zero. □

COROLLARY 3.10. *Let $Q^{(L)}$ be a global minimizer of $\tilde{F}_{LG}$, (1.6), in the space (1.5). Then there exists a set of Lebesgue measure zero, possibly empty, Ω_0 in Ω such that the eigenvectors of $Q^{(L)}$ are smooth at all points $x \in \Omega \setminus \Omega_0$. The*

uniaxial-biaxial, isotropic-uniaxial or isotropic-biaxial interfaces are contained in Ω_0.

PROOF. The global minimizer $Q^{(L)} \in C^\infty(\Omega; S_0)$. The eigenvectors of $Q^{(L)}$ have the same degree of regularity as $Q^{(L)}$ on sets $K \subset \Omega$, where $Q^{(L)}$ has a constant number of distinct eigenvalues i.e. where $Q^{(L)}$ is either biaxial or uniaxial or isotropic, [Nom73], but not necessarily otherwise [Kat76]. If $Q^{(L)}$ is uniaxial everywhere then $\Omega_0 = \emptyset$. If $Q^{(L)}$ is either uniaxial or isotropic on the whole of Ω (i.e. $\tilde{\beta}(Q^{(L)}) = 0$ in Ω), with $Q^{(L)} \neq 0$ at some point in Ω, then let $\tilde{\Omega} = \{x \in \Omega, Q^{(L)}_{ij} = 0, i, j = 1, 2, 3\}$ denote the zero-set of $Q^{(L)}$. Let us observe that $\tilde{\Omega} = \left(|Q|^2\right)^{-1}(0)$ and $|Q|^2$ is an analytic function. By an argument similar to the proof of Proposition 3.9 and since $Q(x) \neq 0$ for at least one point $x \in \Omega$, we have that $\tilde{\Omega}$ has measure zero and we take $\Omega_0 \stackrel{def}{=} \tilde{\Omega}$.

If $Q^{(L)}$ is biaxial somewhere then Proposition 3.9 shows that the set of points where $\tilde{\beta}(Q) = 0$ has measure zero. We denote this set by Ω_0 and observe that $\Omega \backslash \Omega_0$ is an open set and the eigenvectors have the same regularity as $Q^{(L)}$ on $\Omega \setminus \Omega_0$, see [Nom73]. □

CHAPTER 3

Well-posedness of a dynamical model: Q-tensors and Navier–Stokes

In this chapter we consider the Beris and Edwards model as described in Section 4 but restrict ourselves to the case $\xi = 0$. This means that the molecules are such that they only tumble in a shear flow, but are not aligned by such a flow. In this case the system (4.3) reduces to:

$$\begin{cases} (\partial_t + u_\gamma \cdot \partial_\gamma) Q_{\alpha\beta} - \Omega_{\alpha\gamma} Q_{\gamma\beta} + Q_{\alpha\gamma}\Omega_{\gamma\beta} \\ = \Gamma\Big(L\Delta Q_{\alpha\beta} - aQ_{\alpha\beta} + b[Q_{\alpha\gamma}Q_{\gamma\beta} - \frac{\delta_{\alpha\beta}}{d}\operatorname{tr}(Q^2)] - cQ_{\alpha\beta}\operatorname{tr}(Q^2)\Big) \\ \partial_t u_\alpha + u_\beta \partial_\beta u_\alpha = \nu\Delta u_\alpha + \partial_\alpha p - L\partial_\beta \left(\partial_\alpha Q_{\zeta\delta}\partial_\beta Q_{\zeta\delta} - \frac{\delta_{\alpha\beta}}{d}\partial_\lambda Q_{\zeta\delta}\partial_\lambda Q_{\zeta\delta}\right) \\ + L\partial_\beta \left(Q_{\alpha\gamma}\Delta Q_{\gamma\beta} - \Delta Q_{\alpha\gamma} Q_{\gamma\beta}\right) \\ \partial_\gamma u_\gamma = 0 \end{cases} \tag{0.1}$$

in $\mathbb{R}^d$, $d = 2, 3$.

We also need to assume from now on that

$$c > 0 \tag{0.2}$$

This assumption is necessary from a modelling point of view (see [Maj10],[MZ10]) so that the energy $\mathcal{F}$ (see next section, relation (1.1)) is bounded from below, and it is also necessary for having global solutions (see Proposition 1.2 and its proof). The presentation follows ideas from [PZ].

1. The dissipation and apriori estimates

Let us denote the free energy of the director fields:

$$\mathcal{F}(Q) = \int_{\mathbb{R}^d} \frac{L}{2}|\nabla Q|^2 + \frac{a}{2}\operatorname{tr}(Q^2) - \frac{b}{3}\operatorname{tr}(Q^3) + \frac{c}{4}\operatorname{tr}^2(Q^2)\, dx \tag{1.1}$$

In the absence of the flow, when $u = 0$ in the equations (0.1), the free energy is a Lyapunov functional of the system. If $u \neq 0$ we still have a Lyapunov functional for (0.1) but this time one that includes the kinetic energy of the system. More precisely we have:

Proposition 1.1. *The system (0.1) has a Lyapunov functional:*

$$E(t) \overset{def}{=} \frac{1}{2}\int_{\mathbb{R}^d} |u|^2(t,x)\, dx + \int_{\mathbb{R}^d} \frac{L}{2}|\nabla Q|^2(t,x) + \frac{a}{2}\operatorname{tr}(Q^2(t,x)) - \frac{b}{3}\operatorname{tr}(Q^3(t,x)) + \frac{c}{4}\operatorname{tr}^2(Q^2(t,x))\, dx \tag{1.2}$$

If $d = 2, 3$ and (Q, u) is a smooth solution of (0.1) such that $Q \in L^\infty(0,T;H^1(\mathbb{R}^d)) \cap L^2(0,T;H^2(\mathbb{R}^d))$ and $u \in L^\infty(0,T;L^2(\mathbb{R}^d)) \cap L^2(0,T;H^1(\mathbb{R}^d))$ then, for all $t < T$, we have:

$$\frac{d}{dt}E(t) = -\nu \int_{\mathbb{R}^d} |\nabla u|^2 \, dx$$
$$- \Gamma \int_{\mathbb{R}^d} \operatorname{tr}\left(L\Delta Q - aQ + b[Q^2 - \frac{\operatorname{tr}(Q^2)}{3} Id] - cQ \operatorname{tr}(Q^2) \right)^2 dx \leq 0 \quad (1.3)$$

Proof. We multiply the first equation in (0.1) to the right by

$$-\left(L\Delta Q - aQ + b[Q^2 - \frac{\operatorname{tr}(Q^2)}{3} Id] - cQ \operatorname{tr}(Q^2) \right),$$

take the trace, integrate over $\mathbb{R}^d$ and by parts and sum with the second equation multiplied by u and integrated over $\mathbb{R}^d$ and by parts (let us observe that because of our assumptions on Q and u we do not have boundary terms, when integrating by parts). We obtain:

$$\frac{d}{dt} \int_{\mathbb{R}^d} \frac{1}{2}|u|^2 + \frac{L}{2}|\nabla Q|^2 + \frac{a}{2}\operatorname{tr}(Q^2) - \frac{b}{3}\operatorname{tr}(Q^3) + \frac{c}{4}\operatorname{tr}^2(Q^2)\, dx$$
$$+\nu \int_{\mathbb{R}^d} |\nabla u|^2 \, dx + \Gamma \int_{\mathbb{R}^d} \operatorname{tr}\left(L\Delta Q - aQ + b[Q^2 - \frac{\operatorname{tr}(Q^2)}{3} Id] - cQ \operatorname{tr}(Q^2) \right)^2 dx$$
$$= \underbrace{\int_{\mathbb{R}^d} u \cdot \nabla Q_{\alpha\beta} \left(-aQ_{\alpha\beta} + b[Q_{\alpha\gamma}Q_{\gamma\beta} - \frac{\delta_{\alpha\beta}}{3}\operatorname{tr}(Q^2)] - cQ_{\alpha\beta}\operatorname{tr}(Q^2)) \right) dx}_{\mathcal{I}}$$
$$+ \underbrace{\int_{\mathbb{R}^d} (-\Omega_{\alpha\gamma}Q_{\gamma\beta} + Q_{\alpha\gamma}\Omega_{\gamma\beta}) \left(-aQ_{\alpha\beta} + b[Q_{\alpha\delta}Q_{\delta\beta} - \frac{\delta_{\alpha\beta}}{3}\operatorname{tr}(Q^2)] - cQ_{\alpha\beta}\operatorname{tr}(Q^2)) \right) dx}_{\mathcal{II}}$$
$$+ \underbrace{L \int_{\mathbb{R}^d} u_\gamma Q_{\alpha\beta,\gamma} \Delta Q_{\alpha\beta} \, dx}_{\mathcal{A}} \underbrace{- \frac{L}{2} \int_{\mathbb{R}^d} u_{\alpha,\gamma} Q_{\gamma\beta} \Delta Q_{\alpha\beta} \, dx}_{\mathcal{B}}$$

$$\underbrace{+\frac{L}{2} \int_{\mathbb{R}^d} u_{\gamma,\alpha} Q_{\gamma\beta} \Delta Q_{\alpha\beta}}_{\mathcal{C}} \, dx \underbrace{+\frac{L}{2} \int_{\mathbb{R}^d} Q_{\alpha\gamma} u_{\gamma,\beta} \Delta Q_{\alpha\beta} \, dx}_{\mathcal{C}} \underbrace{- \frac{L}{2} \int_{\mathbb{R}^d} Q_{\alpha\gamma} u_{\beta,\gamma} \Delta Q_{\alpha\beta} \, dx}_{\mathcal{B}}$$
$$\underbrace{+L \int_{\mathbb{R}^d} Q_{\gamma\delta,\alpha} Q_{\gamma\delta,\beta} u_{\alpha,\beta} \, dx}_{\mathcal{AA}} - L \int_{\mathbb{R}^d} Q_{\alpha\gamma} \Delta Q_{\gamma\beta} u_{\alpha,\beta} \, dx + L \int_{\mathbb{R}^d} \Delta Q_{\alpha\gamma} Q_{\gamma\beta} u_{\alpha,\beta} \, dx$$

$$= \underbrace{-L\int_{\mathbb{R}^d} u_{\alpha,\gamma}Q_{\gamma\beta}\Delta Q_{\alpha\beta}\,dx}_{2\mathcal{B}} \underbrace{+L\int_{\mathbb{R}^d} u_{\gamma,\alpha}Q_{\gamma\beta}\Delta Q_{\alpha\beta}\,dx}_{2\mathcal{C}}$$

$$\underbrace{-L\int_{\mathbb{R}^d} Q_{\alpha\gamma}\Delta Q_{\gamma\beta}u_{\alpha,\beta}\,dx}_{\mathcal{C}\mathcal{C}} \underbrace{+L\int_{\mathbb{R}^d} \Delta Q_{\alpha\gamma}Q_{\gamma\beta}u_{\alpha,\beta}\,dx}_{\mathcal{B}\mathcal{B}} = 0 \tag{1.4}$$

where $\mathcal{I} = 0$ (since $\nabla \cdot u = 0$), $\mathcal{I}\mathcal{I} = 0$ (since $Q_{\alpha\beta} = Q_{\beta\alpha}$) and for the second equality we used

$$\underbrace{\int_{\mathbb{R}^d} u_\gamma Q_{\alpha\beta,\gamma}\Delta Q_{\alpha\beta}\,dx}_{\mathcal{A}} + \underbrace{\int_{\mathbb{R}^d} Q_{\gamma\delta,\alpha}Q_{\gamma\delta,\beta}u_{\alpha,\beta}\,dx}_{\mathcal{A}\mathcal{A}} = \int_{\mathbb{R}^d} u_\gamma Q_{\alpha\beta,\gamma}\Delta Q_{\alpha\beta}\,dx$$

$$-\int_{\mathbb{R}^d} Q_{\gamma\delta,\alpha}Q_{\gamma\delta,\beta\beta}u_\alpha\,dx - \int_{\mathbb{R}^d} Q_{\gamma\delta,\alpha\beta}Q_{\gamma\delta,\beta}u_\alpha\,dx = \int_{\mathbb{R}^d} \frac{1}{2}Q_{\gamma\delta,\alpha}Q_{\gamma\delta,\alpha}u_{\alpha,\alpha}\,dx = 0 \tag{1.5}$$

while for the last equality in (1.4) we used $2\mathcal{B} + \mathcal{B}\mathcal{B} = 2\mathcal{C} + \mathcal{C}\mathcal{C} = 0$. □

In the following we assume that there exists a smooth solution of (0.1) and obtain estimates on the behaviour of various norms:

PROPOSITION 1.2. *Let (Q, u) be a smooth solution of (0.1), with restriction (0.2), and smooth initial data $(\bar{Q}(x), \bar{u}(x))$, that decays fast enough at infinity so that we can integrate by parts in space (for any $t \geq 0$) without boundary terms.*

(i) If $\bar{Q} \in L^p$ for some $p \geq 2$ we have

$$\|Q(t,\cdot)\|_{L^p} \leq e^{Ct}\|\bar{Q}\|_{L^p}, \forall t \geq 0 \tag{1.6}$$

with $C = C(a, b, c, p, \Gamma)$.

(ii) For $d = 2, 3$ (and $(\bar{Q}, \bar{u})$ so that the right hand side of the expression below is finite) we have:

$$\|u(t,\cdot)\|_{L^2}^2 + 2\nu\int_0^t \|\nabla u(s,\cdot)\|_{L^2}^2\,ds + L\|\nabla Q(t,\cdot)\|_{L^2}^2 + \Gamma L^2\int_0^t \|\Delta Q(s,\cdot)\|_{L^2}^2\,ds$$
$$\leq \|u(0,\cdot)\|_{L^2}^2 + \|\nabla Q(0,\cdot)\|_{L^2}^2 + Ce^{Ct}\left(\|Q(0,\cdot)\|_{L^2}^2 + \|Q(0,\cdot)\|_{L^6}^6\right) \tag{1.7}$$

with the constant $C = C(a, b, c, d, L, \Gamma)$.

PROOF. (i) Multiplying the first equation in (0.1) by $2pQ\,\mathrm{tr}^{p-1}(Q^2)$ and taking the trace we obtain:

$$(\partial_t + u\cdot\nabla)\,\mathrm{tr}^p(Q^2) = \Gamma\Big(2pL\Delta Q_{\alpha\beta}Q_{\alpha\beta}\,\mathrm{tr}^{p-1}(Q^2) - 2pa\,\mathrm{tr}^p(Q^2)$$
$$+2pb\,\mathrm{tr}(Q^3)\,\mathrm{tr}^{p-1}(Q^2) - 2pc\,\mathrm{tr}^{p+1}(Q^2)\Big)$$

Let us observe that for Q a traceless, symmetric, 3×3 matrix we have:

$$\mathrm{tr}(Q^3) \leq \frac{3\varepsilon}{8}\,\mathrm{tr}^2(Q^2) + \frac{1}{\varepsilon}\,\mathrm{tr}(Q^2), \forall \varepsilon > 0 \tag{1.8}$$

Indeed, if Q has the eigenvalues $x, y, -x-y$ then $\operatorname{tr}(Q^3) = -3xy(x+y)$, $\operatorname{tr}(Q^2) = 2(x^2+y^2+xy)$ and the inequality (1.8) follows.

Integrating over $\mathbb{R}^d$, integrating by parts (we have no boundary terms because of our assumption), as well as using that $\nabla \cdot u = 0$, together with (1.8)(where $\varepsilon = \frac{4c}{3b}$) and the assumption $c > 0$ we obtain:

$$\partial_t \int_{\mathbb{R}^d} \operatorname{tr}^p(Q^2)\,dx \leq \underbrace{-2p\Gamma L \int_{\mathbb{R}^d} \nabla Q_{\alpha\beta} \nabla Q_{\alpha\beta} \operatorname{tr}^{p-1}(Q^2))\,dx}_{\leq 0}$$

$$\underbrace{-4p(p-1)\Gamma L \int_{\mathbb{R}^d} Q_{\alpha\beta,\gamma} Q_{\alpha\beta} Q_{\delta\lambda,\gamma} Q_{\delta\lambda} \operatorname{tr}^{p-2}(Q^2)\,dx}_{\leq 0} + C \int_{\mathbb{R}^d} \operatorname{tr}^p(Q^2)\,dx \tag{1.9}$$

where the constant C depends on a, b, c, p and Γ. Thus we have

$$\int_{\mathbb{R}^d} \operatorname{tr}^p(Q^2(t,x))\,dx \leq e^{Ct} \int_{\mathbb{R}^d} \operatorname{tr}^p(Q^2(0,x))\,dx \tag{1.10}$$

with $C = C(a, b, c, p, \Gamma)$.

(ii) Relation (1.3) implies

$$\begin{aligned}
\frac{L}{2}\|\nabla Q(t,\cdot)\|_{L^2}^2 + \frac{1}{2}\|u(t,\cdot)\|_{L^2}^2 + \nu \int_0^t \|\nabla u(s,\cdot)\|_{L^2}^2\,ds + \Gamma L^2 \int_0^t \|\Delta Q(s,\cdot)\|_{L^2}^2 ds \\
\leq C \int_{\mathbb{R}^d} \operatorname{tr}(Q^2(t,x)) + \operatorname{tr}^2(Q^2(t,x))\,dx + C \int_{\mathbb{R}^d} \operatorname{tr}(Q^2(0,x)) \\
+ \operatorname{tr}^2(Q^2(0,x))\,dx + \frac{L}{2}\|\nabla Q(0,\cdot)\|_{L^2} + \frac{1}{2}\|u(0,\cdot)\|_{L^2}^2 \\
+ \Gamma \int_0^t \int_{\mathbb{R}^d} \operatorname{tr}\Big(L\Delta Q\big(aQ - bQ^2 + cQ\operatorname{tr}(Q^2)\big)\Big)\,dx\,ds \\
+ \Gamma \int_0^t \int_{\mathbb{R}^d} \operatorname{tr}\Big(\big(aQ - bQ^2 + cQ\operatorname{tr}(Q^2)\big)L\Delta Q\Big)\,dx\,ds
\end{aligned}$$

In the last inequality we use Holder inequality to estimate ΔQ in L^2 and absorb it in the left hand side while the terms without gradients are estimated using (1.10) and interpolation between the L^2 and L^6 norms. □

2. Weak solutions

A pair (Q, u) is called a weak solution of the system (0.1), subject to initial data

$$Q(0,x) = \bar{Q}(x) \in L^2(\mathbb{R}^d),\ u(0,x) = \bar{u}(x) \in L^2(\mathbb{R}^d), \nabla \cdot \bar{u} = 0 \text{ in } \mathcal{D}'(\mathbb{R}^d) \tag{2.1}$$

if $Q \in L^\infty_{loc}(\mathbb{R}_+; H^1) \cap L^2_{loc}(\mathbb{R}_+; H^2)$, $u \in L^\infty_{loc}(\mathbb{R}_+; L^2) \cap L^2_{loc}(\mathbb{R}_+; H^1)$ and for every compactly supported $\varphi \in C^\infty([0,\infty) \times \mathbb{R}^d; S_0)$, $\psi \in C^\infty([0,\infty) \times \mathbb{R}^d; \mathbb{R}^d)$ with

$\nabla \cdot \psi = 0$ we have

$$\int_0^\infty \int_{\mathbb{R}^d} (-Q \cdot \partial_t \varphi - \Gamma L \Delta Q \cdot \varphi) - Q \cdot u \nabla_x \varphi - \Omega Q \cdot \varphi + Q\Omega \cdot \varphi \, dx \, dt$$
$$= \int_{\mathbb{R}^d} \bar{Q}(x) \cdot \varphi(0,x) \, dx + \Gamma \int_0^\infty \int_{\mathbb{R}^d} \Big\{ -aQ + b[Q^2 - \frac{\operatorname{tr}(Q^2)}{d} Id] - cQ \operatorname{tr}(Q^2) \Big\} \cdot \varphi \, dx \, dt \tag{2.2}$$

and

$$\int_0^\infty \int_{\mathbb{R}^d} -u \partial_t \psi - u_\alpha u_\beta \partial_\alpha \psi_\beta + \nu \nabla u \nabla \psi \, dt \, dx - \int_{\mathbb{R}^d} \bar{u}(x) \psi(0,x) \, dx$$
$$= L \int_0^\infty \int_{\mathbb{R}^d} Q_{\gamma\delta,\alpha} Q_{\gamma\delta,\beta} \psi_{\alpha,\beta} - Q_{\alpha\gamma} \Delta Q_{\gamma\beta} \psi_{\alpha,\beta} + \Delta Q_{\alpha\gamma} Q_{\gamma\beta} \psi_{\alpha,\beta} \, dx \, dt. \tag{2.3}$$

PROPOSITION 2.1. *For $d = 2, 3$ there exists a weak solution (Q, u) of the system (0.1), with restriction (0.2), subject to initial conditions (2.1). The solution (Q, u) is such that $Q \in L^\infty_{loc}(\mathbb{R}_+; H^1) \cap L^2_{loc}(\mathbb{R}_+; H^2)$ and $u \in L^\infty_{loc}(\mathbb{R}_+; L^2) \cap L^2_{loc}(\mathbb{R}_+; H^1)$.*

PROOF. We define the mollifying operator

$$\widehat{J_n f}(\xi) = 1_{[\frac{1}{n}, n]}(|\xi|) \hat{f}(\xi)$$

and consider the system:

$$\begin{cases} \partial_t Q^{(n)} + J_n\Big(\mathcal{P} J_n u^n \nabla J_n Q^{(n)}\Big) - J_n\Big(\mathcal{P} J_n \Omega^n J_n Q^{(n)}\Big) + J_n\Big(J_n Q^{(n)} \mathcal{P} J_n \Omega^n\Big) = \\ \Gamma L \Delta J_n Q^{(n)} + \Gamma\Big(- a J_n Q^{(n)} + b[J_n (J_n Q^{(n)})^2 - \frac{\operatorname{tr}(J_n (J_n Q^{(n)})^2)}{d} Id] \\ - c J_n \left(J_n Q^{(n)} \operatorname{tr}(J_n Q^{(n)})^2 \right) \Big) \\ \partial_t u^n + \mathcal{P} J_n (\mathcal{P} J_n u^n \nabla \mathcal{P} J_n u^n) = \\ - L \mathcal{P} J_n (\nabla \cdot \big(\operatorname{tr}(\nabla J_n Q^{(n)} \nabla J_n Q^{(n)}) - \frac{1}{d} |\nabla J_n Q^{(n)}|^2 Id \big)) \\ + L \mathcal{P} (\nabla \cdot J_n \left(J_n Q^{(n)} \Delta J_n Q^{(n)} - \Delta J_n Q^{(n)} J_n Q^{(n)} \right)) + \nu \Delta \mathcal{P} J_n u^n \end{cases}$$

where $\mathcal{P}$ denotes the Leray projector onto divergence-free vector fields.

The system above can be regarded as an ordinary differential equation in L^2 verifying the conditions of the Cauchy-Lipschitz theorem. Thus it admits a unique maximal solution $(Q^{(n)}, u^n) \in C^1([0, T_n); L^2(\mathbb{R}^d; \mathbb{R}^{d \times d}) \times L^2(\mathbb{R}^d, \mathbb{R}^d))$.

Using Parseval's theorem and the definition of J_n one can easily prove:

LEMMA 2.2. *For $f, g \in L^2(\mathbb{R}^d)$ we have:*

$$J_n^2 f = J_n f \tag{2.4}$$

$$\int_{\mathbb{R}^d} J_n f(x) g(x) \, dx = \int_{\mathbb{R}^d} f(x) J_n g(x) \, dx \tag{2.5}$$

$$(\mathcal{P} J_n)^2 = \mathcal{P} J_n \tag{2.6}$$

Moreover if $f \in H^k(\mathbb{R}^d)$ we have:

$$J_n(D^\alpha f) = D^\alpha J_n f \tag{2.7}$$

for any multi-index $\alpha = (\alpha_1, \ldots, \alpha_d)$, $\alpha_i \in \{0, 1, \ldots, k\}$ with $|\alpha| = \sum_{i=1}^d \alpha_i \leq k$.

Taking into account (2.4) and (2.6) the pair $(J_nQ^{(n)}, \mathcal{P}J_nu^n)$ is also a solution of (2.4). By uniqueness we have $(J_nQ^{(n)}, \mathcal{P}J_nu^n) = (Q^{(n)}, u^n)$ hence $(Q^{(n)}, u^n) \in C^1([0,T_n), H^\infty)$ and $(Q^{(n)}, u^n)$ satisfy the system:

$$\begin{cases} \partial_t Q^{(n)} + J_n\big(u^n \nabla Q^{(n)}\big) - J_n\big(\Omega^n Q^{(n)} - Q^{(n)}\Omega^n\big) = \Gamma L \Delta Q^{(n)} \\ +\Gamma\Big(- aQ^{(n)} + b[J_n(Q^{(n)})^2 - \frac{\operatorname{tr}(J_n(Q^{(n)}Q^{(n)}))}{d} Id] - cJ_n\left(Q^{(n)} \operatorname{tr}(Q^{(n)})^2\right)\Big) \\ \partial_t u^n + \mathcal{P}J_n(u^n \nabla u^n) = -L\mathcal{P}J_n(\nabla \cdot \big(\operatorname{tr}(\nabla Q^{(n)} \nabla Q^{(n)}) - \frac{1}{d}|\nabla Q^{(n)}|^2 Id\big)) \\ +L\mathcal{P}(\nabla \cdot J_n\left(Q^{(n)}\Delta Q^{(n)} - \Delta Q^{(n)} Q^{(n)}\right)) + \nu \Delta u^n \end{cases} \tag{2.8}$$

We can argue as in the proof of the apriori estimates and the same estimates hold for the approximating system (2.8). Indeed, using the previous lemma and the fact that for $(Q^{(n)}, u^{(n)})$ a solution of the previous system we have:

$$J_nQ^{(n)} = Q^{(n)},\ \mathcal{P}J_nu^{(n)} = u^{(n)} \tag{2.9}$$

we see that the cancelations that were used in Proposition 1.1 and Proposition 1.2 will hold for the approximate system (2.8). To show this we just exemplify one cancelation (namely the analogue of cancelation (1.5)) where for simplicity we drop the superscript (n) from $(u^{(n)}, Q^{(n)})$:

$$\begin{aligned} &\int_{\mathbb{R}^d} J_n(u_\gamma Q_{\alpha\beta,\gamma}) \Delta J_n Q_{\alpha\beta}\, dx + \int_{\mathbb{R}^d} J_nQ_{\gamma\delta,\alpha} J_nQ_{\gamma\delta,\beta} J_n^2 u_{\alpha,\beta}\, dx \\ &\qquad \overset{(2.9),(2.5)}{=} \int_{\mathbb{R}^d} J_nu_\gamma J_nQ_{\alpha\beta,\gamma} \Delta J_n^2 Q_{\alpha\beta}\, dx \\ -&\int_{\mathbb{R}^d} J_nQ_{\gamma\delta,\alpha} J_nQ_{\gamma\delta,\beta\beta} J_n^2 u_\alpha\, dx - \int_{\mathbb{R}^d} J_nQ_{\gamma\delta,\alpha\beta} J_nQ_{\gamma\delta,\beta} J_n^2 u_\alpha\, dx \\ &\qquad \overset{(2.4)}{=} \int_{\mathbb{R}^d} \frac{1}{2} J_nQ_{\gamma\delta,\alpha} J_nQ_{\gamma\delta,\alpha} J_n^2 u_{\alpha,\alpha}\, dx = 0 \end{aligned}$$

We multiply the first equation in (2.8) by $Q^{(n)}$, take the trace, integrate over $\mathbb{R}^d$ and using Lemma 2.2 and the cancelations described in the first part of Proposition 1.2, with $p = 2$ we obtain:

$$\|Q^{(n)}(t)\|_{L^2} \le Ce^{Ct} \|\bar{Q}\|_{L^2} \tag{2.10}$$

with C a constant independent of n.

We multiply the first equation in (2.8) by $-L\Delta Q^{(n)}$, take the trace, integrate over $\mathbb{R}^d$ and by parts, and add this to the second equation in (2.8) multiplied by u, integrated over $\mathbb{R}^d$ and by parts. Using Lemma 2.2 and the cancelations described in Proposition 1.1 we obtain:

$$\frac{d}{dt}\int_{\mathbb{R}^d}\frac{L}{2}|\nabla Q^{(n)}|^2+\frac{1}{2}|u^n|^2\,dx+\nu\int_{\mathbb{R}^d}|\nabla u^n|^2\,dx+\Gamma L\int_{\mathbb{R}^d}|\Delta Q^{(n)}|^2\,dx$$
$$=\Gamma L\left(-a\int_{\mathbb{R}^d}|\nabla Q^{(n)}|^2\,dx+2b\int_{\mathbb{R}^d}\mathrm{tr}(\nabla Q^{(n)}\odot\nabla Q^{(n)}Q^{(n)})dx\right)$$
$$-c\int_{\mathbb{R}^d}|\nabla Q^{(n)}|^2\,\mathrm{tr}\left((Q^{(n)})^2\right)\,dx-\Gamma L\left(\frac{c}{2}\int_{\mathbb{R}^d}|\nabla\,\mathrm{tr}\left((Q^{(n)})^2\right)|^2\,dx\right)\qquad(2.11)$$

where we denoted $(\nabla Q\odot\nabla Q)_{kl}\overset{\text{def}}{=}Q_{ij,k}Q_{ij,l},i,j,k,l=1,2,3$ (where we use again the summation convention, i.e. we sum over repeated indices).

We can estimate:

$$\int_{\mathbb{R}^d}\mathrm{tr}(\nabla Q^{(n)}\odot\nabla Q^{(n)}Q^{(n)})dx\le\frac{c}{2}\int_{\mathbb{R}^d}|\nabla Q^{(n)}|^2\,\mathrm{tr}\left((Q^{(n)})^2\right)\,dx+\tilde{C}\int_{\mathbb{R}^d}|\nabla Q^{(n)}|^2\,dx$$

for a constant $\tilde{C}$ depending only on c. Using this estimate into (2.11) and combining with Gronwall as well as using bounds (2.10) we get the following apriori bounds:

$$\sup_n\|Q^{(n)}\|_{L^2(0,T;H^2)\cap L^\infty(0,T;H^1)}<\infty$$
$$\sup_n\|u^n\|_{L^\infty(0,T;L^2)\cap L^2(0,T;H^1)}<\infty\qquad(2.12)$$

for any $T<\infty$. The estimates allow us to conclude that $T_n=\infty$.

The pair $(Q^{(n)},u^n)$ is also a weak solution of the approximating system (2.8) hence for every compactly supported $\varphi\in C^\infty([0,\infty)\times\mathbb{R}^d;S_0)$, $\psi\in C^\infty([0,\infty)\times\mathbb{R}^d;\mathbb{R}^d)$ with $\nabla\cdot\psi=0$ we have:

$$\int_0^\infty\int_{\mathbb{R}^d}(-Q^{(n)}\cdot\partial_t\varphi-\Gamma L\Delta Q^{(n)}\cdot\varphi)-J_n\big(Q^{(n)}\cdot u^n\big)\nabla_x\varphi-J_n\big(\Omega^nQ^{(n)}\big)\cdot\varphi\,dx\,dt$$
$$+\int_0^\infty\int_{\mathbb{R}^d}J_n\big(Q^{(n)}\Omega^n\big)\cdot\varphi\,dx\,dt=\int_{\mathbb{R}^d}\bar{Q}(x)\cdot\varphi(0,x)\,dx$$
$$+\Gamma\int_0^\infty\int_{\mathbb{R}^d}\{-aQ^{(n)}+b[J_n(\left(Q^{(n)}\right)^2)-\frac{\mathrm{tr}\left(J_n(\left(Q^{(n)}\right)^2)\right)}{d}Id]\}\cdot\varphi\,dx\,dt$$
$$-c\Gamma\int_0^\infty\int_{\mathbb{R}^d}Q^{(n)}\,\mathrm{tr}(J_n(Q^{(n)})^2)\}\cdot\varphi\,dxdt\qquad(2.13)$$

and

$$\int_0^\infty\int_{\mathbb{R}^d}-u^n\partial_t\psi-J_n(u^n_\alpha u^n_\beta)\partial_\alpha\psi_\beta+\nu\nabla u^n\nabla\psi\,dx\,dt-\int_{\mathbb{R}^d}\bar{u}(x)\psi(0,x)\,dx$$
$$=L\int_0^\infty\int_{\mathbb{R}^d}\left\{J_n\left(Q^{(n)}_{\gamma\delta,\alpha}Q^{(n)}_{\gamma\delta,\beta}\right)\psi_{\alpha,\beta}-J_n\left(Q^{(n)}_{\alpha\gamma}\Delta Q^{(n)}_{\gamma\beta}-\nu\Delta Q^{(n)}_{\alpha\gamma}Q^{(n)}_{\gamma\beta}\right)\psi_{\alpha,\beta}\right\}dx\,dt\qquad(2.14)$$

We consider the solutions of (2.8) and taking into account the bounds (2.12) we get, by classical compactness and weak convergence arguments, that there exists a

$Q \in L^\infty_{loc}(\mathbb{R}_+; H^1) \cap L^2_{loc}(\mathbb{R}_+; H^2)$ and a $u \in L^\infty_{loc}(\mathbb{R}_+; L^2) \cap L^2_{loc}(\mathbb{R}_+; H^1)$ so that, on a subsequence, we have:

$$\begin{aligned} Q^{(n)} \rightharpoonup Q \text{ in } L^2(0,T;H^2) \text{ and } Q^{(n)} \to Q \text{ in } L^2(0,T;H^{2-\varepsilon}_{loc}), \forall \varepsilon > 0 \\ Q^{(n)}(t) \rightharpoonup Q(t) \text{ in } H^1 \text{ for all } t \in \mathbb{R}_+ \\ u^n \rightharpoonup u \text{ in } L^2(0,T;H^1) \text{ and } u^n \to u \text{ in } L^2(0,T;H^{1-\varepsilon}_{loc}), \forall \varepsilon > 0 \\ u^n(t) \rightharpoonup u(t) \text{ in } L^2 \text{ for all } t \in \mathbb{R}_+ \end{aligned} \tag{2.15}$$

These convergences allow us to the pass to the limit in the weak solutions (2.13),(2.14) to obtain a weak solution of (0.1), namely (2.2),(2.3). The term that is the most difficult to treat in passing to the limit is the last term in (2.14), namely

$$\begin{aligned} L\int_0^\infty \int_{\mathbb{R}^d} J_n \left(Q^{(n)}_{\alpha\gamma} \Delta Q^{(n)}_{\gamma\beta} - \Delta Q^{(n)}_{\alpha\gamma} Q^{(n)}_{\gamma\beta} \right) \psi_{\alpha,\beta} \, dx \, dt \\ = L\int_0^\infty \int_{\mathbb{R}^d} \left(Q^{(n)}_{\alpha\gamma} \Delta Q^{(n)}_{\gamma\beta} - \Delta Q^{(n)}_{\alpha\gamma} Q^{(n)}_{\gamma\beta} \right) \cdot J_n \psi_{\alpha,\beta} \, dx \, dt. \end{aligned}$$

Recalling that ψ is compactly supported we have that there exists a time $T > 0$ so that $\psi(t,x) = J_n\psi(t,x) = 0, \forall t > T, x \in \mathbb{R}^d, n \in \mathbb{N}$. Taking into account that ψ is compactly supported and the convergences (2.15) one can easily pass to the limit the terms $\partial_\beta J_n \psi_\alpha Q^{(n)}_{\alpha\gamma}$ and $\partial_\beta J_n \psi_\alpha Q^{(n)}_{\gamma\beta}$ strongly in $L^2(0,T;L^2)$. Indeed we have:

$$\partial_\beta J_n \psi_\alpha Q^{(n)}_{\alpha\gamma} - \partial_\beta \psi_\alpha Q_{\alpha\gamma} = \underbrace{\left(\partial_\beta J_n \psi_\alpha - \partial_\beta \psi_\alpha \right) Q^{(n)}_{\alpha\gamma}}_{\mathcal{I}} + \underbrace{\partial_\beta \psi_\alpha \left(Q^{(n)}_{\alpha\gamma} - Q_{\alpha\gamma} \right)}_{\mathcal{II}} \tag{2.16}$$

and the first term, $\mathcal{I}$, converges to 0, strongly in $L^2(0,T;L^2)$ because ψ is smooth and compactly supported, hence $\partial_\beta J_n \psi - \partial_\beta \psi$ converges to zero in any $L^q(0,T;L^p)$ and $Q^{(n)}$ is bounded in L^∞ in time and L^p in space ($1 < p < \infty$ if $d = 2$ and $2 \le p \le 6$ if $d = 3$, due to the bounds (2.12)). On the other hand the second term $\mathcal{II}$ converges strongly to zero in $L^2(0,T;L^2)$ because of (2.15) and the fact that ψ is compactly supported.

Relations (2.15) give that $\Delta Q^{(n)}_{\gamma\beta}$, $\Delta Q^{(n)}_{\alpha\gamma}$ converges weakly in $L^2(0,T;L^2)$. Thus we get convergence to the limit term

$$\begin{aligned} L\int_0^\infty \int_{\mathbb{R}^d} (\Delta Q_{\gamma\beta})(\partial_\beta \psi_\alpha Q_{\alpha\gamma}) dx dt - L\int_0^\infty \int_{\mathbb{R}^d} (\Delta Q_{\alpha\gamma})(\partial_\beta \psi_\alpha Q_{\gamma\beta}) dx dt \\ = L\int_0^T \int_{\mathbb{R}^d} (\Delta Q_{\gamma\beta})(\partial_\beta \psi_\alpha Q_{\alpha\gamma}) dx dt - L\int_0^T \int_{\mathbb{R}^d} (\Delta Q_{\alpha\gamma})(\partial_\beta \psi_\alpha Q_{\gamma\beta}) dx dt. \end{aligned} \tag{2.17}$$

□

3. Strong solutions

In this section we restrict ourselves to dimension two and show that starting from an initial data with some higher regularity, we can obtain more regular solutions. More precisely, we have:

THEOREM 3.1. *Let $(\bar{Q}, \bar{u}) \in H^2(\mathbb{R}^2) \times H^1(\mathbb{R}^2)$. There exists a global a solution $(Q(t,x), u(t,x))$ of the system (0.1), with restriction (0.2), subject to initial conditions*

$$Q(0,x) = \bar{Q}(x),\ u(0,x) = \bar{u}(x)$$

and $Q \in L^2_{loc}(\mathbb{R}_+; H^3(\mathbb{R}^2)) \cap L^\infty_{loc}(\mathbb{R}_+; H^2(\mathbb{R}^2))$, $u \in L^2_{loc}(\mathbb{R}_+; H^2(\mathbb{R}^2) \cap L^\infty_{loc}(\mathbb{R}_+; H^1)$.

Moreover, we have:

$$L\|\nabla Q(t,\cdot)\|^2_{H^1(\mathbb{R}^2)} + \|u(t,\cdot)\|^2_{H^1(\mathbb{R}^2)} \leq Ce^{e^{ct}}\left(1 + \|\bar{Q}\|_{H^2(\mathbb{R}^2)} + \|\bar{u}\|_{H^1(\mathbb{R}^2)}\right) \quad (3.1)$$

where the constant C depends only on $\bar{Q}, \bar{u}$, a, b, c, Γ and L.

The proof of the theorem is mainly based on H^1 energy estimates and the following cancelation(that is also used implicitly in showing the dissipation of the energy in Proposition 1.1):

LEMMA 3.2. *For any symmetric matrices $Q', Q \in \mathbb{R}^{d\times d}$ and $\Omega_{\alpha\beta} = \frac{1}{2}(u_{\alpha,\beta} - u_{\beta,\alpha}) \in \mathbb{R}^{d\times d}$ we have*

$$\int_{\mathbb{R}^d} \operatorname{tr}\left((\Omega Q' - Q'\Omega)\Delta Q\right) dx - \int_{\mathbb{R}^d} \partial_\beta(Q'_{\alpha\gamma}\Delta Q_{\gamma\beta} - \Delta Q_{\alpha\gamma}Q'_{\gamma\beta})u_\alpha\, dx = 0$$

PROOF. We note that

$$\begin{aligned}\int_{\mathbb{R}^d} \operatorname{tr}\left((\Omega Q' - Q'\Omega)\Delta Q\right) dx &= \int_{\mathbb{R}^d} \Omega_{\alpha\gamma}Q'_{\gamma\beta}\Delta Q_{\beta\alpha} - Q'_{\alpha\gamma}\Omega_{\gamma\beta}\Delta Q_{\beta\alpha} \\ &= \int_{\mathbb{R}^d} \Omega_{\alpha\gamma}Q'_{\gamma\beta}\Delta Q_{\beta\alpha} + \Omega_{\beta\gamma}Q'_{\gamma\alpha}\Delta Q_{\alpha\beta} = 2\int_{\mathbb{R}^d} \operatorname{tr}\left(\Omega Q'\Delta Q\right) dx \\ &= \underbrace{\int_{\mathbb{R}^d} u_{\alpha,\beta}Q'_{\beta\gamma}\Delta Q_{\gamma\alpha}\, dx}_{\mathcal{I}_1} - \underbrace{\int_{\mathbb{R}^d} u_{\beta,\alpha}Q'_{\beta\gamma}\Delta Q_{\gamma\alpha}\, dx}_{\mathcal{I}_2} \end{aligned} \quad (3.2)$$

and on the other hand

$$-\int_{\mathbb{R}^d} \partial_\beta(Q'_{\alpha\gamma}\Delta Q_{\gamma\beta})u_\alpha = \int_{\mathbb{R}^d} Q'_{\alpha\gamma}\Delta Q_{\gamma\beta}\partial_\beta u_\alpha = \int_{\mathbb{R}^d} Q'_{\beta\gamma}\Delta Q_{\gamma\alpha}\partial_\alpha u_\beta = I_2$$

and also

$$\int_{\mathbb{R}^d} \partial_\beta(\Delta Q_{\alpha\gamma}Q'_{\gamma\beta})u_\alpha = -\int_{\mathbb{R}^d} Q'_{\beta\gamma}\Delta Q_{\gamma\alpha}\partial_\beta u_\alpha = -I_1$$

which finishes the proof. □

REMARK 3.3. *The main point in the proof of the theorem is to use the previous lemma to eliminate the highest derivatives in u in the first equation of the system (0.1) and the highest derivatives in Q in the second equation of the system.*

PROOF. We start by assuming that the system (0.1) has smooth solutions and obtain apriori estimates. Applying ∂_k to the first equation in (0.1) we obtain:

$$\begin{aligned}\partial_t\partial_k Q - \Gamma L\Delta\partial_k Q - \partial_k\Omega Q + Q\partial_k\Omega &= -\partial_k(u\nabla Q) \\ &+ \partial_k\left[-aQ + b(Q^2 - \frac{\operatorname{tr}(Q^2)}{d}Id) - cQ\operatorname{tr}(Q^2)\right] + \Omega\partial_k Q - \partial_k Q\Omega \end{aligned} \quad (3.3)$$

Multiplying by $-L\Delta\partial_k Q$, summing over repeated index k, integrating over $\mathbb{R}^d$ and by parts we obtain:

$$\begin{aligned}\frac{L}{2}\frac{d}{dt}\|\Delta Q\|^2 + L\|\Delta\nabla Q\|_{L^2} + L\int_{\mathbb{R}^d}\partial_k\Omega Q\Delta\partial_k Q\,dx - L\int_{\mathbb{R}^d} Q\partial_k\Omega\Delta\partial_k Q \\ = L\underbrace{(\partial_k(u\nabla Q),\Delta\partial_k Q)}_{\mathcal{I}_1} - L\underbrace{\left(\partial_k(-aQ+bQ^2-cQ\,\mathrm{tr}(Q^2)\right),\Delta\partial_k Q)}_{\mathcal{I}_2} \\ - L\underbrace{\int_{\mathbb{R}^d}\Omega\partial_k Q\partial_k\Delta Q\,dx}_{\mathcal{I}_3} + L\underbrace{\int_{\mathbb{R}^d}\partial_k Q\Omega\partial_k\Delta Q\,dx}_{\mathcal{I}_4} \qquad (3.4)\end{aligned}$$

Applying formally ∂_k to the second equation in (0.1) we obtain:

$$\begin{aligned}&\partial_t\partial_k u - \nu\Delta\partial_k u - L\nabla\cdot(Q\Delta\partial_k Q - \Delta\partial_k QQ) \\ &= -\partial_k(u\nabla u) - L\partial_k\nabla\cdot\left(\nabla Q\odot\nabla Q - \frac{|\nabla Q|^2}{d}Id\right) - L\nabla\cdot(\partial_k Q\Delta Q - \Delta Q\partial_k Q)\end{aligned} \qquad (3.5)$$

Multiplying the last equation by $\partial_k u$ summing over repeated index k, integrating over $\mathbb{R}^d$ and by parts we obtain:

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|\nabla u\|_{L^2}^2 + \nu\|\nabla^2 u\|_{L^2}^2 + L\int_{\mathbb{R}^d} Q\Delta\partial_k Q\nabla\partial_k u\,dx - L\int_{\mathbb{R}^d}\Delta\partial_k QQ\nabla\partial_k u\,dx \\ = \underbrace{(\partial_k(u\nabla u),\partial_k u)}_{\mathcal{J}_1} + L\underbrace{\int_{\mathbb{R}^d}(\nabla Q\odot\nabla Q)(\nabla^2 u)\,dx}_{\mathcal{J}_2} \\ + L\underbrace{\int_{\mathbb{R}^d}\partial_k Q\Delta Q\cdot\partial_k\nabla u\,dx}_{\mathcal{J}_3} - L\underbrace{\int_{\mathbb{R}^d}\Delta Q\partial_k Q\partial_k\nabla u}_{\mathcal{J}_4} \qquad (3.6)\end{aligned}$$

Summing (3.4) and (3.6) and using Lemma 3.2 we obtain:

$$\frac{L}{2}\frac{d}{dt}\|\Delta Q\|_{L^2}^2 + L\|\Delta\nabla Q\|_{L^2}^2 + \frac{1}{2}\frac{d}{dt}\|\nabla u\|_{L^2}^2 + \nu\|\nabla^2 u\|_{L^2}^2 = \sum_{k=1}^4\mathcal{I}_i + \sum_{k=1}^4\mathcal{J}_k \qquad (3.7)$$

In order to estimate the terms in the equation above we recall the interpolation estimate (Ladyzhenskaya) that we will use repeatedly without further comment:

$$\|f\|_{L^4(\mathbb{R}^2)}^2 \le C\|f\|_{L^2(\mathbb{R}^2)}\|\nabla f\|_{L^2(\mathbb{R}^2)}$$

with the constant C independent of $f\in H^1(\mathbb{R}^2)$.

Then we can estimate:

$$\begin{aligned}|\mathcal{I}_1| = |\left(\partial_k(u\nabla Q), \Delta\partial_k Q\right)| &\le |\left(\partial_k u\nabla Q, \Delta\partial_k Q\right)| + |(u\nabla\partial_k Q, \Delta\partial_k Q)| \\ &\le \|\nabla u\|_{L^4}\|\nabla Q\|_{L^4}\|\Delta\nabla Q\|_{L^2} + \|u\|_{L^4}\|\nabla^2 Q\|_{L^4}\|\Delta\nabla Q\|_{L^2} \\ &\le \|\nabla Q\|_{L^2}^{\frac12}\|\Delta Q\|_{L^2}^{\frac12}\|\nabla u\|_{L^2}^{\frac12}\|\nabla^2 u\|_{L^2}^{\frac12}\|\Delta\nabla Q\|_{L^2} \\ &\quad + \|u\|_{L^2}^{\frac12}\|\nabla u\|_{L^2}^{\frac12}\|\nabla^2 Q\|_{L^2}^{\frac12}\|\Delta\nabla Q\|_{L^2}^{\frac32} \\ &\le C\|\nabla Q\|_{L^2}^2\|\Delta Q\|_{L^2}^2\|\nabla u\|_{L^2}^2 + C\|u\|_{L^2}^2\|\nabla u\|_{L^2}^2\|\nabla^2 Q\|_{L^2}^2 \\ &\quad + \frac{\nu}{100}\|\nabla^2 u\|_{L^2}^2 + \frac{L}{100}\|\Delta\nabla Q\|_{L^2}^2\end{aligned}$$

$$\begin{aligned}|\mathcal{I}_2| &= |\left(\partial_k(-aQ + bQ^2 - cQ\,\mathrm{tr}(Q^2)\right), \Delta\partial_k Q)| \\ &\le |a|\|\Delta Q\|_{L^2}^2 + 2|b|\|Q\|_{L^4}\|\nabla Q\|_{L^4}\|\Delta\nabla Q\|_{L^2} + 3|c|\|Q\|_{L^4}^2\|\nabla Q\|_{L^4}\|\Delta\nabla Q\|_{L^2} \\ &\le |a|\|\Delta Q\|_{L^2}^2 + 2|b|\|Q\|_{L^4}\|\nabla Q\|_{L^2}^{\frac12}\|\Delta Q\|_{L^2}^{\frac12}\|\Delta\nabla Q\|_{L^2} \\ &\quad + 3|c|\|Q\|_{L^4}^2\|\nabla Q\|_{L^2}^{\frac12}\|\Delta Q\|_{L^2}^{\frac12}\|\Delta\nabla Q\|_{L^2} \\ &\le |a|\|\Delta Q\|_{L^2}^2 + C\|Q\|_{L^4}^4\|\nabla Q\|_{L^2}^2 + \|\Delta Q\|_{L^2}^2 + \frac{L}{100}\|\Delta\nabla Q\|_{L^2}^2\end{aligned}$$

$$\begin{aligned}|\mathcal{I}_3| = |\int_{\mathbb{R}^d}\Omega\partial_k Q\partial_k\Delta Q\,dx| &\le \||\nabla u\|_{L^4}\|\nabla Q\|_{L^4}\|\Delta\nabla Q\|_{L^2} \\ &\le \|\nabla Q\|_{L^2}^{\frac12}\|\Delta Q\|_{L^2}^{\frac12}\|\nabla u\|_{L^2}^{\frac12}\|\nabla^2 u\|_{L^2}^{\frac12}\|\Delta\nabla Q\|_{L^2} \\ &\le C\|\nabla Q\|_{L^2}^2\|\Delta Q\|_{L^2}^2\|\nabla u\|_{L^2}^2 + \frac{\nu}{100}\|\nabla^2 u\|_{L^2}^2 + \frac{L}{100}\|\Delta\nabla Q\|_{L^2}^2\end{aligned}$$

One can immediately see that $\mathcal{I}_4$ can be estimated exactly as $\mathcal{I}_3$.

$$\begin{aligned}|\mathcal{J}_1| = |(\partial_k u\nabla u, \partial_k u)| + |(u\partial_k\nabla u, \partial_k u)| &\le \|\nabla u\|_{L^4}^2\|\nabla u\|_{L^2} + \|u\|_{L^4}\|\nabla u\|_{L^4}\|\Delta u\|_{L^2} \\ &\le \|\nabla u\|_{L^2}\|\Delta u\|_{L^2} + \|u\|_{L^2}^{\frac12}\|\nabla u\|_{L^2}\|\Delta u\|_{L^2}^{\frac32} \\ &\le C\|\nabla u\|_{L^2}^2\|\nabla u\|_{L^2}^2 + \|u\|_{L^2}^2\|\nabla u\|_{L^2}^2\|\nabla u\|_{L^2}^2 + \frac{\nu}{100}\|\Delta u\|_{L^2}^2\end{aligned} \tag{3.8}$$

$$\begin{aligned}|\mathcal{J}_2| \le \|\nabla Q\|_{L^4}^2\|\Delta u\|_{L^2} &\le \|\nabla Q\|_{L^2}\|\Delta Q\|_{L^2}\|\Delta u\|_{L^2} \\ &\le C\|\nabla Q\|_{L^2}^2\|\Delta Q\|_{L^2}^2 + \frac{\nu}{100}\|\Delta u\|_{L^2}^2\end{aligned} \tag{3.9}$$

$$\begin{aligned}|\mathcal{J}_3| &\le \|\nabla Q\|_{L^4}\|\Delta Q\|_{L^4}\|\nabla u\|_{L^2} \le |\nabla Q\|_{L^2}^{\frac12}\|\Delta Q\|_{L^2}\|\Delta\nabla Q\|_{L^2}^{\frac12}\|\Delta u\|_{L^2} \\ &\le C\|\nabla Q\|_{L^2}^2\|\Delta Q\|_{L^2}^2\|\Delta Q\|_{L^2}^2 + \frac{L}{100}\|\Delta\nabla Q\|_{L^2}^2 + \frac{\nu}{100}\|\Delta u\|_{L^2}^2\end{aligned} \tag{3.10}$$

One can immediately see that $\mathcal{J}_4$ can be estimated exactly as $\mathcal{J}_3$.
Using the last estimates in (3.7) we obtain:

$$\frac{d}{dt}(\frac{L}{2}\|\Delta Q\|_{L^2}^2+\frac{1}{2}\|\nabla u\|_{L^2}^2)+L\|\Delta\nabla Q\|_{L^2}^2+\nu\|\nabla^2 u\|_{L^2}^2$$
$$\leq Cf(t)\|\nabla u\|_{L^2}^2+C\tilde{f}(t)\|\Delta Q\|_{L^2}^2+Cg(t)+\frac{\nu}{2}\|\nabla^2 u\|_{L^2}^2+\frac{L}{2}\|\Delta\nabla Q\|_{L^2}^2 \qquad (3.11)$$

where

$$f(t)\stackrel{\text{def}}{=}\|\nabla Q\|_{L^2}^2\|\Delta Q\|_{L^2}^2+\|\nabla u\|_{L^2}^2(1+\|u\|_{L^2}^2)+\|u\|_{L^2}^2$$
$$\tilde{f}(t)\stackrel{\text{def}}{=}1+\|u\|_{L^2}^2\|\nabla u\|_{L^2}^2+\|\nabla Q\|_{L^2}^2+\|\nabla Q\|_{L^2}^2\|\Delta Q\|_{L^2}^2$$
$$g(t)\stackrel{\text{def}}{=}\|Q\|_{L^4}^4\|\nabla Q\|_{L^2}^2.$$

Proposition 2.1 shows that $f,\tilde{f},g\in L^1(0,T)$ for any $T>0$ and

$$\|f\|_{L^1(0,T)},\|\tilde{f}\|_{L^1(0,T)},\|g\|_{L^1(0,T)}\leq Ce^{CT}$$

with C depending only on the initial data. We denote $h(t)\stackrel{\text{def}}{=}\frac{L}{2}\|\Delta Q\|_{L^2}^2(t)+\frac{1}{2}\|\nabla u(t)\|_{L^2}^2$ and then (3.11) implies:

$$h'(t)\leq C(f(t)+\tilde{f}(t))h(t)+g(t).$$

Multiplying the last relation by $e^{-\int_0^t f(s)+\tilde{f}(s)\,ds}$, using that $f(s),\tilde{f}\geq 0$ and integrating on $[0,T]$, as well as taking into account the $L^1(0,T)$ exponential bounds on $f,\tilde{f}$ and g we obtain the claimed relation (3.1).

Let us recall that up to now we have assumed that the system (0.1) has smooth solutions and we have obtained estimates on these solutions. However, the same estimates as before can be obtained for the approximate system (2.8) used in the previous section, system that has smooth solutions as we saw. Indeed, the cancelations that are used in obtaining the previous estimates, also hold for the approximate system, thanks to the repeated use of Lemma 2.2 as shown in the proof of Proposition 2.1.

Once the apriori estimates as above are obtained one can easily pass to the limit, weakly in the approximate solutions (as in Proposition 2.1 but at a higher regularity this time) to obtain strong solutions as above.

In order to obtain the rate of increase of strong norms, as shown in (3.1) one needs to do a further approximation, namely to approximate the initial data by smoother initial data, say $(\bar{Q}^l,\bar{u}^l)\in H^3\times H^2$ with $(\bar{Q}^l,\bar{u}^l)\to(\bar{Q},\bar{u})\in H^2\times H^1$ and arguing similarly as above (i.e. differentiating the system twice and obtaining apriori estimates) one obtains the above doubly exponential estimates for the quantity $L\|\nabla Q^{(n,l)}(t,\cdot)\|_{H^2(\mathbb{R}^2)}^2+\|u^{(n,l)}(t,\cdot)\|_{H^2(\mathbb{R}^2)}^2$ where $Q^{(n,l)}(t),u^{(n,l)}$ are solutions of the approximate system (2.8) with initial data (Q^l,u^l). Then passing strongly to the limit $n\to\infty$ (locally in space initially) in the weaker norms H^1 in for $\nabla Q^{n,l}$ and $u^{(n,l)}$ we obtain the estimates (3.1) for Q^l,u^l corresponding to the smoother data $(\bar{Q}^l,\bar{u}^l)\in H^3\times H^2$). We can let now $l\to\infty$ to obtain the claimed estimate (3.1). □

Bibliography

[Bal06] J. M. Ball, *Graduate lecture notes on the q-tensor theory, unpublished lecture notes*, 2006.

[BBH93] F. Bethuel, M. Brezis, and F. Hélein, *Asymptotics for the minimization of a Ginzburg-Landau functional*, Calc. Var. Partial Differential Equations **1** (1993), no. 2, 123–148.

[BE94] A. N. Beris and B. J. Edwards, *Thermodynamics of flowing systems with internal microstructure.*, Oxford Engineering Science Series, no. 36, Oxford University Press, Oxford, New York, 1994.

[BM10] J. M. Ball and A. Majumdar, *Nematic liquid crystals : from Maier-Saupe to a continuum theory*, Mol. Cryst. Liq. Cryst **525** (2010), 1–11.

[BZ] J. M. Ball and A. Zarnescu, *Orientability and energy minimization in liquid crystal models*, Arch. Ration. Mech. Anal., in press.

[CEG+06] S. G. Cloutier, J. N. Eakin, R. S. Guico, M. E. Sousa, G. P. Crawford, and J. M. Xu, *Molecular self-organization in cylindrical nanocavities.*, Phys. Rev. E **73** (2006), 051703.

[CFTZ07] P. Constantin, C. Fefferman, E. S. Titi, and A. Zarnescu, *Regularity of coupled two-dimensional nonlinear Fokker-Planck and Navier-Stokes systems*, Comm. Math. Phys. **270** (2007), no. 3, 789–811.

[CL95] Y. Chen and F.-H. Lin, *Remarks on approximate harmonic maps*, Comment. Math. Helv. **70** (1995), no. 1, 161–169.

[CM01] J.-Y. Chemin and N. Masmoudi, *About lifespan of regular solutions of equations related to viscoelastic fluids*, SIAM J. Math. Anal. **33** (2001), no. 1, 84–112 (electronic).

[DOY] C. Denniston, E. Orlandini, and J. M. Yeomans, *Lattice Boltzmann simulations of liquid crystals hydrodynamics*, Phys. Rev. E. **63**, no. 5, 056702.

[Eri90] J. L. Ericksen, *Liquid crystals with variable degree of orientation*, Arch. Rational Mech. Anal. **113** (1990), no. 2, 97–120.

[Eva98] L. C. Evans, *Partial differential equations*, Graduate Studies in Mathematics, vol. 19, American Mathematical Society, Providence, RI, 1998. MR 1625845 (99e:35001)

[Fra58] F. C. Frank, *On the theory of liquid crystals.*, Disc. Faraday Soc. **25** (1958), 1.

[Fri22] G. Friedel, *Les états mésomorphes de la matière*, Ann.Phys. (Paris) **18** (1922), 273–474.

[Fri58] A. Friedman, *On the regularity of the solutions of nonlinear elliptic and parabolic systems of partial differential equations*, J. Math. Mech. **7**

(1958), 43–59.

[Gen72] P. G. De Gennes, *Types of singularities permitted in ordered phase*, Comptes Rendus Hebdomadaires des Seances de L'academie des Sciences, Serie B **275** (1972), no. 9, 319.

[Gen74] ———, *The physics of liquid crystals.*, Clarendon Press, Oxford, 1974.

[Kat76] T. Kato, *Perturbation theory for linear operators*, second ed., Springer-Verlag, Berlin, 1976, Grundlehren der Mathematischen Wissenschaften, Band 132.

[Les68] F. M. Leslie, *Some constitutive equations for liquid crystals*, Arch. Rational Mech. Anal. **28** (1968), no. 4, 265–283. MR 1553506

[Lin91] F.-H. Lin, *On nematic liquid crystals with variable degree of orientation*, Comm. Pure Appl. Math. **44** (1991), no. 4, 453–468. MR 1100811 (92g:58024)

[Lin96] T. C. Lin, *Ginzburg-landau vortices in superconductors and defects in biaxial nematic liquid crystals*, Ph.D. thesis, New York University, 1996.

[LL95] F.-H. Lin and C. Liu, *Nonparabolic dissipative systems modeling the flow of liquid crystals*, Comm. Pure Appl. Math. **48** (1995), no. 5, 501–537. MR 1329830 (96a:35154)

[LL00] ———, *Existence of solutions for the Ericksen-Leslie system*, Arch. Ration. Mech. Anal. **154** (2000), no. 2, 135–156. MR 1784963 (2003a:76014)

[LLW10] F.-H. Lin, J. Lin, and C. Wang, *Liquid crystal flows in two dimensions*, Arch. Ration. Mech. Anal. **197** (2010), no. 1, 297–336. MR 2646822 (2011c:35411)

[LM00] P. L. Lions and N. Masmoudi, *Global solutions for some Oldroyd models of non-Newtonian flows*, Chinese Ann. Math. Ser. B **21** (2000), no. 2, 131–146. MR 1763488 (2001c:76012)

[LP94] F.-H. Lin and C.-C. Poon, *On Ericksen's model for liquid crystals*, J. Geom. Anal. **4** (1994), no. 3, 379–392. MR 1294333 (96c:35183)

[LR99] F.-H. Lin and T. Rivière, *Complex Ginzburg-Landau equations in high dimensions and codimension two area minimizing currents*, J. Eur. Math. Soc. (JEMS) **1** (1999), no. 3, 237–311. MR 1714735 (2000g:49048)

[LW08] F. Lin and C. Wang, *The analysis of harmonic maps and their heat flows*, World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2008. MR 2431658 (2011a:58030)

[LZZ08] F.-H. Lin, P. Zhang, and Z. Zhang, *On the global existence of smooth solution to the 2-D FENE dumbbell model*, Comm. Math. Phys. **277** (2008), no. 2, 531–553. MR 2358294 (2008m:82095)

[Maj10] A. Majumdar, *Equilibrium order parameters of nematic liquid crystals in the Landau-de Gennes theory*, European J. Appl. Math. **21** (2010), no. 2, 181–203. MR 2596543 (2011c:82077)

[Mas08] N. Masmoudi, *Well-posedness for the FENE dumbbell model of polymeric flows*, Comm. Pure Appl. Math. **61** (2008), no. 12, 1685–1714. MR 2456183 (2010m:35396)

[MN04] N. J. Mottram and C. Newton, *Introduction to Q-tensor theory*, Tech. report, University of Strathclyde, Department of Mathematics, 2004.

[MZ10] A. Majumdar and A. Zarnescu, *Landau-De Gennes theory of nematic liquid crystals: the Oseen-Frank limit and beyond*, Arch. Ration. Mech. Anal. **196** (2010), no. 1, 227–280. MR 2601074 (2011g:82122)

[Nom73] K. Nomizu, *Characteristic roots and vectors of a differentiable family of symmetric matrices*, Linear and Multilinear Algebra **1** (1973), no. 2, 159–162. MR 0335536 (49 #317)

[NZ] L. Nguyen and A. Zarnescu, *Refined approximation of Landau de Gennes energy minimizers*, arXiv:1006.5689.

[PWS75] E. B. Priestley, P. Wojtowicz, and P. Sheng, *Introduction to liquid crystals*, Plenum, New York, 1975.

[PZ] M. Paicu and A. Zarnescu, *Energy dissipation and regularity for a coupled Navier-Stokes and Q-tensor system*, Arch. Ration. Mech. Anal., in press.

[Rap73] A. Rapini, *Les propriétés des cristaux liquides et leurs applications*, Progress in Solid State Chemistry **8** (1973), 337–386.

[Sch84] R. M. Schoen, *Analytic aspects of the harmonic map problem*, Seminar on nonlinear partial differential equations (Berkeley, Calif., 1983), Math. Sci. Res. Inst. Publ., vol. 2, Springer, New York, 1984, pp. 321–358. MR 765241 (86b:58032)

[Sch09] M. E. Schonbek, *Existence and decay of polymeric flows*, SIAM J. Math. Anal. **41** (2009), no. 2, 564–587. MR 2507462 (2010d:76012)

[SKH95] A. M. Sonnet, A. Kilian, and S. Hess, *Alignment tensor versus director: Description of defects in nematic liquid-crystals*, Phys. Rev. E **52** (1995), no. 1, 718–722.

[SMV04] A. M. Sonnet, P. L. Maffettone, and E. G. Virga, *Continuum theory for nematic liquid crystals with tensorial order.*, J. Non-Newtonian Fluid Mech. **119** (2004), 51–59.

[Tay11] M. E. Taylor, *Partial differential equations III. Nonlinear equations*, second ed., Applied Mathematical Sciences, vol. 117, Springer, New York, 2011. MR 2744149

[TDY03] G. Toth, C. Denniston, and J. M. Yeomans, *Hydrodynamics of domain growth in nematic liquid crystals*, Phys. Rev. E **67** (2003), 051705.

[Vir94] E. G. Virga, *Variational theories for liquid crystals*, Applied Mathematics and Mathematical Computation, vol. 8, Chapman & Hall, London, 1994. MR 1369095 (97m:73001)

[ZK08] S. Zhang and S. Kumar, *Carbon nanotubes as liquid crystals*, Small **4** (2008), no. 9, 1270–1283.

APPENDIX A

Representations of Q and the biaxiality parameter $\beta(Q)$

PROPOSITION 0.1. *A matrix $Q \in S_0$ can be represented in the form*

$$Q = s(n \otimes n - \frac{1}{3}Id) + r(m \otimes m - \frac{1}{3}Id) \tag{0.1}$$

with n and m unit-length eigenvectors of Q, $n \cdot m = 0$ and

$$0 \leq r \leq \frac{s}{2} \quad or \ \frac{s}{2} \leq r \leq 0 \tag{0.2}$$

The scalar order parameters r and s are piecewise linear combinations of the eigenvalues of Q.

PROOF. From the spectral decomposition theorem we have

$$Q = \lambda_1 n_1 \otimes n_1 + \lambda_2 n_2 \otimes n_2 + \lambda_3 n_3 \otimes n_3 \tag{0.3}$$

where $\lambda_1, \lambda_2, \lambda_3$ are eigenvalues of Q and n_1, n_2, n_3 are the corresponding unit eigenvectors, pairwise perpendicular. We have $I = \sum_{i=1}^{3} n_i \otimes n_i$ and the tracelessness condition implies that $\lambda_1 + \lambda_2 + \lambda_3 = 0$. Thus

$$Q = \lambda_1 n_1 \otimes n_1 + \lambda_2 n_2 \otimes n_2 - (\lambda_1 + \lambda_2)(Id - n_1 \otimes n_1 - n_2 \otimes n_2)$$

We consider six regions $R_i^+, i = 1, \ldots, 6$ in the (λ_1, λ_2) - plane which cover exactly half of the whole plane. This corresponds to the representation (0.1) with $0 \leq r \leq \frac{s}{2}$. The other half of the plane is covered by the regions $R_i^-, i = 1, \ldots, 6$, (which are obtained by reflecting R_i^+ through the origin $(0, 0)$) and the regions R_i^- correspond to the representation (0.1),with $r, s \leq 0$.

We let $R_1^+ = \{(\lambda_1, \lambda_2) \in \mathbb{R}^2, -2\lambda_1 \leq \lambda_2, \lambda_1 \leq 0\}$. In this case $r \stackrel{def}{=} 2\lambda_1 + \lambda_2$ and $s \stackrel{def}{=} 2\lambda_2 + \lambda_1$ with $n \stackrel{def}{=} n_2, m \stackrel{def}{=} n_1$. One can directly verify that for r, s thus defined, we have

$$r = 2\lambda_1 + \lambda_2 \leq \frac{s}{2} = \lambda_2 + \frac{\lambda_1}{2}.$$

Interchanging λ_1 with λ_2 in the definition of r and s and m with n, we obtain the region $R_2^+ = \{(\lambda_1, \lambda_2); \lambda_2 \geq -\lambda_1/2; \lambda_2 \leq 0\}$.

Let $R_3^+ = \{(\lambda_1, \lambda_2) \in \mathbb{R}^2, \lambda_2 \leq 0, \lambda_2 \geq \lambda_1\}$. Taking $r \stackrel{def}{=} \lambda_2 - \lambda_1, s \stackrel{def}{=} -2\lambda_1 - \lambda_2$, $n \stackrel{def}{=} n_3, m \stackrel{def}{=} n_2$, one can check that

$$r = \lambda_2 - \lambda_1 \leq \frac{s}{2} = -\lambda_1 - \frac{\lambda_2}{2}.$$

The region R_4^+ is obtained from interchanging λ_1 and λ_2 in the definitions of r and s and letting $m = n_1, n = n_3$.

We have $R_5^+ = \{(\lambda_1, \lambda_2) \in \mathbb{R}^2, \lambda_1 \leq 0, -2\lambda_1 \geq \lambda_2 \geq -\lambda_1\}$ with $r \stackrel{def}{=} -2\lambda_1 - \lambda_2, s \stackrel{def}{=} \lambda_2 - \lambda_1$, $n = n_2$ and $m \stackrel{def}{=} n_3$. Again, it is straightforward to check that

$$r = -2\lambda_1 - \lambda_2 \leq \frac{\lambda_2}{2} - \frac{\lambda_1}{2}.$$

Interchanging λ_1 with λ_2 in the definition of r,s and letting $n = n_1, m = n_3$ we obtain the region R_6^+.

Finally the remaining half of the (λ_1, λ_2)-plane is covered by the regions R_i^- (obtained from R_i^+ by changing the signs of the inequalities and keeping the definitions of r, s, n and m unchanged). For example, R_1^- is defined to be

$$R_1^- = \left\{(\lambda_1, \lambda_2) \in \mathbb{R}^2; \lambda_1 \geq 0,\ 2\lambda_1 \leq -\lambda_2\right\}$$

with $r = 2\lambda_1 + \lambda_2$ and $s = 2\lambda_2 + \lambda_1$. One can then directly check that

$$\frac{s}{2} \leq r \leq 0.$$

The remaining five regions R_i^- for $i = 2 \ldots 6$ are defined analogously. □

REMARK 0.2. *The representation formula (0.1) is known in the literature [MN04]. In Proposition 0.1, we state that it suffices to consider the two cases given by (0.2); we have not found references for this fact.*

We can also state a second representation formula for admissible $Q \in S_0$.

PROPOSITION 0.3. *Given $Q \in S_0$, where*

$$Q = s(n \otimes n - \frac{1}{3}Id) + r(m \otimes m - \frac{1}{3}Id)$$

and $0 \leq r \leq \frac{s}{2}$ or $\frac{s}{2} \leq r \leq 0$, Q can be equivalently expressed as

$$Q = S(n \otimes n - \frac{1}{3}Id) + R\,(m \otimes m - p \otimes p) \tag{0.4}$$

where n, m and p are unit-length and pairwise perpendicular eigenvectors of Q and the scalar order parameters S, R are given by

$$S = s - \frac{r}{2} \qquad R = \frac{r}{2}. \tag{0.5}$$

Moreover, $0 \leq r \leq \frac{s}{2}$ implies that $0 \leq R \leq \frac{S}{3}$ and $\frac{s}{2} \leq r \leq 0$ implies that $\frac{S}{3} \leq R \leq 0$.

PROOF. Given $Q = s(n \otimes n - \frac{1}{3}Id) + r(m \otimes m - \frac{1}{3}Id) \in S_0$, we can equivalently express it as

$$Q = \frac{2s - r}{3} n \otimes n + \frac{2r - s}{3} m \otimes m - \frac{s + r}{3} p \otimes p \tag{0.6}$$

where we use the fact that $Id = n \otimes n + m \otimes m + p \otimes p$. After some re-arrangement, (0.6) is equivalent to

$$Q = \frac{2s - r}{2}(n \otimes n - \frac{1}{3}Id) + \frac{r}{2}\,(m \otimes m - p \otimes p)\,. \tag{0.7}$$

We set $S = \frac{2s-r}{2}$ and $R = \frac{r}{2}$ and these relations can be inverted to obtain

$$r = 2R;\ s = S + R. \tag{0.8}$$

One can now readily check that the case $0 \leq r \leq \frac{s}{2}$ translates to $0 \leq R \leq \frac{S}{3}$ and the case $\frac{s}{2} \leq r \leq 0$ translates to $\frac{S}{3} \leq R \leq 0$. □

REMARK 0.4. *When using the representation formula (0.4), it suffices to consider two cases, namely* $0 \leq R \leq \frac{S}{3}$ *or* $\frac{S}{3} \leq R \leq 0$.

For a $Q \in S_0$ the biaxiality parameter $\beta(Q)$ is defined in the physical literature to be

$$\beta(Q) = 1 - 6\frac{\left(\operatorname{tr} Q^3\right)^2}{\left(\operatorname{tr} Q^2\right)^3} \tag{0.9}$$

The significance of $\beta(Q)$ as a measure of biaxiality is due to the following

LEMMA 0.5. *Let* $Q \in S_0 \setminus \{0\}$.

(i) The biaxiality parameter $\beta(Q) \in [0,1]$ *and* $\beta(Q) = 0$ *if and only if* Q *is uniaxial i.e. if* Q *is of the form,* $Q = s\left(n \otimes n - \frac{1}{3}Id\right)$ *for some* $s \in \mathbb{R}\setminus\{0\}, n \in \mathbb{S}^2$.

(ii) The ratio $\frac{r}{s}$ *can be bounded in terms of the biaxiality parameter,* $\beta(Q)$, *where* $(s,\ r)$ *are the scalar order parameters in Proposition 0.1 . These bounds are given by*

$$\frac{1}{2}\left(1 - \sqrt{1 - \sqrt{\beta}}\right) \leq \frac{r}{s} \leq \frac{1}{2}. \tag{0.10}$$

Equivalently,

$$\frac{1 - \sqrt{1 - \sqrt{\beta}}}{3 + \sqrt{1 - \sqrt{\beta}}} \leq \frac{R}{S} \leq \frac{1}{3} \tag{0.11}$$

where $(S,\ R)$ *are the order parameters in Proposition 0.3. Further* $\beta(Q) = 1$ *if and only if* $r = \frac{s}{2}$ *or if and only if* $\frac{R}{S} = \frac{1}{3}$.

(iii) For an arbitrary $Q \in S_0$, *we have that*

$$-\frac{|Q|^3}{\sqrt{6}}\left(1 - \frac{\beta}{2}\right) \leq \operatorname{tr} Q^3 \leq \frac{|Q|^3}{\sqrt{6}}\left(1 - \frac{\beta}{2}\right). \tag{0.12}$$

PROOF. (i) The quantity $\beta(Q)$ is known as the biaxiality parameter in the liquid crystal literature and it is well-known that $\beta(Q) \in [0,1]$. We present a simple proof here for completeness.

Following Proposition 0.1, we represent an arbitrary $Q \in S_0$ as

$$Q = s\left(n \otimes n - \frac{1}{3}Id\right) + r\left(m \otimes m - \frac{1}{3}Id\right) \quad 0 \leq r \leq \frac{s}{2} \text{ or } \frac{s}{2} \leq r \leq 0. \tag{0.13}$$

Since $6\frac{\left(\operatorname{tr} Q^3\right)^2}{\left(\operatorname{tr} Q^2\right)^3} \geq 0$, the inequality $\beta(Q) \leq 1$ is trivial. To show $\beta(Q) \geq 0$, we use the representation (0.13) to express $\operatorname{tr} Q^3$ and $\operatorname{tr} Q^2$ in terms of the order parameters s and r.

$$\operatorname{tr} Q^3 = \frac{1}{9}\left(2s^3 + 2r^3 - 3s^2r - 3sr^2\right)$$

$$\operatorname{tr} Q^2 = \frac{2}{3}\left(s^2 + r^2 - sr\right) \tag{0.14}$$

A straightforward calculation shows that

$$\left(\operatorname{tr} Q^3\right)^2 = \frac{1}{81}\left(4s^6 + 4r^6 - 12s^5r - 12sr^5 + 26s^3r^3 - 3s^4r^2 - 3s^2r^4\right)$$

and

$$\left(\operatorname{tr} Q^2\right)^3 = \frac{8}{27}\left(s^6 + r^6 - 3s^5 r - 3sr^5 - 7s^3 r^3 + 6s^2 r^4 + 6s^4 r^2\right).$$

One can then directly verify that

$$\left(\operatorname{tr} Q^2\right)^3 - 6\left(\operatorname{tr} Q^3\right)^2 = 2s^2 r^2 (s-r)^2 \geq 0 \tag{0.15}$$

as required. It follows immediately from (0.15) that $\beta(Q) = 0$ if and only if either $s = 0, r = 0$ or $s = r$. From (0.13), the three cases, $s = 0, r = 0$ and $s = r$, correspond to uniaxial nematic states (in fact all uniaxial states can be described by one of these three conditions) and therefore, $\beta(Q) = 0$ if and only if Q is uniaxial.

(ii) From Proposition 0.1, it suffices to consider Q-tensors with either $0 \leq r \leq \frac{s}{2}$ or $\frac{s}{2} \leq r \leq 0$. Let $\gamma = \frac{r}{s}$, then $\gamma \in \left[0, \frac{1}{2}\right]$ for the two cases under consideration. The biaxiality parameter, $\beta(Q)$, can be expressed in terms of the ratio γ as follows

$$\frac{\left(2 - 3\gamma - 3\gamma^2 + 2\gamma^3\right)^2}{\left(1 - \gamma + \gamma^2\right)^3} = 4(1 - \beta). \tag{0.16}$$

From (0.15), we have that

$$\left(2 - 3\gamma - 3\gamma^2 + 2\gamma^3\right)^2 = 4\left(1 - \gamma + \gamma^2\right)^3 - 27\gamma^2 (1 - \gamma)^2, \tag{0.17}$$

which in turn, yields the following equality

$$\frac{27\gamma^2 (1 - \gamma)^2}{\left(1 - \gamma + \gamma^2\right)^3} = 4\beta. \tag{0.18}$$

Noting that for $\gamma = \left[0, \frac{1}{2}\right]$, the polynomial $1 - \gamma + \gamma^2 \geq \frac{3}{4}$, we obtain the following upper bound

$$\beta \leq 16\gamma^2 (1 - \gamma)^2 \tag{0.19}$$

From Proposition 0.3, it suffices to consider Q-tensors with either $0 \leq R \leq \frac{S}{3}$ or with $\frac{S}{3} \leq R \leq 0$. The bounds (0.11) follow from noting that

$$r = 2R, \ s = S + R$$

and the ratio $\frac{R}{S} \in \left[0, \frac{1}{3}\right]$.

One can readily see from (0.10) that $\beta(Q) = 1$ if and only if $\frac{r}{s} = \frac{1}{2}$. The ratio $\frac{r}{s} = \frac{1}{2}$ corresponds to $\frac{R}{S} = \frac{1}{3}$ and the claims in (ii) now follow. (iii) From the definition of the biaxiality parameter in (0.9), we necessarily have that

$$\operatorname{tr} Q^3 = \pm \frac{|Q|^3}{\sqrt{6}} \sqrt{1 - \beta(Q)}. \tag{0.20}$$

It is easily checked that

$$\sqrt{1 - \beta} \leq 1 - \frac{\beta}{2} \tag{0.21}$$

The bounds (0.12) follow from combining (0.20) and (0.21). □

APPENDIX B

Properties of the bulk term $f_B(Q)$

PROPOSITION 0.1. *Consider the bulk energy density $f_B(Q)$ given by*

$$f_B(Q) = -\frac{a^2}{2}\operatorname{tr} Q^2 - \frac{b^2}{3}\operatorname{tr} Q^3 + \frac{c^2}{4}\left(\operatorname{tr} Q^2\right)^2. \tag{0.1}$$

Then $f_B(Q)$ attains its minimum for uniaxial Q-tensors of the form

$$Q = s_+\left(n \otimes n - \frac{1}{3}\right), \tag{0.2}$$

where $n : \Omega \to S^2$ is a unit eigenvector of Q and

$$s_+ = \frac{b^2 + \sqrt{b^4 + 24a^2c^2}}{4c^2}. \tag{0.3}$$

PROOF. We recall that for a symmetric, traceless matrix Q of the form

$$Q = \sum_{i=1}^{3} \lambda_i e_i \otimes e_i,$$

$\operatorname{tr} Q^n = \sum_{i=1}^{3} \lambda_i^n$ subject to the tracelessness condition so that the bulk energy density f_B in (0.1) only depends on the eigenvalues λ_1, λ_2 and λ_3. Then the stationary points of the bulk energy density f_B are given by the stationary points of the function $f : \mathbb{R}^3 \to \mathbb{R}$ defined by

$$f(\lambda_1, \lambda_2, \lambda_3) = -\frac{a^2}{2}\sum_{i=1}^{3}\lambda_i^2 - \frac{b^2}{3}\sum_{i=1}^{3}\lambda_i^3 + \frac{c^2}{4}\left(\sum_{i=1}^{3}\lambda_i^2\right)^2 - 2\delta\sum_{i=1}^{3}\lambda_i. \tag{0.4}$$

where we have recast f_B in terms of the eigenvalues and introduced a Lagrange multiplier δ for the tracelessness condition.

The equilibrium equations are given by a system of three algebraic equations

$$\frac{\partial f}{\partial \lambda_i} = 0 \Leftrightarrow -a^2\lambda_i - b^2\lambda_i^2 + c^2\left(\sum_{k=1}^{3}\lambda_k^2\right)\lambda_i = 2\delta \quad \text{for } i = 1\ldots 3, \tag{0.5}$$

or equivalently

$$(\lambda_i - \lambda_j)\left[-a^2 - b^2(\lambda_i + \lambda_j) + c^2\sum_{k=1}^{3}\lambda_k^2\right] = 0 \quad 1 \le i < j \le 3. \tag{0.6}$$

Let $\{\lambda_i\}$ be a solution of the system (0.5) with three distinct eigenvalues $\lambda_i \neq \lambda_2 \neq \lambda_3$. We consider equation (0.6) for the pairs (λ_1, λ_2) and (λ_1, λ_3). This yields two

equations

$$-a^2 - b^2(\lambda_1 + \lambda_2) + c^2 \sum_{k=1}^{3} \lambda_k^2 = 0$$
$$-a^2 - b^2(\lambda_1 + \lambda_3) + c^2 \sum_{k=1}^{3} \lambda_k^2 = 0 \tag{0.7}$$

from which we obtain

$$-b^2(\lambda_2 - \lambda_3) = 0, \tag{0.8}$$

contradicting our initial hypothesis $\lambda_2 \neq \lambda_3$. We, thus, conclude that a stationary point of the bulk energy density must have at least two equal eigenvalues and therefore correspond to either a uniaxial or isotropic liquid crystal state.

We consider an arbitrary uniaxial state given by $(\lambda_1, \lambda_2, \lambda_3) = \left(\frac{2s}{3}, -\frac{s}{3}, -\frac{s}{3}\right)$ and the corresponding Q-tensor is $Q = s\left(e_1 \otimes e_1 - \frac{1}{3}Id\right)$. The function f_B is then a quartic polynomial in the order parameter s ie.

$$f_B(s) = \frac{s^2}{27}\left(-9a^2 - 2b^2 s + 3c^2 s^2\right) \tag{0.9}$$

and the stationary points are solutions of the algebraic equation $\frac{df_B}{ds} = 0$,

$$\frac{df_B}{ds} = \frac{1}{27}\left(-18a^2 s - 6b^2 s^2 + 12c^2 s^3\right) = 0. \tag{0.10}$$

The cubic equation (0.10) admits three solutions;

$$s = 0 \quad \text{and } s_\pm = \frac{b^2 \pm \sqrt{b^4 + 24a^2c^2}}{4c^2} \tag{0.11}$$

where

$$f_B(0) = 0 \quad \text{and} \quad f_B(s_+) < f_B(s_-) < 0. \tag{0.12}$$

Symmetry considerations show that we obtain the same set of stationary points for the remaining two uniaxial choices. The global minimizer is, therefore, a uniaxial Q-tensor of the form

$$Q = s_+\left(n \otimes n - \frac{1}{3}Id\right),\ n \in \mathbb{S}^2 \tag{0.13}$$

where s_+ has been defined in (0.3). □

(Eds.) P. Kaplický

Topics in mathematical modeling and analysis

Published by
MATFYZPRESS
Publishing House of the Faculty of Mathematics and Physics
Charles University, Prague
Sokolovská 83, CZ – 186 75 Praha 8
as the 390 publication

The volume was typeset by the authors using LaTeX

Printed by
Reproduction center UK MFF
Sokolovská 83, CZ – 186 75 Praha 8

First edition

Praha 2012

ISBN 978-80-7378-196-5

Jindřich Nečas

Jindřich Nečas was born in Prague on December 14$^{\text{th}}$, 1929. He studied mathematics at the Faculty of Natural Sciences at the Charles University from 1948 to 1952. After a brief stint as a member of the Faculty of Civil Engineering at the Czech Technical University, he joined the Czechoslovak Academy of Sciences where he served as the Head of the Department of Partial Differential Equations. He held joint appointments at the Czechoslovak Academy of Sciences and the Charles University from 1967 and became a full time member of the Faculty of Mathematics and Physics at the Charles University in 1977. He spent the rest of his life there, a significant portion of it as the Head of the Department of Mathematical Analysis and the Department of Mathematical Modeling.

His initial interest in continuum mechanics led naturally to his abiding passion to various aspects of the applications of mathematics. He can be rightfully considered as the father of modern methods in partial differential equations in the Czech Republic, both through his contributions and through those of his numerous students. He has made significant contributions to both linear and non-linear theories of partial differential equations. That which immediately strikes a person conversant with his contributions is their breadth without the depth being compromised in the least bit. He made seminal contributions to the study of Rellich identities and inequalities, proved an infinite dimensional version of Sard's Theorem for analytic functionals, established important results of the type of Fredholm alternative, and most importantly established a significant body of work concerning the regularity of partial differential equations that had a bearing on both elliptic and parabolic equations. At the same time, Nečas also made important contributions to rigorous studies in mechanics. Notice must be made of his work, with his collaborators, on the linearized elastic and inelastic response of solids, the challenging field of contact mechanics, a variety of aspects of the Navier–Stokes theory that includes regularity issues as well as important results concerning transonic flows, and finally non-linear fluid theories that include fluids with shear-rate dependent viscosities, multi-polar fluids, and finally incompressible fluids with pressure dependent viscosities.

Nečas was a prolific writer. He authored or co-authored eight books. Special mention must be made of his book "Les méthodes directes en théorie des équations elliptiques" which has already had tremendous impact on the progress of the subject and will have a lasting influence in the field. He has written a hundred and forty seven papers in archival journals as well as numerous papers in the proceedings of conferences all of which have had a significant impact in various areas of applications of mathematics and mechanics.

Jindřich Nečas passed away on December 5$^{\text{th}}$, 2002. However, the legacy that Nečas has left behind will be cherished by generations of mathematicians in the Czech Republic in particular, and the world of mathematical analysts in general.

Jindřich Nečas Center for Mathematical Modeling

The Nečas Center for Mathematical Modeling is a collaborative effort between the Faculty of Mathematics and Physics of the Charles University, the Institute of Mathematics of the Academy of Sciences of the Czech Republic and the Faculty of Nuclear Sciences and Physical Engineering of the Czech Technical University.

The goal of the Center is to provide a place for interaction between mathematicians, physicists, and engineers with a view towards achieving a better understanding of, and to develop a better mathematical representation of the world that we live in. The Center provides a forum for experts from different parts of the world to interact and exchange ideas with Czech scientists.

The main focus of the Center is in the following areas, though not restricted only to them: non-linear theoretical, numerical and computer analysis of problems in the physics of continua; thermodynamics of compressible and incompressible fluids and solids; the mathematics of interacting continua; analysis of the equations governing biochemical reactions; modeling of the non-linear response of materials.

The Jindřich Nečas Center conducts workshops, house post-doctoral scholars for periods up to one year and senior scientists for durations up to one term. The Center is expected to become world renowned in its intended field of interest.

ISBN 978-80-7378-196-5