

BIRKHAÜSER

---

**ISNM**  
**International Series of Numerical Mathematics**

---

**Volume 158**

---

*Managing Editors:*  
K.-H. Hoffmann, München  
G. Leugering, Erlangen

*Associate Editors:*  
R. H. W. Hoppe, Augsburg / Houston  
C. Zhiming, Beijing

*Honorary Editor:*  
J. Todd, Pasadena †

# **Optimal Control of Coupled Systems of Partial Differential Equations**

Karl Kunisch  
Günter Leugering  
Jürgen Sprekels  
Fredi Tröltzsch  
Editors

Birkhäuser  
Basel · Boston · Berlin

Editors:

Karl Kunisch  
Institut für Mathematik und  
Wissenschaftliches Rechnen  
Universität Graz  
Heinrichstrasse 36  
8010 Graz  
Austria  
Email: karl.kunisch@uni-graz.at

Günter Leugering  
LS Angewandte Mathematik II  
Institut für Angewandte Mathematik  
Universität Erlangen-Nürnberg  
Martensstrasse 3  
91058 Erlangen  
Germany  
Email: leugering@am.uni-erlangen.de

Jürgen Sprekels  
Abt. Partielle Differentialgleichung  
Weierstraß-Institut für  
Angewandte Analysis und Stochastik  
Humboldt-Universität zu Berlin  
Mohrenstrasse 39  
10117 Berlin  
Germany  
Email: sprekels@wias-berlin.de

Fredi Tröltzsch  
Institut für Mathematik  
Fak. II Mathematik & Naturwissenschaften  
TU Berlin  
Strasse des 17. Juni 136  
10623 Berlin  
Germany  
Email: troeltz@math.tu-berlin.de

2000 Mathematics Subject Classification: Primary 49-XX, 35-XX; Secondary: 14J70, 26A16, 30F45, 35A15, 35B37, 35Bxx, 35D05, 35J85, 35L05, 35L55, 35L99, 35Q30, 35Q35, 35Q40, 47J40, 49J20, 65J15, 65K10, 49K20, 49L20, 65N15, 49Q10, 70K70, 72C02, 73K12, 74B05, 74K20, 74K25, 74K30, 74P05, 76D05, 76D07, 76N10, 76N15, 78A55, 80A20, 90C22, 90C25, 90C31, 90C90, 93A30, 93B05, 93B12, 93C20, 93D20, 93-XX, 76D55, 49K20, 49J20, 49K20, 35J65, 35B37, 93B05, 93B07

Library of Congress Control Number: 2009930314

Bibliographic information published by Die Deutsche Bibliothek.  
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie;  
detailed bibliographic data is available in the Internet at <http://dnb.ddb.de>

ISBN 978-3-7643-8922-2 Birkhäuser Verlag AG, Basel - Boston - Berlin

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in other ways, and storage in data banks. For any kind of use permission of the copyright owner must be obtained.

© 2009 Birkhäuser Verlag AG  
Basel · Boston · Berlin  
P.O. Box 133, CH-4010 Basel, Switzerland  
Part of Springer Science+Business Media

Printed on acid-free paper produced from chlorine-free pulp. TCF ∞  
Printed in Germany

ISBN 978-3-7643-8922-2  
9 8 7 6 5 4 3 2 1

e-ISBN 978-3-7643-8923-9  
[www.birkhauser.ch](http://www.birkhauser.ch)

# Contents

Preface .....	vii
<i>G. Avalos, I. Lasiecka and R. Triggiani</i>	
Beyond Lack of Compactness and Lack of Stability of a Coupled Parabolic-hyperbolic Fluid-structure System .....	1
<i>C. Brandenburg, F. Lindemann, M. Ulbrich and S. Ulbrich</i>	
A Continuous Adjoint Approach to Shape Optimization for Navier Stokes Flow .....	35
<i>E. Casas and F. Tröltzsch</i>	
Recent Advances in the Analysis of State-con- strained Elliptic Optimal Control Problems .....	57
<i>N. Cîndea and M. Tucsnak</i>	
Fast and Strongly Localized Observation for a Perturbed Plate Equation .....	73
<i>M.C. Delfour</i>	
Representations, Composition, and Decomposition of $C^{1,1}$ -hypersurfaces .....	85
<i>J.C. De Los Reyes and K. Kunisch</i>	
On Some Nonlinear Optimal Control Problems with Vector-valued Affine Control Constraints .....	105
<i>P.-É Druet</i>	
Weak Solutions to a Model for Crystal Growth from the Melt in Changing Magnetic Fields .....	123
<i>M. Gugat</i>	
Lavrentiev Prox-regularization Methods for Optimal Control Problems with Pointwise State Constraints .....	139
<i>K. Ito, K. Kunisch and Q. Zhang</i>	
Nonlinear Feedback Solutions for a Class of Quantum Control Problems .....	155

<i>I. Lasiecka and A. Tuffaha</i>	
Optimal Feedback Synthesis for Bolza Control Problem Arising in Linearized Fluid Structure Interaction .....	171
<i>E. Özkaya and N.R. Gauger</i>	
Single-step One-shot Aerodynamic Shape Optimization .....	191
<i>P.I. Plotnikov, E.V. Ruban and J. Sokolowski</i>	
Shape Differentiability of Drag Functional for Compressible Navier-Stokes Equations .....	205
<i>J.-P. Raymond and M. Vanninathan</i>	
Null-controllability for a Coupled Heat-Finite-dimensional Beam System .....	221
<i>T.I. Seidman</i>	
Feedback Modal Control of Partial Differential Equations .....	239
<i>J. Sprekels and D. Tiba</i>	
Optimization Problems for Thin Elastic Structures .....	255
<i>M. Stingl, M. Kočvara and G. Leugering</i>	
A New Non-linear Semidefinite Programming Algorithm with an Application to Multidisciplinary Free Material Optimization .....	275
<i>D. Wachsmuth and A. Rösch</i>	
How to Check Numerically the Sufficient Optimality Conditions for Infinite-dimensional Optimization Problems .....	297
<i>J.-P. Zolésio</i>	
Hidden Boundary Shape Derivative for the Solution to Maxwell Equations and Non Cylindrical Wave Equations .....	319

# Preface

The articles contained in this volume are related to presentations at the international “Conference on Optimal Control of Coupled Systems of Partial Differential Equations”, held at the “Mathematisches Forschungsinstitut Oberwolfach” from March, 2 to 8, 2008. The contributions by internationally well-known scientists in the field of Applied Mathematics cover various topics as controllability, feedback-control, optimality systems, model-reduction techniques, analysis and optimal control of flow problems and fluid-structure interactions, as well as problems of shape and topology optimization. The applications considered range from the optimization and control of quantum mechanical systems, the optimal design of airfoils, optimal control of crystal growth, the optimization of shape and topology in engineering to switching or hybrid systems. The applications are thus across all time and length scales, and range from smooth to non-smooth models.

The field of optimization and control of systems governed by partial differential equations and variational inequalities is a very active area of research in Applied Mathematics and in particular in numerical analysis, scientific computing and optimization with a growing impact on engineering applications. In return, the field benefits from fascinating and challenging applications in that new mathematical often multiscale-modeling and new numerical tools as well as novel optimization results and corresponding iterative strategies are required in order to handle these problems. In particular, it becomes amply clear that constraints have to be taken into account, both on the control- and design-variables as well as on the state variables and the domains of their governing models. Moreover, structure exploiting discretizations, adaptive and multilevel methods become predominant in large-scale applications.

The aim of the conference and hence the aim of this book was to bring together mathematicians and engineers working on challenging problems in order to mark the state-of-the-art and point to future developments. Consequently, the book addresses researchers in the area of optimization and control of infinite-dimensional systems, typically represented by partial differential equations, who are interested in both theory and numerical simulation of such systems.

The editors express their gratitude to the contributors of this volume, the Oberwolfach Institute, and the Birkhäuser Verlag for publishing this volume. They also thank F. Hante for support in the editing procedure.

K. Kunisch, G. Leugering, J. Sprekels and F. Tröltzsch

# Beyond Lack of Compactness and Lack of Stability of a Coupled Parabolic-hyperbolic Fluid-structure System

George Avalos, Irena Lasiecka and Roberto Triggiani

**Abstract.** In this paper we shall derive certain qualitative properties for a partial differential equation (PDE) system which comprises (parabolic) Stokes fluid flow and a (hyperbolic) elastic structure equation. The appearance of such coupled PDE models in the literature is well established, inasmuch as they mathematically govern many physical phenomena; e.g., the immersion of an elastic structure within a fluid. The coupling between the distinct hyperbolic and parabolic dynamics occurs at the boundary interface between the media. In [A-T.1] semigroup well-posedness on the associated space of finite energy was established for solution variables  $\{w, w_t, u\}$ , say, where,  $[w, w_t]$  are the respective displacement and velocity of the structure, and  $u$  the velocity of the fluid (there is also an associated pressure term,  $p$ , say). One problem with this fluid-structure semigroup setup is that, due to the definition of the domain  $\mathcal{D}(\mathcal{A})$  of the generator  $\mathcal{A}$ , there is no immediate implication of smoothing in the  $w$ -variable (i.e., its resolvent  $R(\lambda, \mathcal{A})$  is not compact on this component space). Thus, one is presented with the basic question of whether smooth initial data (I.C.) will give rise to higher regularity of the solutions. Accordingly, one main result described here states that said mechanical displacement, fluid velocity, and pressure term do in fact enjoy a greater regularity if, in addition to the I.C.  $\{w_0, w_1, u_0\} \in \mathcal{D}(\mathcal{A})$ , one also has  $w_0$  in  $(H^2(\Omega_s))^d$ . A second problem of the model is the inherent lack of long time stability. In this connection, a second result described here provides for uniform stabilization of the fluid-structure dynamics, by means of the insertion of a damping term at the interface between the two media.

**Mathematics Subject Classification (2000).** Primary 35Q30; Secondary 73C02; 73K12; 76D07; 93.

**Keywords.** Fluid-structure interaction, higher regularity.

---

First author: Research partially supported by the National Science Foundation under grant DMS-0606776. Second and third author: Research partially supported by the National Science Foundation under Grant DMS-0104305, and by the Army Research Office under Grant DAAD19-02-1-0179.

## 1. The coupled PDE model and its abstract version. A review from [A-T.1] of well-posedness and basic regularity theory

**Physical model.** Throughout,  $\Omega$  will be an open bounded domain in  $\mathbb{R}^d$ ,  $d = 2, 3$ , with sufficiently smooth boundary  $\partial\Omega$ . The present paper is focused on a fluid-structure interaction problem defined on  $\Omega$ , as it has arisen in the applied science and mathematical literature. See the 1969 monograph [Li.1, p. 120], which, in turn, states: “Problems of the type here considered occur in Biology [C-R.1].” Further literature will be given below. The model consists of a fluid-like equation (the Navier-Stokes equation in the velocity field and the pressure) defined on a bounded doughnut-like, exterior sub-domain  $\Omega_f$  of  $\Omega$ , which is suitably coupled with an elastic structure equation defined on an interior sub-domain  $\Omega_s$  of  $\Omega$ . A boundary interaction occurs between the two distinct dynamics at the common boundary  $\Gamma_s = \partial\Omega_s$ , of  $\Omega_s$  and  $\Omega_f$ . In short we have  $\Omega = \Omega_s \cup \Omega_f$ , and  $\overline{\Omega}_s \cap \overline{\Omega}_f = \partial\Omega_s \equiv \Gamma_s$ . The exterior boundary of  $\Omega_f$  will be denoted by  $\Gamma_f$ ; see Figure 1.

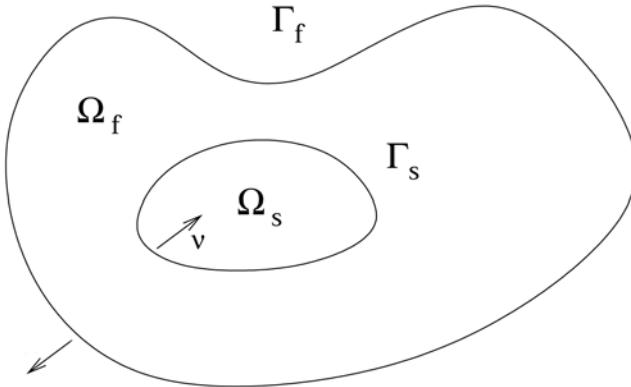


FIGURE 1. The Physical Model

**Mathematical PDE model.** In this paper, we make two simplifications:

- (i) in the structure domain  $\Omega_s$  we consider the pure  $d$ -dimensional wave equation (instead of the more cumbersome and physically more appropriate  $d$ -elastic equation: this is not mathematically crucial);
- (ii) in the fluid domain  $\Omega_f$  we take a linear version of the Navier-Stokes problem.

This is done mostly for reasons of clarity. In a subsequent paper, we intend to cover the technically more demanding elastic wave equation on  $\Omega_s$ , as well as the full (nonlinear) Navier-Stokes model in  $\Omega_f$ . Hereafter,  $u = [u_1, \dots, u_d]$  is a  $d$ -dimensional velocity field; the scalar-valued  $p$  denotes pressure;  $w = [w_1, \dots, w_d]$  is a  $d$ -dimensional displacement field. Moreover,  $\nu$  denotes throughout the unit normal vector *which is outward with respect to  $\Omega_f$*  (hence inward with respect to  $\Omega_s$  on  $\Gamma_s$ ): see Figure 1.

The fluid-structure interaction problem to be studied in the present paper is the following linear problem:

$$\left\{ \begin{array}{l} u_t - \Delta u + \nabla p \equiv 0 \quad \text{in } (0, T] \times \Omega_f \equiv Q_f \\ \operatorname{div} u \equiv 0 \quad \text{in } Q_f \\ w_{tt} - \Delta w + w \equiv 0 \quad \text{in } (0, T] \times \Omega_s \equiv Q_s \\ \text{B.C.} \left\{ \begin{array}{l} u|_{\Gamma_f} \equiv 0 \quad \text{on } (0, T] \times \Gamma_f \equiv \Sigma_f \\ u \equiv w_t \quad \text{on } (0, T] \times \Gamma_s \equiv \Sigma_s \\ \frac{\partial u}{\partial \nu} - \frac{\partial w}{\partial \nu} = p\nu \quad \text{on } \Sigma_s \end{array} \right. \\ \text{I.C. } u(0, \cdot) = u_0, \quad w(0, \cdot) = w_0, \quad w_t(0, \cdot) = w_1, \quad \text{on } \Omega. \end{array} \right. \quad \begin{array}{l} (1.1a) \\ (1.1b) \\ (1.1c) \\ (1.1d) \\ (1.1e) \\ (1.1f) \\ (1.1g) \end{array}$$

**Abstract model of Problem (1.1).** The Navier-Stokes (linear) part (1.1a) contains two unknowns; namely, the velocity field and the pressure. In the present coupled case of problem (1.1), because of the (non-homogeneous) boundary coupling (1.1e-f), it is *not* possible to use the classical, standard idea of N-S problems with *no-slip boundary conditions* to eliminate the pressure: that is, by applying the Leray projector on the equation from  $(L_2(\Omega))^d$  onto the classical space  $\{f \in (L_2(\Omega))^d; \operatorname{div} f \equiv 0 \text{ in } \Omega; f \cdot \nu = 0 \text{ on } \partial\Omega_f\}$  (see [C-F.1, p. 7]). Instead, the paper [A-T.1] (as well as paper [A-T.4], where the  $d$ -dimensional wave equation (1.1c) is replaced by the system of dynamic elasticity) eliminated the pressure by a completely different strategy. Following the idea of [Tr.1], papers [A-T.1], [A-T.4] identify a “suitable” elliptic problem for the pressure  $p$ , to be solved for  $p$  in terms of  $u$  and  $w$ .

*The elimination of  $p$ , by its explicit expression in terms of  $u$  and  $w$ .* A key idea of [A-T.1], [A-T.4] germinates from the observation that the pressure  $p(t, x)$  solves the following elliptic problem on  $\Omega_f$  in  $x$ , for each  $t$ :

$$\left\{ \begin{array}{l} \Delta p \equiv 0 \quad \text{in } (0, T] \times \Omega_f \equiv Q_f; \\ p = \frac{\partial u}{\partial \nu} \cdot \nu - \frac{\partial w}{\partial \nu} \cdot \nu \quad \text{on } (0, T] \times \Gamma_s \equiv \Sigma_s; \\ \frac{\partial p}{\partial \nu} = \Delta u \cdot \nu \quad \text{on } (0, T] \times \Gamma_f \equiv \Sigma_f. \end{array} \right. \quad \begin{array}{l} (1.2a) \\ (1.2b) \\ (1.2c) \end{array}$$

In fact, (1.2a) is obtained by taking the divergence  $\operatorname{div}$  across Eqn. (1.1a), and using  $\operatorname{div} u_t \equiv 0$  in  $Q_f$  by (1.1b), as well as  $\operatorname{div} \Delta u = \Delta \operatorname{div} u \equiv 0$  in  $Q_f$ . Next, the B.C. (1.2b) on  $\Gamma_s$  is obtained by taking the inner product of Eqn. (1.1f) with  $\nu$ . Finally, the B.C. (1.2c) on  $\Gamma_f$  is obtained by taking the inner product of Eqn. (1.1a) restricted on  $\Gamma_f$ , with  $\nu$ , using  $u|_{\Gamma_f} \equiv 0$  by (1.1d), so that on  $\Gamma_f$ ,  $\nabla p \cdot \nu = \Delta u \cdot \nu|_{\Gamma_f}$ . This then results in (1.2c).

*Explicit solution of problem (1.2) for p.* We set

$$p = p_1 + p_2 \quad \text{in } Q_f, \quad (1.2d)$$

where  $p_1$  and  $p_2$  solve the following problems:

$$\left\{ \begin{array}{ll} \Delta p_1 \equiv 0 & \text{in } Q_f; \\ p_1 \equiv \frac{\partial u}{\partial \nu} \cdot \nu - \frac{\partial w}{\partial \nu} \cdot \nu & \text{on } \Sigma_s; \end{array} \right. \quad (1.3a)$$

$$\left\{ \begin{array}{ll} \frac{\partial p_1}{\partial \nu} \Big|_{\Sigma_f} \equiv 0 & \text{on } \Sigma_f; \\ \end{array} \right. \quad (1.3b)$$

$$\left. \begin{array}{l} \Delta p_2 \equiv 0 \quad \text{in } Q_f; \\ p_2 = 0 \quad \text{on } \Sigma_s; \\ \frac{\partial p_2}{\partial \nu} \Big|_{\Sigma_f} = \Delta u \cdot \nu \quad \text{on } \Sigma_f. \end{array} \right. \quad (1.3c)$$

Accordingly, we define the following “Dirichlet” and “Neumann” maps  $D_s$  and  $N_f$ :

$$h \equiv D_s g \iff \left\{ \begin{array}{ll} \Delta h = 0 & \text{in } \Omega_f; \\ h = g & \text{on } \Gamma_s; \end{array} \right. \quad (1.4a)$$

$$\psi \equiv N_f \mu \iff \left\{ \begin{array}{ll} \psi \equiv 0 & \text{on } \Gamma_s; \\ \frac{\partial \psi}{\partial \nu} = \mu & \text{on } \Gamma_f. \end{array} \right. \quad (1.4b)$$

$$\left. \begin{array}{l} \Delta \psi \equiv 0 & \text{in } \Omega_f; \\ \frac{\partial \psi}{\partial \nu} = \mu & \text{on } \Gamma_f. \end{array} \right. \quad (1.4c)$$

Elliptic theory gives that  $D_s$  and  $N_f$  are well defined and possess the following regularity:

$$D_s : \text{ continuous } H^\rho(\Gamma_s) \rightarrow H^{\rho+\frac{1}{2}}(\Omega_f), \quad \rho \in \mathbb{R}; \quad (1.5a)$$

$$N_f : \text{ continuous } H^\rho(\Gamma_f) \rightarrow H^{\rho+\frac{3}{2}}(\Omega_f), \quad \rho \in \mathbb{R} \quad (1.5b)$$

[L-M.1]. Accordingly, in view of the respective problems (1.4), we write the solutions  $p_1, p_2$  in (1.3), and subsequently  $p$  in (1.2d), as

$$p_1 = D_s \left[ \left( \frac{\partial u}{\partial \nu} \cdot \nu - \frac{\partial w}{\partial \nu} \cdot \nu \right)_{\Sigma_s} \right]; \quad p_2 = N_f[(\Delta u \cdot \nu)_{\Sigma_f}] \quad \text{in } Q_f; \quad (1.6)$$

$$p = p_1 + p_2 = \Pi_1 w + \Pi_2 u \quad (1.7a)$$

$$= D_s \left[ \left( \frac{\partial u}{\partial \nu} \cdot \nu - \frac{\partial w}{\partial \nu} \cdot \nu \right)_{\Sigma_s} \right] + N_f[(\Delta u \cdot \nu)_{\Sigma_f}] \quad \text{in } Q_f; \quad (1.7b)$$

$$\Pi_1 w = -D_s \left[ \left( \frac{\partial w}{\partial \nu} \cdot \nu \right)_{\Sigma_s} \right]; \quad \Pi_2 u \equiv D_s \left[ \left( \frac{\partial u}{\partial \nu} \cdot \nu \right)_{\Sigma_s} \right] + N_f[(\Delta u \cdot \nu)_{\Sigma_f}] \quad \text{in } Q_f; \quad (1.8)$$

hence via (1.7a–b):

$$\nabla p = -G_1 w - G_2 u = \nabla \Pi_1 w + \nabla \Pi_2 u \quad (1.9a)$$

$$= \nabla \left( D_s \left[ \left( \frac{\partial u}{\partial \nu} \cdot \nu - \frac{\partial w}{\partial \nu} \cdot \nu \right)_{\Sigma_s} \right] \right) + \nabla (N_f[(\Delta u \cdot \nu)_{\Sigma_f}]) \quad \text{in } Q_f; \quad (1.9b)$$

$$G_1 w = -\nabla \Pi_1 w = \nabla \left\{ D_s \left[ \left( \frac{\partial w}{\partial \nu} \cdot \nu \right)_{\Sigma_s} \right] \right\} \text{ in } Q_f; \quad (1.10)$$

$$G_2 u = -\nabla \Pi_2 u = -\nabla \left\{ D_s \left[ \left( \frac{\partial u}{\partial \nu} \cdot \nu \right)_{\Sigma_s} \right] + N_f[(\Delta u \cdot \nu)_{\Sigma_f}] \right\} \text{ in } Q_f. \quad (1.11)$$

The linear maps  $G_1$  and  $G_2$  in (1.9)–(1.11) are introduced mostly for notational convenience. Eqns. (1.7), (1.9) have managed to eliminate the pressure  $p$ , and, more pertinently, its gradient  $\nabla p$ , by expressing them in terms of the two key variables: the velocity field  $u$  and the wave solution  $w$ . Using (1.9a), we accordingly rewrite the original model (1.1a–g) as

$$\left\{ \begin{array}{l} \left\{ \begin{array}{ll} u_t &= \Delta u + G_1 w + G_2 u \\ \operatorname{div} u &\equiv 0 \end{array} \right. & \text{in } Q_f \\ \left. \begin{array}{ll} w_{tt} &= \Delta w - w \\ \text{B.C.} \left\{ \begin{array}{ll} u|_{\Gamma_f} &\equiv 0 \\ u &\equiv w_t \end{array} \right. & \text{on } \Sigma_f \\ \text{I.C. } u(0, \cdot) = u_0; \, w(0, \cdot) = w_0, \, w_t(0, \cdot) = w_1 & \text{in } \Omega, \end{array} \right. & \text{in } Q_s \\ & \text{on } \Sigma_s \end{array} \right. \quad (1.12)$$

only in terms of  $u$  and  $w$ , where the pressure  $p$  has been eliminated, as desired.

*Abstract model of system (1.12).* The abstract model of system (1.12) is given by

$$\begin{aligned} \frac{d}{dt} \begin{bmatrix} w \\ w_t \\ u \end{bmatrix} &= \begin{bmatrix} 0 & I & 0 \\ \Delta - I & 0 & 0 \\ G_1 & 0 & \Delta + G_2 \end{bmatrix} \begin{bmatrix} w \\ w_t \\ u \end{bmatrix} \\ &= \mathcal{A} \begin{bmatrix} w \\ w_t \\ u \end{bmatrix}; \end{aligned} \quad (1.13a)$$

$$[w(0), w_t(0), u(0)] = [w_0, w_1, u_0] \in \mathcal{H}, \quad (1.13b)$$

where  $\mathcal{H}$  will be the finite energy space, to be specified below, which will be associated with the fluid-structure model (1.12) (see Theorem 1.2 below).

*The operator  $\mathcal{A}$ .* Recalling (1.10), (1.11) prompts the introduction of the operator

$$\mathcal{A} \equiv \begin{bmatrix} 0 & I & 0 \\ \Delta - I & 0 & 0 \\ G_1 & 0 & \Delta + G_2 \end{bmatrix} \quad (1.14a)$$

$$= \begin{bmatrix} 0 & I & 0 \\ \Delta - I & 0 & 0 \\ \nabla \left\{ D_s \left[ \left( \frac{\partial \cdot}{\partial \nu} \cdot \nu \right)_{\Gamma_s} \right] \right\} & 0 & a_{33} \end{bmatrix}; \quad (1.14b)$$

$$a_{33} = \Delta - \nabla \left\{ D_s \left[ \left( \frac{\partial \cdot}{\partial \nu} \cdot \nu \right)_{\Gamma_s} \right] + N_f [((\Delta \cdot) \cdot \nu)_{\Gamma_f}] \right\};$$

$$\mathcal{H} \supset \mathcal{D}(\mathcal{A}) \rightarrow \mathcal{H}. \quad (1.14c)$$

The finite energy space  $\mathcal{H}$  of well-posedness for problem (1.1a–g), or its abstract version (1.13)–(1.14) is defined by [A-T.1], [A-T.4]:

$$\mathcal{H} \equiv (H^1(\Omega_s))^d \times (L_2(\Omega_s))^d \times \tilde{H}_f; \quad (1.15a)$$

$$(f_1, f_2)_{H^1(\Omega_s)} = \int_{\Omega_s} [\nabla f_1 \cdot \nabla \bar{f}_2 + f_1 \cdot \bar{f}_2] d\Omega_s; \quad (1.15b)$$

$$\tilde{H}_f = \{f \in (L_2(\Omega_f))^d : \operatorname{div} f \equiv 0 \text{ in } \Omega_f; f \cdot \nu \equiv 0 \text{ on } \Gamma_f\}, \quad (1.16)$$

$\tilde{H}_f$  endowed with the  $L^2(\Omega_f)$  inner product.

The domain  $\mathcal{D}(\mathcal{A})$  of  $\mathcal{A}$  will be identified below. To this end, we find it convenient to introduce a function  $\pi$ , whose indicated regularity was ascertained in [A-T.1], [A-T.4].

**The scalar harmonic function  $\pi$ .** Henceforth, with reference to (1.14b), for  $[v_1, v_2, f] \in \mathcal{D}(\mathcal{A})$ , we introduce the harmonic function  $\pi = \pi(v_1, f)$ :

$$\pi \equiv D_s \left[ \left( \frac{\partial f}{\partial \nu} \cdot \nu \right)_{\Gamma_s} \right] + N_f [(\Delta f \cdot \nu)_{\Gamma_f}] - D_s \left[ \left( \frac{\partial v_1}{\partial \nu} \cdot \nu \right)_{\Gamma_s} \right] \in L_2(\Omega_f) \quad (1.17)$$

(compare with (1.7b) for the dynamic problem). According to the definition of the Dirichlet map  $D_s$  and Neumann map  $N_f$  given in (1.4a–c),  $\pi = \pi(v_1, f)$  in (1.17) can be equivalently given as the solution of the following elliptic problem (compare

with (1.2a-b-c) for the dynamic problem):

$$\left\{ \begin{array}{ll} \Delta\pi \equiv 0 & \text{in } \Omega_f; \\ \pi = \frac{\partial f}{\partial \nu} \cdot \nu - \frac{\partial v_1}{\partial \nu} \cdot \nu \in H^{-\frac{1}{2}}(\Gamma_s) & \text{on } \Gamma_s; \\ \frac{\partial \pi}{\partial \nu} = \Delta f \cdot \nu \in H^{-\frac{3}{2}}(\Gamma_f) & \text{on } \Gamma_f; \end{array} \right. \quad (1.18a)$$

$$\left[ \begin{array}{c} v_1 \\ v_2 \\ f \end{array} \right] \in \mathcal{D}(\mathcal{A}). \quad (1.18b)$$

$$(1.18c)$$

It then follows from  $\mathcal{A}$  in (1.14b) via the function  $\pi$  defined in (1.17) that

$$\mathcal{A} \left[ \begin{array}{c} v_1 \\ v_2 \\ f \end{array} \right] = \left[ \begin{array}{c} v_2 \\ \Delta v_1 - v_1 \\ \Delta f - \nabla \pi \end{array} \right] \equiv \left[ \begin{array}{c} v_1^* \\ v_2^* \\ f^* \end{array} \right] \in \mathcal{H}, \quad \left[ \begin{array}{c} v_1 \\ v_2 \\ f \end{array} \right] \in \mathcal{D}(\mathcal{A}). \quad (1.19)$$

**The domain  $\mathcal{D}(\mathcal{A})$ .** The domain  $\mathcal{D}(\mathcal{A})$  of  $\mathcal{A}$  is inferred from Theorem 2.1 of [A-T.1], which considers the existence and regularity issues of solutions to the coupled problems arising from imposing identity (1.19), along with the corresponding (coupled) boundary conditions dictated by the dynamics (1.12a-f).

**Proposition 1.1.** (a) *The domain  $\mathcal{D}(\mathcal{A})$  of the operator  $\mathcal{H} \supset \mathcal{D}(\mathcal{A}) \rightarrow \mathcal{H}$  in (1.19) is characterized as follows:  $\{v_1, v_2, f\} \in \mathcal{D}(\mathcal{A})$  if and only if the following properties hold true:*

$$(a_1) \quad v_1 \in (H^1(\Omega_s))^d \text{ with } \Delta v_1 \in (L_2(\Omega_s))^d, \\ \text{so that } \frac{\partial v_1}{\partial \nu} \Big|_{\Gamma_s} \in (H^{-\frac{1}{2}}(\Gamma_s))^d; \quad (1.20)$$

$$(a_2) \quad v_2 \in (H^1(\Omega_s))^d; \quad (1.21)$$

$$(a_3) \quad f \in (H^1(\Omega_f))^d \cap \tilde{H}_f, \text{ with } \Delta f - \nabla \pi \in \tilde{H}_f, \\ \text{where } \pi(v_1, f) \in L_2(\Omega_f) \text{ is the harmonic} \\ \text{function defined by (1.17) or (1.18a-c);} \quad (1.22)$$

$$(a_4) \quad \frac{\partial f}{\partial \nu} \Big|_{\Gamma_s} \in (H^{-\frac{1}{2}}(\Gamma_s))^d \text{ and } \pi|_{\Gamma_s} \in H^{-\frac{1}{2}}(\Gamma_s); \quad (1.23)$$

$$(a_5) \quad \frac{\partial f}{\partial \nu} \Big|_{\Gamma_s} = \left[ \frac{\partial v_1}{\partial \nu} + \pi \nu \right]_{\Gamma_s} \in (H^{-\frac{1}{2}}(\Gamma_s))^d; \quad (1.24)$$

$$(a_6) \quad f|_{\Gamma_f} = 0; \quad v_2|_{\Gamma_s} = f|_{\Gamma_s} \in (H^{\frac{1}{2}}(\Gamma_s))^d; \quad [\Delta f \cdot \nu]_{\Gamma_f} \in H^{-\frac{3}{2}}(\Gamma_f). \quad (1.25)$$

We can now, finally, state the basic well-posedness result for model (1.1a–g), or its abstract version (1.13)–(1.14).

**Main well-posedness results on the energy space  $\mathcal{H}$ .** Reference [A-T.1] (as well as [A-T.4] for the system of dynamic elasticity) establishes the following basic well-posedness result on the energy space  $\mathcal{H}$ .

**Theorem 1.2.** [A-T.1] *With reference to model (1.1a–g) or its abstract version (1.13)–(1.14), the following results hold true.*

(1) *The map  $\{w_0, w_1, u_0\} \rightarrow \{w(t), w_t(t), u(t)\}$  defines a strongly continuous contraction semigroup  $e^{\mathcal{A}t}$  on the energy space  $\mathcal{H}$  defined in (1.15), where the domain  $\mathcal{D}(\mathcal{A})$  of the maximal dissipative generator  $\mathcal{A}$  is identified in Proposition 1.1. The dissipativity relation of  $\mathcal{A}$  is, more specifically,*

$$Re \left( \mathcal{A} \begin{bmatrix} v_1 \\ v_2 \\ f \end{bmatrix}, \begin{bmatrix} v_1 \\ v_2 \\ f \end{bmatrix} \right)_{\mathcal{H}} = - \int_{\Omega_f} |\nabla f|^2 d\Omega_f \leq 0, \quad [v_1, v_2, f] \in \mathcal{D}(\mathcal{A}). \quad (1.26)$$

*Thus, for initial data  $y_0 = \{w_0, w_1, u_0\} \in \mathcal{H}$  as in (1.1g), there is a unique solution of the abstract Cauchy problem (1.13)–(1.14), which satisfies the regularity*

$$\{w(\cdot; y_0), w_t(\cdot; y_0), u(\cdot; y_0)\}$$

$$\in C([0, T]; \mathcal{H} = (H^1(\Omega_s))^d \times (L_2(\Omega_s))^d \times \tilde{H}_f), \quad y_0 = \{w_0, w_1, u_0\} \in \mathcal{H}; \quad (1.27)$$

*and, moreover, still with  $y_0 = \{w_0, w_1, u_0\} \in \mathcal{H}$ , and  $0 \leq s \leq t \leq T$ , the following dissipativity identity obtains:*

$$\|e^{\mathcal{A}t} y_0\|_{\mathcal{H}}^2 + 2 \int_s^t \|\nabla u(\tau; y_0)\|_{\Omega_f}^2 d\tau = \|e^{\mathcal{A}s} y_0\|_{\mathcal{H}}^2, \quad 0 \leq s \leq t. \quad (1.28)$$

*In particular, with  $y_0 = \{w_0, w_1, u_0\} \in \mathcal{H}$ ,*

$$u(\cdot; y_0) \in L_2(0, T; (H^1(\Omega_f))^d) \quad (1.29)$$

*[as it follows from (1.28) by Poincaré inequality via (1.1d)], hence*

$$u(\cdot; y_0)|_{\Gamma_s} = w_t(\cdot; y_0)|_{\Gamma_s} \in L_2(0, T; (H^{\frac{1}{2}}(\Gamma_s))^d). \quad (1.30)$$

(2) *Next, let  $y_0 = \{w_0, w_1, u_0\} \in \mathcal{D}(\mathcal{A})$ , characterized in Proposition 1.1. This identifies  $p \in L_2(\Omega_f)$  via (1.17) or (1.18). Then*

$$e^{\mathcal{A}\cdot} y_0 = \{w(\cdot; y_0), w_t(\cdot; y_0), u(\cdot; y_0)\} \in C([0, T]; \mathcal{D}(\mathcal{A})); \quad (1.31a)$$

*and thus, via Proposition 1.1,*

$$\{w, w_t, u\} \in C([0, T]; (H^1(\Omega_s))^d \times (H^1(\Omega_s))^d \times (H^1(\Omega_f))^d); \quad (1.31b)$$

$$\{p(\cdot; y_0), p(\cdot; y_0)|_{\Gamma_s}\} \in C([0, T]; L_2(\Omega_f) \times H^{-\frac{1}{2}}(\Gamma_s)), \quad (1.32a)$$

$$\nabla p \in C([0, T]; (H^{-1}(\Omega_f))^d), \quad (1.32b)$$

where the harmonic pressure  $p$  is given by the following expression (see (1.17) or equivalently the boundary value problem (1.2)):

$$p = D_s \left\{ \left[ \left( \frac{\partial u}{\partial \nu} - \frac{\partial w}{\partial \nu} \right) \cdot \nu \right]_{\Gamma_s} \right\} + N_f \{ [\Delta u \cdot \nu]_{\Gamma_f} \}, \quad (1.33)$$

and, moreover, with  $\hat{y}_0 = \mathcal{A}y_0 \in \mathcal{H}$ :

$$\begin{bmatrix} w_t(\cdot; y_0) \\ w_{tt}(\cdot; y_0) \\ u_t(\cdot; y_0) \end{bmatrix} = e^{\mathcal{A}t} \mathcal{A}y_0 = \begin{bmatrix} w(\cdot; \hat{y}_0) \\ w_t(\cdot; \hat{y}_0) \\ u(\cdot; \hat{y}_0) \end{bmatrix} \in C([0, T]; \mathcal{H}); \quad (1.34)$$

$$\|e^{\mathcal{A}t} \mathcal{A}y_0\|_{\mathcal{H}}^2 + 2 \int_s^t \|\nabla u_t(\tau; y_0)\|_{\Omega_f}^2 d\tau = \|e^{\mathcal{A}s} \mathcal{A}y_0\|_{\mathcal{H}}^2, \quad 0 \leq s \leq t. \quad (1.35)$$

In particular,

$$u_t(\cdot; y_0) \in L_2(0, T; (H^1(\Omega_f))^d). \quad (1.36)$$

(3) The Hilbert space adjoint  $\mathcal{A}^* : \mathcal{H} \supset \mathcal{D}(\mathcal{A}^*) \rightarrow \mathcal{H}$ , with  $\mathcal{D}(\mathcal{A}^*) = \mathcal{D}(\mathcal{A})$ , is likewise maximal dissipative. Analogous regularity properties hold for the adjoint problem, where the adjoint operator  $\mathcal{A}^*$  is given explicitly in [A-T.1].

*Remark 1.3.* In view of the characterization of the domain  $\mathcal{D}(\mathcal{A})$  of  $\mathcal{A}$  in Proposition 1.1, the regularity result (1.34) for  $[w_t, w_{tt}, u_t]$  is much stronger than that in [D-G-H-L.1, Thm. 3.2, p. 647], that requires

$$y_0 = \{w_0, w_1, u_0\} \in (H^2(\Omega_s))^d \times (H^2(\Omega_s))^d \times (H^2(\Omega_f))^d.$$

□

*Remark 1.4.* Here we point out the relevance and the implications of the problem considered in this paper, whose main result is Theorem 2.1 below. By Proposition 1.1, a point  $\{v_1, v_2, f\} \in \mathcal{D}(\mathcal{A})$  carries a smoothing of one unit (as measured in the scale of Sobolev spaces) of the second and third components:  $v_2 \in (H^1(\Omega_s))^d$  and  $f \in (H^1(\Omega_f))^d$ , over the basic regularity of the second and third component spaces  $(L_2(\Omega_s))^d \times \tilde{H}_f$  of the energy space  $\mathcal{H}$  in (1.15a). However, for  $\{v_1, v_2, f\} \in \mathcal{D}(\mathcal{A})$ , the first component carries no additional smoothing over the corresponding first component space  $(H^1(\Omega_s))^d$  of  $\mathcal{H}$ . [The resolvent  $R(\lambda, \mathcal{A})$  of the generator  $\mathcal{A}$  is not compact in the first component space [A-T.1].] This raises the following question: how can one generate smoother solutions, if starting with I.C. in  $\mathcal{D}(\mathcal{A})$  is not enough for the first component  $w$ , the structure displacement. This issue is important in the application of energy methods, which involve computations requiring higher regularity of the solutions over that of the basic finite energy space  $\mathcal{H}$ , starting from I.C. which are still dense in  $\mathcal{H}$ . One such case occurs when studying the uniform stabilization problem [A-T.2], [A-T.3]. The present paper addresses this issue: it provides a class of I.C., still dense in  $\mathcal{H}$ , which guarantees a higher regularity of the solution  $\{w, w_t, u\}$  across the board. □

## 2. The main result of higher regularity

Because of Theorem 1.2, we know that fluid-structure solution

$$[w, w_t, u, p] \in C([0, T]; (H^1(\Omega_s))^d \times (H^1(\Omega_s))^d \times (H^1(\Omega_f))^d \times L_2(\Omega_f));$$

$$u \in L_2(0, T; (H^1(\Omega_f))^d), \quad (2.0)$$

for initial  $[w_0, w_1, u_0] \in \mathcal{D}(\mathcal{A})$ , continuously. We will ultimately show that the mechanical displacement, fluid velocity and pressure term enjoy greater regularity, if we assume a regularity of one unit more in  $w_0$ . In fact, the main result of the first part of this paper is as follows:

**Theorem 2.1.** *Let the initial data  $\{w_0, w_1, u_0\} \in D(\mathcal{A})$  in (1.1a–g) further satisfy  $w_0 \in (H^2(\Omega_s))^d$ . Then the solution  $[w, w_t, u, p]$  of problem (1.1a–g) satisfies the following extra spatial regularity, with continuous dependence on the data:*

$$(a) \quad w \in L_\infty(0, T; (H^2(\Omega_s))^d); \quad (2.1)$$

$$(b) \quad u \in L_2(0, T; (H^2(\Omega_f))^d); \quad (2.2)$$

$$(c) \quad p \in L_2(0, T; H^1(\Omega_f)). \quad (2.3)$$

For the most part, we shall only sketch the proofs of the main and supporting results; the full details may be found in [A-L-T.1].

## 3. High-level initial conditions.

### Regularity in the tangential direction

Broadly, the proof of Theorem 2.1 is based on two main steps. Step 1, the most demanding, consists of obtaining the regularity of the solution  $\{w, w_t, u, p\}$  in the tangential direction; that is, when acted upon by a smooth first-order differential operator which is tangential on the boundary  $\Gamma_s \cup \Gamma_f$ . This is done in Theorem 3.1.1 below. Its proof is indicated in the present Section 3. Step 2 – and carried out in Section 4 – deduces then the regularity of the relevant quantities in the normal direction, by use of the equations (1.1a–b).

#### 3.1. Slashing the variables $u$ and $w$ by a first-order operator $\mathcal{B}$ on $\Omega$ , tangential to the boundary $\Gamma_f \cup \Gamma_s$

We now initiate a space regularity analysis. It consists of two steps: In this section we analyze regularity in the tangential direction, while an analysis of regularity in the normal direction will be carried out in the next Section 4. Here we follow the

pattern of, e.g., [L-L-T.1, p. 162, p. 166]. To this end, for  $\ell = 1, \dots, d-1$ , let

$$\begin{aligned} \mathcal{B} = \mathcal{B}_\ell &= \sum_{i=1}^d b_{\ell i}(\xi) \frac{\partial}{\partial \xi_i} = b_\ell(\xi) \cdot \nabla = \text{first-order, scalar differential} \\ &\text{operator with smooth coefficients } b_\ell(\cdot) = \{b_{\ell,i}(\cdot)\} \text{ on } \overline{\Omega_f \cup \Omega_s}, \\ &\text{assumed to be tangential to } \Gamma_s \cup \Gamma_f; \text{ that is, satisfying} \\ b_\ell \cdot \nu &= \sum_{i=1}^d b_{\ell,i} \nu_i = 0 \text{ on } \Gamma_s \cup \Gamma_f; \nu = [\nu_1, \dots, \nu_d] \text{ being the unit} \\ &\text{normal vector on } \Gamma_f \cup \Gamma_s, \text{ outward with respect to } \Omega_f \text{ (Fig. 1).} \end{aligned} \quad (3.1.1)$$

(For  $\ell = 1, \dots, d-1$ , such an operator  $\mathcal{B} = \mathcal{B}_\ell$ , say on  $\Gamma_s$ , may be thought of as the pre-image, under diffeomorphism via partition of unity from  $\Omega_s$  into the half-space  $\mathbb{R}_+^d = \{(x, y) : x > 0, y \in \mathbb{R}^{d-1}\}$  of the tangential derivative  $D_{y_\ell}$ ,  $\ell = 1, \dots, d-1$ , on the boundary  $x = 0$  of  $\mathbb{R}^d$  [L-L-T.1, footnote, p. 162].) Of course, when  $d = 2$ , then  $D_{y_\ell} = D_y$ .

We next convert the scalar operator  $\mathcal{B}$  into a vector form, as usual, by setting

$$\begin{aligned} \mathcal{B}u &= \mathcal{B}[u_1, \dots, u_d]^{\text{tr}} = [\mathcal{B}u_1, \dots, \mathcal{B}u_d]^{\text{tr}}, \\ \mathcal{B}w &= \mathcal{B}[w_1, \dots, w_d]^{\text{tr}} = [\mathcal{B}w_1, \dots, \mathcal{B}w_d]^{\text{tr}}. \end{aligned}$$

Thus, the vector operator  $[\mathcal{B}_1, \dots, \mathcal{B}_{d-1}]$  corresponds to the tangential gradient  $\nabla_y = \left[ \frac{\partial}{\partial y_1}, \dots, \frac{\partial}{\partial y_{d-1}} \right]$  in  $\mathbb{R}^n$ . Finally, we apply the (vectorial) operator  $\mathcal{B}$  across problem (1.1a–g), and define new variables

$$\tilde{u} \equiv \mathcal{B}u; \quad \tilde{p} \equiv \mathcal{B}p; \quad \tilde{w} \equiv \mathcal{B}w, \quad (3.1.2)$$

In these new variables, we will then obtain the following problem:

$$\left\{ \begin{array}{ll} \tilde{u}_t - \Delta \tilde{u} + \nabla \tilde{p} &= K_f(u, p) & \text{in } Q_f \\ \text{div } \tilde{u} &= [\text{div}, \mathcal{B}] \cdot u & \text{in } Q_f \\ \tilde{w}_{tt} - \Delta \tilde{w} + \tilde{w} &= K_s(w) & \text{in } Q_s \\ \text{B.C.} \begin{cases} \tilde{u} \equiv 0 & \text{in } \Sigma_f \\ \tilde{u} = \tilde{w}_t & \text{in } \Sigma_s \\ \frac{\partial \tilde{u}}{\partial \nu} - \frac{\partial \tilde{w}}{\partial \nu} = \tilde{p}\nu + (\text{div } \nu)(\tilde{w} - \tilde{u}) + p\tilde{\nu} & \text{in } \Sigma_s \end{cases} & & \\ \text{I.C. } \tilde{w}_0 = \mathcal{B}w_0; \quad \tilde{w}_1 \equiv \mathcal{B}w_1; \quad \tilde{u}_0 \equiv \mathcal{B}u_0 & & \text{in } \Omega, \end{array} \right. \quad \begin{array}{l} (3.1.3a) \\ (3.1.3b) \\ (3.1.3c) \\ (3.1.3d) \\ (3.1.3e) \\ (3.1.3f) \\ (3.1.3g) \end{array}$$

where  $K_f(u, p)$ ,  $K_s(w)$ ,  $[\text{div}, \mathcal{B}]$  are the following commutators

$$K_f(u, p) \equiv [\mathcal{B}, \Delta]u - [\mathcal{B}, \nabla]p \text{ on } Q_f; \quad K_s(w) \equiv [\mathcal{B}, \Delta]w \text{ on } Q_s; \quad (3.1.4a)$$

$$[\text{div}, \mathcal{B}] \cdot u \equiv [[\partial_{x_1}, \mathcal{B}], \dots, [\partial_{x_d}, \mathcal{B}]] \cdot u \equiv [\partial_{x_1}, \mathcal{B}]u_1 + \dots + [\partial_{x_d}, \mathcal{B}]u_d \text{ on } Q_f. \quad (3.1.4b)$$

Justification of (3.1.3) is given in Appendix A. The main result of Section 3 is the following regularity theorem for the problem (3.1.3a–g) in the slashed variables  $\{\tilde{w}, \tilde{w}_t, \tilde{u}, \tilde{p}\}$ .

**Theorem 3.1.1.** *Let initial data  $\{w_0, w_1, u_0\} \in \mathcal{D}(\mathcal{A})$  in (1.1a–g) further satisfy  $w_0 \in (H^2(\Omega_s))^d$ . Then, with reference to (3.1.1) we have the following: With  $\mathcal{B} = \mathcal{B}_\ell$ ,  $\ell = 1, \dots, d-1$ , the solution  $[\tilde{w}, \tilde{w}_t, \tilde{u}, \tilde{p}] = [\mathcal{B}w, \mathcal{B}w_t, \mathcal{B}u, \mathcal{B}p]$  of problem (3.1.3a–g) satisfies*

$$\begin{aligned} & \|\tilde{w}\|_{L_\infty(0,T;(H^1(\Omega_s))^d)} + \|\tilde{w}_t\|_{L_\infty(0,T;(L_2(\Omega_s))^d)} \\ & + \|\tilde{u}\|_{L_2(0,T;(H^1(\Omega_f))^d)} + \|\tilde{p}\|_{L_2(0,T;L^2(\Omega_f))} \\ & \leq C_T \left( \|w_0, w_1, u_0\|_{\mathcal{D}(\mathcal{A})} + \|w_0\|_{(H^2(\Omega_s))^d} \right). \end{aligned} \quad (3.1.5a)$$

In other words, if  $\nabla_y$  denotes the tangential gradient, then

$$\begin{aligned} & \|\nabla_y w\|_{L_\infty(0,T;((H^1(\Omega_s))^{(d-1)} \times d))} + \|\nabla_y w_t\|_{L_\infty((0,T;((L_2(\Omega_s))^{(d-1)} \times d)))} \\ & + \|\nabla_y u\|_{L_2(0,T;(H^1(\Omega_f))^{(d-1)} \times d))} + \|\nabla_y p\|_{L_2(0,T;(L_2(\Omega_f))^{d-1})} \\ & \leq C_T \left( \|w_0, w_1, u_0\|_{\mathcal{D}(\mathcal{A})} + \|w_0\|_{(H^2(\Omega_s))^d} \right). \end{aligned} \quad (3.1.5b)$$

The following subsections are devoted to showing the main intermediary steps in the proof of Theorem 3.1.1.

**A preliminary identity and estimate of slashed problem (3.1.3a–g).** We introduce the following notation for  $\{\tilde{w}(t), \tilde{w}_t(t), \tilde{u}(t)\} \in \mathcal{H}$  in (3.1.2):

$$\begin{aligned} E_{\tilde{w}}(t) & \equiv \|[\tilde{w}(t), \tilde{w}_t(t)]\|_{(H^1(\Omega_s))^d \times (L_2(\Omega_s))^d}^2 \\ & \equiv \int_{\Omega_s} [|\tilde{w}(t)|^2 + |\nabla \tilde{w}(t)|^2 + |\tilde{w}_t(t)|^2] d\Omega_s. \end{aligned} \quad (3.1.6)$$

Moreover, notation such as  $\|f\|_\Omega$  will refer to the  $L_2(\Omega)$ -norm of  $f$ . The following identity on problem (3.1.3a–g) is obtained by standard energy methods.

**Proposition 3.1.2.** *With  $\{w_0, w_1, u_0\} \in \mathcal{H}$ , let  $\{\tilde{w}_0, \tilde{w}_1, \tilde{u}_0\}$  be as defined in (3.1.2) (or (3.1.3g)). Then, for all  $t$ ,  $0 < t \leq T$ , the following identity holds true for problem (3.1.3a–f), in the notation of (3.1.6):*

$$\begin{aligned} & E_{\tilde{w}}(t) + \|\tilde{u}(t)\|_{\Omega_f}^2 + 2 \int_0^t \int_{\Omega_f} |\nabla \tilde{u}|^2 d\Omega_f dt \\ & = E_{\tilde{w}}(0) + \|\tilde{u}_0\|_{\Omega_f}^2 + 2 \int_0^t \int_{\Omega_s} K_f(u, p) \cdot \tilde{u} d\Omega_f ds \\ & + 2 \int_0^t \int_{\Omega_s} K_s(w) \cdot \tilde{w}_t d\Omega_s dt + 2 \int_0^t \int_{\Omega_f} \tilde{p}[\operatorname{div}, \mathcal{B}] \cdot u d\Omega_f ds \\ & + 2 \int_0^t \int_{\Gamma_s} (\operatorname{div} \nu)[\tilde{w} - \tilde{u}] \cdot \tilde{u} d\Gamma_s ds + 2 \int_0^t p \tilde{\nu} \cdot \tilde{u} d\Gamma_s ds. \end{aligned} \quad (3.1.7)$$

Identity (3.1.7) is still not the ultimate form we are seeking. To obtain the latter, we shall employ the following moment-type boundary inequality [B-S.1, p. 39], or [Th.1, p. 26]: Let  $\Omega$  be a general bounded domain in  $\mathbb{R}^d$ ,  $d \geq 2$ , with sufficiently smooth boundary  $\partial\Omega$ . Then, there is a constant  $C^* > 0$  such that

$$\|h|_{\partial\Omega}\|_{\partial\Omega} \leq \sqrt{C^*} \|h\|_{\Omega}^{\frac{1}{2}} \|h\|_{1,\Omega}^{\frac{1}{2}}, \text{ for any } h \in H^1(\Omega), \quad (3.1.8)$$

with  $C^*$  independent of  $h$ , where  $\|\cdot\|_{1,\Omega}$  denotes the  $H^1(\cdot)$ -norm. Incorporating this estimate in a majorization of (3.1.7) we then obtain

**Theorem 3.1.3.** *With  $\{w_0, w_1, u_0\} \in \mathcal{H}$ , let  $\{\tilde{w}_0, \tilde{w}_1, \tilde{u}_0\}$  be as defined in (3.1.2) (or (3.1.3g)). Then, for  $0 \leq t \leq T$ , the following inequality holds true for problem (3.1.3a–f), for given  $\epsilon > 0$  arbitrary:*

$$\begin{aligned} E_{\tilde{w}}(t) + \|\tilde{u}(t)\|_{(L_2(\Omega_f))^d}^2 + (2 - 2\epsilon) \int_0^t \int_{\Omega_f} |\nabla \tilde{u}(s)|^2 d\Omega_f ds \\ \leq E_{\tilde{w}}(0) + \|\tilde{u}_0\|_{(L_2(\Omega_f))^d}^2 + 2 \left| \int_0^t \int_{\Omega_f} K_f(u, p) \cdot \tilde{u} d\Omega_f ds \right| \\ + 2 \left| \int_0^t \int_{\Omega_s} K_s(w) \cdot \tilde{w}_t d\Omega_s ds \right| \\ + 2 \left| \int_0^t \int_{\Omega_f} \tilde{p} [\operatorname{div}, \mathcal{B}] \cdot u d\Omega_f ds \right| + 2 \left| \int_0^t \int_{\Gamma_s} p \tilde{\nu} \cdot \tilde{u} d\Gamma_s ds \right| \\ + \epsilon \int_0^t \|\tilde{w}(s)\|_{(H^1(\Omega_s))^d}^2 ds + \frac{C}{\epsilon} \int_0^t \left[ \|\tilde{w}\|_{\Omega_s}^2 + \|\tilde{u}\|_{\Omega_f}^2 \right] ds. \end{aligned} \quad (3.1.9)$$

**Corollary 3.1.4.** *Let  $\{w_0, w_1, u_0\} \in \mathcal{D}(\mathcal{A})$  as given in Proposition 1.1. Then recalling  $\mathcal{B}$  in (3.1.1), (3.1.2), and  $E_{\tilde{w}}(\cdot)$  in (3.1.6), we obtain:*

(a)

$$\tilde{w} \equiv \mathcal{B}w \in C([0, T]; (L_2(\Omega_s))^d); \quad \tilde{w}_t \equiv \mathcal{B}w_t \in C([0, T]; (L_2(\Omega_s))^d); \quad (3.1.10a)$$

$$\tilde{u} \equiv \mathcal{B}u \in C([0, T]; (L_2(\Omega_f))^d); \quad (3.1.10b)$$

$$\int_0^T \left[ \|\tilde{w}\|_{\Omega_f}^2 + \|\tilde{u}\|_{\Omega_f}^2 \right] ds = \mathcal{O} \left\{ \|w_0, w_1, u_0\|_{\mathcal{D}(\mathcal{A})}^2 \right\}. \quad (3.1.11)$$

(b) Assume further that  $w_0 \in (H^2(\Omega_s))^d$ . Then:

$$\{w_0, w_1, u_0\} \in \mathcal{D}(\mathcal{A}), \quad w_0 \in (H^2(\Omega_s))^d \Rightarrow$$

$$\tilde{w}_0 \in (H^1(\Omega_s))^d; \quad \tilde{w}_1 \in (L_2(\Omega_s))^d; \quad \tilde{u}_0 \in (L_2(\Omega_f))^d; \quad (3.1.12)$$

$$E_{\tilde{w}}(0) + \|\tilde{u}_0\|_{\Omega_f}^2$$

$$= \int_{\Omega_s} [\|\mathcal{B}w_0\|^2 + |\nabla(\mathcal{B}w_0)|^2 + |\mathcal{B}w_1|^2] d\Omega_s + \|\mathcal{B}u_0\|_{\Omega_f}^2 \quad (3.1.13)$$

$$= \mathcal{O} \left\{ \| [w_0, w_1, u_0] \|_{\mathcal{D}(\mathcal{A})}^2 + \| w_0 \|_{2, \Omega_s}^2 \right\}. \quad (3.1.14)$$

(c) Accordingly, with  $\{w_0, w_1, u_0\} \in \mathcal{D}(\mathcal{A})$ ,  $w_0 \in (H^2(\Omega_s))^d$ , and with reference to estimate (3.1.9), for  $0 \leq t \leq T$ :

$$\begin{aligned} & \|\tilde{w}(t)\|_{(H^1(\Omega_s))^d}^2 + \|\tilde{w}_t(t)\|_{(L_2(\Omega_s))^d}^2 + \|\tilde{u}(t)\|_{(L_2(\Omega_f))^d}^2 + (2 - 2\epsilon) \int_0^t \int_{\Omega_f} |\nabla \tilde{u}(t)|^2 d\Omega_f ds \\ & \leq 2 \left| \int_0^t \int_{\Omega_f} K_f(u, p) \cdot \tilde{u} d\Omega_f ds \right| + 2 \left| \int_0^t \int_{\Omega_s} K_s(w) \cdot \tilde{w}_t d\Omega_s ds \right| \\ & + 2 \left| \int_0^t \int_{\Omega_f} \tilde{p} [\operatorname{div}, \mathcal{B}] \cdot u d\Omega_f ds \right| + \epsilon \int_0^t \|\tilde{w}(s)\|_{(H^1(\Omega_s))^d}^2 ds \\ & + 2 \left| \int_0^t \int_{\Gamma_s} p \tilde{\nu} \cdot \tilde{u} d\Gamma_s ds \right| + C_{T,\epsilon} \left( \| [w_0, w_1, u_0] \|_{\mathcal{D}(\mathcal{A})}^2 + \| w_0 \|_{2, \Omega_s}^2 \right), \quad 0 \leq t \leq T. \end{aligned} \quad (3.1.15)$$

**Orientation.** In estimate (3.1.15), there are three more terms we shall estimate in the original  $\Omega_f$  by using the *a priori* regularity of Theorem 1.2(2) for  $[w_0, w_1, u_0] \in \mathcal{D}(\mathcal{A})$ . These are: (i) the penultimate and ultimate integrals on the RHS of (3.1.15), one in the interior of  $\Omega_f$  involving the integrand  $\tilde{p} [\operatorname{div}, \mathcal{B}] \cdot u$  and one on the boundary of  $\Gamma_s$  involving  $p \tilde{\nu} \cdot u$ ; and (ii) the integral involving the component  $[\mathcal{B}, \nabla]p$  of the commutator  $K_f(u, p)$  (see (3.1.4a)) in the first integral on the RHS of (3.1.15).

In our next step, we estimate the penultimate integral on the RHS of (3.1.15) involving  $\tilde{p} [\operatorname{div}, \mathcal{B}] \cdot u$ .

**Proposition 3.1.5.** *Let  $\{w_0, w_1, u_0\} \in \mathcal{D}(\mathcal{A})$ . Then, with reference to the third integral on the RHS of (3.1.15), we have*

$$2 \left| \int_0^t \int_{\Omega_f} \tilde{p} [\operatorname{div}, \mathcal{B}] \cdot u d\Omega_f ds \right| \leq \epsilon C \int_0^t \|\nabla \tilde{u}\|_{\Omega_f}^2 ds + C_{\epsilon,T} \| [w_0, w_1, u_0] \|_{\mathcal{D}(\mathcal{A})}^2. \quad (3.1.16)$$

Next, we estimate the two integrals on the RHS of (3.1.15) involving the terms  $[\mathcal{B}, \nabla]p$  and  $p \tilde{\nu} \cdot \tilde{u}$ , the former being part of the commutator  $K_f(u, p)$  in (3.1.4a).

**Proposition 3.1.6.** *Let  $[w_0, w_1, u_0] \in \mathcal{D}(\mathcal{A})$ . Then, with reference to the first integral on the RHS of (3.1.15) involving  $K_f(u, p) = [\mathcal{B}, \Delta]u - [\mathcal{B}, \nabla]p$  on  $Q_f$ , see (3.1.4a), we have for all  $\epsilon > 0$ :*

$$\begin{aligned} & 2 \left| \int_0^t \int_{\Omega_f} [\mathcal{B}, \nabla]p \cdot \tilde{u} d\Omega_f ds \right| + 2 \left| \int_0^t \int_{\Gamma_s} p \tilde{\nu} \cdot \tilde{u} d\Gamma_s ds \right| \\ & \leq \epsilon C \int_0^t \|\nabla \tilde{u}\|_{\Omega_f}^2 d\Omega_f + C_{\beta, \epsilon} \| [w_0, w_1, u_0] \|_{\mathcal{D}(\mathcal{A})}^2. \end{aligned} \quad (3.1.17)$$

Using now estimate (3.1.16) of Proposition 3.1.5 and estimate (3.1.17) of Proposition 3.1.6 on the RHS of inequality (3.1.31), with  $K_f(u, p) = [\mathcal{B}, \Delta]u - [\mathcal{B}, \nabla]p$ , and recalling  $K_s(w) = [\mathcal{B}, \Delta]w$ , we finally obtain:

**Theorem 3.1.7.** *Let  $[w_0, w_1, u_0] \in \mathcal{D}(\mathcal{A})$ ,  $w_0 \in (H^2(\Omega_s))^d$ . Then, with reference to estimate (3.1.15), we obtain for  $0 \leq t \leq T$ :*

$$\begin{aligned} & \|\tilde{w}(t)\|_{(H^1(\Omega_s))^d}^2 + \|\tilde{w}_t(t)\|_{(L_2(\Omega_s))^d}^2 + \|\tilde{u}(t)\|_{L_2(\Omega_f))^d}^2 \\ & + (2 - C\epsilon) \int_0^t \int_{\Omega_f} |\nabla \tilde{u}(t)|^2 d\Omega_f ds \\ & \leq 2 \left| \int_0^t \int_{\Omega_f} [\mathcal{B}, \Delta]u \cdot \tilde{u} d\Omega_f ds \right| + 2 \left| \int_0^t \int_{\Omega_s} [\mathcal{B}, \Delta]w \cdot \tilde{w}_t d\Omega_s ds \right| \\ & + \epsilon \int_0^t \|\tilde{w}\|_{(H^1(\Omega_s))^d}^2 ds + C_{T, \epsilon} \left( \| [w_0, w_1, u_0] \|_{\mathcal{D}(\mathcal{A})}^2 + \| w_0 \|_{(H^2(\Omega_s))^d}^2 \right). \end{aligned} \quad (3.1.18)$$

### 3.2. Analysis of the commutator terms $[\mathcal{B}, \Delta]u$ and $[\mathcal{B}, \Delta]w$ in the half-space

**Orientation.** The commutator  $[\mathcal{B}, \Delta]$ , which appears on the RHS of estimate (3.1.18) of Theorem 3.1.7, as acting on  $u$  or on  $w$ , is of order  $1+2-1=2$ . When acting as  $([\mathcal{B}, \Delta]u, \tilde{u})_{Q_f}$ , we can give a gross analysis as follows. *A priori*, we have  $u \in L_2(0, T; (H^1(\Omega_f))^d)$  via (1.29), for  $[w_0, w_1, u_0] \in \mathcal{H}$  (or  $u \in C([0, T]; (H^1(\Omega_f))^d)$  via (1.31b) for  $[w_0, w_1, u_0] \in \mathcal{D}(\mathcal{A})$ ). Thus  $(*) : [\mathcal{B}, \Delta]u \in L_2(0, T; (H^{-1}(\Omega_f))^d)$  for the second-order operator  $[\mathcal{B}, \Delta]$ . But this regularity  $(*)$  of  $[\mathcal{B}, \Delta]u$  is *not enough* to handle in  $([\mathcal{B}, \Delta]u, \tilde{u})_{Q_f}$  the regularity of  $\tilde{u} \in L_2(0, T; (H^1(\Omega_f))^d)$ , which would be required to have the term  $\epsilon \int_0^T \int_{\Omega_f} |\nabla \tilde{u}|^2 d\Omega_f dt$  – which would arise in estimating the RHS of (3.1.18) – absorbed by the term  $(2 - C\epsilon) \int_0^T \int_{\Omega_f} |\nabla \tilde{u}|^2 d\Omega_f dt$ , which is present on the LHS of (3.1.18). This gross analysis is even more inadequate while considering the term  $(K_s(w), \tilde{w}_t)_{Q_s} \equiv ([\mathcal{B}, \Delta]w, \tilde{w}_t)_{Q_s}$ . *A priori*,  $w \in C([0, T]; (H^1(\Omega_s))^d)$ , hence  $[\mathcal{B}, \Delta]w \in C([0, T]; (H^{-1}(\Omega_s))^d)$ , while the velocity term  $\tilde{w}_t$  is required to be in  $C([0, T]; (L_2(\Omega_s))^d)$ , in order to have the term

$\epsilon \|\tilde{w}_t\|_{C([0,T];(L_2(\Omega_s))^d)}^2$  which would arise in estimating the RHS of (3.1.18) absorbed by the term  $1\|\tilde{w}_t\|_{C([0,T];(L_2(\Omega_f))}$  which comes from the LHS of (3.1.18). Accordingly, a more refined analysis of the commutator terms is needed. This will be carried out in the half-space, where it will be more transparent and precise. In particular, this more refined analysis will permit us to see that the counterpart of the commutator  $[\mathcal{B}, \Delta]$  in the half-space is, yes, a second-order operator, *but only in the tangential direction*: this latter feature will then be instrumental in obtaining the sought-after energy estimate for  $\{\tilde{w}, \tilde{w}_t, \tilde{u}\}$  from estimate (3.1.18).

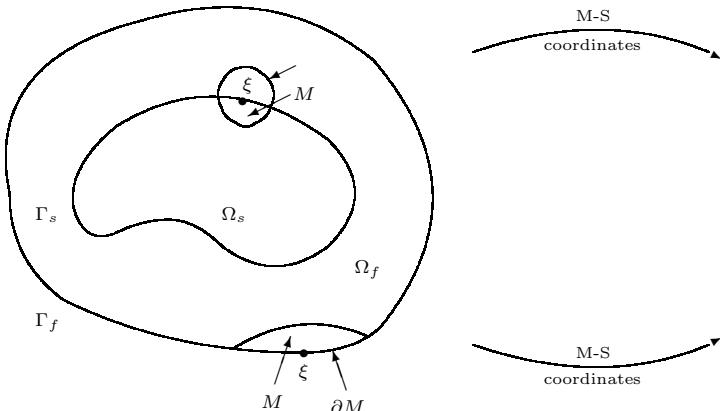
**Reduction to Melrose-Sjöstrand coordinates over a collar domain.** As  $\Delta = \frac{\partial^2}{\partial \xi_1^2} + \dots + \frac{\partial^2}{\partial \xi_d^2}$  in problem (1.1) or (3.1.3) over the original domain  $\Omega = \Omega_f \cup \Omega_s$  is a second-order differential operator on  $\Omega$  with real (principal) symbol  $-(\zeta_1^2 + \dots + \zeta_d^2)$  and with non-characteristic boundary, then near any point  $\xi \in \Gamma_s$ , respectively,  $\xi \in \Gamma_f$ , we may choose [M-S.1, pp. 597–598] local coordinates  $(x, y)$ ,  $x \in \mathbb{R}^1$ ,  $y = [y_1, \dots, y_{d-1}] \in \mathbb{R}^{d-1}$ , centered at  $\xi$ , such that  $\Omega_s$  is locally given by  $-1 \leq x < 0$ ,  $|y| \leq 1$ , and  $\Omega_f$  is locally given by  $0 \leq x < 1$ ,  $|y| \leq 1$  in the first case  $\xi \in \Gamma_s$ ; while  $\Omega_f$  is given locally by  $0 \leq x \leq 1$ ,  $|y| \leq 1$  in the second case  $\xi \in \Gamma_f$ . Moreover, the Laplacian  $\Delta$  is replaced by

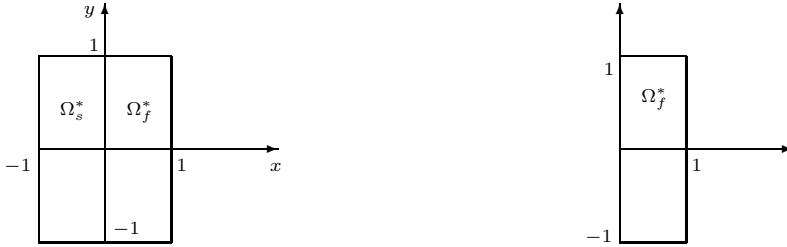
$$\hat{\Delta} = D_x^2 + \rho(x, y) \cdot D_y^2 + (\text{l.o.t. in } D_y) = D_x^2 + \sum_{|\alpha|=2} \rho_\alpha(x, y) D_y^\alpha + \text{l.o.t.}, \quad (3.2.1a)$$

where our notation is as follows:

$$D_x = \frac{\partial}{\partial x}; \quad \nabla_y = \left[ \frac{\partial}{\partial y_1}, \dots, \frac{\partial}{\partial y_{d-1}} \right]; \quad D_y^\alpha = \frac{\partial^{\alpha_1}}{\partial y_1} \cdots \frac{\partial^{\alpha_{d-1}}}{\partial y_{d-1}}, \quad \alpha = (\alpha_1, \dots, \alpha_{d-1}), \quad (3.2.1b)$$

with  $\rho(x, y)$  a vector real and smooth. Also, the term “l.o.t.” denotes here a differential operator of at most first order in  $y$ .





We also recall that passage, under the aforementioned diffeomorphism, from  $\Omega$  to  $\mathbb{R}_d^+$  preserves the norms and the inner products [L-M.1, p. 35]. Thus, henceforth, we may consider problem (3.1.3a–g) as defined in the collar domain

$$\Omega^* = \Omega_f^* \cup \Omega_s^*; \quad \Omega_s^* = \{(x, y) : -1 < x < 0; |y| < 1\},$$

$$\Omega_f^* = \{(x, y) : 0 < x < 1; |y| < 1\}, \quad (3.2.2)$$

where  $\Delta$  is replaced by  $\hat{\Delta}$  as given in  $\Omega^*$  by (3.2.1a) and the vector  $\rho(x, y)$  is real and smooth on the closure  $\text{cl}(\Omega^*)$  of  $\Omega^*$ . Such a new problem over  $\Omega^*$  may be viewed as corresponding to the original problem (3.1.3), defined, however, only over a boundary (collar) subdomain  $M$  of  $\Omega$  and acting on the solution  $\{w, u\}$  having compact support on  $\partial M$  in the case  $\xi \in \Gamma_s$ , and on the internal part  $\partial M \cap \Omega_f$  of  $\partial\Omega$  in the case  $\xi \in \Gamma_f$ , after the change of coordinates  $\xi = (\xi_1, \dots, \xi_d) \in M \rightarrow (x, y) \rightarrow \Omega^*$ .

Consequently, the new problem over  $\Omega^*$  with  $\hat{\Delta}$  given here by (3.2.1a) may be considered for solution  $\{w, u\}$  vanishing as follows:

$$\begin{aligned} \text{for the case } \xi \in \Gamma_s : w \text{ has compact support for } x = -1 \text{ and for } |y| = 1, \\ u \text{ has compact support for } x = 1 \text{ and for } |y| = 1; \end{aligned} \quad (3.2.3)$$

$$\text{for the case } \xi \in \Gamma_f : u \text{ has compact support for } x = 1 \text{ and for } |y| = 1. \quad (3.2.4)$$

As finitely many subdomains such as  $M$  will cover the full collar of  $\Gamma = \Gamma_f \cup \Gamma_s$ , estimates obtained for the new problem over  $\Omega^*$  will provide corresponding estimates of the original problem, in part by the invocation of a partition of unity.

After applying said partition of unity, it will suffice to consider a boundary layer of  $\Omega$ . Indeed, a partition corresponding to the interior of  $\Omega$  will be mapped onto a compactly supported set, in which case, both wave and fluid are decoupled – as the boundary conditions imposed on both wave and fluid are zero (Dirichlet B.C.). For these, standard regularity theory for both wave and fluid then applies.

**The commutator  $[\mathcal{B}, \Delta]$  on the half-space.** In both cases,  $\xi \in \Gamma_s$  and  $\xi \in \Gamma_f$ , the first-order operator  $\mathcal{B}$  of (3.1.1), tangential on the boundaries  $\Gamma_s \cup \Gamma_f$ , may be thought of as the pre-image – under the diffeomorphism via partition of unity from  $\Omega_f$ , resp.,  $\Omega_s$ , into the half-space  $\mathbb{R}_+^d = \{(x, y) : x \in \mathbb{R}_+, y \in \mathbb{R}^{d-1}\}$  (resp.,  $\mathbb{R}_-^d =$

$\{(x, y) : x \in \mathbb{R}_-, y \in \mathbb{R}^{d-1}\}$ ) – of the tangential derivative  $D_{y_\ell}$ ,  $\ell = 1, \dots, d-1$ , on the boundary  $x \equiv 0$  on  $\mathbb{R}^d$  [L-L-T.1, footnote, p. 162]. Thus,

$$\begin{aligned} & \text{the commutator } [\mathcal{B}, \Delta] \text{ on } \Omega \text{ goes into} \\ & \text{the commutator } [D_{y_\ell}, \hat{\Delta}] \text{ on } \Omega^*. \end{aligned} \quad (3.2.5)$$

We can then compute explicitly such commutator: from (3.2.1), with

$$\ell = 1, \dots, d-1 \quad \text{and} \quad D_{y_\ell} = \frac{\partial}{\partial y_\ell},$$

we obtain

$$\begin{aligned} D_{y_\ell} \hat{\Delta} &= D_{y_\ell} [D_x^2 + \rho(x, y) \cdot D_y^2 + \ell.o.t.] \\ &= D_{y_\ell} \left[ D_x^2 + \sum_{|\alpha|=2} \rho_\alpha(x, y) D_y^\alpha + \ell.o.t. \right] \end{aligned} \quad (3.2.6a)$$

$$\begin{aligned} &= \left[ D_x^2 + \sum_{|\alpha|=2} \rho_\alpha(x, y) D_y^\alpha \right] D_{y_\ell} \\ &+ \sum_{|\alpha|=2} \frac{\partial \rho_\alpha(x, y)}{\partial y_\ell} D_y^\alpha + D_{y_\ell} [\ell.o.t.] \end{aligned} \quad (3.2.6b)$$

$$= [D_x^2 + \rho(x, y) \cdot D_y^2] D_{y_\ell} + \rho_{y_\ell}(x, y) \cdot D_y^2 + D_{y_\ell} [\ell.o.t.] \quad (3.2.6c)$$

$$= \hat{\Delta} D_{y_\ell} + \sum_{|\alpha|=2} \frac{\partial \rho_\alpha(x, y)}{\partial y_\ell} D_y^\alpha + [D_{y_\ell}, \ell.o.t.] \quad (3.2.6d)$$

$$= \hat{\Delta} D_{y_\ell} + \rho_{y_\ell}(x, y) \cdot D_y^2 + \ell.o.t. \quad (3.2.6e)$$

Thus

$$[\mathcal{B}, \Delta] \text{ in } \Omega \rightarrow [D_{y_\ell}, \hat{\Delta}] = \rho_y(x, y) \cdot D_y^2 + \ell.o.t. \text{ on } \Omega^* \quad (3.2.7a)$$

$$= \sum_{|\alpha|=2} \frac{\partial \rho_\alpha(x, y)}{\partial y_\ell} D_y^\alpha + \ell.o.t. \text{ on } \Omega^*, \quad (3.2.7b)$$

where  $\ell.o.t.$  denotes a differential operator of at most first order.

Thus, in  $\Omega^*$  the commutator in question is, yes, of second order, but *only in the tangential direction y*, a big advantage – as we shall see in the analysis below – over the gross assessment in the Orientation that  $[\mathcal{B}, \Delta]$  is of order  $1+2-1=2$  in all variables! Thus, estimate (3.1.18) of Theorem 3.1.7 is rewritten in the half-space via (3.2.7) as

**Theorem 3.2.1.** Let  $[w_0, w_1, u_0] \in \mathcal{D}(\mathcal{A})$ ,  $w_0 \in (H^2(\Omega_s^*))^d$ . Then, with  $\tilde{u} = \mathcal{B}u = \mathcal{B}_\ell u$  and  $\tilde{w} = \mathcal{B}w = \mathcal{B}_\ell w$  as in (3.1.2),

$$\begin{aligned} & \|\tilde{w}(t)\|_{1,\Omega_s^*}^2 + \|\tilde{w}_t(t)\|_{\Omega_s^*}^2 + \|\tilde{u}(t)\|_{\Omega_f^*}^2 + (2 - C\epsilon) \int_0^t \int_{\Omega_f^*} |\nabla \tilde{u}(s)|^2 d\Omega_f ds \\ & \leq 2 \left| \int_0^t \int_{\Omega_f^*} (D_{y_\ell} \rho) \cdot (D_y^2 u) \cdot \tilde{u} d\Omega_f^* ds \right| + 2 \left| \int_0^t \int_{\Omega_s^*} (D_{y_\ell} \rho) \cdot (D_y^2 w) \cdot \tilde{w}_t d\Omega_f^* ds \right| \\ & \quad + \epsilon \int_0^t \|\tilde{w}(s)\|_{1,\Omega_s^*}^2 ds + C_T \left( \| [w_0, w_1, u_0] \|_{\mathcal{D}(\mathcal{A})}^2 + \| w_0 \|_{2,\Omega_s^*}^2 \right). \end{aligned} \quad (3.2.8)$$

Proceeding with the analysis, we see that on the half-space, we have  $\tilde{u} = D_{y_\ell} u$ , hence for  $\ell = 1, \dots, d-1$ ,

$$|\nabla \tilde{u}|^2 = |D_x(D_{y_\ell} u)|^2 + |D_y(D_{y_\ell} u)|^2. \quad (3.2.9)$$

For the first integral over  $\int_0^t \int_{\Omega_f^*}$  on the RHS of (3.2.8), we estimate, since  $\rho$  is smooth:

$$\begin{aligned} & 2 \left| \int_0^t \int_{\Omega_f^*} (D_{y_\ell} \rho) \cdot (D_y^2 u) \cdot (D_{y_\ell} u) d\Omega_f^* ds \right| \\ & \leq \epsilon \int_0^t \int_{\Omega_f^*} |D_y^2 u|^2 d\Omega_f^* ds + \frac{C}{\epsilon} \int_0^t \int_{\Omega_f^*} |D_{y_\ell} u|^2 d\Omega_f^* ds \end{aligned} \quad (3.2.10)$$

$$\leq \epsilon \int_0^t \int_{\Omega_f^*} |D_y^2 u|^2 d\Omega_f^* ds + \mathcal{O} \left\{ \| [w_0, w_1, u_0] \|_{\mathcal{D}(\mathcal{A})}^2 \right\}, \quad (3.2.11)$$

where in going from (3.2.10) to (3.2.11), we have invoked once more the *a priori* regularity for  $u$  in (1.31). Similarly, as to the second integral this time over  $\int_0^t \int_{\Omega_s^*}$  on the RHS of (3.2.8), we likewise estimate

$$\begin{aligned} & 2 \left| \int_0^t \int_{\Omega_s^*} (D_y \rho) \cdot (D_y^2 w) \cdot (D_{y_\ell} w_t) d\Omega_s^* ds \right| \\ & \leq \epsilon \int_0^t \int_{\Omega_s^*} |D_y^2 w|^2 d\Omega_s^* ds + \frac{C}{\epsilon} \int_0^t \int_{\Omega_s^*} |D_{y_\ell} w_t|^2 d\Omega_s^* ds \end{aligned} \quad (3.2.12)$$

$$\leq \epsilon \int_0^t \int_{\Omega_s^*} |D_y^2 w|^2 d\Omega_s^* ds + \mathcal{O} \left\{ \| [w_0, w_1, u_0] \|_{\mathcal{D}(\mathcal{A})}^2 \right\}, \quad (3.2.13)$$

recalling, in the last step, the regularity result  $D_{y_\ell} w_t \in C([0, T]; (L_2(\Omega_s^*))^d)$  from (3.31b). Invoking both (3.2.11) and (3.2.13) on the RHS of (3.2.8) yields via (3.2.9)

$$\begin{aligned} & \|D_{y_\ell} w(t)\|_{1,\Omega_s^*}^2 + \|D_{y_\ell} w_t(t)\|_{\Omega_s^*}^2 + \|D_{y_\ell} u(t)\|_{\Omega_f^*}^2 \\ & + (2 - C\epsilon) \int_0^t \int_{\Omega_f^*} [|D_y(D_{y_\ell} u)|^2 + |D_x(D_{y_\ell} u)|^2] d\Omega_f^* ds \\ & \leq \epsilon \int_0^t \int_{\Omega_f^*} |D_y^2 u|^2 d\Omega_f^* ds + \epsilon \int_0^t \int_{\Omega_s^*} |D_y^2 w|^2 d\Omega_s^* ds + \epsilon \int_0^t \|D_{y_\ell} w(s)\|_{1,\Omega_s^*}^2 ds \\ & + C_T \left( \| [w_0, w_1, u_0] \|_{\mathcal{D}(\mathcal{A})}^2 + \| w_0 \|_{2,\Omega_s^*}^2 \right), \quad \ell = 1, \dots, d-1. \end{aligned} \quad (3.2.14a)$$

This estimate is valid for all  $\ell = 1, \dots, d-1$ . Summing up these  $d-1$  estimates we thereby obtain for all  $0 \leq t \leq T$ ,

$$\begin{aligned} & \|D_y w(t)\|_{1,\Omega_s^*}^2 + \|D_y w_t(t)\|_{\Omega_s^*}^2 + \|D_y u(t)\|_{\Omega_f^*}^2 \\ & + (2 - C\epsilon) \int_0^t \int_{\Omega_f^*} [|D_y^2 u|^2 + |D_x(D_y u)|^2] d\Omega_f^* ds \\ & \leq (d-1)\epsilon \int_0^t \int_{\Omega_f^*} |D_y^2 u|^2 d\Omega_f^* ds + (d-1)\epsilon \int_0^t \int_{\Omega_s^*} |D_y^2 w|^2 d\Omega_f^* ds \\ & + (d-1)\epsilon \int_0^t \|D_y w(s)\|_{1,\Omega_s^*}^2 ds + C_T \left( \| [w_0, w_1, u_0] \|_{\mathcal{D}(\mathcal{A})}^2 + \| w_0 \|_{2,\Omega_s^*}^2 \right). \end{aligned} \quad (3.2.14b)$$

But the first term  $(d-1)\epsilon \int_0^t \int_{\Omega_f^*} |D_y^2 u|^2 d\Omega_f^* ds$  on the RHS of (3.2.14b) is absorbed by the corresponding term on the LHS of (3.2.14b) with factor  $(2 - C\epsilon)$ ; moreover, the second term in  $w$  on the RHS of (3.2.14b) is absorbed by the stronger third term. Thus, (3.2.14) yields (after possibly rescaling  $\epsilon > 0$ ):

$$\begin{aligned} & \|D_y w(t)\|_{1,\Omega_s^*}^2 + \|D_y w_t(t)\|_{\Omega_s^*}^2 + \|D_y u(t)\|_{\Omega_f^*}^2 \\ & + (2 - C\epsilon) \int_0^t \int_{\Omega_f^*} [|D_y^2 u|^2 + |D_x(D_y u)|^2] d\Omega_f^* ds \\ & \leq 2\epsilon \int_0^t \|D_y w\|_{1,\Omega_s^*}^2 ds + C_T \left( \| [w_0, w_1, u_0] \|_{\mathcal{D}(\mathcal{A})}^2 + \| w_0 \|_{2,\Omega_s^*}^2 \right). \end{aligned} \quad (3.2.15)$$

Next, setting

$$v(t) \equiv \|D_y w(t)\|_{1,\Omega_s^*}^2; \quad \mathcal{K}_T = C_T \left( \| [w_0, w_1, u_0] \|_{\mathcal{D}(\mathcal{A})}^2 + \| w_0 \|_{2,\Omega_s^*}^2 \right), \quad (3.2.16)$$

we obtain from (3.2.15) dropping three positive terms

$$v(t) \leq \mathcal{K}_T + 2\epsilon \int_0^t v(s) ds, \quad 0 < t \leq T. \quad (3.2.17)$$

Then, Gronwall's inequality implies

$$v(t) \leq \mathcal{K}_T e^{2\epsilon t}, \text{ or for } 0 < t \leq T,$$

$$\|D_y w(t)\|_{1,\Omega_s^*}^2 \leq C_T \left( \| [w_0, w_1, u_0] \|_{\mathcal{D}(\mathcal{A})}^2 + \|w_0\|_{2,\Omega_s^*}^2 \right) e^{2\epsilon t}. \quad (3.2.18)$$

Using (3.2.18) on the RHS of estimate (3.2.15), then yields for  $0 < t \leq T$ :

$$\begin{aligned} & \|D_y w(t)\|_{1,\Omega_s^*}^2 \\ &+ \|D_y w_t(t)\|_{\Omega_s^*}^2 + \|D_y u(t)\|_{\Omega_f^*}^2 + (2 - C\epsilon) \int_0^t \int_{\Omega_f^*} |\nabla(D_y u)|^2 d\Omega_f^* ds \\ &\leq C_T \left( \| [w_0, w_1, u_0] \|_{\mathcal{D}(\mathcal{A})}^2 + \|w_0\|_{2,\Omega_f}^2 \right). \end{aligned} \quad (3.2.19)$$

The counterpart of estimate (3.2.19), this time on the original domain  $\Omega = \Omega_s \cup \Omega_f$  via the diffeomorphism into  $\mathbb{R}^d$ , in terms of the variables  $\tilde{w} = \mathcal{B}w$ ,  $\tilde{w}_t = \mathcal{B}w_t$ ,  $\tilde{u} = \mathcal{B}u$  in (3.1.2), where  $\mathcal{B} = \mathcal{B}_\ell$ ,  $\ell = 1, \dots, d-1$ , is then, still for  $0 < t \leq T$ : For  $\ell = 1, \dots, d-1$ ,

$$\begin{aligned} & \|\tilde{w}(t)\|_{1,\Omega_s}^2 + \|\tilde{w}_t(t)\|_{\Omega_s}^2 + \|\tilde{u}(t)\|_{\Omega_f}^2 + (2 - C\epsilon) \int_0^t \int_{\Omega_f} |\nabla \tilde{u}|^2 d\Omega_f ds \\ &\leq C_T \left( \| [w_0, w_1, u_0] \|_{\mathcal{D}(\mathcal{A})}^2 + \|w_0\|_{2,\Omega_f}^2 \right). \end{aligned} \quad (3.2.20)$$

Then, (3.2.20) proves estimate (3.1.5a–b) of Theorem 3.1.1, save for the pressure term. To handle also  $\tilde{p} = \mathcal{B}p$ , see (3.1.2): In part by a classic result in Stokes theory (e.g., [C-F.1], Prop. 1.7(ii), p. 7) and appropriate usage of the slashed fluid equation (3.1.3a), we can eventually obtain as in [A-L-T.1], the chain

$$\begin{aligned} & \|\tilde{p}\|_{L_2(0,T;L_2(\Omega_f))} \\ &\leq C_1 \left[ \|\nabla \tilde{p}\|_{L_2(0,T;(L_2(\Omega_f))^d)} + \| [w_0, w_1, u_0] \|_{\mathcal{D}(\mathcal{A})} \right] \\ &\leq C_2 \left[ \|\nabla \tilde{u}\|_{L_2(0,T;(L_2(\Omega_f))^{d \times d})} + \| [w_0, w_1, u_0] \|_{\mathcal{D}(\mathcal{A})} \right]. \end{aligned} \quad (3.2.21)$$

This estimate, combined with that in (3.2.20) will now provide the full statement of Theorem 3.1.1, upon summing up in the index  $\ell$ .

## 4. Sketch of proof of Theorem 2.1

### 4.1. Boosting the regularity for the structural component $w$ :

#### Proof of Theorem 2.1(a)

Having established Theorem 3.1, we could then return to the original domain  $\Omega_s$  and carry out on  $\Omega_s$  the arguments of the present subsection. However, alternatively, it may still be more transparent to continue to work on  $\Omega_s^*$ . The analysis of

Section 3 has established that for  $[w_0, w_1, u_0] \in \mathcal{D}(\mathcal{A})$ ,  $w_0 \in (H^2(\Omega_s^*))^d$ , then

$$\nabla_y w(t, x, y) \in L_\infty(0, T; (H^1(\Omega_s^*))^{(d-1) \times d}), \quad (4.1.1a)$$

with continuous dependence on the I.C.; that is

$$\|\nabla_y w\|_{L_\infty(0, T; (H^1(\Omega_s^*))^{(d-1) \times d})}^2 \leq C_T \left( \| [w_0, w_1, u_0] \|_{\mathcal{D}(\mathcal{A})}^2 + \| w_0 \|_{(H^2(\Omega_s^*))^d}^2 \right). \quad (4.1.1b)$$

Hence, by trace theory on  $\Gamma_s^*$ :

$$\nabla_y w(t, 0, y) \in L_\infty(0, T; (H^{\frac{1}{2}}(\Gamma_s^*))^{d \times d}), \quad (4.1.2)$$

that is,

$$w(t, 0, y) \in L_\infty(0, T; (H^{\frac{3}{2}}(\Gamma_s^*))^d), \quad (4.1.3a)$$

$$\| w|_{\Gamma_s^*} \|_{L_\infty(0, T; (H^{\frac{3}{2}}(\Gamma_s^*))^d)} \leq C_T \left( \| [w_0, w_1, u_0] \|_{\mathcal{D}(\mathcal{A})}^2 + \| w_0 \|_{(H^2(\Omega_s^*))^d}^2 \right). \quad (4.1.3b)$$

In terms of the original domain  $\Omega_s$ , (4.1.3) says

$$\| w|_{\Gamma_s} \|_{L_\infty(0, T; (H^{\frac{3}{2}}(\Gamma_s))^d)} \leq C_T \left( \| [w_0, w_1, u_0] \|_{\mathcal{D}(\mathcal{A})}^2 + \| w_0 \|_{(H^2(\Omega_s))^d}^2 \right), \quad (4.1.4)$$

a result that could be shown directly on  $\Omega_s$ , using  $\tilde{w} = \mathcal{B}w$ ,  $\mathcal{B} = \mathcal{B}_\ell$ ,  $\ell = 1, \dots, d-1$ ,  $\mathcal{B}$  a tangential operator on  $\Gamma_s$ . Moreover, the *a priori* regularity (1.34) gives:  $[w_0, w_1, u_0] \in \mathcal{D}(\mathcal{A}) \Rightarrow w_{tt} \in C([0, T]; (L_2(\Omega_s))^d)$ . This, combined with (4.1.4) and *a-fortiori* (1.27) for  $w$ , yields then

$$\begin{aligned} & \| w|_{\Gamma_s} \|_{L_\infty(0, T; (H^{\frac{3}{2}}(\Gamma_s))^d)} + \| w_{tt} + w \|_{C([0, T]; (L_2(\Omega_s))^d)} \\ & \leq C_T \left( \| [w_0, w_1, u_0] \|_{\mathcal{D}(\mathcal{A})} + \| w_0 \|_{2, \Omega_s} \right). \end{aligned} \quad (4.1.5)$$

Thus, we have that, pointwise in time,  $w(t)$  is the solution of the elliptic BVP

$$\begin{cases} \Delta w = w_{tt}(t) + w(t) \in C([0, T]; (L_2(\Omega_s))^d); \\ w = w(t)|_{\Gamma_s} \in L_\infty(0, T; (H^{\frac{3}{2}}(\Gamma_s))^d), \end{cases} \quad (4.1.6a)$$

from which

$$w \in L_\infty(0, T; (H^2(\Omega_s))^d) \quad (4.1.7)$$

now follows from classical elliptic theory. Thus, Theorem 2.1(a), Eqn. (2.1) is established.

#### 4.2. Boosting the regularity for fluid components $\{u, p\}$ :

##### Proof of Theorem 2.1(b)–(c)

In this subsection we shall present two approaches; one obtained in  $\Omega_f^*$  and one obtained in  $\Omega_f$ .

**An approach in  $\Omega_f^*$ .** Under the diffeomorphism, via partition of unity, from the original fluid equation mapped into  $\Omega_f^*$ , the original differential operators  $\Delta$  and  $\nabla$  (in the original variables  $(\xi_1, \dots, \xi_d)$ ), occurring in the original fluid equation (1.1a), are mapped respectively into the M-S form  $\hat{\Delta}$  given by (3.2.1a) and into

the operator  $\hat{\nabla} = [D_x, \rho(x, y)\nabla_y]$ ; see Appendix B of [A-L-T.1]. This way, the original pressure term  $\nabla p$  (in the original variables  $\xi_1, \dots, \xi_d$ ) is mapped into  $\hat{\nabla}p = [D_x p, \rho(x, y)\nabla_y p]$ , where  $\rho(x, y)$  is a smooth coefficient of  $x$  and  $y$ . Thus, in  $Q_f^* = (0, T) \times \Omega_f^*$ , the counterpart of Eqn. (1.1a), originally defined in  $Q_f$ , is now:

$$u_t - \hat{\Delta}u + \hat{\nabla}p = 0,$$

or

$$u_t - D_x^2 u - \sum_{|\alpha|=2} \rho_\alpha(x, y) D_y^\alpha u + \begin{bmatrix} D_x p \\ \rho(x, y) \nabla_y p \end{bmatrix} + \ell.o.t.(u) = 0. \quad (4.2.1)$$

*Step 1.* We shall first obtain the desired estimates for the terms  $D_x^2 u_2, \dots, D_x^2 u_d$ , where  $u = [u_1, \dots, u_d]$ . To this end, we put on one side all the quantities which have already been estimated in (3.2.19) on  $\Omega_f^*$ , ultimately in (3.1.5) of Theorem 3.1.1 in  $Q_f$ . We then re-write (4.2.1) as

$$D_x^2 u - \begin{bmatrix} D_x p \\ 0 \end{bmatrix} = u_t - \sum_{|\alpha|=2} \rho_\alpha(x, y) D_y^\alpha u + \begin{bmatrix} 0 \\ \rho(x, y) \nabla_y p \end{bmatrix} + \ell.o.t.(u) \equiv F. \quad (4.2.2)$$

Next, for initial conditions  $[w_0, w_1, u_0] \in D(\mathcal{A})$  with  $w_0 \in (H^2(\Omega_s))^d$ , the regularity of the forcing term  $F$  on the RHS of (4.2.2) is  $F \in L_2(0, T; (L_2(\Omega_f^*))^d)$ , continuously in the initial conditions:

$$\|F\|_{L_2(0, T; (L_2(\Omega_f^*))^d)} \leq C_T \left( \| [w_0, w_1, u_0] \|_{D(\mathcal{A})} + \| w_0 \|_{2, \Omega_s} \right). \quad (4.2.3)$$

Indeed, to establish (4.2.3), we recall that:  $u_t \in L_2(0, T; (H^1(\Omega_f))^d)$  by (1.36); and moreover from Theorem 3.1.1 (or estimates (3.2.19) and (3.2.21)),

$$\begin{aligned} & \|\nabla_y u\|_{L_2(0, T; (H^1(\Omega_f^*))^{(d-1) \times d})} + \|\nabla_y p\|_{L_2(0, T; (L_2(\Omega_f^*))^d)} \\ & \leq C_T \left( \| [w_0, w_1, u_0] \|_{D(\mathcal{A})} + \| w_0 \|_{2, \Omega_s} \right). \end{aligned}$$

These estimates collectively give (4.2.3). We now return to equation (4.2.2), where  $D_x^2 u = [D_x^2 u_1, \dots, D_x^2 u_d]^{\text{tr}}$ . Then, via (4.2.3) on  $F$ , we obtain the following regularity results:

$$\begin{aligned} (a) \quad & \|D_x^2 u_1 - D_x p\|_{L_2(0, T; L_2(\Omega_f^*))} \\ & \leq C_T \left( \| [w_0, w_1, u_0] \|_{D(\mathcal{A})} + \| w_0 \|_{2, \Omega_s} \right); \end{aligned} \quad (4.2.4)$$

$$\begin{aligned} (b) \quad & \|D_x^2 u_2\|_{L_2(0, T; L_2(\Omega_f^*))} + \dots + \|D_x^2 u_d\|_{L_2(0, T; L_2(\Omega_f^*))} \\ & \leq C_T \left( \| [w_0, w_1, u_0] \|_{D(\mathcal{A})} + \| w_0 \|_{2, \Omega_s} \right). \end{aligned} \quad (4.2.5)$$

*Step 2.* Here, to obtain a similar regularity for component  $D_x^2 u_1$ , we use the fact that  $u$  is solenoidal in  $\Omega_f$ ; see (1.1b):

$$\operatorname{div}(u) = \frac{\partial}{\partial \xi_1} u_1 + \cdots + \frac{\partial}{\partial \xi_d} u_d = 0 \quad \text{in } \Omega_f. \quad (4.2.6)$$

Under the diffeomorphism via partition of unity, the original term  $\operatorname{div}(u)$  (in the  $(\xi_1, \dots, \xi_d)$  variables) is mapped into the term

$$\widehat{\operatorname{div}}(u) = \frac{\partial}{\partial x} u_1 + \frac{\partial}{\partial y_2} u_2 + \cdots + \frac{\partial}{\partial y_{d-1}} u_d + \ell.o.t(u) \quad \text{in } \Omega_f^*,$$

where  $\ell.o.t(u)$  is of zero order; see Appendix B of [A-L-T.1]. Since  $\widehat{\operatorname{div}}(u) = 0$ , we then obtain

$$\frac{\partial}{\partial x} u_1 = - \left( \frac{\partial}{\partial y_1} u_2 + \cdots + \frac{\partial}{\partial y_{d-1}} u_d \right) + \ell.o.t(u) \quad \text{in } \Omega_f^*. \quad (4.2.7)$$

Differentiating both sides of (4.2.7) in  $x$  thus yields

$$D_x^2 u_1 = -D_x \left( \frac{\partial}{\partial y_1} u_2 + \cdots + \frac{\partial}{\partial y_{d-1}} u_d \right) + D_x[\ell.o.t(u)] \quad \text{in } \Omega_f^*. \quad (4.2.8)$$

So, for  $[w_0, w_1, u_0] \in D(\mathcal{A})$ ,  $w_0 \in (H^2(\Omega_s))^d$ , we have, upon combining (4.2.5) with the estimates (3.2.19) on  $\Omega_f^*$  (or Theorem 3.1.1 on  $\Omega_f$ ) and (1.29),

$$\begin{aligned} & \|D_x^2 u_1\|_{L_2(0,T;L_2(\Omega_f^*))} \\ & \leq C_T \left\{ \sum_{j=2}^d \|D_y u_j\|_{L_2(0,T;H^1(\Omega_f^*))} + \|\nabla u\|_{L_2(0,T;(L_2(\Omega_f^*))^{d \times d})} \right\} \end{aligned} \quad (4.2.9)$$

$$\leq C_T \left( \| [w_0, w_1, u_0] \|_{D(\mathcal{A})} + \| w_0 \|_{2,\Omega_s} \right). \quad (4.2.10)$$

Finally, returning to (4.2.4), and applying (4.2.10) thereto, we arrive at

$$\|D_x p\|_{L_2(0,T;L_2(\Omega_f^*))} \leq C_T \left( \| [w_0, w_1, u_0] \|_{D(\mathcal{A})} + \| w_0 \|_{2,\Omega_s} \right). \quad (4.2.11)$$

Now, (4.2.10) along with (4.2.5) provide the desired result for the term  $D_x^2 u$ :

$$\|D_x^2 u\|_{L_2(0,T;(L_2(\Omega_f^*))^d)} \leq C_T \left( \| [w_0, w_1, u_0] \|_{D(\mathcal{A})} + \| w_0 \|_{2,\Omega_s} \right). \quad (4.2.12)$$

Then (4.2.12) for the normal regularity and (3.1.5) for the tangential regularity of component  $u$  establish (2.2) of Theorem 2.1 for  $u$ . Finally, combining (4.2.11) for the normal derivative of  $p$ , and (3.1.5) for the tangential derivative of  $p$ , establishes (2.3) of Theorem 2.1 for the pressure term of system (1.1). Theorem 2.1 is proved.

An alternative approach of Step 2 above in (a collar of the boundary of  $\Omega_f$  may be given [A-L-T.1].

## 5. Uniform stability of the fluid-structure dynamics

### 5.1. The fluid-structure PDE model (1.1a)–(1.1g) under boundary feedback dissipation

Given the seemingly strong dissipation coming from the gradient fluid component of the PDE system (1.1a)–(1.1g) – viz., see the relation in (1.28) – one might be tempted to conjecture that solutions of this fluid-structure model surely decay to the zero state in longtime. However, for any given boundary interface  $\Gamma_s$  there will be at least one eigenvalue of  $\mathcal{A} : \mathcal{H} \rightarrow \mathcal{H}$  on the imaginary axis, thereby precluding the possibility of asymptotic decay. In particular, the point  $\lambda = 0$  will always be an eigenvalue, with readily identifiable one-dimensional eigenspace (see [A-T.1], also [A-T.4], where the spectral analysis is undertaken for the more physically relevant Stokes-Lamé system). For some structural geometries – e.g., when  $\Omega_s$  is a circle in  $\mathbb{R}^2$  or a sphere in  $\mathbb{R}^3$  – there are actually countably infinite eigenvalues on  $i\mathbb{R}$ .

In the best possible geometrical situation (from the point of view of stability); i.e., when  $\lambda = 0$  is the only eigenvalue of  $\mathcal{A}$  on  $i\mathbb{R}$ , a nonstandard argument is invoked in [A-T.1], which ultimately invokes the spectral criterion of [Ar-Ba] and [Ly-Ph], so as to infer strong stability for solutions of the fluid-dynamics PDE (1.1a)–(1.1g), for initial data  $\{w_0, w_1, u_0\} \in [\text{Null}(\mathcal{A})]^\perp$ . (Recall from Proposition 1.1 that the resolvent of  $\mathcal{A}$  is not a compact operator; so classic invariance principles *cannot* be appealed to in order to quickly infer strong decay of fluid-structure solutions.)

Here instead, we are concerned with a notion much stronger than that of asymptotic decay; namely, we are seeking a result of *uniform stability on all of  $\mathcal{H}$* , for solutions of the fluid-structure dynamics under appropriate dissipative feedback. From our remarks above, it is clear that an additional feedback mechanism will be necessary for such decay on all of the energy space  $\mathcal{H}$ .

In view of this necessity, we consider the following fluid structure PDE under additional Neumann boundary dissipation in one of the transmission conditions on  $\Gamma_s$ :

$$\text{PDEs } \left\{ \begin{array}{ll} u_t - \Delta u + \nabla p = 0 & \text{in } (0, T) \times \Omega_f \\ \operatorname{div} u = 0 & \text{in } (0, T) \times \Omega_f \\ w_{tt} - \Delta w + w = 0 & \text{in } (0, T) \times \Omega_s \end{array} \right. \quad (5.1.1)$$

$$\text{B.C. } \left\{ \begin{array}{ll} \frac{\partial u}{\partial \nu} = \frac{\partial w}{\partial \nu} + p\nu & \text{on } (0, T) \times \Gamma_s \\ u|_{\Gamma_f} = 0 & \text{on } (0, T) \times \Gamma_f \\ \left[ w_t - \frac{\partial w}{\partial \nu} \right]_{\Gamma_s} = u|_{\Gamma_s} & \text{on } (0, T) \times \Gamma_s \end{array} \right. \quad (5.1.2)$$

$$\text{I.C. } [w(0), w_t(0), u(0)] = [w_0, w_1, u_0] \in \mathcal{H} \quad (5.1.3)$$

In short, the boundary condition  $w_t|_{\Gamma_s} = u|_{\Gamma_s}$  of (1.1a)–(1.1g) is replaced by  $[w_t - \frac{\partial w}{\partial \nu}]|_{\Gamma_s} = u|_{\Gamma_s}$ , which is consistent with [S.1]. This latter expression, as we shall see, induces an additional structural dissipation.

Let now (feedback) operator  $\mathcal{A}_F : \mathcal{H} \rightarrow \mathcal{H}$  model the dynamics of (5.1.1)–(5.1.3). Then from our preceding remarks, we likewise infer that  $\mathcal{A}_F$  generates a  $C_0$ -semigroup  $\{e^{\mathcal{A}_F t}\}_{t \geq 0} \subset \mathcal{H}$  (as did the original fluid-structure model  $\mathcal{A}$ ; see [A-T.1], [A-T.2]). Thus, analogous to the regularity result in Theorem 1.2, we have the following continuous map for solutions of the PDE (5.1.1)–(5.1.3) (see [A-T.2]):

$$\begin{aligned} [w_0, w_1, u_0] &\in \mathcal{H} \\ \Rightarrow \left\{ [w, w_t, u], u, \frac{\partial w}{\partial \nu} \Big|_{\Gamma_s} \right\} &\in C([0, T]; \mathcal{H}) \times L^2(0, T; [H^1(\Omega_f)]^d \times L^2(\Gamma_s)). \end{aligned} \quad (5.1.4)$$

In particular, to justify the asserted  $L^2$ -in time regularity in (5.1.4), we can invoke a simple energy method, as was employed for relation (1.26), so as to have the dissipative relation for all  $0 \leq s \leq t < \infty$ ,

$$\left\| e^{\mathcal{A}_F t} \begin{bmatrix} w_0 \\ w_1 \\ u_0 \end{bmatrix} \right\|^2 = \left\| e^{\mathcal{A}_F s} \begin{bmatrix} w_0 \\ w_1 \\ u_0 \end{bmatrix} \right\|^2 - 2 \int_s^t \|\nabla u\|_{\Omega_f}^2 d\tau - 2 \int_s^t \left\| \frac{\partial w}{\partial \nu} \right\|_{\Gamma_s}^2 d\tau \quad (5.1.5)$$

cf., (1.28). Moreover, as was shown outright in [A-T.2],  $\sigma(\mathcal{A}_F) \cap i\mathbb{R} = \emptyset$ , and so by an appeal to said spectral criterion in [Ar-Ba], the semigroup  $\{e^{\mathcal{A}_F}\}_{t \geq 0}$  is strongly stable. Subsequently, we can turn our attention to the question of whether the Neumann dissipative term in (5.1.5) gives rise to exponential decay for the semigroup  $\{e^{\mathcal{A}_F t}\}_{t \geq 0}$ , or what is the same, solutions of the boundary damped fluid-structure PDE model (5.1.1)–(5.1.3). In particular, the property of exponential decay being satisfied means there exist positive constants  $C$  and  $\rho$ , say, such that for all  $t > 0$ ,

$$\|e^{\mathcal{A}_F t}\|_{\mathcal{L}(\mathcal{H})} \leq C e^{-\rho t}. \quad (5.1.6)$$

Given the dissipative relation (5.1.5), to establish the uniform decay estimate (5.1.6) for the fluid-structure model, it suffices from a classic argument (see [Bal]) to show the following upper bound for the energy, for some positive constant  $C_T$ :

$$\left\| e^{\mathcal{A}_F T} \begin{bmatrix} w_0 \\ w_1 \\ u_0 \end{bmatrix} \right\|_{\mathcal{H}}^2 \leq C_T \left( \int_0^T \|\nabla u\|_{\Omega_f}^2 dt + \int_0^T \left\| \frac{\partial w}{\partial \nu} \right\|_{\Gamma_s}^2 dt \right). \quad (5.1.7)$$

By means of establishing the *a priori* inequality (5.1.7) we have the following:

**Theorem 5.1.1.** (see [A-T.2]) *For given initial data  $[w_0, w_1, u_0] \in \mathcal{H}$ , the solution of the fluid-structure PDE (5.1.1)–(5.1.3) decays exponentially in time. That is to say, there exist positive constants  $C$  and  $\rho$  such that the solution  $[w, w_t, u]$  of (5.1.1)–(5.1.3) exhibits the decay rate*

$$\|[w(t), w_t(t), u(t)]\|_{\mathcal{H}} \leq C e^{-\rho t} \| [w_0, w_1, u_0] \|_{\mathcal{H}} \quad \text{for all } 0 \leq t \leq T. \quad (5.1.8)$$

The proof of this result in [A-T.2] involves, in part, a “multiplier method” which to some extent is a vector-valued version of that carried out for boundary-controlled (and scalar-valued) wave equations; see, e.g., [Tr.2], which follows the Lyapunov method-based papers [Ch],[Lag]. Note that a key feature of Theorem 5.1.1 is the validity of the decay rate (5.1.8) with *no* geometrical assumptions being imposed upon the boundary interface  $\Gamma_s$ . The “big gun” which allows for this generality is the following microlocal result, which provides for the treatment of (historically troublesome) boundary integrals involving the tangential derivative  $\partial w/\partial\tau$ , these occurring in the course of establishing (5.1.7), via said multiplier method:

**Lemma 5.1.2.** *(See [L-T.2]) Let  $\epsilon > 0$  be arbitrarily small. Let  $z$  solve an arbitrary second-order hyperbolic equation with smooth space-dependent coefficients on  $Q_T \equiv (0, T) \times \Omega$ , where  $\Omega \subset \mathbb{R}^n$  is a smooth bounded domain. Then if  $\Gamma_*$  is a smooth connected segment of boundary  $\partial\Omega$ , we have the estimate*

$$\begin{aligned} & \int_{\epsilon}^{T-\epsilon} \int_{\Gamma_*} \left( \frac{\partial z}{\partial \tau} \right)^2 dt d\partial\Omega \\ & \leq C_T \left( \int_0^T \int_{\Gamma_*} z_t^2 dt d\partial\Omega + \int_0^T \int_{\partial\Omega} \left( \frac{\partial z}{\partial \nu} \right)^2 dt d\partial\Omega + \|z\|_{H^{\frac{1}{2}+\epsilon_0}(Q_T)}^2 \right), \end{aligned} \quad (5.1.9)$$

where parameters  $\epsilon, \epsilon_0 > 0$  are arbitrarily small.

Applying this trace result to the vector-valued function  $\frac{\partial w}{\partial\tau}|$  at the tail end of our multiplier method, we eventually arrive at the preliminary estimate

$$\mathcal{E}(T) \leq C_T \left( \int_0^T \|\nabla u\|_{\Omega_f}^2 dt + \int_0^T \left\| \frac{\partial w}{\partial \nu} \right\|_{\Gamma_s}^2 dt \right) + \text{l.o.t.}(w, w_t), \quad (5.1.10)$$

where  $\text{l.o.t.}(w, w_t)$  denote “lower-order terms,” or measurements of  $\{w, w_t\}$  in a (spatial) topology lower than that of the finite energy  $\mathcal{H}$ . Subsequently, a compactness-uniqueness argument (by contradiction), which uses the classic Holmgren’s result for the uniqueness of the continuation, removes these polluting lower-order terms, so as to establish (5.1.7), and so then the estimate (5.1.8).

## 5.2. Boundary feedback stabilization of a related Stokes-Lamé fluid-structure PDE system

We further note here that one has the exact analogue of the exponential stability result Theorem 5.1.1 for the Stokes-Lamé version of the canonical model (5.1.1)–(5.1.3). This more physically relevant PDE model appears in the aforesaid monograph [Li.1], and subsequently in [D-G-H-L.1]. See also [A-T.4], where a well-posedness and strong stability analysis, akin to that of [A-T.1] undertaken for (1.1a)–(1.1g), is carried out for the Stokes-Lamé system. As for the PDE (5.1.1)–(5.1.3), the geometry on which the fluid-structure interaction evolves will be the union  $\Omega_f \cup \Omega_s$ , where again  $\Omega_f$  denotes the fluid domain, and  $\Omega_s$  the structure domain (see Figure 1).

In presenting this Stokes-Lamé fluid-structure model, it would behoove us to first recall the classical tensor operators which are invoked to mathematically describe the linear (Hookean) system of elasticity on the structural domain  $\Omega_s$  (see, e.g., [Ke]):

1. For  $\omega = [\omega_1, \dots, \omega_n]$ , the *strain tensor*  $\{\epsilon_{ij}\}$  is given by

$$\epsilon_{ij}(\omega) = \frac{1}{2} \left( \frac{\partial \omega_j}{\partial x_i} + \frac{\partial \omega_i}{\partial x_j} \right), \quad 1 \leq i, j \leq n. \quad (5.2.1)$$

2. Subsequently, the *stress tensor* is described by means of Hooke's Law:

$$\sigma_{ij}(\omega) = \lambda \left( \sum_{k=1}^n \epsilon_{kk}(\omega) \right) \delta_{ij} + 2\mu \epsilon_{ij}(\omega), \quad 1 \leq i, j \leq n, \quad (5.2.2)$$

where  $\lambda \geq 0$  and  $\mu > 0$  are the so-called *Lamé's coefficients* of the system. Moreover,  $\delta_{ij}$  denotes as usual the Kronecker delta; i.e.,  $\delta_{ij} = 1$  if  $i = j$  and  $\delta_{ij} = 0$  otherwise.

With the geometry  $\{\Omega_f, \Omega_s\}$  as described above (and again with unit normal  $\nu$  exterior to  $\Omega_f$ ), and the stress-strain relations defined in (5.2.1)–(5.2.2), we are now in a position to describe the fluid-structure interactive PDE which appears in [D-G-H-L.1] and which, as in the canonical model (5.1.1)–(5.1.3), manifests a boundary feedback dissipative term in one of the transmission conditions on  $\Gamma_s$ . In this case, the variables (fluid)  $u(t, x) = [u_1, u_2, \dots, u_d]$  and (structure)  $w(t, x) = [w_1, w_2, \dots, w_d]$  satisfy

$$\text{PDEs } \begin{cases} u_t - \nabla \cdot (\nabla u + \nabla u^T) + \nabla p = 0 & \text{in } (0, T) \times \Omega_f \\ \operatorname{div} u = 0 & \text{in } (0, T) \times \Omega_f \\ w_{tt} - \operatorname{div}(\sigma(w)) + w = 0 & \text{in } (0, T) \times \Omega_s \end{cases} \quad (5.2.3)$$

$$\text{B.C. } \begin{cases} (\nabla u + \nabla u^T) \cdot \nu = \sigma(w) \cdot \nu + p\nu & \text{on } (0, T) \times \Gamma_s \\ u|_{\Gamma_f} = 0 & \text{on } (0, T) \times \Gamma_f \\ [w_t - \sigma(w) \cdot \nu]|_{\Gamma_s} = u|_{\Gamma_s} & \text{on } (0, T) \times \Gamma_s \end{cases} \quad (5.2.4)$$

$$\text{I.C. } [w(0), w_t(0), u(0)] = [w_0, w_1, u_0] \in \mathcal{H}. \quad (5.2.5)$$

In particular, the original boundary transmission condition  $[w_t]|_{\Gamma_s} = u|_{\Gamma_s}$  of [A-T.4] is replaced by  $[w_t - \sigma(w) \cdot \nu]|_{\Gamma_s} = u|_{\Gamma_s}$ . It is shown in [A-T.4] that in the absence of Neumann feedback damping, a spectral pathology occurs, wholly analogous to that of the canonical model (1.1a)–(1.1g): namely,  $\lambda = 0$  is always an eigenvalue of the associated Stokes-Lamé generator, with the possibility of other (at most countably infinite) eigenvalues on  $i\mathbb{R}$ . That is to say, the inherent damping emanating from the symmetric fluid gradient is not enough to insure strong stabilization of both fluid and structure PDE components. However, the presence of the additional boundary term  $[\sigma(w) \cdot \nu]|_{\Gamma_s}$  on  $\Gamma_s$  removes the possibility of point spectrum on the

imaginary axis; consequently, as for the canonical model, one can quickly appeal to [Ar-Ba] so as to infer strong decay. But one should think the more striking result can be had; just as we showed for the canonical model (5.1.1)–(5.1.3), the corresponding solutions of the Stokes-Lamé PDE, with said extra Neumann dissipative feedback, should decay exponentially in time. Indeed, we have the following:

**Theorem 5.2.1.** (see [A-T.3]) *For given initial data  $[w_0, w_1, u_0] \in \mathcal{H}$ , the solution of the fluid-structure PDE (5.2.3)–(5.2.5) decays exponentially in time. That is to say, there exist positive constants  $C$  and  $\rho$  such that the solution  $[w, w_t, u]$  of (5.2.3)–(5.2.5) exhibits the decay rate*

$$\|[w(t), w_t(t), u(t)]\|_{\mathcal{H}} \leq Ce^{-\rho t} \|[w_0, w_1, u_0]\|_{\mathcal{H}} \quad \text{for all } 0 \leq t \leq T. \quad (5.2.6)$$

Our *modus operandi* in the paper [A-T.3] is very much as we detailed for the canonical (5.1.1)–(5.1.3): In [A-T.3] we invoke a multiplier method to establish the energy inequality needed for exponential stability, and which is wholly analogous to (5.1.7). Namely, if  $[w, w_t, u]$  solves the aforesaid Stokes-Lamé system with Neumann boundary dissipation, we must establish the following estimate to infer exponential decay:

$$\|[w(T), w_t(T), u(T)]\|_{\mathcal{H}}^2 \leq C_T \left( \int_0^T \|\nabla u + \nabla u^T\|_{\Omega_f}^2 dt + \int_0^T \|\sigma(w)\nu\|_{\Gamma_s}^2 dt \right). \quad (5.2.7)$$

In the work [A-T.3], the known energy identities for the Lamé system of elasticity are put to good use (see, e.g., [Al-Ko],[Be-La],[Ho]). At some point in the course of this method, there is the need to estimate  $\|D_\tau w\|_{L^2(0,T;L^2(\Gamma_s))}$ , similar to the situation we outlined above for (5.1.1)–(5.1.3). By way of dealing with this tangential gradient, we invoke the trace estimate in [Ho.2] for solutions of the system of elasticity, this estimate being the natural descendant of the wave equation estimate (5.1.10). Eventually, by the means we have sketched out, we reach a point in [A-T.3] at which we derive the following estimate:

$$\begin{aligned} & \|[w(T), w_t(T), u(T)]\|_{\mathcal{H}}^2 \\ & \leq C_T \left( \int_0^T \|\nabla u + \nabla u^T\|_{\Omega_f}^2 dt + \int_0^T \|\sigma(w)\nu\|_{\Gamma_s}^2 dt \right) + \ell.o.t(w, w_t), \end{aligned}$$

where again,  $\ell.o.t(w, w_t)$  denotes polluting lower-order terms. As we did for the proof of Theorem 5.1.1, we wish to complete the derivation of estimate (5.2.7) by invoking a compactness-uniqueness argument. To make the uniqueness part of this argument (by contradiction) work, we invoke the unique continuation result in [E-I-N-T] for systems of elasticity. (Note that compared to the classic Holmgren's theorem, the result in [E-I-N-T] is relatively “state of the art”.)

We close here with a brief announcement concerning nonlinear boundary stabilization of the Stokes-Lamé system, wherein the Neumann dissipative term is

subjected to specified nonlinearities. This work is currently in progress. To describe this problem, we let  $g_i(\cdot)$ ,  $i = 1, \dots, d$ , be functions on the real line which satisfy the following criteria:

- (H1) Each  $g_i(s)$  is continuous and monotone increasing;
- (H2)  $g_i(s)s > 0$  for  $s \neq 0$ ;
- (H3) For each  $i$ ,  $i = 1, \dots, d$ , there exist positive constants  $m_i$  and  $M_i$  such that the following inequality obtains for  $|s| > 1$ :

$$m_i s^2 \leq g_i(s)s \leq M_i s.$$

With these nonlinearities, we now define the map  $\mathcal{G}([\sigma(w) \cdot \nu]_{\Gamma_s})$ , by

$$\mathcal{G}([\sigma(w) \cdot \nu]_{\Gamma_s}) \equiv [g_1([\sigma(w) \cdot \nu]_1), \dots, g_d([\sigma(w) \cdot \nu]_d)], \quad (5.2.8)$$

where  $[\sigma(w) \cdot \nu]_i$  is the  $i^{th}$  component of the boundary term  $[\sigma(w) \cdot \nu]_{\Gamma_s}$ .

Subsequently, we consider the following nonlinear PDE

$$\text{PDEs } \left\{ \begin{array}{ll} u_t - \nabla \cdot (\nabla u + \nabla u^T) + \nabla p = 0 & \text{in } (0, T) \times \Omega_f \\ \operatorname{div}(u) = 0 & \text{in } (0, T) \times \Omega_f \\ w_{tt} - \operatorname{div}(\sigma(w)) + w = 0 & \text{in } (0, T) \times \Omega_s \end{array} \right. \quad (5.2.9)$$

$$\text{B.C. } \left\{ \begin{array}{ll} (\nabla u + \nabla u^T) \cdot \nu = \sigma(w) \cdot \nu + p\nu & \text{on } (0, T) \times \Gamma_s \\ u|_{\Gamma_f} = 0 & \text{on } (0, T) \times \Gamma_f \\ [w_t - \mathcal{G}([\sigma(w) \cdot \nu]_{\Gamma_s})]|_{\Gamma_s} = u|_{\Gamma_s} & \text{on } (0, T) \times \Gamma_s \end{array} \right. \quad (5.2.10)$$

$$\text{I.C. } [w(0), w_t(0), u(0)] = [w_0, w_1, u_0] \in \mathcal{H}. \quad (5.2.11)$$

In short, the boundary feedback of PDE (5.2.3)–(5.2.5) is now allowed to vary in a nonlinear fashion. In this connection, our main result is as follows:

**Theorem 5.2.2.** (see [A-L-T.2]). *If the assumptions (H1)–(H3) are in place, then there exists a  $T_0$  such the solution  $[w, w_t, u]$  of (5.2.9)–(5.2.11) satisfies the following rate of decay for time:*

$$\|[w(t), w_t(t), u(t)]\|_{\mathcal{H}} \leq \mathcal{S} \left( \frac{t}{T_0} - 1 \right) \quad \text{for } t > T_0,$$

where  $\lim_{t \rightarrow \infty} \mathcal{S}(t) = 0$ . Here,  $\mathcal{S}(t)$  is the solution of a first-order differential equation with initial data  $\mathcal{S}(0) = \|[w(0), w_t(0), u(0)]\|_{\mathcal{H}}$ . The (explicitly computable) coefficients of the ODE will depend upon the nonlinearities  $g_i$ ,  $i = 1, \dots, d$ .

The proof of this result in [A-L-T.2] is an adaptation of the well-known algorithm in [La-Ta], wherein it is shown how one can derive explicit rates of decay for semilinear wave equations under nonlinear boundary damping. The key feature here is that for many typical and nonlinearities  $g_i$ , one can actually compute outright the solution  $\mathcal{S}(t)$  of the given ODE, thereby yielding explicit rates of decay for the nonlinear fluid-structure dynamics (5.2.9)–(5.2.11).

## References

- [Al-Ko] F. Alabau and V. Komornik, Boundary observability, controllability and stabilization of linear elastodynamic systems”, in *Contemporary Mathematics* 209: *Optimization Methods in Partial Differential Equations*, Amer. Math. Soc., Providence, RI (1997), 1–8.
- [Ar-Ba] W. Arendt and C.J.K. Batty, Tauberian theorems and stability of one parameter semigroups, *Trans. Amer. Math. Soc.* 306(8) (1988), 837–852.
- [A-L-T.1] G. Avalos, I. Lasiecka and R. Triggiani, Higher regularity of a coupled parabolic-hyperbolic fluid-structure interactive system, invited paper in memory of J. L. Lions, special volume, *Georgian Mathematical Journal*, 15(3) (2008), 403–437
- [A-L-T.2] G. Avalos, I. Lasiecka and R. Triggiani, Uniform decay rates for solutions to a fluid-structure interactive PDE model with nonlinear boundary dissipation, preprint (2008).
- [A-T.1] G. Avalos and R. Triggiani, The coupled PDE system arising in fluid/structure interaction, Part I: Explicit semigroup generator and its spectral properties, *Contemporary Mathematics* 440: *Fluids and Waves*, American Mathematical Society, Providence, RI (2007), 15–54.
- [A-T.2] G. Avalos and R. Triggiani, Uniform stabilization of a coupled PDE system arising in fluid-structure interaction with boundary dissipation at the interface, *Discr. & Contin. Dynam. Systems*, 22(4) (2008), 817–833.
- [A-T.3] G. Avalos and R. Triggiani, Boundary feedback stabilization of a coupled parabolic-hyperbolic Stokes-Lamé PDE system, preprint (2008), submitted.
- [A-T.4] G. Avalos and R. Triggiani, Well-posedness and stability analysis of a coupled Stokes-Lamé system, as a PDE model of certain fluid-structure interactions, *Discr. & Cont. Dynam. Sys.*, to appear.
- [Bal] A.V. Balakrishnan, *Applied Functional Analysis and Applications*, Second Edition, Springer-Verlag, New York (1981).
- [Be-La] Assia Benabdallah and Irena Lasiecka, Exponential decay rates for a full von Karman system of dynamic thermoelasticity, *J. Diff. Eqns.*, 160 (2000), 51–93.
- [B-G-L-T.1] V. Barbu, Z. Grujić, I. Lasiecka, and A. Tuffaha, Smoothness of weak solutions to a nonlinear fluid-structure interaction model, preprint (2008), *Indiana Univ. Math. J.* 57(3) (2008), 1173–1207.
- [B-S.1] S. Brenner and L.R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York (1994).
- [Ch] G. Chen, A note on the boundary stabilization of the wave equation, *SIAM J. Control Optim.*, 19 (1981), 106–113.
- [C-R.1] H. Cohen and S.I. Rubinow, Some mathematical topics in biology, *Proc. Symp. on System Theory*, Polytechnic Press, New York (1965), 321–337.
- [C-F.1] P. Constantin and C. Foias, *Navier-Stokes Equations*, University of Chicago Press, Chicago, London 1989.

- [D-G-H-L.1] Q. Du, M.D. Gunzburger, L.S. Hou, and J. Lee, Analysis of a linear fluid-structure interaction problem, *Discr. & Contin. Dynam. Systems* 9(3) (2003), 633–650.
- [E-I-N-T] M. Eller, V. Isakov, G. Nakamura, and D. Tataru, Uniqueness and stability in the Cauchy Problem for Maxwell's and elasticity systems, Nonlinear Partial Differential Equations and their Applications, Collège de France, Vol. XIV (Paris, 1997/1998), *Stud. Math. Appl.* 31, North-Holland, Amsterdam (2002), 329–349.
- [G-P-V.1] G. Gentili, F. Podestá, E. Vesentini, *Lezioni di Geometria Differenziale*, Boringhieri, 1995.
- [Ho] M.A. Horn, Implications of sharp trace regularity results on boundary stabilization of the system of linear elasticity, *J. Math. Anal. Appl.* 223 (1998), 126–150.
- [Ho.2] M.A. Horn, Sharp trace regularity for the solutions of the equations of dynamic elasticity, *J. Math. Systems, Estim. Control* 8(2), 217–219.
- [Lag] J. Lagnese, Decay of solutions of wave equations in a bounded region with boundary dissipation, *J. Differential Equations* 50 (1983), 163–182.
- [L-L-T.1] I. Lasiecka, J.L. Lions and R. Triggiani, Non-homogeneous boundary value problems for second-order hyperbolic operators, *J. Math. Pure Appl.* 65 (1986), 149–192.
- [La-Ta] I. Lasiecka and D. Tataru, Uniform boundary stabilization of semilinear wave equations with nonlinear boundary damping, *Diff. Int. Eqns.* 6(3) 507–533.
- [L-T.1] I. Lasiecka and R. Triggiani, *Control Theory for Partial Differential Equations I*, Cambridge University Press, New York (2000).
- [L-T.2] I. Lasiecka and R. Triggiani, Uniform stabilization of the wave equation with Dirichlet or Neumann feedback control without geometrical conditions, *Appl. Math. Optim.* 25 (1992), 189–224.
- [Li.1] J.L. Lions, *Quelques Méthodes de Résolution des Problèmes aux Limites Nonlinéaires*, Dunod-Gauthier-Villars, 1969.
- [L-M.1] J.L. Lions and E. Magenes, *Non-homogeneous Boundary Value Problems and Applications*, Vol. 1, Springer-Verlag, 1972.
- [Ly-Ph] Y.I. Lyubich and V.Q. Phong, Asymptotic stability of linear differential equations in Banach spaces, *Studia Mathematica*, LXXXVII (1988), 37–42.
- [Ke] S. Kesavan, *Topics in Functional Analysis and Applications*, John Wiley & Sons, New York (1989).
- [M-S.1] R. Melrose and J. Sjöstrand, Singularities of boundary value problems, *Comm. Pure Appl. Math.* 31 (1978), 593–617.
- [S.1] J. Serrin, *Mathematical Principles of Classical Fluid Mechanics*, 125–263.
- [Te.1] R. Teman, *Navier-Stokes Equations and Non-linear Functional Analysis*, North Holland, Amsterdam, 1978.
- [Th.1] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Lecture Notes in Mathematics, Vol. 1054, Springer, 1984.
- [Tr.1] R. Triggiani, Exponential feedback stabilization of a 2-D linearized Navier-Stokes channel flow by finite-dimensional, wall-normal boundary controllers

with arbitrarily small support, *Contin. & Discr. Dynam. Systems, B*, 8(2) (2007), 279–314.

- [Tr.2] R. Triggiani, Wave equation on a bounded domain with boundary dissipation: An operator approach, *J. Math. Anal. Appl.* 137 (1989), 438–461.

George Avalos

Department of Mathematics  
University of Nebraska-Lincoln  
Lincoln, NE 68588, USA  
e-mail: [gavalos@math.unl.edu](mailto:gavalos@math.unl.edu)

Irena Lasiecka and Roberto Triggiani  
Department of Mathematics  
University of Virginia  
Charlottesville, VA 22903, USA  
e-mail: [i12v@virginia.edu](mailto:i12v@virginia.edu)  
[rt7u@virginia.edu](mailto:rt7u@virginia.edu)

# A Continuous Adjoint Approach to Shape Optimization for Navier Stokes Flow

Christian Brandenburg, Florian Lindemann,  
Michael Ulbrich and Stefan Ulbrich

**Abstract.** In this paper we present an approach to shape optimization which is based on continuous adjoint computations. If the exact discrete adjoint equation is used, the resulting formula yields the exact discrete reduced gradient. We first introduce the adjoint-based shape derivative computation in a Banach space setting. This method is then applied to the instationary Navier-Stokes equations. Finally, we give some numerical results.

**Mathematics Subject Classification (2000).** Primary 76D55; Secondary 49K20.

**Keywords.** Shape optimization, Navier-Stokes equations, PDE-constrained optimization.

## 1. Introduction

In this paper, we consider the optimization of the shape of a body that is exposed to incompressible instationary Navier-Stokes flow in a channel. The developed techniques are quite general and can, without conceptual difficulties, be used to address a wide class of shape optimization problems with Navier-Stokes flow. The goal is to find the optimal shape of the body  $B$ , which is exposed to instationary incompressible fluid, with respect to some quantity of interest, e.g., drag, under constraints on the shape of  $B$ .

In a general setting, the shape optimization problem can be stated in the following way: Minimize an objective functional  $\bar{J}$ , depending on a domain  $\Omega$  and a state  $\tilde{y} = \tilde{y}(\Omega) \in Y(\Omega)$ . The domain  $\Omega$  is contained in a set of admissible domains  $\mathcal{O}_{\text{ad}}$ . Furthermore,  $\tilde{y}$  and  $\Omega$  are coupled by the *state equation*  $\bar{E}(\tilde{y}, \Omega) = 0$ . Thus,

---

We greatly acknowledge support of the Schwerpunktprogramm 1253 sponsored by the German Research Foundation.

the abstract shape optimization problem reads

$$\begin{aligned} \min \quad & \bar{J}(\tilde{y}, \Omega) \\ \text{s.t.} \quad & \bar{E}(\tilde{y}, \Omega) = 0, \quad \Omega \in \mathcal{O}_{\text{ad}}. \end{aligned}$$

The constraint  $\bar{E}(\tilde{y}, \Omega) = 0$  is a partial differential equation defined on  $\Omega$ , which in our case is given by the instationary incompressible Navier-Stokes equations.

Shape optimization is an important and active field of research with many engineering applications, especially in the areas of fluid dynamics and aerodynamics. Detailed accounts of the theory and applications of shape optimization can be found in, e.g., [1, 2, 3, 9, 15, 18]. We use the approach of transformation to a reference domain, as originally introduced by Murat and Simon [17], see also [8]. The domain is then fixed and the design is described by a transformation from a fixed domain to the domain  $\Omega$  corresponding to the current design. This makes optimal control techniques readily applicable. Furthermore, as observed by Guillaume and Masmoudi [8] in the context of linear elliptic equations, discretization and optimization can be made commutable. This means that, if certain guidelines are followed, then the discrete analogue of the continuous adjoint representation of the derivative of the reduced objective function is the exact derivative of the discrete reduced objective function. This allows to circumvent the tedious differentiation of finite element code with respect to the position of the vertices of the mesh. We apply this approach to shape optimization problems governed by the instationary Navier-Stokes equations. On one hand we characterize the appropriate function space for domain transformation in this framework. On the other hand, we focus on the practical implementation of shape optimization methods based on shape derivatives. We show that existing solvers for the state and adjoint equation on the current computational domain  $\Omega_k$  can be used to compute exact shape gradients conveniently for the continuous as well as for the discretized problem. Hence, although shape derivatives are defined through transformation to a reference domain, standard solvers on the transformed domain can be used for its computation.

The outline of this paper is as follows: In Section 2, we will present our approach for the derivative computation in shape optimization in a general setting. These general results will be applied to the instationary incompressible Navier-Stokes equations in Section 3. In Section 4 we present the discretization and stabilization techniques we use to solve the Navier-Stokes equations numerically, which are based on the  $cG(1)dG(0)$  variant of the  $G^2$ -finite-element discretization by Eriksson, Estep, Hansbo, Johnson and others [4, 5]. Moreover, we explain how we apply the adjoint calculus to obtain conveniently exact shape gradients on the discrete level. We will then present numerical results obtained for a model problem in Section 5, where we also briefly discuss the choice of shape transformations and parameterizations. Finally, in Section 6, we will give conclusions and an outlook to future work.

## 2. The shape optimization problem

In this section, we present the framework that we will use for shape derivative computation in a functional analytical setting. We first transform the general shape optimization problem, which is defined on varying domains, into a problem that is defined on a fixed reference domain  $\Omega_{\text{ref}}$ . Then, after introducing the reduced optimization problem on a space  $T$  of transformations of  $\Omega_{\text{ref}}$ , we state optimality conditions and an adjoint based representation for the reduced gradient of the objective function.

### 2.1. Problem formulation on a reference domain

We consider the abstract shape optimization problem given by

$$\begin{aligned} \min \quad & \bar{J}(\tilde{y}, \Omega) \\ \text{s.t.} \quad & \bar{E}(\tilde{y}, \Omega) = 0, \quad \Omega \in \mathcal{O}_{\text{ad}}. \end{aligned}$$

Here,  $\mathcal{O}_{\text{ad}}$  denotes the set of admissible domains  $\Omega \subset \mathbb{R}^d$ ,  $d = 2, 3$ ,  $\bar{J}$  is a real-valued objective function defined on a Banach space  $Y(\Omega)$  of functions defined on  $\Omega \subset \mathbb{R}^d$ ,

$$\bar{J} : \{(\tilde{y}, \Omega) : \tilde{y} \in Y(\Omega), \Omega \in \mathcal{O}_{\text{ad}}\} \rightarrow \mathbb{R},$$

and  $\bar{E}$  is an operator between function spaces  $Y(\Omega)$  and  $Z(\Omega)$  defined over  $\Omega$ ,

$$\bar{E} : \{(\tilde{y}, \Omega) : \tilde{y} \in Y(\Omega), \Omega \in \mathcal{O}_{\text{ad}}\} \rightarrow \{\tilde{z} : \tilde{z} \in Z(\Omega), \Omega \in \mathcal{O}_{\text{ad}}\}$$

with  $\bar{E}(\tilde{y}, \Omega) \in Z(\Omega)$  for all  $\tilde{y} \in Y(\Omega)$  and all  $\Omega \in \mathcal{O}_{\text{ad}}$ .

We now transform the shape optimization problem into a more convenient form. To this end, we consider a reference domain  $\Omega_{\text{ref}} \in \mathcal{O}_{\text{ad}}$  and interpret admissible domains  $\Omega \in \mathcal{O}_{\text{ad}}$  as images of  $\Omega_{\text{ref}}$  under suitable transformations. This is done by introducing a Banach space  $T(\Omega_{\text{ref}}) \subset \{\tau : \Omega_{\text{ref}} \rightarrow \mathbb{R}^d\}$  of bicontinuous transformations of  $\Omega_{\text{ref}}$ . We select a suitable subset  $T_{\text{ad}} \subset T(\Omega_{\text{ref}})$  of admissible transformations. The set  $\mathcal{O}_{\text{ad}}$  of admissible domains is then

$$\mathcal{O}_{\text{ad}} = \{\tau(\Omega_{\text{ref}}) : \tau \in T_{\text{ad}}\}.$$

We assume that

$$\left. \begin{aligned} Y(\Omega_{\text{ref}}) &= \{\tilde{y} \circ \tau : \tilde{y} \in Y(\tau(\Omega_{\text{ref}}))\} \\ \tilde{y} \in Y(\tau(\Omega_{\text{ref}})) &\mapsto y := \tilde{y} \circ \tau \in Y(\Omega_{\text{ref}}) \text{ is a homeomorphism} \end{aligned} \right\} \quad \forall \tau \in T_{\text{ad}}. \quad (\text{A})$$

Then, we can define the following equivalent optimization problem, which is entirely defined on the reference domain:

$$\begin{aligned} \min \quad & J(y, \tau) \\ \text{s.t.} \quad & E(y, \tau) = 0, \quad \tau \in T_{\text{ad}}. \end{aligned} \quad (2.1)$$

Here, the operator  $E : Y(\Omega_{\text{ref}}) \times T(\Omega_{\text{ref}}) \rightarrow Z(\Omega_{\text{ref}})$  is defined such that for all  $\tau \in T_{\text{ad}}$  and  $\tilde{y} \in Y(\tau(\Omega_{\text{ref}}))$  it holds that

$$E(y, \tau) = 0 \iff \bar{E}(\tilde{y}, \tau(\Omega_{\text{ref}})) = 0,$$

where  $y = \tilde{y} \circ \tau$ . The objective function  $J$  is defined in the same fashion.

In the following, we will consequently denote by  $\tilde{y}$  the functions on the physical domain  $\tau(\Omega_{\text{ref}})$  and by  $y$  the corresponding function on the reference domain  $\Omega_{\text{ref}}$ , where

$$y = \tilde{y} \circ \tau.$$

*Remark 2.1.* Let  $\Omega' \supset \bar{\Omega}_{\text{ref}}$  be open and bounded with Lipschitz boundary. If, which is the typical case for elliptic partial differential equations,

$$Y(\Omega) = H_0^1(\tau(\Omega_{\text{ref}})), \quad Z(\Omega) = H^{-1}(\tau(\Omega_{\text{ref}}))$$

with  $\tau \in T_{\text{ad}}$ , then (A) holds if we choose  $T(\Omega_{\text{ref}}) = W^{1,\infty}(\Omega')^d$  and if we require that all  $\tau \in T_{\text{ad}} \subset W^{1,\infty}(\Omega')^d$  are such that  $\tau : \bar{\Omega}_{\text{ref}} \rightarrow \tau(\bar{\Omega}_{\text{ref}})$  is a bi-Lipschitzian mapping.

For other spaces  $Y(\Omega)$  and  $Z(\Omega)$ , it can be necessary to impose further requirements on  $T_{\text{ad}}$ .

If  $\bar{E}$  is given in variational form then the operator  $E$  can be obtained by using the transformation rule for integrals. This will be carried out for the instationary Navier-Stokes equations in section 3.

## 2.2. Reduced problem and optimality conditions

In the following, we will consider the optimization problem on the reference domain (2.1), which has the form

$$\begin{aligned} \min \quad & J(y, \tau) \\ \text{s.t.} \quad & E(y, \tau) = 0, \quad \tau \in T_{\text{ad}}. \end{aligned} \tag{2.1}$$

We denote by  $E_y$  and  $E_\tau$  the partial derivatives of  $E$  with respect to  $y$  and  $\tau$ .

In order to derive first-order optimality conditions, we make the following assumptions:

- (A1)  $T_{\text{ad}} \subset T(\Omega_{\text{ref}})$  is nonempty, closed, convex and assumption (A) holds.
- (A2)  $J : Y(\Omega_{\text{ref}}) \times T(\Omega_{\text{ref}}) \rightarrow \mathbb{R}$  and  $E : Y(\Omega_{\text{ref}}) \times T(\Omega_{\text{ref}}) \rightarrow Z(\Omega_{\text{ref}})$  are continuously Fréchet-differentiable.
- (A3) There exists an open neighborhood  $T'_{\text{ad}} \subset T(\Omega_{\text{ref}})$  of  $T_{\text{ad}}$  and a unique solution operator  $S : T'_{\text{ad}} \rightarrow Y(\Omega_{\text{ref}})$ , assigning to each  $\tau \in T'_{\text{ad}}$  a unique  $y(\tau) \in Y(\Omega_{\text{ref}})$ , such that  $E(y(\tau), \tau) = 0$ .
- (A4) The derivative  $E_y(y(\tau), \tau) \in \mathcal{L}(Y(\Omega_{\text{ref}}), Z(\Omega_{\text{ref}}))$  is continuously invertible for all  $\tau \in T'_{\text{ad}}$ .

Under these assumptions  $y(\tau)$  is continuously differentiable on  $\tau \in T'_{\text{ad}} \supset T_{\text{ad}}$  by the implicit function theorem. Thus, it is reasonable to define the following reduced problem on the space of transformations  $T(\Omega_{\text{ref}})$ :

$$\begin{aligned} \min \quad & j(\tau) := J(y(\tau), \tau) \\ \text{s.t.} \quad & \tau \in T_{\text{ad}} \end{aligned},$$

where  $y(\tau)$  is given as the solution of  $E(y(\tau), \tau) = 0$ .

In the following we will use the abbreviations

$$T_{\text{ref}} := T(\Omega_{\text{ref}}), \quad Y_{\text{ref}} := Y(\Omega_{\text{ref}}), \quad Z_{\text{ref}} := Z(\Omega_{\text{ref}}).$$

In order to derive optimality conditions and to compute the reduced gradient  $j'(\tau)$ , we introduce the Lagrangian function  $\mathcal{L} : Y_{\text{ref}} \times T_{\text{ref}} \times Z_{\text{ref}}^* \rightarrow \mathbb{R}$ ,

$$\mathcal{L}(y, \tau, \lambda) := J(y, \tau) + \langle \lambda, E(y, \tau) \rangle_{Z_{\text{ref}}^*, Z_{\text{ref}}} ,$$

with Lagrange multiplier  $\lambda \in Z_{\text{ref}}^*$ .

Under assumptions (A1)–(A4) a local solution  $(y, \tau) \in Y_{\text{ref}} \times T_{\text{ad}}$  of (2.1) satisfies with an appropriate adjoint state  $\lambda \in Z_{\text{ref}}^*$  the following first-order necessary optimality conditions.

$$\mathcal{L}_\lambda(y, \tau, \lambda) = E(y, \tau) = 0 \quad (\text{state equation})$$

$$\mathcal{L}_y(y, \tau, \lambda) = J_y(y, \tau) + E_y^*(y, \tau)\lambda = 0 \quad (\text{adjoint equation})$$

$$\langle \mathcal{L}_\tau(y, \tau, \lambda), \tilde{\tau} - \tau \rangle_{T_{\text{ref}}^*, T_{\text{ref}}} = \langle J_\tau(y, \tau) + E_\tau^*(y, \tau)\lambda, \tilde{\tau} - \tau \rangle_{T_{\text{ref}}^*, T_{\text{ref}}} \geq 0 \quad \forall \tilde{\tau} \in T_{\text{ad}}.$$

**2.2.1. Adjoint-based shape derivative computation on the reference domain.** By using the adjoint equation the reduced gradient  $j'(\tau)$  can be determined as follows:

1. For given  $\tau$ , find  $y(\tau) \in Y_{\text{ref}}$  by solving the state equation

$$\langle E(y, \tau), \varphi \rangle_{Z_{\text{ref}}, Z_{\text{ref}}^*} = 0 \quad \forall \varphi \in Z_{\text{ref}}^*$$

2. Find the corresponding Lagrange multiplier  $\lambda \in Z_{\text{ref}}^*$  by solving the adjoint equation

$$\langle \lambda, E_y(y, \tau) \varphi \rangle_{Z_{\text{ref}}^*, Z_{\text{ref}}} = -\langle J_y(y, \tau), \varphi \rangle_{Y_{\text{ref}}^*, Y_{\text{ref}}} \quad \forall \varphi \in Y_{\text{ref}} \quad (2.2)$$

3. The reduced gradient with respect to  $\tau$  is now given by

$$\langle j'(\tau), \cdot \rangle_{T_{\text{ref}}^*, T_{\text{ref}}} = \langle \lambda, E_\tau(y, \tau) \cdot \rangle_{Z_{\text{ref}}^*, Z_{\text{ref}}} + \langle J_\tau(y, \tau), \cdot \rangle_{T_{\text{ref}}^*, T_{\text{ref}}}. \quad (2.3)$$

**2.2.2. Adjoint-based shape derivative computation on the physical domain.** For the application of optimization algorithms it is convenient to solve, for a given iterate  $\tau_k \in T(\Omega_{\text{ref}})$ , an equivalent representation of the optimization problem on the domain  $\Omega_k := \tau_k(\Omega_{\text{ref}})$ . To this end, we introduce operators  $\tilde{E}$ ,  $\tilde{J}$  and  $\tilde{j}$ , which differ from  $E$ ,  $J$  and  $j$  only in that the function spaces  $Y$ ,  $Z$  and  $T$  are defined on  $\Omega_k$  instead of  $\Omega_{\text{ref}}$ , i.e.,

$$\tilde{E}(\tilde{y}, \tilde{\tau}) = 0 \iff E(y, \tilde{\tau} \circ \tau_k) = 0, \text{ where } y = \tilde{y} \circ (\tilde{\tau} \circ \tau_k).$$

Then we have the relation

$$\tilde{j}(\tilde{\tau}) = j(\tilde{\tau} \circ \tau_k) = j(\tau) \quad \text{and therefore} \quad \tilde{\tau} \circ \tau_k = \tau, \text{ i.e., } \tilde{\tau} = \tau \circ \tau_k^{-1}.$$

We are thus led to the following procedure for computing the reduced gradient:

1. For  $\text{id} : \Omega_k \rightarrow \Omega_k$ ,  $\text{id}(\tau_k(x)) = \tau_k(x)$ ,  $x \in \Omega_{\text{ref}}$ , find  $\tilde{y}_k \in Y(\Omega_k)$  by solving the state equation

$$\langle \tilde{E}(\tilde{y}_k, \text{id}), \varphi \rangle_{Z(\Omega_k), Z(\Omega_k)^*} = 0 \quad \forall \varphi \in Z(\Omega_k)^*,$$

where  $\tilde{y}_k(\tau_k(x)) = y_k(x)$ ,  $x \in \Omega_{\text{ref}}$ . This corresponds to solving the standard state equation in variational form on the domain  $\Omega_k$ , which in the abstract setting was denoted by  $\bar{E}(\tilde{y}_k, \Omega_k) = 0$ .

2. Find the corresponding Lagrange multiplier  $\tilde{\lambda}_k \in Z(\Omega_k)^*$  by solving the adjoint equation

$$\langle \tilde{\lambda}_k, \tilde{E}_{\tilde{y}}(\tilde{y}_k, \text{id})\varphi \rangle_{Z(\Omega_k)^*, Z(\Omega_k)} = -\langle \tilde{J}_{\tilde{y}}(\tilde{y}_k, \text{id}), \varphi \rangle_{Y(\Omega_k)^*, Y(\Omega_k)} \quad \forall \varphi \in Y(\Omega_k),$$

where  $\tilde{\lambda}_k(\tau_k(x)) = \lambda_k(x)$ ,  $x \in \Omega_{\text{ref}}$ . This corresponds to the solution of the standard adjoint equation on  $\Omega_k$ .

3. The reduced gradient applied to  $V \in T(\Omega_{\text{ref}})$  is now given by

$$\begin{aligned} \langle j'(\tau_k), V \rangle_{T_{\text{ref}}^*, T_{\text{ref}}} &= \langle \tilde{j}'(\text{id}), \tilde{V} \rangle_{T(\Omega_k)^*, T(\Omega_k)} \\ &= \langle \tilde{\lambda}_k, \tilde{E}_{\tilde{\tau}}(\tilde{y}_k, \text{id})\tilde{V} \rangle_{Z(\Omega_k)^*, Z(\Omega_k)} \\ &\quad + \langle \tilde{J}_{\tilde{\tau}}(\tilde{y}_k, \text{id}), \tilde{V} \rangle_{T(\Omega_k)^*, T(\Omega_k)} \end{aligned}$$

for  $\tilde{V} \in T(\Omega_k)$ ,  $\tilde{V} \circ \tau_k = V$ , i.e.,  $\tilde{V} = V \circ \tau_k^{-1}$ . If we define the linear operator

$$B_k \in \mathcal{L}(T(\Omega_{\text{ref}}), T(\Omega_k)), \quad B_k V = V \circ \tau_k^{-1} \quad (2.4)$$

then we have by our previous calculation

$$j'(\tau_k) = B_k^* \tilde{j}'(\text{id}) = B_k^* (\tilde{E}_{\tilde{\tau}}(\tilde{y}_k, \text{id})^* \tilde{\lambda}_k + \tilde{J}_{\tilde{\tau}}(\tilde{y}_k, \text{id})).$$

This procedure yields the exact gradient of the reduced objective function and has the advantage that we are able to use standard PDE-solvers for the state equation and adjoint equation on the domain  $\Omega_k$ , since we evaluate at  $\tilde{\tau} = \text{id}$ .

### 2.3. Derivatives with respect to shape parameters

In practice, the shape of a domain is defined by design parameters  $u \in U$  with a finite- or infinite-dimensional design space  $U$ . Thus, we have a map  $\tau : U \rightarrow T(\Omega_{\text{ref}})$ ,  $u \mapsto \tau(u)$  and a reference control  $u_0 \in U$  with  $\tau(u_0) = \text{id}$ . Derivatives of the reduced objective function  $j(\tau(u))$  at  $u_k$  are obtained using the chain rule. With  $\tau_k = \tau(u_k)$  and  $B_k$  in (2.4) we have

$$\begin{aligned} \langle \frac{d}{du} j(\tau(u_k)), \cdot \rangle_{U^*, U} &= \langle j'(\tau(u_k)), \tau_u(u_k) \cdot \rangle_{T(\Omega_{\text{ref}})^*, T(\Omega_{\text{ref}})} \\ &= \langle \tilde{j}'(\text{id}), (\tau_u(u_k) \cdot) \circ \tau(u_k)^{-1} \rangle_{T(\Omega_k)^*, T(\Omega_k)} \\ &= \langle \tilde{j}'(\text{id}), B_k \tau_u(u_k) \cdot \rangle_{T(\Omega_k)^*, T(\Omega_k)} = \langle \tau_u(u_k)^* B_k^* \tilde{j}'(\text{id}), \cdot \rangle_{U^*, U}. \end{aligned}$$

Overall, this approach provides a flexible framework that can be used for arbitrary types of transformations (e.g., boundary displacements, free form deformation). The idea of using transformations to describe varying domains can be found, e.g., in Murat and Simon [17] and Guillaume and Masmoudi [8].

## 3. Shape optimization for the Navier-Stokes equations

We now apply this approach to shape optimization problems governed by the instationary Navier-Stokes equations for a viscous, incompressible fluid on a bounded domain  $\Omega = \tau(\Omega_{\text{ref}})$  with Lipschitz boundary. According to our convention, we will denote all quantities on the physical domain by  $\tilde{\cdot}$ .

For  $\Omega \subset \mathbb{R}^d$  with spatial dimension  $d = 2$  or  $3$ , let  $\Gamma_D \subset \partial\Omega$  be a nonempty Dirichlet boundary and  $\Gamma_N = \partial\Omega \setminus \Gamma_D$ . We consider the problem

$$\begin{aligned}\tilde{\mathbf{v}}_t - \nu \Delta \tilde{\mathbf{v}} + (\tilde{\mathbf{v}} \cdot \nabla) \tilde{\mathbf{v}} + \nabla \tilde{p} &= \tilde{\mathbf{f}} && \text{on } \Omega \times I \\ \operatorname{div} \tilde{\mathbf{v}} &= 0 && \text{on } \Omega \times I \\ \tilde{\mathbf{v}} &= \tilde{\mathbf{v}}_D && \text{on } \Gamma_D \times I \\ \tilde{p} \tilde{\mathbf{n}} - \nu \frac{\partial \tilde{\mathbf{v}}}{\partial \tilde{\mathbf{n}}} &= 0 && \text{on } \Gamma_N \times I \\ \tilde{\mathbf{v}}(\cdot, 0) &= \tilde{\mathbf{v}}_0 && \text{on } \Omega\end{aligned}$$

where  $\tilde{\mathbf{v}} : \Omega \times I \rightarrow \mathbb{R}^d$  denotes the velocity  $\tilde{\mathbf{v}}(x, t)$  and  $\tilde{p} : \Omega \times I \rightarrow \mathbb{R}$  the pressure  $\tilde{p}(x, t)$  of the fluid at a point  $x$  at time  $t$ ,  $\tilde{\mathbf{n}} : \partial\Omega \rightarrow \mathbb{R}^d$  is the outer unit normal. Here  $I = (0, T)$ ,  $T > 0$  is the time interval and  $\nu > 0$  is the kinematic viscosity; if the equations are written in dimensionless form,  $\nu$  can be interpreted as  $1/Re$  where  $Re$  is the Reynolds number.

We introduce the spaces

$$\begin{aligned}H_D^1(\Omega) &:= \{\tilde{\mathbf{v}} \in H^1(\Omega)^d : \tilde{\mathbf{v}}|_{\Gamma_D} = 0\}, \quad V := \{\tilde{\mathbf{v}} \in H_D^1(\Omega)^d : \operatorname{div} \tilde{\mathbf{v}} = 0\}, \\ H &:= \operatorname{cl}_{L^2}(V), \quad L_0^2(\Omega) := \{\tilde{p} \in L^2(\Omega) : \int_{\Omega} \tilde{p} = 0\},\end{aligned}$$

the corresponding Gelfand triple  $V \hookrightarrow H \hookrightarrow V^*$ , and define

$$W_{2,q}(I; V) := \{\tilde{\mathbf{v}} \in L^2(I; V) : \tilde{\mathbf{v}}_t \in L^q(I; V^*)\}.$$

Now let

$$\tilde{\mathbf{v}}_D \in H^1(\Omega), \quad \operatorname{div} \tilde{\mathbf{v}}_D = 0, \quad \tilde{\mathbf{f}} \in L^2(I; V^*), \quad \tilde{\mathbf{v}}_0 \in H.$$

Under these assumptions the following results are known.

- If  $\Gamma_D = \partial\Omega$ , i.e.,  $\Gamma_N = \emptyset$ , then for  $d = 2$  there exists a unique weak solution  $(\tilde{\mathbf{v}}, \tilde{p})$  with  $\tilde{\mathbf{v}} - \tilde{\mathbf{v}}_D \in W_{2,2}(I; V)$  and  $\tilde{p}(\cdot, t) \in L_0^2(\Omega)$ ,  $t \in I$ . For  $d = 3$  there exists a weak solution  $(\tilde{\mathbf{v}}, \tilde{p})$  with  $\tilde{\mathbf{v}} - \tilde{\mathbf{v}}_D \in W_{2,4/3}(I; V) \cap L^\infty(I; H)$  and  $\tilde{p}(\cdot, t) \in L_0^2(\Omega)$ ,  $t \in I$ , which is not necessarily unique. For the case  $\tilde{\mathbf{v}}_D = 0$  the proofs can be found for example in [19, Ch. III]. These proofs can be extended to  $\tilde{\mathbf{v}}_D \neq 0$  under the above assumptions on  $\tilde{\mathbf{v}}_D$ .
- If  $\Gamma_N \neq \emptyset$  and  $\Gamma_D$  satisfies some geometric properties (for example, all  $x \in \Omega$  can be connected in all coordinate directions by a line segment in  $\Omega$  to a point in  $\Gamma_D$ ) and if a sequence of Galerkin approximations exists that does not exhibit inflow on  $\Gamma_N$  then the same can be shown as for the Dirichlet case: For  $d = 2$  there exists a unique weak solution  $(\tilde{\mathbf{v}}, \tilde{p})$  with  $\tilde{\mathbf{v}} - \tilde{\mathbf{v}}_D \in W_{2,2}(I; V)$  and  $\tilde{p}(\cdot, t) \in L_0^2(\Omega)$ ,  $t \in I$ . For  $d = 3$  there exists a weak solution  $(\tilde{\mathbf{v}}, \tilde{p})$  with  $\tilde{\mathbf{v}} - \tilde{\mathbf{v}}_D \in W_{2,4/3}(I; V) \cap L^\infty(I; H)$  and  $\tilde{p}(\cdot, t) \in L_0^2(\Omega)$ ,  $t \in I$ , which is not necessarily unique. In fact, in the case without inflow on  $\Gamma_N$  all additional boundary terms have the correct sign such that the proofs in [19, Ch. III] for the Dirichlet case can be adapted.

In the case of possible inflow an existence and uniqueness result local in time and for small data global in time can be found in [10].

### 3.1. Weak formulation

In the following we consider the case  $d = 2$  and  $\Gamma_N = \emptyset$  to have a general global existence and uniqueness result at hand. Moreover, to avoid technicalities in formulating the equations, we consider homogeneous boundary data  $\tilde{\mathbf{v}}_D \equiv 0$ . Then we have  $H_D^1(\Omega) = H_0^1(\Omega)^d$  with the above notations.

The classical weak formulation is now: Find  $\tilde{\mathbf{v}} \in W_{2,2}(I; V)$  such that

$$\begin{aligned} & \langle \tilde{\mathbf{v}}_t(\cdot, t), \tilde{\mathbf{w}} \rangle_{V^*, V} + \int_{\Omega} \tilde{\mathbf{v}}(x, t)^T \nabla \tilde{\mathbf{v}}(x, t) \tilde{\mathbf{w}}(x) dx + \int_{\Omega} \nu \nabla \tilde{\mathbf{v}}(x, t) : \nabla \tilde{\mathbf{w}}(x) dx \\ &= \int_{\Omega} \tilde{\mathbf{f}}(x, t)^T \tilde{\mathbf{w}}(x) dx \quad \forall \tilde{\mathbf{w}} \in V \text{ for a.a. } t \in I \\ & \tilde{\mathbf{v}}(\cdot, 0) = \tilde{\mathbf{v}}_0. \end{aligned} \tag{3.1}$$

As mentioned above, for  $\tilde{\mathbf{f}} \in L^2(I; V^*)$  and  $\tilde{\mathbf{v}}_0 \in H$  there exists a unique weak solution  $\tilde{\mathbf{v}} \in W_{2,2}(I; V)$ . The pressure  $\tilde{p}(\cdot, t) \in L_0^2(\Omega)$ ,  $t \in I$ , is now uniquely determined, see [19, Ch. III].

The weak formulation (3.1) is equivalent to the following velocity-pressure formulation: Find  $\tilde{\mathbf{v}} \in W_{2,2}(I; H_0^1(\Omega)^d)$  and  $\tilde{p}(\cdot, t) \in L_0^2(\Omega)$ ,  $t \in I$ , such that

$$\begin{aligned} & \langle \tilde{\mathbf{v}}_t(\cdot, t), \tilde{\mathbf{w}} \rangle_{H^{-1}, H_0^1} + \int_{\Omega} \tilde{\mathbf{v}}(x, t)^T \nabla \tilde{\mathbf{v}}(x, t) \tilde{\mathbf{w}}(x) dx + \int_{\Omega} \nu \nabla \tilde{\mathbf{v}}(x, t) : \nabla \tilde{\mathbf{w}}(x) dx \\ & - \int_{\Omega} \tilde{p}(x, t) \operatorname{div} \tilde{\mathbf{w}}(x) dx = \int_{\Omega} \tilde{\mathbf{f}}(x, t)^T \tilde{\mathbf{w}}(x) dx \quad \forall \tilde{\mathbf{w}} \in H_0^1(\Omega) \text{ for a.a. } t \in I \\ & \int_{\Omega} \tilde{q}(x) \operatorname{div} \tilde{\mathbf{v}}(x, t) = 0 \quad \forall \tilde{q} \in L_0^2(\Omega) \text{ for a.a. } t \in I \\ & \tilde{\mathbf{v}}(\cdot, 0) = \tilde{\mathbf{v}}_0. \end{aligned}$$

To obtain a weak velocity-pressure formulation in space-time, which is convenient for adjoint calculations, we have to ensure that  $\tilde{p} \in L^2(I; L_0^2(\Omega))$ . To this end we assume that the data  $\tilde{\mathbf{f}}$  and  $\tilde{\mathbf{v}}_0$  are sufficiently regular, for example, see [19, Ch. III, Thm. 3.5],

$$\tilde{\mathbf{f}}, \tilde{\mathbf{f}}_t \in L^2(I; V^*), \tilde{\mathbf{f}}(\cdot, 0) \in H, \tilde{\mathbf{v}}_0 \in V \cap H^2(\Omega)^d. \tag{3.2}$$

Define the spaces

$$\begin{aligned} Y(\Omega) &:= W(I; H_0^1(\Omega)^d) \times L^2(I; L_0^2(\Omega)), \\ Z(\Omega) &:= L^2(I; H^{-1}(\Omega)^d) \times L^2(I; L_0^2(\Omega)). \end{aligned} \tag{3.3}$$

Then

$$Z^*(\Omega) := L^2(I; H_0^1(\Omega)^d) \times L^2(I; L_0^2(\Omega))$$

and the weak formulation (3.1) is equivalent to: Find  $(\tilde{\mathbf{v}}, \tilde{p}) \in Y(\Omega)$ , where  $\Omega = \tau(\Omega_{\text{ref}})$  such that

$$\begin{aligned} & \langle (\tilde{\mathbf{w}}, \tilde{q}), \bar{E}((\tilde{\mathbf{v}}, \tilde{p}), \Omega) \rangle_{Z^*(\Omega), Z(\Omega)} \\ &= \int_{\Omega} \tilde{\mathbf{v}}(x, 0)^T \tilde{\mathbf{w}}(x, 0) dx - \int_{\Omega} \tilde{\mathbf{v}}_0^T \tilde{\mathbf{w}}(\cdot, 0) dx \\ &+ \int_I \int_{\Omega} \tilde{\mathbf{v}}_t^T \tilde{\mathbf{w}} dx dt + \int_I \int_{\Omega} \nu \nabla \tilde{\mathbf{v}} : \nabla \tilde{\mathbf{w}} dx dt \\ &+ \int_I \int_{\Omega} \tilde{\mathbf{v}}^T \nabla \tilde{\mathbf{v}} \tilde{\mathbf{w}} dx dt - \int_I \int_{\Omega} \tilde{p} \operatorname{div} \tilde{\mathbf{w}} dx dt \\ &- \int_I \int_{\Omega} \tilde{\mathbf{f}}^T \tilde{\mathbf{w}} dx dt + \int_I \int_{\Omega} \tilde{q} \operatorname{div} \tilde{\mathbf{v}} dx dt = 0 \quad \forall (\tilde{\mathbf{w}}, \tilde{q}) \in Z^*(\Omega). \end{aligned} \quad (3.4)$$

This formulation defines now the state equation operator

$$\bar{E} : \{(\tilde{y}, \Omega) : \tilde{y} \in Y(\Omega), \Omega \in \mathcal{O}_{\text{ad}}\} \rightarrow \{\tilde{z} : \tilde{z} \in Z(\Omega), \Omega \in \mathcal{O}_{\text{ad}}\}.$$

### 3.2. Transformation to the reference domain

In the following we assume that

- (T)  $\Omega_{\text{ref}}$  is a bounded Lipschitz domain and  $\Omega' \supset \bar{\Omega}_{\text{ref}}$  is open and bounded with Lipschitz boundary. Moreover  $T_{\text{ad}} \subset W^{2,\infty}(\Omega')^d$  is bounded such that for all  $\tau \in T_{\text{ad}}$  the mappings  $\tau : \bar{\Omega}_{\text{ref}} \rightarrow \tau(\bar{\Omega}_{\text{ref}})$  are bi-Lipschitzian and satisfy  $\det(\tau') \geq \delta > 0$ , with a constant  $\delta > 0$ . Here,  $\tau'(x) = \nabla \tau(x)^T$  denotes the Jacobian of  $\tau$ .

Moreover, the data  $\tilde{\mathbf{v}}_0, \tilde{\mathbf{f}}$  are given such that

$$\tilde{\mathbf{f}} \in C^1(I; V(\Omega)), \quad \tilde{\mathbf{v}}_0 \in V(\Omega) \cap H^2(\Omega)^d \quad \forall \Omega \in \mathcal{O}_{\text{ad}} = \{\tau(\Omega_{\text{ref}}) : \tau \in T_{\text{ad}}\},$$

i.e., the data  $\tilde{\mathbf{v}}_0, \tilde{\mathbf{f}}_0$  are used on all  $\Omega \in \mathcal{O}_{\text{ad}}$ .

Then assumption (T) ensures in particular (3.2) and assumption (A) holds in the following obvious version for time dependent problems, where the transformation acts only in space.

**Lemma 3.1.** *Let  $T_{\text{ad}}$  satisfy assumption (T). Then the state space  $Y(\Omega)$  defined in (3.3) satisfies assumption (A), more precisely,*

$$\left. \begin{aligned} Y(\Omega_{\text{ref}}) &= \{(\tilde{\mathbf{v}}, \tilde{p})(\tau(\cdot), \cdot) : (\tilde{\mathbf{v}}, \tilde{p}) \in Y(\tau(\Omega_{\text{ref}}))\} \\ (\tilde{\mathbf{v}}, \tilde{p}) \in Y(\tau(\Omega_{\text{ref}})) &\mapsto (\mathbf{v}, p) := (\tilde{\mathbf{v}}, \tilde{p})(\tau(\cdot), \cdot) \in Y(\Omega_{\text{ref}}) \text{ is a homeom.} \end{aligned} \right\} \quad \forall \tau \in T_{\text{ad}}.$$

A proof of this result is beyond the scope of this paper and will be given elsewhere.

Given the weak formulation of the Navier-Stokes equations on a domain  $\tau(\Omega_{\text{ref}})$  we can apply the transformation rule for integrals to obtain a variational

formulation based on the domain  $\Omega_{\text{ref}}$ . Using our convention to write  $\sim$  for a function that is defined on  $\tau(\Omega_{\text{ref}})$  we use the identifications

$$\mathbf{v}(x, t) := \tilde{\mathbf{v}}(\tau(x), t), \quad p(x, t) := \tilde{p}(\tau(x), t),$$

etc. and the identity

$$\nabla_{\tilde{x}} \tilde{z}(\tau(x)) = \tau'(x)^{-T} \nabla_x z(x), \quad x \in \Omega_{\text{ref}}.$$

Using this formalism we get for example

$$\int_I \int_{\tau(\Omega_{\text{ref}})} \nu \nabla \tilde{\mathbf{v}} : \nabla \tilde{\mathbf{w}} \, dx \, dt = \int_I \sum_{i=1}^d \int_{\Omega_{\text{ref}}} \nu \nabla v_i^T \tau'^{-1} \tau'^{-T} \nabla w_i \det \tau' \, dx \, dt.$$

In this way and by using Lemma 3.1 we arrive at the following equivalent form of the weak formulation (3.4) on  $\tau(\Omega_{\text{ref}})$ , which is only based on the domain  $\Omega_{\text{ref}}$ : Find  $(\mathbf{v}, p) \in Y(\Omega_{\text{ref}})$  such that for all  $(\mathbf{w}, q) \in Z^*(\Omega_{\text{ref}})$

$$\begin{aligned} & \langle (\mathbf{w}, q), E((\mathbf{v}, p), \tau) \rangle_{Z^*(\Omega_{\text{ref}}), Z(\Omega_{\text{ref}})} \\ &= \int_{\Omega_{\text{ref}}} \mathbf{v}(x, 0)^T \mathbf{w}(x, 0) \det \tau' \, dx - \int_{\Omega_{\text{ref}}} \tilde{\mathbf{v}}_0(\tau(x))^T \mathbf{w}(x, 0) \det \tau' \, dx \\ &+ \int_I \int_{\Omega_{\text{ref}}} \mathbf{v}_t^T \mathbf{w} \det \tau' \, dx \, dt + \sum_{i=1}^d \int_I \int_{\Omega_{\text{ref}}} \nu \nabla v_i^T \tau'^{-1} \tau'^{-T} \nabla w_i \det \tau' \, dx \, dt \quad (3.5) \\ &+ \int_I \int_{\Omega_{\text{ref}}} \mathbf{v}^T \tau'^{-T} \nabla \mathbf{v} \, \mathbf{w} \det \tau' \, dx \, dt - \int_I \int_{\Omega_{\text{ref}}} p \, \text{tr}(\tau'^{-T} \nabla \mathbf{w}) \det \tau' \, dx \, dt \\ &- \int_I \int_{\Omega_{\text{ref}}} \tilde{\mathbf{f}}(\tau(x), t)^T \mathbf{w} \det \tau' \, dx \, dt + \int_I \int_{\Omega_{\text{ref}}} q \, \text{tr}(\tau'^{-T} \nabla \mathbf{v}) \det \tau' \, dx \, dt = 0. \end{aligned}$$

For  $\tau = \text{id}$  we recover directly the weak formulation (3.4) on the domain  $\Omega = \Omega_{\text{ref}}$ , for general  $\tau \in T_{\text{ad}}$  we obtain an equivalent form of (3.4) on the domain  $\Omega = \tau(\Omega_{\text{ref}})$ .

### 3.3. Objective function

We consider an objective functional  $\bar{J}$  defined on the domain  $\tau(\Omega_{\text{ref}})$  of the type

$$\begin{aligned} \bar{J}((\tilde{\mathbf{v}}, \tilde{p}), \tau(\Omega_{\text{ref}})) &= \int_I \int_{\tau(\Omega_{\text{ref}})} f_1(x, \tilde{\mathbf{v}}(x, t), \nabla \tilde{\mathbf{v}}(x, t), \tilde{p}(x, t)) \, dx \, dt \\ &+ \int_{\tau(\Omega_{\text{ref}})} f_2(x, \tilde{\mathbf{v}}(x, T)) \, dx \quad (3.6) \end{aligned}$$

with  $f_1 : \bigcup_{\tau \in T_{\text{ad}}} \tau(\Omega_{\text{ref}}) \times \mathbb{R}^2 \times \mathbb{R}^{2,2} \times \mathbb{R}$  and  $f_2 : \bigcup_{\tau \in T_{\text{ad}}} \tau(\Omega_{\text{ref}}) \times \mathbb{R}^2 \rightarrow \mathbb{R}$ . Again we transform the objective function to the reference domain  $\Omega_{\text{ref}}$ .

$$\begin{aligned} \bar{J}((\tilde{\mathbf{v}}, \tilde{p}), \tau(\Omega_{\text{ref}})) &= \int_I \int_{\Omega_{\text{ref}}} f_1(\tau(x), \mathbf{v}(x, t), \tau'(x)^{-T} \nabla \mathbf{v}(x, t), p(x, t)) \det \tau' \, dx \, dt \\ &\quad + \int_{\Omega_{\text{ref}}} f_2(\tau(x), \mathbf{v}(x, T)) \det \tau' \, dx \\ &=: J^D((\mathbf{v}, p), \tau) + J^T(\mathbf{v}, \tau) = J((\mathbf{v}, p), \tau). \end{aligned}$$

### 3.4. Adjoint equation

We apply now the adjoint procedure of Subsection 2.2.1 to compute the shape gradient. To this end, we have to compute the Lagrange multipliers  $(\boldsymbol{\lambda}, \mu) \in Z^*(\Omega_{\text{ref}})$  by solving the adjoint system (2.2), which reads in this case

$$\begin{aligned} \langle (\boldsymbol{\lambda}, \mu), E_{(\mathbf{v}, p)}((\mathbf{v}, p), \tau)(\mathbf{w}, q) \rangle_{Z^*(\Omega_{\text{ref}}), Z(\Omega_{\text{ref}})} \\ = -\langle J_{(\mathbf{v}, p)}((\mathbf{v}, p), \tau), (\mathbf{w}, q) \rangle_{Y^*(\Omega_{\text{ref}}), Y(\Omega_{\text{ref}})} \quad \forall (\mathbf{w}, q) \in Y(\Omega_{\text{ref}}) \end{aligned}$$

with the given weak solution  $(\mathbf{v}, p) \in Y(\Omega_{\text{ref}})$  of the state equation. In detail we seek  $(\boldsymbol{\lambda}, \mu) \in Z^*(\Omega_{\text{ref}})$  with

$$\begin{aligned} &- \int_I \int_{\Omega_{\text{ref}}} \mathbf{w}^T \boldsymbol{\lambda}_t \det \tau' \, dt \, dx + \int_{\Omega_{\text{ref}}} \mathbf{w}(x, T)^T \boldsymbol{\lambda}(x, T) \det \tau' \, dx \\ &+ \int_I \sum_{i=1}^d \int_{\Omega_{\text{ref}}} \nu \nabla w_i^T \tau'^{-1} \tau'^{-T} \nabla (\boldsymbol{\lambda})_i \det \tau' \, dx \, dt \\ &+ \int_I \int_{\Omega_{\text{ref}}} (\mathbf{w}^T \tau'^{-T} \nabla \mathbf{v} + \mathbf{v}^T \tau'^{-T} \nabla \mathbf{w}) \boldsymbol{\lambda} \det \tau' \, dx \, dt \\ &- \int_I \int_{\Omega_{\text{ref}}} q \operatorname{tr}(\tau'^{-T} \nabla \boldsymbol{\lambda}) \det \tau' \, dx \, dt + \int_I \int_{\Omega_{\text{ref}}} \mu \operatorname{tr}(\tau'^{-T} \nabla \mathbf{w}) \det \tau' \, dx \, dt \\ &= -\langle J_{(\mathbf{v}, p)}((\mathbf{v}, p), \tau), (\mathbf{w}, q) \rangle_{Y^*(\Omega_{\text{ref}}), Y(\Omega_{\text{ref}})} \quad \forall (\mathbf{w}, q) \in Y(\Omega_{\text{ref}}). \end{aligned} \tag{3.7}$$

For  $\tau = \text{id}$  this is the weak formulation of the usual adjoint system of the Navier-Stokes equations on  $\Omega_{\text{ref}}$ , which reads in strong form

$$\begin{aligned} -\boldsymbol{\lambda}_t - \nu \Delta \boldsymbol{\lambda} - (\nabla \boldsymbol{\lambda})^T \mathbf{v} + (\nabla \mathbf{v}) \boldsymbol{\lambda} - \nabla \mu &= -J_{\mathbf{v}}^D((\mathbf{v}, p), \text{id}) && \text{on } \Omega_{\text{ref}} \times I \\ -\operatorname{div} \boldsymbol{\lambda} &= -J_p^D((\mathbf{v}, p), \text{id}) && \text{on } \Omega_{\text{ref}} \times I \\ \boldsymbol{\lambda} &= 0 && \text{on } \partial \Omega_{\text{ref}} \times I \\ \boldsymbol{\lambda}(\cdot, T) &= -J_{\mathbf{v}}^T(\mathbf{v}, \text{id}) && \text{on } \Omega_{\text{ref}} \end{aligned}$$

For general  $\tau \in T_{\text{ad}}$  the adjoint system (3.7) is equivalent to the usual adjoint system of the Navier-Stokes equations on  $\tau(\Omega_{\text{ref}})$ . A detailed analysis of the adjoint equation of the Navier-Stokes equations can be found in [11, 21].

### 3.5. Calculation of the shape gradient

The derivative of the reduced objective  $j(\tau) := J((\mathbf{v}(\tau), p(\tau)), \tau)$  is now given by (2.3), which reads in our case

$$\begin{aligned} & \langle j'(\tau), \cdot \rangle_{T^*(\Omega_{\text{ref}}), T(\Omega_{\text{ref}})} \\ &= \langle (\boldsymbol{\lambda}, \mu), E_\tau((\mathbf{v}, p), \tau) \cdot \rangle_{Z^*(\Omega_{\text{ref}}), Z(\Omega_{\text{ref}})} + \langle J_\tau((\mathbf{v}, p), \tau), \cdot \rangle_{T^*(\Omega_{\text{ref}}), T(\Omega_{\text{ref}})}. \end{aligned} \quad (3.8)$$

To state this in detail, we have to compute the derivatives of  $E$  and  $J$  with respect to  $\tau$ . Let  $(\mathbf{v}, p)$  and  $(\boldsymbol{\lambda}, \mu)$  be the solution of the Navier-Stokes equations (3.5) and the corresponding adjoint equation (3.7) for given  $\tau \in T_{\text{ad}}$ . Using the formulation (3.5) of  $E$  on the reference domain  $\Omega_{\text{ref}}$  the first term can be expressed as

$$\begin{aligned} & \langle (\boldsymbol{\lambda}, \mu), E_\tau((\mathbf{v}, p), \tau) V \rangle_{Z^*(\Omega_{\text{ref}}), Z(\Omega_{\text{ref}})} \\ &= \int_{\Omega_{\text{ref}}} (\mathbf{v}(x, 0) - \tilde{\mathbf{v}}_0(\tau(x)))^T \boldsymbol{\lambda}(x, 0) \operatorname{tr}(\tau'^{-1} V') \det \tau' dx \\ &\quad - \int_{\Omega_{\text{ref}}} V^T \nabla \tilde{\mathbf{v}}_0(\tau(x)) \boldsymbol{\lambda}(x, 0) \det \tau' dx + \int_I \int_{\Omega_{\text{ref}}} \mathbf{v}_t^T \boldsymbol{\lambda} \operatorname{tr}(\tau'^{-1} V') \det \tau' dx dt \\ &\quad + \sum_{i=1}^d \int_I \int_{\Omega_{\text{ref}}} \nu \nabla v_i^T \tau'^{-1} (\operatorname{tr}(\tau'^{-1} V') I - V' \tau'^{-1} - \tau'^{-T} V'^T) \tau'^{-T} \nabla \boldsymbol{\lambda}_i \det \tau' dx dt \\ &\quad + \int_I \int_{\Omega_{\text{ref}}} \mathbf{v}^T (\operatorname{tr}(\tau'^{-1} V') I - \tau'^{-T} V'^T) \tau'^{-T} \nabla \mathbf{v} \boldsymbol{\lambda} \det \tau' dx dt \\ &\quad + \int_I \int_{\Omega_{\text{ref}}} p (\operatorname{tr}(\tau'^{-T} V'^T \tau'^{-T} \nabla \boldsymbol{\lambda}) - \operatorname{tr}(\tau'^{-T} \nabla \boldsymbol{\lambda}) \operatorname{tr}(\tau'^{-1} V')) \det \tau' dx dt \\ &\quad - \int_I \int_{\Omega_{\text{ref}}} \left( \tilde{\mathbf{f}}(\tau(x), t)^T \operatorname{tr}(\tau'^{-1} V') + V^T \nabla \tilde{\mathbf{f}}(\tau(x), t) \right) \boldsymbol{\lambda} \det \tau' dx dt \\ &\quad - \int_I \int_{\Omega_{\text{ref}}} \mu (\operatorname{tr}(\tau'^{-T} V'^T \tau'^{-T} \nabla \mathbf{v}) - \operatorname{tr}(\tau'^{-T} \nabla \mathbf{v}) \operatorname{tr}(\tau'^{-1} V')) \det \tau' dx dt. \end{aligned}$$

If  $\tilde{\mathbf{v}}$  solves the state equation then the first term vanishes.

The part with the objective functional is given by

$$\begin{aligned} & \langle J_\tau((\mathbf{v}, p), \tau), V \rangle_{T^*(\Omega_{\text{ref}}), T(\Omega_{\text{ref}})} \\ &= \int_I \int_{\Omega_{\text{ref}}} f_1(\tau(x), \mathbf{v}, \tau'(x)^{-T} \nabla \mathbf{v}(x, t), p) \operatorname{tr}(\tau'^{-1} V') \det \tau' dx dt \\ &\quad - \int_I \int_{\Omega_{\text{ref}}} \frac{\partial}{\partial(\nabla \tilde{\mathbf{v}})} f_1(\tau(x), \mathbf{v}, \tau'(x)^{-T} \nabla \mathbf{v}(x, t), p) \tau'^{-T} V'^T \tau'^{-T} \nabla \mathbf{v} \det \tau' dx dt \\ &\quad + \int_I \int_{\Omega_{\text{ref}}} \frac{\partial}{\partial \tilde{x}} f_1(\tau(x), \mathbf{v}, \tau'(x)^{-T} \nabla \mathbf{v}(x, t), p) V \det \tau' dx dt \\ &\quad + \int_{\Omega_{\text{ref}}} f_2(\tau(x), \mathbf{v}(x, T)) \operatorname{tr}(\tau'^{-1} V') \det \tau' dx \\ &\quad + \int_{\Omega_{\text{ref}}} \frac{\partial}{\partial \tilde{x}} f_2(\tau(x), \mathbf{v}(x, T)) V \det \tau' dx. \end{aligned}$$

If  $\tau = \text{id}$ , i.e.,  $\tau(\Omega_{\text{ref}}) = \Omega_{\text{ref}}$  we obtain the following formula for the reduced gradient, where we omit the first term, since  $\mathbf{v}(\cdot, 0) = \tilde{\mathbf{v}}_0(\tau(\cdot))$ .

$$\begin{aligned}
& \langle j'(\text{id}), V \rangle_{T^*(\Omega_{\text{ref}}), T(\Omega_{\text{ref}})} \\
&= - \int_{\Omega_{\text{ref}}} V^T \nabla \tilde{\mathbf{v}}_0(x) \boldsymbol{\lambda}(x, 0) \, dx \\
&\quad + \int_I \int_{\Omega_{\text{ref}}} \mathbf{v}_t^T \boldsymbol{\lambda} \operatorname{div} V \, dx \, dt + \int_I \int_{\Omega_{\text{ref}}} \nu \nabla \mathbf{v} : \nabla \boldsymbol{\lambda} \operatorname{div} V \, dx \, dt \\
&\quad - \sum_{i=1}^d \int_I \int_{\Omega_{\text{ref}}} \nu \nabla \mathbf{v}_i^T (V' + V'^T) \nabla \boldsymbol{\lambda}_i \, dx \, dt \\
&\quad - \int_I \int_{\Omega_{\text{ref}}} \mathbf{v}^T V'^T \nabla \mathbf{v} \boldsymbol{\lambda} \, dx \, dt + \int_I \int_{\Omega_{\text{ref}}} \mathbf{v}^T \nabla \mathbf{v} \boldsymbol{\lambda} \operatorname{div} V \, dx \, dt \\
&\quad + \int_I \int_{\Omega_{\text{ref}}} p \operatorname{tr}(V'^T \nabla \boldsymbol{\lambda}) \, dx \, dt - \int_I \int_{\Omega_{\text{ref}}} p \operatorname{div} \boldsymbol{\lambda} \operatorname{div} V \, dx \, dt \\
&\quad - \int_I \int_{\Omega_{\text{ref}}} \tilde{\mathbf{f}}^T \boldsymbol{\lambda} \operatorname{div} V \, dx \, dt - \int_I \int_{\Omega_{\text{ref}}} V^T \nabla \tilde{\mathbf{f}} \boldsymbol{\lambda} \, dx \, dt \\
&\quad - \int_I \int_{\Omega_{\text{ref}}} \mu \operatorname{tr}(V'^T \nabla \mathbf{v}) \, dx \, dt + \int_I \int_{\Omega_{\text{ref}}} \mu \operatorname{div} \mathbf{v} \operatorname{div} V \, dx \, dt \\
&\quad + \int_I \int_{\Omega_{\text{ref}}} f_1(x, \mathbf{v}, \nabla \mathbf{v}, p) \operatorname{div} V \, dx \, dt \\
&\quad - \int_I \int_{\Omega_{\text{ref}}} \frac{\partial}{\partial(\nabla \tilde{\mathbf{v}})} f_1(x, \mathbf{v}, \nabla \mathbf{v}, p) V'^T \nabla \mathbf{v} \, dx \, dt \\
&\quad + \int_I \int_{\Omega_{\text{ref}}} \frac{\partial}{\partial \tilde{x}} f_1(x, \mathbf{v}, \nabla \mathbf{v}, p) V \, dx \, dt \\
&\quad + \int_{\Omega_{\text{ref}}} f_2(x, \mathbf{v}(x, T)) \operatorname{div} V \, dx \\
&\quad + \int_{\Omega_{\text{ref}}} \frac{\partial}{\partial \tilde{x}} f_2(x, \mathbf{v}(x, T)) V \, dx.
\end{aligned} \tag{3.9}$$

*Remark 3.2.* As already mentioned in Subsection 2.2.2, for computational purposes it is convenient for a given iterate  $\tau_k$  to calculate the reduced gradient on the domain  $\Omega_k$ . As described in detail in 2.2.2 we have to solve the Navier-Stokes equations and the adjoint system on  $\Omega_k$ . Using  $\langle j'(\tau_k), V \rangle_{T^*(\Omega_{\text{ref}}), T(\Omega_{\text{ref}})} = \langle \tilde{j}'(\text{id}), \tilde{V} \rangle_{T^*(\Omega_k), T(\Omega_k)}$  we can take the formula above replacing  $\Omega_{\text{ref}}$  by  $\Omega_k$  and using the corresponding functions defined on  $\Omega_k$ .

Finally, if we assume more regularity for the state and adjoint, we can integrate by parts in the above formula and can represent the shape gradient as a functional on the boundary.

However, we prefer to work with the distributed version (3.8), since it is also appropriate for FE-Galerkin approximations, while the integration by parts

to obtain the boundary representation is not justified for FE-discretizations with  $H^1$ -elements. In addition, (3.8) can also easily be transferred to a boundary representation by using the procedure of Subsection 2.3 with a boundary displacement-to-domain transformation mapping  $u \mapsto \tau(u) \in T_{\text{ad}}$ . For Galerkin discretization the continuous adjoint calculus can then easily be applied on the discrete level.

## 4. Discretization

To discretize the instationary Navier-Stokes equations, we use the  $cG(1)dG(0)$  space-time finite element method, which uses piecewise constant finite elements in time and piecewise linear finite elements in space. The  $cG(1)dG(0)$  method is a variant of the General Galerkin  $G^2$ -method developed by Eriksson, Estep, Hansbo, and Johnson [4, 5].

Let  $\mathcal{I} = \{I_j = (t_{j-1}, t_j] : 1 \leq j \leq N\}$  be a partition of the time interval  $(0, T]$  with a sequence of discrete time steps  $0 = t_0 < t_1 < \dots < t_N = T$  and length of the respective time intervals  $k_j := |I_j| = t_j - t_{j-1}$ .

With each time step  $t_j$ , we associate a partition  $\mathcal{T}_j$  of the spatial domain  $\Omega$  and the finite element subspaces  $V_h^j, P_h^j$  of continuous piecewise linear functions in space.

The  $cG(1)dG(0)$  space-time finite element discretization with stabilization can be written as an implicit Euler scheme:  $\mathbf{v}_h^0 = \mathbf{v}_0$  and for  $j = 1, \dots, N$ , find  $(\mathbf{v}_h^j, p_h^j) \in V_h^j \times P_h^j$  such that

$$\begin{aligned} & (E^j(\mathbf{v}_h, p_h), (\mathbf{w}_h, q_h)) \\ &:= \left( \frac{\mathbf{v}_h^j - \mathbf{v}_h^{j-1}}{k_j}, \mathbf{w}_h \right) + (\nu \nabla \mathbf{v}_h^j, \nabla \mathbf{w}_h) + (\mathbf{v}_h^j \cdot \nabla \mathbf{v}_h^j, \mathbf{w}_h) - (p_h^j, \operatorname{div} \mathbf{w}_h) \\ &\quad + (\operatorname{div} \mathbf{v}_h^j, q_h) + SD_\delta(\mathbf{v}_h^j, p_h^j, \mathbf{w}_h, q_h) - (f, \mathbf{w}_h) = 0 \quad \forall (\mathbf{w}_h, q_h) \in V_h^j \times P_h^j \end{aligned}$$

with stabilization

$$\begin{aligned} SD_\delta(\mathbf{v}_h^j, p_h^j, \mathbf{w}_h, q_h) &= \left( \delta_1(\mathbf{v}_h^j \cdot \nabla \mathbf{v}_h^j + \nabla p_h^j - f), \mathbf{v}_h^j \cdot \nabla \mathbf{w}_h + \nabla q_h \right) \\ &\quad + (\delta_2 \operatorname{div} \mathbf{v}_h^j, \operatorname{div} \mathbf{w}_h) . \end{aligned}$$

The stabilization parameters

$$\delta_1 = \begin{cases} \frac{1}{2}(k_j^{-2} + |\mathbf{v}_h^j|^2 h_j^{-2})^{-1/2} & \text{if } \nu < |\mathbf{v}_h^j| h_j \\ \kappa_1 h_j^2 & \text{otherwise} \end{cases}, \quad \delta_2 = \begin{cases} \kappa_2 h_j & \text{if } \nu < |\mathbf{v}_h^j| h_j \\ \kappa_2 h_j^2 & \text{otherwise} \end{cases}$$

act as a subgrid model in the convection-dominated case  $\nu < |\mathbf{v}_h^j| h_j$ , where  $h_j$  denotes the local (spatial) mesh size at time  $j$  and  $\kappa_1$  and  $\kappa_2$  are constants of unit size.

As discrete objective functional, we consider

$$\begin{aligned} J^h(\mathbf{v}_h, p_h) &= \sum_{j=1}^N k_j \int_{\Omega} f_1(x, \mathbf{v}_h^j, \nabla \mathbf{v}_h^j, p_h^j) dx + \int_{\Omega} f_2(x, \mathbf{v}_h^N) dx \\ &=: J^{D,h}(\mathbf{v}_h, p_h) + J^{T,h}(\mathbf{v}_h^N). \end{aligned}$$

This is exactly  $J(\mathbf{v}_h, p_h)$ , since  $\mathbf{v}_h, p_h$  are piecewise constant in time.

In order to obtain gradients which are exact on the discrete level, we consider the discrete Lagrangian functional based on the  $cG(1)dG(0)$  finite element method, which is given by

$$\mathcal{L}_h(\mathbf{v}_h, p_h, \boldsymbol{\lambda}_h, \mu_h) = J^h(\mathbf{v}_h, p_h) + \sum_{j=1}^N k_j (E^j(\mathbf{v}_h, p_h), (\boldsymbol{\lambda}_h^j, \mu_h^j)).$$

Note again that this is exactly  $\mathcal{L}(\mathbf{v}_h, p_h, \boldsymbol{\lambda}_h, \mu_h)$ , since  $\mathbf{v}_h, p_h, \boldsymbol{\lambda}_h, \mu_h$  are piecewise constant in time.

Now we take the derivatives of the discrete Lagrangian w.r.t. the state variables to obtain the discrete adjoint equation and w.r.t. the shape variables to obtain the reduced gradient.

The discrete adjoint system can be cast in the form of an implicit time-stepping scheme backward in time:

For  $j = N-1, \dots, 0$ , find  $(\boldsymbol{\lambda}_h^j, \mu_h^j) \in V_h^j \times P_h^j$  such that

$$\begin{aligned} &\frac{(\boldsymbol{\lambda}_h^j, \mathbf{w}_h)}{k_j} + (\nu \nabla \boldsymbol{\lambda}_h^j, \nabla \mathbf{w}_h) + (\mu_h^j, \operatorname{div} \mathbf{w}_h) - (q_h, \operatorname{div} \boldsymbol{\lambda}_h^j) \\ &+ (\mathbf{v}_h^j \cdot \nabla \mathbf{w}_h, \boldsymbol{\lambda}_h^j) + (\mathbf{w}_h \cdot \nabla \mathbf{v}_h^j, \boldsymbol{\lambda}_h^j) + SD_{\delta}^*(\mathbf{v}_h^j, p_h^j, \boldsymbol{\lambda}_h^j, \mu_h^j; \mathbf{w}_h, q_h) \\ &= \frac{(\boldsymbol{\lambda}_h^{j+1}, \mathbf{w}_h)}{k_j} - \frac{1}{k_j} \langle J_{\mathbf{v}_h^j}^{D,h}(\mathbf{v}_h, p_h), \mathbf{w}_h \rangle - \frac{1}{k_j} \langle J_{p_h^j}^{D,h}(\mathbf{v}_h, p_h), q_h \rangle \end{aligned}$$

for all  $(\mathbf{w}_h, q_h) \in V_h^j \times P_h^j$ , where the discrete initial adjoint  $(\boldsymbol{\lambda}_h^N, \mu_h^N)$  solves the system

$$\begin{aligned} &\frac{\boldsymbol{\lambda}_h^N \cdot \mathbf{w}_h}{k_N} + (\nu \nabla \boldsymbol{\lambda}_h^N, \nabla \mathbf{w}_h) + (\mu_h^N, \operatorname{div} \mathbf{w}_h) + (\mathbf{v}_h^N \cdot \nabla \mathbf{w}_h, \boldsymbol{\lambda}_h^N) \\ &+ (\mathbf{w}_h \cdot \nabla \mathbf{v}_h^N, \boldsymbol{\lambda}_h^N) - (q_h, \operatorname{div} \boldsymbol{\lambda}_h^N) + SD_{\delta}^*(\mathbf{v}_h^N, p_h^N, \boldsymbol{\lambda}_h^N, \mu_h^N; \mathbf{w}_h, q_h) \\ &= -\frac{1}{k_N} \langle J_{\mathbf{v}_h^N}^{D,h}(\mathbf{v}_h, p_h), \mathbf{w}_h \rangle - \frac{1}{k_N} \langle J_{\mathbf{v}_h^N}^{T,h}(\mathbf{v}_h^N), \mathbf{w}_h \rangle - \frac{1}{k_N} \langle J_{p_h^N}^{D,h}(\mathbf{v}_h, p_h), q_h \rangle \end{aligned}$$

for all  $(\mathbf{w}_h, q_h) \in V_h^N \times P_h^N$ .

The adjoint stabilization term  $SD_{\delta}^*$  is given by

$$\begin{aligned}
 SD_{\delta}^*(\mathbf{v}_h^j, p_h^j, \boldsymbol{\lambda}_h^j, \mu_h^j; \mathbf{w}_h, q_h) \\
 &= \delta_1(\mathbf{w}_h \cdot \nabla \mathbf{v}_h^j, \mathbf{v}_h^j \cdot \nabla \boldsymbol{\lambda}_h^j) + \delta_1(\mathbf{v}_h^j \cdot \nabla \mathbf{w}_h, \mathbf{v}_h^j \cdot \nabla \boldsymbol{\lambda}_h^j) \\
 &+ \delta_1(\mathbf{v}_h^j \cdot \nabla \mathbf{v}_h^j, \mathbf{w}_h \cdot \nabla \boldsymbol{\lambda}_h^j) + \delta_1(\nabla q_h, \nabla \mu_h^j) + \delta_2(\operatorname{div} \mathbf{w}_h, \operatorname{div} \boldsymbol{\lambda}_h^j) \\
 &+ \delta_1(\mathbf{w}_h \cdot \nabla \mathbf{v}_h^j, \nabla \mu_h^j) + \delta_1(\mathbf{v}_h^j \cdot \nabla \mathbf{w}_h, \nabla \mu_h^j) \\
 &+ \delta_1(\nabla q_h, \mathbf{v}_h^j \cdot \nabla \boldsymbol{\lambda}_h^j) + \delta_1(\nabla p_h^j, \mathbf{w}_h \cdot \nabla \boldsymbol{\lambda}_h^j).
 \end{aligned}$$

For simplicity, we have neglected the terms containing the right-hand side  $f$  and the dependence of  $\delta_1$  on  $\mathbf{v}_h^j$ .

To compute shape derivatives on the discrete level we use a transformation space  $T^h(\Omega_{\text{ref}})$  of piecewise linear continuous functions. Then a discrete version of assumption (A) holds, i.e., the finite element space remains after transformation the space of continuous piecewise linear functions in space. The same holds for higher-order finite elements. Therefore, an analogue of (3.9) holds also on the discrete level if a Galerkin method is used and we obtain easily the exact shape derivative, if the adjoint state is computed by the exact discrete adjoint equation stated above. In this way we have obtained the exact shape derivative on the discrete level by using a continuous adjoint approach without the tedious task of computing mesh sensitivities.

## 5. Numerical results

In this section we demonstrate the adjoint shape derivative calculus on a numerical model problem. In particular, we consider an incompressible instationary flow around an object  $B$  for which the drag shall be minimized.

### 5.1. Problem description

The model problem is based on the DFG benchmark of a 2D instationary flow around a cylinder [20], see Figure 1. We prescribe a fixed parabolic inflow profile

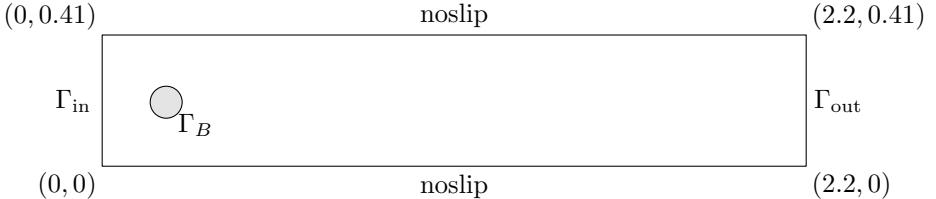


FIGURE 1. DFG-Benchmark flow around a cylinder; sketch of the geometry

on the left boundary  $\Gamma_{\text{in}}$  with  $v_{\text{max}} = 1.5 \text{ m/s}$ , noslip boundary conditions on the top and bottom boundaries, as well as on the object boundary  $\Gamma_B$ , and a free outflow condition on the right boundary  $\Gamma_{\text{out}}$ . The flow is modeled by the

instationary incompressible Navier-Stokes equations, with viscosity  $\nu = 10^{-4}$ . The Navier-Stokes equations are discretized with the  $cG(1)dG(0)$  finite element method presented above, with a fixed time step size  $k = 10^{-2}$  and a triangular spatial mesh with about 4100 vertices and 7900 elements.

The object boundary  $\Gamma_B$  is parameterized using a cubic B-Spline curve with 7 control points for the upper half of  $\Gamma_B$ , which is reflected at the  $y = 0.2$ -axis to obtain a  $y$ -symmetric closed curve. This parameterization allows for apices at the front and rear of the object, while the remaining boundary is  $C^2$ . We impose constraints on the volume of the object  $B$  as well as bound constraints on the control points. Using the coordinates of the control points as design parameters, we arrive at an optimization problem with 14 design variables, 12 of which are free (2 design parameters are fixed as we have to ensure that the  $y$ -coordinates of the first and last B-spline curve points equal 0.2 in order for the curve to be closed).

We compute the mean value of the drag on the object boundary  $\Gamma_B$  over the time interval  $[0, T]$  by using the formula

$$J((\tilde{\mathbf{v}}, \tilde{p}), \Omega) = \frac{1}{T} \int_0^T \int_{\Omega} \left( (\tilde{\mathbf{v}}_t + (\tilde{\mathbf{v}} \cdot \nabla) \tilde{\mathbf{v}} - \tilde{\mathbf{f}})^T \Phi - \tilde{p} \operatorname{div} \Phi + \nu \nabla \tilde{\mathbf{v}} : \nabla \Phi \right) dx dt. \quad (5.1)$$

Here,  $\Phi$  is a smooth function such that with a unit vector  $\phi$  pointing in the mean flow direction holds

$$\Phi|_{\Gamma_B} \equiv \phi, \quad \Phi|_{\partial\Omega \setminus \Gamma_B} \equiv 0 \quad \forall \Omega \in \mathcal{O}_{\text{ad}}.$$

This formula is an alternative formula for the mean value of the drag on  $\Gamma_B$ ,

$$c_d := \frac{1}{T} \int_0^T \int_{\Gamma_B} \mathbf{n} \cdot \sigma(\tilde{\mathbf{v}}, \tilde{p}) \cdot \phi dS,$$

with normal vector  $n$  and stress tensor  $\sigma(\tilde{\mathbf{v}}, \tilde{p}) = \nu \frac{1}{2}(\nabla \tilde{\mathbf{v}} + (\nabla \tilde{\mathbf{v}})^T) - \tilde{p} I$ , and can be obtained through integration by parts. For a detailed derivation, see [12]. Integration by parts in the time derivative shows that (5.1) can also be written as

$$\begin{aligned} J((\tilde{\mathbf{v}}, \tilde{p}), \Omega) &= \frac{1}{T} \int_0^T \int_{\Omega} \left( ((\tilde{\mathbf{v}} \cdot \nabla) \tilde{\mathbf{v}} - \tilde{\mathbf{f}})^T \Phi - \tilde{p} \operatorname{div} \Phi + \nu \nabla \tilde{\mathbf{v}} : \nabla \Phi \right) dx dt \\ &\quad + \frac{1}{T} \int_{\Omega} (\tilde{\mathbf{v}}(x, T) - \tilde{\mathbf{v}}_0(x))^T \Phi(x) dx. \end{aligned} \quad (5.1)$$

Thus the drag functional (5.1) has the form (3.6). Moreover, using the well-known embedding  $Y(\Omega) \hookrightarrow C(I; L^2(\Omega)^d) \times L_0^2(\Omega)$  it is easy to see that  $(\tilde{\mathbf{v}}, \tilde{p}) \in Y(\Omega) \mapsto J((\tilde{\mathbf{v}}, \tilde{p}), \Omega)$  is continuously differentiable if  $\Phi \in W^{1,\infty}(\mathbb{R}^2)^2$ .

Computation of the state, adjoint and shape derivative equations is done using Dolfin [14], which is part of the FEniCS project [7]. The optimization is carried out using the interior point solver IPOPT [22], with a BFGS-approximation for the reduced Hessian.

### 5.2. Choice of shape parameters and shape deformation techniques

One aspect to consider in the implementation of shape optimization algorithms is the choice of the shape parameters and the shape deformation technique. Generally speaking, shape parameterizations and deformations fall into two classes. In the first case, a parameterization directly defines the whole domain, which can be accomplished by using, e.g., free form deformation. In the second case, the parameterization determines the shape of the surface  $\Gamma_B$  of the object  $B$ . Examples for this kind of parameterizations can be B-splines, NURBS, but also the set of boundary points  $\Gamma_B$  itself, if considered in an appropriate function space. Changes in the shape of the boundary  $\Gamma_B$  then have to be transferred to changes of the domain  $\Omega_{\text{ref}}$ . This can be done in various ways, see, e.g., [2].

In our model problem, we have chosen a parameterization of the object boundary  $\Gamma_B$  based on closed cubic B-spline curves [16], where the B-spline control points act as design parameters  $u$ . The transformation of boundary displacements to displacements of the domain is done by solving an elasticity equation, where we prescribe the displacement of the object boundary as inhomogeneous Dirichlet boundary data [2]. The computational domains  $\Omega_k := \Omega(\tau(u_k))$  are obtained as transformations of a triangulation of the domain shown in Figure 1. As described at the end of section 4 we use piecewise linear transformations to ensure a discrete analogue of assumption (A). Then by an analogue of (3.9) together with the discrete adjoint equation we obtained conveniently by a continuous adjoint calculus the exact shape derivative on the discrete level – which we have also checked numerically.

### 5.3. Results

The IPOPT-algorithm needs 15 interior-point iterations for converging to a tolerance of  $10^{-3}$ , altogether needing 17 state equation solves and 16 adjoint solves. The drag value in the optimal shape is reduced by nearly one third in comparison to the initial shape. In the optimal solution, bound constraints for 8 of the design parameters are active, while 6 are inactive. The results of the optimization process are summarized in Table 1.

Figure 2 shows the velocity fields for the initial and optimal shape, with snapshots taken at end time, while Figure 3 shows the computational mesh both for the initial and the optimal shape. Both meshes are obtained by a transformation of the same reference mesh with a circular object, cf. Figure 1, by solving an elasticity equation with fixed displacement of the object boundary.

## 6. Conclusions and outlook

In this paper, we have presented a continuous adjoint approach that can easily be transferred in an exact way to the discrete level, if a Galerkin method in space is used. We use a domain representation of the shape gradient, since a boundary representation requires integration by parts, which is usually not justified on the

iteration	objective	dual infeasibility	linesearch-steps
0	1.2157690e-1	1.69e+0	0
1	1.0209697e-1	1.53e+0	2
2	9.7036722e-2	3.60e-1	1
3	8.7039312e-2	6.44e-1	1
4	8.4563185e-2	5.08e-1	1
5	8.3512670e-2	1.01e-1	1
6	8.2813890e-2	1.22e-1	1
7	8.2516118e-2	8.96e-2	1
8	8.2069666e-2	1.42e-1	1
9	8.2062288e-2	1.39e-1	1
10	8.1995990e-2	1.80e-2	1
11	8.1994727e-2	6.55e-3	1
12	8.1995485e-2	2.76e-3	1
13	8.1995822e-2	2.72e-3	1
14	8.1995966e-2	1.32e-3	1
15	8.1995811e-2	2.66e-5	1

TABLE 1. Optimization Results

discrete level. Nevertheless, adjoint based gradient representations can easily be derived from our gradient representation, e.g., for the boundary shape gradient in function space, but also for shape parameterizations, for example free form deformation or parameterized boundary displacement. The proposed approach allows the solution of the state equation and adjoint equation on the physical domain. Therefore existing solvers of the partial differential equation and its adjoint can be used.

We have applied our approach to the instationary incompressible Navier-Stokes equations. In the context of the stabilized  $cG(1)dG(0)$  method – but also for other Galerkin schemes and other types of partial differential equations – we were able to derive conveniently the exact discrete shape derivative, since our calculus is exact on the discrete level, if some simple rules are followed.

The combination with error estimators and multilevel techniques is subject of current research. Our results indicate that these techniques can reduce the number of optimization iterations on the fine grids and the necessary degrees of freedom significantly. We leave these results to a future paper.

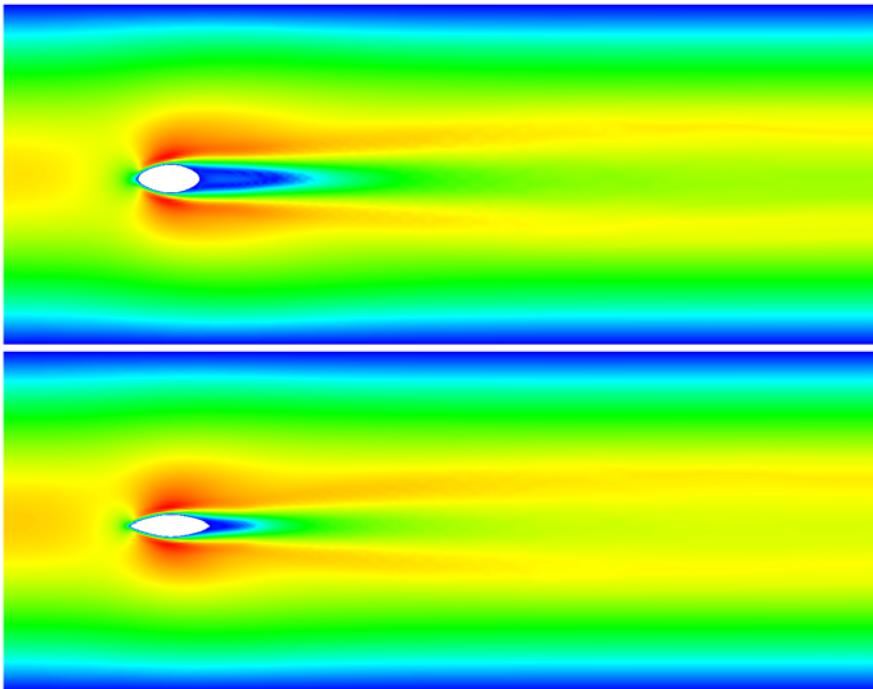


FIGURE 2. Comparison of the velocity fields for the initial and optimal shape

## References

- [1] K.K. Choi and N.-H. Kim, *Structural Sensitivity Analysis and Optimization 1: Linear Systems*, Mechanical Engineering Series, Springer, 2005.
- [2] K.K. Choi and N.-H. Kim, *Structural Sensitivity Analysis and Optimization 2: Non-linear Systems and Applications*, Mechanical Engineering Series, Springer, 2005.
- [3] M.C. Delfour and J.-P. Zolésio, *Shapes and Geometries; Analysis, Differential Calculus, and Optimization*, SIAM series on Advances in Design and Control, 2001.
- [4] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, *Introduction to Adaptive Methods for Differential Equations*, Acta Numerica, pp. 105–158, 1995.
- [5] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, *Computational Differential Equations*, Cambridge University Press, 1996.
- [6] L.C. Evans *Partial Differential Equations*, Graduate Studies in Mathematics, American Mathematical Society, 1998.
- [7] FEniCS, FEniCS Project, <http://www.fenics.org/>, 2007
- [8] P. Guillaume and M. Masmoudi, *Computation of high order derivatives in optimal shape design*, Numer. Math. 67, pp. 231–250, 1994.

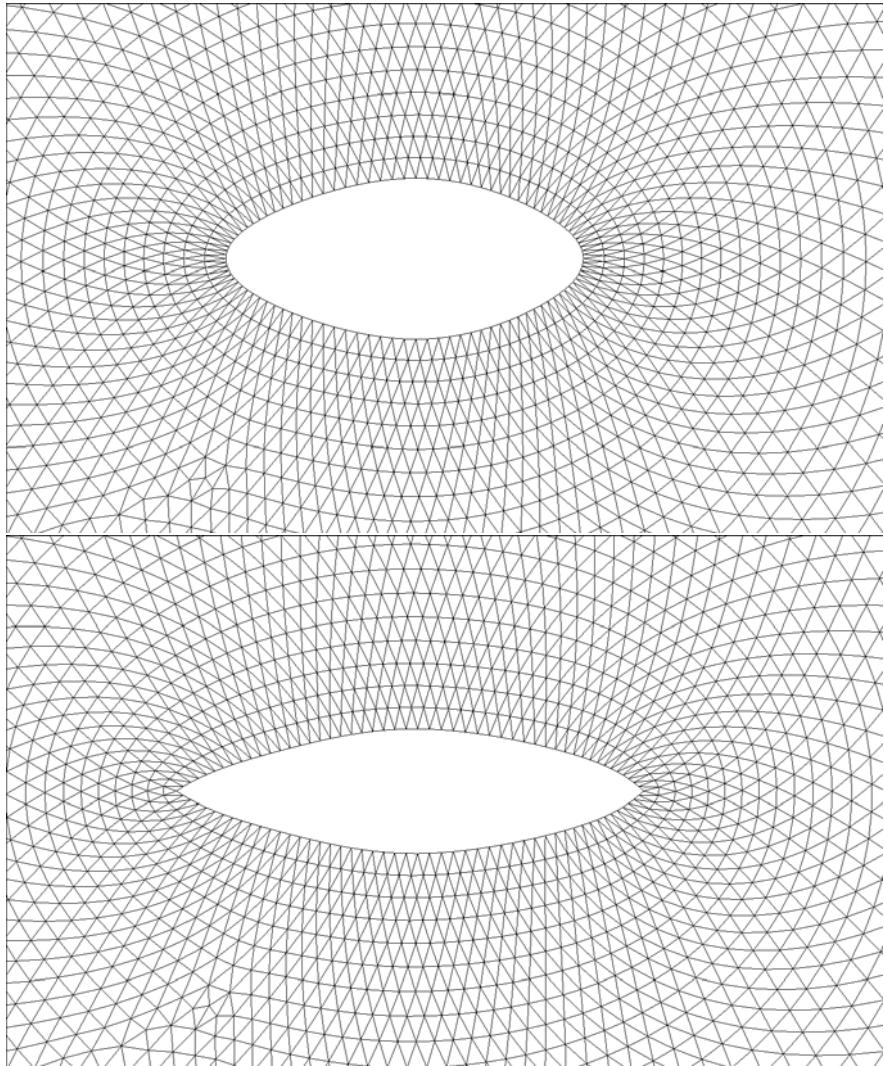


FIGURE 3. Comparison of the meshes for the initial and optimal shape

- [9] J. Haslinger and R.A. E. Mäkinen, *Introduction to Shape Optimization: Theory, Approximation, and Computation*, SIAM series on Advances in Design and Control, 2003.
- [10] J.G. Heywood, R. Rannacher and S. Turek, *Artificial boundaries and flux and pressure conditions for the incompressible Navier-Stokes equations*, Int. J. Numer. Methods Fluids 22, pp. 325–352, 1996.

- [11] M. Hinze and K. Kunisch, *Second order methods for optimal control of time-dependent fluid flow*, SIAM J. Control Optim. 40, pp. 925–946, 2001.
- [12] J. Hoffman and C. Johnson, *Adaptive Finite Element Methods for Incompressible Fluid Flow*, Error estimation and solution adaptive discretization in CFD: Lecture Notes in Computational Science and Engineering, Springer Verlag, 2002.
- [13] J. Hoffman and C. Johnson, *A New Approach to Computational Turbulence Modeling*, Comput. Meth. Appl. Mech. Engrg. 195, pp. 2865–2880, 2006.
- [14] J. Hoffman, J. Jansson, A. Logg, G.N. Wells et al., *DOLFIN*, <http://www.fenics.org/dolfin/>, 2006.
- [15] B. Mohammadi and O. Pironneau, *Applied Shape Optimization for Fluids*, Oxford University Press, 2001.
- [16] M.E. Mortensen, *Geometric Modeling*, Wiley, 1985.
- [17] F. Murat and S. Simon, Etudes de problèmes d'optimal design, Lectures Notes in Computer Science 41, pp. 54–62, 1976.
- [18] J. Sokolowski and J.-P. Zolésio, *Introduction to Shape Optimization*, Series in Computational Mathematic, Springer, 1992.
- [19] R. Temam, *Navier-Stokes Equations: Theory and Numerical Analysis* 3rd Edition, Elsevier Science Publishers, 1984.
- [20] M. Schäfer and S. Turek, *Benchmark Computations of Laminar Flow Around a Cylinder*, Preprints SFB 359, No. 96-03, Universität Heidelberg, 1996.
- [21] M. Ulbrich, *Constrained optimal control of Navier-Stokes flow by semismooth Newton methods*, Systems Control Lett. 48, pp. 297–311, 2003.
- [22] A. Wächter and L.T. Biegler, *On the Implementation of a Primal-Dual Interior Point Filter Line Search Algorithm for Large-Scale Nonlinear Programming*, Math. Program. 106, pp. 25–57, 2006.
- [23] W.P. Ziemer, *Weakly Differentiable Functions*, Graduate Texts in Mathematics, Springer, 1989.

Christian Brandenburg and Stefan Ulbrich  
 Fachbereich Mathematik  
 Technische Universität Darmstadt  
 Schloßgartenstr. 7  
 D-64289 Darmstadt,  
 e-mail: [brandenburg@mathematik.tu-darmstadt.de](mailto:brandenburg@mathematik.tu-darmstadt.de)  
[ulbrich@mathematik.tu-darmstadt.de](mailto:ulbrich@mathematik.tu-darmstadt.de)

Florian Lindemann and Michael Ulbrich  
 Lehrstuhl für Mathematische Optimierung  
 Zentrum Mathematik, M1  
 TU München  
 Boltzmannstr. 3  
 D-85747 Garching bei München  
 e-mail: [lindemann@ma.tum.de](mailto:lindemann@ma.tum.de)  
[mulbrich@ma.tum.de](mailto:mulbrich@ma.tum.de)

# Recent Advances in the Analysis of State-constrained Elliptic Optimal Control Problems

Eduardo Casas and Fredi Tröltzsch

**Abstract.** Pointwise state-constrained control problems associated with semilinear elliptic equations are studied. Theoretical results are derived, which are necessary to carry out the numerical analysis of the control problem. In particular, sufficient second-order optimality conditions, some new regularity results on optimal controls and a sufficient condition for the uniqueness of the Lagrange multiplier associated with the state constraints are presented.

**Mathematics Subject Classification (2000).** 49J20, 49K20, 35J65.

**Keywords.** Optimal control, pointwise state constraints, first- and second-order optimality conditions, Lagrange multipliers, Borel measures.

## 1. Introduction

In this paper, we consider several aspects of state-constrained optimal control problems for semilinear elliptic equations, which seem to be important for a related numerical analysis. For instance, due to the non-convex character of such problems, a reasonable error analysis should be based on second-order sufficient optimality conditions at locally optimal controls. It is known that such conditions are fairly delicate under the presence of state constraints. In [2], second-order sufficient conditions were established, which are, in some sense, closest to associated necessary ones and admit a form similar to the theory of nonlinear programming in finite-dimensional spaces. Here, we briefly discuss this result and show its equivalence to an earlier form stated in [5] that was quite difficult to explain.

Another important pre-requisite for the numerical analysis of control problems is the smoothness that can be expected of optimal controls. We show that

---

The first author was partially supported by the Spanish Ministry of Education and Science under projects MTM2005-06817 and “Ingenio Mathematica (i-MATH)” CSD2006-00032 (Consolider Ingenio 2010).

the optimal control is Lipschitz, if the state constraints are only active at finitely many points. We also present a counterexample that this result is not true for infinitely many active points. On the other hand, we prove the somehow surprising result that optimal controls belong to  $H^1(\Omega)$  no matter how large the active set is. Moreover, we also discuss the uniqueness of the Lagrange multiplier associated with the state-constraints.

## 2. The control problem

Let  $\Omega$  be an open, connected and bounded domain in  $\mathbb{R}^n$ ,  $n = 2, 3$ , with a Lipschitz boundary  $\Gamma$ . In this domain we consider the state equation

$$\begin{cases} -\Delta y + a_0(y) &= u + e \quad \text{in } \Omega, \\ y &= 0 \quad \text{on } \Gamma, \end{cases} \quad (1)$$

where  $a_0$  and  $e$  are fixed functions specified below. In (1), the function  $u$  denotes the control and we will denote by  $y_u$  the solution associated with  $u$ . We will state later the conditions leading to the existence and uniqueness of a solution of (1) in  $C(\bar{\Omega}) \cap H_0^1(\Omega)$ .

The optimal control problem is formulated as follows

$$(P) \begin{cases} \min J(u) = \frac{1}{2} \int_{\Omega} (y_u(x) - y_d(x))^2 dx + \frac{N}{2} \int_{\Omega} u(x)^2 dx \\ \text{subject to } (y_u, u) \in (C(\bar{\Omega}) \cap H_0^1(\Omega)) \times L^\infty(\Omega), \\ \alpha \leq u(x) \leq \beta \quad \text{for a.e. } x \in \Omega, \\ a \leq y_u(x) \leq b \quad \forall x \in K. \end{cases}$$

We impose the following assumptions on the data of the control problem:

(A1) In the whole paper, fixed functions  $y_d, e \in L^2(\Omega)$ , and real numbers  $N > 0$ ,  $\alpha, \beta, a, b$  are given with  $\alpha \leq \beta$  and  $a < b$ . The fixed function  $a_0 : \mathbb{R} \rightarrow \mathbb{R}$  is monotone non-decreasing and locally of class  $C^{2,1}$ .

We introduce the sets of admissible controls  $\mathcal{U}_{\alpha,\beta}$  and the feasible set  $\mathcal{U}_{\text{ad}}$  of the problem by

$$\mathcal{U}_{\alpha,\beta} = \{u \in L^\infty(\Omega) : \alpha \leq u(x) \leq \beta \text{ a.e. in } \Omega\}$$

$$\mathcal{U}_{\text{ad}} = \{u \in \mathcal{U}_{\alpha,\beta} : a \leq y_u(x) \leq b \ \forall x \in K\}.$$

(A2)  $K$  is a compact subset of  $\bar{\Omega}$ . We also assume that either  $K \cap \Gamma = \emptyset$  or  $a < 0 < b$ . Moreover, we define the set

$$\mathcal{Y}_{ab} = \{z \in C(K) : a \leq z(x) \leq b \ \forall x \in K\}.$$

By the formal setting  $a = -\infty$  ( $b = +\infty$ ) we can include the case, where only upper (or lower) bounds on the state are given.

We are able to deal with more general problems, for instance with a more general elliptic differential operator, a nonlinearity  $a_0$  of the form  $a_0(x, y)$ , a non-quadratic objective functional of integral type, bounds depending on  $x$  etc. These

generalizations will be included in a forthcoming paper. Here, to keep the presentation simple, we consider the setting introduced above.

Let us state first the existence and uniqueness of the solution corresponding to the state equation (1).

**Theorem 2.1.** *Under assumption (A1), the equation (1) has a unique solution  $y_u \in H_0^1(\Omega) \cap C(\bar{\Omega})$  for every  $u \in L^2(\Omega)$ . Moreover,  $y_u \in H^2(\Omega)$  if  $\Omega$  is convex and  $y_u \in W^{2,p}(\Omega)$  if  $\Gamma$  is of class  $C^{1,1}$  and  $u, e \in L^p(\Omega)$ , with  $p > n$ .*

It is well known that equation (1) has a unique solution  $y_u \in H_0^1(\Omega) \cap C(\bar{\Omega})$  for every  $u \in L^p(\Omega)$ . A proof of this result can be obtained by the usual cut off process applied to  $a_0$ , then applying a Schauder's fix point theorem combined with the monotonicity of  $a_0$  and  $L^\infty$  estimates for the state; cf. Stampacchia [11]. The continuity of  $y_u$  is proven in [6]. The  $W^{2,p}(\Omega)$  and  $H^2(\Omega)$  estimates can be found in Grisvard [7].

Now the existence of an optimal control can be proved by using standard arguments.

**Theorem 2.2.** *Assume that the set of controls  $\mathcal{U}_{\text{ad}}$  is not empty. Then the control problem (P) has at least one solution.*

In the rest of the paper,  $\bar{u}$  denotes a local minimum of (P) in the sense of the  $L^\infty(\Omega)$ -topology and  $\bar{y}$  will be its associated state. At such a local minimizer, we will assume the linearized Slater condition.

(A3) There exists  $u_0 \in \mathcal{U}_{\alpha,\beta}$  such that

$$a < \bar{y}(x) + z_0(x) < b \quad \forall x \in K, \quad (2)$$

where  $z_0 \in H_0^1(\Omega) \cap C(\bar{\Omega})$  is the unique solution of

$$\begin{cases} -\Delta z + a'_0(\bar{y})z = u_0 - \bar{u} & \text{in } \Omega \\ z = 0 & \text{on } \Gamma. \end{cases} \quad (3)$$

Since  $K$  is compact and  $\bar{y}, z_0 \in C(K)$ , we deduce that (2) is equivalent to the existence of real  $\tau_1, \tau_2 \in \mathbb{R}$  such that

$$a < \tau_1 < \bar{y}(x) + z_0(x) < \tau_2 < b \quad \forall x \in K. \quad (4)$$

### 3. First- and second-order optimality conditions

Before deriving the first-order optimality conditions satisfied by the local minimizer  $\bar{u}$ , we recall some results about the differentiability of the mappings involved in the control problem. For the proofs, the reader is referred to Casas and Mateos [3], where a Neumann boundary condition was considered instead of the Dirichlet condition posed in this paper. The method of proof is very similar and the necessary changes are obvious.

**Theorem 3.1.** *If (A1) holds, then the mapping  $G : L^2(\Omega) \rightarrow C(\bar{\Omega}) \cap H_0^1(\Omega)$ , defined by  $G(u) = y_u$  is of class  $C^2$ . Moreover, for all  $u, v \in L^2(\Omega)$ ,  $z_v = G'(u)v$  is defined as the solution of*

$$\begin{cases} -\Delta z_v + a'_0(y_u)z_v &= v \quad \text{in } \Omega \\ z_v &= 0 \quad \text{on } \Gamma. \end{cases} \quad (1)$$

Finally, for every  $v_1, v_2 \in L^2(\Omega)$ ,  $z_{v_1 v_2} = G''(u)v_1 v_2$  is the solution of

$$\begin{cases} -\Delta z_{v_1 v_2} + a'_0(y_u)z_{v_1 v_2} + a''_0(y_u)z_{v_1}z_{v_2} &= 0 \quad \text{in } \Omega \\ z_{v_1 v_2} &= 0 \quad \text{on } \Gamma, \end{cases} \quad (2)$$

where  $z_{v_i} = G'(u)v_i$ ,  $i = 1, 2$ .

**Remark 3.1.** Let us remark that the assumption  $n \leq 3$  is required to apply the second-order optimality conditions because the differentiability of  $G$  from  $L^2(\Omega)$  to  $C(\bar{\Omega})$  is needed for the proof. This result holds true only for  $n \leq 3$ .

**Theorem 3.2.** Suppose that (A1) holds. Then  $J : L^2(\Omega) \rightarrow \mathbb{R}$  is a functional of class  $C^2$ . Moreover, for every  $u, v, v_1, v_2 \in L^2(\Omega)$

$$J'(u)v = \int_{\Omega} (\varphi_{0u} + Nu) v \, dx \quad (3)$$

and

$$J''(u)v_1 v_2 = \int_{\Omega} [z_{v_1} z_{v_2} + Nv_1 v_2 - \varphi_{0u} a''_0(y_u)z_{v_1}z_{v_2}] \, dx, \quad (4)$$

where  $y_u = G(u)$  and  $\varphi_{0u} \in W_0^{1,s}(\Omega)$ , for all  $s < n/(n-1)$ , is the unique solution of the adjoint problem

$$\begin{cases} -\Delta \varphi + a'_0(y_u)\varphi &= y_u - y_d \quad \text{in } \Omega \\ \varphi &= 0 \quad \text{on } \Gamma, \end{cases} \quad (5)$$

and  $z_{v_i} = G'(u)v_i$ ,  $i = 1, 2$ .

The previous theorem and the next one follow easily from Theorem 3.1 and the chain rule.

**Theorem 3.3.** If (A1) is satisfied, then the mapping  $F : L^2(\Omega) \rightarrow C(K)$ , defined by  $F(u) = y_u(\cdot)|_K$ , is of class  $C^2$ . Moreover, for every  $u, v, v_1, v_2 \in L^2(\Omega)$

$$F'(u)v = z_v(\cdot) \quad (6)$$

and

$$F''(u)v_1 v_2 = z_{v_1 v_2}(\cdot), \quad (7)$$

where  $z_{v_i} = G'(u)v_i$ ,  $i = 1, 2$ , and  $z_{v_1 v_2} = G''(u)v_1 v_2$ .

Before stating the first-order optimality conditions, let us fix some notation. We denote by  $M(K)$  the Banach space of all real and regular Borel measures in  $K$ , which is identified with the dual space of  $C(K)$ . The following result is well known.

**Theorem 3.4.** Let  $\bar{u}$  be a local solution of (P) and suppose that the assumptions (A1)–(A2) are satisfied. Then there exist a measure  $\bar{\mu} \in M(K)$  and a function  $\bar{\varphi} \in W_0^{1,s}(\Omega)$ , for all  $1 \leq s < n/(n-1)$ , such that

$$\begin{cases} -\Delta \bar{\varphi} + a'_0(\bar{y}(x))\bar{\varphi} = \bar{y} - y_d + \bar{\mu} & \text{in } \Omega, \\ \bar{\varphi} = 0 & \text{on } \Gamma, \end{cases} \quad (8)$$

$$\int_K (z(x) - \bar{y}(x)) d\bar{\mu}(x) \leq 0 \quad \forall z \in \mathcal{Y}_{ab}, \quad (9)$$

$$\int_{\Omega} (\bar{\varphi} + N\bar{u})(u - \bar{u}) dx \geq 0 \quad \forall u \in \mathcal{U}_{\alpha,\beta}. \quad (10)$$

**Remark 3.2.** It is well known that, in view of  $a \leq \bar{y}(x) \leq b$ , inequality (9) implies that the support of  $\bar{\mu}$  is in the set  $K_0 = K_a \cup K_b$  with

$$K_a = \{x \in K : \bar{y}(x) = a\} \text{ and } K_b = \{x \in K : \bar{y}(x) = b\}.$$

Moreover the Lebesgue decomposition of  $\bar{\mu} = \mu_+ - \mu_-$  implies that  $\text{supp } \mu_+ \subset K_b$  and  $\text{supp } \mu_- \subset K_a$ . Because of this property and by assumption (A2) we have that  $\text{supp } \bar{\mu} \cap \Gamma = \emptyset$ . Notice that that  $K_a$  and  $K_b$  are closed subsets.

**Remark 3.3.** From (10), it follows

$$\bar{u}(x) = \text{Proj}_{[\alpha,\beta]}(-\frac{1}{N}\bar{\varphi}(x)) = \max\{\alpha, \min\{\bar{\varphi}(x), \beta\}\} \text{ for a.e. } x \in \Omega. \quad (11)$$

Let us formulate the Lagrangian version of the optimality conditions (8)–(10). We define the Lagrange function  $\mathcal{L} : L^2(\Omega) \times M(K) \rightarrow \mathbb{R}$  associated with the problem (P) by

$$\mathcal{L}(u, \mu) = J(u) + \int_K y_u(x) d\mu(x).$$

Using (3) and (6) we find that

$$\frac{\partial \mathcal{L}}{\partial u}(u, \mu)v = \int_{\Omega} (\varphi_u(x) + Nu(x)) v(x) dx, \quad (12)$$

where  $\varphi_u$  with  $\varphi_u \in W_0^{1,s}(\Omega)$ , for all  $1 \leq s < n/(n-1)$ , is the solution of the Dirichlet problem

$$\begin{cases} -\Delta \varphi + a'_0(y_u)\varphi = y_u + \mu & \text{in } \Omega \\ \varphi = 0 & \text{on } \Gamma. \end{cases} \quad (13)$$

Now the inequality (10) along with (12) lead to

$$\frac{\partial \mathcal{L}}{\partial u}(\bar{u}, \bar{\mu})(u - \bar{u}) \geq 0 \quad \forall u \in \mathcal{U}_{\alpha,\beta}. \quad (14)$$

Before stating the sufficient second-order optimality conditions, we provide the expression of the second derivative of the Lagrangian with respect to the control. From (7), we get

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(u, \mu) v_1 v_2 = J''(u) v_1 v_2 + \int_K z_{v_1 v_2}(x) d\mu(x).$$

By (2) and (4), this is equivalent to

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(u, \mu) v_1 v_2 = \int_\Omega [z_{v_1} z_{v_2} + N v_1 v_2 - \varphi_u a_0''(y_u) z_{v_1} z_{v_2}] dx, \quad (15)$$

where  $\varphi_u$  is the solution of (13).

Associated with  $\bar{u}$ , we define the cone of critical directions by

$$C_{\bar{u}} = \{v \in L^2(\Omega) : v \text{ satisfies (16), (17) and (18)}\},$$

$$v(x) = \begin{cases} \geq 0 & \text{if } \bar{u}(x) = \alpha, \\ \leq 0 & \text{if } \bar{u}(x) = \beta, \\ = 0 & \text{if } \bar{\varphi}(x) + N\bar{u}(x) \neq 0, \end{cases} \quad (16)$$

$$z_v(x) = \begin{cases} \geq 0 & \text{if } x \in K_a \\ \leq 0 & \text{if } x \in K_b, \end{cases} \quad (17)$$

$$\int_K z_v(x) d\bar{\mu}(x) = 0, \quad (18)$$

where  $z_v \in H_0^1(\Omega) \cap C(\bar{\Omega})$  satisfies

$$\begin{cases} -\Delta z_v + a_0'(\bar{y}) z_v = v & \text{in } \Omega \\ z_v = 0 & \text{on } \Gamma. \end{cases}$$

The relation (17) expresses the natural sign conditions, which must be fulfilled for feasible directions at active points  $x \in K_a$  or  $K_b$ , respectively. On the other hand, (18) states that the derivative  $z_v$  must be zero whenever the corresponding Lagrange multiplier is non-vanishing. This restriction is needed for second-order sufficient conditions. Compared with the finite-dimensional case, this is exactly what we can expect. Therefore the relations (17)–(18) provide a convenient extension of the usual conditions of the finite-dimensional case.

We should mention that (18) is new in the context of infinite-dimensional optimization problems. In earlier papers on this subject, other extensions to the infinite-dimensional case were suggested. For instance, Maurer and Zowe [10] used first-order sufficient conditions to account for the strict positivity of Lagrange multipliers. Inspired by their approach, in [5] an application to state-constrained elliptic boundary control was suggested by the authors. In terms of our problem, equation (18) was relaxed by

$$\int_K z_v(x) d\bar{\mu}(x) \geq -\varepsilon \int_{\{x: |\bar{\varphi}(x) + N\bar{u}(x)| \leq \tau\}} |v(x)| dx$$

for some  $\varepsilon > 0$  and  $\tau > 0$ , cf. [5, (5.15)]. In the next theorem, which was proven in [2, Theorem 4.3], we will see that this relaxation is not necessary. We obtain a smaller cone of critical directions that seems to be optimal. However, the reader is referred to Theorem 3.6 below, where we consider the possibility of relaxing the conditions defining the cone  $C_{\bar{u}}$ .

**Theorem 3.5.** *Let us assume that (A1) holds. Let  $\bar{u}$  be a feasible control of problem (P),  $\bar{y}$  the associated state and  $(\bar{\varphi}, \bar{\mu}) \in W_0^{1,s}(\Omega) \times M(K)$ , for all  $1 \leq s < n/(n-1)$ , satisfying (8)–(10). Assume further that*

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\mu})v^2 > 0 \quad \forall v \in C_{\bar{u}} \setminus \{0\}. \quad (19)$$

*Then there exist  $\varepsilon > 0$  and  $\delta > 0$  such that it holds*

$$J(\bar{u}) + \frac{\delta}{2}\|u - \bar{u}\|_{L^2(\Omega)}^2 \leq J(u) \quad \text{if } \|u - \bar{u}\|_{L^2(\Omega)} \leq \varepsilon \text{ and } u \in \mathcal{U}_{\text{ad}}. \quad (20)$$

The condition (19) seems to be natural. In fact, under some regularity assumption, we can expect the inequality

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\mu})v^2 \geq 0 \quad \forall v \in C_{\bar{u}}$$

as a necessary condition for local optimality. At least, this is the case when the state constraints are of integral type, see [3], or when  $K$  is a finite set of points, see [1]. In the general case, the necessary second-order optimality conditions for problem (P) remain open for us.

We finish this section by establishing an equivalent condition to (19) that is more convenient for the numerical analysis of problem (P). Let us introduce a cone  $C_{\bar{u}}^\tau$  of critical directions that is bigger than  $C_{\bar{u}}$ . Given  $\tau > 0$ , we denote by  $C_{\bar{u}}^\tau$  the set of elements  $v \in L^2(\Omega)$  satisfying

$$v(x) = \begin{cases} \geq 0 & \text{if } \bar{u}(x) = \alpha, \\ \leq 0 & \text{if } \bar{u}(x) = \beta, \\ = 0 & \text{if } |\bar{\varphi}(x) + N\bar{u}(x)| > \tau, \end{cases} \quad (21)$$

$$z_v(x) = \begin{cases} \geq -\tau\|v\|_{L^2(\Omega)} & \text{if } x \in K_a \\ \leq +\tau\|v\|_{L^2(\Omega)} & \text{if } x \in K_b, \end{cases} \quad (22)$$

$$\int_K z_v(x) d\bar{\mu}(x) \geq -\tau\|v\|_{L^2(\Omega)}. \quad (23)$$

**Theorem 3.6.** *Under the assumption (A1), (19) holds if and only if there exist  $\tau > 0$  and  $\rho > 0$  such that*

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\mu})v^2 \geq \rho\|v\|_{L^2(\Omega)}^2 \quad \forall v \in C_{\bar{u}}^\tau. \quad (24)$$

*Proof.* Since  $C_{\bar{u}} \subset C_{\bar{u}}^\tau$ , it is clear that (24) implies (19). Let us prove by contradiction that (24) follows from (19). Assume that (19) holds but not (24). Then for

any positive integer  $k$  there exists an element  $v_k \in C_{\bar{u}}^{1/k}$  such that

$$\frac{\partial \mathcal{L}}{\partial u}(\bar{u}, \bar{\mu})v_k^2 < \frac{1}{k}\|v_k\|_{L^2(\Omega)}^2. \quad (25)$$

Redefining  $v_k$  as  $v_k/\|v_k\|_{L^2(\Omega)}$  and taking, if necessary, a subsequence denoted in the same way, we can assume that

$$\|v_k\|_{L^2(\Omega)} = 1, \quad v_k \rightharpoonup v \text{ weakly in } L^2(\Omega) \quad \text{and} \quad \frac{\partial \mathcal{L}}{\partial u}(\bar{u}, \bar{\mu})v_k^2 < \frac{1}{k}, \quad (26)$$

and from (21)–(23)

$$v_k(x) = \begin{cases} \geq 0 & \text{if } \bar{u}(x) = \alpha, \\ \leq 0 & \text{if } \bar{u}(x) = \beta, \\ = 0 & \text{if } |\bar{\varphi}(x) + N\bar{u}(x)| > 1/k, \end{cases} \quad (27)$$

$$z_{v_k}(x) = \begin{cases} \geq -1/k & \text{if } x \in K_a \\ \leq +1/k & \text{if } x \in K_b, \end{cases} \quad (28)$$

$$\int_K z_{v_k}(x) d\bar{\mu}(x) \geq -1/k. \quad (29)$$

Since  $z_{v_k} \rightarrow z_v$  strongly in  $H_0^1(\Omega) \cap C(\bar{\Omega})$ , we can pass to the limit in (26)–(29) and get that  $v \in C_{\bar{u}}$  and

$$\frac{\partial \mathcal{L}}{\partial u}(\bar{u}, \bar{\mu})v^2 \leq 0. \quad (30)$$

This is only possible if  $v = 0$ ; see (19). Let us note that the only delicate point to prove that  $v \in C_{\bar{u}}$  is to establish (18). Indeed, (16) and (17) follow easily from (27) and (28). Passing to the limit in (29) we obtain

$$\int_K z_v(x) d\bar{\mu}(x) \geq 0.$$

This inequality, along with (17) and the structure of  $\bar{\mu}$ , implies (18).

Therefore we have that  $v_k \rightarrow 0$  weakly in  $L^2(\Omega)$  and  $z_{v_k} \rightarrow 0$  strongly in  $H_0^1(\Omega) \cap C(\bar{\Omega})$ . Hence, using the expression (15) of the second derivative of the Lagrangian we deduce

$$N = \liminf_{k \rightarrow \infty} N\|v_k\|_{L^2(\Omega)}^2 \leq \liminf_{k \rightarrow \infty} \frac{\partial \mathcal{L}}{\partial u}(\bar{u}, \bar{\mu})v_k^2 = 0,$$

which is a contradiction.  $\square$

#### 4. Regularity of the optimal control

In this section, the existence of the second derivative  $a''_0$  is not needed. We start with the following regularity result for the optimal control, which is well known. We recall that  $K_0$  denotes the set of points, where the state constraints are active.

**Theorem 4.1.** *If  $(\bar{y}, \bar{u}, \bar{\varphi}) \in (H_0^1(\Omega) \cap C(\bar{\Omega})) \times L^\infty(\Omega) \times W_0^{1,s}(\Omega)$  satisfies the optimality system (8)–(10),  $(\bar{y}, \bar{u})$  being a feasible pair for problem (P), then  $\bar{u} \in W^{1,s}(\Omega)$  for all  $s < n/(n-1)$  and  $\bar{u} \in C(\bar{\Omega} \setminus K_0)$ .*

The regularity  $\bar{u} \in W^{1,s}(\Omega)$  follows immediately from (11) and the continuity  $\bar{u} \in C(\bar{\Omega} \setminus K_0)$  is deduced in the same way. This regularity result on the control  $\bar{u}$  can be improved if there is a finite number of points, where the state constraints are active. More precisely, assume that  $K_0 = \{x_j\}_{j=1}^m \subset \Omega$ . Then Remark 3.2 implies that

$$\bar{\mu} = \sum_{j=1}^m \bar{\lambda}_j \delta_{x_j}, \quad \text{with } \bar{\lambda}_j = \begin{cases} \geq 0 & \text{if } y(x_j) = b, \\ \leq 0 & \text{if } y(x_j) = a, \end{cases} \quad (1)$$

where  $\delta_{x_j}$  denotes the Dirac measure concentrated at  $x_j$ . If we denote by  $\bar{\varphi}_j$ ,  $1 \leq j \leq m$ , and  $\bar{\varphi}_0$  the solutions of

$$\begin{cases} -\Delta \bar{\varphi}_j + a'_0(\bar{y}(x)) \bar{\varphi}_j &= \delta_{x_j} \quad \text{in } \Omega, \\ \bar{\varphi}_j &= 0 \quad \text{on } \Gamma, \end{cases} \quad (2)$$

and

$$\begin{cases} -\Delta \bar{\varphi}_0 + a'_0(\bar{y}(x)) \bar{\varphi}_0 &= \bar{y} - y_d \quad \text{in } \Omega, \\ \bar{\varphi}_0 &= 0 \quad \text{on } \Gamma, \end{cases} \quad (3)$$

then the adjoint state associated to  $\bar{u}$  is given by

$$\bar{\varphi} = \bar{\varphi}_0 + \sum_{j=1}^m \bar{\lambda}_j \bar{\varphi}_j. \quad (4)$$

**Theorem 4.2.** *Assume that  $\Gamma$  is of class  $C^{1,1}$ ,  $y_d, e \in L^p(\Omega)$ ,  $p > n$ , and let  $(\bar{y}, \bar{u}, \bar{\varphi}) \in (H_0^1(\Omega) \cap C(\bar{\Omega})) \times L^\infty(\Omega) \times W_0^{1,s}(\Omega)$ , for all  $1 \leq s < n/(n-1)$ , satisfy the optimality system (8)–(10). If the active set consists of finitely many points in  $\Omega$ , i.e.,  $K_0 = \{x_j\}_{j=1}^m \subset \Omega$ , then  $\bar{u}$  belongs to  $C^{0,1}(\bar{\Omega})$  and  $\bar{y}$  to  $W^{2,p}(\Omega)$ .*

Since  $p > n$ , it holds that  $W^{2,p}(\Omega) \subset C^1(\bar{\Omega})$  and therefore  $\bar{\varphi}_0 \in C^1(\bar{\Omega})$ . On the other hand,  $\bar{\varphi}_j(x) \rightarrow +\infty$  when  $x \rightarrow x_j$ , hence  $\bar{\varphi}$  has singularities at the points  $x_j$  where  $\bar{\lambda}_j \neq 0$ . Consequently  $\bar{\varphi}$  cannot be Lipschitz.

Surprisingly, this does not lower the regularity of  $\bar{u}$ : Notice that (11) implies that  $\bar{u}$  is identically equal to  $\alpha$  or  $\beta$  in a neighborhood of  $x_j$ , depending on the sign of  $\bar{\lambda}_j$ . This implies the desired result; see Casas [1] for the details.

Now the question arises if this Lipschitz property remains also valid for an infinite number of points where the pointwise state constraints are active. Unfortunately, the answer is negative. In fact, the optimal control can even fail to be continuous if  $K_0$  is an infinite and numerable set. Let us present an associated

**Counterexample.** We set

$$\Omega = \{x \in \mathbb{R}^2 : \|x\| < \sqrt{2}\}, \quad \bar{y}(x) = \begin{cases} 1 & \text{if } \|x\| \leq 1 \\ 1 - (\|x\|^2 - 1)^4 & \text{if } 1 < \|x\| \leq \sqrt{2}, \end{cases}$$

$$K = \{x^k\}_{k=1}^\infty \cup \{x^\infty\}, \quad \text{where } x^k = \left(\frac{1}{k}, 0\right) \text{ and } x^\infty = (0, 0), \quad \bar{\mu} = \sum_{k=1}^\infty \frac{1}{k^2} \delta_{x^k}.$$

Now we define  $\bar{\varphi} \in W_0^{1,s}(\Omega)$  for all  $1 \leq s < n/(n-1)$  as the solution of the equation

$$\begin{cases} -\Delta \bar{\varphi} = \bar{y} + \bar{\mu} & \text{in } \Omega, \\ \bar{\varphi} = 0 & \text{on } \Gamma. \end{cases} \quad (5)$$

The function  $\bar{\varphi}$  can be decomposed in the form

$$\bar{\varphi}(x) = \bar{\psi}(x) + \sum_{k=1}^{\infty} \frac{1}{k^2} [\psi_k(x) + \phi(x - x^k)],$$

where  $\phi(x) = -(1/2\pi) \log \|x\|$  is the fundamental solution of  $-\Delta$  and  $\bar{\psi}, \psi_k \in C^2(\bar{\Omega})$  satisfy

$$\begin{cases} -\Delta \bar{\psi}(x) = \bar{y}(x) & \text{in } \Omega, \\ \bar{\psi}(x) = 0 & \text{on } \Gamma, \end{cases} \quad \begin{cases} -\Delta \psi_k(x) = 0 & \text{in } \Omega, \\ \psi_k(x) = -\phi(x - x^k) & \text{on } \Gamma. \end{cases}$$

Finally, we set

$$M = \left| \bar{\psi}(0) + \sum_{k=1}^{\infty} \frac{1}{k^2} \psi_k(0) \right| + \sum_{k=1}^{\infty} \frac{1}{k^2} \phi(x^k) + 1, \quad \bar{u}(x) = \text{Proj}_{[-M,+M]}(-\bar{\varphi}(x)) \quad (6)$$

and  $e = -(\bar{u}(x) + \Delta \bar{y}(x))$  and  $a_0 = 0$ . Then  $\bar{u}$  is the unique global solution of the control problem

$$(Q) \quad \begin{cases} \min J(u) = \frac{1}{2} \int_{\Omega} (y_u^2(x) + u^2(x)) dx \\ \text{subject to } (y_u, u) \in (C(\bar{\Omega}) \cap H_0^1(\Omega)) \times L^{\infty}(\Omega), \\ -M \leq u(x) \leq +M \quad \text{for a.e. } x \in \Omega, \\ -1 \leq y_u(x) \leq +1 \quad \forall x \in K, \end{cases}$$

where  $y_u$  is the solution of

$$\begin{cases} -\Delta y = u + e & \text{in } \Omega, \\ y = 0 & \text{on } \Gamma. \end{cases} \quad (7)$$

As a first step to prove that  $\bar{u}$  is a solution of problem, let us check that  $M$  is a real number. Since  $\{\phi(x - x^k)\}_{k=1}^{\infty}$  is bounded in  $C^2(\Gamma)$ , we get that  $\{\psi_k\}_{k=1}^{\infty}$  is also bounded in  $C^2(\bar{\Omega})$ . Therefore, the convergence of the first series of (6) is obvious. The convergence of the second series follows immediately from

$$\sum_{k=1}^{\infty} \frac{1}{k^2} \phi(x^k) = \frac{1}{2\pi} \sum_{k=1}^{\infty} \frac{1}{k^2} \log k < \infty.$$

Problem (Q) is strictly convex and  $\bar{u}$  is a feasible control with associated state  $\bar{y}$  satisfying the state constraints. Therefore, there exists a unique solution characterized by the optimality system. More precisely, the first-order optimality conditions are necessary and sufficient for a global minimum. Let us check that  $(\bar{y}, \bar{u}, \bar{\varphi}, \bar{\mu}) \in H_0^1(\Omega) \cap C(\bar{\Omega}) \times L^{\infty}(\Omega) \times W_0^{1,s}(\Omega) \times M(K)$  satisfies the optimal-

ity system (8)–(10). First, it is clear that  $\bar{y}$  is the state associated with  $\bar{u}$  because of the definition of  $e$ . On the other hand,  $\bar{\varphi}$  is the solution of (5), which is the same as (8) for our example. Relation (10) follows directly from the definition of  $\bar{u}$  given in (6). Finally, because of the definition of  $\bar{\mu}$  and  $K$ , (9) can be written in the form

$$\sum_{k=1}^{\infty} \frac{1}{k^2} z(x^k) \leq \sum_{k=1}^{\infty} \frac{1}{k^2} \quad \forall z \in C(K) \text{ such that } -1 \leq z(x) \leq +1 \quad \forall x \in K,$$

which obviously is satisfied.

Now we prove that  $\bar{u}$  is not continuous at  $x = 0$ . Notice that  $\bar{\varphi}(x^k) = +\infty$  for every  $k \in \mathbb{N}$ , because  $\phi(0) = +\infty$ . Therefore, (6) implies that  $\bar{u}(x^k) = -M$  for every  $k$ . Since  $x^k \rightarrow 0$ , the continuity of  $\bar{u}$  at  $x = 0$  requires that  $u(x) \rightarrow -M$  as  $x \rightarrow 0$ . However, we have for  $\xi^j = (x^j + x^{j+1})/2$  that

$$\lim_{j \rightarrow \infty} \bar{u}(\xi^j) = \bar{\psi}(0) + \sum_{k=1}^{\infty} \frac{1}{k^2} \psi_k(0) + \sum_{k=1}^{\infty} \frac{1}{k^2} \phi(x^k) > -M. \quad (8)$$

Nevertheless, we are able to improve the regularity result of Theorem 4.1.

**Theorem 4.3.** *Suppose that  $\bar{u}$  is a strict local minimum of (P) in the sense of the  $L^2(\Omega)$  topology. We also assume that Assumptions (A1)–(A3) hold. Then  $\bar{u}$  belongs to  $H_0^1(\Omega)$ .*

Let us remark that any global solution of (P) is a local solution of (P) in the sense of  $L^2(\Omega)$ , but we can expect to have more local or global solutions in the sense of  $L^2(\Omega)$ . Theorem 3.5 implies that  $\bar{u}$  is at least a strict local minimum in the sense of  $L^2(\Omega)$ , if the sufficient second-order optimality conditions are satisfied at  $\bar{u}$ . This guarantees that  $\bar{u}$  is the unique global solution in an  $L^2(\Omega)$ -neighborhood, but it does not exclude the case that  $\bar{u}$  is an accumulation point of different local minima.

*Proof of Theorem 4.3.* Let  $\varepsilon_{\bar{u}} > 0$  be chosen such that  $\bar{u}$  is a strict global minimum of (P) in the closed ball  $\bar{B}_{\varepsilon_{\bar{u}}}(\bar{u}) \subset L^2(\Omega)$ . This implies that  $\bar{u}$  is the unique global solution of the problem

$$(P_0) \left\{ \begin{array}{l} \min J(u) \\ \text{subject to } (y_u, u) \in (C(\bar{\Omega}) \cap H_0^1(\Omega)) \times L^\infty(\Omega), \\ \alpha \leq u(x) \leq \beta \quad \text{for a.e. } x \in \Omega, \quad \|u - \bar{u}\|_{L^2(\Omega)} \leq \varepsilon_{\bar{u}} \\ a \leq y_u(x) \leq b \quad \forall x \in K, \end{array} \right.$$

where  $y_u$  is the solution of (1).

Now we take a sequence  $\{x_k\}_{k=1}^\infty$  that is dense in  $K$  and consider the family of control problems

$$(P_k) \left\{ \begin{array}{l} \min J(u) \\ \text{subject to } (y_u, u) \in (C(\bar{\Omega}) \cap H_0^1(\Omega)) \times L^\infty(\Omega), \\ \alpha \leq u(x) \leq \beta \quad \text{for a.e. } x \in \Omega, \quad \|u - \bar{u}\|_{L^2(\Omega)} \leq \varepsilon_{\bar{u}} \\ a \leq y_u(x_j) \leq b, \quad 1 \leq j \leq k. \end{array} \right.$$

It is obvious that  $\bar{u}$  is a feasible control for every problem  $(P_k)$ , hence the existence of a global minimum  $u_k$  of  $(P_k)$  follows easily by a standard argumentation. The proof is split into three steps: First, we prove that the sequence  $\{u_k\}_{k=1}^\infty$  converges to  $\bar{u}$  strongly in  $L^2(\Omega)$ ; in a second step we will check that the linearized Slater condition corresponding to problem  $(P_k)$  holds for all sufficiently large  $k$ . Finally, we deduce the boundedness of  $\{u_k\}_{k=1}^\infty$  in  $H_0^1(\Omega)$ .

*Step 1 – Convergence of  $\{u_k\}_{k=1}^\infty$ .* By taking a subsequence, if necessary, we can suppose that  $u_k \rightharpoonup \tilde{u}$  weakly in  $L^2(\Omega)$ . This implies that  $y_k = y_{u_k} \rightarrow \tilde{y} = y_{\bar{u}}$  strongly in  $H_0^1(\Omega) \cap C(\bar{\Omega})$ . In view of the density of  $\{x_k\}_{k=1}^\infty$  in  $K$  and using the fact that

$$a \leq y(x_j) = \lim_{k \rightarrow \infty} y_k(x_j) \leq b \quad \forall j \geq 1,$$

we get  $a \leq y(x) \leq b$  for every  $x \in K$ . Clearly,  $\tilde{u}$  is a feasible control for problem  $(P_0)$ . Since  $\bar{u}$  is the solution of  $(P_0)$ ,  $u_k$  is a solution of  $(P_k)$  and  $\bar{u}$  is a feasible control for every problem  $(P_k)$ . Therefore,  $J(u_k) \leq J(\bar{u})$  and we get

$$J(\bar{u}) \leq J(\tilde{u}) \leq \liminf_{k \rightarrow \infty} J(u_k) \leq \limsup_{k \rightarrow \infty} J(u_k) \leq \liminf_{k \rightarrow \infty} J(\bar{u}) = J(\bar{u}),$$

which implies that  $\bar{u} = \tilde{u}$  and  $\lim J(u_k) = J(\bar{u})$ . Hence the strong convergence  $u_k \rightarrow \bar{u}$  in  $L^2(\Omega)$  follows from the last equality.

*Step 2 – The linearized Slater condition for  $(P_k)$  is satisfied at  $u_k$ .* This follows from a standard idea: With some  $\varepsilon > 0$ , one takes the control  $u_{0,\varepsilon} = \varepsilon(u_0 - \bar{u}) + \bar{u}$ , where  $u_0$  is taken from the Slater condition (A3). Then it is not difficult to show that the linearized Slater condition for  $(P_k)$  is satisfied for all sufficiently large  $k$ .

*Step 3 –  $\{u_k\}_{k=1}^\infty$  is bounded in  $H_0^1(\Omega)$ .* The strong convergence  $u_k \rightarrow \bar{u}$  in  $L^2(\Omega)$  implies that  $\|u_k - \bar{u}\|_{L^2(\Omega)} < \varepsilon_{\bar{u}}$  holds for all sufficiently large  $k$ . Therefore,  $u_k$  is a local minimum of the problem

$$(Q_k) \left\{ \begin{array}{l} \min J(u) \\ \text{subject to } (y_u, u) \in (C(\bar{\Omega}) \cap H_0^1(\Omega)) \times L^\infty(\Omega), \\ \alpha \leq u(x) \leq \beta \quad \text{for a.e. } x \in \Omega, \\ a \leq y_u(x_j) \leq b, \quad 1 \leq j \leq k. \end{array} \right.$$

Next, we apply Theorem 3.4 and deduce

$$u_k(x) = \text{Proj}_{[\alpha, \beta]} \left( -\frac{1}{N} \varphi_k(x) \right) = \min \left\{ \max \left\{ \alpha, -\frac{1}{N} \varphi_k(x) \right\}, \beta \right\}, \quad (9)$$

with

$$\varphi_k = \varphi_{k,0} + \sum_{j=1}^k \lambda_{k,j} \varphi_{k,j}. \quad (10)$$

Above,  $\{\lambda_{k,j}\}_{j=1}^k$  are the Lagrange multipliers, more precisely

$$\mu_k = \sum_{j=1}^k \lambda_{k,j} \delta_{x_j}, \quad \text{with } \lambda_{k,j} = \begin{cases} \geq 0 & \text{if } y_k(x_j) = b, \\ \leq 0 & \text{if } y_k(x_j) = a. \end{cases} \quad (11)$$

Finally,  $\varphi_{k,0}$  and  $\{\varphi_{k,j}\}_{j=1}^k$  are given by

$$\begin{cases} -\Delta \varphi_{k,0} + a'_0(y_k(x)) \varphi_{k,0} &= y_k - y_d \quad \text{in } \Omega, \\ \varphi_{k,0} &= 0 \quad \text{on } \Gamma, \end{cases} \quad (12)$$

$$\begin{cases} -\Delta \varphi_{k,j} + a'_0(y_k(x)) \varphi_{k,j} &= \delta_{x_j} \quad \text{in } \Omega, \\ \varphi_{k,j} &= 0 \quad \text{on } \Gamma. \end{cases} \quad (13)$$

It is known that the linearized Slater condition implies the boundedness of the Lagrange multipliers

$$\exists C > 0 \text{ such that } \|\mu_k\|_{M(K)} = \sum_{j=1}^k |\lambda_{k,j}| \leq C \quad \forall k. \quad (14)$$

We suppress the associated proof. Now (10), (12) and (13) lead to

$$\begin{cases} -\Delta \varphi_k + a'_0(y_k(x)) \varphi_k &= y_k - y_d + \mu_k \quad \text{in } \Omega, \\ \varphi_k &= 0 \quad \text{on } \Gamma. \end{cases} \quad (15)$$

Define

$$C_{\alpha,\beta} = |\alpha| + |\beta| + 1 \quad \text{and} \quad v_k(x) = \text{Proj}_{[-C_{\alpha,\beta}, +C_{\alpha,\beta}]} \left( -\frac{1}{N} \varphi_k(x) \right).$$

From the last relation and (9) it follows that

$$u_k(x) = \text{Proj}_{[\alpha, \beta]}(v_k(x)).$$

The goal is to prove that  $\{v_k\}_{k=1}^\infty$  is bounded in  $H_0^1(\Omega)$ , which implies the boundedness of  $\{u_k\}_{k=1}^\infty$  in the same space. The last claim is an immediate consequence of

$$|\nabla u_k(x)| \leq |\nabla v_k(x)| \quad \text{for a.e. } x \in \Omega.$$

If  $\{u_k\}_{k=1}^\infty$  is bounded in  $H_0^1(\Omega)$ , then obviously  $\bar{u}$  belongs to  $H_0^1(\Omega)$ .

Let us prove the boundedness of  $\{v_k\}_{k=1}^\infty$  in  $H_0^1(\Omega)$ . Notice that the solution of the Dirichlet problem for  $-\Delta$  and a Lipschitz boundary  $\Gamma$  belongs to  $W_0^{1,r}(\Omega)$  if the right-hand side belongs to  $W^{-1,r}(\Omega)$  for any  $n < r < n + \varepsilon_n$ , with  $\varepsilon_n > 0$  depending on  $n$  and  $n = 2$  or  $3$ ; see Jerison and Kenig [8] and Mateos [9]. Since  $L^2(\Omega) \subset W^{-1,6}(\Omega)$  for  $n \leq 3$ , we have that  $\varphi_{k,0} \in W_0^{1,r}(\Omega)$  holds for all  $r \leq 6$  in the range indicated above. From this, it follows  $\varphi_k \in W_0^{1,s}(\Omega) \cap W^{1,r}(\Omega \setminus S_k)$ , where  $S_k$  is the set of points  $x_j$  such that  $\lambda_{k,j} \neq 0$ . Taking into account that  $v_k$  is constant

in a neighborhood of every point  $x_j \in S_k$ , we obtain  $v_k \in W_0^{1,r}(\Omega) \subset C(\bar{\Omega})$ . Multiplying equation (15) by  $-v_k$  and integrating by parts, we get

$$-\int_{\Omega} (\nabla v_k \cdot \nabla \varphi_k + a'_0(y_k)v_k \varphi_k) dx = -\int_{\Omega} (y_k - y_d)v_k dx - \sum_{j=1}^k \lambda_{k,j} v_k(x_j). \quad (16)$$

The definition of  $v_k$  implies for almost all  $x \in \Omega$

$$\nabla v_k(x) = \begin{cases} -\frac{1}{N} \nabla \varphi_k(x) & \text{if } -C_{\alpha,\beta} \leq -\frac{1}{N} \varphi_k(x) \leq +C_{\alpha,\beta} \\ 0 & \text{otherwise.} \end{cases} \quad (17)$$

Invoking this property in (16) along with the boundedness of  $\{y_k\}_{k=1}^{\infty}$  in  $C(\bar{\Omega})$ , the estimate  $\|v_k\|_{L^\infty(\Omega)} \leq C_{\alpha,\beta}$  and assumptions (A3), (A2) we deduce

$$\lambda_A N \int_{\Omega} |\nabla v_k|^2 dx \leq \|\psi_M\|_{L^2(\Omega)} \|v_k\|_{L^2(\Omega)} + C \sum_{j=1}^k |\lambda_{k,j}| \|v_k\|_{L^\infty(\Omega)} \leq C',$$

which implies that  $\{v_k\}_{k=1}^{\infty}$  is bounded in  $H_0^1(\Omega)$  as required.

## 5. On the uniqueness of the Lagrange multiplier $\bar{\mu}$

Finally, we provide a sufficient condition for the uniqueness of the Lagrange multiplier associated to the state constraints. We also analyze some situations under which the condition is satisfied.

**Theorem 5.1.** *Assume (A1)–(A3) and the existence of some  $\varepsilon > 0$  such that*

$$T : L^2(\Omega_{\varepsilon}) \longrightarrow C(K_0), \quad v \mapsto z_v, \quad \text{has a dense range ,} \quad (1)$$

where

$$\Omega_{\varepsilon} = \{x \in \Omega : \alpha + \varepsilon < \bar{u}(x) < \beta - \varepsilon\},$$

$z_v \in H_0^1(\Omega) \cap C(\bar{\Omega})$  satisfies

$$\begin{cases} -\Delta z_v + a'_0(\bar{y}) z_v &= v \quad \text{in } \Omega \\ z_v &= 0 \quad \text{on } \Gamma, \end{cases} \quad (2)$$

and  $v$  is extended by zero to the whole domain  $\Omega$ . Then there exists a unique Lagrange multiplier  $\mu \in M(K)$  such that (8)–(10) holds.

*Proof.* Assume to the contrary that  $\bar{\mu}_i$ ,  $i = 1, 2$ , are two Lagrange multipliers associated with the state constraints corresponding to the optimal control  $\bar{u}$ . Then (14) holds for  $\bar{\mu} = \bar{\mu}_i$ ,  $i = 1, 2$ . Take an arbitrary  $v \in L^\infty(\Omega_{\varepsilon}) \setminus \{0\}$ . Then we have

$$\alpha \leq u_{\rho}(x) = \bar{u}(x) + \rho v(x) \leq \beta \quad \text{for a.e. } x \in \Omega, \quad \forall |\rho| < \frac{\varepsilon}{\|v\|_{L^\infty(\Omega_{\varepsilon})}},$$

where  $v$  is extended by zero to the whole domain  $\Omega$ . Taking  $u = u_\rho$  in (14), with  $\rho$  positive and negative respectively, and remembering that  $\text{supp } \bar{\mu}_i \subset K_0$  (Remark 3.2), we deduce

$$J'(\bar{u})v + \int_{K_0} z_v(x) d\bar{\mu}_i(x) = \frac{\partial \mathcal{L}}{\partial u}(\bar{u}, \bar{\mu}_i)v = 0, \quad i = 1, 2,$$

which leads to

$$\langle \bar{\mu}_1, Tv \rangle = -J'(\bar{u})v = \langle \bar{\mu}_2, Tv \rangle \quad \forall v \in L^\infty(\Omega_\varepsilon).$$

Since  $L^\infty(\Omega_\varepsilon)$  is dense in  $L^2(\Omega_\varepsilon)$  and  $T(L^2(\Omega_\varepsilon))$  is dense in  $C(K_0)$ , this identity implies  $\bar{\mu}_1 = \bar{\mu}_2$ .  $\square$

**Remark 5.1.** For a finite set  $K = \{x_j\}_{j=1}^n$ , assumption (1) is equivalent to the independence of the gradients  $\{G'_j(\bar{u})\}_{j \in I_0}$  in  $L^2(\Omega_\varepsilon)$ , where the functions  $G_j : L^2(\Omega_\varepsilon) \rightarrow \mathbb{R}$  are defined by  $G_j(u) = g(x_j, y_u(x_j))$  and  $I_0$  is the set of indexes  $j$  corresponding to active constraints. It is a regularity assumption on the control problem at  $\bar{u}$ . This type of assumption was introduced by the authors in [4] to analyze control constrained problems with finitely many state constraints. The first author proved in [1] that, under very general hypotheses, this assumption is equivalent to the Slater condition in the case of a finite number of pointwise state constraints.

**Remark 5.2.** We are able to show that (1) holds under some assumptions on  $\bar{u}$  and the set of points  $K_0$  where the state constraint is active. For instance, assume in addition to (A1)–(A3) that also

1. the Lebesgue measure of  $K_0$  is zero and that
2. there exists  $\varepsilon > 0$  such that, for every open connected component  $\mathcal{A}$  of  $\Omega \setminus K_0$ , the set  $\mathcal{A} \cap \Omega_\varepsilon$  has a nonempty interior.

Then the regularity assumption (1) is satisfied, as it will be proven in a forthcoming paper. We are also able to show that the regularity hypothesis (1) is stronger than the linearized Slater assumption (A3).

**Remark 5.3.** We know  $\bar{u} \in C(\bar{\Omega} \setminus K_0)$  from Theorem 4.1, hence the assumption 2 of the theorem holds if  $\bar{u}$  is not identically equal to  $\alpha$  or  $\beta$  in any open connected component  $\mathcal{A} \subset \Omega \setminus K_0$ . Indeed, since  $\bar{u} \in C(\mathcal{A})$  and  $\bar{u} \not\equiv \alpha$  and  $\bar{u} \not\equiv \beta$  in  $\mathcal{A}$ , there exists  $x_0 \in \mathcal{A}$  such that  $\alpha < \bar{u}(x_0) < \beta$ . Then the continuity of  $\bar{u}$  implies the existence of  $\varepsilon > 0$  such that  $\mathcal{A} \cap \Omega_\varepsilon$  contains a ball  $B_\rho(x_0)$ .

## References

- [1] E. CASAS, Necessary and sufficient optimality conditions for elliptic control problems with finitely many pointwise state constraints, ESAIM:COCV, (to appear).
- [2] E. CASAS, J. DE LOS REYES, AND F. TRÖLTSCH, Sufficient second order optimality conditions for semilinear control problems with pointwise state constraints, SIAM J. Optim., (to appear).

- [3] E. CASAS AND M. MATEOS, *Second order optimality conditions for semilinear elliptic control problems with finitely many state constraints*, SIAM J. Control Optim., 40 (2002), pp. 1431–1454.
- [4] E. CASAS AND F. TRÖLTZSCH, *Second order necessary optimality conditions for some state-constrained control problems of semilinear elliptic equations*, App. Math. Optim., 39 (1999), pp. 211–227.
- [5] E. CASAS, F. TRÖLTZSCH, AND A. UNGER, *Second order sufficient optimality conditions for some state-constrained control problems of semilinear elliptic equations*, SIAM J. Control Optim., 38 (2000), pp. 1369–1391.
- [6] D. GILBARG AND N. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin-Heidelberg-New York, 1977.
- [7] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston-London-Melbourne, 1985.
- [8] D. JERISON AND C. KENIG, *The inhomogeneous Dirichlet problem in Lipschitz domains*, J. Funct. Anal., 130 (1995), pp. 161–219.
- [9] M. MATEOS, *Problemas de control óptimo gobernados por ecuaciones semilineales con restricciones de tipo integral sobre el gradiente del estado*, PhD thesis, University of Cantabria, 2000.
- [10] H. MAURER AND J. ZOWE, *First- and second-order conditions in infinite-dimensional programming problems*, Math. Programming, 16 (1979), pp. 98–110.
- [11] G. STAMPACCHIA, *Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus*, Ann. Inst. Fourier (Grenoble), 15 (1965), pp. 189–258.

Eduardo Casas  
 Departamento de Matemática  
 Aplicada y Ciencias de la Computación  
 E.T.S.I. Industriales y de Telecomunicación  
 Universidad de Cantabria  
 39005 Santander, Spain  
 e-mail: [eduardo.casas@unican.es](mailto:eduardo.casas@unican.es)

Fredi Tröltzsch  
 Institut für Mathematik  
 Technische Universität Berlin  
 D-10623 Berlin, Germany  
 e-mail: [troeltzsch@math.tu-berlin.de](mailto:troeltzsch@math.tu-berlin.de)

# Fast and Strongly Localized Observation for a Perturbed Plate Equation

Nicolae Cîndea and Marius Tucsnak

**Abstract.** The aim of this work is to study the exact observability of a perturbed plate equation. A fast and strongly localized observation result was proven using a perturbation argument of an Euler-Bernoulli plate equation and a unique continuation result for bi-Laplacian.

**Mathematics Subject Classification (2000).** 35B37, 93B05, 93B07.

**Keywords.** Plate equation, observability, perturbation, Carleman estimation.

## 1. Introduction

Various observability and controllability properties for the system of partial differential equations modeling the vibrations of an Euler-Bernoulli plate have been investigated in the literature. In most of the existing references it assumed that the observation region satisfies the geometric optics condition of Bardos, Lebeau and Rauch [1], which is known to be necessary and sufficient for the exact observability of the wave equation (see, for instance, Lasiecka and Triggiani [9], Lebeau [10], Burq and Zworski [2] and references therein). In the case of *internal control*, the first result asserting that exact observability for the Schrödinger equation holds for an *arbitrarily small control region* has been given by Jaffard [7], who shows, in particular, that for systems governed by the Schrödinger equation in a rectangle we have exact internal observability with an arbitrary observation region and in arbitrarily small time. Jaffard's method has been adapted by Komornik [8] to an  $n$ -dimensional context. The similar results for boundary observation have been given in Ramdani, Takahashi, Tenenbaum and Tucsnak [13] and Tenenbaum and Tucsnak [14]. The aim of this work is to extend some of these results, namely those in [7], for the case of an Euler-Bernoulli plate perturbed by a zero-order term. Note that the above-mentioned papers tackling arbitrarily small observation regions use the explicit knowledge of the eigenvalues and of the eigenvectors of the Laplace operator in rectangular domains. Such an information is not available

for the plate equations perturbed by lower-order terms. On the other hand, as far as we know, the method based on Carleman estimates, which is generally used to tackle lower-order terms, does not yield exact observability with arbitrarily small observation region. This is why we consider a different method, in which our problem is tackled as a perturbation of the case considered in [7] and [8], using recent results from Hadd [4] and Tucsnak and Weiss [15].

Let us now give the precise statement of the problem and of the main results. In the remaining part of this work  $n \in \mathbb{N}$  and  $\Omega$  is a rectangular domain in  $\mathbb{R}^n$ , say

$$\Omega = [0, a_1] \times [0, a_2] \times \cdots \times [0, a_n],$$

with  $a_1, a_2, \dots, a_n > 0$ .

We consider the initial and boundary value problem

$$\frac{\partial^2 \eta}{\partial t^2} + \Delta^2 \eta + a\eta = 0, \quad \text{in } \Omega \times (0, \infty) \quad (1.1)$$

$$\eta = \Delta \eta = 0, \quad \text{on } \Gamma \times (0, \infty) \quad (1.2)$$

$$\eta(0) = f, \quad \dot{\eta}(0) = g \quad \text{in } \Omega, \quad (1.3)$$

where  $a \in L^\infty(\Omega)$ ,  $f \in H^2(\Omega) \cap H_0^1(\Omega)$  and  $g \in L^2(\Omega)$ . For  $n = 2$  the above equations model the vibrations of an Euler-Bernoulli plate with a hinged boundary. The output of this system is

$$y(t) = \dot{\eta}(\cdot, t)|_{\mathcal{O}}, \quad (1.4)$$

where  $\mathcal{O}$  is an open subset of  $\Omega$ . Here, and in the remaining part of this paper we denote

$$\dot{\eta} = \frac{\partial \eta}{\partial t}.$$

Our main result is:

**Theorem 1.1.** *For any subset  $\mathcal{O} \subset \Omega$  the system (1.1)–(1.4) is exactly observable in time any time  $\tau > 0$ , i.e., there exists a constant  $k_\tau > 0$  such that*

$$\int_0^\tau \|\dot{\eta}(t)\|_{L^2(\mathcal{O})}^2 dt \geq k_\tau^2 \left( \|f\|_{H^2(\Omega)}^2 + \|g\|_{L^2(\Omega)}^2 \right) \forall f \in H^2(\Omega) \cap H_0^1(\Omega), g \in L^2(\Omega).$$

The above theorem has two consequences concerning exact controllability and uniform stabilizability for the plate equations. The first one follows from a standard duality argument, see for instance, Lions [11].

**Corollary 1.2.** *For any open subset  $\mathcal{O} \subset \Omega$  the following problem*

$$\frac{\partial^2 \eta}{\partial t^2} + \Delta^2 \eta + a\eta + u\chi_{\mathcal{O}} = 0, \quad \text{in } \Omega \times (0, \infty) \quad (1.5)$$

$$\eta = \Delta \eta = 0, \quad \text{on } \Gamma \times (0, \infty) \quad (1.6)$$

$$\eta(0) = f, \quad \dot{\eta}(0) = g \quad \text{in } \Omega, \quad (1.7)$$

is exactly controllable in any time  $\tau > 0$ , i.e., for any  $\begin{bmatrix} f_1 \\ g_1 \end{bmatrix}, \begin{bmatrix} f_2 \\ g_2 \end{bmatrix} \in (H^2(\Omega) \cap H_0^1(\Omega)) \times L^2(\Omega)$  there exists a control  $u \in L^2(\mathcal{O})$  such that

$$\begin{bmatrix} \eta(0) \\ \dot{\eta}(0) \end{bmatrix} = \begin{bmatrix} f_1 \\ g_1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \eta(\tau) \\ \dot{\eta}(\tau) \end{bmatrix} = \begin{bmatrix} f_2 \\ g_2 \end{bmatrix},$$

where by  $\chi_{\mathcal{O}}(x)$  we denote the function that is 1 for  $x \in \mathcal{O}$  and 0 otherwise.

Moreover, from Theorem 1.1 and the general result in Haraux [5] it follows that the system (1.5)–(1.7) can be exponentially stabilized by using a simple feedback. More precisely, the following result holds.

**Corollary 1.3.** *Let  $\mathcal{O}$  be an open subset of  $\Omega$  and let  $a, b \in L^\infty(\Omega, [0, \infty))$  with  $b(x) \geq b_0 > 0$  for almost every  $x \in \mathcal{O}$ . Then the system determined by initial and boundary value problem (1.5)–(1.7) with  $u(x, t) = -b(x)\dot{\eta}(x, t)$ , is exponentially stable, i.e., there exist  $M, \omega > 0$  such that*

$$\|\dot{\eta}(t)\|_{L^2(\Omega)} + \|\eta(t)\|_{H^2(\Omega)} \leq M e^{-\omega t} (\|f\|_{H^2(\Omega)} + \|g\|_{L^2(\Omega)}) \quad (t \geq 0).$$

The plan of this work is as follows. In Section 2 we fix some notation and we recall some basic results. Section 3 contains the proofs of the main results. In section 4 we prove a Carleman estimate for the bi-Laplacian, which has been used for the proof of the main result.

## 2. Notation and preliminaries

In the remaining part of the paper we denote  $H = L^2(\Omega)$  and

$$H_1 = \{\varphi \in H^4(\Omega) | \varphi = \Delta\varphi = 0 \text{ on } \Gamma\}.$$

Let  $A_0 : H_1 \rightarrow H$  be the operator defined by  $A_0\varphi = \Delta^2\varphi$ ,  $\forall \varphi \in H_1$ . Let  $H_{\frac{1}{2}} = H^2(\Omega) \cap H_0^1(\Omega)$ ,  $X = H_{\frac{1}{2}} \times H$ ,  $X_1 = H_1 \times H_{\frac{1}{2}}$  and

$$A : X_1 \rightarrow X, \quad A = \begin{bmatrix} 0 & I \\ -A_0 & 0 \end{bmatrix}.$$

It is well known that  $A$  is skew-adjoint so that, according to Stone's theorem, it generates a strongly continuous group of isometries  $\mathbb{T}$  on  $X$ . By  $\|\cdot\|$  without any index we design the standard norm in  $L^2(\Omega)$ . We denote  $Y = L^2(\mathcal{O})$ , with  $\mathcal{O} \subset \Omega$  an open set. The operator  $C \in \mathcal{L}(X_1, Y)$  corresponding to the observation (1.4) is

$$C \begin{bmatrix} f \\ g \end{bmatrix} = g|_{\mathcal{O}} \quad \left( \begin{bmatrix} f \\ g \end{bmatrix} \in X \right). \quad (2.1)$$

Let  $P_0 \in \mathcal{L}(H)$  be the linear operator defined by  $P_0 f = -af$  for all  $f \in H$ , and  $P \in \mathcal{L}(X)$  given by

$$P = \begin{bmatrix} 0 & 0 \\ P_0 & 0 \end{bmatrix}, \quad P \begin{bmatrix} f \\ g \end{bmatrix} = \begin{bmatrix} 0 \\ P_0 f \end{bmatrix}.$$

We define  $A_P : \mathcal{D}(A_P) \rightarrow X$  by

$$\mathcal{D}(A_P) = \mathcal{D}(A), \quad A_P = A + P. \quad (2.2)$$

We note that

$$\|P\|_{\mathcal{L}(X)} = \sup \left\{ \|P \begin{bmatrix} f \\ g \end{bmatrix}\|_X \right\} = \sup_{\|f\| \leq 1} \|af\| \leq \|a\|_{L^\infty}.$$

We know from Pazy [12] (Theorem 1.1 p. 76) that  $A_P$  is the generator of a strongly continuous semigroup  $\mathbb{T}^P$  satisfying

$$\|\mathbb{T}_t^P\| \leq M e^{\alpha t}, \quad t \geq 0, \quad (2.3)$$

where  $\alpha = \omega + M\|P\|$ , and  $\omega$  and  $M$  are such that  $\|\mathbb{T}_t\| \leq M e^{\omega t}$  for all  $t \geq 0$ .

In this context the problem (1.1)–(1.3) can be written as a first-order equation

$$\dot{z}(t) = A_P z(t), \quad t \geq 0 \quad (2.4)$$

$$z(0) = z_0, \quad (2.5)$$

where  $z(t) = \begin{bmatrix} \eta \\ \dot{\eta}(t) \end{bmatrix}$  and  $z_0 = \begin{bmatrix} f \\ g \end{bmatrix}$ .

The proof of Theorem 1.1 is based on two abstract results, which are stated below. The first one concerns the robustness of the exact observability with respect to bounded small norm perturbations of the generator and it can be proved by a simple duality argument from Theorem 3.3 in [4].

**Proposition 2.1.** *Suppose that  $C \in \mathcal{L}(X_1, Y)$  is an admissible observation operator for  $\mathbb{T}$ . Assume that  $(A, C)$  is exactly observable in time  $\tau > 0$ , i.e., there exists  $k_\tau > 0$  such that*

$$\left( \int_0^\tau \|C\mathbb{T}_t z_0\|^2 dt \right)^{\frac{1}{2}} \geq k_\tau \|z_0\| \quad \forall z_0 \in \mathcal{D}(A).$$

Let  $P \in \mathcal{L}(X)$  and let  $\mathbb{T}^P$  be the strongly continuous semigroup generated by  $A+P$ . If there exists a constant  $\mathcal{K} > 0$  such that

$$\|P\| \leq \mathcal{K}, \quad (2.6)$$

then  $(A+P, C)$  is exactly observable in time  $\tau$ , i.e., there exists  $k_\tau^P > 0$  such that

$$\left( \int_0^\tau \|C\mathbb{T}_t^P z_0\|^2 dt \right)^{\frac{1}{2}} \geq k_\tau^P \|z_0\| \quad \forall z_0 \in \mathcal{D}(A).$$

The second result says, roughly speaking, that for systems with diagonalisable generators that in order to prove the exact observability it is sufficient to check the exact observability of the high frequency part and the observability of eigenvectors. More precisely, we have the following result, borrowed from [15].

**Proposition 2.2.** *Assume that there exists an orthonormal basis  $(\phi_k)_{k \in \mathbb{N}}$  formed of eigenvectors of  $A$  and the corresponding eigenvalues  $(\lambda_k)_{k \in \mathbb{N}}$  satisfy  $\lim \lambda_k = \infty$ . Let  $C \in \mathcal{L}(X_1, Y)$  be an admissible observation operator for  $\mathbb{T}$ . For some bounded set  $J \subset C$  denote*

$$V = \text{span} \{ \phi_k \mid \lambda_k \in J \}^\perp$$

and let  $A_V$  be the part of  $A$  in  $V$ . Let  $C_V$  be the restriction of  $C$  to  $\mathcal{D}(A_V)$ . Assume that  $(A_V, C_V)$  is exactly observable in time  $\tau_0 > 0$  and that  $C\phi \neq 0$  for every eigenvector  $\phi$  of  $A$ . Then  $(A, C)$  is exactly observable in any time  $\tau > \tau_0$ .

### 3. Main results

The proof of Theorem 1.1 follows the same idea like in [15] (Theorem 6.3.2), where a similar result is proved for the waves equation. Also, we will use an appropriate decomposition of  $X$  as a direct sum of invariant subspaces. To obtain this decomposition, we need the following characterization of the eigenvalues and eigenvectors of  $A_P$ .

**Proposition 3.1.** *With the above notation,  $\phi = [\begin{smallmatrix} \varphi \\ \psi \end{smallmatrix}] \in \mathcal{D}(A_P)$  is an eigenvector of  $A_P$ , associated to the eigenvalue  $i\mu$ , if and only if  $\varphi$  is an eigenvector of  $A_0 - P_0$ , associated to the eigenvalue  $\mu^2$ , and  $\psi = i\mu\varphi$ .*

*Proof.* Suppose that  $\mu \in \mathbb{C}$  and  $[\begin{smallmatrix} \varphi \\ \psi \end{smallmatrix}] \in X \setminus \{[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}]\}$ . According to the definition of  $A_P$  this is equivalent to

$$\begin{cases} \psi = i\mu\varphi \\ (-A_0 + P_0)\varphi = i\mu\psi. \end{cases}$$

The above conditions hold iff

$$(-A_0 + P_0)\varphi = -\mu^2\varphi \text{ and } \psi = i\mu\varphi.$$

□

Clearly,  $A_0 - P_0$  is self-adjoint and it has compact resolvent. Then  $A_0 - P_0$  is diagonalisable with an orthonormal basis  $(\varphi_k)_{k \in \mathbb{N}^*}$  of eigenvectors and the corresponding family of real eigenvalues  $(\lambda_k)_{k \in \mathbb{N}^*}$  satisfies  $\lim_{k \rightarrow \infty} |\lambda_k| = \infty$ . Since  $A_0 - P_0 + \|P_0\|I \geq 0$ , it follows that all the eigenvalues  $\lambda$  of  $A_0 - P_0$  satisfy  $\lambda > -\|P_0\|$ . Hence,  $\lim_{k \rightarrow \infty} \lambda_k = \infty$ . Without loss of generality we may assume that the sequence  $(\lambda_k)_{k \in \mathbb{N}^*}$  is non-decreasing. We extend the sequence  $(\varphi_k)$  to a sequence indexed by  $\mathbb{Z}^*$  by setting  $\varphi_k = -\varphi_{-k}$  for every  $k \in \mathbb{Z}_-$ . We introduce the real sequence  $(\mu_k)_{k \in \mathbb{Z}^*}$  by

$$\mu_k = \sqrt{|\lambda_k|} \text{ if } k > 0 \text{ and } \mu_k = -\mu_{-k} \text{ if } k < 0.$$

We denote

$$W_0 = \text{span} \left\{ \begin{bmatrix} \frac{1}{i\text{sign}(k)}\varphi_k \\ \varphi_k \end{bmatrix} \mid k \in \mathbb{Z}^*, \mu_k = 0 \right\}.$$

If  $\text{Ker}(A_0 - P_0) = \{0\}$  then of course  $W_0$  is the zero subspace of  $X$ . Let  $N \in \mathbb{N}^*$  be such that  $\lambda_N > 0$ . We denote

$$W_N = \text{span} \left\{ \begin{bmatrix} \frac{1}{i\mu_k}\varphi_k \\ \varphi_k \end{bmatrix} \mid k \in \mathbb{Z}^*, |k| < N, \mu_k \neq 0 \right\},$$

and define  $Y_N = W_0 + W_N$ . We also introduce the space

$$V_N = \text{clos span} \left\{ \begin{bmatrix} \frac{1}{i\mu_k}\varphi_k \\ \varphi_k \end{bmatrix} \mid |k| \geq N \right\}. \quad (3.1)$$

**Lemma 3.2.** *We have  $X = Y_N \oplus V_N$  and  $Y_N, V_N$  are invariant under  $\mathbb{T}^P$ .*

By  $X = Y_N \oplus V_N$  we mean that  $X = Y_N + V_N$  and  $Y_N \cap V_N = \{0\}$ .

*Proof.* Let  $A_1 : \mathcal{D}(A_0) \rightarrow H$  be defined by

$$A_1 f = \sum_{\lambda_k=0} \langle f, \varphi_k \rangle \varphi_k + \sum_{\lambda_k \neq 0} |\lambda_k| \langle f, \varphi_k \rangle \varphi_k, \quad \forall f \in \mathcal{D}(A_0).$$

Since the family  $(\varphi_k)_{k \in \mathbb{N}^*}$  is an orthonormal basis in  $H$  and each  $\varphi_k$  is an eigenvector of  $A_1$ , it follows that  $A_1$  is diagonalisable. Moreover, since the eigenvalues of  $A_1$  are strictly positive, it follows that  $A_1 > 0$ . Is easy to see that the inner product on  $X$  defined by

$$\left\langle \begin{bmatrix} f_1 \\ g_1 \end{bmatrix}, \begin{bmatrix} f_2 \\ g_2 \end{bmatrix} \right\rangle_1 = \langle A_1^{\frac{1}{2}} f_1, A_1^{\frac{1}{2}} f_2 \rangle + \langle g_1, g_2 \rangle, \quad \forall \begin{bmatrix} f_1 \\ g_1 \end{bmatrix}, \begin{bmatrix} f_2 \\ g_2 \end{bmatrix} \in X,$$

is equivalent to the original one (meaning that it induces a norm equivalent to the original norm). Let  $\mathcal{A}_1$  be the operator on  $X$  defined by

$$\mathcal{D}(\mathcal{A}_1) = H_1 \times H_{\frac{1}{2}}, \quad \mathcal{A}_1 = \begin{bmatrix} 0 & I \\ -A_1 & 0 \end{bmatrix}.$$

We can verify that  $\mathcal{A}_1$  is skew-adjoint on  $X$  (if endowed with the inner product  $\langle \cdot, \cdot \rangle_1$ ). Consequently we obtain that  $Y_N = V_N^\perp$  (with respect to this inner product  $\langle \cdot, \cdot \rangle_1$ ). It follows that  $X = Y_N \oplus V_N$ .

We still have to show that  $V_N$  and  $Y_N$  are invariant subspaces under  $\mathbb{T}^P$ . Since  $V_N$  is the closed span of a set of eigenvectors of  $A_P$ , its invariance under the action of  $\mathbb{T}^P$  is clear. If  $\mu_k = 0$ , then

$$A_P \begin{bmatrix} \frac{1}{i\text{sign}(k)} \varphi_k \\ \varphi_k \end{bmatrix} = \begin{bmatrix} \varphi_k \\ 0 \end{bmatrix} = \frac{1}{2} \left( \begin{bmatrix} \frac{1}{i\text{sign}(k)} \varphi_k \\ \varphi_k \end{bmatrix} + \begin{bmatrix} \frac{1}{i\text{sign}(-k)} \varphi_{-k} \\ \varphi_{-k} \end{bmatrix} \right) \in W_0,$$

so that  $W_0$  is invariant under  $\mathbb{T}^P$ . If  $|k| < N$  and  $\lambda_k < 0$  then

$$(A_0 - P_0)\varphi_k = -\mu_k^2 \varphi_k,$$

so that

$$A_P \begin{bmatrix} \frac{1}{i\mu_k} \varphi_k \\ \varphi_k \end{bmatrix} = \begin{bmatrix} \varphi_k \\ \frac{\mu_k}{i} \varphi_k \end{bmatrix} = i\mu_k \begin{bmatrix} \frac{1}{i\mu_k} \varphi_k \\ -\varphi_k \end{bmatrix} = i\mu_k \begin{bmatrix} \frac{1}{i\mu_{-k}} \varphi_{-k} \\ \varphi_{-k} \end{bmatrix} \in W_N.$$

If  $|k| < N$  and  $\lambda_k > 0$ , then

$$A_P \begin{bmatrix} \frac{1}{i\mu_k} \varphi_k \\ \varphi_k \end{bmatrix} = i\mu_k \begin{bmatrix} \frac{1}{i\mu_k} \varphi_k \\ \varphi_k \end{bmatrix} \in W_N.$$

Thus  $W_N$ , and hence also  $Y_N = W_0 + W_N$ , are invariant for  $\mathbb{T}$ .  $\square$

**Lemma 3.3.** *With the previous notation and (3.1), let  $N \in \mathbb{N}^*$  be such that  $\lambda_N > \|a\|_{L^\infty}$ . Let us denote by  $P_{V_N} \in \mathcal{L}(V_N, X)$  the restriction of  $P$  to  $V_N$ . Then*

$$\|P_{V_N}\| \leq \frac{\|a\|_{L^\infty}}{\sqrt{\lambda_N - \|a\|_{L^\infty}}}. \quad (3.2)$$

*Proof.* Take a finite linear combination of the vectors  $\varphi_k$  with  $k \geq N$ :

$$f = \sum_{k=N}^M \alpha_k \varphi_k, \quad (3.3)$$

so that  $\|f\|^2 = \sum_{k=N}^M |\alpha_k|^2$ . Then

$$\begin{aligned} \|\Delta f\|^2 + \langle af, f \rangle &= \int_{\Omega} \Delta f \Delta \bar{f} \, dx + \int_{\Omega} a(x) f \bar{f} \, dx \\ &= \int_{\Omega} \Delta^2 f \bar{f} + a f \bar{f} \, dx = \int_{\Omega} (A_0 - P_0) f \bar{f} \, dx \\ &= \sum_{k,l=N}^M \alpha_k \bar{\alpha}_l \langle (A_0 - P_0) \varphi_k, \varphi_l \rangle = \sum_{k=N}^M |\alpha_k|^2 \lambda_k \geq \lambda_N \|f\|^2. \end{aligned}$$

From here we see that

$$\|\Delta f\|^2 \geq (\lambda_N - \|a\|_{L^\infty}) \|f\|^2.$$

Now take  $z$  to be a finite linear combination of the eigenvectors of  $A_P$  in  $V_N$ :

$$z \in \text{span} \left\{ \begin{bmatrix} \frac{1}{i\mu_k} \varphi_k \\ \varphi_k \end{bmatrix} \mid |k| \geq N \right\},$$

so that in particular  $z \in V_N$  and  $z = \begin{bmatrix} f \\ g \end{bmatrix}$ , with  $f$  as in (3.3). Therefore

$$\begin{aligned} \|P_{V_N} z\|_X &= \|Pz\|_X = \|af\| \leq \|a\|_{L^\infty} \|f\| \\ &\leq \frac{\|a\|_{L^\infty}}{\sqrt{\lambda_N - \|a\|_\infty}} \|\Delta f\| \leq \frac{\|a\|_{L^\infty}}{\sqrt{\lambda_N - \|a\|_\infty}} \|z\|_X. \end{aligned}$$

Since all the vectors like our  $z$  are dense in  $V_N$ , it follows that the above estimate holds for all  $z \in V_N$ , and this implies the estimate in the lemma.  $\square$

**Lemma 3.4.** *Let  $a \in L^\infty(\Omega)$  and let  $u$  be a function such that*

$$\Delta^2 u + au = \mu^2 u \quad \text{in } \Omega \quad (3.4)$$

$$u = \Delta u = 0 \quad \text{on } \partial\Omega \quad (3.5)$$

and

$$u = 0 \quad \text{in } \mathcal{O}. \quad (3.6)$$

Then  $u = 0$  in all  $\Omega$ .

*Proof.* The proof of this lemma is an direct consequence of the Theorem 4.3 given in the next section. Let denote  $g = (\mu^2 - a)u \in L^2(\Omega)$ . Now we apply Theorem 4.3 for (3.4)–(3.5) and using (3.6) we obtain

$$\begin{aligned} s\lambda^2 \int_{\Omega} |\nabla(\Delta u)|^2 e^{2s\varphi} \, dx + s^4 \lambda^6 \int_{\Omega} |\nabla u|^2 e^{2s\varphi} \, dx + s^6 \lambda^8 \int_{\Omega} |u|^2 \varphi^2 e^{2s\varphi} \, dx \\ \leq C \int_{\Omega} \frac{|g|^2}{\varphi} e^{2s\varphi} \, dx. \end{aligned}$$

After some small calculations we can prove the estimate

$$\int_{\Omega} \frac{|g|^2}{\varphi} e^{2s\varphi} dx \leq C_1(\lambda) \int_{\Omega} (\mu^2 + \|a\|_{L^\infty}^2) |u|^2 \varphi^2 e^{2s\varphi} dx + C_2(\lambda),$$

where  $C_1, C_2$  depend only of  $\lambda$ . Coupling the last two equations and taking  $s \rightarrow \infty$  we obtain that  $u = 0$  in  $\Omega$ .  $\square$

*Proof of Theorem 1.1.* Let  $N \in \mathbb{N}^*$  be such that  $\lambda_N > 0$  and let  $A_N$  and  $C_N$  be the parts of  $A_P$ , respectively of  $C$ , in  $V_N$ , where  $V_N$  has been defined in (3.1). (Thus,  $A_N = (A + P)|_{V_N}$  and  $C_N = C|_{V_N}$ .) We claim that for  $N \in \mathbb{N}^*$  large enough the pair  $(A_N, C_N)$  (with state space  $V_N$ ) is exactly observable in time  $\tau_0$ .

For a given constant  $\mathcal{K} > 0$ , from the estimation (3.2), there exists a  $N \in \mathbb{N}^*$ , big enough, such that

$$\|P_{V_N}\| \leq \mathcal{K}.$$

Because  $(A, C)$  is exactly observable in any time  $\tau > 0$ , using Proposition 2.1 we obtain that  $(A_N, C_N)$  is exactly observable in time  $\tau$  in  $V_N$ .

On the other hand, if  $\phi = \begin{bmatrix} \varphi \\ \psi \end{bmatrix} \in \mathcal{D}(A_P)$  is an eigenvector of  $A_P$ , associated to the eigenvalue  $i\mu$ , such that  $C\phi = 0$  then, according to Proposition 3.1,  $\varphi \in H_1$  is an eigenvector of  $A_0 - P_0$ , associated to the eigenvalue  $\mu^2$ , i.e.,  $\varphi \in H_1$  satisfies

$$\Delta^2 \varphi + a\varphi = \mu^2 \varphi. \quad (3.7)$$

Moreover, the condition  $C\phi = 0$  is equivalent to

$$\varphi = 0 \text{ in } \mathcal{O}.$$

As shown in Lemma 3.4, the only function  $\varphi \in H_1$  satisfying above conditions is  $\varphi = 0$ . Now, from Proposition 2.2 we can conclude that  $(A, C)$  is exactly observable in any time  $\tau > 0$ .  $\square$

#### 4. A global Carleman estimate for bi-Laplacian

In this section we will prove a global Carleman estimate for bi-Laplacian, applying two times a particular case of the global Carleman estimate proved in [6].

Let  $\Omega$  be a nonempty open set of class  $C^2$ . Let  $y \in H^2(\Omega) \cap H_0^1(\Omega)$  be the solution of the problem

$$\Delta y = f, \quad \text{in } \Omega \quad (4.1)$$

$$y = 0, \quad \text{on } \partial\Omega, \quad (4.2)$$

where  $f \in L^2(\Omega)$ . We use the following classic lemma stated in [6], and proved in [3].

**Lemma 4.1.** *Let  $\mathcal{O}$  be an nonempty open set  $\mathcal{O} \subset \Omega$ . Then there exists a function  $\psi \in C^2(\overline{\Omega})$  such that*

$$\psi = 0, \quad \text{on } \partial\Omega \quad (4.3)$$

$$\psi(x) > 0, \quad \forall x \in \Omega \quad (4.4)$$

$$|\nabla\psi(x)| > 0, \quad \forall x \in \overline{\Omega \setminus \mathcal{O}}. \quad (4.5)$$

We consider a weight function

$$\varphi(x) = e^{\lambda\psi(x)}, \quad (4.6)$$

where  $\lambda \in \mathbb{R}$ ,  $\lambda \geq 1$  will be chosen later. The following theorem is a particular case of the Carleman estimate proved by Imanuvilov-Puel in [6] for the general elliptic operators.

**Theorem 4.2.** *Assume that the hypotheses (4.3)–(4.6) are verified and let  $y \in H^2(\Omega) \cap H_0^1(\Omega)$  be the solution of (4.1)–(4.2). Then there exists a constant  $C > 0$  independent of  $s$  and  $\lambda$ , and parameters  $\hat{\lambda} > 1$  and  $\hat{s} > 1$  such that for all  $\lambda \geq \hat{\lambda}$  and for all  $s > \hat{s}$  we have*

$$\begin{aligned} & \int_{\Omega} |\nabla y|^2 e^{2s\varphi} dx + s^2 \lambda^2 \int_{\Omega} |y|^2 \varphi^2 e^{2s\varphi} dx \\ & \leq C \left( \frac{1}{s\lambda^2} \int_{\Omega} \frac{|f|^2}{\varphi} e^{2s\varphi} dx + \int_{\mathcal{O}} (|\nabla y|^2 + s^2 \lambda^2 \varphi^2 |y|^2) e^{2s\varphi} dx \right). \end{aligned} \quad (4.7)$$

Let  $u \in H_1$  be the solution of the problem

$$\Delta^2 u = g, \quad \text{in } \Omega \quad (4.8)$$

$$u = \Delta u = 0, \quad \text{on } \partial\Omega, \quad (4.9)$$

where  $g \in L^2(\Omega)$ .

**Theorem 4.3.** *Let  $\psi \in C^2(\overline{\Omega})$  be a function such that (4.3)–(4.5) are verified, let  $\varphi$  given by (4.6), and let  $u \in H_1$  be the solution of (4.8)–(4.9). Then there exist  $\hat{s} > 1$ ,  $\lambda > 1$  and a constant  $C > 0$  independent of  $s \geq \hat{s}$  such that*

$$\begin{aligned} & s\lambda^2 \int_{\Omega} (|\nabla(\Delta u)|^2 e^{2s\varphi} + s^3 \lambda^4 |\nabla u|^2 e^{2s\varphi} + s^5 \lambda^6 |u|^2 \varphi^2 e^{2s\varphi}) \leq C \left( \int_{\Omega} \frac{|g|^2}{\varphi} e^{2s\varphi} dx \right. \\ & \left. + s\lambda^2 \int_{\mathcal{O}} (|\nabla(\Delta u)|^2 + s^2 \lambda^2 \varphi^2 |\Delta u|^2 + s^3 \lambda^4 |\nabla u|^2 + s^5 \lambda^6 \varphi^2 |u|^2) e^{2s\varphi} dx \right). \end{aligned} \quad (4.10)$$

*Proof.* We denote  $y = \Delta u$ . Then (4.8) and the last part of (4.9) can be written as

$$\Delta y = g, \quad \text{in } \Omega \quad (4.11)$$

$$y = 0, \quad \text{on } \partial\Omega \quad (4.12)$$

Applying the Theorem 4.2 there exist  $s_1 > 1$ ,  $\lambda_1 > 1$  and  $C_1 > 0$  independent of  $s$  and  $\lambda$  such that for all  $s \geq s_1$ ,  $\lambda \geq \lambda_1$  the following estimate is satisfied

$$\begin{aligned} & s\lambda^2 \int_{\Omega} |\nabla y|^2 e^{2s\varphi} dx + s^3\lambda^4 \int_{\Omega} |y|^2 \varphi^2 e^{2s\varphi} dx \\ & \leq C_1 \left( \int_{\Omega} |g|^2 \varphi^{-1} e^{2s\varphi} dx + \int_{\mathcal{O}} (s\lambda^2 |\nabla y|^2 + s^3\lambda^4 \varphi^2 |y|^2) e^{2s\varphi} dx \right). \end{aligned}$$

Replacing  $y$  with  $\Delta u$  in the previous estimate we obtain

$$\begin{aligned} & s\lambda^2 \int_{\Omega} |\nabla(\Delta u)|^2 e^{2s\varphi} dx + s^3\lambda^4 \int_{\Omega} |\Delta u|^2 \varphi^2 e^{2s\varphi} dx \\ & \leq C_1 \left( \int_{\Omega} |g|^2 \varphi^{-1} e^{2s\varphi} dx + \int_{\mathcal{O}} (s\lambda^2 |\nabla(\Delta u)|^2 + s^3\lambda^4 \varphi^2 |\Delta u|^2) e^{2s\varphi} dx \right). \quad (4.13) \end{aligned}$$

Now consider the problem

$$\Delta u = y, \quad \text{in } \Omega \quad (4.14)$$

$$u = 0, \quad \text{on } \partial\Omega, \quad (4.15)$$

and apply the Theorem 4.2. Then there exist  $C_2 > 0$ ,  $s_2 > 1$ ,  $\lambda_2 > 1$  such that for  $s \geq s_2$  and  $\lambda \geq \lambda_2$  we have

$$\begin{aligned} & s\lambda^2 \int_{\Omega} |\nabla u|^2 e^{2s\varphi} dx + s^3\lambda^4 \int_{\Omega} |u|^2 \varphi^2 e^{2s\varphi} dx \\ & \leq C_2 \left( \int_{\Omega} |\Delta u|^2 \varphi^{-1} e^{2s\varphi} dx + \int_{\mathcal{O}} (s\lambda^2 |\nabla u|^2 + s^3\lambda^4 \varphi^2 |u|^2) e^{2s\varphi} dx \right) \\ & \leq C_3 \left( \int_{\Omega} |\Delta u|^2 \varphi^2 e^{2s\varphi} dx + \int_{\mathcal{O}} (s\lambda^2 |\nabla u|^2 + s^3\lambda^4 \varphi^2 |u|^2) e^{2s\varphi} dx \right). \quad (4.16) \end{aligned}$$

We denote  $\lambda = \max\{\lambda_1, \lambda_2\}$  and  $\hat{s} = \max\{s_1, s_2\}$ . For  $s \geq \hat{s}$ , combining (4.13) and (4.16) we have

$$\begin{aligned} & s\lambda^2 \int_{\Omega} |\nabla(\Delta u)|^2 e^{2s\varphi} dx + \frac{s^3\lambda^4}{C_3} \left( s\lambda^2 \int_{\Omega} |\nabla u|^2 e^{2s\varphi} dx + s^3\lambda^4 \int_{\Omega} |u|^2 \varphi^2 e^{2s\varphi} dx \right) \\ & \quad - s^3\lambda^4 \left( \int_{\mathcal{O}} (s\lambda^2 |\nabla u|^2 + s^3\lambda^4 \varphi^2 |u|^2) e^{2s\varphi} dx \right) \\ & \leq C_1 \left( \int_{\Omega} |g|^2 \varphi^{-1} e^{2s\varphi} dx + \int_{\mathcal{O}} (s\lambda^2 |\nabla(\Delta u)|^2 + s^3\lambda^4 \varphi^2 |\Delta u|^2) e^{2s\varphi} dx \right). \quad (4.17) \end{aligned}$$

How  $\lambda$  is fixed in 4.17, we affirm that exists a constant  $C > 0$  such that (4.10) is verified. So, the proof of the theorem is completed.  $\square$

## References

- [1] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control and stabilization of waves from the boundary*, SIAM J. Control. and Optim., 30 (1992), pp. 1024–1065.
- [2] N. Burq and M. Zworski, *Geometric control in the presence of a black box*, J. Amer. Math. Soc., 17 (2004), pp. 443–471.
- [3] A.V. Fursikov and O.Y. Imanuvilov, *Controllability of evolution equations*, vol. 34 of Lecture Notes Series, Seoul National University Research Institute of Mathematics Global Analysis Research Center, Seoul, 1996.
- [4] S. Hadd, *Exact controllability of infinite dimensional systems persists under small perturbations*, J. Evol. Equ., 5 (2005), pp. 545–555.
- [5] A. Haraux, *Une remarque sur la stabilisation de certains systèmes du deuxième ordre en temps*, Portugal. Math., 46 (1989), pp. 245–258.
- [6] O.Y. Imanuvilov and J.-P. Puel, *Global Carleman estimates for weak solutions of elliptic nonhomogeneous Dirichlet problems*, Int. Math. Res. Not., (2003), pp. 883–913.
- [7] S. Jaffard, *Contrôle interne exact des vibrations d'une plaque rectangulaire*, Port. Math., 47 (1990), pp. 423–429.
- [8] V. Komornik, *On the exact internal controllability of a Petrowsky system*, J. Math. Pures Appl. (9), 71 (1992), pp. 331–342.
- [9] I. Lasiecka and R. Triggiani, *Exact controllability and uniform stabilization of Euler-Bernoulli equations with boundary control only in  $\Delta w|_{\Sigma}$* , Boll. Un. Mat. Ital. B (7), 5 (1991), pp. 665–702.
- [10] G. Lebeau, *Contrôle de l'équation de Schrödinger*, J. Math. Pures Appl. (9), 71 (1992), pp. 267–291.
- [11] J.-L. Lions, *Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués. Tome 1*, vol. 8 of Recherches en Mathématiques Appliquées, Masson, Paris, 1988.
- [12] A. Pazy, *Semigroups of linear operators and applications to partial differential equations*, vol. 44 of Applied Mathematical Sciences, Springer-Verlag, New York, 1983.
- [13] K. Ramdani, T. Takahashi, G. Tenenbaum, and M. Tucsnak, *A spectral approach for the exact observability of infinite-dimensional systems with skew-adjoint generator*, J. Funct. Anal., 226 (2005), pp. 193–229.
- [14] G. Tenenbaum and M. Tucsnak, *Fast and strongly localized observation for the Schrödinger equation*, Transactions of the American Mathematical Society, to appear (2008).
- [15] M. Tucsnak and G. Weiss, *Observability and Controllability of Infinite Dimensional Systems*, to appear.

Nicolae Cîndea and Marius Tucsnak  
Université Henri Poincaré Nancy1, Institut Élie Cartan  
Département de Mathématiques (IECN), B.P. 239  
F-54506 Vandoeuvre les Nancy Cedex, France  
e-mail: Nicolae.Cindea@iecn.u-nancy.fr  
Marius.Tucsnak@iecn.u-nancy.fr

# Representations, Composition, and Decomposition of $C^{1,1}$ -hypersurfaces

Michel C. Delfour

**Abstract.** We revisit and expand the *intrinsic* and *parametric* representations of hypersurfaces with application to the theory of thin and asymptotic shells. A central issue is the minimal smoothness of the midsurface to still make sense of asymptotic *membrane shell* and *bending equations* without ad hoc mechanical or mathematical assumptions. This is possible for a  $C^{1,1}$ -midsurface with or without boundary and without local maps, local bases, and Christoffel symbols via the purely intrinsic methods developed by Delfour and Zolésio starting with [14] in 1992. Anicic, Le Dret, and Raoult [1] introduced in 2004 a family of surfaces  $\omega$  that are the image of a connected bounded open Lipschitzian domain in  $\mathbf{R}^2$  by a bi-Lipschitzian mapping with the assumption that the normal field is globally Lipschitzian. From this, they construct a *tubular neighborhood* of thickness  $2h$  around the surface and show that for sufficiently small  $h$  the associated *tubular neighborhood mapping* is bi-Lipschitzian. We prove that such surfaces are  $C^{1,1}$ -surfaces with a bounded measurable second fundamental form. We show that the tubular neighborhood can be completely described by the algebraic distance function to  $\omega$  and that it is generally not a Lipschitzian domain in  $\mathbf{R}^3$  by providing the example of a plate around a flat surface  $\omega$  verifying all their assumptions. Therefore, the  $G_1$ -join of  $K$ -regular patches in the sense of Le Dret [20] generates a new  $K$ -regular patch that is a  $C^{1,1}$ -surface and the join is  $C^{1,1}$ . Finally, we generalize everything to hypersurfaces generated by a bi-Lipschitzian mapping defined on a domain with facets (e.g., for sphere, torus). We also give conditions for the decomposition of a  $C^{1,1}$ -hypersurface into  $C^{1,1}$ -patches.

**Mathematics Subject Classification (2000).** 74K25, 74K30, 74K20, 14J70, 26A16, 30F45.

**Keywords.** Intrinsic representation, parametric representation, hypersurface, oriented distance function, tubular neighborhood,  $G_1$ -join, domains with facets, domain decomposition, shells, smoothness of midsurface.

---

This research has been supported by a discovery grant of the National Sciences and Engineering Research Council of Canada.

## 1. Introduction

How to represent hypersurfaces in the Euclidean space  $\mathbf{R}^N$ ,  $N \geq 1$  an integer, and construct a differential calculus is a central topic of Differential Geometry with a broad spectrum of applications to partial differential equations, optimization, and control on hypersurfaces (heat, wave, elasticity, fluids). Among the many fundamental issues is the minimum smoothness of the underlying hypersurface (e.g., we can make sense of asymptotic membrane shell and bending equations in the theory of shells on a  $C^{1,1}$ -midsurface), the  $G_1$ -join of  $C^{1,1}$ -hypersurfaces or patches, and the domain decomposition of  $C^{1,1}$ -hypersurfaces into  $C^{1,1}$ -patches to generate meshings for finite element approximation. From the analysis point of view, it is always preferable to choose an intrinsic representation and avoid local bases and Christoffel symbols. From a numerical point of view, local bases are unavoidable and are influenced by the preferred description of the surface often dictated by *image processing* considerations.

In this paper we revisit and expand the *intrinsic* and *parametric* representations of hypersurfaces. The first approach goes back to E. De Giorgi [5] to solve the Plateau [21] problem of minimal surfaces. The hypersurface  $\omega$  in  $\mathbf{R}^N$  is viewed as the boundary or a subset of the boundary  $\Gamma$  of an open subset  $\Omega$  of  $\mathbf{R}^N$  whose characteristic function is of bounded variation. This was sufficient to make sense of the surface measure. In the context of the theory of shells, Delfour and Zolésio developed a purely intrinsic approach without local maps, local bases, and Christoffel symbols starting in 1992 with [14] and in a number of subsequent papers [15, 16, 17, 6, 7, 8, 10, 12, 2]. The key ingredient was to use the *oriented distance function*  $b_\Omega$  to the underlying set  $\Omega$  instead of the characteristic function. This function completely describes the surface  $\omega$ : its (outward) normal is the gradient  $\nabla b_\Omega$ , its first, second, third, ..., and  $N$ th fundamental forms are  $\nabla b_\Omega \otimes \nabla b_\Omega$ , its Hessian  $D^2 b_\Omega$ ,  $(D^2 b_\Omega)^2, \dots$  and  $(D^2 b_\Omega)^{N-1}$  restricted to the boundary  $\Gamma$  ([13], [18, Chapter 8, § 5]). In addition, a fairly complete intrinsic theory of Sobolev spaces on  $C^{1,1}$ -surfaces is available in [9].

In the theory of thin shells, the asymptotic model, when it exists, only depends on the choice of the *constitutive law*, the *midsurface*, and the space of solutions that properly handles the loading applied to the shell and the boundary conditions. A central issue is how rough this midsurface can be to still make sense of asymptotic *membrane shell* and *bending equations* without ad hoc mechanical or mathematical assumptions. This is possible for a general  $C^{1,1}$ -midsurface with or without boundary such as a sphere, a torus, or a closed reservoir. Moreover, it can be done without local maps, local bases, and Christoffel symbols. A brief review and sharpened results are given in Section 2.

In the parametric approach (cf. for instance [4]), the surface  $\omega$  is defined as the image of a flat smooth bounded connected domain  $U$  in  $\mathbf{R}^2$  via a  $C^2$ -immersion  $\varphi : U \rightarrow \mathbf{R}^3$ . Anicic, Le Dret, and Raoult [1] relaxed the  $C^2$  assumption by introducing in 2004 a family of surfaces  $\omega$  that are the image of a connected bounded open Lipschitzian domain  $U$  in  $\mathbf{R}^2$  by a bi-Lipschitzian mapping  $\varphi$  with the assumption

that the normal field defined almost everywhere is globally Lipschitzian. Such surfaces are called *K-regular patches* by Le Dret [20]. From this, they construct a *tubular neighborhood*  $\mathbb{S}_h(\omega)$  of thickness  $2h$  around the surface and show that for sufficiently small  $h$  the *tubular neighborhood mapping* is bi-Lipschitzian.

In Section 3, we prove that the surfaces of [1] (or *K-regular patches*) are  $C^{1,1}$ -surfaces with a bounded measurable second fundamental form. It was already known that  $C^{1,1}$ -surfaces have a globally Lipschitzian normal field, but it was not, a priori, clear whether midsurfaces generated in the parametrized set-up of [1] would be strictly rougher than  $C^{1,1}$  or not. Moreover, since a *K-regular patch* does not see the singularities of the underlying bi-Lipschitzian parametrization, the  $G_1$ -join of *K-regular patches* along a join developed in [20] generates a new *K-regular patch* that is a  $C^{1,1}$ -surface and the join is in fact  $C^{1,1}$ . We first generalize everything to hypersurfaces in  $\mathbf{R}^N$ ,  $N \geq 2$ , since the proofs are independent of the dimension. Secondly, we show that such tubular neighborhoods can be completely specified by the *algebraic distance* to  $\omega$  and that they are generally not Lipschitzian domains in  $\mathbf{R}^3$  since their tangential smoothness is not effectively controlled by the assumptions of [1] as illustrated by our example in [11]. Therefore,  $C^{1,1}$  is still the currently available minimum smoothness to make sense of asymptotic *membrane shell* and *bending equations*.

In Section 4 we extend the results of Section 3 to hypersurfaces defined on a *connected domain with facets*. This makes it possible to parametrize surfaces such as a sphere or a torus. We show that under the same assumptions as in Section 3 the resulting hypersurface is  $C^{1,1}$  and that the tubular neighborhood mapping theorem still holds. In Section 5, we generalize the work of [20] on  $G_1$ -joins of *K-regular patches* to the  $G_1$ -joins of  $C^{1,1}$ -patches defined on a domain with facets. Finally, in Section 6, we introduce natural assumptions to decompose a  $C^{1,1}$ -hypersurface into  $C^{1,1}$ -patches by constructing a domain with facets. This construction is of special interest in finite element methods for thin shells.

### 1.1. Notation

For an integer  $N \geq 1$  the inner product and the norm in  $\mathbf{R}^N$  will be written  $x \cdot y$  and  $|x|$ . The transpose of a matrix  $A$  will be denoted  $A^*$  and its image  $\text{Im } A$ . The *complement*  $\{x \in \mathbf{R}^N : x \notin \Omega\}$  and the boundary  $\overline{\Omega} \cap \overline{\mathbb{C}\Omega}$  of a subset  $\Omega$  of  $\mathbf{R}^N$  will be respectively denoted by  $\mathbb{C}\Omega$  or  $\mathbf{R}^N \setminus \Omega$  and by  $\partial\Omega$  or  $\Gamma$ . The *distance* and the *oriented distance function* from a point  $x$  to a subset  $\Omega$  of  $\mathbf{R}^N$  are defined as

$$d_\Omega(x) \stackrel{\text{def}}{=} \inf_{y \in \Omega} |y - x|, \quad b_\Omega(x) \stackrel{\text{def}}{=} d_\Omega(x) - d_{\mathbb{C}\Omega}(x). \quad (1.1)$$

In particular  $d_\Omega = |b_\Omega|$ . The *set of projections* of a point  $x$  onto  $\overline{\Omega}$  will be denoted  $\Pi_\Omega(x)$ . When  $\Pi_\Omega(x)$  is a singleton, the projection will be denoted  $p_\Omega(x)$ . The *h-neighborhood* of  $\Omega$  is defined as  $U_h(\Omega) \stackrel{\text{def}}{=} \{x \in \mathbf{R}^N : d_\Omega(x) < h\}$ .

## 2. Intrinsic representation via the oriented distance function

The basic idea goes back to E. De Giorgi [5] in the fifties to solve the celebrated Plateau [21] problem of minimal surfaces. An  $(N - 1)$ -dimensional hypersurface  $\omega$  is viewed as a subset of the boundary  $\Gamma$  of a set  $\Omega$  in the Euclidean space  $\mathbf{R}^N$  for which the gradient of the set parametrized *characteristic function*  $\chi_\Omega$  is a bounded measure. Such sets are called Caccioppoli [3] sets and the norm of the gradient of  $\chi_\Omega$  in the space of bounded measures coincides with the surface measure of  $\Gamma$ .

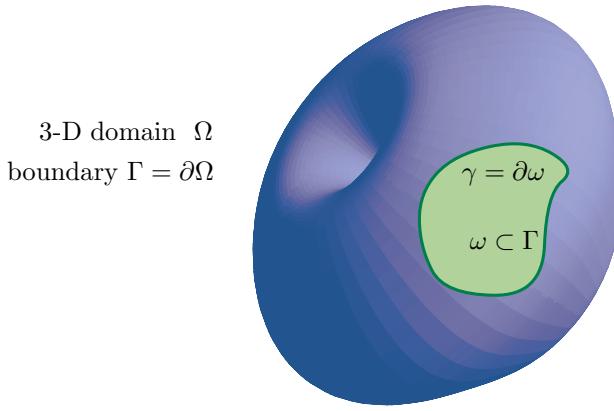


FIGURE 1. Domain  $\omega$  with boundary  $\gamma$

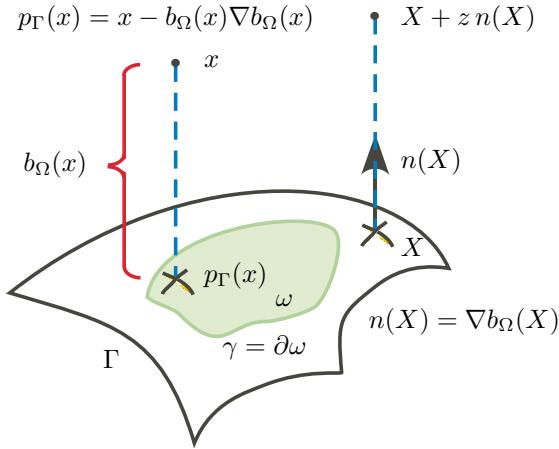
In the context of shells we use as set parametrized function the *oriented distance function*  $b_\Omega$  whose first and second derivatives are directly related to the normal and curvatures. More precisely, for a subset  $\Omega \subset \mathbf{R}^N$  of class  $C^{1,1}$  with a non-empty boundary  $\Gamma \stackrel{\text{def}}{=} \partial\Omega$ ,  $\Gamma$  is a  $C^{1,1}$ -submanifold of codimension one, the normal coincides with  $\nabla b_\Omega$  and the first, second, third, . . . , and  $N$ th fundamental forms are  $\nabla b_\Omega \otimes \nabla b_\Omega$ , the Hessian  $D^2 b_\Omega$ ,  $(D^2 b_\Omega)^2$ , . . . and  $(D^2 b_\Omega)^{N-1}$  restricted to the boundary  $\Gamma$  ([13], [18, Chapter 8, § 5]). In addition, a fairly complete tangential differential calculus and an intrinsic theory of Sobolev spaces on  $C^{1,1}$ -surfaces is available in [9]. We quote the following theorem from [9] that is used to work in curvilinear coordinates in the neighborhood of  $\Gamma$ . It is the intrinsic version of the tubular neighborhood theorem that we shall see in § 3.

**Theorem 2.1.** *Let  $\Omega \subset \mathbf{R}^N$  be a set of class  $C^{1,1}$  such that its boundary  $\Gamma \stackrel{\text{def}}{=} \partial\Omega \neq \emptyset$  be bounded. Then there exists  $h > 0$  such that  $b_\Omega \in C^{1,1}(\overline{U_h(\Gamma)})$ ,*

$$X, z \mapsto T(X, z) \stackrel{\text{def}}{=} X + z\nabla b_\Omega(X) : \Gamma \times ]-h, h[ \rightarrow U_h(\Gamma) \quad (2.1)$$

*is a bi-Lipschitzian bijection, and*

$$T^{-1}(x) = (p_\Gamma(x), b_\Omega(x)). \quad (2.2)$$

FIGURE 2. Bijective bi-Lipschitzian mapping  $T$ 

Assume for the moment that the assumptions of Theorem 2.1 are verified and let  $h > 0$  be such that  $b_\Omega \in C^{1,1}(\overline{U_h(\Gamma)})$ . Given a relatively open subset  $\omega$  of  $\Gamma$ , define the *tubular neighborhood* of thickness  $k$ ,  $0 < k \leq h$ , around  $\omega$

$$S_k(\omega) \stackrel{\text{def}}{=} \{x \in \mathbf{R}^N : |b_\Omega(x)| < k \text{ and } p_\Gamma(x) \in \omega\}. \quad (2.3)$$

By definition,  $S_k(\Gamma) = U_k(\Gamma)$ . But when  $\omega \subsetneq \Gamma$ ,  $U_k(\omega)$  is larger than or equal to  $S_k(\omega)$ .

**Corollary 2.2.** *Let  $\Omega \subset \mathbf{R}^N$  be a set of class  $C^{1,1}$  such that its boundary  $\Gamma \neq \emptyset$  be bounded. Let  $\omega$  be a relatively open subset of  $\Gamma$ . Then there exists  $h > 0$  such that*

$$X, z \mapsto T(X, z) \stackrel{\text{def}}{=} X + z \nabla b_\Omega(X) : \omega \times ]-h, h[ \rightarrow S_h(\omega) \quad (2.4)$$

*is a bi-Lipschitzian bijection and*

$$T^{-1}(x) = (p_\Gamma(x), b_\Omega(x)). \quad (2.5)$$

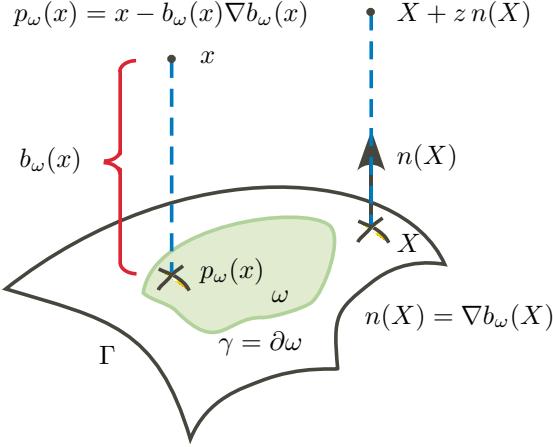
Let  $\gamma$  be the relative boundary of  $\omega$  in  $\Gamma$ . In view of Corollary 2.2, the boundary  $\partial S_k(\omega)$  of  $S_k(\omega)$  is made up of three parts: the *bottom* and *top boundaries*

$$T(\omega, -k) \text{ and } T(\omega, k) \quad (2.6)$$

and the *lateral boundary*

$$\Sigma_k(\gamma) \stackrel{\text{def}}{=} \{x \in \mathbf{R}^N : |b_\Omega(x)| \leq k \text{ and } p_\Gamma(x) \in \gamma\}. \quad (2.7)$$

The top and bottom boundaries  $T(\omega, k)$  and  $T(\omega, -k)$  are  $C^{1,1}$  surfaces with respective normal  $\nabla b_\Omega$  and  $-\nabla b_\Omega$  since the sets  $\{x \in \mathbf{R}^N : b_\Omega(x) < k\}$  and  $\{x \in \mathbf{R}^N : b_\Omega(x) < -k\}$  are still sets of class  $C^{1,1}$ .  $\Sigma_k(\gamma)$  is *normal* to  $\Gamma$ ,  $T(\omega, k)$ , and  $T(\omega, -k)$ .

FIGURE 3. Mapping  $T$  with  $b_\Omega \in C^{1,1}(\mathbb{S}_h(\omega))$ 

The global smoothness assumptions on  $\Gamma$  can be relaxed to a local one in a neighborhood of  $\bar{\omega}$ .

**Theorem 2.3.** *Given  $\Omega \subset \mathbf{R}^N$  with boundary  $\Gamma \neq \emptyset$  and a bounded (relatively) open subset  $\omega$  of  $\Gamma$ , assume that there exists a neighborhood  $N(\omega)$  of  $\bar{\omega}$  such that  $b_\Omega \in C^{1,1}(N(\omega))$ . Then there exists  $\bar{h} > 0$  such that  $b_\Omega \in C^{1,1}(U_{\bar{h}}(\omega))$  and, for all  $h$ ,  $0 < h < \bar{h}$ , the mapping*

$$X, z \mapsto T(X, z) \stackrel{\text{def}}{=} X + z \nabla b_\Omega(X) : \omega \times ]-h, h[ \rightarrow S_h(\omega) \quad (2.8)$$

is a bi-Lipschitzian bijection and its inverse is given by

$$x \mapsto T^{-1}(x) = (p_\Gamma(x), b_\Omega(x)) : S_h(\omega) \rightarrow \omega \times ]-h, h[. \quad (2.9)$$

Under the hypotheses of Theorem 2.3, it is possible to define the *signed distance function* to the hypersurface  $\bar{\omega}$  in the region  $S_{\bar{h}}(\omega)$

$$b_\omega(x) \stackrel{\text{def}}{=} \begin{cases} d_\omega(x), & \text{if } b_\Omega(x) \geq 0 \\ -d_\omega(x), & \text{if } b_\Omega(x) < 0. \end{cases} \quad (2.10)$$

When the projection of a point  $x$  onto  $\bar{\omega}$  is a singleton, we denote it by  $p_\omega(x)$  and necessarily  $p_\omega = p_\Gamma$  on  $S_h(\omega)$ . Note the difference between the oriented distance function  $b_\Omega$  that is always defined everywhere in  $\mathbf{R}^N$  and the signed distance function that is defined only in a region where it is possible to distinguish what is *above* from what is *below*  $\omega$ . Here  $b_\omega = b_\Omega \in C^{1,1}(\overline{S_h(\omega)})$  and  $T$  and  $T^{-1}$  can be rewritten as

$$\begin{aligned} (X, z) \mapsto T(X, z) &= X + z \nabla b_\omega(X) : \omega \times ]-h, h[ \rightarrow S_h(\omega) \\ x \mapsto T^{-1}(x) &= (p_\omega(x), b_\omega(x)) : S_h(\omega) \rightarrow \omega \times ]-h, h[. \end{aligned} \quad (2.11)$$

It is natural to characterize its smoothness in the tangent plane to  $\Gamma$  by specifying the smoothness of  $\Sigma_k(\gamma)$  near  $\gamma$ .

**Definition 2.4** ([9, § 4.5]). Let  $\omega$  be a bounded relatively open subset of  $\Gamma$  which satisfies the assumptions of Theorem 2.1.

- (i) Given an integer  $k \geq 1$  and a real  $0 \leq \lambda \leq 1$ ,  $\gamma$  is  $C^{k,\lambda}$  if there exist  $h > 0$  and  $0 < h' \leq h$  such that the piece  $\Sigma_{h'}(\gamma)$  of the lateral boundary of  $S_h(\omega)$  is  $C^{k,\lambda}$ .
- (ii)  $\gamma$  is Lipschitzian if there exist  $h', 0 < h' \leq h$ , such that  $\Sigma_{h'}(\gamma)$  is Lipschitzian.
- (iii)  $\omega$  is connected if there exists  $h', 0 < h' < h$ , such that  $S_{h'}(\omega)$  is connected.

The definitions correspond to the usual ones in  $\mathbf{R}^N$ . For instance condition (i) is equivalent to say that the oriented distance function  $b_{S_h(\omega)}$  associated with the set  $S_h(\omega)$  has the required smoothness in a neighborhood of  $\Sigma_{h'}(\gamma)$ .

Under the assumptions of Theorem 2.1,  $S_h(\Gamma)$  is  $C^{1,1}$  since  $\Gamma$  has no boundary; for a bounded relatively open subset  $\omega$  of  $\Gamma$  with Lipschitzian relative boundary  $\gamma$ ,  $S_h(\omega)$  is Lipschitzian. In both cases, two versions of Korn's inequality are given in [9, Thms 5.1 and 5.2] and the theory of linear elasticity over a Lipschitzian domain is readily available.

### 3. Parametric hypersurfaces

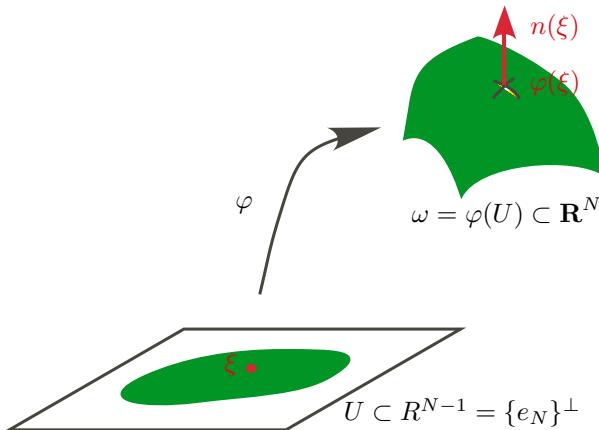


FIGURE 4. Parametric representation in  $\mathbf{R}^N$ .

Let  $\{e_i : 1 \leq i \leq N\}$  be an orthonormal basis in  $\mathbf{R}^N$  and  $A$  an affine subspace of  $\mathbf{R}^N$  of dimension  $N - 1$ . Generalizing [1] from  $N = 3$  to an arbitrary  $N \geq 1$ , let  $U$  be a bounded open domain in an affine subspace  $A$  of  $\mathbf{R}^N$  of dimension  $N - 1$

and let  $\varphi : U \rightarrow \mathbf{R}^N$  be a mapping with the following properties: there exist  $c > 0$  and  $C > 0$  such that

$$\text{Assumption } (H_1): \quad \forall \xi, \zeta \in U, \quad c |\zeta - \xi| \leq |\varphi(\zeta) - \varphi(\xi)| \leq C |\zeta - \xi|, \quad (3.1)$$

where  $c = \inf_{\zeta \neq \xi \in U} |\varphi(\zeta) - \varphi(\xi)| / |\zeta - \xi|$ . By using a translation followed by a rotation, it is always possible to redefine the mapping  $\varphi$  on a subset of the hyperplane  $R^{N-1} = \{e_N\}^\perp$  orthogonal to  $e_N$ .

In view of assumption  $(H_1)$ <sup>1</sup>,  $\omega \stackrel{\text{def}}{=} \varphi(U)$  is a (non self-intersecting) parametric hypersurface in  $\mathbf{R}^N$  of dimension  $N - 1$ . For almost all  $\xi \in U$ ,  $D\varphi(\xi)$ ,  $D\varphi(\xi)_{ij} \stackrel{\text{def}}{=} \partial_j \varphi_i(\xi)$ , exists and

$$\forall V \in R^{N-1}, \quad c |V| \leq |D\varphi(\xi)V| \leq C |V|. \quad (3.2)$$

Therefore,  $D\varphi(\xi) : R^{N-1} \rightarrow \mathbf{R}^N$  is injective and the  $(N - 1)$  column vectors  $\partial_1 \varphi(\xi), \partial_2 \varphi(\xi), \dots, \partial_{N-1} \varphi(\xi)$  are linearly independent in  $\mathbf{R}^N$ . The *surface measure* associated with  $\omega$  is

$$\int_{\omega} dH^{N-1} = \int_U J\varphi d\xi,$$

where  $J\varphi$  is the square root of the sum of the squares of the  $(N - 1) \times (N - 1)$  subdeterminants of  $D\varphi$

$$(J\varphi)^2 = \sum_{i=1}^N \left[ \frac{\partial(\varphi_1, \dots, \varphi_{i-1}, \varphi_{i+1}, \dots, \varphi_N)}{\partial(\xi^1, \dots, \xi^{N-1})} \right]^2.$$

Choose a unit vector  $n(\xi)$  orthogonal to the vectors  $\{\partial_i \varphi(\xi)\}$ ,

$$D\varphi(\xi)^* n(\xi) = 0 \text{ and } |n(\xi)| = 1. \quad (3.3)$$

Then the square matrix  $[D\varphi(\xi) \cdot n(\xi)]$  is invertible and

$$\det[D\varphi(\xi) \cdot n(\xi)] = b(\xi) \cdot n(\xi) \neq 0 \text{ where} \quad (3.4)$$

$$b(\xi)_i = M([D\varphi(\xi) \cdot 0])_{iN}, \quad 1 \leq i \leq N, \quad (3.5)$$

and  $M(A)$  denotes the *matrix of cofactors* of a square matrix  $A$ . In dimension  $N = 3$ ,  $b(\xi)$  coincides with the wedge product  $\partial_1 \varphi(\xi) \wedge \partial_2 \varphi(\xi)$ .

We summarize the main properties using the classical definitions of tangent cone and dual cone. Given  $\Omega \subset \mathbf{R}^N$  and  $x \in \overline{\Omega}$ , denote by  $T_x \Omega$  the *Bouligand's contingent cone* to  $\Omega$  in  $x$ ,

$$T_x \Omega \stackrel{\text{def}}{=} \{v \in \mathbf{R}^N : \exists \{x_n\} \subset \Omega \text{ and } \varepsilon_n \searrow 0 \text{ such that } (x_n - x)/\varepsilon_n \rightarrow v\}, \quad (3.6)$$

and by  $(T_x \Omega)^*$  its *dual cone*  $(T_x \Omega)^* \stackrel{\text{def}}{=} \{y \in \mathbf{R}^N : \forall v \in T_x \Omega, \quad y \cdot v \geq 0\}$ .

---

<sup>1</sup>In contrast with the notation in the theory of shells, the Greek lower case letters  $\omega$  and  $\gamma$  are used for the hypersurface and its boundary. The associated flat reference domain in  $R^{N-1}$  is denoted  $U$ .

**Theorem 3.1** ([11]). *Assume that  $\varphi : U \rightarrow \mathbf{R}^N$  verifies assumption  $(H_1)$  and let  $\tilde{U} = \{\xi \in U : \varphi \text{ is differentiable at } \xi\}$ . Then, at each point  $\xi \in \tilde{U}$ ,*

$$n(\xi) = \pm b(\xi)/|b(\xi)|, \quad \det \left[ D\varphi(\xi) : \frac{b(\xi)}{|b(\xi)|} \right] = |b(\xi)| > 0, \quad J\varphi(\xi) = |b(\xi)|, \quad (3.7)$$

for all  $V \in R^{N-1}$  and  $V^N \in \mathbf{R}$

$$c^2|V|^2 + |V^N|^2 \leq \left| \left[ D\varphi(\xi) : \frac{b(\xi)}{|b(\xi)|} \right] \begin{bmatrix} V \\ V^N \end{bmatrix} \right|^2 \leq C^2|V|^2 + |V^N|^2, \quad (3.8)$$

and

$$T_{\varphi(\xi)}\omega = \text{Im } D\varphi(\xi) = \{n(\xi)\}^\perp \text{ and } (T_{\varphi(\xi)}\omega)^* = \mathbf{R} n(\xi). \quad (3.9)$$

The normal field to  $\omega$  in  $\varphi(\xi)$  is specified by  $n(\xi) = \pm b(\xi)/|b(\xi)|$  in each point of the subset  $\tilde{U}$  of  $U$  where  $\varphi$  is differentiable. We now choose

$$a(\xi) \stackrel{\text{def}}{=} b(\xi)/|b(\xi)|$$

in order to have the determinant of  $[D\varphi : a]$  positive and equal to  $J\varphi$ .

Following [1], it is now assumed that the resulting normal mapping  $a(\xi)$  is uniformly Lipschitz on  $\tilde{U}$ :

$$\text{Assumption } (H_2): \exists \alpha > 0 \text{ such that } \forall \xi, \zeta \in \tilde{U}, |a(\zeta) - a(\xi)| \leq \alpha |\zeta - \xi|. \quad (3.10)$$

Since  $\overline{\tilde{U}} = \overline{U}$ ,  $a$  extends to a unique uniformly Lipschitz function, still denoted  $a$ , on  $\overline{U}$ :  $a$  verifies assumption  $(H_2)$  on  $\overline{U}$ . This very strong assumption “orients” the hypersurface  $\omega$  that no longer “see” the singularities of its bi-Lipschitzian representation. It is a generalization of the classical set-up for  $C^2$ -surfaces used by [4] for shells.

**Theorem 3.2** ([11]). *Assume that  $\varphi$  and  $a$  verify assumptions  $(H_1)$  and  $(H_2)$ . Then*

$$\forall \xi \in U, \quad T_{\varphi(\xi)}\omega = \{a(\xi)\}^\perp \text{ and } (T_{\varphi(\xi)}\omega)^* = \mathbf{R} a(\xi). \quad (3.11)$$

and the parametric hypersurface  $\omega$  has a unique tangent hyperplane in each point with Lipschitzian normal  $a \circ \varphi^{-1}$ .

As in [1], consider for an arbitrary  $k > 0$  the Lipschitz continuous mapping

$$\tilde{\xi} \stackrel{\text{def}}{=} (\xi, \xi^N) \mapsto \Phi(\tilde{\xi}) \stackrel{\text{def}}{=} \varphi(\xi) + \xi^N a(\xi) : \overline{U} \times [-k, k] \rightarrow \mathbf{R}^N \quad (3.12)$$

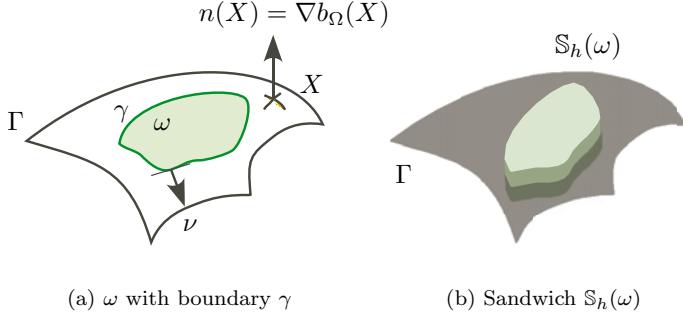
and the associated “sandwich” of thickness  $2h$  around  $\omega$

$$\mathbb{S}_k(\omega) \stackrel{\text{def}}{=} \Phi(U \times ]-k, k[). \quad (3.13)$$

Under assumptions  $(H_1)$  and  $(H_2)$ ,  $\Phi$  is Lipschitzian on  $\overline{U} \times [-k, k]$ ,

$$|\Phi(\xi_2, \xi_2^N) - \Phi(\xi_1, \xi_1^N)| \leq (C + k\alpha)|\xi_2 - \xi_1| + |\xi_2^N - \xi_1^N| \text{ and}$$

$$\overline{\mathbb{S}_k(\omega)} = \Phi(\overline{U} \times [-k, k]).$$

FIGURE 5. Sandwich or tubular neighborhood  $\mathbb{S}_h(\omega)$ 

Associate with  $\Phi$  the intrinsic Lipschitzian mapping

$$(X, z) \mapsto \tilde{\Phi}(X, z) \stackrel{\text{def}}{=} \Phi(\varphi^{-1}(X), z) = X + z a(\varphi^{-1}(X)) : \overline{\omega} \times [-k, k] \rightarrow \mathbf{R}^N. \quad (3.14)$$

Define the *signed distance function* to the hypersurface  $\overline{\omega}$  in the region  $\overline{\mathbb{S}_h(\omega)}$

$$b_\omega(\Phi(\zeta, \zeta^N)) \stackrel{\text{def}}{=} \begin{cases} d_\omega(\Phi(\zeta, \zeta^N)), & \text{if } \zeta^N \geq 0, \\ -d_\omega(\Phi(\zeta, \zeta^N)), & \text{if } \zeta^N < 0. \end{cases} \quad (3.15)$$

When the set projections  $\Pi_\omega(y)$  of  $y$  onto  $\overline{\omega}$  is a singleton, we denote it by  $p_\omega(y)$ .

We now give a constructive proof of Theorem 3.9 in [1], the expression of the inverse of  $\Phi$  in terms of  $p_\omega$  and  $b_\omega$ . To do that we need the following additional assumption on the bounded open subset  $U$  of  $R^{N-1}$ :

$$(H_3) \quad \begin{cases} U \text{ is connected and} \\ \exists C_U \text{ such that } \forall \xi, \zeta \in \overline{U}, \quad d_U(\xi, \zeta) \leq C_U |\xi - \zeta|, \end{cases} \quad (3.16)$$

where  $d_U$  denotes the geodesic distance in  $U$ . This is a weaker assumption than the one used in [1]

$$(H_{3L}) \quad U \text{ is connected and Lipschitzian} \quad (3.17)$$

where they make use of the following lemma:

**Lemma 3.3.** [1, Proposition A.1] *Assume that  $U$  is a bounded, open, connected, and Lipschitzian domain in  $R^{N-1}$ . Then  $U$  verifies assumption  $(H_3)$ .*

We shall see later on that this relaxation will allow us to go from a flat domain  $U$  to a domain with facets and enable us to parametrize all  $C^{1,1}$ -hypersurfaces such as the sphere and the torus in dimension  $N = 3$ .

Recalling that  $\overline{\mathbb{S}_h(\omega)} = \Phi(\overline{U} \times [-h, h])$ , we get the following results.

**Theorem 3.4** ([11]). *Assume that  $(H_1)$ ,  $(H_2)$ , and  $(H_3)$  are verified and let  $h$ ,  $0 < h < \bar{h} \stackrel{\text{def}}{=} c^2/(2C_U^2 C\alpha)$ . The mapping  $\Phi : \overline{U} \times [-h, h] \rightarrow \overline{\mathbb{S}_h(\omega)}$  is bijective and bi-Lipschitzian, and*

$$y \mapsto \Phi^{-1}(y) = (\varphi^{-1}(p_\omega(y)), b_\omega(y)) : \overline{\mathbb{S}_h(\omega)} \rightarrow \overline{U} \times [-h, h]. \quad (3.18)$$

*Remark 3.5.* Assumption  $(H_2)$  on the normal field  $a$  effectively controls the smoothness of the hypersurface  $\omega$  in the normal direction. There is a unique tangent plane and a unique one-dimensional normal field at *every* point *without* the additional assumption  $(H_3)$ . In other words,  $\omega$  ignores the singularities of the mapping  $\varphi$  in the normal direction. Yet, in the tangential direction, the choice of a bi-Lipschitzian parametrization  $\varphi$  of  $\omega$  and the assumptions  $(H_{3L})$  that  $U$  be Lipschitzian and connected are not sufficient to make the *lateral boundary* of  $\mathbb{S}_h(\omega)$  a Lipschitzian hypersurface even in the case of a plate in  $\mathbf{R}^3$  as illustrated for the plate of thickness  $2h$  of the example given in [11]. Results from the *theory of linear elasticity* assuming a Lipschitzian elastic body cannot be directly used. This contradicts the following statement from [1, Remark on page 1290]: “*Note that this result shows that the boundary of the three-dimensional shell is Lipschitz, hence the three-dimensional linearized elasticity problem is well posed*”.

We now complete the characterization of the hypersurface  $\omega$  and make the connection between  $\Phi$  and  $\tilde{\Phi}$  and the intrinsic mapping  $T$  defined by (2.8) in Theorem 2.3 and between the sets  $\mathbb{S}_h(\omega)$  and  $\text{Im } T$ . We also show that the parametric hypersurface is in fact not rougher than  $C^{1,1}$  for which the previous intrinsic theory already applies. No gain is achieved through this parametrization (cf. Theorem 2.3 with the mapping (2.8) replaced by the mapping (2.11)). As a result the constructions and results of the intrinsic theory of thin and asymptotic shells described in § 7 readily applies with  $\Omega$  replaced by  $\omega$ . It is not necessary to do it again even in the parametric case.

**Theorem 3.6** ([11]). *Assume that assumptions  $(H_1)$ ,  $(H_2)$ , and  $(H_3)$  are verified and let  $h$ ,  $0 < h < \bar{h}$ . Denote by  $\Omega$  the open domain  $\Phi(U \times ]-h, h[)$  and by  $\Gamma$  its boundary. The hypersurface  $\omega$  is locally  $C^{1,1}$ , that is, for each  $x \in \omega$ , there exists  $r(x) > 0$  such that  $b_\Omega \in C^{1,1}(B_{r(x)}(x))$  and, hence,  $\omega \cap B_{r(x)}(x)$  is of class  $C^{1,1}$  in  $B_{r(x)}(x)$ . Moreover, its normal and second fundamental form are given by*

$$\nabla b_\Omega|_\omega = a \circ \varphi^{-1} = \nabla b_\omega|_\omega \in C^{0,1}(\overline{\omega})^N \quad (3.19)$$

$$D^2 b_\Omega|_\omega = \{Da[(D\varphi)^* D\varphi]^{-1} (D\varphi)^*\} \circ \varphi^{-1} = D^2 b_\omega|_\omega \in L^\infty(\omega)^{N \times N}. \quad (3.20)$$

Moreover,  $\tilde{\Phi} = T$  on  $\omega \times ]-h, h[$  and  $\text{Im } T = \text{Im } \tilde{\Phi} = \text{Im } \Phi = \mathbb{S}_h(\omega)$ .

## 4. Hypersurfaces defined from a domain with facets

It is not possible to represent a sphere or a torus by a bi-Lipschitzian mapping  $\varphi$  from some domain  $U$  in the hyperplane  $R^{N-1}$ . Yet, we can do so under assumptions  $(H_1)$  to  $(H_3)$  by replacing the “flat” domain  $U$  by a domain with facets. Each

facet will lie in an  $(N - 1)$ -dimensional affine subspace  $A$  of  $\mathbf{R}^N$  allowing different angles between facets. The good news is that all the results in § 3 remain true. We summarize the definitions and main results from [11].



FIGURE 6. Example of a domain  $U$  with facets

**Definition 4.1.** (i) A *facet*  $U$  in  $\mathbf{R}^N$  is a bounded, open, connected subset of an  $(N - 1)$ -dimensional affine subspace of  $\mathbf{R}^N$  such that its geodesic distance satisfies the condition

$$\exists C_U, \forall \zeta, \xi \in U, \quad d_U(\zeta, \xi) \leq C_U |\zeta - \xi|,$$

where  $d_U$  denotes the geodesic distance in  $\overline{U}$ .

(ii) Given  $n$  facets  $U_i$ ,  $1 \leq i \leq n$ , in  $\mathbf{R}^N$  such that

a)  $\forall i \neq j$ ,  $U_i \cap U_j = \emptyset$ ,

b) for all pairs  $i \neq j$  such that  $\overline{U_i} \cap \overline{U_j} \neq \emptyset$ ,  $H_{N-1}(\overline{U_i} \cap \overline{U_j}) = 0$ , where  $H_{N-1}$  is the  $(N - 1)$ -dimensional Hausdorff measure in  $\mathbf{R}^N$ ,

c)  $\overline{\cup_{i=1}^n U_i} = \cup_{i=1}^n \overline{U_i}$ ,

we say that the set

$$U \stackrel{\text{def}}{=} \text{rel int } \cup_{i=1}^n \overline{U_i} \quad (4.1)$$

is a *domain with  $n$  facets*.

From the above definition

$$\overline{\cup_{i=1}^n U_i} = \overline{U} = \cup_{i=1}^n \overline{U_i} \text{ and } H_{N-1} \left( \bigcup_{\substack{i,j=1,\dots,n \\ i \neq j}} \overline{U_i} \cap \overline{U_j} \right) = 0. \quad (4.2)$$

The definition is a little technical since it includes domains with holes of various shapes: all the facets are not necessarily of the polygonal type.

Let  $U$  be a bounded domain in  $\mathbf{R}^N$  with facets and  $\varphi : U \rightarrow \mathbf{R}^N$  be a mapping with the following properties: there exist  $c > 0$  and  $C > 0$  such that

$$\text{Assumption (H}_1\text{): } \forall \xi, \zeta \in U, \quad c |\zeta - \xi| \leq |\varphi(\zeta) - \varphi(\xi)| \leq C |\zeta - \xi|, \quad (4.3)$$

where  $c = \inf_{\zeta \neq \xi \in U} |\varphi(\zeta) - \varphi(\xi)|/|\zeta - \xi|$ . Let  $\varphi_i$  be the restriction of  $\varphi$  to  $\overline{U}_i$ . For almost all  $\xi \in U_i$ , we can construct a normal  $a$  and a surface density in  $U$  and we get a generalization of Theorem 3.1.

**Theorem 4.2.** *Let  $U$  be a bounded domain in  $\mathbf{R}^N$  with facets. Assume that  $\varphi : U \rightarrow \mathbf{R}^N$  verifies assumption  $(H_1)$  and let  $\tilde{U} = \{\xi \in \cup_{i=1}^n U_i : \varphi \text{ is differentiable at } \xi\}$ . Then, at each point  $\xi \in \tilde{U}$ ,*

$$n(\xi) = \pm b(\xi)/|b(\xi)|, \quad \det \left[ D\varphi(\xi) : \frac{b(\xi)}{|b(\xi)|} \right] = |b(\xi)| > 0, \quad J\varphi(\xi) = |b(\xi)|, \quad (4.4)$$

for all  $i$ ,  $V \in \{d_i\}^\perp$ , and  $V^N \in \mathbf{R}$

$$c^2|V|^2 + |V^N|^2 \leq \left| \left[ D\varphi(\xi) : \frac{b(\xi)}{|b(\xi)|} \right] \begin{bmatrix} V \\ V^N \end{bmatrix} \right|^2 \leq C^2|V|^2 + |V^N|^2, \quad (4.5)$$

and

$$T_{\varphi(\xi)}\omega = \text{Im } D\varphi(\xi) = \{n(\xi)\}^\perp \text{ and } (T_{\varphi(\xi)}\omega)^* = \mathbf{R} n(\xi). \quad (4.6)$$

Now assume that the resulting normal mapping  $a(\xi) = b(\xi)/|b(\xi)|$  is uniformly Lipschitz on  $\tilde{U}$ :

$$\text{Assumption } (H_2): \exists \alpha > 0 \text{ such that } \forall \xi, \zeta \in \tilde{U}, |a(\zeta) - a(\xi)| \leq \alpha |\zeta - \xi|. \quad (4.7)$$

From assumptions b) and c),  $\overline{\tilde{U}} = \overline{U}$  and  $a$  extends to a (unique) uniformly Lipschitz function, still denoted  $a$ , on  $\overline{U}$  that verifies assumption  $(H_2)$  on  $\overline{U}$ .

**Theorem 4.3.** *Let  $U$  be a bounded domain in  $\mathbf{R}^N$  with facets. Assume that  $\varphi$  and  $a$  verify assumptions  $(H_1)$  and  $(H_2)$ . Then*

$$\forall \xi \in U, \quad T_{\varphi(\xi)}\omega = \{a(\xi)\}^\perp \text{ and } (T_{\varphi(\xi)}\omega)^* = \mathbf{R} a(\xi). \quad (4.8)$$

and the parametric surface  $\omega$  has a unique tangent hyperplane in each point with Lipschitzian normal  $a \circ \varphi^{-1}$ .

Now define the Lipschitzian mapping  $\Phi$ , the intrinsic Lipschitzian mapping (3.14), and the *signed distance function* to the hypersurface  $\overline{\omega}$  in the region  $\overline{\mathbb{S}_h(\omega)}$  as in § 3. Finally, introduce the following assumption on the underlying domain  $U$  with facets:

$$(H_3) \quad \begin{cases} U \text{ is connected and} \\ \exists C_U \text{ such that } \forall \xi, \zeta \in \overline{U}, \quad d_U(\xi, \zeta) \leq C_U |\xi - \zeta|, \end{cases} \quad (4.9)$$

where  $d_U$  denotes the geodesic distance in  $U$ . We get the generalization of Theorems 3.4 and 3.6 using essentially the same proofs.

**Theorem 4.4.** *Let  $U$  be a bounded domain in  $\mathbf{R}^N$  with facets. Assume that  $(H_1)$ ,  $(H_2)$ , and  $(H_3)$  are verified and let  $h$ ,  $0 < h < \bar{h} \stackrel{\text{def}}{=} c^2/(2C_U^2 C \alpha)$ . The mapping  $\Phi : \overline{U} \times [-h, h] \rightarrow \overline{\mathbb{S}_h(\omega)}$  is bijective and bi-Lipschitzian, and*

$$y \mapsto \Phi^{-1}(y) = (\varphi^{-1}(p_\omega(y)), b_\omega(y)) : \overline{\mathbb{S}_h(\omega)} \rightarrow \overline{U} \times [-h, h]. \quad (4.10)$$

**Theorem 4.5.** Let  $U$  be a bounded domain in  $\mathbf{R}^N$  with facets. Assume that assumptions  $(H_1)$ ,  $(H_2)$ , and  $(H_3)$  are verified and let  $h$ ,  $0 < h < \bar{h}$ . Denote by  $\Omega$  the open domain  $\Phi(U \times ]-h, 0[)$  and by  $\Gamma$  its boundary. The hypersurface  $\omega$  is locally  $C^{1,1}$ , that is, for each  $x \in \omega$ , there exists  $r(x) > 0$  such that  $b_\Omega \in C^{1,1}(B_{r(x)}(x))$  and, hence,  $\omega \cap B_{r(x)}(x)$  is of class  $C^{1,1}$  in  $B_{r(x)}(x)$ . Moreover, its normal and second fundamental form are given by

$$\nabla b_\Omega|_\omega = a \circ \varphi^{-1} = \nabla b_\omega|_\omega \in C^{0,1}(\overline{\omega})^N \quad (4.11)$$

$$D^2 b_\Omega|_\omega = \{Da[(D\varphi)^* D\varphi]^{-1} (D\varphi)^*\} \circ \varphi^{-1} = D^2 b_\omega|_\omega \in L^\infty(\omega)^{N \times N}. \quad (4.12)$$

Moreover,  $\tilde{\Phi} = T$  on  $\omega \times ]-h, h[$  and  $\text{Im } T = \text{Im } \tilde{\Phi} = \text{Im } \Phi = \mathbb{S}_h(\omega)$ .

## 5. $G_1$ -joins of $K$ -regular and $C^{1,1}$ -patches

For completeness we first recall and introduce some definitions.

**Definition 5.1.** Given an open subset  $U$  of  $R^{N-1}$  and a mapping  $\varphi : \overline{U} \rightarrow \mathbf{R}^N$ , we say that the set  $\varphi(\overline{U})$  is *not self-intersecting* if  $\varphi$  is injective.

Following the terminology of [20, Definition 2.4] in dimension three, a  *$K$ -regular patch*  $\overline{\omega}$  is an hypersurface specified by the two mappings  $\varphi$  and  $a$  from  $U \subset R^{N-1} \rightarrow \mathbf{R}^N$  that verify assumptions  $(H_1)$ ,  $(H_2)$ , and  $(H_{3L})$  with constants  $c$ ,  $C$ , and  $\alpha$ . From assumption  $(H_1)$ , such surfaces are not self-intersecting and, from assumptions  $(H_1)$ ,  $(H_2)$ , and  $(H_{3L})$ , they are  $C^{1,1}$ -hypersurfaces by Theorem 3.6. So, we suggest to use the more descriptive and general terminology  *$C^{1,1}$ -patch* that emphasizes the purely geometric property which is not only specific of the theory of shells.

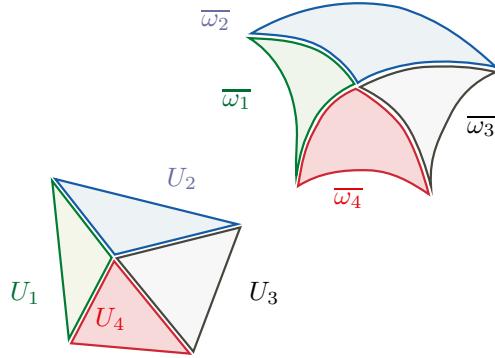
**Definition 5.2.** A  *$C^{1,1}$ -patch* is a parametric hypersurface specified by the two mappings  $\varphi$  and  $a$  from  $U \subset R^{N-1} \rightarrow \mathbf{R}^N$  that verify assumptions  $(H_1)$  to  $(H_3)$  with constants  $c$ ,  $C$ , and  $\alpha$ .

By definition, a  $K$ -regular patch is a  $C^{1,1}$ -patch since assumption  $(H_{3L})$  implies assumption  $(H_3)$ .

One important contribution in the paper of Le Dret is an accurate definition of a  $G_1$ -join [20, Definition 2.6] and the proof that for two contiguous  $K$ -regular patches  $\overline{\omega}_1 = \varphi_1(\overline{U}_1)$  and  $\overline{\omega}_2 = \varphi_2(\overline{U}_2)$  such that  $\overline{\omega}_1 \cup \overline{\omega}_2$  is not self-intersecting with a  $G_1$ -join along  $\overline{U}_1 \cap \overline{U}_2$ , rel int  $\overline{\omega}_1 \cup \overline{\omega}_2$  satisfies assumptions  $(H_1)$ ,  $(H_2)$ , and  $(H_{3L})$  [20, Lemma 3.2], that is it is a  $C^{1,1}$ -hypersurface and the  $G_1$ -join along  $\overline{U}_1 \cap \overline{U}_2$  is in fact a  $C^{1,1}$ -join.

To complete this section we extend this last result to a finite number of  $C^{1,1}$ -patches in  $\mathbf{R}^N$  defined on a domain with facets.

**Theorem 5.3.** Let  $U$  be a bounded connected domain in  $\mathbf{R}^N$  with facets. Assume that the facets  $\omega_i = \varphi_i(U_i)$  specified by  $(U_i, \varphi_i, a_i)$  are  $C^{1,1}$ -patches such that for all  $i \neq j$  such that  $\overline{\omega}_i \cap \overline{\omega}_j \neq \emptyset$

FIGURE 7. From  $C^{1,1}$ -patches  $\{\omega_i\}$  to a globally  $C^{1,1}$ -patch  $\omega$ 

- a) there exists  $C_{ij}$  such that for all  $\xi$  and  $\zeta$  in  $\overline{U}_i \cup \overline{U}_j$ ,  $d_{\overline{U}_i \cup \overline{U}_j}(\xi, \zeta) \leq C_{ij} |\xi - \zeta|$ ,
- b)  $\varphi_i(\xi) = \varphi_j(\xi)$ , for all  $\xi \in \overline{\omega}_i \cap \overline{\omega}_j$ ,
- c)  $a_i(\xi) = a_j(\xi)$ , for all  $\xi \in \overline{\omega}_i \cap \overline{\omega}_j$ ,
- d)  $\overline{\omega}_i \cap \overline{\omega}_j \subset \varphi(\overline{U}_i \cap \overline{U}_j)$ ,
- e) given any sequences  $\{\zeta_{in}\} \subset U_i$  and  $\{\zeta_{jn}\} \subset U_j$  that converge to some point  $\xi \in \overline{U}_i \cap \overline{U}_j$  and a corresponding sequence  $\{\xi_n\} \subset \overline{U}_i \cap \overline{U}_j$  such that  $\xi_n$  lie on the geodesic between  $\zeta_{ni}$  and  $\zeta_{nj}$ , the angle between any limit vectors  $\tau_i$  and  $\tau_j$  of the sequences  $(\varphi_i(\zeta_{in}) - \varphi_i(\xi_n)) / |\zeta_{in} - \xi_n|$  and  $(\varphi_j(\zeta_{jn}) - \varphi_j(\xi_n)) / |\zeta_{jn} - \xi_n|$  is nonzero.

Then

- (i)  $U$  satisfies assumption (H<sub>3</sub>),
- (ii) the maps  $\varphi$  and  $a : \overline{U} \rightarrow \mathbf{R}^N$ ,

$$\varphi(\zeta) = \varphi_i(\zeta), \quad \text{if } \zeta \in \overline{U}_i, \quad a(\zeta) = a_i(\zeta), \quad \text{if } \zeta \in \overline{U}_i,$$

are well defined and Lipschitz continuous on  $\overline{U}$ ,

- (iii)  $\overline{\omega}$  is not self-intersecting ( $\varphi$  is injective),
- (iv)  $\varphi$  satisfies assumption (H<sub>1</sub>) and  $a$  satisfies assumption (H<sub>2</sub>).

In particular,  $\omega = \varphi(U)$  satisfies assumptions (H<sub>1</sub>), (H<sub>2</sub>), and (H<sub>3</sub>), that is  $\omega$  is a  $C^{1,1}$ -patch.

*Remark 5.4.* It is readily checked that [20, Lemma 3.2] is a special case of this Theorem for  $k = 1$ .

## 6. Decomposition of a $C^{1,1}$ -hypersurface into $C^{1,1}$ -patches over a domain with facets

Of course, it is also possible to decompose a  $C^{1,1}$ -hypersurface  $\omega$  into  $C^{1,1}$ -patches defined over a domain with facets when the size of each facet is sufficiently small.

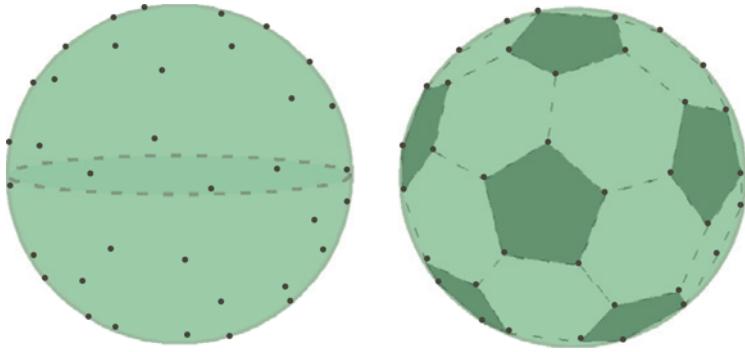


FIGURE 8. Points on the sphere and associated domain with facets for  $N = 3$

The construction follows the following scheme.

(A-1) Assumptions of Theorem 2.3 or assumptions  $(H_1)$  to  $(H_3)$ .

Under assumption (A-1), there exists  $h > 0$  such that the mapping

$$X, z \mapsto T(X, z) = X + z \nabla b_\omega(X) : \overline{\omega} \times [-h, h] \rightarrow \overline{\mathbb{S}_h(\omega)}$$

is bi-Lipschitzian.

If  $\omega$  has no boundary, then we proceed as in [19] but the *convex polytopes* (triangles in dimension  $N = 3$ ) have to be chosen sufficiently small in view of the curvature of the surface.

(A-2) Choose  $N$  neighboring points  $\xi_1, \xi_2, \dots, \xi_N$  on the surface  $\omega$  such that the vectors  $\{\xi_i - \xi_N; i = 1, \dots, N-1\}$  be linearly independent and such that the convex polytope  $\Delta = \text{co}\{\xi_1, \xi_2, \dots, \xi_N\}$  with vertices  $\xi_1, \xi_2, \dots, \xi_N$  lies in  $\mathbb{S}_h(\omega)$ . Denote by  $\nu$  the normal to the affine subspace  $A = \{\xi : (\xi - \xi_N) \cdot \nu = 0\}$  generated by  $\Delta$ .

This defines the patch  $\overline{\omega}_\Delta \stackrel{\text{def}}{=} p_\omega(\Delta)$ . Since  $\Delta \subset \mathbb{S}_h(\omega)$  the mapping

$$\xi \mapsto p_\omega(\xi) : \Delta \rightarrow \overline{\omega}_\Delta \subset \mathbf{R}^3$$

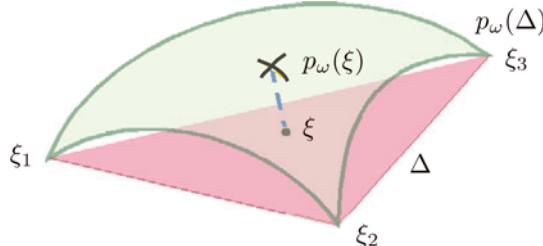


FIGURE 9. Triangle  $\Delta$  and  $C^{1,1}$ -patch  $p_\omega(\Delta)$  for  $N = 3$

is Lipschitzian. To make  $p_\omega$  bi-Lipschitzian on  $\Delta$ , we need to further reduce the size of  $\Delta$ .

(A-3) Assume that the convex polytope  $\Delta$  is sufficiently small so that

$$m \stackrel{\text{def}}{=} \min_{\xi \in \Delta} |n \cdot \nabla b_\omega(\xi)| > 0. \quad (6.1)$$

Under assumption (A-2), for each  $\xi \in \Delta$ ,  $\xi = p_\omega(\xi) + b_\omega(\xi) \nabla b_\omega(p_\omega(\xi))$  and under assumption (A-3), for each  $\xi \in \Delta$ ,

$$b_\omega(\xi) = -\frac{n \cdot (p_\omega(\xi) - \xi_3)}{n \cdot \nabla b_\omega(p_\omega(\xi))}$$

and there exists a constant  $c'$  such that

$$\forall \zeta, \xi \in \Delta, \quad |\xi - \zeta| \leq c' |p_\omega(\xi) - p_\omega(\zeta)|$$

and  $p_\omega : \Delta \rightarrow p_\omega(\Delta)$  is bi-Lipschitzian.

As a result, for a  $C^{1,1}$ -surface without boundary,  $p_\omega(\Delta)$  is a  $C^{1,1}$ -patch since assumption  $(H_1)$  is verified with  $U = \Delta$  and  $\varphi = p_\Omega|_\Delta$ , assumption  $(H_2)$  is verified with  $a = \nabla b_\omega \circ p_\omega$ , and assumption  $(H_3)$  is verified for the triangle  $\Delta$ . We summarize the above discussion in the following theorem.

**Theorem 6.1.** *Let  $\Omega$  be a bounded set of class  $C^{1,1}$  in  $\mathbf{R}^N$  and set  $\omega = \Gamma$ . Around each point  $\hat{\xi} \in \omega$  there exist  $(N-1)$  points  $\{\xi_i \in \omega : 1 \leq i \leq N-1\}$  such that  $\{\xi_i - \hat{\xi} : 1 \leq i \leq N-1\}$  be linearly independent and assumptions (A-1) to (A-3) be verified for the convex polytope*

$$\Delta \stackrel{\text{def}}{=} \text{co}\{\xi_1, \dots, \xi_{N-1}, \hat{\xi}\}$$

with unit normal  $\nu$ . The mapping

$$\xi \mapsto p_\omega(\xi) : \Delta \rightarrow \overline{\omega}_\Delta \subset \mathbf{R}^3$$

is bi-Lipschitzian and

$$p_\omega^{-1}(X) = X - \frac{\nu \cdot (X - \hat{\xi})}{\nu \cdot \nabla b_\omega(X)} \nabla b_\omega(X).$$

When  $\omega$  has a relative boundary  $\gamma$ , then it is necessary to use the assumptions of Theorem 2.3 in order to cover the boundary  $\gamma$  with triangles since some of the vertices may lie outside of  $\omega$ . In that case the patch  $p_\omega(\Delta)$  and the triangle  $\Delta$  should be replaced by the smaller patch  $p_\omega(\Delta) \cap \omega$  and the smaller domain  $U_\Delta \stackrel{\text{def}}{=} p_\omega^{-1}(p_\omega(\Delta) \cap \omega)$  in  $\Delta$  on which assumption  $(H_3)$  must now be imposed in the absence of a specific assumption on the boundary  $\gamma$ .

## 7. Intrinsic theory of thin and asymptotic shells

In order to complete the references in [1] on the theory of shells and to provide a broader perspective to the reader, we briefly recall a few results starting with the key paper [6] on the use of intrinsic methods in the asymptotic analysis of three models of thin shells for an arbitrary linear 3D constitutive law. They all converge to asymptotic shell models that consist of a coupled system of two variational equations. They only differ in their resulting effective constitutive laws. The first equation yields the generally accepted classical *membrane shell equation* and the Love-Kirchhoff terms. The second is a generalized *bending equation*. It explains that convergence results for the 3D models were only established for plates and in the bending dominated case for shells. From the analysis of the three models, the richer  $P(2, 1)$ -model turns out to be the most pertinent since it converges to the right asymptotic model with the right effective constitutive law. We also show in [7] that models of the Naghdi's type can be obtained directly from the  $P(2, 1)$ -model by a simple elimination of variables without introducing the a priori assumption on the stress tensor  $\sigma_{33} = 0$ . Bridges are thrown with classical models using local bases or representations. Those results are completed in [7] with the characterization of the space of solution for the  $P(2, 1)$  thin shell model and the space of solutions of the asymptotic membrane shell equation in [8]. This characterization was only known in the case of the plate and uniformly elliptic shells.

In [10], a new choice of the projection leads to the disappearance of the coupling term in the second asymptotic equation. After reduction of the number of variables, this new choice changes the form of the second equation to achieve the complete decoupling of the membrane and bending equations without the classical plate or bending dominated assumptions. In the second part of [10] we present a dynamical thin shell model for small vibrations and investigate the corresponding dynamical asymptotic model. Those papers complete [6] and make the connection with most existing results in the literature thus confirming the pertinence and the interest of the methods we have developed. Extensions of the  $P(2, 1)$ -model have also been developed for piezoelectric shells [12, 2] where a complete decoupling of the membrane and bending equations is also obtained.

## 8. Conclusions

To conclude we summarize the main points of the paper.

- 1) It is sufficient to replace  $\Omega$  by  $\omega$  and use the already available theory of thin and asymptotic shells in both parametric and intrinsic cases.
- 2) In general the tubular neighborhood is not a Lipschitzian domain and an assumption has to be made on the lateral boundary to use the available Linear Elasticity Theory in Lipschitzian domains.
- 3) The (minimal) smoothness of the midsurface obtained in [1] is in fact  $C^{1,1}$  and the  $G_1$ -joins of  $K$ -regular contiguous patches in [20] are indeed  $C^{1,1}$ .

- 4) By relaxing the Lipschitzian assumption ( $H_{3L}$ ) to the natural condition ( $H_3$ ) on the geodesic distance, it is possible to extend the parametric set up to capture surfaces such as the sphere and the torus by using domains with facets.
- 5) Similarly, the  $G_1$ -join of  $C^{1,1}$ -patches generated from a domain with facets yields a  $C^{1,1}$ -hypersurface. Conversely, we gave a procedure to decompose a  $C^{1,1}$ -hypersurface into  $C^{1,1}$ -patches with potential application to the meshing of surfaces to solve partial differential equations defined on a surface,

## References

- [1] S. Anicic, H. Le Dret, and A. Raoult, *The infinitesimal rigid displacement lemma in Lipschitz co-ordinates and application to shells with minimal regularity*, Math. Meth. Appl. Sci. 27 (2004), 1283–1299.
- [2] M. Bernadou and M.C. Delfour, *Intrinsic models of piezoelectric shells*, in “Proceedings of ECCOMAS 2000” (European Congress on Computational Methods in Applied Sciences and Engineering, Barcelona, Spain, Sept. 11–14, 2000).
- [3] R. Caccioppoli, *Misura e integrazione sugli insiemi dimensionalmente orientati*, Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Nat. **VIII** (1952), no. 12, 3–11; 137–146.
- [4] P.D. Ciarlet, *Mathematical elasticity, vol. III: theory of shells*, North-Holland, Elsevier Science, Amsterdam 2000.
- [5] E. De Giorgi, *Su una teoria generale della misura*, Ann. Mat. Pura Appl. **36** (1954), no. 4, 191–213.
- [6] M.C. Delfour, *Intrinsic differential geometric methods in the asymptotic analysis of linear thin shells*, Boundaries, interfaces and transitions (Banff, AB, 1995), (M.C. Delfour, ed.), pp. 19–90, CRM Proc. Lecture Notes, 13, Amer. Math. Soc., Providence, RI, 1998.
- [7] M.C. Delfour, *Intrinsic  $P(2,1)$  thin shell model and Naghdi’s models without a priori assumption on the stress tensor*, in “Optimal control of partial differential equations” (Chemnitz, 1998), K.H. Hoffmann, G. Leugering, F. Tröltzsch, eds., pp. 99–113, Internat. Ser. Numer. Math., 133, Birkhäuser, Basel, 1999.
- [8] M.C. Delfour, *Characterization of the space of the membrane shell equation for arbitrary  $C^{1,1}$  midsurfaces*, Control and Cybernetics **28** (1999), no. 3, 481–501.
- [9] M.C. Delfour, *Tangential differential calculus and functional analysis on a  $C^{1,1}$  submanifold*, in “Differential-geometric methods in the control of partial differential equations” (Boulder, CO, 1999), R. Gulliver, W. Littman and R. Triggiani, eds., pp. 83–115, Contemp. Math., 268, Amer. Math. Soc., Providence, RI, 2000.
- [10] M.C. Delfour, *Modeling and control of asymptotic shells*, in “Control and Estimation of Distributed Parameter Systems”, W. Desch, F. Kappel, and K. Kunish, eds., pp. 105–120, Int. Series of Numerical Mathematics, Vol 143, Birkhäuser Verlag 2002.
- [11] M.C. Delfour, *Representations of hypersurfaces and minimal smoothness of the mid-surface*, CRM Report 3259, Université de Montréal July 2008.
- [12] M.C. Delfour and M. Bernadou, *Intrinsic asymptotic model of piezoelectric shells*, in “Optimal Control of Complex Structures” (Oberwolfach, 2000), K.-H. Hoffmann,

- I. Lasiecka, G. Leugering, J. Sprekels, F. Tröltzsch (Eds.), pp. 59–72, Internat. Ser. Numer. Math., 139, Birkhäuser Verlag, Basel, 2002.
- [13] M.C. Delfour and J.-P. Zolésio, *Shape analysis via oriented distance functions*, J. Funct. Anal. **123** (1994), no. 1, 129–201.
  - [14] M.C. Delfour and J.-P. Zolésio, *On a variational equation for thin shells*, Control and optimal design of distributed parameter systems (Minneapolis, MN, 1992), (J. Lagnese, D.L. Russell, and L. White, eds.), pp. 25–37, IMA Vol. Math. Appl., 70, Springer, New York, 1995.
  - [15] M.C. Delfour and J.-P. Zolésio, *A boundary differential equation for thin shells*, J. Differential Equations **119** (1995), 426–449.
  - [16] M.C. Delfour and J.-P. Zolésio, *Tangential differential equations for dynamical thin/shallow shells*, J. Differential Equations **128** (1996), 125–167.
  - [17] M.C. Delfour and J.-P. Zolésio, *Differential equations for linear shells: comparison between intrinsic and classical models*, in “Advances in mathematical sciences: CRM’s 25 years” (Montreal, PQ, 1994), (Luc Vinet, ed.), pp. 41–124, CRM Proc. Lecture Notes, 11, Amer. Math. Soc., Providence, RI, 1997.
  - [18] M.C. Delfour and J.-P. Zolésio, *Shapes and geometries. Analysis, differential calculus, and optimization*, Advances in Design and Control, 4. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2001.
  - [19] P.J. Frey and P.L. George, *Mesh Generation: Application to Finite Element*, Second Edition, ISTE Ltd, London and John Wiley and Sons, Inc, Hoboken (NJ), 2008.
  - [20] H. Le Dret, *Well-posedness for Koiter and Naghdi shells with a  $G_1$ -midsurface*, Analysis and Applications **2**, No. 4 (2004), 365–388.
  - [21] J.A.F. Plateau, *Statique expérimentale et théorique des liquides soumis aux seules forces moléculaires*, Gauthier-Villars, Paris, 1873.

Michel C. Delfour  
Centre de recherches mathématiques  
et Département de mathématiques et de statistique  
Université de Montréal  
C.P. 6128, succ. Centre-ville  
Montréal (Qc), Canada H3C 3J7  
e-mail: [delfour@crm.umontreal.ca](mailto:delfour@crm.umontreal.ca)

# On Some Nonlinear Optimal Control Problems with Vector-valued Affine Control Constraints

Juan Carlos De Los Reyes and Karl Kunisch

**Abstract.** We investigate a class of nonlinear optimal control problems with pointwise affine control constraints. Necessary optimality conditions of first order and sufficient second-order conditions are obtained. For the numerical solution of the optimal control problems a semismooth Newton method is proposed. Local superlinear convergence of the infinite-dimensional method is proved. Finally, the properties of the method are tested numerically by controlling the Navier-Stokes equations with affine constraints.

**Mathematics Subject Classification (2000).** 35Q35, 49J20, 65J15, 65K10.

**Keywords.** Optimal control, affine control constraints, semi-smooth Newton methods.

## 1. Introduction

The presence of pointwise control constraints in optimal control problems is of importance if limited control action is allowed to take place. When only one control function is involved, a usual choice consists in imposing pointwise bounds on the control, the so-called box constraints. When multidimensional control functions are involved, however, the box constrained case is just one of the possible practical choices that may be considered. If more complicated or different type of restrictions, such as technological, financial, etc., come into play, then a system of linear pointwise constraints may arise instead of the usual box constraints.

The investigation of optimal control problems with affine constraints for vector-valued control has not been thoroughly carried out yet. Although first-order necessary conditions can be obtained in a straightforward manner, the existence of appropriate Lagrange multipliers has not being studied in depth. With respect to second-order sufficient conditions, results previously obtained for general convex problems can be applied to this case (cf. [1, 7, 12]). However, since the special

structure of the affine constraints is not exploited in such cases, the results may be improved.

For the numerical solution of optimal control problems with affine constraints only few references are available. While box constrained problems are fairly well understood, relatively little research was directed towards devising and analyzing efficient second-order type methods for more general constraints. In [11] the authors consider diagonally dominant systems and prove global convergence of the primal-dual active set strategy applied to this type of problems. Semi-smooth Newton methods for problems with affine constraints have been considered in [6], where optimality conditions were derived and the convergence of the method investigated.

The outline of this paper is as follows. In Section 2 the optimal control problem and its main hypotheses are stated. In Section 3 existence of Lagrange multipliers is proved and a first-order optimality system derived. Second-order sufficient conditions are studied in Section 4. The result avoids the so-called two-norm discrepancy by using a contradiction argument. In Section 5 the superlinear convergence of semi-smooth Newton methods applied to this kind of problems is proved and a semi-smooth Newton algorithm stated. Finally, in Section 6 an optimal control problem of the stationary Navier-Stokes equations with affine constraints is numerically solved and the main properties of the method modified.

## 2. Optimal control problem

Let  $\Omega$  be a bounded domain of  $\mathbb{R}^n$ . We consider the following optimal control problem:

$$\begin{cases} \min J(y) + \frac{\alpha}{2} \|Cu\|_{L^2(\hat{\Omega}, \mathbb{R}^l)}^2 + \frac{\alpha}{2} \|Pu\|_U^2 \\ \text{subject to} \\ e(y, u) = 0 \\ Cu \leq \psi \quad \text{a.e.,} \end{cases} \quad (2.1)$$

where  $\alpha > 0$ ,  $C \in \mathbb{R}^{l \times m}$ ,  $\psi \in L^2(\hat{\Omega}, \mathbb{R}^l)$  and  $P : \mathbb{R}^m \rightarrow \mathbb{R}^m$  is the orthogonal projection onto  $\ker(C)$ . The operator  $e : Y \times U \rightarrow Y'$ , with  $Y$ ,  $U$  Hilbert spaces, is assumed to be of the form:

$$e(y, u) = e_1(y) + e_2 u, \quad (2.2)$$

with  $e_2$  a compact linear operator from  $U$  to  $Y'$  and  $e_1 : Y \rightarrow Y'$  satisfying the conditions in Assumption 2.1 below.

Hereafter we assume that for every  $u \in U$ , there exists a locally unique  $y = y(u) \in Y$  such that  $e(y, u) = 0$ . Moreover, we assume that the corresponding

optimal control problem

$$\begin{cases} \min J(y) + \frac{\alpha}{2} \|Cu\|_{L^2(\hat{\Omega}, \mathbb{R}^l)}^2 + \frac{\alpha}{2} \|Pu\|_U^2 \\ \text{subject to} \\ e_1(y) + e_2 u = 0 \\ Cu \leq \psi \quad \text{a.e.,} \end{cases} \quad (2.3)$$

has a locally unique solution  $(y^*, u^*) \in Y \times U$ .

**Assumption 2.1.** *There exists a neighborhood  $V(y^*)$  of the optimal state  $y^*$  such that:*

- a)  $e_1 : Y \rightarrow Y'$  is twice Fréchet differentiable in  $V(y^*)$ .
- b)  $e_y(y)$  is continuously invertible for each  $y \in V(y^*)$ .
- c)  $e_{yy}$  is Lipschitz continuous in  $V(y^*)$ , i.e., there exists a constant  $L > 0$  such that

$$\|e_{yy}(\bar{y}) - e_{yy}(y^*)\|_{\mathcal{L}(Y \times Y, Y')} \leq L\|\bar{y} - y^*\|_Y, \quad \text{for } \bar{y} \in V(y^*). \quad (2.4)$$

These regularity requirements are needed for the first and second-order optimality conditions as well as for the convergence analysis of the semi-smooth Newton method.

The Hilbert spaces are  $\mathbb{R}^n$ -valued function spaces over a bounded domain  $\Omega \subset \mathbb{R}^n$ , such as  $Y = H^1(\Omega, \mathbb{R}^n)$ . Throughout the space of controls is

$$U = L^2(\hat{\Omega}, \mathbb{R}^m), \quad \hat{\Omega} \subset \Omega \subset \mathbb{R}^n.$$

Further, we choose  $J$  as

$$J(y) = \frac{1}{2}(y, Qy)_Y + (q, y)_Y,$$

where  $Q \in \mathcal{L}(Y, Y)$ ,  $Q \geq 0$  and  $q \in Y$ .

*Example 1.* Let  $\Omega \subset \mathbb{R}^m$ ,  $m \leq 3$ , be a bounded domain. Consider the stationary Navier-Stokes equations

$$\begin{aligned} -\nu \Delta y + (y \cdot \nabla) y + \nabla p &= u && \text{in } \Omega \\ \operatorname{div} y &= 0 && \text{in } \Omega \\ y &= 0 && \text{on } \partial\Omega, \end{aligned}$$

with  $(y \cdot \nabla)y = \sum_{i=1}^m y_i \partial_i y$ ,  $Y = \{H_0^1(\Omega, \mathbb{R}^m) : \operatorname{div} y = 0\}$ ,  $U = L^2(\Omega, \mathbb{R}^m)$ , where  $\nu$  stands for the viscosity coefficient of the fluid,  $y$  for the velocity vector field,  $p$  for the scalar pressure and  $u$  for a distributed body force. The operator  $e_1$  is given by

$$e_1 : Y \rightarrow Y' \quad (2.5)$$

$$y \mapsto \nu(\nabla y, \nabla \cdot)_U - ((y \cdot \nabla)y, \cdot)_U, \quad (2.6)$$

It can be easily verified that the operator  $e_1$  is twice Fréchet differentiable with its first and second derivatives given by

$$e_y(y)w = \nu(\nabla w, \nabla \cdot)_U + ((w \cdot \nabla)y + (y \cdot \nabla)w, \cdot)_U \quad \text{and} \quad e_{yy}(y)[w]^2 = (2(w \cdot \nabla)w, \cdot)_U,$$

respectively. Condition (2.4) follows immediately from the expression for the second derivative. To verify the surjectivity of  $e_y(y)$  let us consider the linearized equation

$$e_y(y)w = \nu(\nabla w, \nabla \cdot)_U + ((w \cdot \nabla)y + (y \cdot \nabla)w, \cdot)_U = \langle g, \cdot \rangle_{Y', Y}, \quad (2.7)$$

with  $g \in Y'$ . We assume that  $\nu$  is sufficiently large so that

$$\nu > \mathcal{M}(y^*) := \sup_{v \in Y} \frac{|(v \cdot \nabla y^*), v)_{\mathbf{L}^2(\Omega)}|}{\|v\|_Y^2}.$$

It can be argued that there exists a neighborhood  $V(y^*) \subset V$  of  $y^*$  such that this inequality remains correct with  $y^*$  replaced by  $y \in V(y^*)$ . Then there exists a unique solution  $w_g$  to the linearized equation (2.7) associated with  $g$  for each  $y \in V(y^*)$ . From the bijectivity of  $e_y(y)$  the continuous invertibility follows. Summarizing, Assumption 2.1 holds for this problem.

Throughout the paper we will use the following assumption with respect to the restriction matrix  $C$ :

**Assumption 2.2.** *The rows  $\{C_i\}_{i \in \mathcal{A}(v, x)}$  are linearly independent in  $\mathbb{R}^m$ , where  $\mathcal{A}(v, x) := \{i : (Cv)_i = \psi_i(x)\}$ , for any  $v \in \mathbb{R}^m$  satisfying  $Cv \leq \psi(x)$  and for a.e.  $x \in \hat{\Omega}$ .*

*Example 2.* For  $U = L^2(\hat{\Omega}, \mathbb{R}^2)$ , the constraints  $u_1 \leq \psi_1, u_2 \leq \psi_2$  result in  $C = I, P = 0$ , which was considered in previous work (see [4, 5]).

*Example 3.* The case  $U = L^2(\Omega, \mathbb{R}^m)$ ,  $u_i \leq 0, -1 \leq \sum_{i=1}^m u_i$  results in  $l = m + 1$ ,  $C = \begin{pmatrix} I \\ -e \end{pmatrix}$ , where  $e = (1, \dots, 1) \in \mathbb{R}^m$ ,  $I$  is the  $m \times m$  identity matrix,  $\psi = (0, \dots, 0, 1)$  and  $P = 0$ . Here Assumption 2.2 is satisfied.

### 3. First-order necessary conditions

In this section, existence of Lagrange multipliers for (2.1) is verified and a first-order optimality system is derived. Note that the differentiability of the control to state mapping in a neighborhood of the optimal solution follows from Assumption 2.1 and the implicit function theorem. From hypothesis b) in Assumption 2.1 the surjectivity of  $e_y(y)$  follows.

Let us set

$$\mathcal{A} = \bigcup_{i=1}^l \mathcal{A}_i$$

with

$$\mathcal{A}_i = \{x \in \hat{\Omega} : C_i u(x) = \psi_i(x)\},$$

and define the inactive set  $\mathcal{I} := \Omega \setminus \mathcal{A}$ .

**Theorem 3.1.** *Let Assumptions 2.1 and 2.2 hold. If  $(y^*, u^*) \in Y \times U$  is a locally unique solution to (2.1), then there exist multipliers  $p \in Y$  and  $\lambda \in L^2(\hat{\Omega}, \mathbb{R}^l)$  such that*

$$e(y^*, u^*) = 0 \quad (3.1)$$

$$(e_y(y^*))^* p = -J'(y^*) \quad (3.2)$$

$$\alpha C^T C u^* + \alpha P u^* + C^T \lambda + e_u^* p = 0 \quad (3.3)$$

$$C u^* \leq \psi, \lambda \geq 0, \lambda^T (C u^* - \psi) = 0 \text{ a.e. in } \hat{\Omega}. \quad (3.4)$$

*Proof.* The first-order necessary and sufficient optimality condition satisfied by  $u^*$  is given by

$$(\alpha C^T C u^* + \alpha P u + e_u^* p, u - u^*)_{L^2(\hat{\Omega})} \geq 0, \text{ for all } C u \leq \psi, \quad (3.5)$$

where

$$\begin{cases} e(y^*, u^*) = 0 \\ (e_y(y^*))^* p = -J'(y^*). \end{cases} \quad (3.6)$$

We define a partitioning of the active set next. Note that by Assumption 2.2 at most  $m$  constraints can be active simultaneously at a.e.  $x \in \hat{\Omega}$ . Let  $\mathcal{P}$  be the set of all subsets of  $\{1, \dots, l\}$  of cardinality  $\leq m$  and set for  $I \in \mathcal{P}$

$$\Omega_I = \{x \in \hat{\Omega} : C_j u^*(x) = \psi_j(x), \text{ for all } j \in I\}.$$

Then  $\hat{\Omega} = \biguplus_{I \in \mathcal{P}} \Omega_I \uplus \mathcal{I}$  and we have

$$\mathcal{A}_i = \{x \in \Omega_I : C_i u(x) = \psi_i(x) \text{ for some } I \in \mathcal{P}\}.$$

We consider the auxiliary problem

$$\begin{cases} \min_{u \in L^2(\hat{\Omega}, \mathbb{R}^m)} J(y) + \|C u\|_{L^2(\hat{\Omega}, \mathbb{R}^l)}^2 + \frac{\alpha}{2} \|P u\|_U^2 \\ \text{subject to:} \\ e(y, u) = 0 \\ C_i u \leq \psi_i \text{ on } \Omega_I \text{ for } i \in I \text{ and all } I \in \mathcal{P}. \end{cases} \quad (P_{\text{aux}})$$

Note that the inequality constraints in  $(P_{\text{aux}})$  can equivalently be expressed as  $C_i u \leq \psi_i$  on  $\mathcal{A}_i$ , for  $i = 1, \dots, l$ . Clearly  $(P_{\text{aux}})$  admits a local unique solution  $\hat{u} \in U$ , since, by hypothesis, (2.1) also does. Associated to  $(P_{\text{aux}})$  we introduce the Lagrangian  $\mathcal{L} : Y \times U \times Y \times Z \rightarrow \mathbb{R}$ , where  $Z = \bigotimes_{I \in \mathcal{P}} L^2(\Omega_I, \mathbb{R}^{\#(I)})$ , with  $\#(I)$  the cardinality of  $I$ ,

$$\begin{aligned} \mathcal{L}(y, u, p, \tilde{\lambda}) &= J(y) + \|C u\|_{L^2(\hat{\Omega}, \mathbb{R}^l)}^2 + \frac{\alpha}{2} \|P u\|_U^2 \\ &\quad + \langle p, e(y, u) \rangle_{Y, Y'} + \sum_{I \in \mathcal{P}} \sum_{i \in I} (\lambda_i^I, C_i u - \psi_i)_{L^2(\Omega_I, \mathbb{R})}. \end{aligned}$$

By Assumptions 2.1 and 2.2 the linearized constraints

$$(e_y(y^*) + e_2, \{(C_i)_{i \in I} : I \in \mathcal{P}\}) : Y \times U \rightarrow Y' \times Z$$

are surjective. Here we identify  $U$  with  $\bigotimes_{I \in \mathcal{P}} L^2(\Omega_I, \mathbb{R}^l) \times L^2(\mathcal{I}, \mathbb{R}^l)$ . Hence there exists  $(p, \{\lambda^I\}_{I \in \mathcal{P}}) \in Y \times Z$  which is a Lagrange multiplier for  $(P_{\text{aux}})$ , i.e.:

$$\left\{ \begin{array}{l} e(\hat{y}, \hat{u}) = 0 \\ (e_y(y^*))^* p = -J'(\hat{y}) \\ \alpha C^T C \hat{u} + \alpha P \hat{u} + e_u^* p + \sum_{I \in \mathcal{P}} \sum_{i \in I} C_i^T \lambda_i^I \chi_{\Omega_I} = 0 \\ C \hat{u} \leq \psi, \text{ in } \hat{\Omega} \\ \lambda_i^I \geq 0, \lambda_i(C_i \hat{u} - \psi_i) = 0, i \in I, I \in \mathcal{P}. \end{array} \right. \quad (3.7)$$

Defining  $\lambda \in L^2(\hat{\Omega}, \mathbb{R}^l)$  by setting

$$\begin{aligned} \lambda_i &= \lambda_i^I \text{ for } i \in I, \text{ and } \lambda_i = 0 \text{ for } i \notin I, & \text{for any } I \in \mathcal{P}, x \in \Omega_I; \\ \lambda_i &= 0 \text{ on } \mathcal{I}, \end{aligned}$$

(3.7) can equivalently expressed as

$$\left\{ \begin{array}{l} e(\hat{y}, \hat{u}) = 0 \\ (e_y(y^*))^* p = -J'(\hat{y}) \\ \alpha C^T C \hat{u} + \alpha P \hat{u} + e_u^* p + C^T \lambda = 0 \\ \lambda \geq 0, C \hat{u} \leq \psi, (\lambda, C \hat{u} - \psi)_{L^2(\hat{\Omega}, \mathbb{R}^l)} = 0. \end{array} \right. \quad (3.8)$$

From (3.8) we obtain for  $Cu \leq \psi$ ,

$$(\alpha C^T C \hat{u} + \alpha P \hat{u} + e_u^* p, u - \hat{u}) = (\lambda, C \hat{u} - Cu) = (\lambda, \psi - Cu) \geq 0.$$

Hence  $\hat{u}$  satisfies the first-order condition (3.5) and therefore  $\hat{u} = u^*$ . System (3.1)–(3.4) follows from (3.8).  $\square$

*Remark 3.2.* Note that from equation (3.3) we have  $\alpha P u^* + P e_u^* p = 0$ .

#### 4. Second-order sufficient optimality condition

In this section we derive a second-order sufficient optimality condition for (2.1). The result makes use of the following cone of critical directions

$$K(u^*) = \left\{ v \in U : (C_j v)(x) \begin{cases} = 0 & \text{if } \lambda_j(x) \neq 0 \\ \leq 0 & \text{if } (Cu^*)_j = \psi_j \text{ and } \lambda_j(x) = 0 \end{cases} \right\},$$

which does not involve strongly active constraints. Moreover, sufficient optimality is obtained without the use of a two-norm discrepancy argument. Rather a technique based solely on the second-order optimality condition (SSC) below is used. The technique was previously applied in [2] to the optimal control of the Navier-Stokes equations with box constraints and in [3] to semilinear state constrained optimal control problems.

For some work concerning second-order conditions for control problems with special kinds of control constraints we refer to [1, 7]. In the cited papers, constraints of the type  $u(x) \in U$ , with  $U$  independent of  $x$  and polygonal, are considered.

In [12] second-order sufficient conditions for control problems with more general convex control constraints are studied. The critical cone used in those cases is, however, not the smallest one. Additionally, the result involves the classical two-norm discrepancy.

**Theorem 4.1.** *Suppose that Assumptions 2.1–2.2 holds and that  $C$  is surjective. Further let  $(y^*, u^*, p^*)$  be a solution of the necessary condition (3.5)–(3.6) and suppose that*

$$\alpha \int_{\hat{\Omega}} |Ch|^2 + \alpha \int_{\hat{\Omega}} |Ph|^2 + (v, Qv) + (p^*, e_{yy}(y^*)[v]^2) > 0 \quad (\text{SSC})$$

holds for every pair  $(v_h, h) \in Y \times K(u^*)$ ,  $(v_h, h) \neq (0, 0)$  that solves the linearized equation

$$e_y(y^*)v_h + e_u h = 0. \quad (4.1)$$

Then there exist  $\varepsilon > 0$  and  $\delta > 0$  such that

$$J(y^*, u^*) + \frac{\delta}{2} \|u - u^*\|_U^2 \leq J(y, u),$$

for every feasible pair  $(y, u)$  such that  $\|u - u^*\|_U \leq \varepsilon$ .

*Proof.* Let us suppose that  $u^*$  does not satisfy the quadratic growth condition. Then there exists a feasible sequence  $\{u_k\}_{k=1}^{\infty} \subset U$  such that

$$\|u_k - u^*\|_U < \frac{1}{k^2} \quad (4.2)$$

and

$$J(y^*, u^*) + \frac{1}{k} \|u_k - u^*\|_U^2 > J(y_k, u_k) = \mathcal{L}(y_k, u_k, p^*) \quad \text{for all } k, \quad (4.3)$$

where  $y_k$  denotes the unique solution of (3.1) associated with  $u_k$ . By defining

$$\rho_k = \|u_k - u^*\|_U \quad \text{and} \quad h_k = \frac{1}{\rho_k} (u_k - u^*).$$

it follows that  $\|h_k\|_U = 1$  and, therefore, we may extract a subsequence, denoted also by  $\{h_k\}$ , such that  $h_k \rightharpoonup h$  weakly in  $U$ . The proof is now given in four steps.

*Step 1:*  $(\frac{\partial \mathcal{L}}{\partial u}(y^*, u^*, p^*)h = 0)$ . From the mean value theorem it follows that

$$\begin{aligned} \mathcal{L}(y_k, u_k, p^*) + \frac{\partial \mathcal{L}}{\partial y}(z_k, u_k, p^*)(y^* - y_k) &= \mathcal{L}(y^*, u_k, p^*) \\ &= \mathcal{L}(y^*, u^*, p^*) + \rho_k \frac{\partial \mathcal{L}}{\partial u}(y^*, w_k, p^*)h_k, \end{aligned}$$

where  $w_k$  and  $z_k$  are points between  $u^*$  and  $u_k$  and  $y^*$  and  $y_k$ , respectively. By (4.3) it follows that

$$\frac{\partial \mathcal{L}}{\partial u}(y^*, w_k, p^*)h_k < \frac{1}{k} \|u_k - u^*\|_U + \frac{1}{\rho_k} \frac{\partial \mathcal{L}}{\partial y}(z_k, u_k, p^*)(y^* - y_k). \quad (4.4)$$

Working on the last term we obtain

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial y}(z_k, u_k, p^*)(y^* - y_k) &= J'(z_k)(y^* - y_k) + \langle p^*, e_y(z_k)(y^* - y_k) \rangle_{Y, Y'} \\ &= J'(z_k)(y^* - y_k) + \langle p^*, e_y(y^*)(y^* - y_k) \rangle_{Y, Y'} \\ &\quad + \langle p^*, e_{yy}(y^*)(z_k - y^*)(y^* - y_k) \rangle_{Y, Y'} \\ &\quad + \langle p^*, (e_{yy}(\zeta_k) - e_{yy}(y^*))(z_k - y^*)(y^* - y_k) \rangle_{Y, Y'}, \end{aligned}$$

with  $\zeta_k = y^* + \xi(z_k - y^*)$ , for some  $\xi \in [0, 1]$ . From the optimality system and Assumption 2.1 we get that

$$\begin{aligned} \left| \frac{\partial \mathcal{L}}{\partial y}(z_k, u_k, p^*)(y^* - y_k) \right| &\leq \|J'(z_k) - J'(y^*)\|_{Y'} \|y^* - y_k\|_Y \\ &\quad + \|p^*\|_Y \|e_{yy}(y^*)\|_{\mathcal{L}(Y \times Y, Y')} \|z_k - y^*\|_Y \|y^* - y_k\|_Y + L \|p^*\|_Y \|z_k - y^*\|_Y^2 \|y^* - y_k\|_Y. \end{aligned}$$

Due to the quadratic nature of  $J$  and since  $h_k \rightharpoonup h$  weakly in  $U$ ,  $w_k \rightarrow u^*$  in  $U$  and  $y_k \rightarrow y^*$  in  $Y$ , we obtain from (4.4) that

$$\frac{\partial \mathcal{L}}{\partial u}(y^*, u^*, p^*)h = \lim_{k \rightarrow \infty} \frac{\partial \mathcal{L}}{\partial u}(y^*, w_k, p^*)h_k \leq 0. \quad (4.5)$$

On the other hand, we know that  $Cu_k(x) \leq \psi(x)$  a.e. in  $\Omega$ , which implies that

$$\frac{\partial \mathcal{L}}{\partial u}(y^*, u^*, p^*)h_k = \rho_k \frac{\partial \mathcal{L}}{\partial u}(y^*, u^*, p^*)(u_k - u^*) \geq 0, \quad (4.6)$$

and consequently

$$\frac{\partial \mathcal{L}}{\partial u}(y^*, u^*, p^*)h = \lim_{k \rightarrow \infty} \frac{\partial \mathcal{L}}{\partial u}(y^*, w_k, p^*)h_k \geq 0.$$

Altogether we obtain that

$$\frac{\partial \mathcal{L}}{\partial u}(y^*, u^*, p^*)h = 0. \quad (4.7)$$

*Step 2:* ( $h \in K(u^*)$ ). The set

$$\{v \in U : (C_j v)(x) \leq 0, \text{ if } (C_j u^*) = \psi_j, \lambda_j(x) = 0, j = 1, \dots, l\}$$

is closed and convex and, therefore, it is weakly sequentially closed. Since each  $h_k$  belongs to this set, then  $h$  also does. From the optimality condition, it follows that  $-\lambda_j(x) C_j h(x) \geq 0$  for all  $j$ , a.e. in  $\Omega$ , which implies that

$$\begin{aligned} 0 &= \frac{\partial \mathcal{L}}{\partial u}(y^*, u^*, p^*)h = (\alpha C^* Cu^* + e_u^* p^*, h)_U \\ &= - \sum_{j=1}^l \int_{\Omega} \lambda_j(x) C_j h(x) dx = \sum_{j=1}^l \int_{\Omega} |\lambda_j(x)| C_j h(x) dx. \end{aligned}$$

Consequently,  $C_j h(x) = 0$  if  $\lambda_j(x) \neq 0$  and, therefore,  $h \in K(u^*)$ .

Step 3: ( $h = 0$ ). From condition (SSC) it suffices to show that

$$\begin{aligned} \frac{\partial^2 \mathcal{L}}{\partial u^2}(y^*, u^*, p^*)h + \frac{\partial^2 \mathcal{L}}{\partial y^2}(y^*, u^*, p^*)v \\ = \alpha \int_{\hat{\Omega}} |Ch|^2 + \alpha \int_{\hat{\Omega}} |Ph|^2 + (v, Qv) + (p^*, e_{yy}(y^*)[v]^2) \leq 0. \end{aligned} \quad (4.8)$$

Using a Taylor expansion of the Lagrangian we get that

$$\begin{aligned} \mathcal{L}(y_k, u_k, p^*) &= \mathcal{L}(y^*, u^*, p^*) + \rho_k \frac{\partial \mathcal{L}}{\partial u}(y^*, u^*, p^*)h_k \\ &\quad + \frac{\rho_k^2}{2} \frac{\partial^2 \mathcal{L}}{\partial u^2}(y^*, u^*, p^*)h_k^2 + \frac{1}{2} \frac{\partial^2 \mathcal{L}}{\partial y^2}(z_k, u^*, p^*)(y_k - y^*)^2, \end{aligned} \quad (4.9)$$

with  $z_k$  an intermediate point between  $y^*$  and  $y_k$ . We therefore get that

$$\begin{aligned} \rho_k \frac{\partial \mathcal{L}}{\partial u}(y^*, u^*, p^*)h_k + \frac{\rho_k^2}{2} \frac{\partial^2 \mathcal{L}}{\partial u^2}(y^*, u^*, p^*)h_k^2 + \frac{\rho_k^2}{2} \frac{\partial^2 \mathcal{L}}{\partial y^2}(y^*, u^*, p^*) \left( \frac{y_k - y^*}{\rho_k} \right)^2 \\ = \mathcal{L}(y_k, u_k, p^*) - \mathcal{L}(y^*, u^*, p^*) \\ + \frac{\rho_k^2}{2} \left[ \frac{\partial^2 \mathcal{L}}{\partial y^2}(y^*, u^*, p^*) - \frac{\partial^2 \mathcal{L}}{\partial y^2}(z_k, u^*, p^*) \right] \left( \frac{y_k - y^*}{\rho_k} \right)^2. \end{aligned} \quad (4.10)$$

Additionally by (4.3),

$$\mathcal{L}(y_k, u_k, p^*) - \mathcal{L}(y^*, u^*, p^*) \leq \frac{\rho_k^2}{k}. \quad (4.11)$$

Since  $u_k \rightarrow u^*$  in  $U$  and  $\|h_k\|_U = 1$ , we obtain from (2.4) that

$$\begin{aligned} &\left| \left[ \frac{\partial^2 \mathcal{L}}{\partial y^2}(y^*, u^*, p^*) - \frac{\partial^2 \mathcal{L}}{\partial y^2}(z_k, u^*, p^*) \right] \left( \frac{y_k - y^*}{\rho_k} \right)^2 \right| \\ &\leq \|p^*\|_Y \|e_{yy}(y^*) - e_{yy}(y_k)\|_{\mathcal{L}(Y^2, Y')} \left\| \frac{y_k - y^*}{\rho_k} \right\|^2 \rightarrow 0 \quad \text{when } k \rightarrow \infty. \end{aligned} \quad (4.12)$$

For the latter we used the fact that, due to the differentiability of the control to state mapping,  $\left\| \frac{y_k - y^*}{\rho_k} \right\|_Y$  is bounded.

Consequently by (4.10),

$$\begin{aligned} &\liminf_{k \rightarrow \infty} \frac{\partial^2 \mathcal{L}}{\partial u^2}(y^*, u^*, p^*)h_k^2 + \liminf_{k \rightarrow \infty} \frac{\partial^2 \mathcal{L}}{\partial y^2}(y^*, u^*, p^*) \left( \frac{y_k - y^*}{\rho_k} \right)^2 \\ &\leq 2 \limsup_{k \rightarrow \infty} \frac{1}{\rho_k^2} (\mathcal{L}(y_k, u_k, p^*) - \mathcal{L}(y^*, u^*, p^*)) - 2 \liminf_{k \rightarrow \infty} \frac{1}{\rho_k} \frac{\partial \mathcal{L}}{\partial u}(y^*, u^*, p^*)h_k. \end{aligned}$$

which implies, since  $\frac{\partial^2 \mathcal{L}}{\partial u^2}(y^*, u^*, p^*)$  is w.l.s.c. and thanks to (4.6), (4.11), that

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(y^*, u^*, p^*)h^2 + \liminf_{k \rightarrow \infty} \frac{\partial^2 \mathcal{L}}{\partial y^2}(y^*, u^*, p^*) \left( \frac{y_k - y^*}{\rho_k} \right)^2 \leq 2 \lim \frac{1}{k} = 0.$$

Additionally,

$$\begin{aligned} \frac{\partial^2 \mathcal{L}}{\partial y^2}(y^*, u^*, p^*) \left( \frac{y_k - y^*}{\rho_k} \right)^2 &= \frac{\partial^2 \mathcal{L}}{\partial y^2}(y^*, u^*, p^*) \left( \frac{y_k - y^*}{\rho_k} - v_{h_k} \right)^2 \\ &+ 2 \frac{\partial^2 \mathcal{L}}{\partial y^2}(y^*, u^*, p^*) \left( \frac{y_k - y^*}{\rho_k} - v_{h_k}, v_{h_k} \right) + \frac{\partial^2 \mathcal{L}}{\partial y^2}(y^*, u^*, p^*) (v_{h_k})^2, \end{aligned}$$

where  $v_{h_k}$  is the solution to (4.1) associated to  $h_k$ , which also corresponds to the derivative of the control-to-state mapping at  $u^*$  in direction  $h_k$ . Due to the differentiability of this mapping, the continuity of the bilinear form  $\frac{\partial^2 \mathcal{L}}{\partial y^2}(y^*, u^*, p^*)$ , and since  $v_{h_k} \rightarrow v_h$  strongly in  $Y$  (by the compactness of  $e_2$ ), we obtain that

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(y^*, u^*, p^*) h^2 + \frac{\partial^2 \mathcal{L}}{\partial y^2}(y^*, u^*, p^*) v_h^2 \leq 2 \lim \frac{1}{k} = 0.$$

Since  $h \in K(u^*)$ , it follows by (SSC) that  $(v_h, h) = (0, 0)$ .

*Step 4: ( $h_k \rightarrow 0$  strongly in  $U$ .)* From the properties of  $C$  and the structure of the cost functional, there exists a constant  $\bar{K} > 0$  such that

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(y^*, u^*, p^*) w^2 \geq \alpha \bar{K} \|w\|_U^2, \text{ for all } w \in U.$$

Since  $h_k \rightharpoonup 0$  weakly in  $U$ , it follows that  $v_{h_k} \rightarrow 0$  strongly in  $Y$  and by (4.10), (4.6), (4.11) and (4.12)

$$\alpha \bar{K} \lim_{k \rightarrow \infty} \sup \|h_k\|_U^2 \leq \lim_{k \rightarrow \infty} \sup \frac{\partial^2 \mathcal{L}}{\partial u^2}(y^*, u^*, p^*) h_k^2 \leq 2 \lim \frac{1}{k} = 0.$$

Thus  $h_k$  converges to 0 strongly. Since  $\|h_k\|_U = 1$ , a contradiction is obtained.  $\square$

## 5. Semismooth Newton method

In this section we turn to the analysis of semismooth Newton methods applied to (3.1)–(3.4). We begin by reformulating the complementarity condition (3.4) as the following operator equation

$$\lambda = \max(0, \lambda + c(Cu - \psi)), \quad (5.1)$$

for any  $c > 0$ , where max is interpreted componentwise and (5.1) must be interpreted in the a.e. in  $\hat{\Omega}$  sense.

Let us hereafter assume that  $C$  is surjective and introduce the Lagrangian

$$\mathcal{L}(y, u, p) = \mathcal{J}(y, u) + \langle p, e(y, u) \rangle_{Y, Y'}.$$

Then  $C^T$  is injective and (3.3) can equivalently be expressed as

$$\begin{cases} \alpha Cu^* + \lambda + D^{-1} C P e_u^* p = 0 \\ \alpha P u^* + P e_u^* p = 0, \end{cases} \quad (5.2)$$

where

$$D = CC^T \in \mathbb{R}^{l \times l}.$$

We therefore obtain that

$$\lambda(x) = -(\alpha Cu^* + D^{-1}Ce_u^*p)(x).$$

Choosing  $c = \alpha$  in (5.1) results in

$$-\alpha Cu - D^{-1}CPe_u^*p = \max(0, -D^{-1}CPe_u^*p - \alpha\psi). \quad (5.3)$$

Considering  $p$  as a function of  $u$  given by equations in (3.1)–(3.2), the optimality system can equivalently be expressed as

$$F(u) = 0, \quad (5.4)$$

where  $F : L^2(\hat{\Omega}; \mathbb{R}^m) \rightarrow L^2(\hat{\Omega}; \mathbb{R}^l) \times L^2(\hat{\Omega}; \mathbb{R}^m)$  is defined by

$$F(u) = \begin{pmatrix} \alpha Cu + D^{-1}Ce_u^*p(u) + \max(0, -D^{-1}Ce_u^*p(u) - \alpha\psi) \\ \alpha Pu + Pe_u^*p(u) \end{pmatrix}. \quad (5.5)$$

Next, we recall the definition of Newton differentiability and a superlinear convergence result for semi-smooth Newton methods [9].

**Definition 5.1.** Let  $X$  and  $Z$  be Banach spaces and  $D \subset X$  an open subset. The mapping  $F : D \rightarrow Z$  is called Newton differentiable on the open subset  $U \subset D$  if there exists a generalized derivative  $G : U \rightarrow L(X, Z)$  such that

$$\lim_{h \rightarrow 0} \frac{1}{\|h\|} \|F(x + h) - F(x) - G(x + h)h\| = 0,$$

for every  $x \in U$ .

**Proposition 5.1.** If  $x^*$  is a solution of  $F(x) = 0$ ,  $F$  is Newton differentiable in an open neighborhood  $U$  containing  $x^*$  and  $\{\|G(y)^{-1}\| : y \in U\}$  is bounded, then the Newton iterations

$$x_{k+1} = x_k - G(x_k)^{-1}F(x_k)$$

converge superlinearly to  $x^*$ , provided that  $\|x_0 - x^*\|_X$  is sufficiently small.

To apply Proposition 5.1 we consider  $X = L^2(\hat{\Omega}, \mathbb{R}^m)$  and  $Z = L^2(\hat{\Omega}, \mathbb{R}^l) \times L^2(\hat{\Omega}, \ker C)$ . In order to define a generalized derivative of  $F$  in the sense of Definition 5.1 we first introduce a generalized derivative for  $\max : L^2(\hat{\Omega}, \mathbb{R}^l) \mapsto L^2(\hat{\Omega}, \mathbb{R}^l)$  by setting

$$(G_m\varphi(x))_i = \begin{cases} 1 & \text{if } \varphi(x)_i > 0 \\ 0 & \text{if } \varphi(x)_i \leq 0. \end{cases} \quad (5.6)$$

From [9] it is known that  $\max : L^q(\hat{\Omega}, \mathbb{R}^l) \mapsto L^2(\hat{\Omega}, \mathbb{R}^l)$  is Newton differentiable with generalized derivative given by (5.6) if  $q > 2$ .

As generalized derivative of  $F$  at  $u$  we choose  $G_F \in \mathcal{L}(L^2(\hat{\Omega}; \mathbb{R}^m), L^2(\hat{\Omega}; \mathbb{R}^l) \times L^2(\hat{\Omega}; \mathbb{R}^m))$  with

$$G_F(u)h = \begin{pmatrix} \alpha Ch + D^{-1}Ce_u^*p'(u; h) - G_m(-D^{-1}Ce_u^*p(u) - \alpha\psi)D^{-1}Ce_u^*p'(u; h) \\ \alpha Ph + Pe_u^*p'(u; h) \end{pmatrix}, \quad (5.7)$$

where  $p'(u; h)$  is solution of

$$\begin{cases} e_y(y)y' + e_u h = 0 \\ e_y^*(y)p'(u; h) = -\mathcal{J}'(y') - (e_y^*)_y(p(u), y'). \end{cases} \quad (5.8)$$

For the subsequent analysis the following additional hypotheses are used:

$$u \rightarrow e_u^* p(u) \text{ is Fréchet differentiable from } L^2(\hat{\Omega}; \mathbb{R}^m) \text{ to } L^q(\Omega; \mathbb{R}^n), \text{ for some } q > 2. \quad (\text{H1})$$

$$e_y(y)^* \text{ is uniformly continuously invertible in } V(y^*). \quad (\text{H2})$$

$$\begin{cases} \text{any solution } (v, h) \in Y \times U \text{ of the linearized equation} \\ e_y(y)v + e_u h = 0 \text{ satisfies, for } y \in V(y^*), \text{ the estimate} \\ |v|_Y \leq \sqrt{K}|h|_U, \text{ with } K \text{ independent of } y. \end{cases} \quad (\text{H3})$$

Hypothesis (H1) guarantees the Newton differentiability of the operator (5.5).

The following stronger second-order condition is also assumed to hold: there exists a constant  $\kappa > 0$  such that

$$\alpha \int_{\hat{\Omega}} |Ch|^2 + \alpha \int_{\hat{\Omega}} |Ph|^2 + (v, Qv) + (p(u^*), e_{yy}(y^*)[v]^2) \geq \kappa \|h\|_U^2 \quad (\text{SSC}')$$

holds for every pair  $(v, h) \in Y \times U$  that solves the linearized equation (4.1).

**Theorem 5.2.** *Let  $C : \mathbb{R}^m \rightarrow \mathbb{R}^l$  be surjective and let Assumptions 2.1–2.2, (H1), (H2), (H3) and (SSC') hold. Then the semi-smooth Newton method applied to  $F(u) = 0$ , with  $F$  given in (5.5) and generalized derivative  $G_F(u)$  as in (5.7) converges locally superlinearly.*

*Proof.* We need to verify the hypotheses of Proposition 5.1. Newton differentiability of  $F$  follows from (H1), the Newton differentiability of the max-function from  $L^q(\hat{\Omega})$  to  $L^2(\hat{\Omega})$  and the chain rule for Newton-differentiable functions, see [10]. It remains to argue uniform boundedness of the inverse of the generalized derivative.

We define for  $i = 1, \dots, l$

$$\tilde{\mathcal{A}}_i = \{x \in \hat{\Omega} : (-D^{-1}Ce_u^*p(u) - \alpha\psi)_i(x) > 0\} \text{ and } \tilde{\mathcal{I}}_i = \hat{\Omega} - \tilde{\mathcal{A}}_i.$$

and the diagonal matrix-valued function  $\chi_{\tilde{\mathcal{A}}} \in L^2(\hat{\Omega}, \mathbb{R}^{m \times n})$  with

$$(\chi_{\tilde{\mathcal{A}}})_{i,i} = \chi_{\tilde{\mathcal{A}}_i} \text{ for } i = 1, \dots, l, \text{ and } (\chi_{\tilde{\mathcal{A}}})_{i,j} = 0, \text{ for } i \neq j,$$

and analogously for  $\chi_{\tilde{\mathcal{I}}}$ . We have to analyze the equation

$$G_F(u)h = (f_1, f_2),$$

for  $(f_1, f_2) \in L^2(\hat{\Omega}; \mathbb{R}^l) \times L^2(\hat{\Omega}; \ker C)$  and  $h \in L^2(\hat{\Omega}; \mathbb{R}^m)$ , which can also be written as

$$\begin{cases} \alpha \chi_{\tilde{\mathcal{A}}} Ch = \chi_{\tilde{\mathcal{A}}} f_1 \\ \alpha \chi_{\tilde{\mathcal{I}}} Ch + \chi_{\tilde{\mathcal{I}}} D^{-1}Ce_u^*p'(u; h) = \chi_{\tilde{\mathcal{I}}} f_1 \\ \alpha Ph + Pe_u^*p'(h) = f_2, \end{cases} \quad (5.9)$$

Let us now consider the following auxiliary problem:

$$\begin{cases} \min \mathcal{J}_a(v, h) = \frac{1}{2}(v, Qv)_Y + \frac{\alpha}{2} |\chi_{\tilde{\mathcal{I}}}(Ch - g_1)|_{L^2(\tilde{\mathcal{I}}, \mathbb{R}^l)}^2 \\ \quad + \frac{\alpha}{2} |Ph - g_2|_U^2 + \frac{1}{2} \langle p, e_{yy}(y)[v]^2 \rangle_{Y, Y'} \\ \text{subject to:} \\ e_y(y, u)v + e_u h = 0 \\ \chi_{\tilde{\mathcal{A}}} Ch = \chi_{\tilde{\mathcal{A}}} g_1 \end{cases} \quad (5.10)$$

with  $(g_1, g_2) = (\frac{1}{\alpha}f_1, \frac{1}{\alpha}f_2)$ . To verify that (5.9) is the optimality condition for (5.10), let us introduce the Lagrangian

$$\mathcal{L} = \mathcal{J}_a(v, h) + \langle q, e_y(y, u)v + e_u h \rangle_{Y, Y'} + \langle \mu, \chi_{\tilde{\mathcal{A}}}(Ch - g_1) \rangle. \quad (5.11)$$

Taking the derivative with respect to  $v$  yields

$$e_y^*(y)q + e_{yy}^*(y)p v = -Qv. \quad (5.12)$$

We define  $\mathcal{E} : Y \times U \rightarrow Y' \times \bigotimes_{i=1}^m L^2(\tilde{\mathcal{A}}_i, \mathbb{R})$  by

$$\mathcal{E}(v, h) = \begin{pmatrix} e_y(y)v + e_u h \\ \chi_{\tilde{\mathcal{A}}}(Ch - g_1) \end{pmatrix}.$$

with  $\ker(\mathcal{E}') = \{(v, h) \in Y \times U : \chi_{\tilde{\mathcal{A}}} Ch = 0, e_y(y)v + e_u h = 0\}$ .

The Hessian of  $\mathcal{J}_a$  is given by

$$\mathcal{J}_a''(\delta v, \delta h)^2 = (\delta v, Q\delta v)_Y + \alpha |\chi_{\tilde{\mathcal{I}}} C \delta h|_{L^2(\tilde{\mathcal{I}}, \mathbb{R}^l)}^2 + \alpha |P\delta h|_U^2 + \langle p, e_{yy}(y)[\delta v]^2 \rangle_{Y, Y'}.$$

For  $(\delta v, \delta h) \in \ker(\mathcal{E}')$  we therefore obtain that

$$\mathcal{J}_a''(\delta v, \delta h)^2 \geq \alpha \int_{\tilde{\Omega}} |C\delta h|^2 + \alpha \int_{\tilde{\Omega}} |P\delta h|^2 + (\delta v, Q\delta v) + \langle p, e_{yy}(y)[\delta v]^2 \rangle_{Y, Y'},$$

which by (SSC') and the Lipschitz continuity of  $e_{yy}$  implies the existence of a constant  $\bar{K} > 0$ , independent of  $u$ , such that

$$\mathcal{J}_a''(\delta v, \delta h)^2 \geq \alpha \bar{K} |\delta h|_U^2 \text{ for all } (\delta v, \delta h) \in \ker(\mathcal{E}'), \quad (5.13)$$

in a neighborhood of  $u^*$ . Additionally, due to (H2) we obtain that

$$\mathcal{J}_a''(\delta v, \delta h)^2 \geq \frac{\alpha \bar{K}}{2K} \|(\delta v, \delta h)\|_{Y \times U}^2 \text{ for all } (\delta v, \delta h) \in \ker(\mathcal{E}'). \quad (5.14)$$

The auxiliary problem is therefore a linear quadratic optimization problem with convex objective function and, consequently, there exists a unique solution to (5.9).

Moreover, since  $\mathcal{E}'(y)$  is surjective, there exist multipliers  $(q, \varphi)$  such that the Lagrangian is stationary at  $(v, h, q, \varphi)$ , i.e.,

$$\begin{cases} e_y(y)v + e_u h = 0, \\ e_y^*(y)q = -Qv - e_{yy}^*(y)p v, \\ \chi_{\tilde{\mathcal{A}}}(Ch^* - g_1) = 0, \\ \alpha C^T \chi_{\tilde{\mathcal{I}}}(Ch^* - g_1) + \alpha(Ph^* - g_2) + e_u^* p^* + C^T \chi_{\tilde{\mathcal{A}}} \varphi = 0 \end{cases} \quad (5.15)$$

Projecting the last equation with respect to  $P$  and  $I - P = C^T(CC^T)^{-1}C$ , we get

$$\begin{cases} e_y(y, u)v + e_u h = 0, \\ e_y^*(y, u)q = -Qv - e_{yy}^*(y, u)p v \\ \chi_{\tilde{\mathcal{A}}}(Ch^* - g_1) = 0, \\ \alpha\chi_{\tilde{\mathcal{I}}} Ch^* + \chi_{\tilde{\mathcal{I}}} D^{-1}Ce_u^* q = \alpha\chi_{\tilde{\mathcal{I}}} g_1 \\ \chi_{\tilde{\mathcal{A}}}\varphi + \chi_{\tilde{\mathcal{A}}} D^{-1}Ce_u^* q = 0 \\ \alpha Ph + Pe_u^* q = \alpha g_2. \end{cases} \quad (5.16)$$

For the bounded invertibility analysis we consider the equivalent system:

$$\begin{cases} \mathcal{J}_a''(v, h) + (\mathcal{E}')^*(q, \varphi) = (0, \alpha C^T \chi_{\tilde{\mathcal{I}}} g_1 + \alpha g_2)_{Y' \times L^2(\hat{\Omega}, \mathbb{R}^m)}^T \\ \mathcal{E}'(v, h) = (0, \chi_{\tilde{\mathcal{A}}} g_1)^T. \end{cases} \quad (5.17)$$

Since  $\mathcal{E}'(y)(v, h)$  is surjective, the following decomposition holds:

$$(v, h) = (v_k, h_k) + (v_r, h_r) \in \ker(\mathcal{E}') \oplus \text{range}(\mathcal{E}')^*$$

From the third equation in (5.17) we obtain that  $\chi_{\mathcal{A}}Ch_r = \chi_{\mathcal{A}}g_1$ , which implies that  $|h_r|_{L^2} \leq K_2|g_1|$  (since  $\chi_{\mathcal{A}}C$  is invertible on  $\text{range}(\mathcal{E}')^*$ ). Also since  $(v_r, h_r)$  satisfies equation  $e_y(y)v_r + e_2h_r = 0$ , we obtain from (H3) that  $|v_r|_Y \leq K|h_r|_U$ . Therefore we obtain the bound

$$|(v_r, h_r)|_{Y \times U} \leq K_1|g_1|_{L^2(\hat{\Omega}, \mathbb{R}^l)} \quad (5.18)$$

From the first equation in (5.17) we obtain, since  $(\mathcal{E}'(y))^*$  is continuously invertible on its range, that

$$(q, \varphi) = (\mathcal{E}'(y))^{-*}[-\mathcal{J}_a''(v, h) + (0, \alpha C^T \chi_{\mathcal{I}} g_1 + \alpha g_2)^T]. \quad (5.19)$$

Moreover, by assumption (H2) there exists  $C > 0$ , independent of  $y$ , such that  $\|(\mathcal{E}'(y))^{-*}\| \leq C$ , for all  $y \in V(y^*)$ . Therefore,

$$|(q, \varphi)|_{Y \times L^2(\mathcal{A})} \leq C(|(v, ah)|_{Y \times U} + \alpha|(g_1, g_2)|_{L^2(\hat{\Omega}, \mathbb{R}^l) \times L^2(\hat{\Omega}, \ker(C))}).$$

Using (5.14) and (5.17) we find

$$\begin{aligned} \frac{\alpha\bar{K}}{2K}|(v_k, h_k)|_{Y \times U}^2 &\leq \langle \mathcal{J}_a''(v_k, h_k), (v_k, h_k) \rangle \\ &= \langle \mathcal{J}_a''(v, h), (v, h) \rangle - 2\langle \mathcal{J}_a''(v_k, h_k), (v_r, h_r) \rangle - \langle \mathcal{J}_a''(v_r, h_r), (v_r, h_r) \rangle \\ &= \alpha(\chi_{\tilde{\mathcal{I}}} g_1, \chi_{\tilde{\mathcal{I}}} Ch)_{L^2(\hat{\Omega}, \mathbb{R}^l)} + \alpha(g_2, h_r)_{L^2(\hat{\Omega}, \mathbb{R}^l)} + \alpha(g_2, \chi_{\tilde{\mathcal{I}}} h_k)_{L^2(\hat{\Omega}, \mathbb{R}^l)} \\ &\quad - (\varphi, \chi_{\tilde{\mathcal{A}}} g_1)_{L^2(\tilde{\mathcal{A}}, \mathbb{R}^l)} - 2\langle \mathcal{J}_a''(v_k, h_k), (v_r, h_r) \rangle - \langle \mathcal{J}_a''(v_r, h_r), (v_r, h_r) \rangle \\ &\leq \alpha(\chi_{\tilde{\mathcal{I}}} g_1, \chi_{\tilde{\mathcal{I}}} Ch)_{L^2(\hat{\Omega}, \mathbb{R}^l)} + \alpha(g_2, h_r)_{L^2(\hat{\Omega}, \mathbb{R}^l)} + \alpha(g_2, \chi_{\tilde{\mathcal{I}}} h_k)_{L^2(\hat{\Omega}, \mathbb{R}^l)} \\ &\quad - (\varphi, \chi_{\tilde{\mathcal{A}}} g_1)_{L^2(\tilde{\mathcal{A}}, \mathbb{R}^m)} - 2\langle \mathcal{J}_a''(v_k, h_k), (v_r, h_r) \rangle - \langle e_{yy}^*(y) \cdot p \cdot v_r, v_r \rangle_{Y' \times Y}. \end{aligned}$$

From (5.18) and (5.19) we obtain the existence of constants  $K_3$  and  $K_4$  such that

$$(\mu, \chi_{\tilde{\mathcal{A}}} g_1)_{L^2(\tilde{\mathcal{A}}, \mathbb{R}^l)} \leq K_3\alpha(|g_1|^2 + |g_2|^2 + \frac{1}{\alpha^2}|g_1|^2) + \frac{\alpha\bar{K}}{8K}|h_k|^2$$

and

$$2\langle \mathcal{J}_a''(v_k, h_k), (v_r, h_r) \rangle \leq K_4 \alpha \left( |g_1|^2 + \frac{1}{\alpha^2} |g_1|^2 \right) + \frac{\alpha \bar{K}}{8K} |v_k|^2.$$

These estimates imply that

$$\begin{aligned} \frac{\alpha \bar{K}}{2K} |(v_k, h_k)|_{Y \times U}^2 &\leq \alpha \|C\| |g_1|(|h_k| + |h_r|) + \alpha |g_2| |h_k| + (v_r, Q v_r) \\ &\quad + \alpha |g_2| |h_r| + \alpha (K_3 + K_4) \left( |g_1|^2 + |g_2|^2 + \frac{1}{\alpha^2} |g_1|^2 \right) + \frac{\alpha \bar{K}}{4K} |v_k|^2. \end{aligned}$$

From (5.18) and (SSC') there exists a constant  $K_5 > 0$  such that

$$|(v_k, h_k)|_{Y \times U}^2 \leq K_5 (|g_1|^2 + |g_2|^2 + \frac{1}{\alpha^2} |g_1|^2),$$

and consequently

$$|(v, h)|_{Y \times U} \leq K_6 \left( |f_1|_{L^2(\hat{\Omega}, \mathbb{R}^l)} + \frac{1}{\alpha} |f_1|_{L^2(\hat{\Omega}, \mathbb{R}^l)} + |f_2|_{L^2(\hat{\Omega}, \ker C)} \right).$$

This estimate implies the a priori bound on the inverse of  $G_F(u)$  uniformly.  $\square$

A complete semi-smooth Newton step for problem (2.3) is then given by the following algorithm.

### Algorithm 5.3 (Semi-smooth Newton method (SSN)).

1. Initialize,  $u_0, k = 0$
2. Solve  $G_F(u_k) \delta u_k = -F(u_k)$ .
3. Set  $u_{k+1} = u_k + \delta u_k$ .
4. Solve  $e(y, u_{k+1}) = f$  for  $y_{k+1}$ .
5. Solve  $(e_y(y_k))^* p = -\mathcal{J}'(y_{k+1})$  for  $p_{k+1} = p(u_{k+1})$ .
5. Stop or set  $k = k + 1$ , goto 2.

Note that the equation in Step 3 may be solved by using Newton's method.

## 6. Numerical experiment

In this section we test the efficiency of Algorithm 5.3 for solving an optimal control problem governed by the stationary Navier-Stokes equations in presence of affine control constraints. For the numerical experiment a forward facing step channel of length 1 and height 0.5 was considered. The fluid enters the channel on the left with Dirichlet boundary condition of parabolic type and leaves the channel on the right with stress free boundary condition. The domain is discretized using a homogeneous staggered grid with step  $h$ . A first-order upwind finite differences scheme is used to approximate the flow equations.

The target of the control problem is to drive the fluid to an almost linear behavior given by the Navier-Stokes flow with Reynolds number equal to 1 and, in this manner, reduce recirculations before and after the step. The  $\Re = 1$  flow was therefore chosen as desired state  $z_d$ .

For the solution of the discretized systems appearing in each semi-smooth Newton step a penalty method was applied (cf. [8, p. 125]). The resulting linear systems in each SSN iteration were solved using MATLAB exact solver. All algorithms were implemented in MATLAB 7.6 and run in an Intel Xeon Quart Core machine with a precision of  $\text{eps} = 2.2204e - 16$ .

The semi-smooth Newton algorithm stops if the  $L^2$ -residuum of the discretized control is lower than a given tolerance, typically set as  $10^{-6}$ . The method is initialized with the solution of the unconstrained optimal control problem.

To verify the main properties of the method we introduce the quantities

$$\varrho_k = \|u_k - u_{k-1}\|_{\mathbf{L}_h^2}, \quad \vartheta_k = \frac{\|u_k - u_{k-1}\|_{\mathbf{L}_h^2}}{\|u_{k-1} - u_{k-2}\|_{\mathbf{L}_h^2}},$$

whose purpose is to evaluate the difference of two consecutive controls and the convergence rate, respectively. The  $L^2$ -norms are evaluated by using a rectangle formula.

For the numerical experiment we consider the optimal control problem (2.1) with

$$C = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad \psi = \begin{pmatrix} 0.8 \\ 0.4 \end{pmatrix}$$

as constraint matrix and vector respectively. The correspondent projection matrix is given by  $P = 0$ . The remaining parameter values are  $\alpha = 0.01$  and  $Re = 800$ . In further numerical tests which will be reported elsewhere the significance of using the controls weighted by  $C$  as in (2.1) will be demonstrated.

With a mesh step size  $h = 1/160$ , the algorithm stops after 4 iterations. The size of the resulting active sets are 968 and 680 for the first and the second constraints, respectively. The constraints and the correspondent multipliers for the optimal solution are depicted in Figure 1. From the graphics the complementarity condition can be verified by inspection.

In Table 1 the convergence history is documented. From the data, superlinear rate of convergence can be inferred. Also a monotonic behavior of the cost functional value can be observed.

Iteration	$ \mathcal{A}_k^1 $	$ \mathcal{A}_k^2 $	$J(y, u)$	$\varrho_k$	$\vartheta_k$
0	0	0	0.00141942	-	-
1	864	592	0.00146985	0.007206	-
2	966	683	0.00146987	1.3343e-5	0.001851
3	968	680	0.00146987	1.5717e-10	1.1779e-5
4	968	680	0.00146987	1.5807e-24	1.005e-14

TABLE 1. Example 1,  $\Re = 800$ ,  $\alpha = 0.01$ ,  $h = 1/160$ .

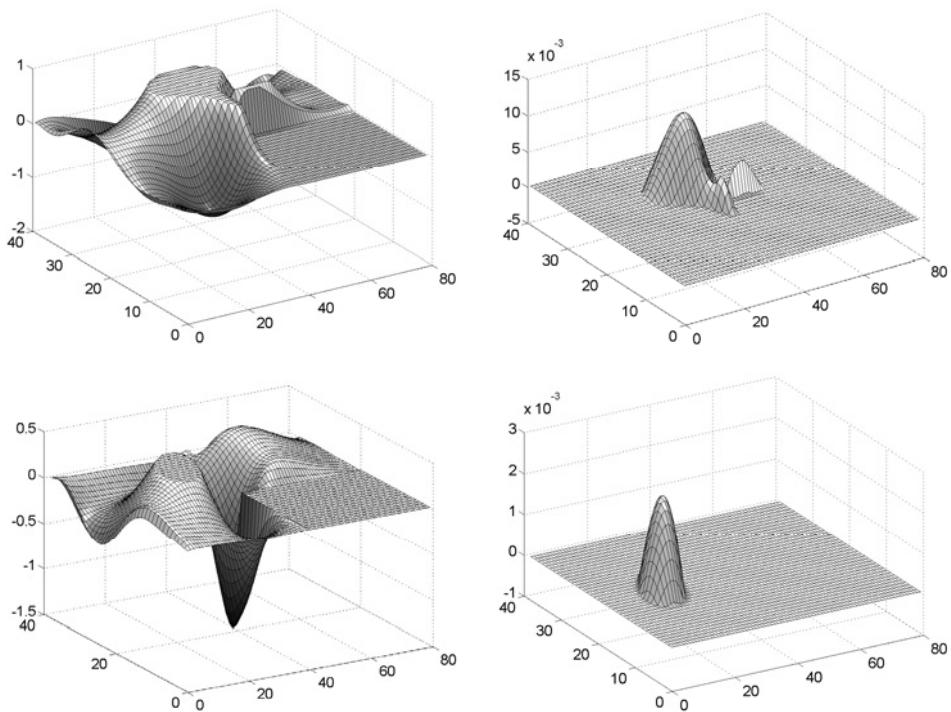


FIGURE 1. Affine constraints (left) and their multipliers (right);  $h = 1/80$

## References

- [1] F. Bonnans, *Second-order analysis for control constrained optimal control problems of semilinear elliptic systems*, Applied Mathematics and Optimization **38** (1998), 303–325.
- [2] Eduardo Casas, Mariano Mateos, and Jean-Pierre Raymond, *Error estimates for the numerical approximation of a distributed control problem for the steady-state Navier–Stokes equations*, SIAM Journal on Control and Optimization **46** (2007), no. 3, 952–982.
- [3] Eduardo Casas, Juan Carlos De Los Reyes, and Fredi Tröltzsch, *Sufficient second-order optimality conditions for semilinear control problems with pointwise state constraints*, SIAM Journal on Optimization **19** (2008), no. 2, 616–643.
- [4] J.-C. De Los Reyes, *Primal-dual active set method for control constrained optimal control of the Stokes equations*, Optimization Methods and Software **21** (2006), no. 2, 267–293.
- [5] J.-C. De Los Reyes and K. Kunisch, *A semi-smooth Newton method for control constrained boundary optimal control of the Navier-Stokes equations*, Nonlinear Analysis. Theory, Methods & Applications **62** (2005), no. 7, 1289–1316.

- [6] ———, *Optimal control of partial differential equations with affine control constraints*, Tech. report, Institute of Mathematics and Scientific Computing, University of Graz, Austria, 2008.
- [7] Joseph C. Dunn, *Second-order optimality conditions in sets of  $l^\infty$  functions with range in a polyhedron*, SIAM Journal on Control and Optimization **33** (1995), no. 5, 1603–1635.
- [8] M. Gunzburger, *Navier-Stokes equations for incompressible flows: finite-element methods*, Handbook of Computational Fluid Mechanics, Academic Press, San Diego, 2000, pp. 99–158.
- [9] M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semismooth Newton method*, SIAM Journal on Optimization **13** (2002), no. 3, 865–888.
- [10] Kazufumi Ito and Karl Kunisch, *The primal-dual active set method for nonlinear optimal control problems with bilateral constraints*, SIAM Journal on Control and Optimization **43** (2004), no. 1, 357–376.
- [11] ———, *Convergence of the primal-dual active set strategy for diagonally dominant systems*, SIAM Journal on Control and Optimization **46** (2007), no. 1, 14–34.
- [12] D. Wachsmuth, *Sufficient second-order optimality conditions for convex control constraints*, Journal of Mathematical Analysis and Applications **319** (2006), no. 1, 228–247.

Juan Carlos De Los Reyes  
 Departamento de Matemática  
 EPN Quito, Ecuador  
 e-mail: jcde losreyes@math.epn.edu.ec

Karl Kunisch  
 Institute for Mathematics and Scientific Computing  
 University of Graz, Austria  
 e-mail: karl.kunisch@uni-graz.at

# Weak Solutions to a Model for Crystal Growth from the Melt in Changing Magnetic Fields

Pierre-Étienne Druet

**Abstract.** We present a model for crystal growth from the melt that accounts for the interaction between melt flow, heating process, and additional applied alternating or travelling magnetic fields. Functional setting and variational formulation are derived for the quasi-stationary approximation of the model.

**Mathematics Subject Classification (2000).** Primary 35D05; Secondary 76D05, 78A55, 80A20.

**Keywords.** Nonlinear system of PDE, Navier-Stokes equations, Maxwell's equations, nonlocal radiation boundary conditions.

## 1. Introduction

In the last years, applied mathematics has discovered in industrial crystal growth a field rich of interesting problems. Due to the high-temperatures and the high costs of experiments that characterize crystal growth, specific knowledge has often to be obtained via *mathematical modeling* and *numerical simulations* (see [Phi03], [KPS04], [Voi01]). In the search for means to systematically improve the production, the tools developed in the mathematical *theory of optimal control* have to be mobilized (see [Mey06], [MPT06], [GM06], [HZ07]). The problems posed in crystal growth are very challenging, since their mathematical modeling leads to strongly coupled systems of nonlinear PDEs for which few results have been stated.

In this paper, we want to introduce a model that aims at describing the heat-transfer mechanisms, the melt flow, and their interaction with applied magnetic fields, in crystal growth from the melt. In the first section, we briefly describe the main physical phenomena. We then introduce in the second section the mathematical model, successively for hydrodynamics, heat-transfer and electromagnetics. In the last section, we propose a natural mathematical setting and a concept of weak solution to the system.

## 2. Czochralski's method in crystal growth. The melt instability

Czochralski's method for the growth of single crystals basically consists in inducing recrystallization of a melted polycrystalline material around a single crystal seed. This idea is nowadays realized at very large scales by the semiconductor industry.

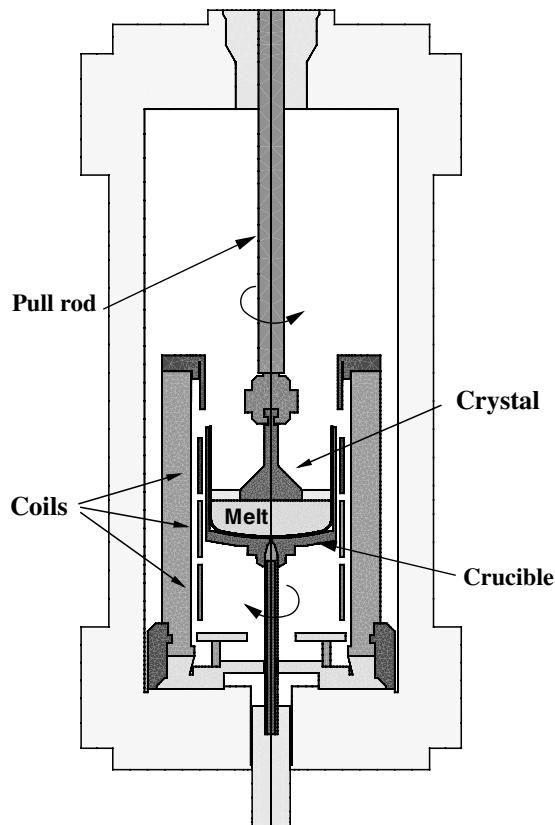


FIGURE 1. Schematic cross-sectional representation of a growth arrangement of the Institute of Crystal Growth (IKZ) Berlin.

Figure 1 represents a typical high-temperatures furnace for the growth of single crystals. The polycrystalline semiconducting material is filled in the crucible. Once the material has been melted, the pull rod at the top of the furnace is used to dip a single crystal seed into the melt. The art of crystal growth consists in adjusting the growth parameters to create a thermodynamical equilibrium near the seed, so that it is able to support a meniscus of liquid. The seed is then lifted slowly and recrystallization can occur through cooling at the contact of the colder

gas phase. The rotation of the pull rod ensures the circular shape of the crystal, the so-called *ingot form*.

For the production of reliable electronic devices, crystals of high quality are required. Determining for this quality are chiefly the thermodynamical parameters at the crystallization interface, such as for instance the shape of the free phase boundary. The *melt motion* in the crucible is also a factor of decisive influence.

The flow in the melt is principally due to thermal convection, originating from the temperature difference between the surface of the melt and the warmer bottom of the crucible. A liquid subject to a temperature gradient is *thermally unstable*, and this is assumed to be responsible for the formation of inhomogeneities in the crystal lattice. The large melt dimensions used in industry are also a factor diminishing the influence of viscous forces, that is of stabilization, on the flow.

Applied magnetic fields are known to provide the possibility to influence unstable flows in melted metals<sup>1</sup>. It has been confirmed in numerous examples that the effect of the magnetic field in such cases amounts to *increasing the viscosity* of the fluid. A basic explanation of the damping effect exerted by magnetic fields on thermally unstable fluids is that the resistivity of the fluid acts as a second viscosity. The Joule effect increases the quantity of heat produced in the fluid<sup>2</sup>, so that thermal instability can only set in at higher temperature gradients.

**Magnetic fields in crystal growth. The project KristMAG.** In the particular area of crystal growth, detailed investigations of different types of magnetic fields and their specific effect are at the center of intense research.

The theoretical practicability of the melt stabilization by magnetic fields is not yet equivalent to technical feasibility, let alone to a rentable use in industry. A field sufficiently strong to show a positive influence on the melt is to realize only at the cost of a hight additional electrical input power.

Some of the open questions related have been recently investigated in the project *KristMAG* (see <http://www.kristmag.com>). In this project, a technological innovation was proposed to make *travelling magnetic fields* for Czochralski crystal growth attractive for the industry (see [Rud07]): the induction coils that usually surround the furnace are replaced by a resistance heater in the furnace, specifically designed to at the same time generate a travelling magnetic field (see Figure 1). In this way of doing, the power used to heat the furnace is cleverly redirected to give control possibilities on the melt.

---

<sup>1</sup>The reference [Cap72], volume III, pages 128–135, describes how a steady state magnetic field is used to increase the stability region of the convective flow in a Bénard cell. A more detailed analysis of the same problem can be found in [Cha81], Chapter I–VI, in particular Chapter III.

<sup>2</sup>[Cap72], or [Cha81], page 160

### 3. Describing the melt flow and the global heat transfer in a crystal growth furnace

In the modeling of crystal growth, one usually distinguishes between *global* and *local* considerations, according to whether the entire furnace is considered (see for instance [Voi01], [KPS04]), or a part of it, typically the system crystal-melt, is decoupled and treated separately ([HZ07], [GM06]). A decoupling the system crystal-melt from the rest of the apparatus is very desirable from the point of view of numerical analysis, since the domain of computation is then substantially smaller, but in the present paper, we take the global point of view and focus on the interaction between heat transfer, melt flow and electromagnetic fields. From this viewpoint, the free boundaries (interface crystal-melt and melt-gas) do not play the most important role and can in first approximation be treated as fixed and flat.

In order to formulate the mathematical problem, we at first need to introduce a description of the geometry that fits realistic situations such as represented on Figure 1. Note that in most situations, it is not realistic to assume that the applied magnetic field is confined to the region of interest for the computation of temperature, the furnace. The geometry considered through this paper thus contains the following ingredients:

1. A simply connected, bounded domain  $\tilde{\Omega} \subset \mathbb{R}^3$  that represent the region of extension of the electromagnetic fields. This domain has the representation  $\overline{\tilde{\Omega}} = \bigcup_{i=0}^m \overline{\tilde{\Omega}_i}$ , where the domains  $\tilde{\Omega}_i$  ( $i = 0, \dots, m$ ) are disjoint, and represent the different materials filling this region.
2. We denote by  $\Omega \subseteq \tilde{\Omega}$  the bounded domain that represent the region of interest for the computation of the temperature (furnace). Setting  $\Omega_i := \tilde{\Omega}_i \cap \Omega$  for  $i = 0, \dots, m$ , we obviously have  $\overline{\Omega} = \bigcup_{i=0}^m \overline{\Omega_i}$ .
3. One of the material in the furnace  $\Omega$ , say  $\Omega_0$ , is transparent and fills a connected cavity. The remaining materials are opaque. We set  $\Omega_{\text{op}} := \Omega \setminus \Omega_0$ . The transparent cavity is *enclosed* in  $\Omega$ , that means, the set  $\mathbb{R}^3 \setminus \Omega_{\text{op}}$  is disconnected.
4. The crucible, containing the melted semiconducting material, is denoted by  $\Omega_1$ .
5. The set of electrical conductors that are located respectively in  $\tilde{\Omega}$  and in  $\Omega$  are respectively denoted by  $\tilde{\Omega}_c$  and  $\Omega_c$ .
6. We denote by  $\tilde{\Omega}_{c_0} \subset \tilde{\Omega}_c$  the conductors in which current is applied. They correspond in Picture 1 to the coils inside of the furnace. Throughout the paper, we also allow for the usual case that induction coils are located outside the furnace, but we restrict for simplicity to the practically relevant case that  $\tilde{\Omega}_{c_0}$  consists of closed current loops.
7. Nonlocal radiation interaction take place at the boundary  $\partial\Omega_0$  of the transparent cavity. We use the usual notation  $\Sigma := \partial\Omega_0$  and  $\Gamma := \partial\Omega$  for the external boundary of the furnace.

The model for the melt that we propose here essentially follows [Voi01]. Global heat transfer is modeled with an approach similar to [KPS04] for computing the heat sources from the Maxwell equations. The model for heat radiation is of wide use in crystal growth and is also described in [KPS04], [Voi01], [Tii97] and other publications. Our main references for modeling the magnetic field is the book [Bos04].

### 3.1. The model for the fluid flow

The melt flow is governed by the full Navier-Stokes equations for a viscous, electrically conducting and heat-conducting fluid. However, it is widely accepted that thermal (natural) convection in liquids can be reasonably described by Boussinesq's approximation (see [GG76] for a general description). According to the Boussinesq model, it is possible to assume that the fluid is *incompressible in the mean*. The velocity  $v$  and the pressure  $p$  in the melt are consequently assumed to satisfy the Navier-Stokes equations in the form

$$\rho_1 \left( \frac{\partial v}{\partial t} + (v \cdot \nabla)v \right) = -\nabla p + \operatorname{div}(2\eta(\theta) D v) + F, \\ \operatorname{div} v = 0, \quad \text{in } ]0, T[ \times \Omega_1. \quad (3.1)$$

where the reference mass density  $\rho_1$  of the fluid is a given constant, the function  $\eta$  denotes the dynamical viscosity of the fluid, which may depend on temperature, and  $D v$  is the rate of strain tensor, with the notations

$$D v = D_{i,j}(v) := \frac{1}{2} \left( \frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right) \quad (i, j = 1, \dots, 3), \quad (3.2)$$

$$D(u, v) := D u : D v := D_{i,j}(u) D_{i,j}(v). \quad (3.3)$$

Here and throughout this paper, we use the convention that repeated indices imply summation over 1, 2, 3.

$F$  denotes the external force, which is twofold. On the one hand the melt flow in Czochralski crystal growth is mainly driven by buoyancy. Denoting by  $\rho$  the mass density of the fluid, and using linear expansion, we write for the thermal expansion of the fluid

$$\rho = \rho(\theta) = \rho_1 (1 - \alpha (\theta - \theta_1)), \quad (3.4)$$

where  $\alpha$  is the thermal expansion coefficient of the fluid,  $\theta$  the absolute temperature, and  $\theta_1$  the reference temperature. Boussinesq's model for the force  $f$  of gravity consists in setting

$$f = f(\theta) := \rho(\theta) \vec{g}, \quad (3.5)$$

where  $\vec{g}$  is the fixed vector of gravity.

On the other hand, since the electrically conducting fluid is in presence of a magnetic field, it is subject to the Lorentz force  $j \times B$ , where  $j$  denotes the vector

of the current density and  $B$  the vector of the magnetic induction. Therefore, the resulting external force is given by

$$\mathbf{F} = f(\theta) + \mathbf{j} \times \mathbf{B}. \quad (3.6)$$

### 3.2. The model for global heat transfer

The main heat transfer phenomena in crystal growth from the melt are

Heat conduction;

Heat convection, in the melt and in the gas that fills the transparent cavity in the furnace;

Heat radiation in the transparent cavity.

There is a common agreement to consider that heat transport in the gas is dominated by radiation, so that the gas atmosphere can be considered as non participating, and the heat convection in the gas can be neglected in the model. This yields a useful simplification of the model, since otherwise one would need the full Navier-Stokes system to describe the gas motion.

The global temperature distribution in the furnace is thus governed by the heat equation

$$\rho c_V \left( \frac{\partial \theta}{\partial t} + \mathbf{v} \cdot \nabla \theta \right) = \operatorname{div}(\kappa(\theta) \nabla \theta) + f, \quad \text{in } ]0, T[ \times \Omega, \quad (3.7)$$

where  $\theta$  denotes the absolute temperature,  $\rho$  denotes the mass density,  $c_V$  is the specific heat at constant volume, and  $\kappa$  is the heat conductivity of the medium that may depend on temperature. Since we neglect the gas convection, we make the simplifying assumption that  $\mathbf{v} \neq 0$  only in the melt.

The heat sources  $f$  result, on the one hand, from the Joule effect in the electrical conductors. On the other hand, heat is produced by viscous friction in the fluid. Therefore

$$f = \frac{|\mathbf{j}|^2}{\mathfrak{s}(\theta)} + 2 \eta(\theta) D(\mathbf{v}, \mathbf{v}), \quad (3.8)$$

where  $\mathfrak{s}$  denotes the temperature dependent electrical conductivity,  $\mathbf{j}$  the density of electrical current supported in the conductors, and  $\mathbf{v}$  the velocity supported in the melt.

**Heat radiation.** Heat radiation is emitted at the surfaces of the opaque bodies that are located in the transparent cavity inside of the furnace. On this surface, the energy balance takes the form

$$\left[ -\kappa(\theta) \frac{\partial \theta}{\partial \vec{n}} \right] = R - J, \quad \text{on } ]0, T[ \times \Sigma, \quad (3.9)$$

where  $R$  is the radiosity (outgoing radiation) and  $J$  is the incoming radiation. The relation (3.9) means that the outgoing conductive heat flux has to balance the energy brought to the surface by radiation. On the other hand, a simple constitutive relation is given by

$$R = \epsilon \sigma \theta^4 + (1 - \epsilon) J, \quad \text{on } ]0, T[ \times \Sigma, \quad (3.10)$$

which means that the outgoing radiation is the sum of the radiation emitted according to the Stefan-Boltzmann law, and of the reflected radiation. In (3.10), the function  $\epsilon$ , that attains value in  $[0, 1]$ , is the emissivity of the body, and  $\sigma$  denotes the Stefan-Boltzmann constant. Following most modeling approaches for global heat transfer [KPS04], [Voi01] in crystal growth, we assume that all materials involved are *diffuse grey*. Accordingly, the material parameters *emissivity* and *reflexivity* depend neither on the angle of incidence nor on the wavelength. We in addition assume that all materials are opaque except for the gas in  $\Omega_0$ , and that no interaction takes place between gas and radiation.

One needs to obtain a second constitutive relation between  $R$  and  $J$ . For two arbitrary points  $z, y$  on the boundary  $\Sigma$  of the transparent cavity  $\Omega_0$  that can see each other, the part of the radiation outgoing at the point  $y$  that attains the point  $z$ , that we can denote  $j_y(z)$ , is given by the inverse square law

$$j_y(z) = \frac{\vec{n}(z) \cdot (y - z) \vec{n}(y) \cdot (z - y)}{\pi |y - z|^4} R(y),$$

where  $\vec{n}$  denotes a unit normal to  $\Sigma$ . We now want to obtain an expression for the total radiation  $J(z)$  incoming at point  $z$ . We introduce the so-called *view factor*, which for pairs of points  $(z, y) \in \Sigma \times \Sigma$  is given by

$$w(z, y) := \begin{cases} \frac{\vec{n}(z) \cdot (y - z) \vec{n}(y) \cdot (z - y)}{\pi |y - z|^4} \Theta(z, y) & \text{if } z \neq y, \\ 0 & \text{if } z = y, \end{cases} \quad (3.11)$$

where the visibility function  $\Theta$  penalizes the presence of opaque obstacles

$$\Theta(z, y) = \begin{cases} 1 & \text{if } ]z, y[ \subset \Omega_0, \\ 0 & \text{else.} \end{cases} \quad (3.12)$$

In (3.12),  $]z, y[$  is an abbreviation for  $\text{conv}(z, y) \setminus \{z, y\}$ . We obtain the total incoming radiation at  $z \in \Sigma$  by setting

$$J = K(R) \quad \text{on } ]0, T[\times\Sigma. \quad (3.13)$$

with the linear integral operator  $K$  defined by

$$(K(R))(z) := \int_{\Sigma} w(z, y) R(y) dS_y \quad \text{for } z \in \Sigma. \quad (3.14)$$

Introducing the linear operator

$$G := (I - K)(I - (1 - \epsilon)K)^{-1} \epsilon, \quad (3.15)$$

it can be shown from (3.10) and (3.13) that the boundary condition (3.9) finds an equivalent formulation in the relation

$$\left[ -\kappa(\theta) \frac{\partial \theta}{\partial \vec{n}} \right] = G(\sigma \theta^4) \text{ on } ]0, T[\times\Sigma, \quad (3.16)$$

where only the unknown  $\theta$  is involved.

### 3.3. The model for electromagnetics

In crystal growth *without* additional applied magnetic fields, a modeling of the electromagnetic inductive and/or resistive heating system is necessary to compute the heat sources (see [KPS04], [LKD07] and the references therein). However, it is often satisfactory to neglect the interaction of the fields generated in this way with the fluid motion. This is of course not anymore the case if such interaction is at the core of the investigation.

A fundamental difficulty for the computation of magnetic fields is the wide range of action. It is seldom realistic to assume that the applied magnetic field is confined to the region of interest for the computation of temperature, the furnace. However, since it is clearly neither necessary to consider extension of the electromagnetic fields to the entire space, we assume that they extend to some *bounded region* which may be *larger than the furnace*. This assumption is central in most numerical models (see [Bos04], Ch. 5 or [Mon03], 13.5).

The electric field  $E$  and the magnetic induction  $B$  satisfy Faraday's law of induction

$$\operatorname{curl} E + \frac{\partial B}{\partial t} = 0, \quad \text{in } ]0, T[ \times \tilde{\Omega}. \quad (3.17)$$

Magnetohydrodynamics, or low-frequency approximation of Maxwell's equations, means that Ampère's law

$$\operatorname{curl} H = j, \quad (3.18)$$

is assumed to be valid in  $]0, T[ \times \tilde{\Omega}$  for the vector  $H$  of the magnetic field strength and the current density  $j$ . In the electrical conductors, Ohm's law is valid in the form

$$j = \mathfrak{s}(\theta) (E + v \times B), \quad \text{in } ]0, T[ \times \tilde{\Omega}_c \quad (3.19)$$

where  $\mathfrak{s}$  denotes the electrical conductivity, and where  $v$  is supported in the melt. The magnetic induction  $B$  satisfies the so-called Gauss law

$$\operatorname{div} B = 0, \quad \text{in } ]0, T[ \times \tilde{\Omega}, \quad (3.20)$$

and the vector field  $D$  of electric displacement has to satisfy the conservation of charge in the nonconductors, that means, in the absence of free charges,

$$\operatorname{div} D = 0, \quad \text{in } ]0, T[ \times (\tilde{\Omega} \setminus \tilde{\Omega}_c). \quad (3.21)$$

We need a constitutive relation between  $B$  and  $H$ , as well as between  $E$  and  $D$ . We consider only linear media, that is

$$B = \mu H, \quad D = \epsilon E, \quad (3.22)$$

where  $\mu$  is the magnetic permeability, and  $\epsilon$  is the electrical permittivity.

At the same time, we have to model the presence of a current source in some parts of the conductors. We denote by  $\tilde{\Omega}_{c_0}$  the conductors where a current source is acting. Typically, these are magnetic coils that surround the furnace. In the case

of Figure 1, the coils are placed inside of the furnace. We discuss two possibilities for modeling the current sources.

**First model:** In the first model one considers that the current is *imposed* by the current source in  $\tilde{\Omega}_{c_0}$ . This model is the natural one if the applied current is a *direct current*, but is also widely used to approximate technical applications with alternating current (see [Bos04], Ch. 5), for example in the case that  $\tilde{\Omega}_{c_0}$  is an inductor that does not belong to the furnace. We have  $\operatorname{curl} H = j_g$  in  $\tilde{\Omega}_{c_0}$ , where  $j_g$  denotes the known density of the given current.

**Second model:** In the second model, one considers that the induction exerted by the system on the conductors  $\tilde{\Omega}_{c_0}$  is not negligible. Therefore, it is not possible to regard the current as being imposed therein.

Observe that from (3.17) and (3.20), it follows that

$$E = -\frac{\partial A}{\partial t} + \nabla \chi, \quad (3.23)$$

with a vector potential  $A$ , and a scalar potential  $\chi$ . The choice of these potential can be fixed with the help of diverse additional conditions called *gauge* (see [Jac99]).

In the second model, it is assumed that only the part  $\mathfrak{s} \nabla \chi$  of the current, originating from an applied voltage, can be considered as imposed. Therefore

$$\mathfrak{s} \nabla \chi \sim j_g \quad \text{in } ]0, T[ \times \tilde{\Omega}_{c_0}. \quad (3.24)$$

where  $j_g$  denotes the known density of the given current. It follows that (3.18) and (3.19) have to be written in the form

$$\operatorname{curl} H = \mathfrak{s}(\theta) \left( -\frac{\partial A}{\partial t} + v \times B \right) + j_g, \quad (3.25)$$

with  $j_g$  supported in  $\tilde{\Omega}_{c_0}$ . The potential  $A$  is related to  $B$  by the relation  $\operatorname{curl} A = B$ .

We make the consistency assumptions that

$$\operatorname{div} j_g = 0 \quad \text{in } ]0, T] \times \tilde{\Omega}_{c_0}, \quad j_g \cdot \vec{n} = 0 \quad \text{on } ]0, T[ \times \partial \tilde{\Omega}_{c_0}, \quad (3.26)$$

which express the conservation of charge in closed current loops.

In both models, we assume that at each time  $t$ , the density of an applied current  $j_g$  in the conductor  $\tilde{\Omega}_{c_0}$  is given in the form

$$j_g(t, x) = \sin(\omega t + \Phi) j_0(x) \quad \text{in } ]0, T[ \times \tilde{\Omega}_{c_0}. \quad (3.27)$$

The parameter  $\omega > 0$  and  $\Phi \in [0, 2\pi]$  are given ( $\omega$  = angular frequency of the imposed alternating current,  $\Phi$  = phase-shift).

### 3.4. Initial and boundary conditions

At the boundary of the melt, we assume that the velocity is imposed, that is

$$v = v_g \quad \text{on } ]0, T[ \times \partial \Omega_1, \quad (3.28)$$

and that  $v_g$  satisfies

$$v_g \cdot \vec{n} = 0 \quad \text{on } ]0, T] \times \partial\Omega_1, \quad (3.29)$$

since free boundaries are neglected. At time zero, we have

$$v(0) = v_0 \quad \text{in } \{0\} \times \Omega_1, \quad (3.30)$$

with the given velocity distribution  $v_0$ .

The heat radiation that occurs at the surface  $\Sigma$  is modeled by the boundary condition

$$\left[ -\kappa(\theta) \frac{\partial \theta}{\partial \vec{n}} \right] = R - J \quad \text{on } ]0, T[\times\Sigma,$$

which can be written in the equivalent form (3.16) as described in the paragraph 3.2. At the outer boundary  $]0, T[\times\Gamma$ , we consider the condition

$$\theta = \theta_g \quad \text{on } ]0, T[\times\Gamma. \quad (3.31)$$

On interfaces between opaque materials, the continuity of the heat flux is assumed. At time zero, we have

$$\theta(0) = \theta_0 \quad \text{in } \{0\} \times \Omega. \quad (3.32)$$

We supply the system (3.7) with the boundary conditions (3.9) and (3.31) and the initial condition (3.32).

The boundary conditions for the electromagnetic fields are the *natural interface conditions*

$$[H \times \vec{n}]_{i,j} = 0, \quad [B \cdot \vec{n}]_{i,j} = 0, \quad [E \times \vec{n}]_{i,j} = 0 \quad \text{on } ]0, T[\times(\partial\tilde{\Omega}_i \cap \partial\tilde{\Omega}_j), \quad (3.33)$$

where  $[\cdot]_{i,j}$  denotes the jump of a quantity across the surface  $]0, T[\times(\partial\tilde{\Omega}_i \cap \partial\tilde{\Omega}_j)$ ,  $i, j = 0, \dots, m$ ,  $i \neq j$ .

We consider that the outer boundary can be modeled as a magnetic shield and set

$$B \cdot \vec{n} = 0, \quad E \times \vec{n} = 0 \quad \text{on } ]0, T[\times\partial\tilde{\Omega}. \quad (3.34)$$

Therefore, we can also assume that the magnetic potential  $A$  is such that

$$A \times \vec{n} = 0 \quad \text{on } ]0, T[\times\partial\tilde{\Omega}. \quad (3.35)$$

Finally, we have the initial condition

$$H(0) = 0 \quad \text{in } \{0\} \times \tilde{\Omega}. \quad (3.36)$$

### 3.5. Quasi-stationary approach

Though a Czochralski crystal growth is essentially time-dependent, the long running times (order of days) are a crux for time-dependent simulations. In this section we present a set of hypotheses under which the model admits *quasi-stationary* so-

lutions that can basically be computed from an elliptic boundary value problem. For simplicity, we assume that

$$j_g(t, x) = \sin(\omega t) j_0(x) \quad \text{in } ]0, T[ \times \tilde{\Omega}_{c_0}, \quad (3.37)$$

that is, we set the phase-shift  $\Phi$  to zero in (3.37).

1. We assume that the applied alternating current  $j_g$  has a characteristic frequency  $\omega > 0$ , which is higher than the typical relaxation times for momentum and heat transfer. At the time-scale of the electromagnetic evolution, for example the interval  $]t, t + 2\pi/\omega[$ , the quantities  $v$ ,  $p$  and the temperature  $\theta$  can be assumed to be stationary.
2. We assume that we are far from the beginning of the evolution, and that the electromagnetic fields are now independent of the initial conditions. We in addition assume that the electromagnetic quantities have reached a *time-harmonic regime*.

Due to the first hypothesis, we can average the equations (3.1) and (3.7) over the interval  $]t, t + 2\pi/\omega[$ , and obtain that

$$\rho_1 (v \cdot \nabla) v = -\nabla p + \operatorname{div}(2\eta(\theta) D v) + f(\theta) + [j \times B]_{\text{av}}, \quad (3.38)$$

and that

$$\rho_1 c_V v \cdot \nabla \theta = \operatorname{div}(\kappa(\theta) \nabla \theta) + 2\eta(\theta) D(v, v) + [\frac{|j|^2}{\mathfrak{s}(\theta)}]_{\text{av}}, \quad (3.39)$$

where

$$[F]_{\text{av}} := \frac{\omega}{2\pi} \int_t^{t+2\pi/\omega} F(s) ds.$$

Due to the second hypothesis, we have

$$j(t, x) = \operatorname{Im}(\tilde{j}(x) \exp(i\omega t)), \quad H(t, x) = \operatorname{Im}(\tilde{H}(x) \exp(i\omega t)),$$

and so on for the other electromagnetic quantities, where  $\tilde{j}$ ,  $\tilde{H}$  are complex valued vector fields, called the *amplitudes* of the time-harmonic fields  $j$ ,  $H$ . It follows that

$$\begin{aligned} [j \times B]_{\text{av}} &= 1/2 (\operatorname{Re}(\tilde{j}) \times \operatorname{Re}(\tilde{B}) + \operatorname{Im}(\tilde{j}) \times \operatorname{Im}(\tilde{B})), \\ [\frac{|j|^2}{\mathfrak{s}}]_{\text{av}} &= \frac{|\operatorname{Re}(\tilde{j})|^2 + |\operatorname{Im}(\tilde{j})|^2}{2\mathfrak{s}(\theta)}. \end{aligned} \quad (3.40)$$

In the time-harmonic setting, the relation (3.17) yields

$$\operatorname{curl} \tilde{E} + i\omega \tilde{B} = 0,$$

whereas the other Maxwell's relations (3.18), (3.19), (3.20), (3.21) and (3.22) are valid for the fields  $\tilde{j}$ ,  $\tilde{H}$ ,  $\tilde{E}$ .

If the boundary data  $v_g$  and  $\theta_g$  are stationary, we can solve these equations in connection to (3.38), (3.39) and to the boundary conditions (3.28), (3.16), (3.31), (3.33), (3.34). We denote this well-posed<sup>3</sup> elliptic boundary value problem by  $(P)$ .

---

<sup>3</sup>See for comparison the model in [RT92]

#### 4. Functional setting. Weak solutions

For the electromagnetic part of the problem, spaces of vector fields with generalized curl and  $\operatorname{div}$  are needed. We first introduce

$$L^2_{\operatorname{curl}}(\tilde{\Omega}; \mathbb{C}^3) := \left\{ H \in [L^2(\tilde{\Omega}; \mathbb{C})]^3 \mid \operatorname{curl} H \in [L^2(\tilde{\Omega}; \mathbb{C})]^3 \right\},$$

where the differential operator  $\operatorname{curl}$  is intended in its generalized sense and applied componentwise to real and imaginary part of a complex-valued vector field. It is well known that  $L^2_{\operatorname{curl}}(\tilde{\Omega}; \mathbb{C}^3)$  is a Hilbert space with respect to the product

$$(H_1, H_2)_{L^2_{\operatorname{curl}}(\tilde{\Omega}; \mathbb{C}^3)} := \int_{\tilde{\Omega}} (\operatorname{curl} H_1 \cdot \operatorname{curl} \overline{H_2} + H_1 \cdot \overline{H_2}),$$

where for  $a \in \mathbb{C}$ , we denote by  $\overline{a}$  the complex number conjugated to  $a$ . A natural context in which to search for the field  $H$  is then the space

$$\mathcal{H}(\tilde{\Omega}) := \left\{ H \in L^2_{\operatorname{curl}}(\tilde{\Omega}; \mathbb{C}^3) \mid \operatorname{curl} H = 0 \text{ in } \tilde{\Omega} \setminus \tilde{\Omega}_c \right\}. \quad (4.1)$$

Obviously, this is a closed linear subspace of  $L^2_{\operatorname{curl}}(\tilde{\Omega}; \mathbb{C}^3)$ .

The appropriate setting for the Navier-Stokes equations is widely known. We need the spaces of real-valued vector fields

$$\begin{aligned} D^{1,2}(\Omega_1) &:= \left\{ u \in [W^{1,2}(\Omega_1)]^3 \mid \operatorname{div} u = 0 \text{ in } \Omega_1 \right\}, \\ D_0^{1,2}(\Omega_1) &:= \left\{ u \in [W_0^{1,2}(\Omega_1)]^3 \mid \operatorname{div} u = 0 \text{ in } \Omega_1 \right\}. \end{aligned} \quad (4.2)$$

For the mathematical setting of the stationary heat equation with radiation boundary condition, we need spaces of functions whose traces are integrable to a higher exponent than the one given by Sobolev's embedding relations. These are the spaces

$$V^{p,q}(\Omega) := \left\{ \theta \in W^{1,p}(\Omega) \mid \gamma(\theta) \in L^q(\Sigma) \right\}, \quad 1 \leq p \leq \infty, 4 \leq q \leq \infty, \quad (4.3)$$

where  $\gamma$  denotes the trace operator. The subscript  $\Gamma$  will indicate the subspace consisting of all functions whose trace vanishes on the boundary part  $\Gamma$ .

With these preliminaries, we can define

**Definition 4.1.** A *weak solution* to the problem  $(P)$  introduced in Section 3.5 is a triple

$$\{v, H, \theta\} \in D^{1,2}(\Omega_1) \times \mathcal{H}(\tilde{\Omega}) \times \bigcap_{1 \leq p < 3/2} V^{p,4}(\Omega),$$

such that  $v = v_g$  on  $\partial\Omega_1$ ,  $\theta = \theta_g$  on  $\Gamma$  and the integral relations

$$\int_{\Omega_1} \rho_1 (v \cdot \nabla) v \cdot \phi + \int_{\Omega_1} \eta(\theta) D(v, \phi) = \int_{\Omega_1} [\operatorname{curl} H \times \mu H]_{\text{av}} \cdot \phi + \int_{\Omega_1} f(\theta) \cdot \phi, \quad (4.4)$$

$$i \int_{\tilde{\Omega}} \mu \omega H \cdot \bar{\psi} + \int_{\tilde{\Omega}} r(\theta) \operatorname{curl} H \cdot \operatorname{curl} \bar{\psi} = \int_{\Omega_1} (v \times \mu H) \cdot \operatorname{curl} \bar{\psi} + \int_{\tilde{\Omega}_{c_0}} r j_g \cdot \operatorname{curl} \bar{\psi}, \quad (4.5)$$

$$\begin{aligned} \int_{\Omega_1} \rho_1 c_V v \cdot \nabla \theta \xi + \int_{\Omega} \kappa(\theta) \nabla \theta \cdot \nabla \xi + \int_{\Sigma} G(\sigma \theta^4) \xi \\ = \int_{\Omega} \left( [r(\theta) |\operatorname{curl} H|^2]_{\text{av}} + \eta(\theta) D(v, v) \chi_{\Omega_1} \right) \xi, \end{aligned} \quad (4.6)$$

are satisfied for all  $\{\phi, \psi, \xi\} \in D_0^{1,2}(\Omega) \times \mathcal{H}(\tilde{\Omega}) \times W_{\Gamma}^{1,q}(\Omega)$  with  $q > 3$ . Here,  $r = \mathfrak{s}^{-1}$  = electrical resistivity.

**Existence result.** Finally, we present a set of assumptions that allow to prove the existence of a weak solution in the sense of Definition 4.1. The full proof cannot be presented here: its essential steps have been carried out in [Dru08] (nonlocal radiation with integrable right-hand side) and [Dru07] (higher integrability of the Lorentz force).

**Theorem 4.2 (Main Theorem).** *Assume that there exist positive constants  $\mathfrak{s}_l, \mathfrak{s}_u, \mu_l, \mu_u, \kappa_l, \kappa_u, \eta_l, \eta_u$  such that*

$$\begin{aligned} 0 < \mathfrak{s}_l \leq \mathfrak{s} \leq \mathfrak{s}_u < +\infty, \quad 0 < \mu_l \leq \mu \leq \mu_u < +\infty, \\ 0 < \kappa_l \leq \kappa \leq \kappa_u < +\infty, \quad 0 < \eta_l \leq \eta \leq \eta_u < +\infty. \end{aligned} \quad (4.7)$$

*Assume that the emissivity on the surface  $\Sigma$ , denoted by  $\epsilon$ , is a measurable function of the position, and that there exists a positive number  $\epsilon_l$  such that*

$$0 < \epsilon_l \leq \epsilon < 1 \quad \text{on } \Sigma. \quad (4.8)$$

*The remaining coefficients are assumed to be positive constants. We require that*

$$\mathfrak{s}_i, \kappa_i, \eta \in C(\mathbb{R}), \quad \mu_i \in \mathcal{C}(\overline{\tilde{\Omega}_i}) \quad \text{for } i = 0, \dots, m, \quad (4.9)$$

*where  $\kappa_i, \mathfrak{s}_i, \mu_i$  denote the restriction of  $\kappa, \mathfrak{s}, \mu$  to  $\tilde{\Omega}_i$ .*

*Assume that the geometry satisfies*

$$\partial\tilde{\Omega}_i, \partial\tilde{\Omega} \in \mathcal{C}^1 \text{ for } i = 0, \dots, m, \quad (4.10)$$

*and that the heterogeneous conducting materials  $\tilde{\Omega}_{k_0} \subseteq \tilde{\Omega}_c$  ( $k_0 \in \{1, \dots, m\}$ ) are separated from each other and from the outer boundary  $\partial\tilde{\Omega}$  by nonconducting material or vacuum.*

*We furthermore require that  $\Sigma \in \mathcal{C}^{1,\delta}$  for some  $\delta > 0$ .*

Assume that the force term  $f(\theta)$  in the Navier-Stokes equations is either globally bounded (truncated), or that the thermal expansion coefficient  $\alpha$  of the fluid is sufficiently small. Assume that the given current  $j_g$  is given by (3.27), satisfies (3.26) and that  $j_0 \in [L^2(\tilde{\Omega}_{c_0})]^3$ . Finally, assume that  $v_g \in D^{1,2}(\Omega_1) \cap L^\infty(\Omega_1)$  satisfies (3.29) and that the number  $\|v_g\|_{[L^\infty(\Omega_1)]^3}$  is sufficiently small. Assume that the imposed temperature  $\theta_g$  belongs to  $W^{1,2}(\Omega) \cap L^\infty(\Omega)$ .

Then, there exists a weak solution to (P) in the sense of Definition 4.1.

## References

- [Bos04] A. Bossavit. *Electromagnétisme en vue de la modélisation*. Springer, Berlin, Heidelberg, New York, 2004.
- [Cap72] F. Cap. *Einführung in die Plasmaphysik I, II, III*. Akademie Verlag, Berlin, 1972.
- [Cha81] S. Chandrasekhar. *Hydrodynamic and hydromagnetic stability*. Dover Publications Inc., New York, 1981.
- [Dru07] P.-E. Druet. Higher integrability of the lorentz force for weak solutions to Maxwell's equations in complex geometries. Preprint 1270 of the Weierstrass Institute for Applied Mathematics and Stochastics, Berlin, 2007. Available in pdf-format at <http://www.wias-berlin.de/publications/preprints/1270>.
- [Dru08] P.-E. Druet. Weak solutions to a stationary heat equation with nonlocal radiation boundary condition and right-hand side in  $L^p$  ( $p \geq 1$ ). To appear in Math. Meth. Appl. Sci., 2008. <http://dx.doi.org/10.1002/mma.1029>.
- [GG76] Donald D. Gray and A. Giorgini. The validity of the Boussinesq approximation for liquids and gases. *Int. J. Heat Mass Transfer*, 19:545–551, 1976.
- [GM06] R. Griesse and A.J. Meir. Modeling of a MHD free surface problem arising in Cz crystal growth. In *Proceedings of the 5th IMACS Symposium on Mathematical Modelling (5th MATHMOD)*, Vienna, 2006.
- [HZ07] M. Hinze and S. Ziegenbalg. Optimal control of the free boundary in a two-phase Stefan problem with flow driven by convection. *Z. Angew. Math. Mech.*, 87:430–448, 2007.
- [Jac99] J.D. Jackson. *Classical Electrodynamics*. John Wiley and Sons, Inc., third edition, 1999.
- [KPS04] O. Klein, P. Philip, and J. Sprekels. Modelling and simulation of sublimation growth in sic bulk single crystals. *Interfaces and Free Boundaries*, 6(1):295–314, 2004.
- [LKD07] C. Lechner, O. Klein, and P.-E. Druet. Development of a software for the numerical simulation of VCz growth under the influence of a traveling magnetic field. *Journal of Crystal Growth*, 303:161–164, 2007.
- [Mey06] C. Meyer. *Optimal Control of Semilinear Elliptic Equations with Applications to Sublimation Crystal Growth*. PhD thesis, Technische Universität Berlin, Germany, 2006.
- [Mon03] P. Monk. *Finite element methods for Maxwell's equations*. Clarendon press, Oxford, 2003.

- [MPT06] C. Meyer, P. Philip, and F. Tröltzsch. Optimal control of a semilinear pde with nonlocal radiation interface conditions. *SIAM Journal On Control and Optimization (SICON)*, 45:699–721, 2006.
- [Phi03] P. Philip. *Transient Numerical Simulation of Sublimation Growth of SiC Bulk Single Crystals. Modeling, Finite Volume Method, Results*. PhD thesis, Department of Mathematics, Humboldt University of Berlin, Germany, 2003. Report No. 22, Weierstrass Institute for Applied Analysis and Stochastics (WIAS), Berlin.
- [RT92] J. Rappaz and R. Touzani. On a two-dimensional magnetohydrodynamic problem, i. modeling and analysis. *RAIRO Modél. Math. Anal., Num.*, 26:347–364, 1992.
- [Rud07] P. Rudolph. Travelling magnetic fields applied to bulk crystal growth from the melt: The step from basic research to industrial scale. To appear in *J. Crystal Growth*, 2007. <http://dx.doi.org/10.1016/j.jcrysGro.2007.11.036>.
- [Tii97] T. Tiihonen. Stefan-Boltzmann radiation on non-convex surfaces. *Math. Meth. in Appl. Sci.*, 20(1):47–57, 1997.
- [Voi01] A. Voigt. *Numerical Simulation of Industrial Crystal Growth*. PhD thesis, Technische Universität München, Germany, 2001.

Pierre-Étienne Druet

Weierstrass Institute for Applied Analysis and Stochastics

Mohrenstrasse 39

D-10117 Berlin, Germany

e-mail: [druet@wias-berlin.de](mailto:druet@wias-berlin.de)

# Lavrentiev Prox-regularization Methods for Optimal Control Problems with Pointwise State Constraints

Martin Gugat

**Abstract.** A Lavrentiev prox-regularization method for optimal control problems with pointwise state constraints is introduced. The convergence of the controls generated by the iterative Lavrentiev prox-regularization algorithm is studied. For a sequence of regularization parameters that converges to zero, strong convergence of the generated control sequence to the optimal control is proved. Due to the prox-character of the proposed regularization, the feasibility of the iterates for a given parameter can be improved compared with the non-prox Lavrentiev regularization.

**Mathematics Subject Classification (2000).** 49J20, 90C31.

**Keywords.** Optimal control, pointwise state constraints, prox-regularization, Lavrentiev regularization, pde constrained optimization, convergence.

## 1. Introduction

In optimal control problems, pointwise state constraints are important since often in the applications, certain restrictions on the state occur. In the optimality systems corresponding to such systems, multipliers appear, which in general can only be represented as measures. However, for the numerical solution of such problems, it is desirable to have multipliers that are as regular as possible. In order to obtain regular multipliers that can be represented by  $L^2$ -functions, the Lavrentiev regularization is introduced, which is studied for example in [5, 3, 7, 8, 6] and in the references cited there. We do not claim to give a complete list of references about this subject here. Due to the regularization, for the auxiliary problems multipliers with  $L^2$ -regularity exist, see [9].

In the Lavrentiev regularization the Lavrentiev regularization parameter  $\lambda$  must converge to  $0+$  to obtain convergence. However, as  $\lambda$  decreases the problems

become more and more difficult to solve. For each fixed  $\lambda > 0$  in general the generated controls are infeasible for the original problem. In this paper we introduce a Lavrentiev prox-regularization method where for a given parameter value  $\lambda$ , the feasibility can be improved (see Section 3.5). In Section 3.4 we show that for a sequence of regularization parameters converging to zero, the new algorithm generates a sequence of controls that converges with respect to the  $L^2$ -norm to the optimal control.

We start by considering the elliptic optimal control problem with pointwise state constraints and without pointwise control constraints (Section 2) and the corresponding Lavrentiev prox-regularization (Section 3). Then we turn to the elliptic optimal control problem with pointwise state and control constraints (Section 4) and the Lavrentiev prox-regularization (Section 5) for this problem.

At the end of the paper we present examples where we compare the convergence of the Lavrentiev prox-regularization method with the non-prox Lavrentiev regularization.

## 2. The elliptic problem without pointwise control constraints

In this section we introduce an elliptic optimal control problem with state constraints.

Let  $N \in \{2, 3\}$  and  $\Omega \subset R^N$  be a bounded domain with  $C^{0,1}$  boundary  $\Gamma$ . Let a desired state  $y_d \in L^\infty(\Omega)$  be given. Let a real number  $\kappa > 0$  be given. Define the objective function

$$J(y, u) = \int_{\Omega} (y - y_d)^2 + \kappa u^2 dx.$$

Let state bounds  $y_a, y_b \in C(\bar{\Omega})$  be given such that  $y_a < y_b$  in  $\bar{\Omega}$  that is the Slater condition holds.

Define the following elliptic optimal control problem with distributed control and pointwise state constraints

$$\mathbf{P} : \left\{ \begin{array}{l} \text{minimize } J(y, u) \text{ subject to} \\ \partial_\nu y = 0 \text{ in } \Gamma \\ Ay = u \text{ in } \Omega \\ y_a \leq y \leq y_b \text{ in } \Omega. \end{array} \right. \quad (1)$$

Here  $\partial_\nu$  denotes the normal derivative and  $A$  is an elliptic differential operator as in [1]. The elliptic control Problem  $\mathbf{P}$  has also been considered in [3], where a complete list of references can be found. As in [3], the notation  $G$  is used for the control to state map that gives the state as a function of the control,  $G : L^2(\Omega) \rightarrow H^1(\Omega) \cap L^\infty(\Omega)$ . The notation  $S$  is used for the control to state map as an operator  $L^2(\Omega) \rightarrow L^2(\Omega)$  which is the composition of  $G$  and the suitable embedding operator.

### 3. Lavrentiev prox-regularization

Let the Lavrentiev regularization parameter  $\lambda > 0$  and  $v \in L^\infty(\Omega)$  be given. Define  $K(u) = J(G(u), u)$ . We consider the regularized problem

$$\mathbf{P}_{\lambda,v} : \begin{cases} \text{minimize } K(u) \text{ subject to} \\ y_a \leq \lambda(u - v) + G(u) \leq y_b \text{ in } \Omega. \end{cases} \quad (2)$$

Concerning the regularity of the multipliers corresponding to the inequality constraints in  $\mathbf{P}_{\lambda,v}$ , we can apply Theorem 2.1 in [3] that states that we find multipliers in  $L^2(\Omega)$ .

We consider the following **Lavrentiev prox-regularization algorithm**:

- Start:** CHOOSE  $u_1 \in L^\infty(\Omega)$  AND  $\lambda_1 > 0$ .
- Step  $k$ :** GIVEN  $u_k \in L^\infty(\Omega)$  AND  $\lambda_k > 0$ , SOLVE  $\mathbf{P}_{\lambda_k, u_k}$ .  
DEFINE  $u_{k+1}$  AS THE SOLUTION OF  $\mathbf{P}_{\lambda_k, u_k}$ . CHOOSE  $\lambda_{k+1} \in (0, \lambda_k]$ .  
GO TO STEP  $k + 1$ .

This is a prox-regularization that has some similarities to the prox-regularization considered in [2]. Note however, that in our case the regularization term appears in the constraint and not in the objective function. In our discussion we use the choice  $u_1 = 0$ . First we show that the iteration is well defined. For  $u_1 = 0$ , problem  $\mathbf{P}_{\lambda_1, u_1}$  is of the form studied in the papers about Lavrentiev regularization [7], [5], [4], hence the corresponding existence results are applicable.

The classical non-prox Lavrentiev regularization corresponds to the definition  $u_{k+1} = 0$  for all  $k$ .

In step  $k$ , the function  $u_{k+1}$  satisfies the state constraint

$$y_a \leq \lambda_k(u_{k+1} - u_k) + G(u_{k+1}) \leq y_b \text{ in } \Omega.$$

Hence  $u_{k+1} \in L^\infty(\Omega)$  and the function

$$\tilde{u}_{k+1} = u_{k+1} + (\lambda_{k+1}I + S)^{-1}\lambda_k(u_{k+1} - u_k)$$

is feasible for  $\mathbf{P}_{\lambda_{k+1}, u_{k+1}}$ . Therefore the iteration is well defined.

#### 3.1. Properties of $\mathbf{P}_{\lambda, u_*}$

Let  $\omega_*$  denote the optimal value of  $\mathbf{P}$ , and  $\omega(\lambda, v)$  denote the optimal value of  $\mathbf{P}_{\lambda,v}$ . Let  $u_*$  be the solution of  $\mathbf{P}$ . Let  $F_*$  denote the admissible set of  $\mathbf{P}$  and  $F(\lambda, v)$  denote the admissible set of  $\mathbf{P}_{\lambda,v}$ . We use the notation  $\|\cdot\| = \|\cdot\|_{L^2(\Omega)}$ .

In the following lemma we given an upper bound for  $\omega(\lambda, u_*)$ .

**Lemma 1.** *We have*

$$\omega(\lambda, u_*) \leq \omega_*. \quad (3)$$

Let  $v_\lambda \in F(\lambda, u_*)$  be a solution of  $\mathbf{P}_{\lambda, u_*}$ , that is  $K(v_\lambda) = \omega(\lambda, u_*)$ . Then

$$\|v_\lambda\| \leq \sqrt{\omega_*}/\sqrt{\kappa} \quad (4)$$

and  $v_\lambda - u_* \in L^\infty(\Omega)$ .

*Proof.* We have  $u_* \in F(\lambda, u_*)$ , thus

$$\omega(\lambda, u_*) \leq K(u_*) = \omega_*.$$

Hence (3) follows. Therefore we have

$$\kappa \|v_\lambda\|^2 \leq K(v_\lambda) = \omega(\lambda, u_*) \leq \omega_*$$

which implies (4). The inequality constraint implies

$$\frac{y_a - G(v_\lambda)}{\lambda} \leq v_\lambda - u_* \leq \frac{y_b - G(v_\lambda)}{\lambda}$$

almost everywhere on  $\Omega$ , hence  $v_\lambda - u_* \in L^\infty(\Omega)$ .  $\square$

In the next lemma we show that  $\omega(\cdot, u_*)$  is continuous at zero.

**Lemma 2.** *Let  $v_\lambda \in F(\lambda, u_*)$  be a solution of  $\mathbf{P}_{\lambda, u_*}$ . Then  $\lim_{\lambda \rightarrow 0+} \|v_\lambda - u_*\| = 0$ . In particular, this implies  $\lim_{\lambda \rightarrow 0+} \omega(\lambda, u_*) = \omega_*$ .*

*Proof.* Lemma 1 implies that the norms  $\|v_\lambda\|$  are uniformly bounded. Moreover we have the inequality  $\limsup_{\lambda \rightarrow 0+} \omega(\lambda, u_*) \leq \omega_*$ . Now we consider a sequence  $\lambda_k \rightarrow 0+$ . Let  $\tilde{v}$  denote a weak limit point of the sequence  $v_{\lambda_k}$ . Then  $\tilde{v} \in F_*$  (see Lemma 3.2 in [3]). Moreover

$$K(\tilde{v}) \leq \liminf_{k \rightarrow \infty} \omega(\lambda_k, u_*) \leq \omega_*.$$

Since  $\tilde{v} \in F_*$ , this implies that  $K(\tilde{v}) = \omega_*$ . Thus the uniqueness of the solution of  $\mathbf{P}$  implies  $\tilde{v} = u_*$ . Hence the sequence  $(v_{\lambda_k})_k$  converges weakly to  $u_*$ . Hence  $\lim_{k \rightarrow \infty} \omega(\lambda_k, u_*) = \omega_*$ . Since the sequence  $(\lambda_k)_k$  was chosen arbitrarily, this implies  $\lim_{\lambda \rightarrow 0+} \omega(\lambda, u_*) = \omega_*$ . Therefore

$$\begin{aligned} \lim_{\lambda \rightarrow 0+} \kappa \|v_\lambda\|^2 &= \lim_{\lambda \rightarrow 0+} K(v_\lambda) - \|S(v_\lambda) - y_d\|^2 \\ &= K(u_*) - \|S(u_*) - y_d\|^2 \\ &= \kappa \|u_*\|^2. \end{aligned}$$

The weak convergence of  $v_\lambda$  to  $u_*$  and the convergence of the norms imply  $\lim_{\lambda \rightarrow 0+} \|v_\lambda - u_*\| = 0$ .  $\square$

The following lemma states that  $(\|\lambda(\lambda I + S)^{-1}\|)_{\lambda > 0}$  is uniformly bounded (see [3]).

**Lemma 3.** *Let  $\|S\|$  denote the operator norm of  $S$  as a map from  $L^2(\Omega)$  to  $L^2(\Omega)$ . For all  $\lambda > 0$  we have the inequality*

$$\|(\lambda I + S)^{-1}\| \leq \frac{1}{\lambda}. \tag{5}$$

*Let  $u \in L^2(\Omega)$  and let  $\lambda_k > 0$  with  $\lim_{k \rightarrow \infty} \lambda_k = 0$ . Then*

$$\lim_{k \rightarrow \infty} \lambda_k \|(\lambda_k I + S)^{-1} u\| = 0.$$

*Proof.* As in inequality (3.3) in [3] we see that for all  $u \in L^2(\Omega)$  we have  $\|\lambda(\lambda I + S)^{-1}u\| \leq \|u\|$ , hence  $\|\lambda(\lambda I + S)^{-1}\| \leq 1$  and assertion follows. For the convenience of the reader we repeat the argument here. Assume that  $u \neq 0$ . Let  $(v_i)_{i=1}^\infty$  denote the set of eigenvectors of  $S$  that forms an orthonormal basis of  $L^2(\Omega)$ . Let  $\mu_i > 0$  denote the corresponding eigenvalues of  $S$ . Then

$$\lambda(\lambda I + S)^{-1}u = \sum_{i=1}^\infty \frac{\lambda}{\lambda + \mu_i} \int_\Omega uv_i dx v_i$$

hence

$$\begin{aligned} \|\lambda(\lambda I + S)^{-1}u\|^2 &= \sum_{i=1}^\infty \frac{\lambda^2}{(\lambda + \mu_i)^2} \left( \int_\Omega uv_i dx \right)^2 \\ &< \sum_{i=1}^\infty \left( \int_\Omega uv_i dx \right)^2 \\ &= \|u\|^2. \end{aligned}$$

This implies

$$\lim_{k \rightarrow \infty} \sum_{i=1}^\infty \frac{\lambda_k^2}{(\lambda_k + \mu_i)^2} \left( \int_\Omega uv_i dx \right)^2 = \sum_{i=1}^\infty \lim_{k \rightarrow \infty} \frac{\lambda_k^2}{(\lambda_k + \mu_i)^2} \left( \int_\Omega uv_i dx \right)^2 = 0, \quad (6)$$

where we can interchange the summation and the limit since the first series in (6) represents a function series with continuous functions that is dominated and therefore converges uniformly.  $\square$

In the next lemma we give a Lipschitz condition for  $\omega(\cdot, u_*)$  for  $\lambda > 0$ .

**Lemma 4.** *Let  $L > 0$ ,  $M_4 \geq 1$  denote real numbers such that for all  $z \in L^2(\Omega)$  with  $\|z\| \leq \sqrt{\frac{\omega_*}{\kappa}}$  and all  $\delta \in L^2(\Omega)$  with*

$$\|\delta\| \leq M_4 \left( \sqrt{\frac{\omega_*}{\kappa}} + \|u_*\| \right)$$

*we have the Lipschitz inequality*

$$K(z + \delta) \leq K(z) + L\|\delta\|.$$

*Then for all  $\lambda_1 > 0$ ,  $\lambda_2 > 0$  with  $\lambda_1/\lambda_2 \leq M_4$  and  $\lambda_2/\lambda_1 \leq M_4$*

$$|\omega(\lambda_2, u_*) - \omega(\lambda_1, u_*)| \leq |\lambda_2 - \lambda_1| L \max \left\{ \frac{1}{\lambda_2}, \frac{1}{\lambda_1} \right\} \left( \|u_*\| + \frac{\sqrt{\omega_*}}{\sqrt{\kappa}} \right). \quad (7)$$

*Proof.* Let  $v_1$  denote the solution of  $\mathbf{P}_{\lambda_1, u_*}$  and let  $v_2$  denote the solution of  $\mathbf{P}_{\lambda_2, u_*}$ . Define the function

$$\tilde{v}_1 = v_1 + (S + \lambda_2 I)^{-1}(\lambda_1 - \lambda_2)(v_1 - u_*) \in F(\lambda_2, u_*).$$

Then due to (5) we have

$$\begin{aligned} \|(S + \lambda_2 I)^{-1}(\lambda_1 - \lambda_2)(v_1 - u_*)\| &\leq \frac{|\lambda_1 - \lambda_2|}{\lambda_2} \|v_1 - u_*\| \\ &\leq \left| \frac{\lambda_1}{\lambda_2} - 1 \right| \left( \|u_*\| + \frac{\sqrt{\omega_*}}{\sqrt{\kappa}} \right) \\ &\leq M_4 \left( \|u_*\| + \frac{\sqrt{\omega_*}}{\sqrt{\kappa}} \right) \end{aligned}$$

and we obtain the inequality

$$\begin{aligned} \omega(\lambda_2, u_*) &\leq K(\tilde{v}_1) \\ &= K(v_1 + (S + \lambda_2 I)^{-1}(\lambda_1 - \lambda_2)(v_1 - u_*)) \\ &\leq K(v_1) + L |\lambda_1 - \lambda_2| \frac{1}{\lambda_2} \|v_1 - u_*\| \\ &= \omega(\lambda_1, u_*) + \frac{L}{\lambda_2} |\lambda_1 - \lambda_2| \left( \|u_*\| + \frac{\sqrt{\omega_*}}{\sqrt{\kappa}} \right) \end{aligned}$$

and analogously

$$\omega(\lambda_1, u_*) \leq \omega(\lambda_2, u_*) + \frac{L}{\lambda_1} |\lambda_1 - \lambda_2| \left( \|u_*\| + \frac{\sqrt{\omega_*}}{\sqrt{\kappa}} \right)$$

and the assertion follows.  $\square$

### 3.2. Uniform boundedness of the solutions of $\mathbf{P}_{\lambda, u}$

In Section 3.1 we have seen that the solutions of  $\mathbf{P}_{\lambda, u_*}$  are uniformly bounded, see (4). In this section we consider the more general problem  $\mathbf{P}_{\lambda, u}$  and give an upper bound for the  $L^2$ -norm of its solution.

**Lemma 5.** *Let  $M_5 > 0$  be given. Let  $u \in L^2(\Omega)$  be given such that  $\|u - u_*\| \leq M_6$ . Let  $\lambda > 0$  and  $u_{p,\lambda}$  denote the solution of  $\mathbf{P}_{\lambda, u}$ . Then the following inequality holds:*

$$\|u_{p,\lambda}\| \leq \hat{C}(M_6)/\kappa,$$

where  $\hat{C}(M_6) = \sup_{v: \|v - u_*\| \leq M_6} K(v)$ . Moreover we have

$$\lim_{\lambda \rightarrow 0+} \|u_{p,\lambda} - u_*\| = 0, \quad \lim_{\lambda \rightarrow 0+} \omega(\lambda, u) = \omega_*.$$

*Proof.* Define the function

$$\begin{aligned} \tilde{u}_* &= (S + \lambda I)^{-1}(S(u_*) + \lambda u) \\ &= u_* + \lambda(S + \lambda I)^{-1}(u - u_*) \in F(\lambda, u). \end{aligned}$$

Then

$$\begin{aligned} \kappa \|u_{p,\lambda}\|^2 &\leq K(u_{p,\lambda}) \\ &= \omega(\lambda, u) \\ &\leq K(\tilde{u}_*) = K(u_* + \lambda(S + \lambda I)^{-1}(u - u_*)) \\ &\leq \sup_{w: \|w - u_*\| \leq M_6} K(u_* + \lambda(S + \lambda I)^{-1}(w - u_*)). \end{aligned}$$

Due to (5) we have  $\|\lambda(S + \lambda I)^{-1}(w - u_*)\| \leq \|w - u_*\| \leq M_6$ , and the uniform boundedness follows. Since  $\lim_{\lambda \rightarrow 0+} \|\lambda(S + \lambda I)^{-1}(u - u_*)\| = 0$  (see the proof of Lemma 3.1 in [3]) the above inequality implies  $\limsup_{\lambda \rightarrow 0+} \omega(\lambda, u) \leq K(u_*) = \omega_*$  and as in the proof of Lemma 2 the last part of the assertion follows.  $\square$

### 3.3. Boundedness of the generated sequence

The iteration of the Lavrentiev prox-regularization method generates a bounded sequence if the regularization parameters are chosen sufficiently small. This can be seen as follows:

**Lemma 6.** *Assume that there exists a slater control  $\bar{\eta} \in L^\infty(\Omega)$  and  $\bar{\epsilon} > 0$  such that*

$$y_a + \bar{\epsilon} \leq G(\bar{\eta}) \leq y_b - \bar{\epsilon}$$

*on  $\bar{\Omega}$ . Assume that in each step,  $\lambda_k$  is chosen such that*

$$\lambda_k \|\bar{\eta} - u_k\|_{L^\infty(\Omega)} \leq \bar{\epsilon}. \quad (8)$$

*Then the sequence  $(u_k)_k$  generated by the Lavrentiev prox-regularization algorithm is bounded.*

**Remark 1.** *Note that condition (8) can easily be satisfied during the iteration by choosing  $\lambda_k$  sufficiently small since the functions  $\eta$  and  $u_k$  are known.*

*Proof.* For all  $k$  we have the inequalities

$$\begin{aligned} G(\bar{\eta}) + \lambda_k(\bar{\eta} - u_k) &\leq y_b - \bar{\epsilon} + \bar{\epsilon} = y_b, \\ G(\bar{\eta}) + \lambda_k(\bar{\eta} - u_k) &\geq y_a + \bar{\epsilon} - \bar{\epsilon} = y_a \end{aligned}$$

hence  $\bar{\eta} \in F(\lambda_k, u_k)$  which implies

$$\kappa \|u_k\|^2 \leq K(u_k) \leq K(\bar{\eta})$$

and the assertion follows.  $\square$

### 3.4. Convergence of the generated sequence

We study now the convergence of the solutions  $(u_k)_k$  for  $k \rightarrow \infty$ .

**Lemma 7.** *Assume that  $u_* \in L^\infty(\Omega)$  and that there exists a slater control  $\bar{\eta} \in L^\infty(\Omega)$  and  $\bar{\epsilon} > 0$  such that*

$$y_a + \bar{\epsilon} \leq G(\bar{\eta}) \leq y_b - \bar{\epsilon}$$

*on  $\bar{\Omega}$ . Assume that in each step,  $\lambda_k$  is chosen such that*

$$\sqrt{\lambda_k} \|\bar{\eta} - u_k\|_{L^\infty(\Omega)} \leq \bar{\epsilon} \quad (9)$$

$$\sqrt{\lambda_k} \|u_* - u_k\|_{L^\infty(\Omega)} \leq \bar{\epsilon}. \quad (10)$$

*If  $\lim_{k \rightarrow \infty} \lambda_k = 0$ , we have*

$$\lim_{k \rightarrow \infty} \|u_k - u_*\| = 0. \quad (11)$$

**Remark 2.** Condition (9) can easily be satisfied during the iteration by choosing  $\lambda_k$  sufficiently small since the functions  $\eta$  and  $u_k$  are known. Condition (10) can only be satisfied if an a priori bound for  $\|u_*\|_{L^\infty(\Omega)}$  is known. For the problem with additional pointwise control constraints, this problem does not occur, see Section 4. Lemma 7 states that if the  $\lambda_k$  decrease sufficiently fast we obtain convergence.

*Proof.* Let  $\tilde{u}$  denote a weak limit point of the sequence  $(u_k)_k$ . Then  $\tilde{u} \in F_*$ . Moreover, we have the inequality

$$\omega_* \leq K(\tilde{u}) \leq \liminf_{k \rightarrow \infty} \omega(\lambda_k, u_k).$$

Define  $\tau_k = \sqrt{\lambda_k}$  and the function  $v_k = (1 - \tau_k)u_* + \tau_k\bar{\eta}$ . Then we have

$$\begin{aligned} G(v_k) + \lambda_k(v_k - u_k) &= (1 - \tau_k)G(u_*) + \tau_k G(\bar{\eta}) \\ &+ \lambda_k((1 - \tau_k)(u_* - u_k) + \tau_k(\bar{\eta} - u_k)) \\ &\leq (1 - \tau_k)y_b + \tau_k(y_b - \bar{\epsilon}) \\ &+ (1 - \tau_k)\sqrt{\lambda_k}\bar{\epsilon} + \tau_k\sqrt{\lambda_k}\bar{\epsilon} \\ &= y_b - \sqrt{\lambda_k}\bar{\epsilon} + \sqrt{\lambda_k}\bar{\epsilon} \\ &= y_b. \end{aligned}$$

On the other hand, we have

$$\begin{aligned} G(v_k) + \lambda_k(v_k - u_k) &= (1 - \tau_k)G(u_*) + \tau_k G(\bar{\eta}) \\ &+ \lambda_k((1 - \tau_k)(u_* - u_k) + \tau_k(\bar{\eta} - u_k)) \\ &\geq (1 - \tau_k)y_a + \tau_k(y_a + \bar{\epsilon}) \\ &- (1 - \tau_k)\sqrt{\lambda_k}\bar{\epsilon} - \tau_k\sqrt{\lambda_k}\bar{\epsilon} \\ &= y_a + \sqrt{\lambda_k}\bar{\epsilon} - \sqrt{\lambda_k}\bar{\epsilon} \\ &= y_a. \end{aligned}$$

Hence  $v_k \in F(\lambda_k, u_k)$ . Moreover,  $\lim_{k \rightarrow \infty} \|v_k - u_*\| = 0$ . Thus we have

$$\begin{aligned} \limsup_{k \rightarrow \infty} \omega(\lambda_k, u_k) &\leq \limsup_{k \rightarrow \infty} K(v_k) \\ &= K(u_*) = \omega_*. \end{aligned}$$

Hence we have  $\lim_{k \rightarrow \infty} \omega(\lambda_k, u_k) = \omega_*$ . This implies that  $K(\tilde{u}) = \omega_*$ . Since  $\tilde{u} \in F_*$ , the uniqueness of the solution of  $\mathbf{P}$  implies  $\tilde{u} = u_*$ . Hence the sequence  $(u_k)_k$  converges weakly to  $u_*$ . Therefore

$$\begin{aligned} \lim_{k \rightarrow \infty} \kappa\|u_k\|^2 &= \lim_{k \rightarrow \infty} K(u_k) - \|S(u_k) - y_d\|^2 \\ &= K(u_*) - \|S(u_*) - y_d\|^2 \\ &= \kappa\|u_*\|^2. \end{aligned}$$

The weak convergence of  $u_k$  to  $u_*$  and the convergence of the norms imply  $\lim_{k \rightarrow \infty} \|u_k - u_*\| = 0$ . □

### 3.5. Constraint violation

For all  $k$  we have the inequalities

$$\begin{aligned} G(u_{k+1}) - y_b &\leq \lambda_k(u_k - u_{k+1}) \\ y_a - G(u_{k+1}) &\leq \lambda_k(u_{k+1} - u_k). \end{aligned}$$

This implies

$$\|(G(u_{k+1}) - y_b)_+\| + \|(y_a - G(u_{k+1}))_+\| \leq \lambda_k \|u_{k+1} - u_k\|. \quad (12)$$

Hence  $b_k = \lambda_k \|u_{k+1} - u_k\|$  is a bound for constraint violation. In the non-prox Lavrentiev regularization we have the corresponding bound  $t_k = \lambda_k \|u_{k+1}\|$ . If  $u_* \neq 0$  and  $\|u_k - u_*\| \rightarrow 0$  we have the inequality

$$\lim_{k \rightarrow \infty} \frac{t_k}{\lambda_k} = \|u_*\| > 0 = \lim_{k \rightarrow \infty} \frac{b_k}{\lambda_k} \quad (13)$$

which indicates that at least asymptotically, the Lavrentiev prox-regularization method yields smaller bounds for constraint violation.

### 3.6. Lipschitz continuity of $\omega(\lambda, \cdot)$

In this section we study the properties of the function  $\omega$ . We show its Lipschitz continuity for a fixed parameter  $\lambda > 0$ .

**Lemma 8.** *Let  $\lambda > 0$  and  $u, v \in L^2(\Omega)$  be given.*

*We define the proximal-point mapping as follows: For  $v \in L^2(\Omega)$  let  $\text{prox}(v)$  denote the solution of problem  $\mathbf{P}_{\lambda,v}$ . Let  $u_p = \text{prox}(u)$  and  $v_p = \text{prox}(v)$ . Define the number*

$$C_0 = \max\{\|S\| (\|S\| \|u_p\| + \|y_d\|) + \kappa \|u_p\|, \|S\| (\|S\| \|v_p\| + \|y_d\|) + \kappa \|v_p\|\}.$$

*Then we have the inequality*

$$\begin{aligned} |\omega(\lambda, u) - \omega(\lambda, v)| &\leq 2\lambda C_0 \|(\lambda I + S)^{-1}\| \|u - v\| \\ &+ \lambda^2 (\|S\|^2 + \kappa) \|(\lambda I + S)^{-1}\|^2 \|u - v\|^2. \end{aligned}$$

*In particular we have*

$$|\omega(\lambda, u) - \omega(\lambda, v)| = O(\|u - v\|). \quad (14)$$

*Proof.* We have  $y_a \leq \lambda(u_p - u) + S(u_p) \leq y_b$  and  $y_a \leq \lambda(v_p - v) + S(v_p) \leq y_b$ . Define  $\tilde{u}_p = u_p + (\lambda I + S)^{-1}\lambda(v - u)$ . Then  $\tilde{u}_p$  is feasible for  $\mathbf{P}_{\lambda,v}$  that is  $\tilde{u}_p \in F(\lambda, v)$ .

Define  $\tilde{v}_p = v_p + (\lambda I + S)^{-1}\lambda(u - v)$ . Then  $\tilde{v}_p$  is feasible for  $\mathbf{P}_{\lambda,u}$  that is  $\tilde{v}_p \in F(\lambda, u)$ . Hence we have the inequalities

$$\begin{aligned} \omega(\lambda, u) = K(u_p) &\leq K(\tilde{v}_p) \\ \omega(\lambda, v) = K(v_p) &\leq K(\tilde{u}_p). \end{aligned}$$

We have

$$\begin{aligned} K(\tilde{u}_p) - K(u_p) &= \int_{\Omega} \kappa (\tilde{u}_p^2 - u_p^2) + (S(\tilde{u}_p) - y_d)^2 - (S(u_p) - y_d)^2 dx \\ &= \int_{\Omega} \kappa \left\{ 2\lambda u_p (\lambda I + S)^{-1}(v - u) + [\lambda(\lambda I + S)^{-1}(v - u)]^2 \right\} \\ &\quad + 2\lambda(S(u_p) - y_d)S((\lambda I + S)^{-1}(v - u)) \\ &\quad + \lambda^2 [S((\lambda I + S)^{-1}(v - u))]^2 dx \end{aligned}$$

and

$$\begin{aligned} K(\tilde{v}_p) - K(v_p) &= \int_{\Omega} \kappa \left\{ 2\lambda v_p (\lambda I + S)^{-1}(u - v) + [\lambda(\lambda I + S)^{-1}(u - v)]^2 \right\} \\ &\quad + 2\lambda(S(v_p) - y_d)S((\lambda I + S)^{-1}(u - v)) \\ &\quad + \lambda^2 [S((\lambda I + S)^{-1}(u - v))]^2 dx. \end{aligned}$$

Hence

$$\begin{aligned} K(u_p) - K(v_p) &\leq K(\tilde{v}_p) - K(v_p) \\ &\leq [2\kappa\lambda\|v_p\| \|(\lambda I + S)^{-1}\| \\ &\quad + 2\lambda\|S\| (\|S\|\|v_p\| + \|y_d\|) \|(\lambda I + S)^{-1}\|] \|u - v\| \\ &\quad + [\kappa\lambda^2 \|(\lambda I + S)^{-1}\|^2 + \lambda^2\|S\|^2 \|(\lambda I + S)^{-1}\|^2] \|u - v\|^2 \\ &= 2(\|S\| (\|S\|\|v_p\| + \|y_d\|) + \kappa\|v_p\|) \lambda \|(\lambda I + S)^{-1}\| \|u - v\| \\ &\quad + (\|S\|^2 + \kappa) \lambda^2 \|(\lambda I + S)^{-1}\|^2 \|u - v\|^2. \end{aligned}$$

Analogously, we obtain the inequality

$$\begin{aligned} K(v_p) - K(u_p) &\leq K(\tilde{u}_p) - K(u_p) \\ &\leq 2(\|S\| (\|S\|\|u_p\| + \|y_d\|) + \kappa\|u_p\|) \lambda \|(\lambda I + S)^{-1}\| \|u - v\| \\ &\quad + (\|S\|^2 + \kappa) \lambda^2 \|(\lambda I + S)^{-1}\|^2 \|u - v\|^2. \end{aligned}$$

This yields

$$\begin{aligned} |K(v_p) - K(u_p)| &\leq \max \{ \|S\| (\|S\|\|u_p\| + \|y_d\|) + \kappa\|u_p\|, \\ &\quad \|S\| (\|S\|\|v_p\| + \|y_d\|) + \kappa\|v_p\| \} \\ &\quad \lambda 2 \|(\lambda I + S)^{-1}\| \|u - v\| \\ &\quad + \lambda^2 (\|S\|^2 + \kappa) \|(\lambda I + S)^{-1}\|^2 \|u - v\|^2. \end{aligned}$$

and the first part of the assertion follows. Lemma 5 implies the statement (14), since  $C_0$  remains uniformly bounded as a function of  $\lambda$ ,  $u$  and  $v$ .  $\square$

#### 4. The elliptic problem with pointwise control constraints

In this section we introduce an elliptic optimal control problem with state constraints. Let  $\kappa > 0$ ,  $N$ ,  $\Omega$ ,  $y_d$ ,  $y_a$ ,  $y_b$  be as before. In addition, let control bounds  $u_a, u_b \in L^\infty(\Omega)$  be given such that  $u_a \leq u_b$  on  $\Omega$ .

Define the following elliptic optimal control problem with distributed control, pointwise state constraints and pointwise control constraints:

$$\mathbf{Q} : \begin{cases} \text{minimize } J(y, u) \text{ subject to} \\ \partial_\nu y = 0 \text{ in } \Gamma \\ Ay = u \text{ in } \Omega \\ y_a \leq y \leq y_b \text{ in } \Omega \\ u_a \leq u \leq u_b \text{ in } \Omega. \end{cases} \quad (15)$$

Note that for a solution  $u_*$  of  $\mathbf{Q}$ , we have  $u_* \in L^\infty(\Omega)$ .

#### 5. Lavrentiev prox-regularization

Let the Lavrentiev regularization parameter  $\lambda > 0$  and  $v \in L^\infty(\Omega)$  be given. We consider the regularized problem

$$\mathbf{Q}_{\lambda, v} : \begin{cases} \text{minimize } K(u) \text{ subject to} \\ y_a \leq \lambda(u - v) + G(u) \leq y_b \text{ in } \Omega \\ u_a \leq u \leq u_b \text{ in } \Omega. \end{cases} \quad (16)$$

Let  $\nu_*$  denote the optimal value of  $\mathbf{Q}$ , and  $\nu(\lambda, v)$  denote the optimal value of  $\mathbf{Q}_{\lambda, v}$ . Let  $F_*$  denote the admissible set of  $\mathbf{Q}$  and  $F(\lambda, v)$  denote the admissible set of  $\mathbf{Q}_{\lambda, v}$ .

We consider the following **Lavrentiev prox-regularization algorithm**:

- Start:** CHOOSE  $u_1 \in L^\infty(\Omega)$  AND  $\lambda_1 > 0$ .
- Step  $k$ :** GIVEN  $u_k \in L^\infty(\Omega)$  AND  $\lambda_k > 0$ , SOLVE  $\mathbf{Q}_{\lambda_k, u_k}$ .  
DEFINE  $u_{k+1}$  AS THE SOLUTION OF  $\mathbf{Q}_{\lambda_k, u_k}$ . CHOOSE  $\lambda_{k+1} \in (0, \lambda_k]$ .  
GO TO STEP  $k + 1$ .

##### 5.1. Uniform boundedness of the feasible sets of $\mathbf{Q}_{\lambda, u}$

Due to the pointwise control constraints, the feasible points of  $\mathbf{Q}_{\lambda, u}$  are uniformly bounded in  $L^\infty(\Omega)$ :

**Lemma 9.** *Let  $v \in F(\lambda, u)$  be a feasible point of  $\mathbf{Q}_{\lambda, u}$ . Then*

$$\|v\|_{L^\infty(\Omega)} \leq \max\{\|u_a\|_{L^\infty(\Omega)}, \|u_b\|_{L^\infty(\Omega)}\}.$$

### 5.2. Well-definedness and convergence of the generated sequence

We study now the convergence of the solutions  $(u_k)_k$  for  $k \rightarrow \infty$ .

**Lemma 10.** *Assume that there exists a slater control  $\bar{\eta} \in L^\infty(\Omega)$  and  $\bar{\epsilon} > 0$  such that*

$$\begin{aligned} u_a &\leq \bar{\eta} \leq u_b, \\ y_a + \bar{\epsilon} &\leq G(\bar{\eta}) \leq y_b - \bar{\epsilon} \text{ almost everywhere on } \bar{\Omega}. \end{aligned}$$

Define  $M = \max\{\|u_a\|_{L^\infty(\Omega)}, \|u_b\|_{L^\infty(\Omega)}\}$ . Assume that in each step,  $\lambda_k$  is chosen such that  $\sqrt{\lambda_k} \leq \bar{\epsilon}/(2M)$ . Then  $\mathbf{Q}$  has a solution, the Lavrentiev prox-regularization algorithm is well defined, and if  $\lim_{k \rightarrow \infty} \lambda_k = 0$ , we have

$$\lim_{k \rightarrow \infty} \|u_k - u_*\| = 0. \quad (17)$$

Moreover, there exists a constant  $C_{10} > 0$  such that for all  $k$

$$|\nu(\lambda_k, u_k) - \nu_*| \leq C_{10} \sqrt{\lambda_k}. \quad (18)$$

*Proof.* First we show the existence of a solution of  $\mathbf{Q}$ . Since  $\bar{\eta}$  is feasible for  $\mathbf{Q}$ , we have  $\nu_* < \infty$ . Let  $m_k$  denote of minimizing sequence for  $\mathbf{Q}$ , that is the points  $m_k$  are feasible for  $\mathbf{Q}$  and  $\lim_{k \rightarrow \infty} K(m_k) = \nu_*$ . Since the sequence  $(m_k)_k$  is bounded in  $L^\infty(\Omega)$ , we can choose a subsequence that converges weakly\* in  $L^\infty(\Omega)$  to a limit point  $\bar{u} \in L^\infty(\Omega)$ . Then this subsequence converges also weakly in  $L^2(\Omega)$  to  $\bar{u}$ . Since the subsequence converges weakly in  $L^2(\Omega)$ , we have  $\nu_* = \liminf_{k \rightarrow \infty} K(m_k) \geq K(\bar{u})$ . Moreover, the weak\* convergence implies that  $\bar{u}$  is feasible for  $\mathbf{Q}$ . Hence  $\bar{u}$  is a solution of  $\mathbf{Q}$ . Due to the strong convexity of the objective function, this solution is uniquely determined.

Now we consider the sequence  $(u_k)$  generated by the algorithm. Define  $\tau_k = \sqrt{\lambda_k}$  and the function  $v_k = (1 - \tau_k)u_* + \tau_k\bar{\eta}$ . Then  $u_a \leq v_k \leq u_b$  and we have

$$\begin{aligned} G(v_k) + \lambda_k(v_k - u_k) &= (1 - \tau_k)G(u_*) + \tau_kG(\bar{\eta}) \\ &\quad + \lambda_k((1 - \tau_k)(u_* - u_k) + \tau_k(\bar{\eta} - u_k)) \\ &\leq (1 - \tau_k)y_b + \tau_k(y_b - \bar{\epsilon}) \\ &\quad + (1 - \tau_k)\lambda_k 2M + \tau_k\lambda_k 2M \\ &\leq y_b - \sqrt{\lambda_k}\bar{\epsilon} + \sqrt{\lambda_k}\bar{\epsilon} = y_b. \end{aligned}$$

On the other hand, we have  $G(v_k) + \lambda_k(v_k - u_k) \geq y_a$ . Hence  $v_k \in F(\lambda_k, u_k)$ . This implies that the iteration is well defined. Moreover,  $\lim_{k \rightarrow \infty} \|v_k - u_*\| = 0$ . Thus we have

$$\begin{aligned} \limsup_{k \rightarrow \infty} \nu(\lambda_k, u_k) &\leq \limsup_{k \rightarrow \infty} K(v_k) \\ &= K(u_*) = \nu_*. \end{aligned}$$

Let  $\tilde{u} \in L^\infty(\Omega)$  denote a weak\* limit point of the sequence  $(u_k)_k$ . Then  $\tilde{u} \in F_*$  and we have

$$K(\tilde{u}) \leq \liminf_{k \rightarrow \infty} \nu(\lambda_k, u_k) \leq \nu_*.$$

Since  $\tilde{u} \in F_*$ , the uniqueness of the solution of  $\mathbf{Q}$  implies  $\tilde{u} = u_*$ . Hence the sequence  $(u_k)_k$  converges weakly\* to  $u_*$ . This implies the equation

$$\begin{aligned}\lim_{k \rightarrow \infty} \kappa \|u_k\|^2 &= \lim_{k \rightarrow \infty} K(u_k) - \|S(u_k) - y_d\|^2 \\ &= K(u_*) - \|S(u_*) - y_d\|^2 = \kappa \|u_*\|^2.\end{aligned}$$

The weak convergence of  $u_k$  to  $u_*$  and the convergence of the norms imply  $\lim_{k \rightarrow \infty} \|u_k - u_*\| = 0$ .

There exists a Lipschitz constant  $C > 0$  such that for all points  $v_1, v_2 \in L^\infty(\Omega)$  with  $\|v_1\|_{L^\infty(\Omega)} \leq M$  and  $\|v_2\|_{L^\infty(\Omega)} \leq M$ , respectively we have  $K(v_1) \leq K(v_2) + C\|v_2 - v_1\|$ . Hence

$$\begin{aligned}\nu(\lambda_k, u_k) &\leq K(v_k) \leq K(u_*) + C\|v_k - u_*\| \\ &\leq K(u_*) + C\tau_k(\|u_*\| + \|\bar{\eta}\|) \leq \nu_* + \sqrt{\lambda_k} \sqrt{\mu(\Omega)} 2MC.\end{aligned}$$

For all  $k > 1$ , the point  $\tilde{u}_{k+1} = (1 - \tau_k)u_{k+1} + \tau_k\bar{\eta}$  is in  $F_*$ . Hence

$$\begin{aligned}\nu_* &\leq K(\tilde{u}_{k+1}) \leq K(u_{k+1}) + C\|\tilde{u}_{k+1} - u_{k+1}\| \\ &\leq \nu(\lambda_k, u_k) + C\tau_k(\|u_{k+1}\| + \|\bar{\eta}\|) \leq \nu(\lambda_k, u_k) + C\sqrt{\lambda_k} \sqrt{\mu(\Omega)} 2M.\end{aligned}$$

Define  $C_{10} = 2MC\sqrt{\mu(\Omega)}$ . Then (18) follows.  $\square$

**Remark 3.** The results about the improvement in constraint violation given in Section 3.5 also apply to the Lavrentiev prox-regularization algorithm for problem  $\mathbf{Q}$  compared with the corresponding non-prox Lavrentiev regularization algorithm where in step  $k$  problem  $\mathbf{Q}_{\lambda_k, 0}$  is solved.

## 6. Examples

In this section we study two examples that allow to compare the performance of the Lavrentiev prox-regularization method and the non-prox Lavrentiev regularization method.

**Example 1.** Consider a problem  $\mathbf{P}$ , where for the solution both inequality constraints are not active, that is we have  $y_a < G(u_*) < y_b$ . In this case,  $u_*$  is an unconstrained local minimal point of  $K$  and the convexity of  $K$  implies that  $\omega_* = K(u_*) = \min_{u \in L^2(\Omega)} K(u)$ . Let  $v \in L^2(\Omega)$  be given. Since  $F(\lambda, v) \subset L^2(\Omega)$ , for all  $\lambda > 0$  we have the inequality

$$\omega(\lambda, v) = \min_{u \in F(\lambda, v)} K(u) \geq \min_{u \in L^2(\Omega)} K(u) = K(u_*).$$

Lemma 1 implies that in this case we have  $\omega(\lambda, u_*) = \omega_*$ . Thus with the choice  $u_1 = u_*$ , the Lavrentiev prox-regularization method generates the constant sequence  $u_k = u_*$  for all  $k$  and all  $\lambda_k > 0$ , even if the sequence  $(\lambda_k)_k$  does not converge to zero.

More generally, the Lavrentiev prox-regularization method finds  $u_*$  in step  $k$  if  $u_* \in F(\lambda_k, u_k)$  which is the case if

$$\lambda_k \leq \min \left\{ \frac{\text{ess inf}_\Omega(y_a - G(u_*))}{\|u_k - u_*\|_{L^\infty(\Omega)}}, \frac{\text{ess inf}_\Omega(G(u_*) - y_b)}{\|u_k - u_*\|_{L^\infty(\Omega)}} \right\}.$$

For the non-prox Lavrentiev regularization method,  $u_*$  is the solution with the parameter  $\lambda_k$  if  $u_* \in F(\lambda_k, 0)$  which is the case if

$$\lambda_k \leq \min \left\{ \frac{\text{ess inf}_\Omega(y_a - G(u_*))}{\|u_*\|_{L^\infty(\Omega)}}, \frac{\text{ess inf}_\Omega(G(u_*) - y_b)}{\|u_*\|_{L^\infty(\Omega)}} \right\}.$$

If  $\|u_* - u_k\|_{L^\infty(\Omega)} < \|u_*\|_{L^\infty(\Omega)}$ , the Lavrentiev prox-regularization method can find  $u_*$  with larger parameter values  $\lambda_k$  than the non-prox Lavrentiev regularization.

**Example 2.** Consider a problem  $\mathbf{P}$ , where for the solution both inequality constraints are active almost everywhere in  $\Omega$ , that is we have  $y_a = G(u_*) = y_b$  and the Slater condition is violated. Assume that  $y_a \in C^2(\Omega)$  satisfies the boundary conditions  $\partial_\nu y_a = 0$  in  $\Gamma$ . In this case, we have  $S(u_*) = y_a$ .

The non-prox Lavrentiev regularization method computes the solution  $u_{k+1}^{NP}$  of  $\mathbf{P}_{\lambda_k, 0}$  for which we have the following equation:  $(\lambda_k I + G)u_{k+1}^{NP} = y_a$ . Hence  $(\lambda_k I + S)(u_{k+1}^{NP} - u_*) = y_a - \lambda_k u_* - y_a = -\lambda_k u_*$ .

This yields

$$u_{k+1}^{NP} - u_* = -\lambda_k(\lambda_k I + S)^{-1}u_*,$$

hence if  $\lambda_k \rightarrow 0$  we have

$$\lim_{k \rightarrow \infty} \|u_{k+1}^{NP} - u_*\| = \lim_{k \rightarrow \infty} \|\lambda_k(\lambda_k I + S)^{-1}u_*\| = 0.$$

The Lavrentiev prox-regularization method computes the solution  $u_{k+1}$  of  $\mathbf{P}_{\lambda_k, u_k}$  for which we have the following equation:  $(\lambda_k I + G)u_{k+1} - \lambda_k u_k = y_a$ . Hence  $(\lambda_k I + S)(u_{k+1} - u_*) = y_a + \lambda_k u_k - \lambda_k u_* - y_a = \lambda_k(u_k - u_*)$ . Thus if  $u_k \neq u_*$  we have

$$\|u_{k+1} - u_*\| = \|\lambda_k(\lambda_k I + S)^{-1}(u_k - u_*)\| < \|u_k - u_*\|$$

(see the proof of Lemma 3) hence the algorithm generates a bounded sequence with strictly decreasing distance to  $u_*$  also if  $(\lambda_k)_k$  does not converge to zero.

We have  $(\lambda_k I + S)(u_{k+1} - u_*) - \lambda_k u_k = -\lambda_k u_*$ , hence if  $\lambda_k \rightarrow 0$  we have

$$\lim_{k \rightarrow \infty} \|[u_{k+1} - \lambda_k(\lambda_k I + S)^{-1}u_k] - u_*\| = \lim_{k \rightarrow \infty} \|\lambda_k(\lambda_k I + S)^{-1}u_*\| = 0.$$

## 7. Conclusion

In this paper we have introduced the Lavrentiev prox-regularization method for an elliptic optimal control problem. The cost for the solution of the parametric auxiliary problems in each step of the method is the same as for the well-known non-prox Lavrentiev regularization method since the auxiliary problems are of exactly the same form. Hence also the same numerical methods can be used for the solution, for example primal-dual active set methods, interior point methods

of semismooth Newton methods, see for examples [3, 4, 6, 8]. Our convergence results indicate that the Lavrentiev prox-regularization method yields improved feasibility for a given regularization parameter  $\lambda_k$ . In other words, we can obtain approximations of the same quality as for the non-prox Lavrentiev regularization method with larger regularization parameters  $\lambda_k$ .

The corresponding Lavrentiev prox-regularization method for optimal control problems of parabolic type as studied in [7] and for elliptic boundary control problems with pointwise state-constraints as studied in [8] will be the subject of future research. Also the case of nonlinear elliptic optimal control problems will be considered as in [5].

## References

- [1] E. Casas and M. Mateos, *Second Order Optimality Conditions for Semilinear Elliptic Control Problems with Finitely Many State Constraints*. SIAM J. Control Optim. Volume 40, Issue 5, pp. 1431–1454, 2002.
- [2] A. Kaplan and R. Tichatschke, *Stable Methods for ill-posed variational problems*. Akademie Verlag, Berlin, 1994.
- [3] C. Meyer and U. Prüfert and F. Tröltzsch, *On two numerical methods for state-constraint elliptic control problems*, Preprint, Technische Universität Berlin 2005.
- [4] C. Meyer *Optimal control of semilinear elliptic equations with application to sublimation crystal growth*, Dissertation, Technische Universität Berlin 2006.
- [5] C. Meyer and F. Tröltzsch, *On an elliptic optimal control problem with pointwise mixed control-state constraints*, Recent Advances in Optimization, Lecture Notes in Economics and Mathematical Systems 563, A. Seeger, Ed., Springer-Verlag, 187–204, 2006.
- [6] C. Meyer and A. Rösch and F. Tröltzsch, *Optimal control of PDEs with regularized pointwise state constraints*, Computational Optimization and Applications 33 (2006), 209–228.
- [7] I. Neitzel and F. Tröltzsch, *On Convergence of Regularization Methods for Nonlinear Parabolic Optimal Control Problems with Control and State Constraints*, submitted.
- [8] F. Tröltzsch and I. Yousept, *A Regularization method for the numerical solution of elliptic boundary control problems with pointwise state-constraints*, to appear in COAP.
- [9] F. Tröltzsch, *Regular Lagrange multipliers for control problems with mixed pointwise control-state-constraints*, SIAM Journal on Optimization 15, 616–634, 2005.

Martin Gugat  
 Lehrstuhl 2 für Angewandte Mathematik  
 Martensstr. 3  
 D-91058 Erlangen  
 e-mail: gugat@am.uni-erlangen.de

# Nonlinear Feedback Solutions for a Class of Quantum Control Problems

Kazufumi Ito, Karl Kunisch and Qin Zhang

**Abstract.** Control of quantum systems described by the linear Schrödinger equation are considered. Control inputs enter through coupling operators and result in a bilinear control system. Feedback control laws are developed for orbit tracking. The asymptotic properties of the feedback laws are analyzed by the LaSalle-type invariance principle. Numerical integration via time-splitting is also investigated and used to demonstrate the feasibility of the proposed feedback laws and to compare their performance.

**Mathematics Subject Classification (2000).** 49L20, 35Q40, 93D20.

**Keywords.** Nonlinear feedback control, quantum control systems, Schrödinger equation.

## 1. Introduction

Consider a quantum system with internal Hamiltonian  $\mathcal{H}_0$  prepared in the initial state  $\Psi_0(x)$ , where  $x$  denotes the relevant spatial coordinate. The state  $\Psi(x, t)$  will be required to satisfy the time-dependent Schrödinger equation. In the presence of an external interaction taken as an electric field modeled by a coupling operator with amplitude  $\epsilon(t) \in \mathbb{R}$  and a time independent dipole moment operator  $\mu$  the controlled Hamiltonian results in  $\mathcal{H} = \mathcal{H}_0 + \epsilon(t)\mu$  and the following dynamical system is obtained:

$$i \frac{\partial}{\partial t} \Psi(x, t) = (\mathcal{H}_0 + \epsilon(t)\mu)\Psi(x, t), \quad \Psi(x, 0) = \Psi_0(x), \quad (1.1)$$

---

Research partially supported by the Army Research Office under DAAD19-02-1-0394, by the Fonds zur Förderung der wissenschaftlichen Forschung under SFB 32, “Mathematical Optimization and Applications in Biomedical Sciences” and the Air Force Office of Scientific Research under FA 9550-06-01-0241.

where  $\mathcal{H}_0$  is a positive, closed, self-adjoint operator in the Hilbert space  $H$ ,  $\mu \in \mathcal{L}(H)$  is self-adjoint, and  $\epsilon \in L^2(0, \infty)$  is the control input. Let  $X$  be the complexified Hilbert space corresponding to  $H$ , so that the inner product on  $X$  is defined by

$$(\Phi, \Psi)_X = (\Phi_1, \Psi_1)_H + (\Phi_2, \Psi_2)_H + i((\Phi_2, \Psi_1)_H - (\Phi_1, \Psi_2)_H),$$

where  $\Phi = (\Phi_1, \Phi_2)$ ,  $\Psi = (\Psi_1, \Psi_2)$ . This representation of  $\Psi$  with two components is used to formulate (1.1) as a bilinear control system (2.1)–(2.2) in  $X$ . Throughout we normalize the initial state by  $|\Psi_0|_X = 1$ .

We consider the control problem of driving the state  $\Psi(t)$  of (1.1) to an orbit  $\mathcal{H}(t)$  of the uncontrolled dynamics

$$i \frac{d}{dt} \mathcal{O}(t) = \mathcal{H}_0 \mathcal{O}(t), \quad (1.2)$$

specifically to the one that corresponds to an eigenstate or the manifold spanned by finite many eigenstates. An element  $\psi \in \text{dom}(\mathcal{H}_0)$  is an eigenstate of  $\mathcal{H}_0$  if  $\mathcal{H}_0 \psi = \lambda \psi$  for  $\lambda > 0$ . Then, the corresponding orbit is given by

$$\mathcal{O}(t) = e^{-i(\lambda t - \theta)} \psi, \quad (1.3)$$

where  $\theta \in [0, 2\pi)$  is the phase factor. We have  $|\mathcal{O}(t)|_X = 1$  if  $\psi$  is normalized as  $|\psi|_H = 1$ . We consider the discrete spectrum case: i.e., it is assumed that the spectrum of  $\mathcal{H}_0$  consists only of eigenvalues  $\{\lambda_k\}$ , and that the family of eigenfunctions  $\{\psi_k\}_{k=1}^\infty$  forms an orthonormal basis of  $X$ . The eigenvalues  $\{\lambda_k\}$  are arranged in increasing order.

We employ a variational approach based on the Lyapunov functional

$$V(t) = V(\Psi(t), \mathcal{O}(t)) = \frac{1}{2} |\Psi(t) - \mathcal{O}(t)|_X^2. \quad (1.4)$$

If  $V(t) \rightarrow 0$  as  $t \rightarrow \infty$  for a controlled orbit  $\Psi$  of (1.1) then  $\Psi(t)$  achieves the tracking of the target orbit  $\mathcal{O}(t)$  asymptotically. Variational approaches were previously discussed in [BCMR, MRT, IK2], for example. The more general case

$$\mathcal{O}(t) = \sum_{k=1}^N A_k e^{-i(\lambda_k t - \theta_k)} \psi_k, \quad (1.5)$$

where  $\{(\lambda_k, \psi_k)\}_{k=1}^N$  are the first  $N$  eigenpairs of  $\mathcal{H}_0$  and  $\sum_{k=1}^N A_k^2 = 1$  is also of our interests. The functional  $V_2(t) = \frac{1}{2}(1 - |(\mathcal{O}(t), \Psi(t))_X|^2)$  is used for [BCMR, MRT] for the manifold tracking, i.e.,  $V_2$  is motivated by the fact that  $V_2(\Psi, \mathcal{O}) = 0$  if and only if  $\Psi = e^{i\theta} \mathcal{O}$  where the phase  $\theta \in [0, 2\pi)$  is arbitrary. We develop a feedback synthesis that achieves the orbit tracking based on the functional  $V$  defined by (1.4). We shall see in Section 2 that  $|\Psi(t)|_X = 1$  for all  $t \geq 0$ . Together with  $|\mathcal{O}(t)|_X = 1$  this implies that the functional  $V$  can equivalently be expressed as

$$V(\Psi(t), \mathcal{O}(t)) = 1 - \text{Re}(\mathcal{O}(t), \Psi(t))_X. \quad (1.6)$$

It will be shown that

$$\frac{d}{dt} V(\Psi(t), \mathcal{O}(t)) = \epsilon(t) \operatorname{Im}(\mathcal{O}(t), \mu\Psi(t))_X. \quad (1.7)$$

We propose the feedback law

$$\epsilon(t) = -\frac{1}{\alpha}(u(t) + \beta \operatorname{sign}(u(t))V(t)^\gamma) = F(\Psi(t), \mathcal{O}(t)), \quad (1.8)$$

where

$$u(t) = \operatorname{Im}(\mathcal{O}(t), \mu\Psi(t))_X, \quad V(t) = V(\Psi(t), \mathcal{O}(t)),$$

for  $\alpha > 0$ ,  $\beta \geq 0$ ,  $\gamma \in (0, 1]$ . From (1.7)

$$\frac{d}{dt} V(\Psi(t), \mathcal{O}(t)) = -\frac{1}{\alpha}(|u(t)|^2 + \beta |u(t)|V(t)^\gamma). \quad (1.9)$$

Note that  $u(t)$  is linear in  $\Psi(t)$  and the second term of the control law (1.8) can switch its sign. The case  $\beta = 0$  is a negative feedback law based on the Lyapunov energy equality (1.7) and was analyzed in [IK2]. It will be shown in Section 6 that the performance of the feedback law (1.8) significantly increases by incorporating the switching control term with  $\beta > 0$ . We believe this switching mechanism can be applied to a wide general class of control systems and feedback synthesis based on the Lyapunov stability.

In this paper we establish well-posedness of the feedback law (1.8) and analyze its asymptotic tracking properties. Sufficient conditions for the orbit tracking will be obtained.

In order to obtain an improved tracking capability we shall also analyze multiple control potentials of the form

$$\mu(t) = \sum_{j=1}^m \epsilon_j(t) \mu_j \quad (1.10)$$

and the corresponding feedback law

$$\epsilon_j(t) = -\frac{1}{\alpha}(u_j(t) + \beta \operatorname{sign}(u_j(t))V(t)^\gamma), \quad u_j(t) = \operatorname{Im}(\mathcal{O}(t), \mu_j\Psi(t))_X.$$

Section 2 is devoted to well-posedness of the dynamical system in open and closed loop form. In Section 3 it is shown that the feedback law  $F$  is optimal in the sense that  $\epsilon(t) = F(\Psi(t), \mathcal{O}(t))$  minimizes

$$\int_0^T \frac{\alpha}{2} \left( \left| \epsilon + \frac{\beta}{\alpha} \operatorname{sign}(u(t)) \right|^2 + \frac{1}{\alpha} \left( \frac{1}{2}|u(t)|^2 + \beta |u(t)|V(t)^\gamma \right) \right) dt + V(\Psi(T), \mathcal{O}(T))$$

where  $u(t) = \operatorname{Im}(\mathcal{O}(t), \mu\Psi(t))_X$ .

An operator splitting method for solving (1.1) is discussed in Section 4. Section 5 is devoted to analyzing the asymptotic tracking properties of the feedback control laws. Section 6 contains numerical experiments that demonstrate the feasibility of the proposed feedback laws. The nonlinear feedback law ( $\beta > 0$ ) significantly improves the tracking performance compared to the linear one ( $\beta = 0$ ).

## 2. Well-posedness

Associated to the closed, positive, self-adjoint operator  $\mathcal{H}_0$  densely defined in the Hilbert space  $H$ , we define the closed linear operator  $A_0$  in  $H \times H$  by

$$A_0 = \begin{pmatrix} 0 & \mathcal{H}_0 \\ -\mathcal{H}_0 & 0 \end{pmatrix}$$

with  $\text{dom}(A_0) = \text{dom}(\mathcal{H}_0) \times \text{dom}(\mathcal{H}_0)$ . Here  $\Psi = (\Psi_1, \Psi_2) \in H \times H$  is identified with  $\Psi = \Psi_1 + i\Psi_2 \in X$ . We note that

$$|(\Psi_1, \Psi_2)|_{H \times H} = |\Psi|_X, \text{ and } (\Phi, \Psi)_{H \times H} = \Re(\Phi, \Psi)_X,$$

and  $X$  is isometrically isomorphic with  $H \times H$  by means of

$$(\Phi, \Psi)_X = (\Phi_1, \Psi_1)_H + (\Phi_2, \Psi_2)_H + i((\Phi_2, \Psi_1)_H - (\Phi_1, \Psi_2)_H),$$

with  $\Phi = \Phi_1 + i\Phi_2$ ,  $\Psi = \Psi_1 + i\Psi_2$ . Throughout this paper this identification will be used.  $A_0$  is skew-adjoint, i.e.,

$$(A_0\Psi, \hat{\Psi})_{H \times H} = -(A_0\hat{\Psi}, \Psi)_{H \times H} \text{ for all } \Psi, \hat{\Psi} \in \text{dom}(A_0).$$

Thus by Stone's theorem [P],  $A_0$  generates  $C_0$ -group on  $X$  and  $|S(t)\Psi_0|_X = |\Psi_0|_X$ . Let  $V = \text{dom}(\mathcal{H}_0^{\frac{1}{2}})$  and  $X_2 = V \times V$ . Then  $\mathcal{H}_0 \in \mathcal{L}(V, V^*)$  with  $V^* = \text{dom}(\mathcal{H}_0^{-\frac{1}{2}})$  where  $V$  is equipped with

$$|\phi|_V^2 = \langle \mathcal{H}_0\phi, \phi \rangle_{V^* \times V}$$

as norm. The restriction of  $S(t)$  to  $X_2$  defines a  $C_0$  group.

Associated to the self-adjoint operator  $\mu \in \mathcal{L}(H)$  we define the skew-adjoint operator

$$B = \begin{pmatrix} 0 & \mu \\ -\mu & 0 \end{pmatrix}.$$

Then for  $\epsilon \in L^2(0, T)$  there exists a unique mild solution  $\Psi(t) \in C(0, T; X)$  to

$$\Psi(t) = S(t)\Psi_0 + \int_0^t S(t-s)\epsilon(s)B\Psi(s)ds, \quad t \in [0, T], \quad (2.1)$$

and

$$\frac{d}{dt}\Psi = A_0\Psi(t) + \epsilon(t)B\Psi(t) \quad \text{in } (\text{dom}(A_0))^*, \quad (2.2)$$

where

$$(\text{dom}(A_0))^* = \text{dom}(\mathcal{H}_0^{-1}) \times \text{dom}(\mathcal{H}_0^{-1}),$$

[IK], Chapter2, [P], Chapter 4. Equivalently

$$\frac{d}{dt}\Psi(t) = -i(\mathcal{H}_0\Psi(t) + \epsilon(t)\mu\Psi(t)) \quad \text{in } \text{dom}(\mathcal{H}_0^{-1})$$

where we identify  $\Psi \in X$  as  $\Psi = \Psi_1 + i\Psi_2$  by  $\Psi = (\Psi_1, \Psi_2) \in H \times H$ . Since  $\mathcal{O}(t)$  defined by (1.3) and (1.5) satisfy  $\mathcal{O}(t) \in C(0, T; \text{dom}(A_0)) \cap C^1(0, T; X)$  and

$$\frac{d}{dt}\mathcal{O}(t) = -i\mathcal{H}_0\mathcal{O}(t) \quad \text{in } H. \quad (2.3)$$

Thus,

$$\begin{aligned} \frac{d}{dt} \operatorname{Re}(\mathcal{O}(t), \Psi(t))_X &= \operatorname{Re}((-i\mathcal{H}_0\mathcal{O}(t), \Psi(t))_X + (\mathcal{O}(t), -i(\mathcal{H}_0\Psi(t) + \epsilon(t)\mu\Psi(t))_X) \\ &= \operatorname{Re}(i\epsilon(t)(\mathcal{O}(t), \mu\Psi(t)))_X = -\epsilon(t) \operatorname{Im}(\mathcal{O}(t), \mu\Psi(t))_X, \end{aligned}$$

which proves (1.7).

By inserting the feedback law (1.8) of  $\epsilon$  in to (2.1), we obtain the closed loop system of the form

$$\Psi(t) = S(t)\Psi_0 + \int_0^t S(t-s)F(\Psi(s), \mathcal{O}(s))B\Psi(s) ds. \quad (2.4)$$

We show that (2.4) has a solution. Let  $\operatorname{sign}(u)$  be the maximal monotone function

$$\operatorname{sign}(u) = \begin{cases} [-1, 1] & u = 0 \\ u/|u| & |u| > 0, \end{cases}$$

and  $\operatorname{sign}_\delta(u)$  be the Yosida approximation of  $\operatorname{sign}(u)$  for  $\delta > 0$ :

$$\operatorname{sign}_\delta(u) = \begin{cases} u/\delta & |u| \leq \delta \\ u/|u| & |u| \geq \delta. \end{cases}$$

Define the operators  $F_\delta$  by

$$F_\delta(\Psi, \mathcal{O}) = -\frac{1}{\alpha}(u + \beta \operatorname{sign}_\delta(u)V^\gamma), \quad u = \operatorname{Im}(\mathcal{O}(t), \mu\Psi). \quad (2.5)$$

Then assuming that  $V \geq c > 0$ ,  $F_\delta$  is Lipschitz continuous and  $|F_\delta| \leq M$  for all  $|\Psi|_X = |\mathcal{O}|_X = 1$ . Thus, it can be proved [IK], [IK2] that

$$\Psi(t) = S(t)\Psi_0 + \int_0^t S(t-s)F_\delta(\Psi(s), \mathcal{O}(s))B\Psi(s) ds$$

has a unique solution  $\Psi_\delta \in C(0, T; X)$ . Moreover

$$V_\delta(t) = V(0) - \int_0^t \frac{1}{\alpha}(|u_\delta(s)|^2 + \beta \operatorname{sign}_\delta(u_\delta)(s)V_\delta(s)^\gamma)u_\delta(s) ds,$$

where

$$V_\delta(s) = 1 - \operatorname{Re}(\Psi_\delta(s), \mathcal{O}(s)), \quad u_\delta(s) = \operatorname{Im}(\mathcal{O}(s), \mu\Psi_\delta(s)).$$

It follows from [BMS], Theorem 3.6 that there exists a subsequence  $\delta$  and  $\Psi \in C(0, T; X)$  for which  $\Psi_\delta$  converges to  $\Psi$  in  $C(0, T; X)$ . Thus  $u_\delta(t) \rightarrow u(t) = \operatorname{Im}(\mathcal{O}(t))(t, \mu\Psi(t))$  strongly in  $L^2(0, T; R)$ . Since  $\operatorname{sign}_\delta(u_\delta(t)) \in L^2(0, T; R)$ , there exists a subsequence of  $\delta$  such that  $\operatorname{sign}_\delta(t) \rightarrow z(t)$  weakly in  $L^2(0, T; R)$ . The sign operator is maximal monotone and hence  $z(t) = \operatorname{sign}(u(t))$ . Since  $V_\delta(t) \rightarrow V(t)$  in  $C(0, T; R)$ ,  $\epsilon_\delta(t) \rightarrow \epsilon(t) = u(t) + \operatorname{sign}(u(t))V(t)^\gamma$  weakly in  $L^2(0, T; R)$ . For all  $\phi \in X$

$$(\Psi_\delta(t), \phi) = (S(t)\Psi_0, \phi) + \int_0^t \epsilon_\delta(t)(S(t-s)B\Psi_\delta(s), \phi) ds$$

and letting  $\delta \rightarrow 0^+$  we have

$$(\Psi(t), \phi) = (S(t)\Psi_0, \phi) + \int_0^t \epsilon(s)(S(t-s)B\Psi(s), \phi) ds,$$

which implies that  $\Psi$  is the mild solution to (2.1) that corresponds to  $\epsilon$ . Moreover, we have

$$V(t) = V(0) - \int_0^t \frac{1}{\alpha} (|u(s)|^2 + \beta |u(s)|V(s)^\gamma) ds. \quad (2.6)$$

### 3. Optimality

We argue that

$$V(t, \Psi) = 1 - (\mathcal{O}(t), \Psi)_{H \times H}$$

satisfies the Hamilton Jacobi equation

$$\begin{aligned} \frac{\partial V}{\partial t} &+ \min_{\epsilon} [V_\Psi(A_0\Psi + \epsilon B\Psi)] \\ &+ \frac{\alpha}{2} \left| \epsilon + \frac{\beta}{\alpha} \text{sign}(u(t))V^\gamma \right|^2 + \frac{1}{\alpha} \left( \frac{1}{2}|u(t)|^2 + |u(t)|V^\gamma \right) = 0, \end{aligned} \quad (3.1)$$

where

$$u(t) = \text{Im}(\mathcal{O}(t), \mu\Psi) = -(\mathcal{O}(t), B\Psi)_{H \times H}$$

and

$$V_\Psi(\Phi) = -(\mathcal{O}(t), \Phi)_{H \times H}.$$

In fact,

$$\begin{aligned} &\frac{\alpha}{2} \left| \epsilon + \frac{\beta}{\alpha} \text{sign}(u(t))V^\gamma \right|^2 + u(t) \left( \epsilon + \frac{\beta}{\alpha} \text{sign}(u(t)) \right) V^\gamma + \frac{1}{2\alpha} |u(t)|^2 \\ &= \frac{\alpha}{2} \left| \epsilon + \frac{1}{\alpha} (u(t) + \beta \text{sign}(u(t))V^\gamma) \right|^2, \end{aligned} \quad (3.2)$$

and thus  $\epsilon^*(t)$  minimizes

$$\frac{\alpha}{2} \left| \epsilon + \frac{\beta}{\alpha} \text{sign}(u(t))V^\gamma \right|^2 + u(t) \left( \epsilon + \frac{\beta}{\alpha} \text{sign}(u(t))V^\gamma \right) + \frac{1}{2\alpha} |u(t)|^2.$$

This implies that

$$\begin{aligned} &\frac{\partial V}{\partial t} + V_\Psi(A_0\Psi + \epsilon^* B\Psi) \\ &+ \frac{\alpha}{2} \left| \epsilon^*(t) + \frac{\beta}{\alpha} \text{sign}(u(t))V^\gamma \right|^2 + u(t) \left( \frac{\beta}{\alpha} \text{sign}(u(t))V^\gamma \right)^2 + \frac{1}{2\alpha} |u(t)|^2 \end{aligned}$$

$$\begin{aligned}
&= -(A_0 \mathcal{O}(t), \Psi)_{H \times H} - (\mathcal{O}(t), A_0 \Psi + \epsilon^*(t) B \Psi)_{H \times H} \\
&\quad + \frac{\alpha}{2} \left| \epsilon^*(t) + \frac{\beta}{\alpha} \text{sign}(u(t)) V^\gamma \right|^2 + u(t) \left( \epsilon^*(t) + \frac{\beta}{\alpha} \text{sign}(u(t)) V^\gamma \right) |^2 \\
&\quad + \frac{1}{2\alpha} |u(t)|^2 \\
&= 0,
\end{aligned}$$

as desired.

We next show that  $\epsilon^*$  minimizes

$$J(\epsilon) = \int_0^T \left( \frac{\alpha}{2} |\epsilon(t) + \frac{\beta}{\alpha} \text{sign}(u(t)) V(t)^\gamma|^2 + \frac{1}{\alpha} \left( \frac{1}{2} |u(t)|^2 + \beta |u(t)| \right) dt \right) + V(\Psi(T), \mathcal{O}(T)),$$

over  $\epsilon \in L^2(0, T)$ . To this end choose any  $\epsilon \in L^2(0, T)$  and let  $\Psi(t) \in C(0, T; X)$  be the solution to (2.1)–(2.2). Since  $\mathcal{O}(t) \in C^1(0, T; X) \cap C(0, T; \text{dom}(A_0))$  we have

$$\frac{d}{dt} V(\mathcal{O}(t), \Psi(t)) = -(A_0 \mathcal{O}(t), \Psi(t))_{H \times H} - (\mathcal{O}(t), A_0 \Psi(t) + \epsilon(t) B \Psi(t))_{H \times H}.$$

Integrating this over  $(0, T)$  and using (3.2) we find

$$\begin{aligned}
&V(\Psi(T), \mathcal{O}(T)) \\
&\quad + \int_0^T \left( \frac{\alpha}{2} \left| \epsilon(t) + \frac{\beta}{\alpha} \text{sign}(u(t)) V(t)^\gamma \right|^2 + \frac{1}{\alpha} \left( \frac{1}{2} |u(t)|^2 + \beta |u(t)| V(t)^\gamma \right) dt \right) \\
&= V(\Psi(0), \mathcal{O}(0)) + \int_0^T \frac{\alpha}{2} \left| \epsilon(t) + \frac{1}{\alpha} (u(t) + \beta \text{sign}(u(t)) V(t)^\gamma) \right|^2 dt
\end{aligned}$$

where  $u(t) = -(\mathcal{O}(t), B \Psi(t))_{H \times H}$ . Hence

$$\epsilon^*(t) = F(\Psi^*(t), \mathcal{O}(t)),$$

where  $\Psi^*(t)$  is the trajectory corresponding to  $\epsilon^*(t)$ , minimizes  $J(\epsilon)$  over  $L^2(0, T)$ .

#### 4. Operator splitting and numerical methods

Since the Hamiltonian is the sum of  $\mathcal{H}_0$  and  $\epsilon(t)\mu$  it is very natural to consider time integration based on the operator splitting method. For the stepsize  $h > 0$  consider the Strang splitting method:

$$\frac{\hat{\Psi}_{k+1} - \hat{\Psi}_k}{h} = \epsilon_k B \frac{\hat{\Psi}_{k+1} + \hat{\Psi}_k}{2}, \quad \hat{\Psi}_k = S \left( \frac{h}{2} \right) \Psi_k, \quad \Psi_{k+1} = S \left( \frac{h}{2} \right) \hat{\Psi}_{k+1}, \quad (4.1)$$

where

$$\epsilon_k = \frac{1}{h} \int_{kh}^{(k+1)h} \epsilon(s) ds.$$

For time integration of the controlled Hamiltonian we employ the Crank-Nicolson scheme since it is a norm preserving scheme. In fact, since  $B$  is skew adjoint

$$\left( \frac{\Psi_{k+1} - \hat{\Psi}_k}{h}, \Psi_{k+1} + \hat{\Psi}_k \right)_X = 0,$$

and thus  $|\Psi_{k+1}|_X^2 = |\hat{\Psi}_k|_X^2$ . The Strang splitting is of second order as time-integration. The following theorem addresses the convergence of the scheme (4.1):

**Theorem 4.1.** *If we define  $\Psi_h(t) = \Psi_k$  on  $[kh, (k+1)h)$ , then*

$$|\Psi_h(t) - \Psi(t)|_X \rightarrow 0 \text{ uniformly in } t \in [0, T]$$

where  $\Psi(t)$ ,  $t \geq 0$ , satisfies

$$\Psi(t) = S(t)\Psi_0 + \int_0^t S(t-s)\epsilon(s)B\Psi(s) ds.$$

*Proof.* Define the one step transition operator

$$\Psi_{k+1} = T_h(t)\Psi_k$$

by

$$T_h(t) = S\left(\frac{h}{2}\right) \left(I - \frac{\epsilon_k h}{2}B\right)^{-1} \left(I + \frac{\epsilon_k h}{2}B\right) S\left(\frac{h}{2}\right) \Psi. \quad (4.2)$$

Then,  $|T_h(t)\Psi|_X = |\Psi|_X$  and  $T_h(t)$  is norm preserving. Define the difference quotient of  $T_h(t)$  by

$$A_h(t)\Psi = \frac{T_h(t)\Psi - \Psi}{h} = S\left(\frac{h}{2}\right) \frac{J_{h/2}(\epsilon_k B) - I}{h/2} S\left(\frac{h}{2}\right) \Psi + \frac{S(h)\Psi - \Psi}{h} \quad (4.3)$$

where

$$J_{h/2}(\epsilon_k B) = \left(I - \frac{\epsilon_k h}{2}B\right)^{-1}.$$

It follows from the Chernoff theorem [IK] that if the consistency

$$|A_h(t)\Psi - (A_0\Psi + \epsilon(t)B\Psi)|_X \rightarrow 0 \text{ as } h \rightarrow 0^+. \quad (4.4)$$

holds for  $\Psi \in \text{dom}(A)$ , then  $|\Psi_h(t) - \Psi(t)|_X \rightarrow 0$  uniformly in  $t \in [0, T]$ . In fact, since for  $\Psi \in X$

$$\lim_{h \rightarrow 0^+} \frac{J_{h/2}(\epsilon_k B) - I}{h/2} \Psi = \epsilon(t)B\Psi$$

and for  $\Psi \in \text{dom}(A)$

$$\lim_{h \rightarrow 0^+} \frac{S(h)\Psi - \Psi}{h} = A_0\Psi,$$

it follows from (4.3) that the consistency (4.4) holds for  $\Psi \in \text{dom}(A)$  and  $\epsilon \in C(0, T)$ .

Note that

$$\Psi_{k+1} = S(h)\Psi_k + hS\left(\frac{h}{2}\right) \epsilon_k J_{h/2}(\epsilon_k B) S\left(\frac{h}{2}\right) \Psi_k$$

and thus

$$\Psi^m = S(mh)\Psi_0 + \sum_{k=1}^m h S((m-k)h)S\left(\frac{h}{2}\right) \epsilon_k B J_{h/2}(\epsilon_k B) S\left(\frac{h}{2}\right) \Psi_{k-1}.$$

Thus, letting  $h \rightarrow 0$  in this expression,  $\Psi(t) \in C(0, T; X)$  satisfies (2.1).  $\square$

Suppose that for (4.1) there exists an  $\epsilon_k$  on  $[kh, (k+1)h]$  such that for  $\mathcal{O}_{k+1/2} = S\left(\frac{h}{2}\right)\mathcal{O}_k$

$$\begin{aligned} \epsilon_k &= F(\Psi_{k+1/2}, \mathcal{O}_{k+1/2}) = \frac{1}{\alpha}(u_{k+1/2} + \beta \operatorname{sign}(u_{k+1/2})V_k^\gamma), \\ u_{k+1/2} &= (\mathcal{O}_{k+1/2}, B\Psi_{k+1/2}), \quad \Psi_{k+1/2} = \frac{\hat{\Psi}_{k+1} + \hat{\Psi}_k}{2}. \end{aligned} \quad (4.5)$$

Then  $\Psi_k$  satisfies the closed loop system

$$\begin{aligned} \frac{\hat{\Psi}_{k+1} - \hat{\Psi}_k}{h} &= \epsilon_k B \frac{\hat{\Psi}_{k+1} + \hat{\Psi}_k}{2}, \quad \hat{\Psi}_k = S\left(\frac{h}{2}\right) \Psi_k, \\ \epsilon_k &= F(\Psi_{k+1/2}, \mathcal{O}_{k+1/2}), \quad \Psi_{k+1} = S\left(\frac{h}{2}\right) \hat{\Psi}_{k+1}. \end{aligned} \quad (4.6)$$

Since

$$V\left(S\left(\frac{h}{2}\right) \hat{\Psi}_{k+1}, S\left(\frac{h}{2}\right) \mathcal{O}_{k+1/2}\right) = V(\hat{\Psi}_{k+1}, \mathcal{O}_{k+1/2}),$$

the discrete analog of (2.6)

$$V(\Psi_{k+1}, \mathcal{O}_{k+1}) = V(\Psi_k, \mathcal{O}_k) + \frac{1}{\alpha}(|u_k|^2 + \beta |u_k| V(Psi_k, \mathcal{O}_k))^\gamma$$

holds for the closed loop (4.6).

Now, we show that there exists a unique  $\epsilon_k$  that satisfies (4.5). Let

$$\chi(u) = u + \beta \operatorname{sign}(u).$$

Then, it is equivalent to finding  $\epsilon \in \mathbb{R}$  that satisfies

$$\chi^{-1}(\alpha \epsilon) = (B\hat{\Psi}(\epsilon), \mathcal{O}_{k+1/2}), \quad (4.7)$$

where  $\hat{\Psi}(\epsilon)$  is the solution to

$$\frac{\hat{\Psi} - \hat{\Psi}_k}{h} = \epsilon B \frac{\hat{\Psi} - \hat{\Psi}_k}{2}.$$

Note that

$$\hat{\Psi} - \hat{\Psi}_k = h\epsilon B \left(I - \frac{h\epsilon}{2} B\right)^{-1} \hat{\Psi}_k$$

and thus

$$(B\hat{\Psi}(\epsilon), \mathcal{O}_{k+1/2}) = (B\hat{\Psi}_k, \mathcal{O}_{k+1/2}) + h\epsilon(B^2 \left(I - \frac{h\epsilon}{2} B\right)^{-1} \hat{\Psi}_k, \mathcal{O}_{k+1/2}).$$

Since

$$B^2 = \begin{pmatrix} -\mu^2 & 0 \\ 0 & -\mu^2 \end{pmatrix}$$

one can assume that there exists  $c > 0$  such that for all  $k$

$$(B^2 \hat{\Psi}_k, \mathcal{O}_{k+1/2}) \leq -c.$$

Thus,

$$\left( B^2 \left( I - \frac{h\epsilon}{2} B \right)^{-1} \hat{\Psi}_k, \mathcal{O}_{k+1/2} \right) \leq -\frac{c}{2},$$

if  $h > 0$  is sufficiently small and hence (4.7) has a unique solution.

## 5. Asymptotic tracking

The objective of this section is to analyze the asymptotic properties of the controlled system (1.1). Let  $\mathcal{O}$  be of the form

$$\mathcal{O}(t) = e^{-i(\lambda_{k_0} t - \hat{\theta})} \psi_{k_0}$$

for some eigenpair  $(\lambda_{k_0}, \psi_{k_0})$  of  $\mathcal{H}_0$  and phase  $\hat{\theta}$ . We assume that

$$\mu_{k_0}^k = (\psi_{k_0}, \mu \psi_k)_X \neq 0 \text{ for all } k = 1, 2, \dots, \quad (5.1)$$

and that

$$\{S(t)\Psi_0 : t \geq 0\} \text{ be relatively compact in } H \times H. \quad (5.2)$$

Assumption (5.2) holds, for example if  $\text{dom}(\mathcal{H}_0)$  is relatively compact in  $H$  and  $\psi_0 \in V \times V$ . In case the domain  $\Omega$  on which  $\mathcal{H}_0$  is defined is unbounded, we may assume that  $W = V \cap L^p(\Omega)$ ,  $p > 2$ , is compactly embedded in  $H = L^2(\Omega)$ . Then, if  $\Psi_0 \in W \times W$  and  $S(t)$  leaves  $W \times W$  invariant [IK2], we have (5.2). Since  $V(t) \geq 0$ , it follows from (2.6) that either  $V(t) \rightarrow \infty$  or  $\int_0^\infty |u(t)| dt < \infty$ . We also assume that

$$\int_0^\infty |\epsilon(t)| dt < \infty. \quad (5.3)$$

This assumption holds if either we use the regularized feedback law (2.5) for arbitrary  $\delta > 0$  or  $\beta = 0$ . Thus,

$$\lim_{t \rightarrow \infty} \int_0^t S(t-s) \epsilon(s) B \Psi(s) ds \text{ exists.}$$

It follows that  $\{\int_0^t S(t-s) \epsilon(s) B \Psi(s) : t \geq 0\}$  is compact in  $H \times H$ . Together with (5.2) we conclude that  $\{\Psi(t) : t \geq 0\}$  is relatively compact. We shall proceed with the asymptotic analysis utilizing assumptions (5.1)–(5.3) and summarize the results in a theorem at the end.

Since  $\{\Psi(t) : t \geq 0\}$  and  $\{\mathcal{O}(t) : t \geq 0\}$  are relatively compact in  $X$  there exists a sequence  $\{t_n\} \rightarrow \infty$  and elements  $\Psi_\infty \in X$ ,  $\mathcal{O}_\infty \in X$  such that

$$\lim_{n \rightarrow \infty} \Psi(t_n) = \Psi_\infty \text{ and } \lim_{n \rightarrow \infty} \mathcal{O}(t_n) = \mathcal{O}_\infty, \quad (5.4)$$

in particular,  $\Psi_\infty, \mathcal{O}_\infty$  are in the  $\omega$ -limit points of the flow defined by (2.2) and (2.3), respectively. Since  $\epsilon \in L^2(0, \infty)$  it follows from (2.1) that  $\Psi(t_n + \tau) \rightarrow S(\tau)\Psi_\infty$  and analogously  $\mathcal{O}(t_n + \tau) \rightarrow S(\tau)\mathcal{O}_\infty$  uniformly with respect to  $\tau \in (0, \infty)$ . Here  $S(\tau)\Psi_\infty$  and  $S(\tau)\mathcal{O}_\infty$  are the mild solutions to

$$\begin{aligned}\frac{d}{dt}\Psi_\infty(t) &= A_0\Psi_\infty(t), \quad \Psi_\infty(0) = \Psi_\infty, \\ \frac{d}{dt}\mathcal{O}_\infty(t) &= A_0\mathcal{O}_\infty(t), \quad \mathcal{O}_\infty(0) = \mathcal{O}_\infty.\end{aligned}$$

Moreover

$$\Psi_\infty(\tau) = \sum_{k=1}^{\infty} A_k e^{-i(\lambda_k \tau - \theta_k)} \psi_k, \quad \mathcal{O}_\infty(\tau) = e^{-i(\lambda_{k_0} \tau - \tilde{\theta}_{k_0})} \psi_{k_0},$$

with  $0 \leq \theta_k, \tilde{\theta}_{k_0} < \pi$  and  $\sum |A_k|^2 = 1$ . Since

$$u(t_n + \cdot) = \text{Im}(\mathcal{O}(t_n + \cdot)), \quad \mu\Psi(t_n + \cdot) \rightarrow 0 \text{ in } L^2(0, \infty), \text{ as } t_n \rightarrow \infty,$$

we have

$$u(\tau) = \text{Im}(\mathcal{O}_\infty(\tau), \mu\Psi_\infty(\tau)) = 0, \text{ for } \tau \geq 0. \quad (5.5)$$

It follows now that

$$\begin{aligned}u(\tau) &= \text{Im} \left( \sum_{k=1}^{\infty} A_k e^{i((\lambda_k - \lambda_{k_0})\tau - \theta_k + \tilde{\theta}_{k_0})} \mu_{k_0}^k \right) \\ &= \sum_{k=1}^{\infty} \mu_{k_0}^k A_k \left( \cos(\theta_k - \tilde{\theta}_{k_0}) \sin((\lambda_k - \lambda_{k_0})\tau) - \sin(\theta_k - \tilde{\theta}_{k_0}) \cos((\lambda_k - \lambda_{k_0})\tau) \right) \\ &= 0.\end{aligned} \quad (5.6)$$

Suppose that the family

$$\{\cos((\lambda_k - \lambda_{k_0})\tau), \sin((\lambda_k - \lambda_{k_0})\tau)\} \text{ is } \omega\text{-independent in } L^2(0, T). \quad (5.7)$$

Here a family of elements  $\{\varphi_k\}_{k=-\infty}^{\infty}$  in  $L^2(0, T)$ ,  $T > 0$  is called  $\omega$ -independent, if  $\sum_{k=-\infty}^{\infty} c_k \varphi_k = 0$  implies that  $c_k = 0$  for all  $k$ . If (5.7) holds, then  $\mu_{k_0}^k A_k = 0$  for  $k \neq k_0$  and  $\mu_{k_0}^{k_0} A_{k_0} - \sin(\theta_k - \tilde{\theta}_{k_0}) = 0$ . Thus by (5.1),  $A_k = 0$  for  $k \neq k_0$ . Moreover, since  $|\Psi_\infty| = 1$ , we have  $\theta_{k_0} = \tilde{\theta}_{k_0}$  and  $A_{k_0} = 1$ . Here the case  $A_{k_0} = -1$  can be excluded since it implies that

$$V(\Psi_\infty(\tau), \mathcal{O}_\infty(\tau)) = 1 + \text{Re}(e^{-i(\lambda_{k_0} \tau - \theta_{k_0})} \psi_{k_0}, e^{-i(\lambda_{k_0} \tau - \theta_{k_0})} \psi_{k_0})_X = 2,$$

and

$$V(\Psi_0, \mathcal{O}(0)) = 1 - \text{Re}(e^{i\tilde{\theta}_{k_0}} \psi_{k_0}, \Psi_0)_X = 1 - \text{Re}(e^{i\tilde{\theta}_{k_0}} (\psi_{k_0}, \Psi_0)_X) < 2,$$

since  $\tilde{\theta}_{k_0} \in [0, \pi]$ . Hence,  $A_{k_0} = -1$  is impossible due to  $\frac{d}{dt} V(\Psi(t), \mathcal{O}(t)) \leq 0$ .

Since the  $\omega$ -limit pair  $(\Psi_\infty, \mathcal{O}_\infty)$  was arbitrary it follows from (1.4) that  $\lim_{t \rightarrow \infty} V(\Psi(t), \mathcal{O}(t)) = 0$ , i.e.,  $\Psi(t)$  asymptotically approaches the orbit  $\mathcal{O}(t)$ . We summarize the above discussion in the following result.

**Theorem 5.1.** Assume that (5.1), (5.2) and (5.7) hold. Then

$$\lim_{t \rightarrow \infty} V(\Psi(t), \mathcal{O}(t)) = 0,$$

for the feedback law  $F$ .

The following lemma addresses condition (5.7).

**Lemma 5.1.** If there exists a constant  $\delta > 0$  such that  $|\lambda_k + \lambda_\ell - 2\lambda_{k_0}| \geq \delta$  for all  $k, \ell \geq 1$  with  $\ell \neq k_0$ , and  $|\lambda_k - \lambda_\ell| \geq \delta$  for all  $k \neq \ell$ , then  $\{e^{i(\lambda_k - \lambda_{k_0})\tau}\} \cup \{e^{-i(\lambda_k - \lambda_{k_0})\tau}\}_{k \neq k_0}$  is  $\omega$ -independent for sufficiently large  $T > 0$ .

*Proof.* Let  $\{\mu_\ell\}_{\ell \in I}$  be a real number sequence defined by

$$\mu_k = \lambda_k - \lambda_{k_0}, \quad k \geq 1, \quad \mu_{-k} = -(\lambda_k - \lambda_{k_0}), \quad k \neq k_0.$$

It follows from the assumption that

$$|\mu_m - \mu_\ell| \geq \delta, \quad m \neq \ell$$

From Ingham's theorem [I], if  $T > \frac{2\pi}{\delta}$ , there exists a constant  $c$ , depending on  $T$  and  $\delta > 0$  such that

$$c \sum_{m \in I} |a_m|^2 \leq \int_0^T |f(\tau)|^2 d\tau \quad \text{for} \quad f(\tau) = \sum_{m \in I} a_m e^{i\mu_m \tau}. \quad \square$$

**Remark 5.1.** For the harmonic oscillator case we have

$$\mathcal{H}_0 \psi = -\frac{d^2}{dx^2} \psi + x^2 \psi, \quad x \in \mathbb{R} = \Omega.$$

The eigenpairs  $\{(\lambda_k, \psi_k)\}_{k=1}^\infty$  are given by

$$\lambda_k = 2k - 1, \quad \psi_k(x) = \hat{c} H_{k-1}(x) e^{-\frac{x^2}{2}},$$

where  $H_k$  is the Hermite polynomial of degree  $k$  and  $\hat{c}$  is a normalizing factor. In this case we have

$$\lambda_{k_0-\ell} - \lambda_{k_0} = -(\lambda_{k_0+\ell} - \lambda_{k_0}), \quad 1 \leq \ell \leq k_0 - 1,$$

and the gap condition  $|\lambda_k + \lambda_\ell - 2\lambda_{k_0}| > \delta$  is not satisfied. Thus,  $\int_0^T |u(\tau)|^2 d\tau = 0$  implies

$$\operatorname{Im} (A_{k_0+\ell} e^{i(\lambda_\ell \tau - \theta_{k_0+\ell} + \tilde{\theta}_{k_0})} \mu_{k_0}^{k_0+\ell} + A_{k_0-\ell} e^{-i(\lambda_\ell \tau - \theta_{k_0-\ell} + \tilde{\theta}_{k_0})} \mu_{k_0}^{k_0-\ell}) = 0$$

for  $1 \leq \ell < k_0$ . That is,  $A_{k_0+\ell}$  and  $A_{k_0-\ell}$  are not necessarily zero and thus  $\Psi_\infty(\tau)$  is distributed over energy levels  $1 \leq \ell \leq 2k_0 - 1$ .

### 5.1. Degenerated case

We now turn to the case when the gap condition  $|\lambda_k + \lambda_\ell - 2\lambda_{k_0}| > \delta$  is violated. Then more than one control operator  $\mu$  is required and we consider (1.10). For  $V(\Psi, \mathcal{O}) = 1 - \text{Re}(\mathcal{O}, \Psi)_X$  we find

$$\frac{d}{dt}V(\Psi(t), \mathcal{O}(t)) = \sum_{j=1}^m \epsilon_j \text{Im}(\mathcal{O}(t), \mu_j \Psi(t))_X,$$

which suggests feedback laws of the form

$$\epsilon_j(t) = -\frac{1}{\alpha}(u_j(t) + \beta \text{sign}(u_j(t))V(t)^\gamma), \quad u_j(t) = \text{Im}(\mathcal{O}(t), \mu_j \Psi(t)). \quad (5.8)$$

As shown above, we have

$$V(t) - V(0) = -\frac{1}{\alpha} \int_0^t \sum_{j=1}^m (|u_j(s)|^2 + \beta |u_j(s)|V(s)^\gamma) ds.$$

In the following discussion we assume (5.2), i.e., that  $\{S(t)\Psi_0 : t \geq 0\}$  is compact. Then using the same arguments as above for all  $\omega$ -limits  $\{\Psi_\infty(\tau) : \tau \geq 0\}$  we have

$$u_j(\tau) = \text{Im}(\mathcal{O}_\infty(\tau), \mu_j \Psi_\infty(\tau)) = 0, \quad \text{for } \tau \geq 0, j = 1, \dots, m.$$

Thus,

$$\text{Im} \left( \sum_{k=1}^{\infty} A_k e^{i((\lambda_k - \lambda_{k_0})\tau - \theta_k + \tilde{\theta}_{k_0})} (\mu_j)_{k_0}^k \right) = 0, \quad \text{for } j = 1, \dots, m,$$

where

$$(\mu_j)_{k_0}^k = (\psi_{k_0}, \mu_j \psi_k)_X.$$

We henceforth consider the case  $m = 2$ . Suppose that  $\lambda_{\bar{k}} + \lambda_{\bar{\ell}} - 2\lambda_{k_0} = 0$  for a single pair  $(\bar{k}, \bar{\ell})$ ,  $\bar{\ell} \neq k_0$ , and that otherwise (5.7) holds. Then  $\lambda_{\bar{k}} - \lambda_{k_0} = -(\lambda_{\bar{\ell}} - \lambda_{k_0})$  and we have

$$\text{Im} \left( A_{\bar{k}} e^{i((\lambda_{\bar{k}} - \lambda_{k_0})\tau - \theta_{\bar{k}} + \tilde{\theta}_{k_0})} (\mu_j)_{k_0}^{\bar{k}} + A_{\bar{\ell}} e^{i(-(\lambda_{\bar{k}} - \lambda_{k_0})\tau - \theta_{\bar{\ell}} + \tilde{\theta}_{k_0})} (\mu_j)_{k_0}^{\bar{\ell}} \right) = 0, \quad (5.9)$$

for  $j = 1, 2$ . If

$$\text{rank} \begin{pmatrix} (\mu_1)_{k_0}^{\bar{k}} & (\mu_1)_{k_0}^{\bar{\ell}} \\ (\mu_2)_{k_0}^{\bar{k}} & (\mu_2)_{k_0}^{\bar{\ell}} \end{pmatrix} = 2, \quad (5.10)$$

then from (5.9), it follows that  $A_{\bar{k}} = A_{\bar{\ell}} = 0$ . If moreover

$$\text{for each } k \text{ there exists } j \in \{1, 2\} \text{ such that } (\mu_j)_{k_0}^k \neq 0, \quad (5.11)$$

then  $A_k = 0$  for all  $k \neq k_0$ ,  $A_{k_0} = 1$  and  $\theta_{k_0} = \tilde{\theta}_{k_0}$ . As a consequence we have  $\lim_{t \rightarrow \infty} V(\Psi(t), \mathcal{O}(t)) = 0$ .

In general let

$$\lambda_{k_i} + \lambda_{\ell_i} - 2\lambda_{k_0} = 0$$

for multiple pairs  $(k_i, \ell_i)$  with  $\ell_i \neq k_0$ . If

$$\text{rank} \begin{pmatrix} (\mu_1)_{k_0}^{k_i} & (\mu_1)_{k_0}^{\ell_i} \\ (\mu_2)_{k_0}^{k_i} & (\mu_2)_{k_0}^{\ell_i} \end{pmatrix} = 2 \quad (5.12)$$

for each  $i$ , then  $A_{k_i} = A_{\ell_i} = 0$ , and in particular  $A_k = 0$  for all  $k$ . If in addition (5.11) holds then, again  $\lim_{t \rightarrow \infty} V(\Psi(t), \mathcal{O}(t)) = 0$ .

## 6. Numerical tests

In this section we demonstrate the feasibility of the proposed feedback laws using a test example. We set  $H = L^2(0, 1)$  and

$$\mathcal{H}_0 \psi = \sum_{k=1}^{\infty} \lambda_k (\psi, \psi_k)_H \psi_k,$$

where  $\psi_k(x) = \sqrt{2} \sin(k\pi x)$  and  $\lambda_k = k\pi$ . The control Hamiltonians are given by

$$(\mu_i \Psi)(x) = b_i(x) \Psi(x), \quad x \in (0, 1),$$

with  $i = 1, 2$ . For computations we truncated the expansion of  $\mathcal{H}_0$  at  $N = 99$ , so that

$$S_N(h) \Psi_0 = \sum_{k=1}^N e^{-i\lambda_k h} (\Psi_0, \psi_k) \psi_k.$$

To integrate the control Hamiltonian term the collocation method was used in the form

$$(B_i^N \psi)(x_n^N) = b_i(x_n^N) \psi(x_n^N), \quad i = 1, 2,$$

where  $x_n^N = \frac{n}{N}$ ,  $1 \leq n \leq N - 1$ . Thus, we implemented the feedback law based on the Strang splitting method in the form

$$\begin{aligned} \Psi^{k+1} &= S_N \left( \frac{h}{2} \right) \mathcal{F}_N \left( I - \frac{\epsilon_1^k h}{2} B_1^N - \frac{\epsilon_2^k h}{2} B_2^N \right)^{-1} \left( I + \frac{\epsilon_1^k h}{2} B_1^N + \frac{\epsilon_2^k h}{2} B_2^N \right) S_N \left( \frac{h}{2} \right) \\ \epsilon_i^k &= F_i(\Psi^{k+1/2}, \mathcal{O}^{k+1/2}), \quad i = 1, 2, \end{aligned}$$

where  $\mathcal{F}_N$  and  $\mathcal{F}_N^{-1}$  are the discrete Fourier sine transform and its inverse transform, respectively, and  $B_i^N$  is the diagonal matrix with diagonal

$$(b_i(x_1^N), \dots, b_i(x_{N-1}^N)) \quad \text{for each } i = 1, 2.$$

Well-posedness of this implicit method was discussed in Section 4 for given  $\beta > 0$  and  $\gamma \in [0, 1]$ . The numerical tests that we report on are obtained with  $h = 0.01$ ,  $\alpha = 1/500$  and

$$b_1(x) = (x - .5) + 1.75(x - .5)^2, \quad b_2(x) = 2.5(x - .5)^3 - 2.5(x - .5)^4.$$

These control potentials satisfy the rank condition of Section 5. They are selected by minimizing the tracking time by trial and error tests. Figure 1 provides a comparison for the orbit tracking performance  $V = \frac{1}{2} |\Psi(t) - \mathcal{O}(t)|_X^2$  between

different  $\beta$ -values and different powers  $\gamma$  for  $V$ . As  $\beta$  increases, the performance  $V$  significantly improves and the 10% performance level is achieved in a much shorter horizon. By decreasing the power of  $V$ , the performance of  $V$  improves also, and more rapidly in the beginning of the time horizon.

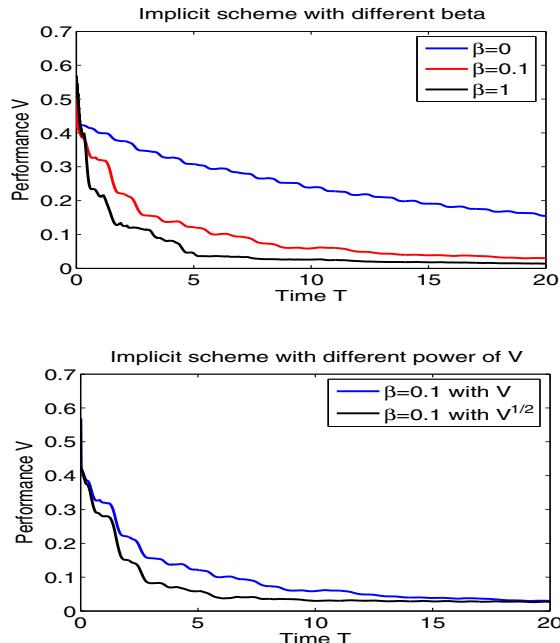


FIGURE 1. Performance comparison

## References

- [BCMR] K. Beauchard, J.M. Coron, M. Mirrahimi, and P. Rouchon, Implicit Lyapunov control of finite-dimensional Schrödinger equations, preprint.
- [BMS] J.M. Ball, J.M. Marsden and M. Slemrod: Control of bilinear systems, SIAM J. Control and Optimization 20(1982), 575–597.
- [I] A.E. Ingham, Some trigonometrical inequalities with applications to the theory of series, Math. Z. 41(1936), 367–379.
- [IK] K. Ito and F. Kappel, *Evolution Equations and Approximations*, World Scientific, New Jersey, 2002.
- [IK1] K. Ito and K. Kunisch, Optimal bilinear control of an abstract Schrödinger equation. SIAM J. Control Optim. 46 (2007), 274–287.
- [IK2] K. Ito and K. Kunisch, Asymptotic Properties of Feedback Solutions for a Class of Quantum Control Problems, SIAM J. Control Optim., submitted.

- [MRT] M. Mirrahimi, P. Rouchon, and G. Turinici, Lyapunov control of bilinear Schrödinger equations, *Automatica*, 41(2005), 1987–1994.
- [P] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Berlin, 1983.

Kazufumi Ito and Qin Zhang  
Department of Mathematics  
North Carolina State University  
Raleigh, North Carolina, 27695-8205, USA

Karl Kunisch  
Institut für Mathematik  
Karl-Franzens-Universität Graz  
A-8010 Graz, Austria

# Optimal Feedback Synthesis for Bolza Control Problem Arising in Linearized Fluid Structure Interaction

Irena Lasiecka and Amjad Tuffaha

**Abstract.** Bolza boundary control problem defined for linearized fluid structure interaction model is considered. The aim of this paper is to develop an optimal feedback control synthesis based on Riccati theory. The main mathematical challenge of the problem is caused by unbounded action of control forces which, in turn, give rise to Riccati equations with unbounded coefficients and singular behavior of the gain operator. This class of problems has been recently studied via the so-called Singular Estimate Control Systems (SECS) theory, which is based on the validity of the so-called Singular Estimate (SE) [4, 27, 32]. It is shown that the fluid structure interaction does satisfy Singular Estimate (SE) condition. This is accomplished by showing that the maximal abstract parabolic regularity is transported, onto the wave dynamics, via hidden hyperbolic regularity of the boundary traces on the interface. The established Singular Estimate allows for the application of recently developed general theory which, in turn, implies well-posedness of feedback synthesis and of the associated Riccati Equation. Blow up rates of optimal control and of the feedback operator at the terminal time are provided.

**Mathematics Subject Classification (2000).** 35Bxx, 35B37.

**Keywords.** Fluid Structure Interaction, Boundary Control, Singular Estimate Control System, Riccati Equation, Feedback Control, Hyperbolic Trace Theory, Maximal Parabolic Regularity.

## 1. Introduction

The mathematical model consists of linearized Navier-Stokes equation defined on an open domain  $\Omega_f$  coupled with an elastic equation defined on another domain  $\Omega_s$ , with boundary conditions matching velocities and normal stresses on the boundary  $\Gamma_s$  which separates the two open domains  $\Omega_f$  and  $\Omega_s$ . Various versions of this model that describes the elastic motion of a solid fully immersed in a viscous

incompressible fluid, have received an extensive coverage in the literature [26, 19, 37, 24, 16, 15, 34].

We consider a boundary control system of this fluid structure interaction model, with the objective of developing a feedback optimal control, acting as a force on the interface between the two media. The construction is based on a solution to the appropriate Riccati equation.

It is known that Riccati theory is a very powerful tool for designing and computing feedback control for finite-dimensional systems. In response to numerous technological applications, where correct modeling requires an infinite-dimensional state space (as in PDE theory), an extension of the Riccati theory becomes available for infinite-dimensional control systems modeled by PDEs [42, 7, 41, 20, 17, 43, 31] and references therein. Though actual computations of feedback controller are performed on *finite-dimensional* structures, the role of infinite-dimensional Riccati theory needs not be defended. In fact, as documented by a large body of literature, rigorous infinite-dimensional theory is responsible for stability and consistency estimates obtained for finite-dimensional approximation of Riccati equations [25, 18, 35, 36, 39]. The infinite-dimensional Riccati theory was first developed for control systems generated by strongly continuous semi-groups with *bounded* control operators. However, the optimal control theory has an additional degree of complexity when studying systems with *unbounded control actions*, the latter arise in boundary or point control. The mathematical difficulty in extending the Riccati theory available for bounded controls, has to do with the fact that the so-called gain (feedback) operator along with the coefficients in the associated Riccati equation to the system might not be well defined. Therefore, a standard like treatment of an unbounded control system via a Riccati equation and their finite-dimensional approximations ought to be justified with a sound theoretical framework. The framework has indeed been laid out in the case of systems generated by analytic semigroups [7, 22, 23, 17, 1, 31], where the theory has acquired a reasonable degree of maturity and completeness. More recently the analytic approach has been extended to a class of systems referred to as Singular Estimate Control Systems (SECS) which captures partial or approximate analytic dynamics [4, 29, 27, 13, 2, 12, 32] combined with hyperbolic structures. Coupled PDE systems which do not have analytic generators, but rather combine hyperbolic and parabolic effects are classical prototypes for the SECS systems. A prime example of such is structural acoustic interaction for which SECS theory has been laid out in [4]. Since then a rather rich theory pertaining to SECS systems has been developed over the last decade or so. Of particular interest to this work is Bolza Riccati SECS theory which involves penalization of the terminal state [32]. It is known that Bolza problems, even in the case of analytic dynamics, do lead to singular behavior of optimal controls and of Riccati feedback operators.

The general formulation of SECS (Singular Estimate Control Systems) class is as follows: Consider the dynamics

$$y_t = \mathcal{A}y + \mathcal{B}g \in [\mathcal{D}(\mathcal{A}^*)]' \quad (1)$$

with state space  $\mathcal{H}$  and a control space  $U$  while  $\mathcal{A}$  is a generator of a strongly continuous semigroup  $e^{\mathcal{A}t}$  on  $\mathcal{H}$ , and  $\mathcal{B}$  is an unbounded control operator such that  $\mathcal{B} \in \mathcal{L}(U \rightarrow [D(\mathcal{A}^*)]')$ . The additional singular estimate (SE) condition is the following one: with some  $0 \leq \gamma < 1$ .

$$|\mathcal{O}e^{\mathcal{A}t}\mathcal{B}g|_{\mathcal{Z}} \leq \frac{C}{t^\gamma}|g|_U \quad (2)$$

where  $\mathcal{O}$  denotes selected observations of the system defined via bounded operators from the state space  $\mathcal{H}$  into the observed space  $\mathcal{Z}$ . The control problem considered is to minimize any functional of the general form

$$J(y, g) = \int_0^T [|Ry|_{\mathcal{Z}}^2 + |g|_U^2]ds + |Gy(T)|_W^2 \quad (3)$$

over a set of controls  $g \in L_2([0, T]; U)$  where  $R \in \mathcal{L}(\mathcal{H}, \mathcal{Z})$  and  $G \in \mathcal{L}(\mathcal{H}, W)$  are bounded operators on suitable Hilbert spaces. In the context of control problem (3) relevant observations operators in (2) are the following:  $\mathcal{O} = R$  and  $\mathcal{O} = G$ .

**Remark 1.1.** Note that singular estimate (2) is automatically satisfied for analytic semigroups and unbounded control operators  $\mathcal{B}$  which are relatively bounded with respect to the generator  $\mathcal{A}$  [23, 22, 17, 31]. Thus SECS systems provide a proper generalization of control systems governed by analytic semigroups.

Our aim in this work is to show that the control system under consideration falls into the class of singular estimate control systems for which a satisfactory Riccati theory is available. In particular, we will show that the system satisfies singular estimate condition (2) with  $\gamma = 1/4 + \epsilon$  and the conditions laid out in [32] allow for an application of the results on existence, regularity of optimal control and, most importantly, feedback characterization of the control via solutions to a Riccati equation. The optimal control along with the feedback gain operator, while well defined and bounded for transient times, become singular at the terminal time. This singularity is quantified by the algebraic blow up rate of the order  $(T - t)^{-1/4 - \epsilon}$ .

## 2. The control problem

### 2.1. Formulation

Let  $\Omega \in \mathbb{R}^3$  be a bounded domain with an interior region  $\Omega_s$  and an exterior region  $\Omega_f$ . The boundary  $\Gamma_f$  is the outer boundary of the domain  $\Omega$  while  $\Gamma_s$  is the boundary of the region  $\Omega_s$  which also borders the exterior region  $\Omega_f$  and where the interaction of the two systems take place. Let  $u$  be a function defined on  $\Omega_f$  representing the velocity of the fluid while the scalar function  $p$  represents the pressure. Additionally, let  $w, w_t$  be the displacement and velocity functions of the solid  $\Omega_s$ . We also denote by  $\nu$  the unit outward normal vector with respect to the domain  $\Omega_s$ . The boundary-interface control is represented by  $g \in L_2([0, T]; L_2(\Gamma_s))$  and is active on the boundary  $\Gamma_s$ . We work under the assumption of small but rapid

oscillations of the solid, hence the interface  $\Gamma_s$  is assumed static see [26, 37, 15] for more modeling details.

Given control  $g \in L_2([0, T]; L_2(\Gamma_s))$ ,  $(u, w, w_t, p)$  satisfy the following system:

$$\left\{ \begin{array}{ll} u_t - \Delta u + Lu + \nabla p = 0 & Q_f \equiv \Omega_f \times [0, T] \\ \operatorname{div} u = 0 & Q_f \equiv \Omega_f \times [0, T] \\ w_{tt} - \operatorname{div} \sigma(w) = 0 & Q_s \equiv \Omega_s \times [0, T] \\ u(0, \cdot) = u_0 & \Omega_f \\ w(0, \cdot) = w_0, w_t(0, \cdot) = w_1 & \Omega_s \\ w_t = u & \Sigma_s \equiv \Gamma_s \times [0, T] \\ u = 0 & \Sigma_f \equiv \Gamma_f \times [0, T] \\ \sigma(w) \cdot \nu = \epsilon(u) \cdot \nu - p \nu - g & \Sigma_s \equiv \Gamma_s \times [0, T] \end{array} \right. \quad (4)$$

where the elastic stress tensor  $\sigma$  and the strain tensor  $\epsilon$ , respectively, are given by

$$\sigma_{ij}(u) = \lambda \sum_{k=1}^{k=3} \epsilon_{kk}(u) \delta_{ij} + 2\mu \epsilon_{ij}(u), \quad \lambda, \mu > 0, \quad \text{and} \quad \epsilon_{ij}(u) = \frac{1}{2} \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right).$$

The term  $Lu$  is a linearization of the convective term in Navier Stokes  $(u \cdot \nabla)u$  and is defined as

$$Lu = (\nabla v)u + (v \nabla)u \quad (5)$$

where  $v$  is a time-independent smooth vector function  $\in [C^\infty(\Omega_f)]^n$  with the property  $\operatorname{div} v = 0$ .

**Notation** Throughout the paper  $\mathcal{H} \equiv H \times H^1(\Omega_s) \times L_2(\Omega_s)$  where

$$H \equiv \{u \in L_2(\Omega_f) : \operatorname{div} u = 0, u \cdot \nu|_{\Gamma_f} = 0\}$$

will denote the energy space for the system. Note that all Sobolev spaces  $H^s$  and  $L_2$  spaces pertaining to  $u$  and  $w$  are in fact  $(H^s)^n$ ,  $(L_2)^n$ ,  $n = 2, 3$  and only for simplicity we omit the exponent  $n$ .

In addition we will use the following notation:

$$V \equiv \{v \in H^1(\Omega_f) : \operatorname{div} v = 0, u|_{\Gamma_f} = 0\}$$

$$(u, v) = \int_{\Omega} uv d\Omega, \quad \langle u, v \rangle = \int_{\Gamma_s} uv d\Gamma_s, \quad D_i = \frac{\partial}{\partial x_i}.$$

The space  $V$  is topologized with respect to the inner product given by:

$$(u, v)_{1,f} \equiv \int_{\Omega_f} \epsilon(u) \epsilon(v) d\Omega_f.$$

We denote the induced norm by  $| \cdot |_{1,\Omega_f}$ , and that is equivalent to the usual  $H^1(\Omega_f)$  norm via Korn's inequality and Poincaré's inequality

$$|u|_{1,\Omega_f} = [\int_{\Omega_f} |\epsilon u|^2 d\Omega_f]^{1/2}.$$

$H^1(\Omega_s)$  is topologized with respect to the inner product given by

$$(w, z)_{1,s} \equiv \int_{\Omega_s} wz + \int_{\Omega_s} \sigma(w) \epsilon(z).$$

We denote by  $|\cdot|_{1,\Omega_s}$  the induced norm by inner product above

$$|w|_{1,\Omega_s}^2 = \int_{\Omega_s} \sigma(w) \epsilon(w) d\Omega_s + |w|_{0,\Omega_s}^2.$$

This is equivalent to the usual  $H^1(\Omega_s)$  norm by Korn's inequality.

## 2.2. Weak solutions

We consider a weak solution to the system (4) defined to be  $(u, w, w_t) \in C([0, T]; H \times H^1(\Omega_s) \times L_2(\Omega_s)) = C([0, T]; \mathcal{H})$  and such that

- $(u_0, w_0, w_1) \in H \times H^1(\Omega_s) \times L_2(\Omega_s)$
- $u \in L_2([0, T]; V)$
- $\sigma(w) \cdot \nu \in L_2([0, T]; H^{-1/2}(\Gamma_s))$  and  $w_t|_{\Gamma_s} = u|_{\Gamma_s} \in L_2([0, T]; H^{1/2}(\Gamma_s))$
- The following variational equality holds a.e. in  $t \in (0, T)$

$$\begin{cases} (u_t, \phi)_{\Omega_f} + (\epsilon(u), \epsilon(\phi))_{\Omega_f} + (Lu, \phi)_{\Omega_f} + \langle \sigma(w) \cdot \nu + g, \phi \rangle = 0 \\ (w_{tt}, \psi)_{\Omega_s} + (\sigma(w), \epsilon(\psi))_{\Omega_s} - \langle \sigma(w) \cdot \nu, \psi \rangle = 0 \end{cases} \quad (6)$$

for all test functions  $\phi \in V$  and  $\psi \in H^1(\Omega_s)$ .

The following well-posedness result has been established in [15]

**Theorem 2.1.** *Let  $g \in L_2([0, T], H^{-1/2}(\Gamma_s))$  and  $(u_0, w_0, w_1) \in \mathcal{H}$  there exists a unique weak solution  $(u, w, w_t) \in C([0, T], \mathcal{H})$  and such that*

$$\begin{aligned} & |u(t)|_{0,\Omega_f}^2 + |w(t)|_{1,\Omega_s}^2 + |w_t(t)|_{0,\Omega_s}^2 + \int_0^t [|u(s)|_{1,\Omega_f}^2 + |\sigma(w) \cdot \nu|_{-1/2,\Gamma_s}^2] ds \\ & \leq Ce^{\omega t} [|u(0)|_{0,\Omega_f}^2 + |w(0)|_{1,\Omega_s}^2 + |w_t(0)|_{0,\Omega_s}^2 + |g|_{L_2([0,T];H^{-1/2}(\Gamma_s))}^2]. \end{aligned} \quad (7)$$

**Remark 2.1.** Note that the definition of weak solutions postulates trace regularity  $\sigma(w) \cdot \nu \in L_2([0, T]; H^{-1/2}(\Gamma_s))$  which does not follow from the interior regularity of solutions. This is, in fact, "hidden regularity" on the boundary. Thus, the proof of existence of weak solutions must involve obtaining the information on boundary regularity of the normal stresses. For this step methods of microlocal analysis were used (see [15]).

## 2.3. Control objective

The control problem to be considered is of Bolza type: we wish to minimize the following functional over all  $g \in L_2([0, T]; \Gamma_s)$ .

$$J(u, g) = \int_0^T [|g(t)|_{L_2(\Gamma_s)}^2 + |\mathcal{R}u(t)|_{L_2(\Omega_f)}^2] ds + |u(T, \cdot)|_{L_2(\Omega_f)}^2 \quad (8)$$

where  $\mathcal{R} \in \mathcal{L}(L_2(\Omega_f))$ .

**Remark 2.2.** The structure of the functional cost can be generalised. Indeed, it is not necessary that the terminal (Bolza) observation operator is equal to the identity. However, it is the identity that epitomizes the main intricacy of the problem. It would be much simpler to study the problem with compact or "smoothing" terminal observation. In such cases one would not have singularity in the feedback representation.

### 3. Semigroup formulation

In order to be able to apply available abstract results, we represent the solution to (4) via semigroup framework as an abstract equation of the form:

$$y_t = \mathcal{A}_L y + \mathcal{B}g, \quad y_0 \in \mathcal{H} \quad (9)$$

where

$$\mathcal{A}_L = \begin{pmatrix} A-L & AN\sigma(\cdot)\nu & 0 \\ 0 & 0 & I \\ 0 & \operatorname{div} \sigma & 0 \end{pmatrix}, \quad \mathcal{B} = \begin{pmatrix} AN \\ 0 \\ 0 \end{pmatrix}. \quad (10)$$

Here  $A : V \rightarrow V'$  is defined by

$$(Au, \phi) = -(\epsilon(u), \epsilon(\phi)), \quad \forall \phi \in V \quad (11)$$

and the Neumann map  $N : L_2(\Gamma_s) \rightarrow H$  defined by

$$Ng = h \Leftrightarrow \{(\epsilon(h), \epsilon(\phi)) = \langle g, \phi \rangle, \forall \phi \in V\}. \quad (12)$$

It follows immediately from Lax-Milgram Theorem that the map  $A \in \mathcal{L}(V \rightarrow V')$  and the map  $N$  enjoys the following regularity property:

$$N \in \mathcal{L}(H^{-1/2}(\Gamma_s) \rightarrow V \subset H^1(\Omega_f)). \quad (13)$$

This allows to consider the operator  $A$  (denoted by the same symbol) as acting on  $H$  with the domain  $D(A) \equiv \{u \in V; |(\epsilon(u), \epsilon(\phi))| \leq C|\phi|_H\}$ .  $A$  is self adjoint, negative and generates an analytic semigroup  $e^{At}$  on  $H$ . Fractional powers of  $-A$ , denoted by  $A^\alpha$  are then well defined, [38]. In particular

$$|A^\alpha e^{At}|_{\mathcal{L}(H)} \leq Ct^{-\alpha}, \quad 0 < t \leq 1. \quad (14)$$

In addition, the perturbation of  $A - L$  still generates an analytic semigroup since  $L$  is compact from  $\mathcal{D}(A)$  to  $H$ , see [21]. Let the space  $X$  be the trace space corresponding to  $V$  and  $X'$  its dual (with respect to  $L_2$  inner product). Elements of  $X$  are defined as  $X \equiv \{z = \phi|_{\Gamma_s}, \phi \in V\}$ . As a consequence, elements of  $X$  are in  $H^{1/2}(\Gamma_s)$  and they satisfy boundary compatibility relation  $\int_{\Gamma_s} g \cdot \nu = 0$ .

It was shown in [15] that the operator  $\mathcal{A}_L$  given by (10) and defined on  $\mathcal{D}(\mathcal{A}_L) \subset \mathcal{H} \rightarrow \mathcal{H}$  with

$$\begin{aligned} \mathcal{D}(\mathcal{A}_L) = \{y \in \mathcal{H} : u \in V, (A - L)u + AN\sigma(w).\nu \in H; z \in H^1(\Omega_s), \\ \operatorname{div} \sigma(w) \in L_2(\Omega_s); z|_{\Gamma_s} = u|_{\Gamma_s}\} \end{aligned}$$

generates a strongly continuous semigroup  $e^{\mathcal{A}_L t} \in \mathcal{L}(\mathcal{H})$ .

**Remark 3.1.** *There is another approach available in recent literature that leads to semigroup solutions of fluid structure interaction models. This has been pursued in [5, 6] (also references therein) where the “generator” is explicitly constructed via non-local Green maps. The main difference between these two approaches is that our framework has explicit boundary conditions involving strain stresses on the boundary, which are defined for weak solutions (see Definition 6). In contrast, the approach taken in [4, 5] does not exhibit boundary traces at the level of weak solutions. These become apparent only for strong solutions. The analysis in this paper critically relies on the notion of boundary traces defined for finite energy solutions.*

## 4. Main results

We first recall an abstract result from [32], which provides Riccati theory pertinent to SECS control systems.

**Theorem 4.1.** *Let  $U$ ,  $Y$ ,  $Z$  and  $W$  be given Hilbert spaces. Spaces  $U$  and  $Y$  denote, respectively, control and state spaces while  $Z$  and  $W$  are observation spaces. We consider the dynamics governed by the state equation with a state  $y(t) \in Y$  and control  $g(t) \in U$ :*

$$y_t = \mathcal{A}y + \mathcal{B}g; \text{ on } [\mathcal{D}(\mathcal{A}^*)]'; \quad y(s) = y_s \in Y. \quad (15)$$

*The control problem is to minimize  $J(g, y, s, y_s)$  subject to the state equation (4) over all  $g \in L_2([s, T]; U)$*

$$J(g, y, s, y_s) = \int_s^T |Ry(t)|_Z^2 + |g(t)|_U^2 dt + |Gy(T)|_W^2 \quad (16)$$

*under the following assumptions:*

- (a)  $\mathcal{A}$  is a generator of a strongly continuous semigroup denoted by  $e^{\mathcal{A}t}$  on the Hilbert space  $Y$ .
- (b) The control operator  $\mathcal{B}$  is a linear operator from  $U \rightarrow [\mathcal{D}(\mathcal{A}^*)]'$ , satisfying the condition  $R(\lambda, \mathcal{A})\mathcal{B} \in \mathcal{L}(U, Y)$ , for some  $\lambda \in \rho(\mathcal{A})$  where  $R(\lambda, \mathcal{A})$  is the resolvent of  $\mathcal{A}$  and  $\rho(\mathcal{A})$  is the resolvent set.
- (c) Singular Estimate Control condition: There exists  $\gamma < 1$  and a constant  $C > 0$  such that  $|\Re^{\mathcal{A}t}\mathcal{B}u|_Z \leq \frac{C}{t^\gamma}|u|_U$  and  $|Ge^{\mathcal{A}t}\mathcal{B}u|_W \leq \frac{C}{t^\gamma}|u|_U$  for all  $0 < t \leq 1$ .
- (d)  $R \in \mathcal{L}(Y, Z)$  and the operator:  $G \in \mathcal{L}(Y, W)$  is such that the operator  $GL_T : L_2([0, T]; U) \rightarrow W$  is closeable where  $L_T$  is the control to state map at time  $T$ .

*Then for any initial state  $y_s \in Y$  there exists a unique optimal control*

$$g^0(t, s, y_s) \in L_2([s, T]; U)$$

*and optimal trajectory*

$$y^0(t, s, y_s) \in C([0, T], D(\mathcal{A}^*))' \quad \text{with} \quad Ry^0(t, s, y_s) \in L_2([0, T], Z)$$

*such that*

$$J(g^0, y^0, s, y_s) = \min_{u \in L_2([s, T], U)} J(g, y(g), s, y_s).$$

*Moreover, there exists a selfadjoint positive operator  $P(t) \in \mathcal{L}(Y)$ ,  $t \in [0, T]$  such that  $(P(t)x, x)_Y = J(g^0, y^0, t, x)$ . In addition, the following properties hold:*

- (i) *The optimal control  $g^0(t)$  is continuous on  $[s, T]$  but has a singularity of order gamma at the terminal time. More specifically the following estimate holds*

$$|g^0(t, s, y_s)|_U \leq \frac{C}{(T-t)^\gamma} |y_s|_Y, \quad s \leq t < T. \quad (17)$$

- (ii) The optimal output  $y^0(t)$  is continuous on  $[s, T]$  when  $\gamma < 1/2$  with values in the observation space  $Z$ , but has a singularity of order  $2\gamma - 1$  at the terminal time when  $\gamma \geq 1/2$ . The following estimate holds:

$$|Ry^0(t, s, y_s)|_Z \leq \frac{C}{(T-t)^{2\gamma-1+\epsilon}} |y_s|_Y, \quad s \leq t < T. \quad (18)$$

- (iii)  $P(t)$  is continuous on  $[0, T]$  and  $P(t) \in \mathcal{L}(Y, L_\infty([0, T]; Y))$ .  
(iv)  $\mathcal{B}^*P(t)$  exhibits the following singularity

$$|\mathcal{B}^*P(t)x|_U \leq \frac{C|x|_Y}{(T-t)^\gamma}, \quad 0 \leq t < T. \quad (19)$$

- (v)  $g^0(t, s, y_s) = -\mathcal{B}^*P(t)y^0(t, s, y_s), \quad s \leq t < T.$  (20)

- (vi)  $P(t)$  satisfies the Riccati Differential equation with  $t < T$ ,  $x, y \in \mathcal{D}(\mathcal{A})$

$$\begin{aligned} & \langle P_t x, y \rangle_Y + \langle \mathcal{A}^*P(t)x, y \rangle_Y + \langle P(t)\mathcal{A}x, y \rangle_Y + \langle Rx, Ry \rangle_Z \\ &= \langle \mathcal{B}^*P(t)x, \mathcal{B}^*P(t)y \rangle_U \end{aligned} \quad (21)$$

$$\lim_{t \rightarrow T} P(t)x = G^*Gx \quad \forall x \in Y. \quad (22)$$

- (vii) The solution of the Riccati equation above is unique within the class of positive and self adjoint operators such that (19) holds with  $\gamma < \frac{1}{2}$ .

**Remark 4.1.** Theorem 4.1 generalizes to SECS systems results that are available in the analytic case [1, 23, 31].

The main result of this paper is the following Theorem pertaining to the model in (4) with the functional cost given by (8).

**Theorem 4.2.** In reference to the model in (4) and the control problem in (8), for every initial condition  $y_0 = [u_0, w_0, w_1] \in \mathcal{H}$ , there exists a unique optimal control  $g^0(t, \cdot) \in L_2([0, T]; \Gamma_s)$  and a corresponding optimal state

$$y^0(t, \cdot) = [u^0(t, \cdot), w^0(t, \cdot), w_t^0(t, \cdot)] \in C([0, T]; H \times H^1(\Omega_s) \times L_2(\Omega_s)) \quad (23)$$

such that  $J(g^0, y^0) = \min_{g \in L_2([0, T]; \Gamma_s)} J(g, y)$ . Moreover,

- the optimal control  $g^0$  satisfies singular estimate given by (17) with blow up rate  $\gamma = 1/4 + \epsilon$ , where  $\epsilon$  is positive and can be taken arbitrarily small.
- There exists a positive selfadjoint  $P(t) \in \mathcal{L}(\mathcal{H})$  such that

$$J(g^0, y^0) = (P(0)y_0, y_0)_\mathcal{H}.$$

- In addition with  $\mathcal{B}$  given in (10), the feedback-gain operator  $\mathcal{B}^*P(t) \in \mathcal{L}(\mathcal{H} \rightarrow L_2(\Gamma_s))$  for all  $0 \leq t < T$  and at the terminal point  $t = T$  blows up with the rate given by

$$|\mathcal{B}^*P(t)y|_{L_2(\Gamma_s)} \leq \frac{C|y|_\mathcal{H}}{(T-t)^{1/4+\epsilon}}.$$

In addition, the optimal feedback synthesis (20) holds.

- The operator  $P(t)$  is a unique solution of Riccati Differential Equation satisfied for  $t < T$  with ,  $x = (x_1, x_2, x_3), y = (y_1, y_2, y_3) \in \mathcal{D}(\mathcal{A}_L)$

$$\begin{aligned} & \langle P_t x, y \rangle_{\mathcal{H}} + \langle \mathcal{A}_L^* P(t)x, y \rangle_{\mathcal{H}} + \langle P(t)\mathcal{A}_L x, y \rangle_{\mathcal{H}} + (\mathcal{R}x_1, \mathcal{R}y_1)_{\Omega_f} \\ &= \langle \mathcal{B}^* P(t)x, \mathcal{B}^* P(t)y \rangle_{\Gamma_s} \end{aligned} \quad (24)$$

with the terminal condition

$$\lim_{t \rightarrow T} (P(t)x)_1 = x_1 \text{ in } H \quad \forall x \in \mathcal{H} \quad (25)$$

and operators  $\mathcal{A}_L, \mathcal{B}$  defined in (10).

The main mathematical difficulty of the problem studied is due to the mismatch of parabolic and hyperbolic regularity occurring at the interface. The strategy pursued in this work is to transport, via the interface, maximal (abstract) parabolic regularity [28, 17] resulting from fluid component onto the wave dynamics. The generalization of so-called “hidden” [30] regularity of the waves boundary traces plays a pivotal role in this transfer.

The remainder of this paper is devoted to the discussion of the proof of Theorem 4.2. The proof is based on the following two main technical ingredients: (i) sharp hyperbolic-like regularity theory for interface traces of solutions to fluid structure interaction, and (ii) abstract maximal parabolic regularity [28, 17] applied to boundary value problem driven by the fluid component.

## 5. Sketch of the Proof of Theorem 4.2

We shall provide below main steps in the proof of the theorem, while full account of technical details is given in [33].

The proof of Theorem 4.2 is based on application of the abstract result given by Theorem 4.1. To accomplish this we need to verify the assumptions imposed by that theorem. This, in turn, involves verification of several properties of the generator of the semigroup operators  $\mathcal{A}$  and of the control operator  $\mathcal{B}$ .

**Generator  $\mathcal{A}_L$**  It was shown in [15], that  $\mathcal{A}_L$  generates a  $c_0$  semigroup if  $L : V \rightarrow V'$  is locally Lipschitz and satisfies the condition for any  $\delta > 0$ :

$$|(Lu, u)| \leq \delta |u|_V^2 + C_\delta |u|_H^2. \quad (26)$$

The linear operator  $Lu = (\nabla v)u + (v.\nabla)u$  is bounded when acting from  $V \rightarrow V'$  since  $v$  is smooth, and indeed satisfies the condition above. This is sufficient [15] for the establishment of  $c_0$  semigroup solutions for the system, which are generated by the operator  $\mathcal{A}_L$  whose action is given by (10) and the domain given by:

$$\begin{aligned} \mathcal{D}(\mathcal{A}_L) = \{y \in \mathcal{H} : u \in V, (A - L)u + AN\sigma(w).\nu \in H; z \in H^1(\Omega_s), \\ \text{div } \sigma(w) \in L_2(\Omega_s); z|_{\Gamma_s} = u|_{\Gamma_s}\}. \end{aligned}$$

**Remark 5.1.** We note that the domain  $\mathcal{D}(\mathcal{A}_L)$  is not compact in  $\mathcal{H}$ . This has been noticed in [5] in the context of studying strong stability of uncontrolled model [6, 5]

**Control operator  $\mathcal{B}$ .** The operator  $\mathcal{B}$ , given in (10), is unbounded when acting from  $L_2(\Gamma_s)$  to  $\mathcal{H}$  since  $AN$  is an unbounded operator from  $L_2(\Gamma_s) \rightarrow H$  though bounded from  $L_2(\Gamma_s) \rightarrow V'$ .

We begin by asserting that the control operator  $B$  is relatively bounded with respect to the generator.

**Proposition 5.1.** *There exists  $\omega > 0$  such that  $R(\lambda, \mathcal{A}_L)\mathcal{B} \in \mathcal{L}(\mathcal{H})$ , where  $\lambda > \omega$ .*

*Proof.* Writing  $(\mathcal{A}_L - \lambda)Y = f = (f_1, f_2, f_3)$  leads to:

$$\begin{aligned} Au + AN\sigma(w).\nu + Lu - \lambda u &= f_1 \\ z - \lambda w &= f_2 \\ \operatorname{div} \sigma(w) - \lambda z &= f_3 \\ z|_{\Gamma_s} = u|_{\Gamma_s}, u|_{\Gamma_f} &= 0. \end{aligned}$$

Since  $\mathcal{A}_L$  generates a  $c_0$  semigroup, there exists  $\omega > 0$  such that  $\mathcal{A}_L - \lambda I$  is injective for all  $\lambda > \omega$ . Setting  $f = \mathcal{B}g = (ANg, 0, 0)$  gives:

$$\begin{aligned} -(\epsilon(u), \epsilon(\phi)_f - \langle \sigma(w).\nu, \phi \rangle - (Lu, \phi)_f - \lambda(u, \phi)_f) &= -\langle g, \phi \rangle, \forall \phi \in V \\ -(\sigma(w), \epsilon(\psi))_s + \langle \sigma(w).\nu, \psi \rangle - \lambda(w, \psi)_s &= 0, \forall \psi \in H^1(\Omega_s). \end{aligned} \quad (27)$$

Taking  $\phi = u, \psi = \lambda w$  and using the fact  $u|_{\Gamma_s} = z|_{\Gamma_s} = \lambda w|_{\Gamma_s}$  we add the two equations to get

$$|\epsilon(u)|_{0,\Omega_f}^2 + \lambda|u|_{0,\Omega_f}^2 + \lambda(\sigma(w), \epsilon(w))_s + \lambda^2|w|_{0,\Omega_s}^2 + (Lu, u)_f = \langle g, u \rangle. \quad (28)$$

Using inequality (5) we obtain

$$|u|_V^2 + |w|_{1,\Omega_s}^2 - \delta|u|_V^2 + (\lambda - C_\delta)|u|_H^2 \leq K|g|_{H^{-1/2}(\Gamma_s)}^2 + 1/2|u|_V^2.$$

Choose  $\delta$  so that  $1/2 - \delta > 0$  and then  $\lambda > 0$  so that  $\lambda - C_\delta > 0$  and thus:

$$|u|_V^2 + |w|_{1,\Omega_s}^2 \leq C|g|_{H^{-1/2}(\Gamma_s)}^2.$$

Since  $z = \lambda w$  we also obtain the estimate

$$|z|_{1,\Omega_s}^2 \leq C\lambda^2|g|_{H^{-1/2}(\Gamma_s)}^2. \quad \square$$

The most critical, and also the most technical property is Singular Estimate condition satisfied for the pair  $\mathcal{A}, \mathcal{B}$ .

In order to state the result we define a scale of Hilbert spaces parameterized by the parameter  $\alpha \geq 0$ :  $\mathcal{H}_{-\alpha} \equiv H \times H^{1-\alpha}(\Omega_s) \times H^{-\alpha}(\Omega_s)$ . Note that with the above notation  $\mathcal{H} = \mathcal{H}_0$ .

**Theorem 5.2.** *The semigroup  $e^{\mathcal{A}_L t}$  and control operator  $\mathcal{B}$  satisfy the following Singular Estimate: (SE) for every  $g \in L_2(\Gamma_s)$  and  $t \leq T_0$ , and  $\alpha > 0$ .*

$$|e^{\mathcal{A}_L t}\mathcal{B}g|_{\mathcal{H}_{-\alpha}} \leq \frac{C}{t^{1/4+\epsilon}}|g|_{L_2(\Gamma_s)}.$$

The sketch of the proof of Theorem 5.2 is given in Section 6.

**Completion of the proof of Theorem 4.2** Assuming validity of Theorem 5.2 the proof of the Theorem 4.2 is completed as follows.

The conclusion of the Theorem 4.2 follows from Theorem 4.1, Theorem 5.2 and regularity estimate in (7).

Assumptions (a) and (b) of Theorem 4.1 are satisfied on the strength of the results presented in Section 4, including Proposition 5.1. Assumption (d) follows from the fact that  $R(\lambda, \mathcal{A}_L)\mathcal{B} \in \mathcal{L}(\mathcal{H})$  and from the structure of the operator  $G$  (projection). The most critical assumption is singular estimate in part (c). The aforementioned follows from Theorem 5.2. To see this we first note that Theorem 5.2 implies that with  $G = (I, 0, 0)$  we have that

$$|Ge^{\mathcal{A}_L t}\mathcal{B}g|_W = |Ge^{\mathcal{A}_L t}\mathcal{B}g|_{0, \Omega_f} \leq |e^{\mathcal{A}_L t}\mathcal{B}g|_{\mathcal{H}_{-\alpha}} \leq \frac{C}{t^{1/4+\epsilon}} |g|_{L_2(\Gamma_s)}$$

Similarly for all  $\alpha < 1$

$$|Re^{\mathcal{A}_L t}\mathcal{B}g|_{\mathcal{H}} \leq |e^{\mathcal{A}_L t}\mathcal{B}g|_{\mathcal{H}_{-\alpha}} \leq \frac{C}{t^{1/4+\epsilon}} |g|_{L_2(\Gamma_s)}$$

Thus, the hypothesis (c) in Theorem 4.1 is satisfied with  $\gamma = 1/4 + \epsilon$ .

Finally, the regularity of the trajectories is stronger than postulated in theorem 4.1, but this follows from the estimate in (7). The proof of Theorem 4.2 is completed.

## 6. Sketch of the proof of Theorem 5.2

### 6.1. Supporting Propositions

In order to carry the proof of Theorem 5.2 we need several auxiliary results. These are given below.

**6.1.1. Properties of the Neumann map.** In what follows we shall establish additional properties of the map  $N$  defined in (12) and a useful PDE interpretation of solutions to (6). To this end we define for each  $(u, p) \in V \times L_2(\Omega_f)$  the fluid tensor  $\mathcal{T}(u, p)$  given by

$$\mathcal{T}(u, p) \equiv \epsilon(u) - pI.$$

**Proposition 6.1.** *Let  $g \in H^{-1/2}(\Gamma_s)$ . There exists  $p \in L_2(\Omega_f)$  such that  $h \equiv Ng$  satisfies distributionally*

$$\operatorname{div}(\mathcal{T}(h, p)) = 0 \quad \Omega_f \tag{29}$$

$$\operatorname{div} h = 0 \quad \Omega_f \tag{30}$$

$$\mathcal{T}(h, p) \cdot \nu = g \quad \text{in } H^{-1/2}(\Gamma_s) \tag{31}$$

$$h = 0 \quad \Gamma_f. \tag{32}$$

*Proof.* The form  $(\epsilon(h), \epsilon(\phi))$  is  $V$  elliptic and continuous on  $V \times V$  and hence by Lax-Milgram given  $g \in H^{-1/2}(\Gamma_s)$  there exists  $h \in V$  such that:  $(\epsilon(h), \epsilon(\phi)) = \langle g, \phi \rangle_{\Gamma_s}$  for every  $\phi \in V$ . Hence, the map  $N$  is well defined as a bounded linear operator from  $H^{-1/2}(\Gamma_s) \rightarrow H^1(\Omega_f) \cap H = V$ .

Let  $Ng = h$ , for some  $g \in H^{-1/2}(\Gamma_s)$  then

$$(\epsilon(h), \epsilon(\phi))_f = \langle g, \phi \rangle$$

for all  $\phi \in V$ . Consider all  $\phi \in H_0^1(\Omega_f) \cap V$  thus

$$(\operatorname{div} \epsilon(h), \phi)_f = 0.$$

This implies by De Rham's theorem that there exists  $p \in L_2(\Omega_f)$  such that  $\operatorname{div} \mathcal{T}(h, p) = 0$  in the sense of distributions. Since  $\mathcal{T}(h, p) \in L_2(\Omega_f)$ , the distributional derivative  $\operatorname{div} \mathcal{T}(h, p)$  coincides with a function equal to zero. Hence, we also have  $\operatorname{div} \mathcal{T}(h, p) = 0$  a.e., which relation reconstructs the equation in Proposition 4. This also shows (by Divergence Theorem) that for all  $\phi \in V$

$$\begin{aligned} \langle \mathcal{T}(h, p) \cdot \nu, \phi \rangle_{\Gamma_s} &= (\epsilon(h) - pI, \epsilon(\phi))_f \\ &\leq |\phi|_V [|h|_V + |p|_{L_2(\Omega_f)}] \end{aligned}$$

consequently  $\mathcal{T}(h, p) \cdot \nu \in X'$ . On the other hand, definition of the map  $N$  implies for every  $\phi \in V$  that

$$\langle g, \phi \rangle_{\Gamma_s} = (\epsilon(h), \epsilon(\phi))_f = (\epsilon(h), \epsilon(\phi))_f - (pI, \epsilon\phi)_f = \langle \mathcal{T}(h, p) \cdot \nu, \phi \rangle_{\Gamma_s}.$$

Here we used the fact that  $(pI, \epsilon\phi)_f = (p, \operatorname{div} \phi)_f = 0$ . Hence

$$\langle g - \mathcal{T}(h, p) \cdot \nu, \phi \rangle_{\Gamma_s} = 0, \forall \phi \in V$$

and by the definition of  $X$ ,  $\langle g - \mathcal{T}(h, p) \cdot \nu, z \rangle = 0, \forall z \in X$ , so  $g - \mathcal{T}(h, p) \cdot \nu$  belongs to the normal cone in  $X$  which can be identified with  $\{\lambda = k\nu, k \in R\}$ . Redefining the pressure  $p$  by adding suitable constant gives the boundary conditions and the PDE form of the map  $N$  asserted in the proposition.  $\square$

Now, we turn to higher regularity of the Neumann map  $N$ . The analogous result is known in the case of Dirichlet boundary conditions [44]. However, in the case of Neumann boundary conditions, due to the presence of pressure on the boundary, the issue is more subtle. More specifically, we show that

**Proposition 6.2.** *Let  $g \in H^{1/2}(\Gamma_s)$  then  $Ng = h \in H^2(\Omega_f) \cap V$  and  $p \in H^1(\Omega_f)$ .*

*Proof.* To accomplish this, we follow the strategy of Agmon-Douglis-Nirenberg where it suffices to consider the PDE in the neighborhood of the boundary  $\Gamma_s$  (interior regularity is straightforward) which is accomplished via a partition of unity. We then differentiate in the tangential direction by introducing the tangential differential operator  $\mathcal{S}$  with respect to the boundary  $\Gamma_s$  to obtain the local problem in a collar neighborhood of the boundary  $\Gamma_s$ . The operator  $\mathcal{S} = \sum_{i=1}^n b_i(x) \frac{\partial}{\partial x_i}$  is a first-order operator (time-independent) with  $b_i$  smooth in  $\Omega$ , such that  $\mathcal{S}$  is tangent to  $\Gamma_s$ . Applying  $\mathcal{S}$  to the system and denoting by  $[D, \mathcal{S}]$  the commutator of  $\mathcal{S}$  with an operator  $D$ . Let  $\mathcal{S}h = \hat{h}$  and  $\mathcal{S}p = \hat{p}$  and applying  $\mathcal{S}$  to the system

$$\operatorname{div} \epsilon(\hat{h}) - \nabla \hat{p} = [\operatorname{div} \epsilon, \mathcal{S}]h - [\nabla, \mathcal{S}]p \quad \Omega_f \tag{33}$$

$$\epsilon(\hat{h}) \cdot \nu = \hat{g} + \hat{p}\nu + p\hat{\nu} + [\epsilon \cdot \nu, \mathcal{S}]h \quad \Gamma_s \tag{34}$$

$$\operatorname{div} \hat{h} \in L_2(\Omega_f). \tag{35}$$

Integrating against  $\hat{h}$  we obtain

$$-(\epsilon \hat{h}, \epsilon \hat{h}) - \langle \hat{g}, \hat{h} \rangle - \langle p \hat{v}, \hat{h} \rangle - \langle [\epsilon \cdot \nu, \mathcal{S}] h, \hat{h} \rangle + (\hat{p}, \operatorname{div} \hat{h}) = ([\operatorname{div} \epsilon, \mathcal{S}] h, \hat{h}) + ([\nabla, \mathcal{S}] p, \hat{h}).$$

Now,  $\hat{g} \in H^{-1/2}(\Gamma_s)$  and  $p|_{\Gamma_s} \in H^{-1/2}(\Gamma_s)$  while the commutator  $[\operatorname{div} \epsilon, \mathcal{S}]$  is a second-order differential operator and both  $[\nabla, \mathcal{S}]$ ,  $[\epsilon \cdot \nu, \mathcal{S}]$  are first-order differential operators. Hence, we estimate the  $V$  norm of  $\hat{h}$  as follows

$$\begin{aligned} |\epsilon \hat{h}|_{0,\Omega_f}^2 &\leq [|\hat{g}|_{-1/2,\Gamma_s} + |p|_{-1/2,\Gamma_s} + |h|_{-1/2,\Gamma_s}] |\hat{h}|_{1/2,\Gamma_s} \\ &\quad + |p|_{0,\Omega_f} |\mathcal{S} \operatorname{div} \hat{h}|_{0,\Omega_f} + |h|_{-1,\Omega_f} |\hat{h}|_{1,\Omega_f} + |p|_{-1,\Omega_f} |\hat{h}|_{1,\Omega_f}. \end{aligned} \quad (36)$$

Since  $h$  is divergence free  $\operatorname{div} \hat{h} = \Sigma_{i,j} D_i b_j D_j h_i$ , so  $\mathcal{S} \operatorname{div} \hat{h} = \nabla \hat{h}$  and

$$\begin{aligned} |\epsilon \hat{h}|_{0,\Omega_f}^2 &\leq \frac{C}{\delta} \left[ |\hat{g}|_{-1/2,\Gamma_s}^2 + |p|_{-1/2,\Gamma_s}^2 + |h|_{-1/2,\Gamma_s}^2 \right. \\ &\quad \left. + |p|_{0,\Omega_f}^2 + |h|_{-1,\Omega_f}^2 + |p|_{-1,\Omega_f}^2 \right] + \delta |\hat{h}|_{1,\Omega_f}. \end{aligned}$$

Finally, choosing  $\delta = 1/2$  and using the fact that the  $H^1$  norm of  $h$  and  $L_2$  norm of  $p$  are continuously dependent on the  $H^{1/2}(\Gamma_s)$  of the trace  $g$  by Proposition 6.1

$$|\hat{h}|_{1,\Omega_f}^2 \leq K |g|_{1/2,\Gamma_s}^2.$$

Hence,  $\hat{h} \in H^1(\Omega_f)$ . Let  $v = D_\tau h \equiv \nabla h \cdot \tau$  then  $v \in H^1(\Omega_f)$  and  $D_\tau v, D_\nu v \in L_2(\Omega_f)$ . To show  $h \in H^2(\Omega_f)$  it remains to show  $D_{\nu,\nu}^2 h \in L_2(\Omega_f)$ . Let  $z \equiv D_\nu h$ , and notice that  $\operatorname{div} z \in L_2(\Omega_f)$  which follows from  $h$  being divergence free.

Let  $\nu = (n_1, n_2, n_3)$  the unit normal vector to the boundary  $\Gamma_s$  while  $\tau$  and  $\kappa$  two linearly independent tangential unit vectors to the boundary  $\Gamma_s$  at a given point  $m \in \Gamma_s$ . Let  $q$  be any point in  $\Omega_f$  such that  $q$  has the same  $x$  and  $y$  coordinates as  $m$ . With  $\nu, \tau$  and  $\kappa$  at  $q$  defined to be those at  $m$ , we rewrite  $z$  at any such  $q$  as

$$z = (z \cdot \nu) \nu + (z \cdot \tau) \tau + (z \cdot \kappa) \kappa.$$

Choosing tangential and normal coordinates,  $\tau = \frac{1}{\sqrt{n_3^2 + n_1^2}}(n_3, 0, -n_1)$  and  $\kappa = \frac{1}{\sqrt{n_3^2 + n_1^2}}(-n_1 n_2, n_3^2 + n_1^2, -n_3 n_2)$ .

Since all the tangential derivatives  $D_\tau z, D_\kappa z \in L_2(\Omega_f)$  and  $\tau, \kappa$  are  $C^1$  functions while  $\operatorname{div} z \in L_2(\Omega_f)$ , one obtains

$$D_\nu(z_1 n_1) + D_\nu(z_2 n_2) + D_\nu(z_3 n_3) \in L_2(\Omega_f).$$

With  $\nu \in C^1$  and  $z \in L_2(\Omega_f)$  we then have

$$n_1 D_\nu z_1 + n_2 D_\nu z_2 + n_3 D_\nu z_3 \in L_2(\Omega_f). \quad (37)$$

On the other hand, since we know the tangential derivative of  $h$  is  $\in H^1(\Omega_f)$ , we also have  $D_\nu(z \cdot \tau) = D_{\nu\nu}(h \cdot \tau) \in L_2(\Omega_f)$  and similarly  $D_\nu(z \cdot \kappa) \in L_2(\Omega_f)$ . Thus:

$$\frac{n_3}{\sqrt{n_3^2 + n_1^2}} D_\nu z_1 - \frac{n_1}{\sqrt{n_3^2 + n_1^2}} D_\nu z_3 \in L_2(\Omega_f) \quad (38)$$

$$\frac{-n_1 n_2}{\sqrt{n_3^2 + n_1^2}} D_\nu z_1 + \sqrt{n_3^2 + n_1^2} D_\nu z_2 - \frac{n_3 n_2}{\sqrt{n_3^2 + n_1^2}} D_\nu z_3 \in L_2(\Omega_f). \quad (39)$$

These three equations 37, 38 and 39 produce a system  $MD_\nu z = b \in [L_2(\Omega_f)]^3$

$$M = \begin{pmatrix} \frac{n_1}{n_3} & n_2 & -\frac{n_3}{n_1} \\ -\frac{n_3}{\sqrt{n_3^2+n_1^2}} & 0 & -\frac{\sqrt{n_3^2+n_1^2}}{\sqrt{n_3^2+n_1^2}} \\ -\frac{n_1 n_2}{\sqrt{n_3^2+n_1^2}} & \sqrt{n_3^2+n_1^2} & -\frac{n_3 n_2}{\sqrt{n_3^2+n_1^2}} \end{pmatrix}.$$

The coefficients of the matrix are continuous and its determinant is 1, which gives the desired result of  $D_\nu z \in L_2(\Omega_f)$ .

Therefore,  $h \in H^2(\Omega_f) \cap V$ . It also follows that  $p \in H^1(\Omega_f)$  and  $\epsilon(h).\nu \in H^{1/2}(\Gamma_s)$ . In addition, the following estimate is implied:

$$|h|_{H^2(\Omega_f)} \leq C[|h|_{L_2(\Omega_f)} + |D_\tau h|_{H^1(\Omega_f)} + |D_\kappa h|_{H^1(\Omega_f)} + |D_\nu h|_{H^1(\Omega_f)}] \leq C|g|_{H^{1/2}(\Gamma_s)}.$$

We conclude that  $N$  is bounded from  $H^{1/2}(\Gamma_s) \rightarrow H^2(\Omega_f)$ .

By interpolation,  $N \in \mathcal{L}(H^s(\Gamma_s) \cap X' \rightarrow H^{s+3/2}(\Omega_f) \cap V)$ , for  $-1/2 \leq s \leq 1/2$ . The above regularity can be extended to a full range of  $s$ , but this will not be needed in this paper.  $\square$

Some further properties of  $N$ , relatively straightforward consequences of Proposition 6.1 and Proposition 6.2, are given below.

### Proposition 6.3.

1. The map  $N^*Au = -u|_{\Gamma_s}$ ,  $u \in V$  where the adjoint is computed with respect to  $L_2$  topology.
2.  $N \in \mathcal{L}(L_2(\Gamma_s) \rightarrow \mathcal{D}(A^{3/4-\epsilon})) \cap \mathcal{L}(H^{-1/2}(\Gamma_s) \rightarrow \mathcal{D}(A^{1/2}))$  for all  $\epsilon > 0$ .

**Remark 6.1.** We note that the method of the proof of Proposition 6.2 also leads to higher regularity of the map  $A^{-1}$ . Indeed, one obtains  $A^{-1} : H \rightarrow H^2(\Omega_f) \cap V$  (see [14]). The above regularity, along with interpolation, allow us to identify domains of fractional powers of  $A$  as

$$\begin{aligned} D(A^\theta) &\sim H^{2\theta}(\Omega_f), \quad 0 \leq \theta < 3/4 \\ D(A^\theta) &\subset H^{2\theta}(\Omega_f), \quad \theta \in [0, 1]. \end{aligned} \tag{40}$$

## 6.2. Proof of the Singular Estimate

**6.2.1. Supporting Lemmas.** We begin with the following regularity result established for boundary traces of dynamic wave equation [15].

**Lemma 6.4.** Let  $w_0, w_1 \in H^{\alpha+1}(\Omega_s) \times H^\alpha(\Omega_s)$  with  $0 \leq \alpha \leq 1/4$  and let  $f \in L_2([0, T]; H^{1/2}(\Gamma_s))$  and  $w$  the solution to the wave equation

$$\begin{cases} w_{tt} - \operatorname{div} \sigma(w) = 0 & Q_s \equiv \Omega_s \times [0, T] \\ w(0, \cdot) = w_0, w_t(0, \cdot) = w_1 & \Omega_s \\ w_t = f & \Sigma_s \equiv \Gamma_s \times [0, T]. \end{cases} \tag{41}$$

Then  $w$  can be decomposed into  $w_1 + w_2$  such that  $\sigma(w_1).\nu \in C([0, T]; H^{-1/2}(\Gamma_s))$  and  $\sigma(w_2).\nu \in L_2(\Sigma_s) = L_2([0, T] \times \Gamma_s)$ . If further  $f \in H^\alpha(\Sigma_s)$  then  $\sigma(w_2).\nu \in$

$H^\alpha(\Sigma_s)$ . Moreover, we have the following estimates

$$|\sigma(w_1) \cdot \nu|_{C([0,T];H^{-1/2}(\Gamma_s))}^2 \leq K[|w_0|_{1,\Omega_s}^2 + |w_1|_{0,\Omega_s}^2 + |f|_{L_2([0,T];H^{1/2}(\Gamma_s))}^2] \quad (42)$$

$$|\sigma(w_2) \cdot \nu|_{H^\alpha(\Sigma_s)}^2 \leq K[|w_0|_{1+\alpha,\Omega_s}^2 + |w_1|_{\alpha,\Omega_s}^2 + |f|_{H^\alpha(\Sigma_s)}^2]. \quad (43)$$

We next state the improved boundary regularity for the velocity field  $u$ .

**Lemma 6.5.** *Consider the uncontrolled system (6) with  $g = 0$ . If in addition the initial condition  $[u_0, w_0, w_1] \in L_2(\Omega_f) \times H^{1+\alpha}(\Omega_s) \times H^\alpha(\Omega_s)$  for  $0 < \alpha < 1/4$ , then  $u|_{\Gamma_s} \in H^\alpha(\Sigma_s)$  and the following estimate holds*

$$|u|_{H^\alpha(\Sigma_s)}^2 \leq C_T(|u_0|_{0,\Omega_f}^2 + |w_0|_{1+\alpha,\Omega_s}^2 + |w_1|_{\alpha,\Omega_s}^2). \quad (44)$$

The proof relies on the decomposition given by Lemma 6.4, Proposition 6.3 and abstract parabolic maximal regularity methods from [28]. The details are given in [33].

### 6.2.2. Proper proof of Theorem 5.2.

*Proof.* It is equivalent to prove the following estimate for every  $y_0 = [u_0, w_0, w_1] \in H \times H^{1+\alpha}(\Omega_s) \times H^\alpha(\Omega_s) = \mathcal{H}_\alpha$ :

$$|\mathcal{B}^* e^{\mathcal{A}_L^* t} \begin{pmatrix} u_0 \\ w_0 \\ w_1 \end{pmatrix}|_{\mathcal{H}} \leq \frac{C}{t^{1/4+\epsilon}} |y_0|_{H \times H^{1+\alpha}(\Omega_s) \times H^\alpha(\Omega_s)} = \frac{C}{t^{1/4+\epsilon}} |y_0|_{\mathcal{H}_\alpha}. \quad (45)$$

This term represents the solution  $[\hat{u}, \hat{w}, \hat{w}_t]$  to the adjoint system of (6), when the initial condition is  $u_0, w_0, w_1 \in H \times H^{1+\alpha}(\Omega_s) \times H^\alpha(\Omega_s)$ . Here, the semigroup  $e^{\mathcal{A}_L^* t}$  gives the solution to the equation  $\hat{y}_t = \mathcal{A}_L^* \hat{y}$ ,  $\hat{y} = (\hat{u}, \hat{w}, \hat{w}_t)$  expressed below

$$\left\{ \begin{array}{l} (\hat{u}_t, \phi)_f = -(\epsilon(\hat{u}), \epsilon(\phi))_f - (L^* \hat{u}, \phi)_f + \langle \sigma \hat{w} \cdot \nu, \phi \rangle \\ (\hat{w}_{tt}, \psi)_s = (\operatorname{div} \sigma(\hat{w}), \psi)_s \\ \hat{w}_t|_{\Gamma_s} = -\hat{u}|_{\Gamma_s} \end{array} \right\} \quad (46)$$

for all  $\phi \in V, \psi \in H^1(\Omega_s)$ . The system above is regularity-wise equivalent to the system in (6) with  $g = 0$ . Moreover,  $\mathcal{A}_L^*$  also generates a  $c_0$  semigroup on  $\mathcal{H}$  using the same argument as that used to show that  $\mathcal{A}_L$  generates a  $c_0$  semigroup since  $L^* : V \rightarrow V'$  satisfies the same condition as  $L$ , see [15]. Hence, the same regularity as in (7) holds for the solution  $\hat{y} = [\hat{u}, \hat{w}, \hat{z}]$  to the adjoint system.

$$|\hat{y}(t)|_{\mathcal{H}}^2 + \int_0^t |\hat{u}|_{1,\Omega_f}^2 + |\sigma(\hat{w}) \cdot \nu|_{-1/2,\Gamma_s}^2 ds \leq C e^{\omega t} |\hat{y}_0|_{\mathcal{H}}^2. \quad (47)$$

In addition, results of Lemma 6.4 and of Lemma 6.5 are valid with  $(u, w)$  replaced by  $(\hat{u}, \hat{w})$ . In order to establish (45) we compute the adjoint  $\mathcal{B}^* e^{\mathcal{A}_L^* t}$  obtaining

$$\mathcal{B}^* e^{\mathcal{A}_L^* t} \begin{pmatrix} u_0 \\ w_0 \\ w_1 \end{pmatrix} = [N^* A, 0, 0] e^{\mathcal{A}_L^* t} \begin{pmatrix} u_0 \\ w_0 \\ w_1 \end{pmatrix} = N^* A \hat{u}|_{\Gamma_s} = \hat{u}|_{\Gamma_s}.$$

It is sufficient then to estimate the norm of  $\hat{u}(t)|_{\Gamma_s}$  in  $L_2(\Gamma_s)$  for solutions of (46). We denote  $\hat{u}(t)|_{\Gamma_s} = U_1(t) + U_2(t) + U_3(t) + U_4(t)$  where the terms  $U_i$  are defined as follows:

$$U_1 = N^* Ae^{At} u_0 \quad (48)$$

$$U_2 = \int_0^t N^* Ae^{A(t-s)} AN\sigma(\hat{w}_1)(s, \cdot) \nu ds \quad (49)$$

$$U_3 = \int_0^t N^* Ae^{A(t-s)} AN\sigma(\hat{w}_2)(s, \cdot) \nu ds \quad (50)$$

$$U_4 = \int_0^t N^* Ae^{A(t-s)} L^* \hat{u}(\cdot, s) ds. \quad (51)$$

**Estimate for  $U_1$ .** The term  $U_1$  is precisely the source of the singular estimate and it is estimated on the strength of Proposition 6.3 and (14)

$$|U_1(t)|_{L_2(\Gamma_s)} = |N^* Ae^{At} u_0|_{L_2(\Gamma_s)} \leq |N^* A^{3/4-\epsilon} e^{At} A^{1/4+\epsilon} u_0|_{L_2(\Gamma_s)} \leq \frac{C}{t^{1/4+\epsilon}} |y_0|_{\mathcal{H}}.$$

**Estimate for  $U_2$ .** Estimating  $U_2$  and  $U_3$  involves using properties of  $A$  and the Neumann map  $N$  and the estimates from Lemmas (6.5), (6.4):

$$\begin{aligned} |U_2(t)|_{L_2(\Gamma_s)} &\leq \int_0^t \frac{C}{(t-s)^{3/4+\epsilon}} |A^{1/2} N\sigma(\hat{w}_1)(s, \cdot) \nu|_{L_2(\Omega_f)} ds \\ &\leq C t^{1/4-\epsilon} |\sigma(\hat{w}_1) \cdot \nu|_{C([0, T]; H^{-1/2}(\Gamma_s))} \leq C_T |y_0|_{\mathcal{H}} \end{aligned}$$

where we have used Proposition 6.3 and

$$|N^* Ae^{A(t)} A^{1/2}|_{H \rightarrow L_2(\Gamma_s)} \leq \frac{C}{t^{3/4+\epsilon}}.$$

**Estimating  $U_3$ .** Note that  $H^{\alpha+1, \alpha/2+1/2}(\Gamma_s \times [0, T]) \subset C([0, T]; L_2(\Gamma_s))$  by Sobolev embedding theorems in one dimension. On the other hand  $U_3$  is the restriction on the boundary  $\Gamma_s$  of parabolic solutions driven by Neumann data  $\sigma(\hat{w}_2)$ . Thus  $U_3$  satisfies the following estimate

$$\begin{aligned} |U_3(t)|_{L_2(\Gamma_s)} &\leq |U_3|_{H^{\alpha+1, \alpha/2+1/2}(\Sigma_s) \times [0, T]} \\ &\leq K |\sigma(\hat{w}_2) \cdot \nu|_{H^{\alpha, \alpha/2}(\Sigma_s)} \leq K |\sigma(w_2) \cdot \nu|_{H^\alpha(\Sigma_s)}. \end{aligned}$$

We next apply the estimate (43) from Lemma 6.4 and estimate in Lemma 6.5 to obtain:

$$\begin{aligned} |U_3(t)|_{L_2(\Gamma_s)} &\leq K [\|\hat{u}\|_{H^\alpha(\Sigma_s)} + |y_0|_{H \times H^{1+\alpha}(\Omega_s) \times H^\alpha(\Omega_s)}] \\ |U_3(t)|_{L_2(\Gamma_s)} &\leq K_T |y_0|_{H \times H^{1+\alpha}(\Omega_s) \times H^\alpha(\Omega_s)} = K_T |y_0|_{\mathcal{H}_\alpha} \end{aligned}$$

with any  $0 < \alpha$ .

**Estimate for  $U_4$ .** In estimating  $U_4(t)$  we evoke again Proposition 6.3 and (47)

$$\begin{aligned} |U_4(t)|_{L_2(\Gamma_s)} &\leq \int_0^t |N^* A e^{A(t-s)} L^* \hat{u}(s, \cdot)|_{L_2(\Gamma_s)} ds \\ &\leq K \int_0^t |A^{1/4+\epsilon} e^{A(t-s)} L^* u(s, \cdot)|_{L_2(\Omega_f)} ds \\ &\leq K \int_0^t \frac{C}{(t-s)^{1/4+\epsilon}} |u|_V ds \leq C_T |u|_{L_2([0,T];V)} \leq C |y_0|_{\mathcal{H}}. \end{aligned}$$

Collecting the estimates of  $U_1, U_2, U_3$  and  $U_4$  leads to

$$|\mathcal{B}^* e^{\mathcal{A}_L^* t} y_0|_{L_2(\Gamma_s)} = |\hat{u}(t)|_{L_2(\Gamma_s)} \leq \frac{C}{t^{1/4+\epsilon}} |y_0|_{\mathcal{H}_\alpha}.$$

By duality, the above implies  $|e^{\mathcal{A}_L t} \mathcal{B} g|_{\mathcal{H}_{-\alpha}} \leq C t^{-1/4-\epsilon} |g|_{L_2(\Gamma_s)}$ .  $\square$

## 7. Conclusions and questions

- We have shown that the control problem defined for fluid structure interaction with the functional cost given by (8) admits optimal feedback synthesis expressed via Riccati operator. The gain operator is regular up to terminal point and it blows up with the rate  $(T-t)^{1/4+}$  at the terminal point. The result obtained depends on the validity of Singular Estimate Theorem 5.2. This, in turn, relies on parabolic-hyperbolic mixing that involves transfer of maximal parabolic regularity on one hand, and propagation of hyperbolic hidden regularity, on the other hand.
- One could generalize the result of this paper by considering more general functional penalizing also solid components of the structure. This could be done in a straightforward manner provided that the penalization is controlled by  $\mathcal{H}_{-\alpha}$  norms with  $\alpha > 0$ . Whether one can take  $\alpha = 0$  remains at present an open question. One way to address this question is by exploiting a more microscopic Riccati theory developed in [2]. This theory replaces global singular estimate condition by a local version of it and adds another regularity condition. However, the final result obtained is weaker, as it provides potentially unbounded gain operator  $B^* P(t)$  (though densely defined). It is possible (conjectural at this stage) that this theory could be applied to the model under consideration. The success of this approach depends on feasibility of obtaining additional regularity properties of the system (as it was accomplished for system of thermoelasticity in [3]).
- We note that the variational formulation of weak solutions to fluid structure interaction model is amenable to “friendly” numerical implementations. The test functions considered are “independent”, as they are not required to satisfy compatibility conditions on the interface. This is in contrast with other variational approaches which reinforce total matching on the interface [26, 24, 16]. An advantage of having unconstraint test functions is eminent

when designing FEM approximations for the control problem [44, 25, 35, 36]. Functions  $\phi$  and  $\psi$  do not need to match on  $\Gamma_s$ , thus discrete structures are not required to satisfy any extraneous conditions.

- In the spirit of the comment above, it would be interesting to pursue numerical analysis of the feedback control problem considered in this paper. There are many techniques available for effective solving of Riccati equations [35, 8, 9, 18] and references therein. Of particular interest is numerical verification of the blow-up asymptotics at the terminal point of the process.

## References

- [1] P. ACQUISTAPACE AND B. TERRENI Classical solutions of non-autonomous Riccati equations arising in parabolic boundary control problems, *Appl. Math. Optimiz.* 39, pp. 361–409, 2000.
- [2] P. ACQUISTAPACE, F. BUCCI, I. LASIECKA, Optimal boundary control and Riccati theory for abstract dynamics motivated by hybrid systems of PDEs. *Advances Differential Equations*, 10 (12), pp. 1389–1436, (2005).
- [3] P. ACQUISTAPACE, F. BUCCI, I. LASIECKA, A trace regularity result for thermoelastic equations with applications to optimal boundary control. *J. Math. Anal. Appl.*, vol. 310, pp. 262–277, 2005.
- [4] G. AVALOS, Differential Riccati equations for the active control of a problem in structural acoustic , *J. Optim. Theory, Appl.*, 91, pp. 695–728, 1996.
- [5] G. AVALOS, The strong stability and instability of a fluid-structure semigroup, *Appl. Math. and Optim.*, vol. 55, pp. 163–184 , 2007.
- [6] G. AVALOS AND R. TRIGGIANI, The coupled PDE system arising in fluid structure interactions, *AMS Contemporary mathematics*, Fluids and Waves, vol. 440, pp. 15–55, 2007.
- [7] A.V. BALAKRISHNAN, Applied Functional Analysis, Springer Verlag, 1975.
- [8] H.T. BANKS AND K. KUNISCH, The linear regulator problem for parabolic systems, *Siam J. Control*, vol. 22, pp. 684–698, 1984.
- [9] H.T. BANKS AND K. ITO, A numerical algorithm for optimal feedback gains in high-dimensional linear quadratic regulator problems. *Siam J. Control*, vol. 29, pp. 499–515, 1991.
- [10] F. BUCCI, Control theoretic properties of structural acoustic models with thermal effects. I Singular Estimates. *Journal Evolution Equations*, vol. 7, pp. 387–414, 2007.
- [11] V. BARBU & I. LASIECKA, R. TRIGGIANI, Extended Algebraic Riccati Equations in the Abstract Hyperbolic Case, *Nonlinear Analysis*, 40, pp. 105–129, 2000.
- [12] F. BUCCI, I. LASIECKA, R. TRIGGIANI, Singular estimate and uniform stability of coupled systems of hyperbolic-parabolic PDE's, *Abstract and Applied Analysis*, vol. 7, pp. 169–236, 2002.
- [13] F. BUCCI AND I. LASIECKA, Singular estimates and Riccati theory for thermoelastic plate models with boundary thermal controls. *Dynamics of Continuous, Discrete and Impulsive Systems*, vol. 11, pp. 545–568, 2004.

- [14] V. BARBU, Z. GRUJIC, I. LASIECKA & A. TUFFAH, Smoothness of Weak Solutions to a nonlinear fluid-structure interaction model, *Indiana Journal of Mathematics*, Vol. 57, No. 3, pp. 1173–1207, 2008.
- [15] V. BARBU, Z. GRUJIC, I. LASIECKA & A. TUFFAH, Weak solutions for nonlinear fluid-structure interaction, *AMS Contemporary Mathematics: Recent Trends in Applied Analysis*, Volume 440, pp. 55–81, 2007.
- [16] H. BEIRAO DA VEIGA, On the Existence of Strong Solutions to a Coupled Fluid-Structure Evolution Problem, *Journal of Mathematical Fluid Mechanics*, 6, pp. 21–52, 2004.
- [17] A. BENSOUSSAN, G. DA PRATO , M.C. DELFOUR & S.K. MITTER, *Representation and Control of Infinite Dimensional Systems*, Birkhäuser, 1993.
- [18] J. BURNS AND K. HULSING, Numerical methods for approximating functional gains in LQR boundary control problems, *Mathematical and Computer Modeling*, vol. 33, pp. 89–100, 2001.
- [19] D. COUTAND & S. SHKOLLER, Motion of an elastic solid inside an incompressible viscous fluid, *Arch. Ration. Mech. Anal.*, 176, no. 1, pp. 25–102, 2005.
- [20] R. CURTAIN AND H. ZWART, *Introduction to Infinite Dimensional Linear Systems*, Springer Verlag, 1995.
- [21] W. DESCH & S. SCHAPPACHER, Some Perturbation Results for Analytic Semigroups, *Mathematische Annalen*, Volume 281, pp. 157–162, 1988.
- [22] G. DA PRATO AND A. ICHIKAWA , Riccati equations with unbounded coefficients, *Annali di Mat. Pura et Appl.*, 140, pp. 209–221, 1985.
- [23] F. FLANDOLI, Riccati equations arising in an optimal control problem with distributed parameters, *Siam J. Control*, 22, pp. 76–86, 1984.
- [24] M.A. FERNANDEZ & M. MOUBACHIR, An exact Block-Newton algorithm for solving fluid-structure interaction problems, *C.R. Acad. Sci Paris*, Ser. I 336, pp. 681–686, 2003.
- [25] S. GIBSON, The Riccati integral equations for optimal control problems in Hilbert spaces, *Siam J. Control*, vol. 17, pp. 637–665, 1979.
- [26] Q. DU, M.D. GUNZBURGER, L.S. HOU & J. LEE, Analysis of a Linear Fluid-Structure Interaction Problem, *Discrete Continuous Dynamical Systems*, 9 no. 3, pp. 633–650, 2003.
- [27] I. LASIECKA: NSF-CMBS Lecture Notes; *Mathematical Control Theory of Coupled PDE's*, SIAM, 2002. with Unbounded Controls Continuous and Approximations.
- [28] I. LASIECKA, *Unified theory for abstract parabolic problems – a semigroup approach*, Appl. Math. Optim. 6, 1980, pp. 287–333.
- [29] I. LASIECKA & R. TRIGGIANI , *Optimal control and Differential Riccati Equations under Singular estimates for  $e^{At}B$  in the absence of analyticity*; Advances in Dynamics and Control, Special Volume dedicated to A.V. Balakrishnan, Chapman and Hall/CRC Press, pp. 271–309, 2004.
- [30] I. LASIECKA, J.L. LIONS & R. TRIGGIANI, Non-Homogeneous Boundary Value Problems for Second Order Hyperbolic Operators, *Journal de Mathématique Pure et Appliquée* 65, pp. 149–192, 1986.

- [31] I. LASIECKA & R. TRIGGIANI *Control Theory for Partial Differential Equations: Continuous and Approximations Theories*, Volume I; Cambridge. 2000.
- [32] I. LASIECKA & A. TUFFAH, Riccati Equations for the Bolza Problem arising in boundary/point control problems governed by  $c_0$  semigroups satisfying a singular estimate, *JOTA*, vol. 136, pp. 229–246, 2008.
- [33] I. LASIECKA AND A. TUFFAH Riccati theory and singular estimates for Bolza control problem arising in linearised fluid structure interactions, *Systems and Control Letters*, to appear.
- [34] J.L. LIONS *Quelques méthodes de résolution des problèmes aux limites nonlinéaires*, Dunod. Paris, 1969.
- [35] K. MORRIS AND C. NAVASCA, Solutions of algebraic Riccati equations arising in control of PDE's, *Control and Boundary Analysis*, Marcel Dekker, 2004.
- [36] K. MORRIS, Design of finite dimensional controllers for infinite-dimensional systems by approximations. *J. Math. Systems, Estimation and Control*, vol. 4, pp. 1–30, 1994.
- [37] M. MOUBACHIR J. & ZOLESI, *Moving Shape Analysis and Control: applications to Fluid Structure Interactions*, Chapman & Hall/CRC, 2006.
- [38] A. PAZY; *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, 1983.
- [39] I. ROSEN AND C. WANG, A multilevel technique for the approximate solution of operator Lyapunov and algebraic Riccati equations, *Siam J. Numer. Analysis*, vol. 32, pp. 514–541, 1994.
- [40] D. RUSSELL. Controllability and stabilizability theory for partial differential equations, *Siam Review*, pp. 639–739, 1978.
- [41] D. RUSSELL, Mathematical models for the elastic beam and their control theoretic properties, *Semigroups Theory and Applications*, Pittman Research Notes, vol. 152, pp. 177–217, 1986.
- [42] D. RUSSELL, Quadratic performance criteria in boundary control of linear symmetric hyperbolic systems, *Siam J. Control*, vol. 11, pp. 475–509, 1973.
- [43] O. STAFFANS, *Wellposed Linear Systems*, Cambridge Univ. Press, 2005.
- [44] R. TEMAM *Navier-Stokes Equations*, Studies in Mathematics and its Applications, North-Holland, 1977.

Irena Lasiecka

Department of Mathematics

University of Virginia

Charlottesville, VA 22901, USA

Amjad Tuffaha

Department of Mathematics

University Southern California

Los Angeles, CA, USA

# Single-step One-shot Aerodynamic Shape Optimization

Emre Özkan and Nicolas R. Gauger

**Abstract.** In this paper we consider the shape optimization of a transonic airfoil whose aerodynamic properties are calculated by a structured Euler solver. The optimization strategy is based on a one-shot technique in which pseudo time-steps of the primal and the adjoint solver are iterated simultaneously with design corrections done on the airfoil geometry. The adjoint solver which calculates the necessary sensitivities is based on discrete adjoints and derived by using reverse mode of automatic differentiation. A new preconditioner which is derived by considering an augmented Lagrangian formulation of the optimization problem is employed in order to achieve bounded retardation of the overall optimization process. A design example of drag minimization for an RAE2822 airfoil under transonic flight conditions is included.

**Mathematics Subject Classification (2000).** Primary 99Z99; Secondary 00A00.

**Keywords.** Aerodynamic shape optimization, one-shot optimization, transonic airfoil optimization, automatic differentiation, fixed point solver.

## 1. Introduction

Computational Fluid Dynamics (CFD) is nowadays an essential part of aerodynamic design processes. During the past decades, CFD simulations evolved from basic inviscid potential solvers to Navier-Stokes solvers with complex turbulence and transition models. Along with the tremendous increase of computational power, CFD simulations, performed on meshes with several millions of grid points, are already state of the art. Mathematicians and engineers work on the goal to integrate efficiently CFD simulations into optimization strategies, even though derivative free optimization methods are preferred in industry, because of their simplicity. But derivative free optimization methods need several hundreds of function evaluations, even in the case of only a few design variables, and therefore they are inefficient, because in aerodynamics the simulation part is expensive in terms of computational costs. This is the reason why in detailed design one should

prefer deterministic gradient-based optimization strategies. In our work, we focus on a special class of gradient-based methods, the so-called one-shot methods, where pseudo-time steps of the CFD solver and the design changes are performed simultaneously.

Let us consider the following optimization problem:

$$\min_u f(u, y) \text{ s.t. } c(u, y) = 0 \text{ (e.g., } f(y, u) = C_d(y, u)). \quad (1.1)$$

Here,  $u \in U$  denotes the vector of design variables defining the shape of the airfoil, and  $y \in Y$  denotes the vector of state variables. In our application, the cost function  $f(y, u)$  is the drag coefficient  $C_d$ . We assume that  $Y$ ,  $U$  and their Cartesian product  $X = Y \times U$  are Hilbert spaces.  $c(u, y) = 0$  is a PDE that governs the fluid flow around the airfoil, e.g., the compressible Euler equations. The aim is to decrease the drag force exerted on the airfoil and to satisfy the primal feasibility, i.e., the flow field satisfies the compressible Euler equations. For simplicity, we focus on this unconstrained drag reduction case. Lift and pitching moment constraints are expected to be an extension of our work in the future.

The governing equations and the applied CFD solver will be introduced in detail in Section 2.

As the compressible Euler equations cannot be easily solved numerically due to the appearance of high nonlinearities, one usually uses quasi-unsteady formulations which are solved by explicit finite volume schemes stabilized by artificial dissipation and Runge-Kutta time integration. These pseudo timestepping schemes are most efficient in combination with geometric multigrid. That is to say, that our state equation  $c(y, u) = 0$  is solved by a contractive fixed point iteration  $y_{k+1} = G(y_k, u)$ , i.e.,  $\|G_y\| \leq \rho < 1$ .

Consequently, we can rewrite the optimization problem (1.1) as

$$\min_u f(u, y) \text{ s.t. } y = G(u, y) \quad (1.2)$$

and we define the Lagrangian function

$$L(u, y, \bar{y}) = f(y, u) + (G(y, u) - y)^T \bar{y} = N(y, \bar{y}, u) - y^T \bar{y}, \quad (1.3)$$

while  $\bar{y}$  denotes the adjoint state vector (or co-state vector). Furthermore, we call

$$N(y, \bar{y}, u) := f(y, u) + G(y, u)^T \bar{y} \quad (1.4)$$

the shifted Lagrangian.

If we derive the *KKT* conditions for the problem (1.2), a KKT point  $(y_*, \bar{y}_*, u_*)$  has to satisfy the following conditions:

$$\begin{aligned} y_* &= G(y_*, u_*) \\ \bar{y}_* &= N_y(y_*, \bar{y}_*, u_*)^T = f_y(y_*, u_*)^T + G_y(y_*, u_*)^T \bar{y}_* \\ 0 &= N_u(y_*, \bar{y}_*, u_*)^T = f_u(y_*, u_*)^T + G_u(y_*, u_*)^T \bar{y}_*. \end{aligned} \quad (1.5)$$

Rather than first fully converging the primal state using

$$y_{k+1} = G(y_k, u) \rightarrow \text{primal feasibility at } y_* \quad (1.6)$$

and then fully converging the dual state applying

$$\bar{y}_{k+1} = N_y(y, \bar{y}_k, u) \rightarrow \text{dual feasibility at } \bar{y}_* \quad (1.7)$$

before finally performing an “outer” optimization loop

$$u_{k+1} = u_k - B_k^{-1} N_u(y, \bar{y}, u_k) \rightarrow \text{optimality at } u_* , \quad (1.8)$$

we suggest an extended single-step one-shot iteration of the form

$$\begin{bmatrix} y_{k+1} \\ \bar{y}_{k+1} \\ u_{k+1} \end{bmatrix} = \begin{bmatrix} G(y_k, u_k) \\ N_y(y_k, \bar{y}_k, u_k)^\top \\ u_k - B_k^{-1} N_u(y_k, \bar{y}_k, u_k)^\top \end{bmatrix}. \quad (1.9)$$

For computing the optimization correction  $u_{k+1} - u_k$ , one has to choose a suitable preconditioner  $B_k$ .

In our present work, we employed a new type of preconditioner based on the descent of an augmented Lagrangian function, cf. [8].

We will introduce this preconditioner in detail in Section 3.

Throughout this paper we also aim to satisfy a bounded cost deterioration [14], such that the cost of the coupled iteration (1.9), i.e., the one-shot optimization process, is proportional to the cost of a single state simulation:

$$\frac{\text{cost of optimization}}{\text{cost of simulation}} = c. \quad (1.10)$$

From this it follows, that the cost of an optimization is independent of the number of shape parameters, i.e., the size of the vector  $u$ . As far as calculations of derivative vectors are concerned, we recommend the use of automatic differentiation (AD) tools instead of applying continuous adjoint calculus or finite differences.

In Section 4, we will consider several methods for calculating derivatives as well as pros and cons of the different methods.

Finally, in Section 5 we present numerical results, that have been achieved for the drag minimization under transonic 2D Euler flow.

## 2. The aerodynamic design chain and its flow solver

In this section we state the governing equations of the flow field around the airfoil and the corresponding CFD solver. We will also briefly cover all the computational steps between the shape parametrization of the airfoil and the aerodynamic coefficients and forces.

### 2.1. Governing equations and boundary conditions

Since we are interested in drag reduction in transonic flow regime, the compressible Euler equations are an appropriate choice. They are capable of describing the (inviscid) shocks, which are the main sources of the pressure drag.

Even though the flow is not unsteady, the solution is obtained by integrating the (quasi-)unsteady Euler equations in time until a steady state is reached. Note

that these time steps do not physically mean anything, therefore they are called pseudo-timesteps.

For 2D flow, the compressible Euler equations in cartesian coordinates read:

$$\frac{\partial w}{\partial t} + \frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} = 0 \quad \text{with} \quad f = \begin{bmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ \rho uH \end{bmatrix} \quad \text{and} \quad g = \begin{bmatrix} \rho v \\ \rho vu \\ \rho v^2 + p \\ \rho vH \end{bmatrix}, \quad (2.1)$$

where  $w$  is the vector of conserved variables  $\{\rho, \rho u, \rho v, \rho E\}$  ( $\rho$  is the density,  $u$  and  $v$  are the velocity components and  $E$  denotes the energy).

As far as the boundary conditions are concerned, we assume the Euler slip condition on the wall ( $\vec{n}^T \vec{v} = 0$ ) and free stream conditions at the farfield. For a perfect gas holds

$$p = (\gamma - 1)\rho \left( E - \frac{1}{2}(u^2 + v^2) \right) \quad (2.2)$$

and

$$\rho H = \rho E + p \quad (2.3)$$

for the pressure  $p$  and the enthalpy  $H$ .

The pressure coefficient  $C_p$  and the drag coefficient  $C_d$  are defined as

$$C_p := \frac{2(p - p_\infty)}{\gamma M_\infty^2 p_\infty}, \quad (2.4)$$

$$C_d := \frac{1}{C_{\text{ref}}} \int_C C_p (n_x \cos \alpha + n_y \sin \alpha) dl. \quad (2.5)$$

## 2.2. Shape parametrization

In aerodynamic shape optimization, there are mainly two ways of doing the shape updates: Either parameterizing the shape itself or parameterizing shape deformations. In [10] these possibilities are investigated in detail. In the following, we take the second approach, such that an initial airfoil shape is deformed by some set of basis functions that are scaled by certain design parameters. Here, the basic idea of shape deformation is to evaluate these basis functions scaled with certain design parameters and to deform the camberline of the airfoil accordingly. Then, the new shape is simply obtained by using the deformed camberline and the initial thickness distribution. The result is a surface deformation that maintains the airfoil thickness.

We have chosen Hicks-Henne functions, which are widely used in airfoil optimization. The Hicks-Henne functions are defined as

$$h_{a,b} : [0, 1] \rightarrow [0, 1] : h_{a,b}(x) = \left( \sin \pi x^{\frac{\log 0.5}{\log a}} \right)^b, \quad (2.6)$$

where  $b$  is fixed as 3.0 and  $a$  varies from  $\frac{3}{n+5}$  to  $\frac{n+3}{n+5}$  and  $n$  being the number of design parameters. These function have the positive property that they are defined in the interval  $[0, 1]$  with a peak position at  $a$  and they are analytically smooth at zero and one.

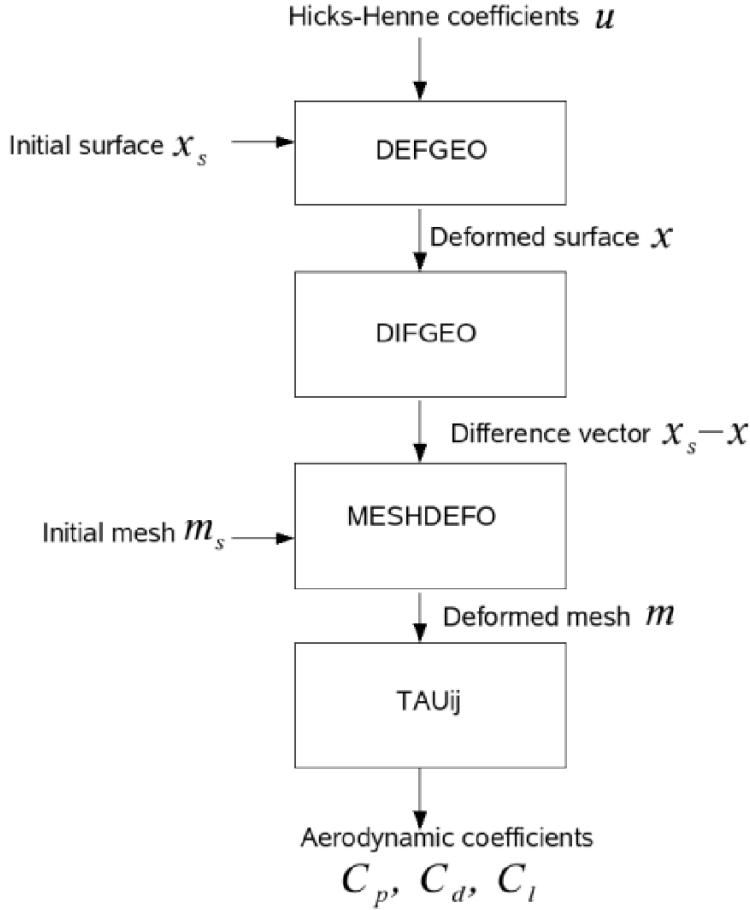


FIGURE 1. Design chain

The normalized airfoil shape is deformed by using Hicks-Henne functions multiplied by the design parameters  $u_i$ :

$$\Delta \text{camber}(x) = \sum u_i h \{a, b\} (x) \text{ and } \text{camber}(x) + = \Delta \text{camber}(x) . \quad (2.7)$$

### 2.3. Design chain

After deforming the airfoil geometry, the next task is to deform the initial mesh. To do this, we use a modular approach by using several tools. Firstly, the airfoil geometry  $x_s$  is deformed by using the tool **defgeo**. Afterwards, a difference vector  $dx = x_s - x$  is calculated by the tool **difgeo** and finally another tool, called **meshdefo**, performs a mesh deformation by using this difference vector. This approach is also very advantageous in terms of gradient computations, since we have

to differentiate only the simple structured mesh and shape deformation tools, instead of complex mesh generators. The design chain between the design parameters (Hicks-Henne coefficients  $u_i$ ) and the aerodynamic coefficients is illustrated in Figure 1.

#### 2.4. CFD solver

The numerical solution of (2.1) is computed by the TAUij code, which is a structured quasi 2D version of the TAU code, developed at the German Aerospace Center (DLR). For the spatial discretization the MAPS+ [11] scheme is used. To achieve second-order accuracy, gradients are used to reconstruct the values of variables at the cell faces. A slip wall and a farfield boundary condition are applied. For the pseudo timestepping, a fourth-order Runge-Kutta scheme is applied. To accelerate the convergence, local time stepping, explicit residual smoothing and a multigrid method are used. The code TAUij is written in C and comprises approximately 6000 lines of code distributed over several files.

### 3. The suitable preconditioner $B_k$

In this section we will briefly introduce a suitable preconditioner in order to achieve the goal of bounded cost deterioration. The reader might take a look at [7] and [8] for more details on the derivation. Griewank et al. suggest to look for descent on the merit function of the augmented Lagrangian

$$L^a(y, \bar{y}, u) = \frac{\alpha}{2} \|G(y, u) - y\|^2 + \frac{\beta}{2} \|N_y(y, \bar{y}, u)^T - \bar{y}\|^2 + N(y, \bar{y}, u) - \bar{y}^T y , \quad (3.1)$$

where the two weighting coefficients  $\alpha$  and  $\beta$  are strictly positive reals. The augmented terms represent the primal and the adjoint state residuals.

In [8] it has been proved that  $L^a$  is an exact penalty function if the so-called correspondence condition

$$\alpha\beta\Delta G_y^T \Delta G_y \succ I + \beta N_{yy}, \text{ with } \Delta G_y = I - G_y , \quad (3.2)$$

is fulfilled.

In the same paper it is also demonstrated that the step increment vector  $s$  of the extended iteration (1.9), which is defined as

$$s(y, \bar{y}, u) = \begin{bmatrix} \Delta y \\ \Delta \bar{y} \\ \Delta u \end{bmatrix} = \begin{bmatrix} G(y, u) - y \\ N_y(y, \bar{y}, u)^T - \bar{y} \\ -B^{-1}N_u(y, \bar{y}, u)^T \end{bmatrix} , \quad (3.3)$$

yields descent on  $L^a$  for all large positive preconditioners  $B$  if the descent condition

$$\alpha\beta\Delta \bar{G}_y \succ (I + \frac{\beta}{2}N_{yy})(\Delta \bar{G}_y)^{-1}(I + \frac{\beta}{2}N_{yy}), \text{ with } \Delta \bar{G}_y = \frac{1}{2}(\Delta G_y + \Delta G_y^T) , \quad (3.4)$$

is satisfied.

Once the weighting coefficients  $\alpha$  and  $\beta$  are chosen such that

$$\sqrt{\alpha\beta}(1 - \rho) > 1 + \frac{\beta}{2} \|N_{yy}\| , \quad (3.5)$$

both inequalities (3.2) and (3.4) are fulfilled and thus  $L^a$  is an exact penalty function on which the increment vector  $s$  yields descent for a sufficiently large preconditioner  $B$ .

Furthermore, in [8] it has been proven that any preconditioner fulfilling

$$B \succeq B_0 \equiv \frac{1}{\sigma}(\alpha G_u^T G_u + \beta N_{yu}^T N_{yu}) , \quad (3.6)$$

while

$$\sigma = 1 - \rho - \frac{(1 + \frac{\|N_{yu}\|}{2}\beta)^2}{\alpha\beta(1 - \rho)} , \quad (3.7)$$

yields descent on the augmented Lagrangian  $L^a$ .

A reasonable way to find a suitable  $B$  is to solve

$$\min_{\Delta u} L^a(y + \Delta y, \bar{y} + \Delta \bar{y}, u + \Delta u) \quad (3.8)$$

and then identifying  $B$  from the obtained solution of  $\Delta u = -B^{-1}N_u^T$ . Furthermore, from (3.8) and by considering a quadratic approximation of  $L^a$ , we obtain the following optimization problem:

$$\min_{\Delta u} s^T \nabla L^a(y, \bar{y}, u) + \frac{1}{2} s^T \nabla^2 L^a(y, \bar{y}, u) s , \quad (3.9)$$

where  $s$  is the increment vector (3.3). Then (3.9) leads to the equivalent minimization problem

$$\min_{\Delta u} \varphi(\Delta u) , \quad (3.10)$$

where  $\varphi$  is the quadratic function given as follows:

$$\begin{aligned} \varphi(\Delta u) &= \Delta u^T (\nabla_u L^a + \nabla_{uy} L^a \Delta y + \nabla_{u\bar{y}} L^a \Delta \bar{y}) + \frac{1}{2} \Delta u^T \nabla_{uu} L^a \Delta u \\ &\approx \Delta u^T \nabla_u L^a(y + \Delta y, \bar{y} + \Delta \bar{y}, u) + \frac{1}{2} \Delta u^T \nabla_{uu} L^a \Delta u . \end{aligned} \quad (3.11)$$

Therefore, whenever  $\nabla_{uu} L^a$  is positive definite, the solution  $\Delta u$  of the minimization problem (3.10) is defined by

$$\Delta u = -\nabla_{uu}^{-1} L^a(y, \bar{y}, u) \nabla_u L^a(y + \Delta y, \bar{y} + \Delta \bar{y}, u) . \quad (3.12)$$

And thus  $B \approx \nabla_{uu} L^a$ . Moreover, at feasibility ( $\Delta y = 0$  and  $\Delta \bar{y} = 0$ ), we obtain

$$\Delta u = -B^{-1}N_u(y, \bar{y}, u)^T , \text{ where } B = \alpha G_u^T G_u + \beta N_{yu}^T N_{yu} + N_{uu} . \quad (3.13)$$

Note, that for the exact Hessian holds

$$\nabla_{uu} L^a = \alpha G_u^T G_u + \beta N_{yu}^T N_{yu} + N_{uu} + \alpha(G - y)^T G_{uu} + \beta(N_y^T - \bar{y})^T N_{yuu} . \quad (3.14)$$

When primal and dual feasibility are satisfied, then the last two terms are zero, since  $G = y$  and  $N_y^T = \bar{y}$ , and we get the expression (3.13) for  $B$ .

Since the computation of  $B$  derived from (3.13) involves matrix derivatives that may lead to expensive calculations, we aim to find an approximation by using

BGFS updates [12] rather than computing it exactly for each iteration. Since  $B \approx \nabla_{uu} L^a$ , we have

$$B\Delta u = \nabla_{uu} L^a(y, \bar{y}, u)\Delta u \approx \nabla_u L^a(y, \bar{y}, u + \Delta u) - \nabla_u L^a(y, \bar{y}, u). \quad (3.15)$$

Thus, we may employ the above approximation as a secant equation into the update of  $H$  ( $B^{-1}$ ). Therefore, we may impose

$$H_{k+1}R_k = \Delta u_k, \text{ where } R_k := \nabla_u L^a(y_k, \bar{y}_k, u_k + \Delta u_k) - \nabla_u L^a(y_k, \bar{y}_k, u_k). \quad (3.16)$$

The secant equation (3.16) has a solution only if

$$R_k^T \Delta u_k > 0 \quad (3.17)$$

is satisfied. Therefore, we check this condition in all iterations and make a BFGS update, whenever it is satisfied. Otherwise, we simply set  $B = I$ . As far as the BFGS update is concerned, there is no need to make an update of  $B$  and then inverse it; we can directly update the inverse of it by using the formula [12]

$$H_{k+1} = (I - r_k \Delta u_k R_k^T) H_k (I - r_k R_k^T \Delta u_k) + r_k \Delta u_k \Delta u_k^T, \quad (3.18)$$

where  $r_k = \frac{1}{r_k^T \Delta u_k}$ . Now, the only difficulty left is to calculate the term  $\nabla_u L^a$ . In the next section, we will address the computation of this term by automatic differentiation (AD).

## 4. Gradient computation and implementation issues

In this section, firstly, we consider several approaches for computing gradient vectors. Then, we will address implementation issues concerning the coupled iteration (1.9) and the computation of the preconditioner  $B$ .

One of the key points in aerodynamic shape optimization with gradient-based methods is the computation of the derivatives. Since we are interested in satisfying the bounded deterioration as stated in Section 1, we will turn our attention to methods in which the computational cost is independent of the number of design variables (i.e.,  $\dim(u)$ ). Continuous adjoint approaches as well as the use of the reverse mode of automatic differentiation (reverse AD) yield this important property. One of the difficulties with continuous adjoint approaches is their complex mathematical treatment and in particular their implementation, which is error prone. Instead, in order to exploit the domain specific experience and expertise invested in the simulation tools, we propose to equip them in an automated fashion with adjoints by the use of AD tools.

Therefore, we make use of the AD tool ADOL-C [2]. The freeware package ADOL-C is a tool in order to differentiate computer programs written in C or C++. Reverse and forward modes are available as well as the capability to derive first- and higher-order derivatives. Since ADOL-C is based on operator overloading, it is necessary to tape all operations that are done on the active variables (the variables which are to differentiate with respect to the selected independent parameters) on memory or disk. Usually a buffer size is defined by the user and the tapes which

are larger than this predefined size are written on the disk instead of the memory. Reading and writing operations from the disk is time consuming. This is why memory should be used instead of storage on disk, whenever it is possible. Since we use a one-shot approach rather than a hierarchical approach, we need to tape only one pseudo-time step in each iteration instead of the whole time-stepping. This is of course very advantageous, since the tape sizes would be extremely large, even for the case of rather coarse meshes, because the tape size of a primal iterate would be multiplied by the number of pseudo-timesteps. Nevertheless, in [4] it is demonstrated how to overcome this kind of drawbacks in cases of hierarchical approaches by the so-called reverse accumulation of adjoints [5].

As previously stated, for the coupled iteration, we need to evaluate several derivative vectors

$$\nabla_u L^a = \alpha \Delta y^T G_u + \beta \Delta \bar{y}^T N_{yu} + N_u , \quad (4.1)$$

in order to update the design vectors  $u$ .

Furthermore, we need to evaluate the terms  $N_y$  for the update of the adjoint states  $\bar{y}$ .

Note, that all expressions in (4.1) are either vectors or matrix vector products. Several subroutines of ADOL-C allow us to calculate these matrix vector products easily by using the reverse mode of AD for the first-order terms and reverse on tangent for the second-order term. These routines are namely

```
int fos_reverse(tag,m,n,u,z)
```

for the first-order terms and

```
int lagra_hess_vec(tag,m,n,x,v,u,h)
```

for the second-order term  $\beta \Delta \bar{y}^T N_{yu}$ . In this connection, we have the declarations

```
short int tag // tape identification
int m          // number of dependent variables
int n          // number of independent variables
double x[n]    // independent vector
double v[n]    // tangent vector
double u[m]    // range weight vector
double h[n]    // resulting second-order derivatives
double z[m]    // resulting adjoint vector
```

while  $h = u^T \nabla^2 F(x)v$  and  $z^T = u^T F'(x)$ . For detailed information about the ADOL-C subroutines, the reader might refer to [2].

For the differentiation, we simply set the independent vector  $[X] = [u; y]$  and the dependent vector  $[Y] = [N; y]$  and correspondingly calculate  $N$  inside the routine that returns the goal functional  $C_d$ . It should also be mentioned that, apart from  $N_y$ , the derivatives with respect to the design parameters  $u$  are propagated within the design chain in the reverse order as vector matrix products. In addition to the flow solver, the other programs of the design chain, namely **meshdefo**, **difgeo**, **defgeo**, have to be differentiated, too. In [1], this reverse propagation of the adjoint

vectors is covered in detail, and comparisons of the resulting adjoint sensitivities versus finite differences are also illustrated.

In order to demonstrate the backward propagation of the derivative vectors, which depend on  $u$ , we consider the last term  $N_u$  of the gradient  $\nabla_u L^a$ . By use of the chain rule, we express this term as

$$\frac{\partial N}{\partial u} = \frac{\partial N}{\partial m} \frac{\partial m}{\partial dx} \frac{\partial dx}{\partial x} \frac{\partial x}{\partial u}. \quad (4.2)$$

Starting with the first term from the left, the adjoint vector is propagated in reverse order by vector matrix products. Therefore, we firstly calculate  $\frac{\partial N}{\partial m} \frac{\partial m}{\partial dx}$  and then proceed in the same way with the other terms.

## 5. Numerical results

We apply now the single-step one-shot method and the related preconditioner  $B_k$ , which were derived in the previous sections, to the drag reduction of a RAE2822 airfoil under transonic inviscid flow conditions. As design variables we choose 20 Hicks-Henne parameters, as mentioned in Section 2.

For the sake of simplicity, we do not update the values  $\alpha$  and  $\beta$  of the augmented Lagrangian in each one-shot iteration, but we keep them constant through the whole optimization process. If we recall the descent condition (3.4), we see that the choice of large numbers for the weighting coefficients  $\alpha$  and  $\beta$  assures descent in  $L^a$ . However, making these coefficients unnecessarily large, slows down the coupled iteration. On the other hand, assigning small values to these coefficients boosts up the speed of convergence. However, the state and adjoint state vectors, which do not satisfy some certain levels of primal and dual feasibility, cause erroneous sensitivity calculations. Hence, the coupled one-shot iteration might diverge and the whole optimization process may not be stable. Therefore, a compromise should be made, while selecting the values for  $\alpha$  and  $\beta$ .

In order to determine these coefficients, the following method is proposed: By deriving  $\alpha$  and  $\beta$  from the descent condition, we may minimize  $\alpha$  as a function of  $\beta$ , such that

$$\min_{\beta} \sqrt{\alpha} \equiv \frac{1 + \frac{\theta}{2}\beta}{(1 - \rho)\sqrt{\beta}} \text{ with } \theta \equiv \|N_{yy}\|, \quad (5.1)$$

which leads to the following values:

$$\beta = \frac{\theta}{2} \text{ and } \alpha = \frac{2\theta}{(1 - \rho)^2}. \quad (5.2)$$

Here, we might assume  $\|N_{yy}\| = 1$ . This assumption is tested and justified in [6] at least for contractive fixed point solvers based on elliptic PDEs.

As a stopping criteria, we choose  $|\Delta u| < \epsilon$ , where  $\epsilon$  is a user defined tolerance. For our particular application we have chosen  $\epsilon = 0.0001$ .

As flow conditions, we have an inflow Mach number of  $M_\infty = 0.73$  and an angle of attack of  $\alpha = 2^\circ$ .

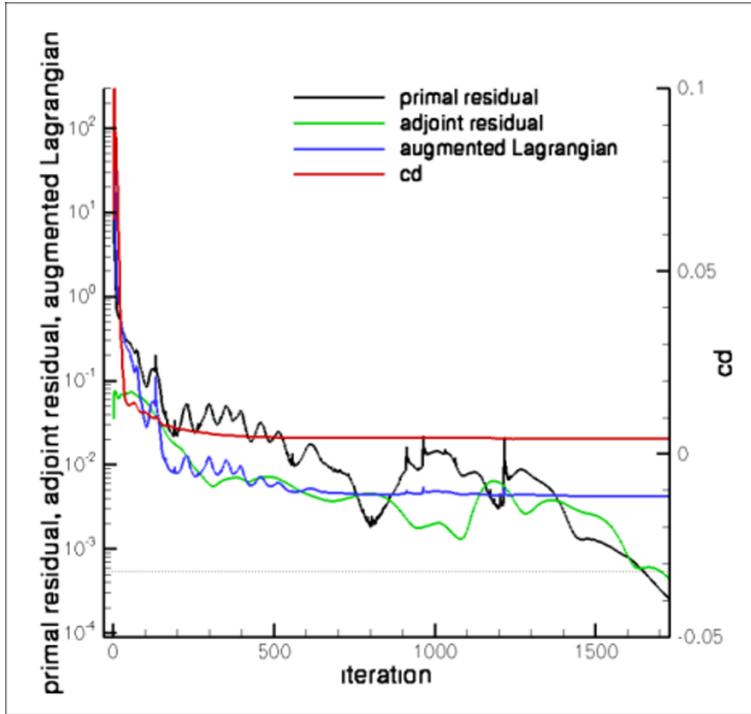


FIGURE 2. Optimization history

Within the first 30 iterations, in order to smooth out possible oscillatory effects, caused by the initialization of the flow field, we do only updates of the state and the adjoint state, without changes of the airfoil geometry. Then, from the 31st iteration on, we do the single-step one-shot iteration as stated in (1.9).

Figure 2 shows the optimization histories of the augmented Lagrangian, the cost functional  $C_d$ , the primal as well as the adjoint state residual. We observe, that after approximately 1600 iterations, the coupled iteration converges and the drag coefficient is reduced drastically.

Figure 3 shows the pressure distributions and geometries of the initial and optimized airfoils. For the optimized airfoil we observe, that the initial shock could be completely eliminated. It is well known, that this is only fulfilled for the correct physical optimum.

Figure 4 shows the convergence history of the flow calculation, just for the initial RAE2822 airfoil. After approximately 400 iterations, the single iterating primal state solver reaches the same level of convergence as the primal states in the coupled iteration, i.e., during the single-step one-shot approach.

Consequently, we just measure a deterioration factor of 4 from the simulation to the one-shot optimization.

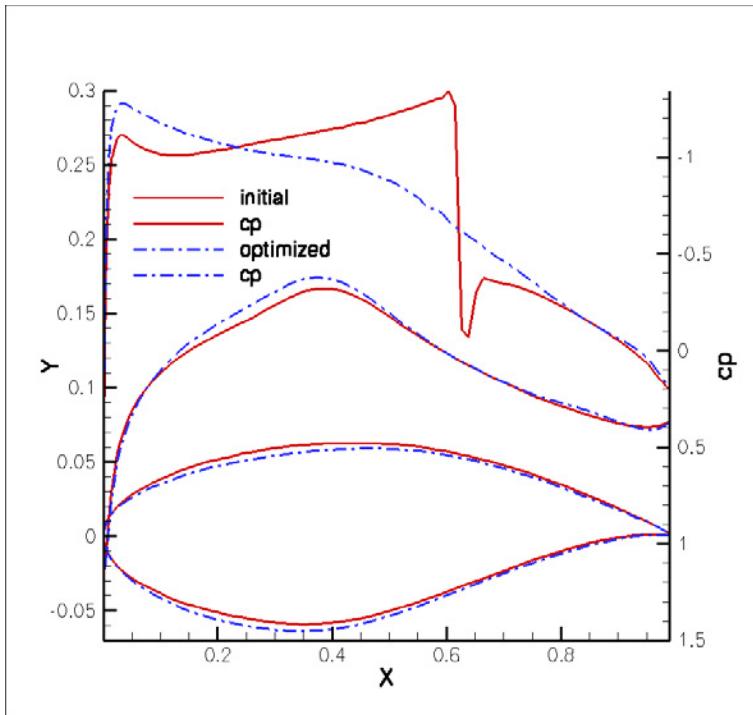


FIGURE 3. Pressure distributions and geometries of initial and optimized airfoils

## 6. Conclusion

We presented the development of a single-step one-shot approach for the efficient transition from simulation to optimization. This methodology is in principle applicable to all areas of scientific computing, where large scale governing equations involving discretized PDEs are treated by custom made fixed point solvers. To exploit the domain specific experience and expertise invested in these simulation tools, we proposed to extend them in an automated fashion by the use of AD tools. A new preconditioner which is derived by considering an augmented Lagrangian formulation of the optimization problem was employed in order to achieve bounded retardation of the overall optimization process. In particular, we focused on an application in aerodynamics to optimize the RAE2822 airfoil under transonic flow conditions. It turned out, that with the suggested single-step one-shot approach an optimization could be performed for this case by the numerical costs of just 4 flow calculations.

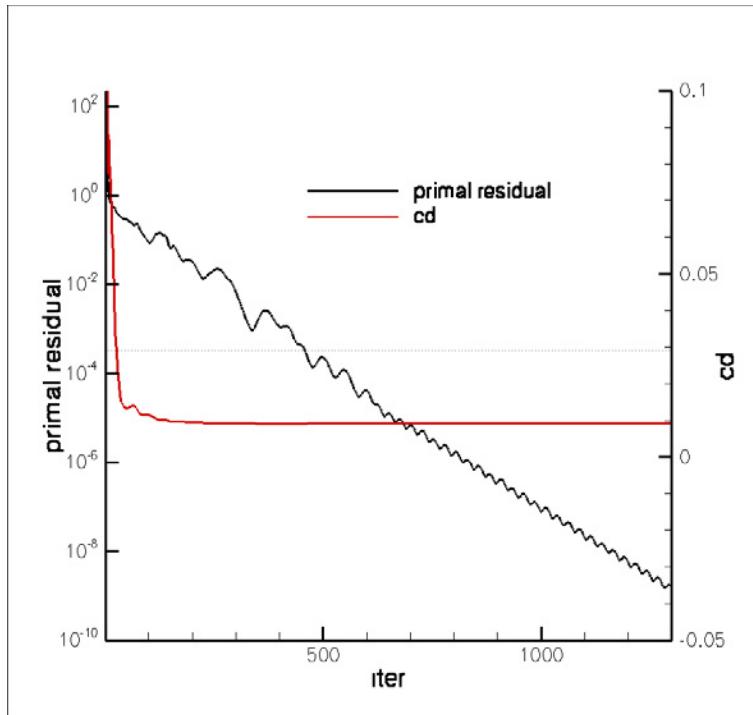


FIGURE 4. Convergence history of flow calculation for initial RAE2822 airfoil

### Acknowledgment

The authors gratefully acknowledge the support of the DFG Priority Program 1253 entitled Optimization With Partial Differential Equations.

### References

- [1] GAUGER N.R., WALTHER A., MOLDENHAUER C., WIDHALM M.: *Automatic Differentiation of an Entire Design Chain for Aerodynamic Shape Optimization*, Notes on Numerical Fluid Mechanics and Multidisciplinary Design, Vol. 96, pp. 454–461, 2007.
- [2] GRIEWANK A., JUEDES D., UTKE, J.: *ADOL-C: A package for the automatic differentiation of algorithms written in C/C++*. ACM Trans. Math. Softw., 22:131–167, 1996.
- [3] HAZRA S.B., SCHULZ V., BREZILLON J., GAUGER N.R.: *Aerodynamic shape optimization using simultaneous pseudo-time stepping*, Journal of Computational Physics, 204(1):46–64, 2005.
- [4] SCHLENKRICH S., WALTHER A., GAUGER N.R., HEINRICH R.: *Differentiating fixed point iterations with ADOL-C: Gradient calculation for fluid dynamics*, In H.G.

- Bock, E. Kostina, H.X. Phu, and R. Rannacher, editors, *Modeling, Simulation and Optimization of Complex Processes – Proceedings of the Third International Conference on High Performance Scientific Computing 2006*, pp. 499–508, 2008.
- [5] CHRISTIANSON B.: *Reverse accumulation and attractive fixed points*, Optimization Methods and Software, 3:311–326, 1994.
  - [6] HAMDI A., GRIEWANK A.: *Alternating approach for solving constrained optimization problem*, 2007.
  - [7] HAMDI A., GRIEWANK A.: *Reduced quasi-Newton method for simultaneous design and optimization*, Preprint-Number SPP1253-11-02, 2008.
  - [8] HAMDI A., GRIEWANK A.: *Properties of an augmented Lagrangian for design optimization*, submitted to Optimization Methods and Software, Preprint-Number SPP1253-11-01, 2008.
  - [9] GAUGER N.R., BREZILLON J.: *The continuous adjoint approach in aerodynamic shape optimization*, MEGAFLOW – Numerical Flow Simulation for Aircraft Design, Notes on Numerical Fluid Mechanics and Multidisciplinary Design, Vol. 89, pp. 181–193, Springer, 2005.
  - [10] MOHAMMADI B. PIRONNEAU O.: *Applied shape optimization for fluids*, Oxford Univ. Press, 2002.
  - [11] ROSSOW C.C.: *A flux splitting scheme for compressible and incompressible flows*, Journal of Computational Physics, 164:104–122, 2000.
  - [12] NOCEDAL J., WRIGHT S.J.: *Numerical Optimization*, Springer Series in Operational Research, 1999.
  - [13] GRIEWANK A.: *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*, SIAM, 2000.
  - [14] GAUGER N.R., GRIEWANK A., RIEHME J.: *Extension of fixed point PDE solvers for optimal design by one-shot method – With first applications to aerodynamic shape optimization*, European Journal of Computational Mechanics (REMN), Vol. 17, pp. 87–102, 2008.

Emre Özkaya  
 Humboldt-Universität zu Berlin  
 Institut für Mathematik  
 Unter den Linden 6  
 D-10099 Berlin, Germany  
 e-mail: ozkaya@math.hu-berlin.de

Nicolas R. Gauger  
 DLR Braunschweig und  
 Humboldt-Universität zu Berlin  
 Institut für Mathematik  
 Unter den Linden 6  
 D-10099 Berlin, Germany  
 e-mail: nicolas.gauger@dlr.de

# Shape Differentiability of Drag Functional for Compressible Navier-Stokes Equations

P.I. Plotnikov, E.V. Ruban and J. Sokolowski

**Abstract.** Compressible, stationary Navier-Stokes (N-S) equations are considered. It is shown, that the model is well posed, i.e., there exist weak solutions in bounded domains, subject to inhomogeneous boundary conditions. The shape sensitivity analysis is performed in the case of small perturbations of the so-called *approximate solutions*. The approximate solutions are determined from Stokes problem. The differentiability of solutions with respect to the coefficients of differential operators is obtained, therefore, the shape differentiability of the drag functional can be shown. The shape gradient of the drag functional is derived in the classical and useful for computations form, an appropriate adjoint state is introduced to this end. The proposed method of shape sensitivity analysis is general, and can be used to establish the well-posedness for distributed and boundary control problems as well as for inverse problems in the case of the state equations in the form of compressible N-S equations.

**Mathematics Subject Classification (2000).** Primary: 76N10; 35Q30; Secondary: 35A15; 76N15.

**Keywords.** Shape optimization, compressible Navier-Stokes equations, drag minimization, transport equations, necessary optimality conditions.

## 1. Preliminaries

Shape optimization for compressible N-S equations is considered in the literature to be important for applications, we refer the reader, e.g., to [9] for a review, and to [10] for general framework in incompressible case. The results presented in the paper lead, in particular, to the first-order optimality conditions for a class of shape optimization problems for compressible N-S equations.

The shape optimization for compressible N-S equations is a field of active research, e.g., in aerodynamics. The main difficulty in analysis of such optimization problems is the mathematical modeling, i.e., the lack of the existence results

for inhomogeneous boundary value problems in bounded domains [18]. The authors already proved the existence of an optimal shape for drag minimisation in three spatial dimensions under the Mosco convergence of admissible domains and assuming that the family of admissible domains is nonempty [17]. This is the general result on the compactness of the set of solutions to N-S equations for the admissible family of obstacles, we refer the reader to [14]–[21] for further details. The shape differentiability of solutions to N-S equations with respect to boundary perturbations is shown in [19], and leads to the optimality system for the shape optimisation problem under considerations.

### 1.1. Function spaces

In this paragraph we assemble some technical results which are used throughout of the paper. Function spaces play a central role, and we recall some notations, fundamental definitions and properties, which are classical. The proofs of some results given here can be found, e.g., in [19]. For our applications we need the results in three spatial dimensions, therefore, the space dimension stands  $d = 3$  in the paragraph on the embedding theorems.

Let  $\Omega$  be the whole space  $\mathbb{R}^3$  or a bounded domain in  $\mathbb{R}^3$  with the boundary  $\partial\Omega$  of class  $C^1$ . For an integer  $l \geq 0$  and for an exponent  $r \in [1, \infty)$ , we denote by  $H^{l,r}(\Omega)$  the Sobolev space endowed with the norm  $\|u\|_{H^{l,r}(\Omega)} = \sup_{|\alpha| \leq l} \|\partial^\alpha u\|_{L^r(\Omega)}$ . For real  $0 < s < 1$ , the fractional Sobolev space  $H^{s,r}(\Omega)$  is obtained by the interpolation between  $L^r(\Omega)$  and  $H^{1,r}(\Omega)$ , and consists of all measurable functions with the finite norm

$$\|u\|_{H^{s,r}(\Omega)} = \|u\|_{L^r(\Omega)} + |u|_{s,r,\Omega},$$

where

$$|u|_{s,r,\Omega}^r = \int_{\Omega \times \Omega} |x - y|^{-d-rs} |u(x) - u(y)|^r dx dy. \quad (1)$$

In the general case, the Sobolev space  $H^{l+s,r}(\Omega)$  is defined as the space of measurable functions with the finite norm  $\|u\|_{H^{l+s,r}(\Omega)} = \sup_{|\alpha| \leq l} \|\partial^\alpha u\|_{H^{s,r}(\Omega)}$ . For  $0 < s < 1$ , the Sobolev space  $H^{s,r}(\Omega)$  is, in fact the interpolation space  $[L^r(\Omega), H^{1,r}(\Omega)]_{s,r}$ .

Furthermore, the notation  $H_0^{l,r}(\Omega)$ , with an integer  $l$ , stands for the closed subspace of the space  $H^{l,r}(\Omega)$  of all functions  $u \in L^r(\Omega)$  which being extended by zero outside of  $\Omega$  belong to  $H^{l,r}(\mathbb{R}^3)$ .

Denote by  $\mathcal{H}_0^{0,r}(\Omega)$  and  $\mathcal{H}_0^{1,r}(\Omega)$  the subspaces of  $L^r(\mathbb{R}^3)$  and  $H^{1,r}(\mathbb{R}^3)$ , respectively, of all functions vanishing outside of  $\Omega$ . Obviously  $\mathcal{H}_0^{1,r}(\Omega)$  and  $H_0^{1,r}(\Omega)$  are isomorphic topologically and algebraically and we can identify them. However, we need the interpolation spaces  $\mathcal{H}_0^{s,r}(\Omega)$  for non-integers, in particular for  $s = 1/r$ .

**Definition 1.1.** *For all  $0 < s \leq 1$  and  $1 < r < \infty$ , we denote by  $\mathcal{H}_0^{s,r}(\Omega)$  the interpolation space  $[\mathcal{H}_0^{0,r}(\Omega), \mathcal{H}_0^{1,r}(\Omega)]_{s,r}$  endowed with one of two equivalent norms [19] defined by interpolation method.*

It follows from the definition of interpolation spaces that  $\mathcal{H}_0^{s,r}(\Omega) \subset H^{s,r}(\mathbb{R}^3)$  and for all  $u \in \mathcal{H}_0^{s,r}(\Omega)$ ,

$$\|u\|_{H^{s,r}(\mathbb{R}^3)} \leq c(r,s)\|u\|_{\mathcal{H}_0^{s,r}(\Omega)}, \quad u = 0 \text{ outside } \Omega. \quad (2)$$

In other words,  $\mathcal{H}_0^{s,r}(\Omega)$  consists of all elements  $u \in H^{s,r}(\Omega)$  such that the extension  $\bar{u}$  of  $u$  by 0 outside of  $\Omega$  have the finite  $[\mathcal{H}_0^{0,r}(\Omega), \mathcal{H}_0^{1,r}(\Omega)]_{s,r}$ -norm. We identify  $u$  and  $\bar{u}$  for the elements  $u \in \mathcal{H}_0^{s,r}(\Omega)$ . With this identification it follows that  $H_0^{1,r}(\Omega) \subset \mathcal{H}_0^{s,r}(\Omega)$  and the space  $C_0^\infty(\Omega)$  is dense in  $\mathcal{H}_0^{s,r}(\Omega)$ . It is worthy to note that for  $0 < s < 1$  and for  $1 < r < \infty$ , the function  $\bar{u}$  belongs to the space  $H^{s,r}(\mathbb{R}^3)$  if and only if  $u \in H^{s,r}(\Omega)$  and  $\text{dist}(x, \partial\Omega)^{-s}u \in L^r(\Omega)$ . We also point out that the interpolation space  $\mathcal{H}_0^{s,r}(\Omega)$  coincides with the Sobolev space  $H_0^{s,r}(\Omega)$  for  $s \neq 1/r$ . Recall that the standard space  $H_0^{s,r}(\Omega)$  is the completion of  $C_0^\infty(\Omega)$  in the  $H^{s,r}(\Omega)$ -norm.

*Embedding theorems.* For  $sr > d$  and  $0 \leq \alpha < s - r/d$ , the embedding  $H^{s,r}(\Omega) \hookrightarrow C^\alpha(\Omega)$  is continuous and compact. In particular, for  $sr > d$ , the Sobolev space  $H^{s,r}(\Omega)$  is a commutative Banach algebra, i.e., for all  $u, v \in H^{s,r}(\Omega)$ ,

$$\|uv\|_{H^{s,r}(\Omega)} \leq c(r,s)\|u\|_{H^{s,r}(\Omega)}\|v\|_{H^{s,r}(\Omega)}. \quad (3)$$

If  $sr < d$  and  $t^{-1} = r^{-1} - d^{-1}s$ , then the embedding  $H^{s,r}(\Omega) \hookrightarrow L^t(\Omega)$  is continuous. In particular, for  $\alpha \leq s$ ,  $(s - \alpha)r < d$  and  $\beta^{-1} = r^{-1} - d^{-1}(s - \alpha)$ ,

$$\|u\|_{H^{\alpha,\beta}(\Omega)} \leq c(r,s,\alpha,\beta,\Omega)\|u\|_{H^{s,r}(\Omega)}. \quad (4)$$

It follows from (2) that all the embedding inequalities remain true for the elements of the interpolation space  $\mathcal{H}_0^{s,r}(\Omega)$ .

*Duality.* We define

$$\langle u, v \rangle = \int_{\Omega} u v \, dx \quad (5)$$

for any functions such that the right-hand side make sense. For  $r \in (1, \infty)$ , each element  $v \in L^{r'}(\Omega)$ ,  $r' = r/(r-1)$ , determines the functional  $L_v$  of  $(\mathcal{H}_0^{s,r}(\Omega))'$  by the identity  $L_v(u)\langle u, v \rangle$ . We introduce the  $(-s, r')$ -norm of an element  $v \in L^{r'}(\Omega)$  to be by definition the norm of the functional  $L_v$ , that is

$$\|v\|_{\mathcal{H}^{-s,r'}(\Omega)} = \sup_{\substack{u \in \mathcal{H}_0^{s,r}(\Omega) \\ \|u\|_{\mathcal{H}_0^{s,r}(\Omega)}=1}} |\langle u, v \rangle|. \quad (6)$$

Let  $\mathcal{H}^{-s,r'}(\Omega)$  denote the completion of the space  $L^{r'}(\Omega)$  with respect to  $(-s, r')$ -norm. For an integer  $s$ ,  $\mathcal{H}^{-s,r'}(\Omega)$  is topologically and algebraically isomorphic to  $(H_0^{s,r}(\Omega))'$ . The same conclusion holds true for all  $s \in (0, 1)$ . Moreover, we can identify  $\mathcal{H}^{-s,r'}(\Omega)$  with the interpolation space  $[L^{r'}(\Omega), H_0^{-1,r'}(\Omega)]_{s,r}$ , see, e.g., [19]. With this denotations we have the duality principle

$$\|u\|_{\mathcal{H}_0^{s,r}(\Omega)} \sup_{\substack{v \in C_0^\infty(\Omega) \\ \|v\|_{\mathcal{H}^{-s,r'}(\Omega)}=1}} |\langle u, v \rangle|. \quad (7)$$

With applications to the theory of N-S equations in mind, we introduce the smaller dual space defined as follows. We identify the function  $v \in L^{r'}(\Omega)$  with the functional  $L_v \in (H^{s,r}(\Omega))'$  and denote by  $\mathbb{H}^{-s,r'}(\Omega)$  the completion of  $L^{r'}(\Omega)$  in the norm

$$\|v\|_{\mathbb{H}^{-s,r'}(\Omega)} := \sup_{\substack{u \in H^{s,r}(\Omega) \\ \|u\|_{H^{s,r}(\Omega)}=1}} |\langle u, v \rangle|. \quad (8)$$

In the sense of this identification the space  $C_0^\infty(\Omega)$  is dense in the interpolation space  $\mathbb{H}^{-s,r}(\Omega)$ . It follows immediately from the definition that

$$\mathbb{H}^{-s,r'}(\Omega) \subset (H^{s,r}(\Omega))' \subset \mathcal{H}^{-s,r'}(\Omega).$$

For an arbitrary bounded domain  $\Omega \subset \mathbb{R}^3$  with a Lipschitz boundary, we introduce the Banach spaces

$$X^{s,r} = H^{s,r}(\Omega) \cap H^{1,2}(\Omega), \quad Y^{s,r} = H^{s+1,r}(\Omega) \cap H^{2,2}(\Omega), \quad Z^{s,r} = \mathcal{H}^{s-1,r}(\Omega) \cap L^2(\Omega)$$

equipped with the norms

$$\begin{aligned} \|u\|_{X^{s,r}} &= \|u\|_{H^{s,r}(\Omega)} + \|u\|_{H^{1,2}(\Omega)}, \quad \|u\|_{Y^{s,r}} \|u\|_{H^{1+s,r}(\Omega)} + \|u\|_{H^{2,2}(\Omega)}, \\ \|u\|_{Z^{s,r}} &= \|u\|_{\mathcal{H}^{s-1,r}(\Omega)} + \|u\|_{L^2(\Omega)}. \end{aligned}$$

It can be easily seen that the embeddings  $Y^{s,r} \hookrightarrow X^{s,r} \hookrightarrow Z^{s,r}$  are compact and for  $sr > 3$ , each of the spaces  $X^{s,r}$  and  $Y^{s,r}$  is a commutative Banach algebra.

## 1.2. Stationary Navier-Stokes equations

We restrict ourselves to the inhomogeneous boundary value problems for compressible, stationary N-S equations. Such a modeling is considered in [14]-[19]. In particular, the well-posedness for inhomogeneous boundary value problems of elliptic-hyperbolic type is shown in [19]. Analysis is performed for small perturbations of the approximate solutions, which are determined from the Stokes problem. The existence and uniqueness of solutions close to approximate solution are proved, and in addition, the differentiability of solutions with respect to the coefficients of differential operators is shown in [19]. The results on the well-posedness of nonlinear problem are interesting on its own, and are used to obtain the shape differentiability of the drag functional for incompressible N-S equations. The shape gradient of the drag functional is derived in the classical and useful for computations form, an appropriate adjoint state is introduced to this end. The shape derivatives of solutions to the N-S equations are given by smooth functions, however the shape differentiability is shown in a weak norm. The method of analysis proposed in [19] is general, and can be used to establish the well-posedness for distributed and boundary control problems as well as for inverse problems in the case of the state equations in the form of compressible N-S equations. The differentiability of solutions to the N-S equations with respect to the data leads to the first-order necessary conditions for a broad class of optimization problems.

## 2. Shape optimisation for Navier-Stokes equations

We present an example of shape optimization in aerodynamics. Mathematical analysis of the drag minimization problem for compressible N-S equations can be found, e.g., in [17] on the domain continuity of solutions, and in [19] on the shape differentiability of the drag functional.

*Mathematical model in the form of N-S equations.* We assume that the viscous gas occupies the double-connected domain  $\Omega = B \setminus S$ , where  $B \subset \mathbb{R}^3$ , is a hold-all domain with the smooth boundary  $\Sigma = \partial B$ , and  $S \subset B$  is a compact obstacle. Furthermore, we assume that the velocity of the gas coincides with a given vector field  $\mathbf{U} \in C^\infty(\mathbb{R}^3)^3$  on the surface  $\Sigma$ . In this framework, the boundary of the flow domain  $\Omega$  is divided into the three subsets, inlet  $\Sigma_{\text{in}}$ , outgoing set  $\Sigma_{\text{out}}$ . In its turn the compact  $\Gamma = \Sigma_0 \cap \Sigma$  splits the surface  $\Sigma$  into three disjoint parts  $\Sigma = \Sigma_{\text{in}} \cup \Sigma_{\text{out}} \cup \Gamma$ . The problem is to find the velocity field  $\mathbf{u}$  and the gas density  $\varrho$  satisfying the following equations along with the boundary conditions

$$\Delta \mathbf{u} + \lambda \nabla \operatorname{div} \mathbf{u} = R \varrho \mathbf{u} \cdot \nabla \mathbf{u} + \frac{R}{\epsilon^2} \nabla p(\varrho) \text{ in } \Omega, \quad \operatorname{div}(\varrho \mathbf{u}) = 0 \text{ in } \Omega, \quad (9)$$

$$\mathbf{u} = \mathbf{U} \text{ on } \Sigma, \quad \mathbf{u} = 0 \text{ on } \partial S, \quad \varrho = \varrho_0 \text{ on } \Sigma_{\text{in}}, \quad (10)$$

where the pressure  $p = p(\varrho)$  is a smooth, strictly monotone function of the density,  $\epsilon$  is the Mach number,  $R$  is the Reynolds number,  $\lambda$  is the viscosity ratio, and  $\varrho_0$  is a positive constant.

*Drag minimization.* One of the main applications of the theory of compressible viscous flows is the optimal shape design in aerodynamics. The classical sample is the problem of the minimization of the drag of airfoil travelling in atmosphere with uniform speed  $\mathbf{U}_\infty$ . Recall that in our framework the hydro-dynamical force acting on the body  $S$  is defined by the formula,

$$\mathbf{J}(S) = - \int_{\partial S} \left( \nabla \mathbf{u} + (\nabla \mathbf{u})^* + (\lambda - 1) \operatorname{div} \mathbf{u} \mathbf{I} - \frac{R}{\epsilon^2} p \mathbf{I} \right) \cdot \mathbf{n} dS .$$

In a frame attached to the moving body the drag is the component of  $\mathbf{J}$  parallel to  $\mathbf{U}_\infty$ ,

$$J_D(S) = \mathbf{U}_\infty \cdot \mathbf{J}(S), \quad (11)$$

and the lift is the component of  $\mathbf{J}$  in the direction orthogonal to  $\mathbf{U}_\infty$ . For the fixed data, the drag can be regarded as a functional depending on the shape of the obstacle  $S$ . The minimization of the drag and the maximization of the lift are between shape optimization problems of some practical importance.

We present the shape differentiability of the drag functional with respect to the boundary variations, in the framework of the so-called approximate solutions of the Navier-Stokes equations. One of the technical difficulties in the analysis is the lack of uniqueness of solutions for the N-S equations under considerations.

### 2.1. Shape sensitivity analysis

We start with description of our framework for shape sensitivity analysis, or more general, for well-posedness of compressible NSE. To this end we choose the vector

field  $\mathbf{T} \in C^2(\mathbb{R}^3)^3$  vanishing in the vicinity of  $\Sigma$ , and define the mapping

$$y = x + \varepsilon \mathbf{T}(x), \quad (12)$$

which describes the perturbation of the shape of the obstacle. We refer the reader to [25] for more general framework and results in shape optimization. For small  $\varepsilon$ , the mapping  $x \rightarrow y$  takes diffeomorphically the flow region  $\Omega$  onto  $\Omega_\varepsilon = B \setminus S_\varepsilon$ , where the perturbed obstacle  $S_\varepsilon = y(S)$ . Let  $(\bar{\mathbf{u}}_\varepsilon, \bar{\varrho}_\varepsilon)$  be solutions to problem (9) in  $\Omega_\varepsilon$ . After substituting  $(\bar{\mathbf{u}}_\varepsilon, \bar{\varrho}_\varepsilon)$  into the formulae for  $\mathbf{J}$ , the drag becomes the function of the parameter  $\varepsilon$ . Our aim is, in fact, to prove that this function is well defined and differentiable at  $\varepsilon = 0$ . This leads to the first-order shape sensitivity analysis for solutions to compressible N-S equations. It is convenient to reduce such an analysis to the analysis of dependence of solutions with respect to the coefficients of the governing equations. To this end, we introduce the functions  $\mathbf{u}_\varepsilon(x)$  and  $\varrho_\varepsilon(x)$  defined in the unperturbed domain  $\Omega$  by the formulae

$$\mathbf{u}_\varepsilon(x) = \mathbf{N}\bar{\mathbf{u}}_\varepsilon(x + \varepsilon \mathbf{T}(x)), \quad \varrho_\varepsilon(x) = \bar{\varrho}_\varepsilon(x + \varepsilon \mathbf{T}(x)),$$

where

$$\mathbf{N}(x) = [\det (\mathbf{I} + \varepsilon \mathbf{T}'(x))(\mathbf{I} + \varepsilon \mathbf{T}'(x))]^{-1}. \quad (13)$$

is the adjugate matrix of the Jacobi matrix  $\mathbf{I} + \varepsilon \mathbf{T}'$ . Furthermore, we also use the notation  $\mathbf{g}(x) = \sqrt{\det \mathbf{N}}$ . It is easily to see that the matrices  $\mathbf{N}(x)$  depends analytically upon the small parameter  $\varepsilon$  and

$$\mathbf{N} = \mathbf{I} + \varepsilon \mathbf{D}(x) + \varepsilon^2 \mathbf{D}_1(\varepsilon, x), \quad (14)$$

where  $\mathbf{D} = \operatorname{div} \mathbf{T} \mathbf{I} - \mathbf{T}'$ . Calculations show that for  $\mathbf{u}_\varepsilon, \varrho_\varepsilon$ , the following boundary value problem is obtained

$$\Delta \mathbf{u}_\varepsilon + \nabla \left( \lambda \mathbf{g}^{-1} \operatorname{div} \mathbf{u}_\varepsilon - \frac{R}{\varepsilon^2} p(\varrho_\varepsilon) \right) = \mathcal{A} \mathbf{u}_\varepsilon + R \mathcal{B}(\varrho_\varepsilon, \mathbf{u}_\varepsilon, \mathbf{u}_\varepsilon) \text{ in } \Omega, \quad (15a)$$

$$\operatorname{div} (\varrho_\varepsilon \mathbf{u}_\varepsilon) = 0 \text{ in } \Omega, \quad (15b)$$

$$\mathbf{u}_\varepsilon = \mathbf{U} \text{ on } \Sigma, \quad \mathbf{u}_\varepsilon = 0 \text{ on } \partial S, \quad (15c)$$

$$\varrho_\varepsilon = \varrho_0 \text{ on } \Sigma_{\text{in}}. \quad (15d)$$

Here, the linear operator  $\mathcal{A}$  and the nonlinear mapping  $\mathcal{B}$  are defined in terms of  $\mathbf{N}$ ,

$$\begin{aligned} \mathcal{A}(\mathbf{u}) &= \Delta \mathbf{u} - \mathbf{N}^{-1} \operatorname{div} (\mathbf{g}^{-1} \mathbf{N} \mathbf{N}^* \nabla (\mathbf{N}^{-1} \mathbf{u})), \\ \mathcal{B}(\varrho, \mathbf{u}, \mathbf{w}) &= \varrho (\mathbf{N}^*)^{-1} (\mathbf{u} \nabla (\mathbf{N}^{-1} \mathbf{w})). \end{aligned} \quad (16)$$

## 2.2. Transport to the fixed domain by the change of variables

In this section we derive equations (15). We will write  $\mathbf{u}(y)$  and  $\varrho(y)$ ,  $y \in \Omega$ , and set

$$y = x + \varepsilon \mathbf{T}(x), \quad \mathbf{M}(x) = \mathbf{I} + \varepsilon \mathbf{T}'(x), \quad \tilde{\mathbf{u}}(x) = \mathbf{u}(y(x)), \quad \varrho_\varepsilon(x) = \varrho(y(x)).$$

Thus we get  $\mathbf{u}_\varepsilon = N\tilde{\mathbf{u}}$ . The Jacobi matrix  $\mathbf{M}$  is connected with the matrix  $\mathbf{N}$  by the relations

$$\det \mathbf{M} = (\det N)^{1/2} \equiv \mathbf{g}, \quad \mathbf{M} = \mathbf{g}\mathbf{N}^{-1} \quad (17)$$

For any function  $\phi \in C^1(\Omega)$  we have  $\nabla_y \phi = (\mathbf{M}^*)^{-1} \nabla_x \tilde{\phi}$ , where  $\tilde{\phi}(x) = \phi(y(x))$ . It follows from this that the identities

$$\begin{aligned} \int_{\tilde{\Omega}} (\operatorname{div}_y \mathbf{u})(y(x)) \tilde{\phi}(x) \det \mathbf{M} dx &= \int_{\tilde{\Omega}} (\operatorname{div}_y \mathbf{u})(y) \phi(y) \det \mathbf{M} dy = - \int_{\tilde{\Omega}} \mathbf{u} \cdot \nabla_y \phi dy \\ &= - \int_{\tilde{\Omega}} \tilde{\mathbf{u}} \cdot (\mathbf{M}^*)^{-1} \nabla_x \tilde{\phi}(x) \det \mathbf{M} dx = \int_{\tilde{\Omega}} \operatorname{div}_x ((\det \mathbf{M}) \mathbf{M}^{-1} \tilde{\mathbf{u}}) \tilde{\phi}(x) dx \end{aligned}$$

hold true for all  $\phi \in C_0^\infty(\Omega)$ . On the other hand, by virtue of (17) we have  $(\det \mathbf{M}) \mathbf{M}^{-1} \tilde{\mathbf{u}} = \mathbf{u}_\varepsilon(x)$ . This leads to the equalities

$$\begin{aligned} (\operatorname{div}_y \mathbf{u})(y(x)) &= \mathbf{g}^{-1} \operatorname{div}_x (\mathbf{N} \tilde{\mathbf{u}}(x)) \equiv \mathbf{g}^{-1} \operatorname{div}_x \mathbf{u}_\varepsilon(x), \\ \operatorname{div}_y (\varrho \mathbf{u})(y(x)) &= \mathbf{g}^{-1} \operatorname{div}_x (\varrho_\varepsilon \mathbf{u}_\varepsilon), \end{aligned} \quad (18)$$

which imply the modified mass balance equation (15b). From (18) and the identity  $(\mathbf{M}^*)^{-1} = \mathbf{g}^{-1} \mathbf{N}^*$  we obtain

$$\nabla \left( \lambda \operatorname{div} \mathbf{u} - \frac{R}{\epsilon^2} p(\varrho) \right) = \mathbf{g}^{-1} \mathbf{N}^* \nabla \left( \lambda \mathbf{g}^{-1} \operatorname{div} \mathbf{u}_\varepsilon - \frac{R}{\epsilon^2} p(\varrho_\varepsilon) \right). \quad (19)$$

Combining (18) with the identity  $\Delta = \operatorname{div} \nabla$  we obtain

$$\begin{aligned} \Delta \mathbf{u}(y) &= \mathbf{g}^{-1} \operatorname{div} (\mathbf{N}(\mathbf{M}^*)^{-1} \nabla \tilde{\mathbf{u}}) \\ &= \mathbf{g}^{-1} \operatorname{div} (\mathbf{g}^{-1} \mathbf{N} \mathbf{N}^* \nabla (\mathbf{N}^{-1} \mathbf{u}_\varepsilon)) = \mathbf{g}^{-1} \mathbf{N}^* (\Delta \mathbf{u}_\varepsilon - \mathcal{A}(\mathbf{u}_\varepsilon)) \end{aligned} \quad (20)$$

Next note that the components  $(\mathbf{u} \nabla \mathbf{u})_i$  of the vector  $\mathbf{u} \nabla \mathbf{u}$  satisfy the equalities

$$(\mathbf{u} \nabla \mathbf{u})_i = \mathbf{u} \cdot \nabla_y u_i = \tilde{\mathbf{u}} \cdot ((\mathbf{M}^*)^{-1} \nabla \tilde{u}_i) = \mathbf{g}^{-1} \mathbf{N} \tilde{\mathbf{u}} \cdot \nabla \tilde{u}_i = \mathbf{g}^{-1} \mathbf{u}_\varepsilon \cdot \nabla (\mathbf{N}^{-1} \mathbf{u}_\varepsilon)_i$$

This gives

$$\varrho \mathbf{u} \nabla \mathbf{u} = \mathbf{g}^{-1} \mathbf{N}^* \mathcal{B}(\varrho_\varepsilon, \mathbf{u}_\varepsilon, \mathbf{u}_\varepsilon). \quad (21)$$

Substituting (19)–(21) into mass balance equation (9) and multiplying both sides of the resulting equality by  $\mathbf{g}(\mathbf{N}^*)^{-1}$  we obtain modified equation (15a).

The specific structure of the matrix  $\mathbf{N}$  does not play any particular role in the further analysis. Therefore, we consider a general problem of the existence, uniqueness and dependence on coefficients of the solutions to equations (15) under the assumption that  $\mathbf{N}$  is a given matrix-valued function which is close, in an appropriate norm, to the identity mapping  $\mathbf{I}$  and coincides with  $\mathbf{I}$  in the vicinity of  $\Sigma$ . By abuse of notations, we write simply  $\mathbf{u}$  and  $\varrho$  instead of  $\mathbf{u}_\varepsilon$  and  $\varrho_\varepsilon$ , when studying the well-posedness and dependence on  $\mathbf{N}$ . Before formulation of main results we write the governing equation in more transparent form using the change of

unknown functions proposed by M. Padula. To do so we introduce *the effective viscous pressure*

$$q = \frac{R}{\epsilon^2} p(\varrho) - \lambda g^{-1} \operatorname{div} \mathbf{u},$$

and rewrite equations (15) in the equivalent form

$$\Delta \mathbf{u} - \nabla q = \mathcal{A}(\mathbf{u}) + R\mathcal{B}(\varrho, \mathbf{u}, \mathbf{u}) \text{ in } \Omega, \quad (22a)$$

$$\operatorname{div} \mathbf{u} = a\sigma_0 p(\varrho) - \frac{gq}{\lambda} \text{ in } \Omega, \quad (22b)$$

$$\mathbf{u} \cdot \nabla \varrho + g\sigma_0 p(\varrho) \varrho = \frac{gq}{\lambda} \varrho \text{ in } \Omega, \quad (22c)$$

$$\mathbf{u} = \mathbf{U} \text{ on } \Sigma, \quad \mathbf{u} = 0 \text{ on } \partial S, \quad (22d)$$

$$\varrho = \varrho_0 \text{ on } \Sigma_{in}. \quad (22e)$$

where  $\sigma_0 = R/(\lambda\epsilon^2)$ . In the new variables  $(\mathbf{u}, q, \varrho)$  the expression for the force  $\mathbf{J}$  reads

$$\mathbf{J} = - \int_{\Omega} [\mathbf{g}^{-1} (\mathbf{N}^* \nabla (\mathbf{N}\mathbf{u}) + \nabla (\mathbf{N}\mathbf{u})^* \mathbf{N} - \operatorname{div} \mathbf{u}) - q - R\varrho \mathbf{u} \otimes \mathbf{u}] \mathbf{N}^* \nabla \eta dx. \quad (23)$$

where  $\eta \in C^\infty(\Omega)$  is an arbitrary function, which is equal to 1 in an open neighborhood of the obstacle  $S$  and 0 in a vicinity of  $\Sigma$ . The value of  $\mathbf{J}$  is independent of the choice of the function  $\eta$ .

### 3. Perturbations of the approximate solutions

We assume that  $\lambda \gg 1$  and  $R \ll 1$ , which corresponds to almost incompressible flow with low Reynolds number. In such a case, the *approximate solutions* to problem (22) can be chosen in the form  $(\varrho_0, \mathbf{u}_0, q_0)$ , where  $\varrho_0$  is a constant in boundary condition (22e), and  $(\mathbf{u}_0, q_0)$  is a solution to the boundary value problem for the Stokes equations,

$$\begin{aligned} \Delta \mathbf{u}_0 - \nabla q_0 &= 0, \quad \operatorname{div} \mathbf{u}_0 = 0 \text{ in } \Omega, \\ \mathbf{u}_0 &= \mathbf{U} \text{ on } \Sigma, \quad \mathbf{u}_0 = 0 \text{ on } \partial S, \quad \Pi q_0 = q_0. \end{aligned} \quad (24)$$

In our notations  $\Pi$  is the projector,

$$\Pi u = u - \frac{1}{\operatorname{meas} \Omega} \int_{\Omega} u dx.$$

Equations (24) can be obtained as the limit of equations (22) for the passage  $\lambda \rightarrow \infty$ ,  $R \rightarrow 0$ . It follows from the standard elliptic theory that for the boundary  $\partial\Omega \in C^\infty$ , we have  $(\mathbf{u}_0, q_0) \in C^\infty(\Omega)$ . We look for solutions to problem (22) in the form

$$\mathbf{u} = \mathbf{u}_0 + \mathbf{v}, \quad \varrho = \varrho_0 + \varphi, \quad q = q_0 + \lambda\sigma_0 p(\varrho_0) + \pi + \lambda m, \quad (25)$$

with the unknowns functions  $\vartheta = (\mathbf{v}, \pi, \varphi)$  and the unknown constant  $m$ . Substituting (25) into (22) we obtain the following boundary problem for  $\vartheta$ ,

$$\begin{aligned} \Delta \mathbf{v} - \nabla \pi &= \mathcal{A}(\mathbf{u}) + R\mathcal{B}(\varrho, \mathbf{u}, \mathbf{u}) \text{ in } \Omega, \\ \operatorname{div} \mathbf{v} &= \mathfrak{g} \left( \frac{\sigma}{\varrho_0} \varphi - \Psi[\vartheta] - m \right) \text{ in } \Omega, \\ \mathbf{u} \cdot \nabla \varphi + \sigma \varphi \Psi_1[\vartheta] + m \mathfrak{g} \varrho &\text{ in } \Omega, \\ \mathbf{v} = 0 \text{ on } \partial\Omega, \quad \varphi = 0 \text{ on } \Sigma_{\text{in}}, \quad \Pi \pi = \pi, \end{aligned} \tag{26a}$$

where

$$\begin{aligned} \Psi_1[\vartheta] &= \mathfrak{g} \left( \varrho \Psi[\vartheta] - \frac{\sigma}{\varrho_0} \varphi^2 \right) + \sigma \varphi (1 - \mathfrak{g}), \quad \Psi[\vartheta] \frac{q_0 + \pi}{\lambda} - \frac{\sigma}{p'(\varrho_0) \varrho_0} H(\varphi), \\ \sigma &= \sigma_0 p'(\varrho_0) \varrho_0, \quad H(\varphi) = p(\varrho_0 + \varphi) - p(\varrho_0) - p'(\varrho_0) \varphi, \end{aligned}$$

the vector field  $\mathbf{u}$  and the function  $\varrho$  are given by (25). Finally, we specify the constant  $m$ . In our framework, in contrast to the case of homogeneous boundary problem, the solution to such a problem is not trivial. Note that, since  $\operatorname{div} \mathbf{v}$  is of the null mean value, the right-hand side of equation (26a)<sub>3</sub> must satisfy the compatibility condition

$$m \int_{\Omega} \mathfrak{g} dx = \int_{\Omega} \mathfrak{g} \left( \frac{\sigma}{\varrho_0} \varphi - \Psi[\vartheta] \right) dx,$$

which formally determines  $m$ . This choice of  $m$  leads to essential mathematical difficulties. To make this issue clear note that in the simplest case  $\mathfrak{g} = 1$  we have  $m = \varrho_0^{-1} \sigma (\mathbf{I} - \Pi) \varphi + O(|\vartheta|^2, \lambda^{-1})$ , and the principal linear part of the governing equations (26a) becomes

$$\begin{pmatrix} \Delta & -\nabla & 0 \\ \operatorname{div} & 0 & -\frac{\sigma}{\varrho_0} \\ 0 & 0 & \mathbf{u} \nabla + \sigma \end{pmatrix} \begin{pmatrix} \mathbf{v} \\ \pi \\ \varphi \end{pmatrix} + \begin{pmatrix} 0 \\ m \\ -m \varrho_0 \end{pmatrix} \sim \begin{pmatrix} \Delta \mathbf{v} - \nabla \pi \\ \operatorname{div} \mathbf{v} - \frac{\sigma}{\varrho_0} \Pi \varphi \\ \mathbf{u} \nabla \varphi + \sigma \Pi \varphi \end{pmatrix}$$

Hence, the question of solvability of the linearized equations derived for (26) can be reduced to the question of solvability of the boundary value problem for nonlocal transport equation

$$\mathbf{u} \nabla \varphi + \sigma \Pi \varphi = f,$$

which is very difficult because of the loss of maximum principle. In fact, this question is concerned with the problem of the control of the total gas mass in compressible flows. Recall that the absence of the mass control is the main obstacle for proving the global solvability of inhomogeneous boundary problems for compressible N-S equations, we refer to [8] for discussion. In order to cope with this difficulty we write the compatibility condition in a sophisticated form, which allows us to control the total mass of the gas. To this end we introduce the auxiliary function  $\zeta$  satisfying the equations

$$-\operatorname{div}(\mathbf{u} \zeta) + \sigma \zeta = \sigma \mathfrak{g} \text{ in } \Omega, \quad \zeta = 0 \text{ on } \Sigma_{\text{out}}, \tag{26b}$$

and fix the constant  $m$  as follows

$$m = \varkappa \int_{\Omega} (\varrho_0^{-1} \Psi_1[\vartheta] \zeta - \mathbf{g} \Psi[\vartheta]) dx, \quad \varkappa = \left( \int_{\Omega} \mathbf{g}(1 - \zeta - \varrho_0^{-1} \zeta \varphi) dx \right)^{-1}. \quad (26c)$$

In this way the auxiliary function  $\zeta$  becomes an integral part of the solution to problem (26).

### 3.1. Existence and uniqueness theory

Denote by  $E$  the closed subspace of the Banach space  $Y^{s,r}(\Omega)^3 \times X^{s,r}(\Omega)^2$  in the following form

$$E = \{\vartheta = (\mathbf{v}, \pi, \varphi) : \mathbf{v} = 0 \text{ on } \partial\Omega, \quad \varphi = 0 \text{ on } \Sigma_{\text{in}}, \quad \Pi\pi = \pi\}, \quad (27)$$

and denote by  $\mathcal{B}_\tau \subset E$  the closed ball of radius  $\tau$  centered at 0. Next, note that for  $sr > 3$ , elements of the ball  $\mathcal{B}_\tau$  satisfy the inequality

$$\|\mathbf{v}\|_{C^1(\Omega)} + \|\pi\|_{C(\Omega)} + \|\varphi\|_{C(\Omega)} \leq c_e(r, s, \Omega) \|\vartheta\|_E \leq c_e \tau, \quad (28)$$

where the norm in  $E$  is defined by

$$\|\vartheta\|_E = \|\mathbf{v}\|_{Y^{s,r}(\Omega)} + \|\pi\|_{X^{s,r}(\Omega)} + \|\varphi\|_{X^{s,r}(\Omega)}.$$

**Theorem 3.1.** *Assume that the surface  $\Sigma$  and given vector field  $\mathbf{U}$  satisfy emergent field conditions. Furthermore, let  $\sigma^*$ ,  $\tau^*$  be given constants determined in [19], and let positive numbers  $r$ ,  $s$ ,  $\sigma$  satisfy the inequalities*

$$1/2 < s \leq 1, \quad 1 < r < 3/(2s-1), \quad sr > 3, \sigma > \sigma^*. \quad (29)$$

*Then there exists  $\tau_0 \in (0, \tau^*]$ , depending only on  $\mathbf{U}, \Omega, r, s, \sigma$ , such that for all*

$$\tau \in (0, \tau_0], \quad \lambda^{-1}, R \in (0, \tau^2], \quad \|\mathbf{N} - \mathbf{I}\|_{C^2(\Omega)} \leq \tau^2, \quad (30)$$

*problem (26), with  $\mathbf{u}_0$  given by (24), has a unique solution  $\vartheta \in \mathcal{B}_\tau$ . Moreover, the auxiliary function  $\zeta$  and the constants  $\varkappa, m$  admit the estimates*

$$\|\zeta\|_{X^{s,r}} + |\varkappa| \leq c, \quad |m| \leq c\tau < 1, \quad (31)$$

*where the constant  $c$  depends only on  $\mathbf{U}, \Omega, r, s$  and  $\sigma$ .*

### 3.2. Material derivatives of solutions

Theorem 3.1 guarantees the existence and uniqueness of solutions to problem (26) for all  $\mathbf{N}$  close to the identity matrix  $\mathbf{I}$ . The totality of such solutions can be regarded as the mapping from  $\mathbf{N}$  to the solution of the N-S equations. The natural question is the smoothness properties of this mapping, in particular its differentiability. With application to shape optimization problems in mind, we consider the particular case where the matrices  $\mathbf{N}$  depend on the small parameter  $\varepsilon$  and have representation (14). We assume that  $C^1$  norms of the matrix-valued functions  $\mathbf{D}$  and  $\mathbf{D}_1(\varepsilon)$  in (14) have a majorant independent of  $\varepsilon$ . By virtue of Theorem 3.1, there are the positive constants  $\varepsilon_0$  and  $\tau$  such that for all sufficiently small  $R, \lambda^{-1}$  and  $\varepsilon \in [0, \varepsilon_0]$ , problem (26) with  $\mathbf{N} = \mathbf{N}(\varepsilon)$  has a unique solution  $\vartheta(\varepsilon) = (\mathbf{v}(\varepsilon), \pi(\varepsilon), \varphi(\varepsilon)), \zeta(\varepsilon), m(\varepsilon)$ , which admits the estimate

$$\|\vartheta(\varepsilon)\|_E + |m(\varepsilon)| \leq c\tau, \quad \|\zeta(\varepsilon)\|_{X^{s,r}} \leq c, \quad (32)$$

where the constant  $c$  is independent of  $\varepsilon$ , and the Banach space  $E$  is defined by (27). Denote the solution for  $\varepsilon = 0$  by  $(\vartheta(0), m(0), \zeta(0))$  by  $(\vartheta, m, \zeta)$ , and define the finite differences with respect to  $\varepsilon$

$$(\mathbf{w}_\varepsilon, \omega_\varepsilon, \psi_\varepsilon) \varepsilon^{-1}(\vartheta - \vartheta(\varepsilon)), \quad \xi_\varepsilon = \varepsilon^{-1}(\zeta - \zeta(\varepsilon)), \quad n_\varepsilon = \varepsilon^{-1}(m - m(\varepsilon)).$$

Formal calculations shows that the limit  $(\mathbf{w}, \omega, \psi, \xi, n) = \lim_{\varepsilon \rightarrow 0} (\mathbf{w}_\varepsilon, \omega_\varepsilon, \psi_\varepsilon, \xi_\varepsilon, n_\varepsilon)$  is a solution to linearized equations

$$\begin{aligned} \Delta \mathbf{w} - \nabla \omega &= R \mathcal{C}_0(\mathbf{w}, \psi) + \mathcal{D}_0(\mathbf{D}) \text{ in } \Omega, \\ \operatorname{div} \mathbf{w} &= b_{21}^0 \psi - b_{22}^0 \omega + b_{23}^0 n + b_{30}^0 \mathfrak{d} \text{ in } \Omega, \\ \mathbf{u} \nabla \psi + \sigma \psi &= -\mathbf{w} \cdot \nabla \varphi + b_{11}^0 \psi + b_{12}^0 \omega + b_{13}^0 n + b_{10}^0 \mathfrak{d} \text{ in } \Omega, \\ &\quad - \operatorname{div}(\mathbf{u} \xi) + \sigma \xi = \operatorname{div}(\zeta \mathbf{w}) + \sigma \mathfrak{d} \text{ in } \Omega, \\ \mathbf{w} &= 0 \text{ on } \partial \Omega, \quad \psi = 0 \text{ on } \Sigma_{\text{in}}, \quad \xi = 0 \text{ on } \Sigma_{\text{out}}, \\ \omega - \Pi \omega &= 0, \quad n = \varkappa \int_{\Omega} (b_{31}^0 \psi + b_{32}^0 \omega + b_{34}^0 \xi + b_{30}^0 \mathfrak{d}) dx, \end{aligned} \quad (33)$$

where  $\mathfrak{d} = 1/2 \operatorname{Tr} \mathbf{D}$ , the variable coefficients  $b_{ij}^0$  and the operators  $\mathcal{C}_0$ ,  $\mathcal{D}_0$ , are defined by the formulae

$$\begin{aligned} b_{11}^0 &= \Psi[\vartheta] - \varrho H'(\varphi) + m - \frac{2\sigma}{\varrho_0} \varphi, \quad b_{12}^0 = \lambda^{-1} \varrho, \quad b_{13}^0 = \varrho, \\ b_{10}^0 &= \varrho \Psi[\vartheta] - \frac{\sigma}{\varrho_0} \varphi^2 - \sigma \varphi + m \varrho, \quad b_{21}^0 = \frac{\sigma}{\varrho_0} \psi_0 + H'(\varphi), \\ b_{22}^0 &= -\lambda^{-1}, \quad b_{23}^0 = -1, \quad b_{20}^0 \sigma \varphi \varrho_0^{-1} - \Psi[\vartheta] - m, \\ b_{31}^0 &= \varrho_0^{-1} \zeta \left( \Psi[\vartheta] - \varrho H'(\varphi) - \frac{2\sigma}{\varrho_0} \varphi \right) - H'(\varphi) + m \varrho_0^{-1} \zeta, \\ b_{32}^0 &= (\lambda \varrho_0)^{-1} \varrho \zeta b_{12}^0 + \lambda^{-1}, \quad b_{34}^0 = \varrho_0^{-1} \Psi_1[\vartheta] + m(1 + \varrho_0^{-1} \varphi), \\ b_{30}^0 &= \varrho_0^{-1} \zeta (\mathfrak{d}_0 - m \varrho) + \Psi[\vartheta] - m(1 - \zeta - \varrho_0^{-1} \zeta \varphi), \end{aligned} \quad (34)$$

$$\mathcal{C}_0(\psi, \mathbf{w}) = R \psi \mathbf{u} \nabla \mathbf{u} + R \varrho \mathbf{w} \nabla \mathbf{u}, + R \varrho \mathbf{u} \nabla \mathbf{w}, \quad (35)$$

$$\begin{aligned} \mathcal{D}_0(\mathbf{D}) &= R \mathbf{u} \nabla(\mathbf{D} \mathbf{u}) + R \mathbf{D}^*(\mathbf{u} \nabla \mathbf{u}) \\ &\quad + \operatorname{div} \left( (\mathbf{D} + \mathbf{D}^*) \nabla \mathbf{u} - \frac{1}{2} \operatorname{Tr} \mathbf{D} \nabla \mathbf{u} \right) - \mathbf{D} \Delta \mathbf{u} - \Delta(\mathbf{D} \mathbf{u}). \end{aligned} \quad (36)$$

The justification of the formal procedure meets the serious problems, since the smoothness of solutions to problem (26) is not sufficient for the well-posedness of problem (33) in the standard weak formulation. In order to cope with this difficulty we define *very weak solutions* to problem (33). The construction of such solutions is based on the following lemma [19]. The lemma is given in  $\mathbb{R}^3$ , for our application  $d = 3$ .

**Lemma 3.2.** *Let  $\Omega \subset \mathbb{R}^3$  be a bounded domain with the Lipschitz boundary, let exponents  $s$  and  $r$  satisfy the inequalities  $sr > d$ ,  $1/2 \leq s \leq 1$  and  $\varphi, \zeta \in H^{s,r}(\Omega) \cap$*

$H^{1,2}(\Omega)$ ,  $\mathbf{w} \in \mathcal{H}_0^{1-s,r'}(\Omega) \cap H_0^{1,2}(\Omega)$ . Then there is a constant  $c$  depending only on  $s, r$  and  $\Omega$ , such that the trilinear form

$$\mathfrak{B}(\mathbf{w}, \varphi, \varsigma) = - \int_{\Omega} \varsigma \mathbf{w} \cdot \nabla \varphi \, dx$$

satisfies the inequality

$$|\mathfrak{B}(\mathbf{w}, \varphi, \varsigma)| \leq c \|\mathbf{w}\|_{\mathcal{H}_0^{1-s,r'}(\Omega)} \|\varphi\|_{H^{s,r}(\Omega)} \|\varsigma\|_{H^{s,r}(\Omega)}, \quad (37)$$

and can be continuously extended to  $\mathfrak{B} : \mathcal{H}_0^{1-s,r'}(\Omega)^3 \times H^{s,r}(\Omega)^2 \mapsto \mathbb{R}$ . In particular, we have  $\varsigma \nabla \varphi \in \mathcal{H}^{s-1,r}(\Omega)$  and  $\|\varsigma \nabla \varphi\|_{H^{1-s,r}(\Omega)} \leq c \|\varphi\|_{H^{s,r}(\Omega)} \|\varsigma\|_{H^{s,r}(\Omega)}$ .

**Definition 3.3.** The vector field  $\mathbf{w} \in \mathcal{H}_0^{1-s,r'}(\Omega)^3$ , functionals  $(\omega, \psi, \xi) \in \mathbb{H}^{-s,r'}(\Omega)^3$  and constant  $n$  are said to be a weak solution to problem (33), if  $\langle \omega, 1 \rangle = 0$  and the identities

$$\begin{aligned} & \int_{\Omega} \mathbf{w} \left( \mathbf{H} - R\varrho \nabla \mathbf{u} \cdot \mathbf{h} + R\varrho \nabla \mathbf{h}^* \mathbf{u} \right) dx - \mathfrak{B}(\mathbf{w}, \varphi, \varsigma) - \mathfrak{B}(\mathbf{w}, v, \zeta) \\ & + \langle \omega, G - b_{12}^0 \varsigma - b_{22}^0 g - \varkappa b_{32}^0 \rangle + \langle \psi, F - b_{11}^0 \varsigma - b_{21}^0 g - \varkappa b_{31}^0 - R\mathbf{u} \cdot \nabla \mathbf{u} \cdot \mathbf{h} \rangle \\ & \quad + \langle \xi, M - \varkappa b_{34}^0 \rangle + n(1 - \langle 1, b_{13}^0 \varsigma \rangle) \\ & = \langle \mathfrak{d}, b_{10}^0 \varsigma + b_{20}^0 g + \varkappa b_{30}^0 + \sigma v \rangle + \langle \mathcal{D}_0, \mathbf{h} \rangle. \end{aligned} \quad (38)$$

hold true for all  $(\mathbb{H}, G, F, M) \in (C^\infty(\Omega))^6$  such that  $G = \Pi G$ . Here  $\mathfrak{d} = 1/2 \operatorname{Tr} \mathbf{D}$ , the test functions  $\mathbf{h}$ ,  $g$ ,  $\varsigma$ ,  $v$  are defined by the solutions to adjoint problems

$$\Delta \mathbf{h} - \nabla g = \mathbb{H}, \quad \operatorname{div} \mathbf{h} = G, \quad \mathcal{L}^* \varsigma = F, \quad \mathcal{L} v = M \text{ in } \Omega, \quad (39)$$

$$\mathbf{h} = 0 \text{ on } \partial\Omega, \quad \Pi g = g, \varsigma = 0 \text{ on } \Sigma_{\text{out}}, \quad v = 0 \text{ on } \Sigma_{\text{in}}. \quad (40)$$

We are now in a position to formulate the third main result of this paper.

**Theorem 3.4.** Under the above assumptions,

$$\begin{aligned} \mathbf{w}_\varepsilon & \rightarrow \mathbf{w} \text{ weakly in } \mathcal{H}_0^{1-s,r'}(\Omega), \quad n_\varepsilon \rightarrow n \text{ in } \mathbb{R}, \\ \psi_\varepsilon & \rightarrow \psi, \quad \omega_\varepsilon \rightarrow \omega, \quad \xi_\varepsilon \rightarrow \xi \text{ (*)-weakly in } \mathbb{H}^{-s,r'}(\Omega) \text{ as } \varepsilon \rightarrow 0, \end{aligned} \quad (41)$$

where the limits, vector field  $\mathbf{w}$ , functionals  $\psi, \omega, \xi$ , and the constant  $n$  are given by the weak solution to problem (33).

Note that the matrices  $\mathbf{N}(\varepsilon)$  defined by equalities (13) meet all requirements of Theorem 3.4, and in the special case we have in representation (14)

$$\mathbf{D}(x) = \operatorname{div} \mathbf{T}(x) \mathbf{I} - \mathbf{T}'(x). \quad (42)$$

Therefore, Theorem 3.4, together with the formulae (11) and (23), imply the existence of the shape derivative for the drag functional at  $\varepsilon = 0$ . Straightforward calculations lead to the following result.

**Theorem 3.5.** *Under the assumptions of Theorem 3.4, there exists the shape derivative*

$$\frac{d}{d\varepsilon} J_D(S_\varepsilon) \Big|_{\varepsilon=0} = L_e(\mathbf{T}) + L_u(\mathbf{w}, \omega, \psi),$$

where the linear forms  $L_e$  and  $L_u$  are defined by the equalities

$$\begin{aligned} L_e(\mathbf{T}) &= \int_{\Omega} \operatorname{div} \mathbf{T} (\nabla \mathbf{u} + \nabla \mathbf{u}^* - \operatorname{div} \mathbf{u} \mathbf{I}) \mathbf{U}_\infty dx \\ &\quad - \int_{\Omega} [\nabla \mathbf{u} + \nabla \mathbf{u}^* - \operatorname{div} \mathbf{u} - q \mathbf{I} - R \varrho \mathbf{u} \otimes \mathbf{u}] \mathbf{D} \nabla \eta \cdot \mathbf{U}_\infty dx \\ &\quad - \int_{\Omega} [\mathbf{D}^* \nabla \mathbf{u} + \nabla \mathbf{u}^* \mathbf{D} + \nabla (\mathbf{D} \mathbf{u}) + \nabla (\mathbf{D} \mathbf{u})^*] \nabla \eta \cdot \mathbf{U}_\infty dx \end{aligned}$$

and

$$\begin{aligned} L_u(\mathbf{w}, \omega, \psi) &= \int_{\Omega} \mathbf{w} [\Delta \eta \mathbf{U}_\infty + R \varrho (\mathbf{u} \cdot \nabla \eta) \mathbf{U}_\infty + R \varrho (\mathbf{u} \cdot \mathbf{U}_\infty) \nabla \eta] dx \\ &\quad + \langle \omega, \nabla \eta \cdot \mathbf{U}_\infty \rangle + R \langle \psi, (\mathbf{u} \cdot \nabla \eta) (\mathbf{u} \cdot \mathbf{U}_\infty) \rangle. \end{aligned}$$

While  $L_e$  depends directly on the vector field  $\mathbf{T}$ , the linear form  $L_u$  depends on the weak solution  $(\mathbf{w}, \psi, \omega)$  to problem (33), thus depends on the *direction*  $\mathbf{T}$  in a very implicit manner, which is inconvenient for applications. In order to cope with this difficulty, we define the *adjoint state*  $\mathbf{Y} = (\mathbf{h}, g, \varsigma, v, l)^\top$  given as a solution to the linear equation

$$\mathfrak{L}\mathbf{Y} - \mathfrak{U}\mathbf{Y} - \mathfrak{V}\mathbf{Y} = \Theta, \quad (43)$$

supplemented with boundary conditions (40). Here the operators  $\mathfrak{L}$ ,  $\mathfrak{U}$ ,  $\mathfrak{V}$  and the vector field  $\Theta$  are defined by

$$\begin{aligned} \mathfrak{L} &= \begin{pmatrix} \Delta & -\nabla & 0 & 0 & 0 \\ \operatorname{div} & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathcal{L}^* & 0 & 0 \\ 0 & 0 & 0 & \mathcal{L} & 0 \\ 0 & 0 & -\mathbb{B}_{13} & 0 & 1 \end{pmatrix}, \quad \mathfrak{U} = \begin{pmatrix} 0 & 0 & -\nabla \varphi & -\zeta \nabla & 0 \\ 0 & 0 & \Pi_{21} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \\ \mathfrak{V} &= \begin{pmatrix} R \varrho (\nabla \mathbf{u} - \mathbf{u} \nabla) & 0 & 0 & 0 & 0 \\ 0 & -\lambda^{-1} \Pi & 0 & 0 & \varkappa \Pi b_{32}^0 \\ R \mathbf{u} \cdot \nabla \mathbf{u} & b_{12}^0 & b_{11}^0 & 0 & \varkappa b_{31}^0 \\ 0 & 0 & 0 & 0 & \varkappa b_{34}^0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \end{aligned}$$

$$\begin{aligned} \Theta &= (\Delta \eta \mathbf{U}_\infty + R \varrho (\nabla \eta \otimes \mathbf{U}_\infty + \mathbf{U}_\infty \otimes \nabla \eta) \mathbf{u}, \Pi (\nabla \eta \cdot \mathbf{U}_\infty), R (\mathbf{u} \nabla \eta) (\mathbf{u} \mathbf{U}_\infty), 0, 0), \\ \Pi_{2i}(\cdot) &= \Pi(b_{2i}^0(\cdot)), \quad \mathbb{B}_{13}(\cdot) = \langle 1, b_{13}^0(\cdot) \rangle \end{aligned}$$

The following theorem guarantees the existence of the adjoint state and gives the expression of the shape derivative for the drag functional in terms of the vector field  $\mathbf{T}$ .

**Theorem 3.6.** *Let a given solution  $\vartheta \in \mathcal{B}_\tau$ ,  $(\zeta, m) \in X^{s,r} \times \mathbb{R}$ , to problem (26) meets all requirements of Theorem 3.1. Then there exists positive constant  $\tau_1$  (depending only on  $\mathbf{U}$ ,  $\Omega$  and  $r, s$ ) such that, if  $\tau \in (0, \tau_1]$  and  $R\lambda^{-1} \leq \tau_1^2$ , then there exists a unique solution  $\mathbf{Y} \in (Y^{s,r})^3 \times (X^{s,r})^3 \times \mathbb{R}$  to problem (43), (40). The form  $L_u$  has the representation*

$$L_u(\mathbf{w}, \psi, \omega) = \int_{\Omega} [\operatorname{div} \mathbf{T}(b_{10}^0 \zeta + b_{20}^0 g + \sigma v + \varkappa b_{30}^0 l) + \mathcal{D}_0(\operatorname{div} \mathbf{T} - \mathbf{T}') \mathbf{h}] dx \quad (44)$$

where the coefficients  $b_{ij}^0$  and the operator  $\mathcal{D}_0$  are defined by the formulae (34), (36).

## References

- [1] J.A. Bello, E. Fernandez-Cara, J. Lemoine, J. Simon *The differentiability of the drag with respect to variations of a Lipschitz domain in a Navier-Stokes flow*, SIAM J. Control. Optim. 35, No. 2, (1997), 626–640.
- [2] E. Feireisl *Dynamics of Viscous Compressible Fluids* (Oxford University Press, Oxford 2004).
- [3] E. Feireisl, A.H. Novotný, H. Petzeltová *On the domain dependence of solutions to the compressible Navier-Stokes equations of a barotropic fluid* Math. Methods Appl. Sci. 25 (2002), no. 12, 1045–1073.
- [4] E. Feireisl *Shape optimization in viscous compressible fluids* Appl. Math. Optim. 47(2003), 59–78.
- [5] J.G. Heywood, M. Padula *On the uniqueness and existence theory for steady compressible viscous flow* in Fundamental directions in mathematical fluids mechanics, 171–189, Adv. Math. Fluids Mech., Birkhäuser, Basel (2000).
- [6] Jae Ryong Kweon, R.B. Kellogg *Compressible Navier-Stokes equations in a bounded domain with inflow boundary condition* SIAM J. Math. Anal., 28, N1, 35 no. 1, 94–108 (1997).
- [7] Jae Ryong Kweon, R.B. Kellogg *Regularity of solutions to the Navier-Stokes equations for compressible barotropic flows on a polygon*. Arch. Ration. Mech. Anal., 163, N1, 36–64 (2000).
- [8] P.L. Lions *Mathematical topics in fluid dynamics, Vol. 2, Compressible models* (Oxford Science Publication, Oxford 1998).
- [9] B. Mohammadi, O. Pironneau *Shape optimization in fluid mechanics* Ann. Rev. Fluid Mech., 36, 255–279, (2004) Ann. Reviews, Palo Alto, CA.
- [10] Marwan Moubarac, J.-P. Zolésio *Moving Shape Analysis And Control: Applications to Fluid Structure Interactions* Chapman & Hall/CRC, Boca Raton 2006.
- [11] A. Novotný, M. Padula *Existence and Uniqueness of Stationary solutions for viscous compressible heat conductive fluid with large potential and small non-potential external forces* Siberian Math. Journal, 34, 1993, 120–146.
- [12] A. Novotný, I. Straškraba *Introduction to the mathematical theory of compressible flow* Oxford Lecture Series in Mathematics and its Applications, Vol. 27. Oxford University Press, Oxford, 2004.
- [13] O.A. Oleinik, E.V. Radkevich *Second order equation with non-negative characteristic form* American Math. Soc., Providence, Rhode Island Plenum Press. New York-London (1973).

- [14] P.I. Plotnikov, J. Sokolowski *On compactness, domain dependence and existence of steady state solutions to compressible isothermal Navier-Stokes equations* J. Math. Fluid Mech. 7(2005), no. 4, 529–573.
- [15] P.I. Plotnikov, J. Sokolowski *Concentrations of solutions to time-discretized compressible Navier-Stokes equations* Communications in Mathematical Physics Volume 258, Number 3, 2005, 567–608.
- [16] P.I. Plotnikov, J. Sokolowski *Stationary Boundary Value Problems for Navier-Stokes Equations with Adiabatic Index  $\nu < 3/2$* , Doklady Mathematics Vol. 70, No. 1, 2004, 535–538. Translated from Doklady Akademii Nauk, Volume 397, Nos. 1–6, 2004.
- [17] P.I. Plotnikov, J. Sokolowski *Domain dependence of solutions to compressible Navier-Stokes equations* SIAM J. Control Optim., Volume 45, Issue 4, 2006, pp. 1147–1539.
- [18] P.I. Plotnikov, J. Sokolowski *Stationary solutions for Navier-Stokes equations for diatomic gases*. Russian Mathematical Surveys, Russian Math. Surveys 62:3 561593, (2007) RAS, Uspekhi Mat. Nauk 62:3 117148.
- [19] P.I. Plotnikov, E.V. Ruban, J. Sokolowski *Inhomogeneous boundary value problems for compressible Navier-Stokes equations: well-posedness and sensitivity analysis*. SIAM J. Math Analysis, **40**:3, (2008), 1152–1200.
- [20] P.I. Plotnikov, E.V. Ruban, J. Sokolowski *Inhomogeneous boundary value problems for compressible Navier-Stokes and transport equations*, Journal des Mathématiques Pure et Appliquées, electronic (2009).
- [21] P.I. Plotnikov, J. Sokolowski *Stationary Boundary Value Problems for Compressible Navier-Stokes Equations*, Handbook of Differential Equations, Volume 6, Elsevier, Edited by M. Chipot, 2008, 313–410.
- [22] H. Schlichting *Boundary-layer theory*, (McGraw-Hill series in mechanical engineering) New York: McGraw-Hill, 1955, 535 p.
- [23] J. Simon *Domain variation for drag in Stokes flow*, in Control and Estimation of Distributed Parameter Systems, Internat. N Ser. Numer. Math. 91, F. Kappel, K. Kuninisch, and W. Schappacher, eds., Birkhäuser, Basel, 1989, 361–378.
- [24] T. Slawig *A formula for the derivative with respect to domain variations in Navier-Stokes flow based on an embedding domain method* SIAM J. Control Optim., 42, No. 2, 495–512 (2003).
- [25] J. Sokolowski, J.-P. Zolésio *Introduction to Shape Optimization. Shape Sensitivity Analysis*. Springer Series in Computational Mathematics Vol. 16, Springer Verlag, (1992).

P.I. Plotnikov and E.V. Ruban  
 Lavrentyev Institute of Hydrodynamics,  
 Siberian Division of Russian Academy of Sciences,  
 Lavrentyev pr. 15,  
 Novosibirsk 630090, Russia  
 e-mail: [plotnikov@hydro.nsc.ru](mailto:plotnikov@hydro.nsc.ru)

J. Sokolowski  
 Institut Elie Cartan, Laboratoire de Mathématiques  
 Université Henri Poincaré Nancy 1, B.P. 239,  
 F-54506 Vandoeuvre lès Nancy Cedex, France  
 e-mail: [Jan.Sokolowski@iecn.u-nancy.fr](mailto:Jan.Sokolowski@iecn.u-nancy.fr)

# Null-controllability for a Coupled Heat-Finite-dimensional Beam System

Jean-Pierre Raymond and Muthusamy Vanninathan

**Abstract.** A model representing a coupling between a heat conducting medium and a finite-dimensional approximation of a beam equation is considered. We establish a Carleman inequality for this model. Next we deduce a null-controllability result with an internal control in the conducting medium and there is no control in the structure equation.

**Mathematics Subject Classification (2000).** 93C20, 93B05.

**Keywords.** Controllability, coupled systems, heat-structure model.

## 1. Introduction

Let  $\Omega$  be a two-dimensional rectangular domain defined by  $\Omega = ]0, L[\times]0, 1[$ , with boundary  $\partial\Omega = \Gamma_e \cup \Gamma_b \cup \Gamma_p$ , where the different parts are given by  $\Gamma_e = ]0, L[\times\{0\}$ ,  $\Gamma_b = ]0, L[\times\{1\}$  and  $\Gamma_p = (\{0\} \cup \{L\}) \times ]0, 1[$ . We shall use the notation  $x = (x_1, x_2)$  for  $x \in \overline{\Omega}$ . We also introduce  $Q = \Omega \times ]0, T[$ ,  $\Gamma = \Gamma_e \cup \Gamma_b$ ,  $\Sigma = \Gamma \times ]0, T[$ ,  $\Sigma_e = \Gamma_e \times ]0, T[$ ,  $\Sigma_b = \Gamma_b \times ]0, T[$ , and  $\Sigma_p = \Gamma_p \times ]0, T[$ . Let us consider a system coupling the heat equation in the domain  $\Omega$  with a beam equation at its boundary  $\Gamma_b$ :

$$\begin{aligned} \phi' - \Delta\phi &= f \quad \text{in } Q, \quad \phi = 0 \quad \text{on } \Sigma_e, \quad \phi = z' \quad \text{on } \Sigma_b, \\ \phi(0, x_2, t) &= \phi(L, x_2, t) \\ \text{and } \partial_n\phi(0, x_2, t) &= -\partial_n\phi(L, x_2, t) \quad \text{for } (x_2, t) \in ]0, 1[\times]0, T[, \\ \phi(0) &= \phi^0 \quad \text{in } \Omega, \\ z'' - \beta z_{x_1 x_1} + \alpha z_{x_1 x_1 x_1 x_1} &= -\partial_n\phi \quad \text{on } \Sigma_b, \\ z(0, t) &= z(L, t), \quad z_{x_1}(0, t) = z_{x_1}(L, t), \quad \text{for } t \in ]0, T[, \\ z_{x_1 x_1}(0, t) &= z_{x_1 x_1}(L, t), \quad z_{x_1 x_1 x_1}(0, t) = z_{x_1 x_1 x_1}(L, t) \quad \text{for } t \in ]0, T[, \\ z(x_1, 0) &= z^0(x_1) \quad \text{and} \quad z'(x_1, 0) = z^1(x_1) \quad \text{in } ]0, L[, \end{aligned} \tag{1.1}$$

where  $z_{x_1}$ ,  $z_{x_1x_1}$ ,  $z_{x_1x_1x_1}$  and  $z_{x_1x_1x_1x_1}$  denote the different partial derivatives of  $z$  with respect to  $x_1$ , while  $z'$  and  $z''$  denote the different partial derivatives of  $z$  with respect to  $t$ . The symbol  $\partial_n$  denotes the normal derivative at the boundary. Let us notice that we have taken periodic boundary conditions on  $\Gamma_p$  both for  $\phi$  and  $z$ . The coefficient  $\alpha > 0$  and  $\beta \geq 0$  are the adimensional rigidity and stretching coefficients of the beam. For simplicity we shall set  $\alpha = \beta = 1$ . We have here considered the heat operator as a simplified model of the Stokes operator, this later providing a more realistic model.

To the authors knowledge, the null-controllability of system (1.1) with a localized control acting in the heat equation and another one in the beam equation is an open problem. Here we are interested in a null-controllability result for a system in which the beam equation is replaced by a finite-dimensional approximation. For that let us consider, for a fixed  $N$ , a family of smooth functions  $(\zeta_1, \dots, \zeta_N)$ , orthonormal in  $L^2(\Gamma_b)$ , with the periodicity conditions stated above. For example, we can consider the first  $N$ -functions of an orthonormal basis in  $L^2(\Gamma_b)$  constituted of eigenfunctions  $\zeta \in H^4(\Gamma_b)$  satisfying

$$\begin{aligned} -\zeta_{x_1x_1} + \zeta_{x_1x_1x_1x_1} &= \lambda\zeta \quad \text{on } \Gamma_b, \\ \zeta(0) &= \zeta(L), \quad \zeta_{x_1}(0) = \zeta_{x_1}(L), \\ \zeta_{x_1x_1}(0) &= \zeta_{x_1x_1}(L), \quad \zeta_{x_1x_1x_1}(0) = \zeta_{x_1x_1x_1}(L). \end{aligned}$$

By setting  $z(x_1, t) = \sum_{i=1}^N r_i(t)\zeta_i(x_1)$ , and following a Galerkin approximation procedure, the system is reduced to

$$\begin{aligned} \phi' - \Delta\phi &= f \quad \text{in } Q, \quad \phi = 0 \quad \text{on } \Sigma_e, \quad \phi = r' \cdot \zeta \quad \text{on } \Sigma_b, \\ \phi(0, x_2, t) &= \phi(L, x_2, t) \\ \text{and } \partial_n\phi(0, x_2, t) &= -\partial_n\phi(L, x_2, t) \quad \text{for } (x_2, t) \in ]0, 1[ \times ]0, T[, \\ \phi(0) &= \phi^0 \quad \text{in } \Omega, \\ r'' + Ar &= - \int_{\Gamma_b} \partial_n\phi \zeta \quad \text{in } ]0, T[, \\ r(0) = r^0 \quad \text{and} \quad r'(0) = r^1 &= r^1 \quad \text{in } \mathbb{R}^N, \end{aligned} \tag{1.2}$$

where  $\zeta = (\zeta_1, \dots, \zeta_N)$  and  $r = (r_1, \dots, r_N)$ . In this model

$$A = \left( \int_{\Gamma_b} (\zeta_{i,x_1}\zeta_{j,x_1} + \zeta_{i,x_1x_1}\zeta_{j,x_1x_1}) dx_1 \right)_{1 \leq i, j \leq N} \in \mathbb{R}^{N \times N}$$

is a symmetric positive definite matrix. Denoting by  $\ell_m$  and  $\ell_M$  the lowest and the highest eigenvalues of  $A$ , we have  $\ell_m I \leq A \leq \ell_M I$ .

The main result of the paper is the following theorem which is a null-controllability result for (1.2) with an internal control in the heat conducting medium  $\Omega$ . We do not require any control on the beam part of the model.

**Theorem 1.** *Let  $\omega$  be an arbitrary nonempty open subset, relatively compact in  $\Omega$ . For all  $\phi^0 \in L^2(\Omega)$ ,  $r^0 \in \mathbb{R}^N$  and  $r^1 \in \mathbb{R}^N$  there exists a function  $u \in L^2(Q)$  such*

that the solution of (1.2) with  $f = u\chi_{\omega \times (0,T)}$  obeys

$$\phi(T) = 0, \quad r(T) = 0, \quad r'(T) = 0.$$

( $\chi_{\omega \times (0,T)}$  is the characteristic function of  $\omega \times (0,T)$ .)

We have already studied in [10] a model similar to (1.2) in which  $\Omega$  is a smooth domain in  $\mathbb{R}^2$  (not necessarily a rectangle), and in which the structure equation is replaced by an oscillator equation

$$r'' + r = - \int_{\Gamma_i} \partial_n \phi n \quad \text{in } ]0, T[, \quad (1.3)$$

modeling the vibration of a tube of boundary  $\Gamma_i$ , surrounded by  $\Omega$ ,  $n$  being the unit normal to  $\Gamma_i$  outward  $\Omega$ . Thus in that model  $r(t)$  belongs to  $\mathbb{R}^2$ . In [10]  $\phi$  satisfies a Dirichlet homogeneous boundary condition on  $\partial\Omega \setminus \Gamma_i$ .

The proof of the null-controllability stated in [10] is based on Carleman estimates. The technique used in [10] to establish Carleman estimates is very similar to the one used in [7]. But as explained in [10] the main difference between the Carleman estimate proved in [7] and the one we obtain in [10] is that the new term  $\int_0^T \rho_{\Gamma_i}^{-2s} |r|^2$  appears in the RHS of Carleman estimates ( $\rho_{\Gamma_i}^{-2s}$  is a weight function only depending on  $t$ ). This is due to the term  $r$  in equation (1.3). For system (1.2) the term  $\int_0^T \rho_{\Gamma_b}^{-2s} |r|^2$  will appear in the Carleman inequality. It is due to the presence  $Ar$  in the structure equation, taking into account the deformation of the structure. Such terms are not present in models coupling a fluid equation with the motion of a rigid body as in [1] or [7].

In order to prove Theorem 1, this new term has to be estimated by the LHS of Carleman estimates. This is done in [10] by proving an appropriate weighted energy estimate for the structure equation. We shall see in a forthcoming work that such an estimate is not sufficient to deal with more complicated models like Stokes equations coupled with a finite-dimensional approximation of a structure equation [11]. For such models a more general compactness argument has to be used. The main objective of the present paper is to introduce this compactness argument in Carleman estimates established for system (1.2). In order to present this compactness argument we first have to prove a preliminary Carleman inequality. Since the calculations are very similar to the ones in [10], we do not repeat them. We recall without proof some inequalities obtained in [10], and we give details only when the calculations are different.

Let us recall that null-controllability of systems coupling a fluid equation with the motion of a rigid body is studied in [1, 7]. In [7] the solid is a disk, whereas in [1] some symmetry assumption is assumed on the solid [1, Assumption (1.9)]. The derivation of Carleman inequalities for parabolic problems can be found in [5] and [6]. Because of the coupling with a hyperbolic part in the present model, proving Carleman type inequality poses a technical challenge which we have overcome in [10].

The plan of the article is as follows: As is known, the proof of Carleman inequality involves a transformation of system (1.2) via a change of variables.

The transformed operator is the heat operator conjugated by exponential. Here due to the boundary conditions coupling the heat equation with the structure equation we have to make a particular choice for the function appearing in the exponential. We introduce the test function  $\eta$  appearing in the exponential and the transformed system in Section 2. As a starting point of our calculations we recall a first inequality obtained in [10] and we explain how to treat the boundary terms for our model in Section 3. This treatment of boundary terms is somewhat different from the one in [10]. New estimates for  $r$  are derived in Section 4. These new estimates are next used in Section 5 to obtain an improved Carleman inequality thanks to a compactness argument. The corresponding Carleman inequality for the original system (1.2) is given in Section 6. The proof of Theorem 1 is provided in Section 7.

Throughout the paper, we use the usual summation convention with respect to repeated indices. Various constants independent of parameters  $(s, \lambda)$  and the solution are generically denoted by  $C$ , unless stated otherwise.

## 2. Preliminary result and transformed system

Let  $V$  be the space defined by

$$V = \left\{ \phi \in H^1(\Omega) \mid \phi = 0 \text{ on } \Gamma_e, \quad \phi(0, x_2) = \phi(L, x_2) \text{ for } x_2 \in ]0, 1[ \right\},$$

and denote by  $V'$  the topological dual of  $V$ . The space  $V$  will be equipped with the norm

$$\phi \mapsto \left( \int_{\Omega} |\nabla \phi|^2 dx \right)^{1/2},$$

denoted by  $\|\cdot\|_V$  (the same kind of notation will be used for other Banach spaces). Let us remark that this norm is equivalent to the usual  $H^1(\Omega)$ -norm. The norm in  $\mathbb{R}^N$  will be simply denoted by  $|\cdot|$ . The inner product of  $s \in \mathbb{R}^N$  and  $\sigma \in \mathbb{R}^N$  is denoted by  $s \cdot \sigma$ .

Well-posedness of system (1.2) is straightforward and it can be established using energy estimates, for instance. Indeed, multiplying (1.2) by  $(\phi, r')$ , we get the energy identity:

$$\begin{aligned} & \|\phi(t)\|_{L^2(\Omega)}^2 + |A^{1/2}r(t)|^2 + |r'(t)|^2 + 2 \int_0^t \int_{\Omega} |\nabla \phi|^2 \\ &= 2 \int_0^t \int_{\Omega} f\phi + \|\phi(0)\|_{L^2(\Omega)}^2 + |A^{1/2}r^0|^2 + |r^1|^2. \end{aligned}$$

Using this, we can prove the following result.

**Theorem 2.** *Let  $f \in L^2(0, T; L^2(\Omega))$ ,  $\phi^0 \in L^2(\Omega)$ ,  $r^0 \in \mathbb{R}^N$  and  $r^1 \in \mathbb{R}^N$ . Then there is a unique solution  $(\phi, r) \in C([0, T]; L^2(\Omega)) \cap L^2(0, T; V) \times C^1([0, T]; \mathbb{R}^N)$*

to the system (1.2) satisfying the energy inequality

$$\begin{aligned} & \|\phi\|_{C([0,T];L^2(\Omega))} + \|\phi\|_{L^2(0,T;V)} + \|\phi'\|_{L^2(0,T;H^{-1}(\Omega))} + \|r\|_{C^1([0,T];\mathbb{R}^N)} \\ & \leq C \left\{ \|f\|_{L^2(0,T;L^2(\Omega))} + \|\phi^0\|_{L^2(\Omega)} + |r^0| + |r^1| \right\}. \end{aligned}$$

Carleman inequality for the system (1.2) is stated in section 6. To prove this inequality, we have to transform the system (1.2) to a new system via a change of variables. We begin by listing the properties of the test function  $\eta$  which is used in defining the change variables. These properties are used at various stages of our computations below.

**Lemma 3.** Suppose that  $\omega_0 \subset\subset \omega \subset\subset \Omega$ . Then there exist a function  $\eta \in C^4(\overline{\Omega})$  and positive constants  $C_{\Gamma_e}$  and  $C_{\Gamma_b}$  such that

- $\eta(x) > 0$  for all  $x \in \overline{\Omega}$ ,
- $\eta(x) = C_{\Gamma_e}$  and  $\partial_n \eta \leq 0$  for all  $x \in \Gamma_e$ ,
- $\eta(x) = C_{\Gamma_b}$ ,  $\partial_n \eta = -1$ , and  $\Delta \eta(x) = 0$ , for all  $x \in \Gamma_b$ ,
- $\eta$  together with its partial derivatives with respect to  $x_1$ , up to fourth order, satisfy periodic boundary conditions on  $\Gamma_p$ ,
- $|\nabla \eta(x)| > 0$  for all  $x \in \overline{\Omega} \setminus \omega_0$ .

*Proof.* The proof of [10, Lemma 3.1] can be adapted to the present geometrical setting.  $\square$

With a large parameter  $\lambda \geq 1$ , we introduce the function

$$\alpha(x) = e^{\lambda K_1} - e^{\lambda \eta(x)} \quad \forall x \in \overline{\Omega}, \quad (2.1)$$

where  $K_1 > 0$  is a constant, with  $K_1 > \max_{x \in \overline{\Omega}} |\eta(x)|$  and  $\eta$  is the function obeying the conditions in Lemma 3. We set

$$\beta(x, t) = \frac{\alpha(x)}{t^k(T-t)^k}, \quad \rho(x, t) = e^{\beta(x, t)},$$

where the constant  $k$  is chosen such that  $k \geq 2$ . Since  $\eta$  is constant on  $\Gamma_e$  and on  $\Gamma_b$ , the functions  $\beta(\cdot, t)$  and  $\rho(\cdot, t)$  are also constants there. In the following, we set

$$\rho_{\Gamma_b}(t) = \rho(\cdot, t)|_{\Gamma_b}.$$

With another large parameter  $s \geq 1$ , we also define the functions

$$f_s(x, t) = \rho^{-s}(x, t)f(x, t) \quad \text{and} \quad \psi(x, t) = \rho^{-s}(x, t)\phi(x, t). \quad (2.2)$$

Notice that (since  $\beta \rightarrow \infty$  as  $t \rightarrow 0^+$  or as  $t \rightarrow T^-$ )  $\psi(\cdot, 0) = \psi(\cdot, T) = 0$  in  $\Omega$ . An easy calculation shows that

$$\begin{aligned} \nabla \phi &= \nabla(e^{s\beta}\psi) = e^{s\beta}(\nabla \psi + s\psi \nabla \beta), \\ \partial_n \phi &= e^{s\beta}(\partial_n \psi + s\psi \partial_n \beta) = \rho_{\Gamma_b}^s \partial_n \psi + s r' \cdot \zeta \partial_n \beta \quad \text{on } \Sigma_b. \end{aligned}$$

Thus the coupled system (1.2) can be rewritten in terms of  $(\psi, r)$  as follows:

$$\begin{aligned} M_1\psi + M_2\psi &= g_s = f_s + s(\Delta\beta)\psi, \quad \text{in } Q, \\ \psi &= 0 \quad \text{on } \Sigma_e, \quad \psi = \rho_{\Gamma_b}^{-s} r' \cdot \zeta \quad \text{on } \Sigma_b, \\ \psi(0, x_2, t) &= \psi(L, x_2, t), \\ \text{and } \partial_n\psi(0, x_2, t) &= -\partial_n\psi(L, x_2, t) \quad \text{for } (x_2, t) \in ]0, 1[ \times ]0, T[, \\ \psi(0) &= \psi(T) = 0 \quad \text{in } \Omega, \\ r'' + Ar &= -\left(\rho_{\Gamma_b}^s \int_{\Gamma_b} \partial_n\psi \zeta + s \int_{\Gamma_b} r' \cdot \zeta \partial_n\beta \zeta\right) \quad \text{in } (0, T), \\ r(0) &= r^0 \text{ and } r'(0) = r^1, \end{aligned} \tag{2.3}$$

where

$$M_1\psi = \psi' - 2s\nabla\beta \cdot \nabla\psi \quad \text{and} \quad M_2\psi = s\beta'\psi - \Delta\psi - s^2|\nabla\beta|^2\psi. \tag{2.4}$$

### 3. Carleman inequality (I)

In this section, we recall the first version of the Carleman inequality for the transformed system (2.3) obtained in [10]. We next give estimates for the boundary terms appearing in the first Carleman estimate because their treatment is different from the calculations made in [10]. Writing the transformed equation satisfied by  $\psi$  in the form  $M_1\psi + M_2\psi = g_s$  is a crucial aspect of the proof.

From the first equation of the system (2.3) it follows that

$$\|M_1\psi\|_{L^2(Q)}^2 + \|M_2\psi\|_{L^2(Q)}^2 + 2(M_1\psi, M_2\psi)_{L^2(Q)} = \|g_s\|_{L^2(Q)}^2. \tag{3.1}$$

As in [10], the cross term  $2(M_1\psi, M_2\psi)_{L^2(Q)}$  can be rewritten as follows

$$2(M_1\psi, M_2\psi)_{L^2(Q)} = I_1 + I_2 + I_3,$$

where

$$\begin{aligned} I_1 &= 2 \int_Q (s\beta'\psi - \Delta\psi - s^2|\nabla\beta|^2\psi) \psi', \quad I_2 = 4s \int_Q (\nabla\beta \cdot \nabla\psi) \Delta\psi, \\ I_3 &= 4s \int_Q (s^2|\nabla\beta|^2\psi - s\beta'\psi) (\nabla\beta \cdot \nabla\psi). \end{aligned} \tag{3.2}$$

With calculations similar to those in [10], we can prove the following estimate

$$\begin{aligned} &\|M_1\psi\|_{L^2(Q)}^2 + \|M_2\psi\|_{L^2(Q)}^2 + s^3\lambda^4 \int_Q \frac{e^{3\lambda\eta}}{t^{3k}(T-t)^{3k}} |\psi|^2 + J_1 + J_2 + J_3 + J_4 \\ &\leq C \left\{ \|f_s\|_{L^2(Q)}^2 + s^3\lambda^4 \int_{\omega \times (0, T)} \frac{e^{3\lambda\eta}}{t^{3k}(T-t)^{3k}} |\psi|^2 \right. \\ &\quad \left. + s\lambda \int_0^T \frac{e^{\lambda\eta}|_{\Gamma_b}}{t^k(T-t)^k} \rho_{\Gamma_b}^{-s} r' \cdot \int_{\Gamma_b} \partial_n \psi \zeta \right\}, \end{aligned}$$

where

$$\begin{aligned}
J_1 &= s\lambda \int_{\Sigma_b} \frac{e^{\lambda\eta}}{t^k(T-t)^k} |\partial_n \psi|^2, \quad J_2 = s \int_0^T \rho_{\Gamma_b}^{-s} r' \Delta \beta|_{\Gamma_b} \cdot \int_{\Gamma_b} \partial_n \psi \zeta, \\
J_3 &= -s \int_{\Sigma_b} \partial_n(\Delta \beta) |\psi|^2, \\
J_4 &= 2 \int_0^T \left( \left( s r' \beta' + s \int_{\Gamma_b} (r' \cdot \zeta) \partial_n \beta \zeta + Ar \right) \rho_{\Gamma_b}^{-s} \cdot \int_{\Gamma_b} \partial_n \psi \zeta \right) \\
&\quad + 2 \int_0^T \left| \int_{\Gamma_b} \partial_n \psi \zeta \right|^2 + 2s^3 \int_{\Sigma_b} (\partial_n \beta)^3 |\psi|^2 - 2s^2 \int_{\Sigma_b} \beta' \partial_n \beta |\psi|^2.
\end{aligned}$$

**Treatment of boundary terms in  $J_4$ .** The effect of the fluid-solid interaction in our model is felt in the treatment of boundary terms which are different from the ones in other classical models. We will estimate these boundary terms below. Let us begin by naming the different terms in  $J_4$  as follows:

$$\begin{aligned}
T_1 &= 2 \int_0^T \left| \int_{\Gamma_b} \partial_n \psi \zeta \right|^2, \\
T_2 &= 2s^3 \int_{\Sigma_b} (\partial_n \beta)^3 |\psi|^2, \\
T_3 &= 2 \int_0^T \left( \left( s r' \beta' + s \int_{\Gamma_b} r' \cdot \zeta \partial_n \beta \zeta + Ar \right) \rho_{\Gamma_b}^{-s} \cdot \int_{\Gamma_b} \partial_n \psi \zeta \right), \\
T_4 &= -2s^2 \int_{\Sigma_b} \beta' \partial_n \beta |\psi|^2 = 2s^2 \lambda k \int_{\Sigma_b} \frac{(T-2t)}{t^{2k+1}(T-t)^{2k+1}} (e^{\lambda K_1} - e^{\lambda\eta}) e^{\lambda\eta} |\psi|^2.
\end{aligned}$$

First let us consider  $T_2$  which can be expressed as (since  $\psi = \rho_{\Gamma_b}^{-s} r' \cdot \zeta$  on  $\Sigma_b$ )

$$T_2 = 2s^3 \lambda^3 \int_0^T \frac{e^{3\lambda\eta}|_{\Gamma_b}}{t^{3k}(T-t)^{3k}} \rho_{\Gamma_b}^{-2s} \int_{\Gamma_b} |r' \cdot \zeta|^2.$$

Since  $(\zeta_1, \dots, \zeta_N)$  is a family of orthonormal functions in  $L^2(\Gamma_b)$ , we have

$$\int_{\Gamma_b} |r' \cdot \zeta|^2 dx = \int_{\Gamma_b} \left| \sum_{i=1}^N r'_i(t) \zeta_i \right|^2 dx = \sum_{i=1}^N |r'_i(t)|^2 = |r'(t)|^2.$$

This allows us to write:

$$T_2 = 2s^3 \lambda^3 \int_0^T \frac{e^{3\lambda\eta}|_{\Gamma_b}}{t^{3k}(T-t)^{3k}} \rho_{\Gamma_b}^{-2s} |r'|^2.$$

Next, we can estimate  $T_3$  in the following way:

$$\begin{aligned} |T_3| &\leq \frac{1}{4} \int_0^T \left| \int_{\Gamma_b} \partial_n \psi \zeta \right|^2 + 4 \int_0^T \left| s r' \beta' + s \int_{\Gamma_b} r' \cdot \zeta \partial_n \beta \zeta + Ar \right|^2 \rho_{\Gamma_b}^{-2s} \\ &\leq \frac{1}{8} T_1 + 12s^2 e^{2\lambda K_1} T^2 \int_0^T \frac{k^2}{t^{2k+2}(T-t)^{2k+2}} \rho_{\Gamma_b}^{-2s} |r'|^2 \\ &\quad + 12s^2 \lambda^2 \int_0^T \frac{e^{2\lambda\eta} |\Gamma_b|}{t^{2k}(T-t)^{2k}} \rho_{\Gamma_b}^{-2s} |\Gamma_b|^2 |r'|^2 + 12\ell_M \int_0^T \rho_{\Gamma_b}^{-2s} |r|^2. \end{aligned}$$

By choosing  $s$  large enough (depending on  $\lambda$ ) and choosing  $k \geq 2$ ,

$$|T_3| \leq \frac{1}{8} T_1 + \frac{1}{8} T_2 + 12\ell_M \int_0^T \rho_{\Gamma_b}^{-2s} |r|^2.$$

Next, we can estimate  $J_2$  as follows :

$$\begin{aligned} |J_2| &= s \left| \int_0^T \rho_{\Gamma_b}^{-s} r' \Delta \beta|_{\Gamma_b} \cdot \int_{\Gamma_b} \partial_n \psi \zeta \right| \\ &\leq \frac{1}{4} \int_0^T \left| \int_{\Gamma_b} \partial_n \psi \zeta \right|^2 + s^2 \int_0^T \rho_{\Gamma_b}^{-2s} |\Delta \beta|_{\Gamma_b}|^2 |r'|^2 \\ &\leq \frac{1}{8} T_1 + C s^2 \lambda^4 \int_0^T \frac{e^{2\lambda\eta} |\Gamma_b|}{t^{2k}(T-t)^{2k}} \rho_{\Gamma_b}^{-2s} |r'|^2. \end{aligned}$$

Once again we see that for large  $s$  (depending on  $\lambda$ ) we have  $|J_2| \leq \frac{1}{8} T_1 + \frac{1}{8} T_2$ . To estimate  $J_3$ , we express it as

$$J_3 = -s \int_{\Sigma_b} \partial_n (\Delta \beta) \rho_{\Gamma_b}^{-2s} |r' \cdot \zeta|^2$$

in which we use the estimate (for  $\lambda$  large)

$$|\partial_n (\Delta \beta)| \leq C \lambda^3 \frac{e^{\lambda\eta} |\Gamma_b|}{t^k (T-t)^k} \text{ on } \Sigma_b.$$

This easily leads to  $|J_3| \leq \frac{1}{8} T_2$  for  $s$  large (depending on  $\lambda$ ). Analogous arguments establish that

$$|T_4| \leq C s^2 \lambda e^{\lambda K_1} \int_0^T \frac{e^{\lambda\eta} |\Gamma_b|}{t^{2k+1} (T-t)^{2k+1}} \rho_{\Gamma_b}^{-2s} \int_{\Gamma_b} |r' \cdot \zeta|^2 \leq \frac{1}{8} T_2$$

for  $s$  large (depending on  $\lambda$ ). Assembling these estimates together, we obtain

$$|T_3| + |T_4| + |J_2| + |J_3| \leq \frac{1}{4} T_1 + \frac{1}{2} T_2 + C \int_0^T \rho_{\Gamma_b}^{-2s} |r|^2.$$

Hence

$$J_2 + J_3 + J_4 \geq \frac{3}{4} T_1 + \frac{1}{2} T_2 - C \int_0^T \rho_{\Gamma_b}^{-2s} |r|^2.$$

Our next task is to estimate  $T_1$  from below. To this end, we use (2.3) and write

$$\int_{\Gamma_b} \partial_n \psi \zeta = -\rho_{\Gamma_b}^{-s} (r'' + Ar) - s \rho_{\Gamma_b}^{-s} \int_{\Gamma_b} (r' \cdot n) \partial_n \beta \zeta.$$

Therefore

$$\begin{aligned} \left| \int_{\Gamma_b} \partial_n \psi \zeta \right|^2 &= \rho_{\Gamma_b}^{-2s} \left| r'' + Ar + s \lambda \frac{e^{\lambda \eta}|_{\Gamma_b}}{t^k(T-t)^k} \int_{\Gamma_b} (r' \cdot \zeta) \zeta \right|^2 \\ &\geq \frac{1}{2} \rho_{\Gamma_b}^{-2s} |r''|^2 - 2\ell_M \rho_{\Gamma_b}^{-2s} |r|^2 - 2s^2 \lambda^2 \rho_{\Gamma_b}^{-2s} \frac{e^{2\lambda \eta}|_{\Gamma_b}}{t^{2k}(T-t)^{2k}} |\Gamma_b|^2 |r'|^2, \end{aligned}$$

using the elementary inequality  $|a+b|^2 \geq \frac{1}{2}|a|^2 - |b|^2$ . It follows then, for  $s, \lambda$  large, that

$$\frac{T_1}{2} \geq \frac{1}{2} \int_0^T \rho_{\Gamma_b}^{-2s} (|r''|^2 + |r|^2) - \frac{1}{4} T_2 - C \int_0^T \rho_{\Gamma_b}^{-2s} |r|^2.$$

As a consequence, we have

$$\frac{3}{4} T_1 + \frac{1}{2} T_2 \geq \frac{1}{4} T_1 + \frac{1}{4} T_2 + \frac{1}{2} \int_0^T \rho_{\Gamma_b}^{-2s} (|r''|^2 + |r|^2) - C \int_0^T \rho_{\Gamma_b}^{-2s} |r|^2.$$

Thus the final estimate of the boundary terms is as follows:

$$\begin{aligned} J_2 + J_3 + J_4 &\geq \frac{1}{2} \int_0^T \left| \int_{\Gamma_b} \partial_n \psi \zeta \right|^2 + \frac{1}{2} s^3 \lambda^3 \int_0^T \frac{e^{3\lambda \eta}|_{\Gamma_b}}{t^{3k}(T-t)^{3k}} \rho_{\Gamma_b}^{-2s} |r'|^2 \\ &\quad + \frac{1}{2} \int_0^T \rho_{\Gamma_b}^{-2s} (|r''|^2 + |r|^2) - C \int_0^T \rho_{\Gamma_b}^{-2s} |r|^2. \end{aligned}$$

**Carleman inequality (I).** Grouping together various estimates obtained, we can summarize the main inequality of Section 3

$$\begin{aligned} &\|M_1 \psi\|_{L^2(Q)}^2 + \|M_2 \psi\|_{L^2(Q)}^2 + s^3 \lambda^4 \int_Q \frac{e^{3\lambda \eta}}{t^{3k}(T-t)^{3k}} |\psi|^2 \\ &\quad + \int_0^T \left| \int_{\Gamma_b} \partial_n \psi \zeta \right|^2 + s \lambda \int_{\Sigma_b} \frac{e^{\lambda \eta}}{t^k(T-t)^k} |\partial_n \psi|^2) \\ &\quad + \int_0^T \rho_{\Gamma_b}^{-2s} (|r''|^2 + |r|^2) + s^3 \lambda^3 \int_0^T \frac{e^{3\lambda \eta}|_{\Gamma_b}}{t^{3k}(T-t)^{3k}} \rho_{\Gamma_b}^{-2s} |r'|^2 \\ &\leq C \left\{ \|f_s\|_{L^2(Q)}^2 + s^3 \lambda^4 \int_{\omega \times (0,T)} \frac{e^{3\lambda \eta}}{t^{3k}(T-t)^{3k}} |\psi|^2 + \int_0^T \rho_{\Gamma_b}^{-2s} |r|^2 \right\}. \end{aligned} \tag{3.3}$$

Let us notice that  $s \lambda \int_{\Sigma_b} \frac{e^{\lambda \eta}}{t^k(T-t)^k} |\partial_n \psi|^2$  dominates the term  $\int_0^T \left| \int_{\Gamma_b} \partial_n \psi \zeta \right|^2$  which can be dropped out.

#### 4. Estimate of $r$

Our goal in the next two sections is to strengthen the above inequality (3.3) by removing the term  $\int_0^T \rho_{\Gamma_b}^{-2s} |r|^2$  from the RHS. This signifies that the observability of the whole system is possible without making any observation on the solid. A priori this is not obvious for the following reason: even though the same term is present in the LHS of (3.3) there is no large parameter  $s$  or  $\lambda$  multiplying it. We need to use additional properties of our system. More precisely, we exploit the fact that the state space of the “solid part” of the model is of finite dimension. Our goal will be achieved in two steps. As a first step, we prove in this section an intermediate inequality (4.1) written down below. The final inequality will be established in the next section (see (5.5)).

Let  $E$  be the vector space of solutions to system (2.3) obtained by varying  $(f_s, r^0, r^1)$ . Consider the following subspace of  $E$ :

$$E_b = \left\{ (\psi, r) \in E \mid r(T/2) = 0 \right\}.$$

We see that  $E_b$  is of infinite dimension and is of codimension  $\leq N$ . In the following arguments, we will suppose that  $E_b$  is of codimension  $= N$  (other cases can be treated in a similar manner). In such a case, there exist  $(\hat{\psi}^\ell, \hat{r}^\ell) \in E$ , with  $\ell \in \{1, \dots, N\}$ , such that

$$\hat{r}^\ell(T/2) = \vec{e}_\ell \quad \text{where} \quad \vec{e}_\ell = (\delta_{i,\ell})_{1 \leq i \leq N}.$$

( $\delta_{i,\ell}$  is the so-called Kronecker symbol.) Let  $E_0$  be the space spanned by  $\{\hat{r}^\ell \mid \ell \in \{1, \dots, N\}\}$ , and  $E_f$  be the subspace spanned by  $\{(\hat{\psi}^\ell, \hat{r}^\ell) \mid \ell \in \{1, \dots, N\}\}$  so that we have

$$E = E_b \oplus E_f.$$

Let us denote by  $\pi_f : E \rightarrow E_f$  the mapping defined by

$$\pi_f(\psi, r) = \sum_{\ell=1}^N (r(T/2) \cdot \vec{e}_\ell) (\hat{\psi}^\ell, \hat{r}^\ell).$$

Observe that  $(\psi, r) - \pi_f(\psi, r) \in E_b$  for all  $(\psi, r) \in E$ . Further we set  $\pi_0(\psi, r) = r$  for all  $(\psi, r) \in E$ , and we define  $\pi : E \rightarrow E_0$  by  $\pi = \pi_0 \circ \pi_f$ . Then, we have

$$\pi(\psi, r) = \sum_{\ell=1}^N (r(T/2) \cdot \vec{e}_\ell) \hat{r}^\ell.$$

**Lemma 4.** *If  $(\psi, r) \in E_b$ , then*

$$\int_0^T \rho_{\Gamma_b}^{-2s} |r|^2 \leq C_1(T) \int_0^T \rho_{\Gamma_b}^{-2s} |r'|^2 \leq C_2(T) \int_0^T \frac{e^{3\lambda\eta} |\Gamma_b|}{t^{3k}(T-t)^{3k}} \rho_{\Gamma_b}^{-2s} |r'|^2,$$

for some  $C_2(T) \geq C_1(T) > 0$ .

*Proof.* Since

$$\frac{e^{3\lambda\eta}|\Gamma_b|}{t^{3k}(T-t)^{3k}} \geq C > 0,$$

the second inequality is obvious and so we focus on the first one. We have (since  $r(T/2) = 0$ )

$$r(t) = - \int_t^{T/2} r'(\tau) d\tau \quad \text{for all } 0 \leq t \leq T/2,$$

and (since  $\rho_{\Gamma_b}^{-s}$  is increasing on  $[0, T/2]$ )

$$\rho_{\Gamma_b}^{-s}(t)|r(t)| \leq \rho_{\Gamma_b}^{-s}(t) \int_t^{T/2} |r'(\tau)| d\tau \leq \int_t^{T/2} \rho_{\Gamma_b}^{-s}(\tau)|r'(\tau)| d\tau$$

for all  $0 \leq t \leq T/2$ . Thus we obtain

$$\int_0^{T/2} \rho_{\Gamma_b}^{-2s}(t)|r(t)|^2 dt \leq \frac{T^2}{4} \int_0^{T/2} \rho_{\Gamma_b}^{-2s}(\tau)|r'(\tau)|^2 d\tau.$$

Similarly we have (using the fact that  $\rho_{\Gamma_b}^{-s}$  is decreasing on  $[T/2, T]$ )

$$\int_{T/2}^T \rho_{\Gamma_b}^{-2s}(t)|r(t)|^2 dt \leq \frac{T^2}{4} \int_{T/2}^T \rho_{\Gamma_b}^{-2s}(\tau)|r'(\tau)|^2 d\tau.$$

The proof is completed by adding the above two inequalities.  $\square$

With these preparations, we can now consider the inequality (3.3) and estimate the last term of the right-hand side of the inequality as follows: Writing  $r = r - \pi(\psi, r) + \pi(\psi, r)$  and noting that  $r - \pi(\psi, r) \in E_b$ , we have by Lemma 4

$$\begin{aligned} \int_0^T \rho_{\Gamma_b}^{-2s}|r|^2 &\leq 2C_2(T) \int_0^T \frac{e^{3\lambda\eta}|\Gamma_b|}{t^{3k}(T-t)^{3k}} \rho_{\Gamma_b}^{-2s}|r'|^2 \\ &\quad + 2C_2(T) \int_0^T \frac{e^{3\lambda\eta}|\Gamma_b|}{t^{3k}(T-t)^{3k}} \rho_{\Gamma_b}^{-2s}|(\pi(\psi, r))'|^2 \\ &\quad + 2 \int_0^T \rho_{\Gamma_b}^{-2s}|\pi(\psi, r)|^2. \end{aligned}$$

Note that the first term can be absorbed in the left-hand side of (3.3) by choosing  $\lambda$  large. More precisely, we have

$$CC_2(T) \int_0^T \frac{e^{3\lambda\eta}|\Gamma_b|}{t^{3k}(T-t)^{3k}} \rho_{\Gamma_b}^{-2s}|r'|^2 \leq \frac{1}{2}s^3\lambda^3 \int_0^T \frac{e^{3\lambda\eta}|\Gamma_b|}{t^{3k}(T-t)^{3k}} \rho_{\Gamma_b}^{-2s}|r'|^2,$$

for  $\lambda$  large.

As in [10], from (3.3), we can also derive an estimate for

$$s^{-1} \int_Q \xi^{-1} (|\psi'|^2 + |\Delta\psi|^2).$$

Thus estimate (3.3) gives **Carleman inequality (II)**:

$$\begin{aligned} & s^{-1} \int_Q \xi^{-1} (|\psi'|^2 + |\Delta\psi|^2) + \|M_1\psi\|_{L^2(Q)}^2 + \|M_2\psi\|_{L^2(Q)}^2 \\ & + s\lambda^2 \int_Q \frac{e^{\lambda\eta}}{t^k(T-t)^k} |\nabla\psi|^2 \\ & + s^3\lambda^4 \int_Q \frac{e^{3\lambda\eta}}{t^{3k}(T-t)^{3k}} |\psi|^2 + s\lambda \int_{\Sigma_b} \frac{e^{\lambda\eta}}{t^k(T-t)^k} |\partial_n\psi|^2 \\ & + \int_0^T \rho_{\Gamma_b}^{-2s} (|r''|^2 + |r|^2) + s^3\lambda^3 \int_0^T \frac{e^{3\lambda\eta}|_{\Gamma_b}}{t^{3k}(T-t)^{3k}} \rho_{\Gamma_b}^{-2s} |r'|^2 \\ & \leq C \left\{ \int_Q |f_s|^2 + s^3\lambda^4 \int_{\omega \times (0,T)} \frac{e^{3\lambda\eta}}{t^{3k}(T-t)^{3k}} |\psi|^2 + \int_0^T \rho_{\Gamma_b}^{-2s} |\pi(\psi, r)|^2 \right. \\ & \quad \left. + \int_0^T \frac{e^{3\lambda\eta}|_{\Gamma_b}}{t^{3k}(T-t)^{3k}} \rho_{\Gamma_b}^{-2s} |(\pi(\psi, r))'|^2 \right\}, \end{aligned} \tag{4.1}$$

where  $\xi(x, t) = \frac{e^{\lambda\eta}}{t^k(T-t)^k}$ .

## 5. Compactness argument and Carleman inequality (III)

The aim in this section is to show that we can strengthen the inequality (4.1) by removing the last two terms from RHS of (4.1). To this end, we set

$$\begin{aligned} I(\psi, r) = & s^{-1} \int_Q \xi^{-1} (|\psi'|^2 + |\Delta\psi|^2) + \int_Q |M_1\psi|^2 + \int_Q |M_2\psi|^2 \\ & + s\lambda^2 \int_Q \frac{e^{\lambda\eta}}{t^k(T-t)^k} |\nabla\psi|^2 + s^3\lambda^4 \int_Q \frac{e^{3\lambda\eta}}{t^{3k}(T-t)^{3k}} |\psi|^2 \\ & + s\lambda \int_{\Sigma_b} \frac{e^{\lambda\eta}}{t^k(T-t)^k} |\partial_n\psi|^2 \\ & + \int_0^T \rho_{\Gamma_b}^{-2s} (|r''|^2 + |r|^2) + s^3\lambda^3 \int_0^T \frac{e^{3\lambda\eta}|_{\Gamma_b}}{t^{3k}(T-t)^{3k}} \rho_{\Gamma_b}^{-2s} |r'|^2, \end{aligned}$$

$$J(\psi, r) = K(\psi, r) + \int_0^T \rho_{\Gamma_b}^{-2s} |\pi(\psi, r)|^2 + s^3\lambda^3 \int_0^T \frac{e^{3\lambda\eta}|_{\Gamma_b}}{t^{3k}(T-t)^{3k}} \rho_{\Gamma_b}^{-2s} |(\pi(\psi, r))'|^2,$$

and

$$K(\psi, r) = \int_Q |f_s|^2 + s^3\lambda^4 \int_{\omega \times (0,T)} \frac{e^{3\lambda\eta}}{t^{3k}(T-t)^{3k}} |\psi|^2.$$

In the previous section, we have proved that there exists a constant  $C > 0$  such that

$$I(\psi, r) \leq C J(\psi, r). \tag{5.1}$$

(From now on, we do not vary the parameters  $(s, \lambda)$  and fix them so that the above inequality holds.) We want to improve this estimate by showing that there exists a constant  $C(s, \lambda) > 0$ , depending on  $s$  and  $\lambda$ , such that

$$I(\psi, r) \leq C(s, \lambda) K(\psi, r). \quad (5.2)$$

This is the **Carleman inequality (III)** that we have for the system (2.3). To prove the inequality (5.2), we argue by contradiction. We suppose that there exists a sequence  $(\psi_j, r_j)_j$  associated with the data  $(f_j, r_j^0, r_j^1)$  such that

$$I(\psi_j, r_j) = 1 \quad \text{and} \quad \lim_{j \rightarrow \infty} K(\psi_j, r_j) = 0.$$

We can assume that there exists a pair  $(\psi, r) \in L^2_{\text{loc}}(Q) \times L^2_{\text{loc}}(0, T)$  and that – after extraction of a subsequence – the sequence  $(\psi_j, r_j)_j$  enjoys the following convergence properties in the indicated weighted spaces:

$\psi'_j \rightharpoonup \psi'$	for the weak topology of $L^2(\xi^{-1}; Q)$ ,
$\Delta\psi_j \rightharpoonup \Delta\psi$	for the weak topology of $L^2(\xi^{-1}; Q)$ ,
$\nabla\psi_j \rightharpoonup \nabla\psi$	for the weak topology of $L^2(e^{\lambda\eta}t^{-k}(T-t)^{-k}; Q)$ ,
$\psi_j \rightharpoonup \psi$	for the weak topology of $L^2(e^{3\lambda\eta}t^{-3k}(T-t)^{-3k}; Q)$ ,
$\partial_n\psi_j \rightharpoonup \partial_n\psi$	for the weak topology of $L^2(e^{\lambda\eta} \Gamma_b t^{-k}(T-t)^{-k}; \Sigma_b)$ ,
$r''_j \rightharpoonup r''$	for the weak topology of $L^2(\rho_{\Gamma_b}^{-2s}; (0, T))$ ,
$r_j \rightharpoonup r$	for the weak topology of $L^2(\rho_{\Gamma_b}^{-2s}; (0, T))$ ,
$r'_j \rightharpoonup r'$	for the weak topology of $L^2(\rho_{\Gamma_b}^{-2s}t^{-3k}(T-t)^{-3k}; (0, T))$ .

In the next two subsections, we will deduce that  $\psi \equiv 0$ ,  $r \equiv 0$ , and that

$$\int_0^T \rho_{\Gamma_b}^{-2s} |\pi(\psi_j, r_j)|^2 + s^3 \lambda^3 \int_0^T \frac{e^{3\lambda\eta} |\Gamma_b|}{t^{3k}(T-t)^{3k}} \rho_{\Gamma_b}^{-2s} |(\pi(\psi_j, r_j))'|^2 \rightarrow 0. \quad (5.3)$$

It follows then that  $J(\psi_j, r_j) \rightarrow 0$ . From (5.1), we conclude that  $I(\psi_j, r_j) \rightarrow 0$ . This is in contradiction to  $I(\psi_j, r_j) = 1$  and proves (5.2).

### 5.1. Passage to the limit in problem (2.3)

To prove that  $\psi \equiv 0$  and  $r \equiv 0$ , we first show that we can pass to the limit in system (2.3). To pass to the limit in the equation

$$M_1\psi_j + M_2\psi_j = \rho^{-s}f_j + s(\Delta\beta)\psi_j,$$

we use the  $L^2$ -estimate on  $(M_1\psi_j)_j$  and  $(M_2\psi_j)_j$ . Hence the subsequences  $(M_1\psi_j)_j$  and  $(M_2\psi_j)_j$  weakly converge in  $L^2(Q)$ . To identify their limits, it is enough to take test functions in  $\mathcal{D}(Q)$  and to pass to the limit. Thanks to the above convergence we get

$$M_1\psi_j \xrightarrow{L^2(Q)} M_1\psi, \quad M_2\psi_j \xrightarrow{L^2(Q)} M_2\psi,$$

$$s(\Delta\beta)\psi_j \rightharpoonup s(\Delta\beta)\psi \text{ weakly in } L^2(e^{3\lambda\eta}t^{-3k}(T-t)^{-3k}; Q).$$

Next we use

$$K(\psi_j, r_j) \rightarrow 0.$$

This shows that  $\rho^{-s} f_j \rightarrow 0$  in  $L^2(Q)$  and  $\psi = 0$  in  $\omega \times (0, T)$ . With this information, we see that

$$M_1\psi + M_2\psi = s(\Delta\beta)\psi \quad \text{in } Q, \quad \text{and} \quad \psi = 0 \quad \text{in } \omega \times (0, T).$$

The passage to the limit in the boundary conditions on  $\Sigma$  is easily done because we have weak convergence of  $(\psi_j)_j$  towards  $\psi$  in the space  $L^2(e^{3\lambda\eta} t^{-3k}(T-t)^{-3k}; 0, T; H^1(\Omega))$ .

To pass to the limit in the equation satisfied by  $r_j$ , we use the weak convergence of  $(\partial_n \psi_j)_j$  towards  $\partial_n \psi$  in the space  $L^2(e^{\lambda\eta} |\Gamma_b| t^{-k}(T-t)^{-k}; \Sigma_b)$ .

This proves that  $(\psi, r)$  satisfies the system (2.3) with  $f_s = 0$ .

To deduce that  $\psi \equiv 0$  and  $r \equiv 0$ , we pass from  $\psi$  to  $\phi = \rho^s \psi$ . We see that  $(\phi, r)$  satisfies the system (1.2) with  $f = 0$ . In addition, we have  $\phi \equiv 0$  in  $\omega \times (0, T)$ . Applying the unique continuation principle for the heat equation [12], we obtain  $\phi = 0$  in  $Q$ , and hence  $\psi = 0$  in  $Q$ . Going back to the system satisfied by  $(\psi, r)$ , we deduce successively that  $r' = 0$ ,  $r'' = 0$  and  $r = 0$ . In particular, we have shown that

$$\begin{aligned} r_j &\rightharpoonup 0 && \text{for the weak topology of } L^2(\rho_{\Gamma_b}^{-2s}; (0, T)), \\ r'_j &\rightharpoonup 0 && \text{for the weak topology of } L^2(\rho_{\Gamma_b}^{-2s} t^{-3k}(T-t)^{-3k}; (0, T)). \end{aligned} \tag{5.4}$$

## 5.2. Proof of (5.3)

We equip the space

$$H = \left\{ r \in H_{\text{loc}}^1(0, T; \mathbb{R}^N) \mid \|r\|_{L^2(\rho_{\Gamma_b}^{-2s}; (0, T))} + \|r'\|_{L^2(\rho_{\Gamma_b}^{-2s} t^{-3k}(T-t)^{-3k}; (0, T))} < \infty \right\}$$

with the norm

$$\|r\|_H = \|r\|_{L^2(\rho_{\Gamma_b}^{-2s}; (0, T))} + \|r'\|_{L^2(\rho_{\Gamma_b}^{-2s} t^{-3k}(T-t)^{-3k}; (0, T))}.$$

The mapping

$$r \longmapsto r(T/2)$$

is continuous from  $H$  into  $\mathbb{R}^N$  since

$$|r(T/2)| \leq C(\|r'\|_{L^2(\rho_{\Gamma_b}^{-2s} t^{-3k}(T-t)^{-3k}; (0, T))} + \|r\|_{L^2(\rho_{\Gamma_b}^{-2s}; (0, T))}).$$

Therefore it is also compact. Due to (5.4),  $|r_j(T/2)| \rightarrow 0$  (or at least a subsequence). The proof of (5.3) is complete.

Grouping all the previous results, let us summarize, for the reader's convenience, the estimate that we have established so far on the system (2.3): For  $\lambda$

sufficiently large, there is  $s_0(\lambda) > 0$  such that, for  $s \geq s_0(\lambda)$ , we have

$$\begin{aligned} & \int_Q \xi^{-1}(|\psi'|^2 + |\Delta\psi|^2) + \int_Q |M_1\psi|^2 + \int_Q |M_2\psi|^2 + \int_Q \frac{e^{\lambda\eta}}{t^k(T-t)^k} |\nabla\psi|^2 \\ & + \int_{\Sigma_b} \frac{e^{\lambda\eta}}{t^k(T-t)^k} |\partial_n\psi|^2 + \int_Q \frac{e^{3\lambda\eta}}{t^{3k}(T-t)^{3k}} |\psi|^2 \\ & + \int_0^T \rho_{\Gamma_b}^{-2s} (|r''|^2 + |r|^2) + \int_0^T \frac{e^{3\lambda\eta}|_{\Gamma_b}}{t^{3k}(T-t)^{3k}} \rho_{\Gamma_b}^{-2s} |r'|^2 \\ & \leq C(s, \lambda) \left\{ \int_Q |f_s|^2 + \int_{\omega \times (0, T)} \frac{e^{3\lambda\eta}}{t^{3k}(T-t)^{3k}} |\psi|^2 \right\}. \end{aligned} \quad (5.5)$$

Since the constant  $C(s, \lambda)$  is not explicitly known in terms of  $s$  and  $\lambda$ , we have dropped out the constant weights  $s^\alpha \lambda^\beta$  in front of each term of the above inequality.

## 6. Carleman inequality (IV)

The purpose here is to translate the Carleman inequality (5.5) from the transformed system (2.3) to original system (1.2). This procedure is very classical and we refer for example to [10]. It yields the following estimate on  $(\phi, r)$ , solution of the original system (1.2).

**Theorem 5.** *Consider the coupled system (1.2). Then there exist positive constants  $\lambda_0$  and  $s_0(\lambda)$  such that the following inequality holds for all  $\lambda \geq \lambda_0, s \geq s_0(\lambda)$  and for all solutions  $(\phi, r)$  of the system (1.2):*

$$\begin{aligned} & \int_Q \rho^{-2s} \xi^{-1} (|\phi'|^2 + |\Delta\phi|^2) + \int_Q \rho^{-2s} \frac{e^{\lambda\eta}}{t(T-t)} |\nabla\phi|^2 \\ & + \int_Q \rho^{-2s} \frac{e^{3\lambda\eta}}{t^{3k}(T-t)^{3k}} |\phi|^2 \\ & + \int_0^T \rho_{\Gamma_b}^{-2s} (|r''|^2 + |r|^2) + \int_0^T \frac{e^{3\lambda\eta}|_{\Gamma_b}}{t^{3k}(T-t)^{3k}} \rho_{\Gamma_b}^{-2s} |r'|^2 \\ & \leq C(s, \lambda) \left\{ \int_Q \rho^{-2s} |\phi' - \Delta\phi|^2 + \int_{\omega \times (0, T)} \rho^{-2s} \frac{e^{3\lambda\eta}}{t^{3k}(T-t)^{3k}} |\phi|^2 \right\}. \end{aligned} \quad (6.1)$$

## 7. Null-controllability result

In this section, we establish null-controllability of our original system (1.2), as a consequence of the Carleman inequality (6.1), with a forcing term in the form:

$$f(x, t) = u(x, t)\chi_{\omega \times (0, T)}(x, t),$$

where  $\omega \subset \subset \Omega$ .

The null-controllability result for the “Heat-Finite-Dimensional Beam System” follows from an observability inequality for the adjoint system

$$\begin{aligned} -y' - \Delta y &= 0 \quad \text{in } Q, \quad y = 0 \quad \text{on } \Sigma_e, \quad y = -q' \cdot \zeta \quad \text{on } \Sigma_b, \\ y(0, x_2, t) &= y(L, x_2, t) \\ \text{and } \partial_n y(0, x_2, t) &= -\partial_n y(L, x_2, t) \quad \text{for } (x_2, t) \in ]0, 1[ \times ]0, T[, \\ y(T) &= y_T \quad \text{in } \Omega, \\ q'' + Aq &= - \int_{\Gamma_b} \partial_n y \zeta \quad \text{in } (0, T), \\ q(T) &= q_T^0, \quad \text{and} \quad q'(T) = q_T^1 \quad \text{in } \mathbb{R}^N. \end{aligned}$$

It is obvious that  $(y, q)$  also obeys the Carleman inequality (6.1).

Thanks to this inequality, the following observability inequality can be established (see, e.g., [10, Lemma 13.1])

$$\|y(0)\|_{L^2}^2 + |A^{1/2}q(0)|^2 + |q'(0)|^2 \leq C \int_{\omega \times (0, T)} \rho^{-2s} \frac{e^{3\lambda\eta}}{t^{3k}(T-t)^{3k}} |y|^2. \quad (7.1)$$

To derive null-controllability from the above observability inequality we use a classical penalization method consisting in solving the optimal control problem

$$(\mathcal{P}_\varepsilon) \quad \inf \left\{ J_\varepsilon(\phi, r, v) \mid (\phi, r, v) \text{ obeys (1.2) with } f = \chi_{\omega \times (0, T)} \sqrt{w} v \right\}$$

where

$$J_\varepsilon(\phi, r, v) = \frac{1}{2\varepsilon} \int_Q |\phi(T)|^2 + \frac{1}{2\varepsilon} |A^{1/2}r(T)|^2 + \frac{1}{2\varepsilon} |r'(T)|^2 + \frac{1}{2} \int_Q |v|^2,$$

and the function  $w$  is the weight appearing in the Carleman inequality (6.1) and in the observability inequality (7.1):

$$w(x, t) = \rho^{-2s} \frac{e^{3\lambda\eta}}{t^{3k}(T-t)^{3k}}.$$

Notice that if  $v \in L^2(Q)$  then  $u = \sqrt{w} v$  also belongs to  $L^2(Q)$ .

It is easy to prove that problem  $(\mathcal{P}_\varepsilon)$  admits a unique solution  $(\phi_\varepsilon, r_\varepsilon, v_\varepsilon)$ , and that the optimal control  $v_\varepsilon$  can be characterized by

$$v_\varepsilon = -\chi_{\omega \times (0, T)} \sqrt{w} y_\varepsilon,$$

where  $(y_\varepsilon, q_\varepsilon)$  is the solution to the adjoint system

$$\begin{aligned} -y'_\varepsilon - \Delta y_\varepsilon &= 0 \quad \text{in } Q, \quad y_\varepsilon = 0 \quad \text{on } \Sigma_e, \quad y_\varepsilon = -q'_\varepsilon \cdot \zeta \quad \text{on } \Sigma_b, \\ y_\varepsilon(0, x_2, t) &= y_\varepsilon(L, x_2, t), \\ \text{and } \partial_n y_\varepsilon(0, x_2, t) &= -\partial_n y_\varepsilon(L, x_2, t) \quad \text{for } (x_2, t) \in ]0, 1[ \times ]0, T[, \\ y_\varepsilon(T) &= \frac{1}{\varepsilon} \phi_\varepsilon(T) \quad \text{in } \Omega, \\ q''_\varepsilon + A q_\varepsilon &= - \int_{\Gamma_b} \partial_n y_\varepsilon \zeta \quad \text{in } (0, T), \\ q_\varepsilon(T) &= -\frac{1}{\varepsilon} r_\varepsilon(T), \quad \text{and } q'_\varepsilon(T) = -\frac{1}{\varepsilon} r'_\varepsilon(T) \quad \text{in } \mathbb{R}^N. \end{aligned}$$

With integration by parts we can prove that the optimal solution  $(\phi_\varepsilon, r_\varepsilon, v_\varepsilon)$  to problem  $(\mathcal{P}_\varepsilon)$  and the corresponding adjoint state  $(y_\varepsilon, q_\varepsilon)$  satisfies

$$\begin{aligned} &\frac{1}{\varepsilon} \int_\Omega |\phi_\varepsilon(T)|^2 + \frac{1}{\varepsilon} |A^{1/2} r_\varepsilon(T)|^2 + \frac{1}{\varepsilon} |r'(T)|^2 + \int_{\omega \times (0, T)} w |y_\varepsilon|^2 \\ &= \int_\Omega \phi^0 y_\varepsilon(0) - A q_\varepsilon(0) \cdot r^0 - q'_\varepsilon(0) \cdot r^1 \\ &\leq C \left( \|\phi^0\|_{L^2(\Omega)}^2 + |A^{1/2} r^0|^2 + |r^1|^2 \right)^{1/2} \\ &\quad \times \left( \|y_\varepsilon(0)\|_{L^2(\Omega)}^2 + |A^{1/2} q_\varepsilon(0)|^2 + |q'_\varepsilon(0)|^2 \right)^{1/2} \\ &\leq C \left( \|\phi^0\|_{L^2(\Omega)}^2 + |A^{1/2} r^0|^2 + |r^1|^2 \right)^{1/2} \left( \|\chi_{\omega \times (0, T)} \sqrt{w} y_\varepsilon\|_{L^2(Q)}^2 \right)^{1/2}. \end{aligned}$$

Therefore, we have

$$\begin{aligned} &\frac{1}{\varepsilon} \int_\Omega |\phi_\varepsilon(T)|^2 + \frac{1}{\varepsilon} |A^{1/2} r_\varepsilon(T)|^2 + \frac{1}{\varepsilon} |r'_\varepsilon(T)|^2 + \int_{\omega \times (0, T)} w |y_\varepsilon|^2 \\ &\leq C \left( \|\phi^0\|_{L^2(\Omega)}^2 + |A^{1/2} r^0|^2 + |r^1|^2 \right). \end{aligned} \tag{7.2}$$

and the sequence  $(v_\varepsilon)_\varepsilon = (-\chi_{\omega \times (0, T)} \sqrt{w} y_\varepsilon)_\varepsilon$  is bounded in  $L^2(Q)$ .

Without loss of generality, we can assume that the sequence  $(v_\varepsilon)_\varepsilon$  converges to some  $v \in L^2(Q)$  for the weak topology of  $L^2(Q)$ . From that and with Theorem 2, we first deduce that  $(\phi_\varepsilon, r_\varepsilon)_\varepsilon$  converges, for the weak topology of  $(L^2(0, T; V) \cap H^1(0, T; H^{-1}(\Omega))) \times H^1(0, T; \mathbb{R}^N)$ , to the solution  $(\phi, r)$  of system (1.2) corresponding to  $f = \chi_{\omega \times (0, T)} \sqrt{w} v$ . Moreover, due to (7.2),  $(\phi_\varepsilon(T))_\varepsilon$  converges to zero in  $L^2(\Omega)$ ,  $(r_\varepsilon(T))_\varepsilon$  converges to zero in  $\mathbb{R}^N$ , and  $(r'_\varepsilon(T))_\varepsilon$  converges to zero in  $\mathbb{R}^N$ . Thus  $(\phi(T), r(T), r'(T)) = 0$  and  $u = \sqrt{w} v$  is a required control providing solution to our null-controllability problem.  $\square$

**Acknowledgements.** The authors have been supported by CEFIPRA within the project 3701-1 “Control of systems of partial differential equations”.

## References

- [1] M. Boulakia and A. Osses, Local null controllability of a two-dimensional fluid-structure interaction problem, *ESAIM COCV*, 14 (1) (2008), 1–42.
- [2] C. Conca, J. Planchard, B. Thomas, and M. Vanninathan, *Problèmes mathématiques en couplage fluide-structure*, Eyrolles, Paris, 1994.
- [3] C. Conca, J. Planchard, and M. Vanninathan, *Fluids and periodic structures*, Masson and J. Wiley, Paris, 1995.
- [4] O.Yu. Èmanuilov, Controllability of parabolic equations, *Mat. Sbornik*, 186 (6) (1995), 109–132.
- [5] A.V. Fursikov, Optimal Control of distributed systems, Theory and Applications, Translations of Mathematical Monographs # 187, AMS Providence RI, 2000.
- [6] A.V. Fursikov and O.Yu. Imanuilov, Controllability of evolution equations, Lecture Notes series 34, Seoul National University, Research Institute of Mathematics, Global Analysis Research Centre, Seoul 1996.
- [7] O. Imanuilov and T. Takahashi, Exact controllability of a fluid-rigid body system, *J. Math. Pures Appl.*, 87 (2007), 408–437.
- [8] J.-L. Lions, *Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués*, Masson, Paris, 1988.
- [9] J.-P. Raymond and M. Vanninathan, Exact controllability in fluid-solid structure: the Helmholtz model, *ESAIM Control Optim. Calc. Var.*, 11(2) (2005), 180–203 (electronic).
- [10] J.-P. Raymond and M. Vanninathan, Null Controllability in a Heat-Solid Structure Model, *Applied Math. Optim.*, 2008, DOI 10.1007/s00245-008-9053-x.
- [11] J.-P. Raymond and M. Vanninathan, Null Controllability in a Fluid-Solid Structure Model, in preparation.
- [12] J.C. Saut and B. Scheurer, Unique Continuation for some evolution equations, *J. Diff. Equations*, 66 (1987), 118–139.

Jean-Pierre Raymond

Université de Toulouse, UPS, Institut de Mathématiques  
F-31062 Toulouse Cedex 9, France

*and*

CNRS, Institut de Mathématiques, UMR 5219  
F-31062 Toulouse Cedex 9, France  
e-mail: [raymond@math.univ-toulouse.fr](mailto:raymond@math.univ-toulouse.fr)

Muthusamy Vanninathan  
TIFR-CAM, Post Bag 6503  
Bangalore 560065, India  
e-mail: [vanni@math.tifrbng.res.in](mailto:vanni@math.tifrbng.res.in)

# Feedback Modal Control of Partial Differential Equations

Thomas I. Seidman

**Abstract.** For hybrid systems in which control consists of selection from a discrete finite set of *modes*, a somewhat unfamiliar formulation is needed for analysis of the possibility of closed loop (feedback) control. We are here concerned to examine the desiderata for such feedback from the viewpoint of descriptive modeling of implementation in a PDE context. A principal result is global existence, in an appropriate sense, for the implemented closed loop control system. A problem of transport on a graph is then presented to show how the relevant hypotheses might be satisfied in a PDE example.

**Mathematics Subject Classification (2000).** 93A30, 93B12, 47J40, 70K70.

**Keywords.** Modeling, multiscale, hybrid systems, switching, modes, discontinuities, differential equations, feedback, Zeno phenomena.

## 1. Introduction

Consider a collection of partial differential equations which we take, somewhat arbitrarily, to have the form

$$\dot{x} = \mathbf{A}_j x + f_j(x) \quad (j \in \mathcal{J}) \quad (1.1)$$

where each  $\mathbf{A}_j$  is a suitable differential operator. Now imagine a system whose evolution is governed, over interswitching intervals, by one or another of these; we call this a *hybrid system* and take the discrete *modal index*  $\mathbf{j} = \mathbf{j}(t)$  to be a component of the system state along with the ‘continuous component’  $x$ . We will be considering the modal transitions  $[\mathbf{j}(t-) = j] \rightsquigarrow [\mathbf{j}(t+) = j']$  as our control mechanism for the system. In particular, we will be concerned with the possibility of closed loop operation of such a control system.

Such hybrid systems are already a much-studied area of interest in the lumped parameter (ordinary differential equation) context, although much of the underlying theory remains open. Although one of the early analyses of such systems [7] was, indeed, motivated by a PDE example (not in a control context), we note that

very little has been done so far for the analysis of switching control for distributed parameter problems governed by partial differential equations, despite the fact that significant applications come easily to mind, involving the use of such familiar ON/OFF control devices as valves and pumps, light switches and thermostats, traffic signals, etc.

[One might also consider systems with  $\mathbf{j} = \mathbf{j}(t, s)$  pointwise, compare [9] where this becomes a free boundary problem for the set  $\{(t, s) : \mathbf{j}(t, s) = j\}$ . Here we will restrict our attention to situations where the switching may be viewed as global with a finite number of modes – as is the case, e.g., for traffic signals if we take each mode as specifying the configuration of signal states for the entire road network. As with hybrid ODEs, the index  $j$  will be a function of  $t$  alone.] Many, but not all, of the relevant aspects of the analysis are then independent of dimensionality.

In the context of feedback for PDEs, the regularity of the sensor inputs being considered may be significant even to know that solutions exist. Our formulation reflects a concern for the modeling of such systems. This will be very much a question of time scales: we are assuming that the switching itself takes place on a time scale more rapid than our modeling concerns but that the interswitching intervals are on the scale of interest.

While other considerations may also be of interest – e.g., controllability or stabilization to a small region – we here envision three canonical results:

**Theorem 1.** *Under appropriate hypotheses, treating  $\mathbf{j}(\cdot)$  as data, the system will be well posed in some suitable sense.*

Again treating open-loop control with a suitable cost functional,

**Theorem 2.** *Under appropriate hypotheses there exists an optimal control  $\mathbf{j}$ , minimizing the cost. For the autonomous infinite horizon problem this can be obtained by a kind of feedback.*

Modifying that notion of *feedback* to be based on suitable sensors,

**Theorem 3.** *Under appropriate hypotheses the feedback controlled system will be well posed in some suitable sense.*

Much of the paper is devoted to explaining what these should mean.

## 2. The formulation

We begin by noting an important distinction between *descriptive* and *prescriptive* modes of modeling: the first is what a scientist does in trying to understand the various patterns arising in the world; the second is what a composer or an engineer does in designing (artificial) patterns for various purposes. E.g., in viewing leonine behavior the first is the modality of the naturalist while the second is the approach of a lion tamer. In this section we provide a formal prescriptive model for lion tamers, while aware that a naturalist's comments will later be complementary in describing how lions can behave.

The formal elements of the underlying system consist of

(E)

- a. a finite *index set*  $\mathcal{J}$  and a corresponding set of *state spaces*  $\mathcal{X}_j$ .
- b. a set of *information spaces*  $\mathcal{Y}_j$  with *sensor maps*  $Y_j : \mathcal{X}_j \rightarrow \mathcal{Y}_j$ .
- c. a nonempty *action set*  $\mathcal{A}_j(y) \subset \mathcal{J}$  for each  $j \in \mathcal{J}$ ,  $y \in \mathcal{Y}_j$ .
- d. a continuous *transition map*  $\mathbf{f} : [j, x] \mapsto [j', x']$  with  $x' \in \mathcal{X}_{j'}$  defined when  $j \in \mathcal{J}$ ,  $x \in \mathcal{X}_j$  and  $j' \in \mathcal{A}_j(Y_j(x))$ .
- e. a set of *dynamical systems = modes*  $\pi_j$ , each satisfying the *causality condition*

$$\pi_j(t, s, \xi) = \pi_j(t, r, \pi_j(r, s, \xi)) \quad \text{for } t \geq r \geq s, \quad \xi \in \mathcal{X}_j \quad (2.1)$$

We assume throughout that each  $\mathcal{X}_j, \mathcal{Y}_j$  is a complete metric space and that each  $\pi_j$  and  $\mathbf{f}$  is continuous; at this point we impose no continuity requirements on  $Y_j$ . [We expect the modes of (E)-e. to be given as in (1.1) so continuity of  $\pi_j$  just means well-posedness. Note, however, that any relevant boundary data is then to be included in specification of the mode.]

We may anthropomorphize the feedback as a *controller* who knows the current mode and sensor values  $j = \mathbf{j}(t)$  and  $y = \mathbf{y}(t) = Y_j(\mathbf{x}(t))$  and, based on this, continually selects the mode. Of course the controller's choices at any moment are restricted to the available *control actions*: to remain in the current mode  $j$  or to make a transition  $j \curvearrowright j'$  on the fast scale; the *switching rules* are just that this selection always be taken from the action set  $\mathcal{A}_j(y)$ . It will also be convenient to introduce the sets

$$\mathcal{S}^j := \{y \in \mathcal{Y}_j : j \in \mathcal{A}_j(y)\}, \quad \mathcal{C}^{j \curvearrowright j'} := \{y \in \mathcal{Y}_j : j' \in \mathcal{A}_j(y)\}, \quad (2.2)$$

noting that the required nonemptiness of each  $\mathcal{A}_j(y)$  ensures that

$$\mathcal{S}^j \cup \mathcal{B}^j = \mathcal{Y}_j \quad \text{where } \mathcal{B}^j = \bigcup_{j' \neq j} \mathcal{C}^{j \curvearrowright j'}. \quad (2.3)$$

We refer to the specification of the sets  $\{\mathcal{S}^j, \mathcal{B}^{j \curvearrowright j'}\}$  as the *switching diagram* for mode  $j$  and to these collectively ( $j \in \mathcal{J}$ ) as the controlling *feedback diagram* for the system.

Intuitively, the operation of such a feedback controlled system should produce finitely many switching times  $\{t_\nu : \nu = 1, \dots, N\}$  in any time interval  $[0, T]$  with a modal transition  $j_\nu \curvearrowright j_{\nu+1}$  at each  $t_\nu$  – thus partitioning  $[0, T]$  into the *interswitching intervals*  $\mathcal{I}_\nu = [t_{\nu-1}, t_\nu]$  with  $0 = t_0 \leq t_1 \leq \dots$ . A *solution* of such a feedback system on the time interval  $[0, T]$  would then be a triple of functions  $[\mathbf{j}(\cdot), \mathbf{x}(\cdot), \mathbf{y}(\cdot)]$  such that

(S)

- a.  $\mathbf{j}(\cdot)$  is piecewise constant with  $\mathbf{j}(t) = j_\nu \in \mathcal{J}$  for  $t$  in each interswitching interval  $\mathcal{I}_\nu = [t_{\nu-1}, t_\nu]$ ; at each  $t$  one will have  $\mathbf{x}(t) \in \mathcal{X}_{\mathbf{j}(t)}$  and  $\mathbf{y}(t) = Y_{\mathbf{j}(t)}(\mathbf{x}(t)) \in \mathcal{Y}_{\mathbf{j}(t)}$ .
- b. the switching times are discrete: finitely many in any  $[0, T]$ .
- c. switching  $[j_{\nu-1} \curvearrowright j_\nu$  at  $t_\nu]$  occurs only if  $j_\nu \in \mathcal{A}_{j_{\nu-1}}(\mathbf{y}(t_\nu-))$  while for  $t$  in the interior of  $\mathcal{I}_\nu$  one must have  $\mathbf{y}(t) \in \mathcal{S}^{j_{\nu-1}}$ .
- d. at each switching time  $t_\nu$

$$\mathbf{x}(t_\nu+) = \mathbf{f}(\mathbf{x}(t_\nu-); j_{\nu-1} \curvearrowright j_\nu). \quad (2.4)$$

- e. on each interswitching interval  $\mathcal{I}_\nu$  we have  $\mathbf{x}(t) \in \mathcal{X}_{j_{\nu-1}}$  with

$$\mathbf{x}(t) = \pi_{j_{\nu-1}}(t - t_{\nu-1}, \mathbf{x}(t_{\nu-1}+)). \quad (2.5)$$

Deferring further discussion to the next section, note that (S)-a. will admit the possibility of degenerate interswitching intervals ( $t_\nu = t_{\nu-1}$ ), for which we formally take  $\mathbf{x}(t_\nu-) = \mathbf{x}(t_{\nu-1}+)$  with no evolution. The occurrence of infinitely many transitions within a finite period is known as a *Zeno phenomenon* and this possibility would be a major technical difficulty for the theory; (S)-b. requires that this does not occur in the problems we consider.

**Remark 2.1.** We note a few generalizations which can be included within the framework of (E), (S).

We have formulated the feedback to depend only on the current sensor values  $Y_j(t) \in \mathcal{Y}_j$ , without memory. Note, however, that we can, e.g., treat a Luenberger observer by introducing it as a state component  $\hat{\mathbf{y}}$  adjoined to each  $\mathcal{X}_j, \mathcal{Y}_j$  with suitably defined dynamics involving the current sensor values and  $id_{\hat{\mathbf{y}}}$  adjoined to each  $Y_j$ .

One reason to exclude Zeno phenomena is to avoid potential difficulties with a recursive use of (S)-d.,e. in constructing solutions. In the proof of Theorem 2 below, the positive switching costs  $c(j \curvearrowright j')$  enforce this exclusion automatically in optimization. Such costs are a well-known practical reality (often corresponding to the residual effects of rapid scale transients in chattering; compare [8, 2]). It is therefore a common practice to introduce *dead time* following (some) transitions  $k \curvearrowright k'$ , temporarily preventing a repetition of the mode  $k$  or of that transition. We can include such dead time in the formulation by introducing a state component  $z \in \mathcal{Z} = \mathbb{R}^J$ , satisfying  $z_{k \curvearrowright k'} = 1$  for the relevant switching indices  $(k \curvearrowright k') \in \mathcal{J}' \subset \mathcal{J} \times \mathcal{J}$ , and adjoining  $\mathcal{Z}$  to each  $\mathcal{X}_j, \mathcal{Y}_j$ ; as part of  $\mathbf{f}(\cdot; k \curvearrowright k')$  one then resets  $z_{k \curvearrowright k'}$  to 0. Now one obtains the dead time effect by deleting  $k$  from each  $\mathcal{A}_j(y, z)$  or deleting  $k'$  from  $\mathcal{A}_k(y, z)$  until  $z_{k \curvearrowright k'}$  reaches its threshold. Note that being able to treat this kind of resetting of  $z$  to handle dead time is now our principal reason for retaining a transition map  $\mathbf{f}$  as part of (E).

There are also problems for which we do expect the possible occurrence of such behavior. It may then be convenient to view this behavior as a whole, defining within our framework a *chattering mode* (or idealized as a *sliding mode*).  $\square$

Without further hypotheses it is easy to construct examples for which no global solutions exist at all: for example, the switching rules as given might produce a sequence of switching times  $t_\nu \nearrow t_* < T$  – violating **(S)-b.** and also with no way to obtain a continuation after  $t_*$ . Also, since  $\mathcal{A}_j(y)$  need not be a singleton, we cannot expect that **(S)** will determine solution evolution uniquely when solutions do exist. Typically the sets  $\mathcal{C}^{j \rightsquigarrow j'}$  will be *switching surfaces* with the trajectories transverse to these and  $y$  leaving  $\mathcal{S}_j$  so switching is forced. We must, however, allow for the alternative possibility that  $y$ , continuing from  $y \in \mathcal{C}^{j \rightsquigarrow j'}$  using mode  $j$ , would remain (at least briefly) in  $\mathcal{S}_j$  and the choice would be genuine. Such *anomalous points* are a major technical difficulty for this theory and we will further discuss their effect later, noting here only that this is an inherent source of non-uniqueness for solutions since we will be accepting *both* possible choices as legitimate. In the next sections we re-examine the formulation above in the light of possible implementation and impose hypotheses ensuring existence.

### 3. Modeling and interpretation

Mathematical models are always created, selected, and analyzed with a purpose and we keep this functionality at the forefront of our present concern: convenience is one of the major desiderata in the selection of appropriate models. While control theory is inherently a prescriptive approach to the world in which we may be inclined to ignore the descriptive aphorism, “*Natura non facit saltus*” (“*Nature does not make jumps,*” attributed to Newton, Leibniz, Linnaeus, . . .), we recognize that any control design is useful only as implemented:

*A prescriptive model should be a descriptive model of its implementation*

so we must have some concern that the nominal behavior of these discontinuous systems is consistent with their actual behavior. In this section we complement the prescriptive formulation **(E)**, **(S)** with some interpretive comments on the construction from this point of view, clarifying our choices of assumptions.

The fundamental principle of such interpretation is that hybrid systems are a simplified description of multiscale problems in which the transitions  $j \rightsquigarrow j'$  which we are describing as ‘instantaneous’ are actually taking place on a faster time scale than we wish to model; see, e.g., [10]. [If  $\mathcal{X}_{j'} = \mathcal{X}_j$ , the transition function  $\mathbf{f} : \mathbf{x}(t-) \rightsquigarrow \mathbf{x}(t+)$  might then simply reflect the result of state evolution on the rapid scale.] As with any modeling, success means that we have taken into account those aspects whose effects are inescapable without treating details which can be ignored. Since our description is then an idealization of the world, we are inclusive in the consideration of mild solutions, so our version of well-posedness will require that

*The limit of solutions will itself be accepted as a solution,*

i.e., the solution set depends upper semicontinuously on the data.

Note that we are permitting degenerate interswitching intervals  $\mathcal{I}_\nu$ , with  $t_{\nu-1} = t_\nu$ . This might simply correspond to the possibility, which we want to include here, that distinct effects can occur simultaneously on the modeling scale, meaning only that we cannot determine priority without resolving aspects of the rapid behavior which we are content to leave hidden from us; we do insist that the sequencing, particularly that of the associated modes  $j_\nu$ , be preserved since this priority may be significant in determining the subsequent evolution on our modeling scale. In such a situation we cannot predict the outcome definitively with the information available. On the other hand, by accepting the alternatives as equally valid solutions we are able to say that,

*“What happens must be one of these possibilities.”*

(to within the level of approximation corresponding to the usual model uncertainty). An arbitrary selection might provide uniqueness, but lacking a selection principle justifiable from considerations of the unknown rapid behavior we are primarily concerned not to exclude any genuine possibility and so reject such an artificial uniqueness as spurious. This is done in much the same spirit as the acceptance of ‘weak’ or ‘mild’ or ‘generalized’ solutions since at worst these are idealized versions of genuine possibilities and this idealization may not permit us the luxury of restricting our attention to ‘classical solutions.’ We will refer to the times and the situations giving this ambiguity as *anomalous points*. [A related possibility would be a cascade with several transitions  $j \curvearrowright j' \cdots \curvearrowright \bar{j}$  occurring as a sequence on the fast scale; it is always possible, but perhaps inconvenient, to replace this by an equivalent compound single switching event  $j \bowtie \bar{j}$ .]

In view of the above,  $\mathbf{j}(\cdot)$  need not be a ‘function’ on  $[0, T]$  in the usual sense. However, we can think of it simply as a finite *modal sequence* of pairs  $(j, \tau)_\nu \in \mathcal{J} \times \mathbb{R}_+$  with  $\tau_\nu$  the length of the  $\nu$ -th interswitching interval  $\mathcal{I}_\nu$  so  $\sum_\nu \tau_\nu = T$ ; one recovers the switching times as  $t_\nu = \tau_1 + \cdots + \tau_\nu$  and recovers  $\mathbf{j}(t)$ , when  $t$  is not a switching time, by (S)-c. Abusing notation somewhat, we continue to denote these by  $\mathbf{j}(\cdot)$ . We topologize the set  $\mathcal{MS}[0, T]$  of all such modal sequences on  $[0, T]$  as follows:

**Definition 3.1.**  $[\mathbf{j}^m \rightarrow \mathbf{j}]$  in  $\mathcal{MS}[0, T]$  means that each  $j_\nu^m \equiv j_\nu$  for large  $m$  and each  $\tau_\nu^m \rightarrow \tau_\nu$  in  $\mathbb{R}_+$  subject to the constraint  $\sum_\nu \tau_\nu^m = T$ .

Somewhat similarly, a solution  $\mathbf{x}(\cdot)$  would not be a ‘function’ on  $[0, T]$  even if there were no change in state spaces: we retain, at any switching time  $t$ , both values  $\mathbf{x}(t-), \mathbf{x}(t+)$  and, even in contexts with degenerate interswitching intervals, include both when discussing a corresponding *trajectory*  $[[\mathbf{x}]] = \{\mathbf{x}(t) : t \in [0, T]\}$ . We view the switching as occupying time on a more rapid scale – so the transition map  $f$  might represent evolution on that rapid scale – but we make no attempt to include more of the course of this evolution as a connecting part of the trajectory. With this treatment we note, from the continuity of each  $\pi_j$ , that the trajectory  $[[\mathbf{x}]]$  for any solution  $\mathbf{x}(\cdot)$  on any  $[0, T]$  will be compact in  $\cup_j \mathcal{X}_j$ .

On the other hand, there might be a still slower time scale on which the switchings we are here describing become a rapidly repetitive *chattering mode*, averaging as a *sliding mode*, switching infinitely often within a finite period. These situations are certainly important and have been treated extensively (cf., e.g., [3, 11, 1, 8]), but they are not our present concern and, as is essential for our treatment here, we will adopt hypotheses bounding the number of pairs in any modal sequence as above for any bounded period  $[0, T]$ ; compare (S)-b. which forbids Zeno phenomena for feedback solutions.

We turn now to considering the *open loop problem* in which a fixed modal sequence  $\mathbf{j}$  is specified as data. [We continue to use (E), (S), but note that (E)-b.,c. are here irrelevant: effectively we are taking each  $\mathcal{A}_j$  independent of  $y$  in defining ‘admissibility’ of  $\mathbf{j}$ , so  $Y_j$  is not needed.]

**Theorem 1.** *Let an admissible  $\mathbf{j}(\cdot)$  be given as data and suppose suitable initial data  $\xi$  given in  $\mathcal{X}_{j_1}$ . Then there is a unique solution of the open loop problem specified by (E), (S) and this depends continuously, in an appropriate sense, on the specified  $\mathbf{j}$  and  $\xi$ .*

*Proof.* Existence is immediate, recursively constructed uniquely by alternately using (S)-d.,e. starting with  $\mathbf{x}(0) = \xi$  and  $\mathbf{x}(t) = \pi_{j_1}(t, 0, \xi)$  on  $\mathcal{I}_1 = [0, t_2]$ , etc., so we need only verify continuous dependence. Our definition of convergence  $\mathbf{j}^m \rightarrow \mathbf{j}$  means that only the interswitching times  $\tau_\nu^m$  change with  $m$  so, recalling the assumed continuity of the transition maps and dynamical systems involved, the same recursion also shows that  $\mathbf{x}^m(t_\nu^m \pm) \rightarrow \mathbf{x}(t_\nu \pm)$  (even taking  $\xi^m \rightarrow \xi$  and even if some interswitching intervals become degenerate in the limit). We similarly get  $\mathbf{x}^m(t) \rightarrow \mathbf{x}(t)$  for any  $t$  in the interior of an interswitching interval for  $\mathbf{j}$  and assume that any ‘appropriate sense’ for convergence of the solutions will follow from this, e.g., we exclude the use of an  $L^\infty$  topology for solutions.  $\square$

**Remark 3.2.** The statement and proof above are ambiguous as to the total interval but we may think of this as finite  $[0, T]$  and, as usual with  $T$  arbitrary, this also provides the result on  $[0, \infty)$ .

We now set

$$\begin{aligned}\mathcal{MS}^N &= \{\mathbf{j} \in \mathcal{MS}[0, T] : \text{there are at most } N \text{ switches}\}, \\ \mathcal{K}^N(\xi) &= \{[[\mathbf{x}]] : \mathbf{x}(\cdot) \text{ corresponds to } \mathbf{j} \in \mathcal{MS}^N, \mathbf{x}(0) = \xi\}.\end{aligned}$$

It is easy to see that each of the subsets  $\mathcal{MS}^N[0, T]$  will be compact in  $\mathcal{MS}[0, T]$ . We have already noted that each individual trajectory  $[[\mathbf{x}]]$  is compact and, from the discussion of continuous dependence in the proof above, we now see that each  $\mathcal{K}^N(\xi)$  is compact.  $\square$

Still in the setting of the open-loop problem, but now in a context of infinite horizon optimal control, we consider choice of the modal sequence so as to minimize

a cost functional of the form

$$\Psi = \int_0^\infty e^{-\beta t} c_{\mathbf{j}(t)}(\mathbf{x}(t)) dt + \sum_{\nu=2}^{\infty} e^{-\beta t_\nu} c(j_{\nu-1} \curvearrowright j_\nu). \quad (3.1)$$

We wish to show that the inf defining the *value function*

$$V_j(\xi) = \inf\{\Psi[\mathbf{j}, \xi] : \mathbf{j}(0) = j, \mathbf{x}(0) = \xi\} \quad (3.2)$$

is actually an attained minimum.

**Theorem 2.** *Assume that each running cost  $c_j(\cdot) \geq 0$  of (3.1) is continuous; suppose  $j_1 = j$  and suitable initial data  $\xi$  are given in  $\mathcal{J}, \mathcal{X}_j$ . Let each switching cost  $c(j \curvearrowright j') > 0$  and assume there is some  $\mathbf{j}_*$  for which  $\Psi$  is finite. Then there is a modal sequence (switching control)  $\mathbf{j} = \mathbf{j}^*$  for which  $\Psi = \Psi[\mathbf{j}, \xi]$  attains its minimum  $V_j(\xi)$ . This minimum cost depends lower semicontinuously on  $\xi \in \mathcal{X}_j$ .*

*Proof.* The set  $\{\mathbf{j} : \Psi < \infty\}$  is nonempty by assumption so we can consider a minimizing sequence  $\mathbf{j}^m$ :  $\Psi^m = \Psi[\mathbf{j}^m, \xi] \rightarrow \inf\{\Psi\} = V_j(\xi)$ . For arbitrary  $T < \infty$ , the switching costs then ensure a bound on the number of transitions during  $[0, T]$  so we may extract a convergent subsequence; further extracting subsequences we can assume  $\mathbf{j}^m \rightarrow \mathbf{j}^*$  on every bounded interval. Theorem 1 applies to the problem on each  $[0, T]$ , showing the corresponding solutions converge  $\mathbf{x}^m \rightarrow \mathbf{x}^*$  there ‘in a suitable sense.’ From the form of (3.1) we easily see this implies convergence of the restricted costs:

$$\Psi^m \Big|_{[0, T]} \rightarrow \Psi^* \Big|_{[0, T]} \quad \text{so} \quad \Psi^* \Big|_{[0, T]} \leq \Psi^m \Big|_{[0, T]} + \varepsilon \leq \Psi^m + \varepsilon \rightarrow V_j(\xi).$$

Letting  $T \rightarrow \infty$ , this shows that  $\Psi^* \leq V_j(\xi)$  so  $V_j(\xi)$  is a min with minimizer  $\mathbf{j}^*$ . If  $\mathbf{j}^m$  is the minimizer for  $\xi^m \rightarrow \xi$ , then we can extract a convergent subsequence as above to get  $\mathbf{j}^m \rightarrow \mathbf{j}^*$  and see

$$V_j(\xi) \leq \Psi[\mathbf{j}^*, \xi] \leq \liminf_m \Psi[\mathbf{j}^m, \xi^m] = \liminf_m V_j(\xi^m). \quad \square$$

[It is not difficult to see that  $V_j(\cdot)$  is actually continuous if each  $\pi_j$  is locally uniformly continuous.]

## 4. Modeling feedback

Suppose we consider the optimization problem of Theorem 2 for autonomous dynamical systems so autonomy of the system makes the value function  $V$  independent of any starting time and

$$V_j(\xi) = \Psi^* \Big|_{[0, \tau]} + e^{-\beta \tau} V_{\mathbf{j}^*(\tau)}(\mathbf{x}^*(\tau)) \quad (4.1)$$

for each  $\tau > 0$ , where  $\mathbf{j}^*, \mathbf{x}^*$  are optimal as in the proof of Theorem 2. We would like to recover the optimal switching control from  $V$ , allowing for the possibility

that this need not be unique. The possibility of a transition  $j \curvearrowright j' \neq j$  when  $\mathbf{x}(t) = \xi$  just means that  $j'$  is in  $\mathcal{A}_j$  and

$$\begin{aligned} & \left[ \begin{array}{l} \text{some optimal } \mathbf{j} \text{ starting at } (j, \xi) \\ \text{immediately switches } j \curvearrowright j' \neq j \end{array} \right] - \\ & c(j \curvearrowright j') + V_{j'}(\mathbf{f}(\xi; j \curvearrowright j')) = V_j(\xi) \end{aligned} \quad (4.2)$$

where equality just means that the optimal value can be attained with a switch to  $j'$ . On the other hand, comparing with (4.1) and (3.1), we see that

$$\begin{aligned} & \left[ \begin{array}{l} \text{some optimal } \mathbf{j} \text{ starting at } (j, \xi) \\ \text{continues in mode } j \end{array} \right] - \\ & \int_0^\tau e^{\beta t} c_j(\pi_j(t, \xi)) dt + e^{\beta \tau} V_j(\pi_j(\tau, \xi)) = V_j(\xi) \quad (\text{some } \tau > 0). \end{aligned} \quad (4.3)$$

**Remark 4.1.** *From this we observe that:*

Let  $\mathcal{J}, \mathbf{f}, \pi_j$  be as in Theorem 2; assume each  $\pi_j$  is autonomous. Set  $\mathcal{Y}_j = \mathcal{X}_j$ ,  $Y_j = id(\mathcal{X}_j)$ , and

$$\mathcal{A}_j(\xi) = \{j \text{ if (4.3), } j' \text{ if (4.2)}\} \quad (4.4)$$

for  $j \in \mathcal{J}$ ,  $\xi \in \mathcal{X}_j$  to complete the specification **(E)**. Let  $(\mathbf{j}, \mathbf{x})$ , starting with  $(j_1, \xi)$ , be as in Theorem 1.

Then the pair  $(\mathbf{j}, \mathbf{x})$  is optimal for the switching control problem of Theorem 2 if and only if it is a feedback solution as in **(S)**.

*Proof.* Clearly any optimal control satisfies **(S)** with (4.4). Conversely, by connectedness and the continuity of  $\mathbf{y}(\cdot) = \mathbf{x}(\cdot)$ , such a solution of **(S)** satisfies (4.1) on each nondegenerate interswitching interval  $\mathcal{I}_\nu$  (hence) and on  $[0, T]$  by induction on  $\nu$ , hence is optimal.  $\square$

This is a primary motivation for taking **(S)** as defining the general structure of feedback we consider here, while noting, for example, that we cannot always expect to have full-state feedback as in Remark 4.1 and would necessarily implement only finitely many sensors. Thus, we consider the evolution of a solution for **(S)** as an independent problem, with the elements of **(E)** somewhat general. Purely for expository convenience, however, we assume henceforth that  $\mathcal{X}_j, \mathcal{Y}_j, Y_j$  are each independent of  $j \in \mathcal{J}$  and that the dynamical systems  $\pi_j$  are autonomous.

Recall that the sensor maps  $Y_j$  and the resulting sensor output  $\mathbf{y}(\cdot)$  played no role in Theorems 1 and 2, but the regularity to be expected of these is now a significant concern in being able to evaluate  $\mathbf{y}$  pointwise in  $t$  so the conditions of **(S)** make sense. This regularity and its interaction with the avoidance of Zeno behavior – i.e., with **(S)**-b. – constitute the essential technical difficulties in analyzing this feedback structure. For Theorem 1, **(S)**-b. was already an admissibility hypothesis on the given  $\mathbf{j}$  and in the proof of Theorem 2 this was a consequence of the assumed positivity of the switching costs. For a general feedback we will need new hypotheses; we begin by assuming the feedback diagram and sensor map satisfy the following set of hypotheses.

**(H<sub>1</sub>)**

- a. each  $\mathcal{C}^{j \curvearrowright j'}$  is closed in  $\mathcal{X}$  and  $\mathcal{S}^j \supset [\mathcal{Y} \setminus \mathcal{B}^j]$ .
- b.  $Y$  is set-valued with  $Y(\xi)$  finite and nonempty for each  $\xi \in \mathcal{X}$ .
- c.  $Y$  is upper-semicontinuous, i.e., if one has  $y_k \in Y(x_k)$  with  $x_k \rightarrow \bar{x}$  in  $\mathcal{X}$ , then there is a subsequence  $(y_{k(\ell)})$  converging to some  $\bar{y} \in Y(\bar{x})$ .
- d. cascades of the form  $j \curvearrowright j$  are forbidden: i.e., there exists no sequence of pairs  $(j, \xi)_{\nu=1}^{\bar{\nu}}$  with  $j_1 = j_{\bar{\nu}}$  such that  

$$Y(\xi_\nu) \cap \mathcal{C}^{j_\nu \curvearrowright j_{\nu+1}} \neq \emptyset, \quad \xi_{\nu+1} = \mathbf{f}(\xi_\nu; j_\nu \curvearrowright j_{\nu+1}).$$

It is precisely at this point that our considerations will depend in an essential way on the particular PDE setting since we have in mind, at least as an idealization, that our sensors will be point evaluations in the spatial domain of (1.1). For the operation of a thermostat, where (1.1) becomes a heat equation, one has more than enough regularity that this causes no difficulty (provided the sensor location is separated from the furnace/AC). For a transport equation, however, the occurrence of modal switching can be expected to introduce spatial discontinuities which propagate to the sensors and cause temporal discontinuities in  $\mathbf{y}(\cdot)$ ; it then becomes a delicate problem (cf. [5]) to provide a space  $\mathcal{X}$  which allows for this and at the same time gives both continuity of the dynamics and adequate regularity of  $\mathbf{y}(\cdot)$ .

We now provide an additional hypothesis which, along with **(H<sub>1</sub>)**, will suffice to give **(S)-b.** in showing the existence of solutions for the feedback problem. This hypothesis **(H<sub>2</sub>)** is rather technical, but, as an example, we will later show how to verify these hypotheses for transport on a graph.

**(H<sub>2</sub>)**

- There exists  $\bar{\tau} > 0$  such that for each  $\xi \in \mathcal{X}$ ,  $T \geq \bar{\tau}$ , and  $N'$  there exists  $N = N(\xi, N', T)$  such that:  
if  $T - \bar{\tau} < T' < T$  and  $\mathbf{j}|_{[0, T - \bar{\tau}]} \in \mathcal{MS}^{N'}$ , then there are no more than  $N$  points of  $\bar{\Xi} = \overline{\{\xi \in \mathcal{X} : \#Y(\xi) \neq 1\}}$  in the trajectory  $\{\mathbf{x}(t) : t \in [0, T']\}$ .

[While we have formulated this hypothesis to obtain a context of piecewise continuous  $\mathbf{y}(\cdot)$ , one might expect that a rather similar treatment could be formulated for, e.g.,  $\mathbf{y}(\cdot)$  of bounded variation.]

**Theorem 3.** *Assume we have **(E)** satisfying **(H<sub>1</sub>)**, and **(H<sub>2</sub>)**. Then, for any given  $(j, \xi) \in \mathcal{J} \times \mathcal{X}$ , there is  $[\mathbf{j}(\cdot), \mathbf{x}(\cdot), \mathbf{y}(\cdot)]$ , a global solution of the feedback problem starting with  $(j, \xi)$ .*

*Proof.* It is convenient to restrict our attention to ‘skittish solutions,’ which switch whenever that is allowable under the switching rules of **(S)**. By Zorn’s Lemma one has existence of a maximally defined skittish solution  $[\mathbf{j}(\cdot), \mathbf{x}(\cdot), \mathbf{y}(\cdot)]$  whose domain necessarily has one of the forms  $[0, 0]$ ,  $[0, T_*]$ ,  $[0, T_*]$ , or  $\mathbb{R}_+ = [0, \infty)$ ; we wish to show this can only be  $[0, \infty)$ .

From (2.3), we have initially either  $\mathbf{y}(0) \cap \mathcal{B}^j \neq \emptyset$  and proceed with a maximal finite cascade  $j \rightsquigarrow \bar{j}$  or have  $\mathbf{y}(0) = Y(\xi)$  in  $\mathcal{S}^j \setminus \mathcal{B}^j$ . Since **(H<sub>1</sub>)**-a. gives  $\mathcal{S}^j \setminus \mathcal{B}^j = \mathcal{Y} \setminus \mathcal{B}^j$  open and **(H<sub>1</sub>)**-c. ensures a solution can remain for some (small) interswitching interval in mode  $j$ . In the former case, the cascade ends with  $j' \rightsquigarrow \bar{j}$  leaving  $\mathbf{x} = \bar{\xi}$  with  $\mathbf{y} = \bar{\eta} = Y(\bar{\xi}) \notin \mathcal{B}^{\bar{j}}$  (or the cascade could have continued); by (2.3) we then have  $\bar{\eta} \in \mathcal{S}^{\bar{j}}$  and the solution could be extended. In either case, then, the domain  $[0, 0]$  is inconsistent with maximality. Similarly, a domain  $[0, T_*]$  is also inconsistent with maximality since we could restart the problem at  $T_*$  and use the same argument.

Next suppose the maximal domain were of the form  $[0, T_*]$ .

Since  $[\mathbf{j}(\cdot), \mathbf{x}(\cdot), \mathbf{y}(\cdot)]$  is a solution on every subinterval, either there is a last switching  $\cdot \rightsquigarrow j_*$  at  $t_* < T_*$  or the sequence of switching times  $(t_\nu)$  converges to  $T_*$ , violating **(S)**-b. on  $[0, T_*]$  itself. In the former case,  $t \mapsto \mathbf{x}(t) = \pi_{j_*}(t - t_*, \mathbf{x}(t_*+))$  is continuous on  $[0, T_*]$  and either  $\mathbf{y}(T_*) = Y(\mathbf{x}(T_*)) \in \mathcal{S}^{j_*}$  – so the solution continues through  $T_*$  in mode  $j_*$  by **(H<sub>1</sub>)**-a. – or  $\mathbf{y}(T_*) \cap \mathcal{B}^{j_*} \neq \emptyset$  so one can switch and the solution can be extended at least to  $[0, T_*]$ ; either of these possibilities contradicts the maximality of  $[0, T_*]$ .

In the latter case, with  $t_\nu \rightarrow T_*$ , the maximally defined  $\mathbf{j}(\cdot)$  necessarily consists of an infinite sequence of nondegenerate interswitching intervals of length  $\tau_\mu > 0$  (with  $\sum_{\mu=1}^{\infty} \tau_\mu = T_*$ ) separated by maximal cascades  $j_{\mu-1} \rightsquigarrow j_\mu$ . Choose any  $T' \in (T_* - \bar{\tau}, T_*)$ , let  $N'$  bound the number of switchings in  $\mathbf{j}(\cdot)$  on  $[0, T']$ , and set  $N = N(\xi, N', T_*)$  as in **(H<sub>2</sub>)**. Now consider any one of the interswitching intervals  $\mathcal{I} = \mathcal{I}_\mu = [t', t'']$  (i.e.,  $t' = t_{\mu-1}, t'' = t_\mu$ ) with  $t' \geq T'$  on which  $\mathbf{j} \equiv j = j_\mu$ . By **(S)**-c, this must be initiated with  $\mathbf{x}(t'-) = \xi^1$  producing a maximal cascade  $j_{\mu-1} = j^1 \rightsquigarrow \cdots \rightsquigarrow j^n = j$  with  $Y(\xi^\nu) \in \mathcal{C}^{j^\nu \rightsquigarrow j^{\nu+1}}$  and  $\xi^{\nu+1} = \mathbf{f}(\xi^\nu; j^\nu \rightsquigarrow j^{\nu+1})$  for  $\nu = 1, \dots, n-1$  as in **(H<sub>1</sub>)**-d. Assuming no points of  $\Xi$  occur in this sequence (or during  $\mathcal{I}$ ) so  $Y$  is simply a continuous single-valued function there, one can show easily that the set  $S$  of points in  $\mathcal{K}$  which can initiate this particular sequence (as  $\xi^1$ ) is closed and in  $\mathcal{K}^{N'}$ , so compact in  $\mathcal{X}$ . Thus, iterating  $\mathbf{f}$ , the set  $S'$  of points terminating the sequence (as  $\xi^n \xi_+$ ) is also compact and  $S'' = Y(S')$  is compact in  $\mathcal{Y}$  – with  $S'' \cap \mathcal{B}^j = \emptyset$ , as the cascade is maximal. Hence there is a minimal distance from  $S''$  to  $\mathcal{B}^j$ . We must have  $\mathbf{y}(t'') \in \mathcal{B}^j$  to end  $\mathcal{I}$  by initiating another transition and note that  $[t \mapsto Y(\pi_j(t - t', \xi^n_+))]$  is uniformly continuous on  $\mathcal{I}$  so there is a minimal time required to make this transit; with only finitely many possibilities for the cascade, this time  $\tau_*$  may be taken as the same for all so the length  $\tau_\nu$  of such an interswitching interval is bounded below by  $\tau_*$  and there can be at most  $\bar{\tau}/\tau_*$  such intervals. We have no lower bound on the length of those interswitching intervals involving points of  $\Xi$ , but the number of these is bounded by our technical hypothesis **(H<sub>2</sub>)**, contradicting the assumption above of an infinite sequence  $\{\mathcal{I}_\mu\}$ .

Thus, the maximal domain must be  $[0, \infty)$ ; as desired, the maximally defined skittish solution is global. Of course, this need not be unique and there may also be additional (non-skittish) global solutions. Note also, from this proof, that if we

are given, on a bounded domain, any  $[\mathbf{j}(\cdot), \mathbf{x}(\cdot), \mathbf{y}(\cdot)]$  satisfying **(S)** there, then it can be extended to a global solution.  $\square$

**Example 4.2.** As a first example, consider a thermostat-controlled heating system. For the simplest case, one would have a single point-evaluation sensor:  $Y : \xi \mapsto \eta = \xi(p)$  with  $p$  given in the spatial region  $\Omega$  and  $\xi \in \mathcal{X} = C(\Omega)$ . We take the effect of the control in the boundary flux so (1.1) becomes the heat equation for the temperature distribution  $\mathbf{x}(t, \cdot)$

$$\mathbf{x}_t = \Delta \mathbf{x} \text{ on } \Omega, \quad \mathbf{x}_\nu = \alpha \mathbf{x} + v_j \text{ at } \partial\Omega \quad (4.5)$$

defining  $\pi_j$  for the two modes  $j \in \mathcal{J} = \{0, 1\}$  denoting OFF/ON. [Here the flux difference  $v_1 - v_0$  gives the effect of the furnace or AC.] We have no jumps in the state itself when the thermostat switches so  $\mathbf{f} = id_{\mathcal{X}}$ . The well-posedness of (4.5) is standard and **(H<sub>1</sub>)**-c.,d. as well as **(H<sub>2</sub>)** are immediate since  $Y$  is single-valued and continuous.

Now let  $\eta_*$  be our setpoint, the desired temperature, and allow a margin  $\pm\delta$  with  $\delta > 0$ . Then switching is determined by

$$\begin{aligned} \mathcal{A}_0(y) &= \begin{cases} \{0\} & \text{if } y > \eta_* - \delta \\ \{0, 1\} & \text{if } y = \eta_* - \delta \\ \{1\} & \text{if } y < \eta_* - \delta \end{cases} \quad \mathcal{A}_1(y) = \begin{cases} \{0\} & \text{if } y > \eta_* + \delta \\ \{0, 1\} & \text{if } y = \eta_* + \delta \\ \{1\} & \text{if } y < \eta_* + \delta \end{cases} \\ \text{so} \quad \mathcal{C}^{0\sim 1} &= (-\infty, \eta_* - \delta], \quad \mathcal{S}^0 = [\eta_* - \delta, \infty), \\ &\quad \mathcal{C}^{1\sim 0} = [\eta_* + \delta, \infty), \quad \mathcal{S}^1 = (-\infty, \eta_* + \delta]. \end{aligned}$$

I.e., the furnace turns ON when temperature (at the thermostat) falls below  $\eta_* - \delta$  and goes OFF when it rises above  $\eta_* + \delta$ . [The resulting transducer:  $y(\cdot) \mapsto \mathbf{j}(\cdot)$  is precisely the hysteretic *non-ideal relay* of [6, section 28.2], well defined except for the possible ambiguity of anomalous points.] We have **(H<sub>1</sub>)**-a.,c.,d. trivially; with  $\delta > 0$ , **(H<sub>1</sub>)**-b. holds as  $\mathcal{C}^{0\sim 1} \cap \mathcal{C}^{1\sim 0} = \emptyset$ , and **(H<sub>1</sub>)**-e. holds as  $y(\cdot)$  is continuous here with each  $\mathcal{S}^j \setminus \mathcal{B}^j$  open.

Taking  $\delta > 0$  is implicit in the usual design of thermostats and we note that our hypotheses fail for the idealized thermostat with  $\delta = 0$ . In that setting one has a (pointwise) functional map:  $y \mapsto j$  and convexifying when  $y = \eta_*$  (compare [4, 3]) one does obtain existence, although with the possibility of Zeno-ness in the form of sliding modes: ON/OFF oscillation of the furnace on the rapid scale.  $\square$

**Example 4.3.** We conclude with a more demanding example, considering transport on a graph with feedback modal control: descriptively, we imagine reacting chemical species being transported by a solvent, moving as plug flow along the pipe segments  $\{E_m : m \in \mathcal{M}\}$  of a network. These single-segment problems are then coupled at each node  $N_n$  of the resulting graph  $\Gamma$  through the allocation of incoming flux, including exogenous sources, to outgoing segments, including external outputs). Our presentation here largely follows the more detailed treatment in [5].

The state  $\mathbf{x}(t)$  in this example will be the densities (concentrations)  $u(t, \cdot)$  of conserved species of interest, taken in a suitable state space  $\mathcal{X}$  of vector functions

on  $\Gamma$ . The map  $Y : \mathcal{X} \rightarrow \mathcal{Y} = \mathbb{R}^K$  is given by evaluations  $y_k = u_{i(k)}(\cdot, \bar{s}_k)$  at specified sensor points  $\bar{s}_k \in \Gamma$ . We have  $\mathbf{f} = id_{\mathcal{X}}$  here and initially let the feedback diagram be subject only to (2.3), **(H<sub>1</sub>)**-a., and  $\mathcal{C}^{j \curvearrowright j'} \cap \dots \cap \mathcal{C}^{j'' \curvearrowright j} = \emptyset$ , which is here equivalent to **(H<sub>1</sub>)**-b.

For simplicity of exposition we assume an incompressible carrier (solvent) and uniform cross-sectional area  $\alpha_m$  in each pipe segment  $E_m$  and input end  $0_m$  independent of the mode; the transport is produced by the action (specified by  $j$ ) of a pump at  $0_m$ . The flow velocity  $v_m^j$  will then be constant on  $E_m$  and, again for simplicity we assume  $v_m^j$  is also constant in  $t$ . The evolution  $\pi_j$  of the system is now determined by these flow velocities. First, we have a set of convection/reaction equations: on each of the individual edges

$$u_t + v_m^j u_s = f(u) \quad \text{on } E_m \quad (4.6)$$

and will use the classical *method of characteristics* to construct solutions:

Let  $\omega(t; \omega_0)$  be the solution of the ordinary differential equation

$$\omega' = f(\omega), \quad \omega(0) = \omega_0, \quad (4.7)$$

Given  $(t, s)$ , track back along the characteristic  $\sigma(\tau) = s - [t - \tau]v$  to an initialization point  $(\tau_0, \sigma_0)$  – either  $\tau = \tau_0 \leq t$  is a starting time (i.e., 0 or the most recent switching time) with  $\sigma_0 = \sigma(\tau_0) \in E_m$  or else  $\sigma_0 = 0_m$  with  $\tau_0 \leq \tau \leq t$ . Now set  $u(t, s) = \omega(t - \tau; \omega_*)$  where  $\omega_*$  is the given data at  $(\tau_0, \sigma_0)$ .

The construction of  $\pi_j$  is then completed by the nodal coupling, specifying the input data  $u_{*m}(\cdot)$  to each pipe. For each node  $N_n$  we have input edges  $\mathcal{M}_n^+$  and output edges  $\mathcal{M}_n^-$  (with  $\cup_n \mathcal{M}_n^- = \mathcal{M} = \cup_n \mathcal{M}_n^+$ ). Clearly the assigned flow velocities must satisfy the consistency condition

$$[\text{flux in}]_n^j = \sum_{m \in \mathcal{M}_n^-} \alpha_m v_m^j = \sum_{m \in \mathcal{M}_n^+} \alpha_m v_m^j = [\text{flux out}]_n^j = \Phi_n^j \quad (4.8)$$

Assuming perfect mixing at the node, the vector of combined input concentrations at  $N_n$  of the chemical species will be

$$U_n(\tau) = \frac{\sum \{\alpha_m v_m^j u_m(\tau, 1_m) : m \in \mathcal{M}_n^-\}}{\sum \{\alpha_m v_m^j : m \in \mathcal{M}_n^-\}} \quad (4.9)$$

and the required input data to  $E_m$  is then given by

$$u(\tau, 0_m) = u_*(\tau) = U_n(\tau) \quad \text{for } m \in \mathcal{M}_n^+ \quad (4.10)$$

[This must be modified in the case of exogenous sources, for which one can permit some choice in the formulation.]

We take this construction along characteristics as defining our notion of solution for (4.6) and so the definition of  $\pi_j$ .

Our major technical difficulty in this example is to specify and topologize the state space  $\mathcal{X}$  so as to verify the hypotheses **(H<sub>1</sub>)** and **(H<sub>2</sub>)** while maintaining

continuity of this  $\pi_j$ . From our solution construction and the continuity of  $\omega(\cdot, \cdot)$ , we see that discontinuities will propagate along the characteristics (including across nodes) and can be created only at nodes at switching times. We expect, then, that the state  $\mathbf{x}(t) = u(t, \cdot)$  will be a piecewise continuous function and will take  $\mathcal{X} = \mathcal{PC}$  to be a suitable space of such functions.

As with modal sequences  $\mathbf{j}(\cdot)$ , it is possible to create degenerate ‘intervals of continuity’ – allowing  $u(t, \cdot)$  to be continuous on an interval  $[s, s']$ , take a value on the degenerate interval  $[s', s'']$  with  $s'' = s'$ , and then again continuous on  $[s'', s''']$ . This could occur if discontinuities propagating through edges  $E_m$  and  $E_{m'}$  incoming to the same node  $N_n$  arrive simultaneously; in view of our modeling considerations we interpret ‘simultaneously’ as meaning ‘indistinguishably close’ – although possibly distinct on the rapid time scale so we retain both possibilities with the alternative intermediate values. These degenerate intervals correspond to a fine spatial scale, comparable to the rapid time scale. In view of this possibility we must careful with the interpretation of the sensor map  $Y$ , taking this to be set-valued when such a subinterval coincides with one of the sensor points  $\bar{s}_k$ .

This suggests our characterization of an element of  $\mathcal{X} = \mathcal{PC}$ : for each  $m$  one has a vector-valued piecewise continuous functions on closed subintervals, including possible finite sequences of degenerate subintervals as with  $\mathcal{MS}$  and then, much as with Definition 3.1, we topologize this as follows:

**Definition 4.4.**  $[u^k \rightarrow u]$  in  $\mathcal{PC}$  if, for each  $E_m$ , the number of subintervals is eventually fixed, the dividing endpoints converge, and the functions on them (normalized to domain  $[0, 1]$  with values on degenerate subintervals taken as constants) converge in the sense of  $C[0, 1]$ .

One easily sees that the problem is well posed in this setting:  $\pi_j$  is continuous from  $\mathbb{R}_+ \times \mathcal{PC}$  to  $\mathcal{PC}$ . As suggested earlier, we use point evaluations to define  $Y : \mathcal{PC} \rightarrow \mathcal{Y} = \mathbb{R}^K$  by

$$Y(\xi) = [\xi_{i(1)}(\bar{s}_1), \dots, \xi_{i(K)}(\bar{s}_K)] \quad (\xi \in \mathcal{PC}) \quad (4.11)$$

with the provision that: if a discontinuity of  $\xi$  occurs at one of the sensor points  $\bar{s}_k$  so  $\xi(\cdot)$  has both left- and right-hand values there (perhaps even more values if this involves a degenerate subinterval), then  $y_k(\xi)$  becomes the set of all relevant values; this clearly gives **(H<sub>1</sub>)**-b.,c. It is not difficult to construct the feedback diagram to give **(H<sub>1</sub>)**-a.,d. – e.g., taking  $\mathcal{S}^j$  to contain the open set  $\mathcal{Y} \setminus \mathcal{B}^j$ , perhaps adjoining (as anomalous points) other points  $\eta \in \mathcal{B}^j$  for which  $Y^{-1}(\eta)$  contains  $\xi$  from which one might wish to extend the solution in mode  $j$  – and we assume this.

In order to satisfy **(H<sub>2</sub>)** we require that the sensor points are separated from the actuators, i.e., from the input nodes where discontinuities might be created. Thus, we will assume there is some  $\bar{\tau} > 0$  such that

$$[\bar{s}_k - 0_m] / v_m^j \geq \bar{\tau} \quad \text{for all } j \in \mathcal{J}, \bar{s}_k \in E_m, k = 1, \dots, K. \quad (4.12)$$

With this assumption, no discontinuity created after time  $t = T$  could possibly be propagated along characteristics to arrive at any sensor before  $t = T + \bar{\tau}$ . If

we have bounded by  $N'$  the number of switchings in  $\mathbf{j}(\cdot)$  up to  $T_* - \bar{\tau}$ , then our dynamics and the graph geometry bound both the number of spatial discontinuities arriving to any sensor point, creating a point of  $\Xi$ , up to  $T_* - \bar{\tau}$  and the number of discontinuities in  $\mathbf{x}(T_* - \bar{\tau})$ , viewed now as an ‘initial’ state, and so bounds the number which can arrive to a sensor point, creating a new point of  $\Xi$ , by any time  $T' < T_*$ ). This total bound is then  $N(\xi, N', T)$  and we have verified **(H<sub>2</sub>)**.  $\square$

### Acknowledgment

Many thanks to G. Leugering for his support in this work.

## References

- [1] M. di Bernardo, C.J. Budd, A.R. Champneys, and P. Kowalczyk, *Piecewise-smooth Dynamical Systems*, (AMS #163) Springer-Verlag, London (2008).
- [2] I. Capuzzo-Dolcetta and L.C. Evans, *Optimal switching for ordinary differential equations*, SIAM J. Cont. Opt. **22**, 1984, pp. 143–161.
- [3] A.F. Filippov, *Differential Equations with Discontinuous Right-hand Sides*, Nauka, Moscow (1985) [*transl.* Kluwer, Dordrecht (1988)].
- [4] K. Glashoff and J. Sprekels, *An application of Glicksberg’s theorem to set-valued integral equations arising in the theory of thermostats*, SIAM J. Math. Anal. Appl. **12**, 1981, pp. 477–486.
- [5] F. Hante, G. Leugering, and T.I. Seidman, *Modeling and analysis of modal switching in networked transport systems*, Appl. Math. and Optimization, to appear.
- [6] M.A. Krasnosel’skiĭ and A.V. Pokrovskii, *Systems with Hysteresis*, Nauka, Moscow (1983) [*transl.* Springer-Verlag, Berlin (1989)].
- [7] T.I. Seidman, *Switching systems: thermostats and periodicity* (report **MRR-83-07**), UMBC, 1983. ([http://userpages.umbc.edu/~seidman/ss\\_83.pdf](http://userpages.umbc.edu/~seidman/ss_83.pdf))  
[modified, as *Switching systems, I*, Control and Cybernetics **19**, 1990, pp. 63–92.]
- [8] T.I. Seidman, *The residue of model reduction*, in *Hybrid Systems III. Verification and Control*, (LNCS #1066; R. Alur, T.A. Henzinger, E.D. Sontag, eds.), Springer-Verlag, Berlin (1996), pp. 201–207.
- [9] T.I. Seidman, *A convection/reaction/switching system*, Nonlinear Anal. – TMA **67**, pp. 2060–2071, 2007.
- [10] T.I. Seidman, *Aspects of modeling with discontinuities*, Int’l. J. Evolution Eqns. **3**, 2008, pp. 129–143.
- [11] V.I. Utkin, *Sliding Modes and their Application in Variable Structure Systems*, Mir, Moscow, 1978.

Thomas I. Seidman  
 Department of Mathematics and Statistics  
 University of Maryland Baltimore County  
 Baltimore, MD 21250, USA  
 e-mail: [seidman@math.umbc.edu](mailto:seidman@math.umbc.edu)

# Optimization Problems for Thin Elastic Structures

Jürgen Sprekels and Dan Tiba

**Abstract.** We discuss shape optimization problems and variational methods for fundamental mechanical structures like beams, plates, arches, curved rods, and shells.

**Mathematics Subject Classification (2000).** Primary 49Q10; Secondary 35J85.

**Keywords.** Kirchhoff–Love arches, Naghdi curved rods and shells, the control variational method.

## 1. Introduction

This work is a survey on recent results concerning thickness and shape optimization problems associated with linear elasticity models of thin bodies like beams, plates, arches, curved rods, and shells. We investigate subjects like existence, uniqueness, optimality conditions, approximation, and numerical experiments. Our approach is also strongly related to the so-called *control variational method*. In the case of Kirchhoff–Love arches, this approach even yields the explicit solution of the model.

The plan of the paper is as follows: in Section 2, simplified beam and plate models are analyzed, and the control variational method is briefly introduced. Section 3 is devoted to Kirchhoff–Love arches and their optimization. In Sections 4 and 5, generalized Naghdi models for curved rods and shells, respectively, are described together with optimization and control variational methods.

Finally, we mention that a general background and complete explanations of much of the presented material can be found in the recent monograph by Neittaanmäki, Sprekels, and Tiba [11]. Further references of interest will be indicated throughout the text.

---

The second author acknowledges the support of CNCSIS Romania under grant PCE 1192-09.

## 2. Beams and plates

We start with a thickness optimization problem for a simplified model of a simply supported plate. For a given open set  $\Omega$  having a sufficiently smooth boundary  $\partial\Omega$ , we consider the problem:

$$\text{Min} \left\{ \int_{\Omega} u(x) dx \right\}, \quad (2.1)$$

subject to

$$\Delta(u^3 \Delta y) = f \quad \text{in } \Omega, \quad (2.2)$$

$$y = \Delta y = 0 \quad \text{on } \partial\Omega, \quad (2.3)$$

$$0 < m \leq u(x) \leq M \quad \text{a.e. in } \Omega, \quad (2.4)$$

$$y \in C, \quad (2.5)$$

where  $C \subset L^2(\Omega)$  is nonempty and closed. The dimension of  $\Omega$  is arbitrary, with the plate model corresponding to  $\Omega \subset \mathbb{R}^2$  (and the beam model to  $\Omega \subset \mathbb{R}$ ). Here,  $u \in L^\infty(\Omega)_+$  is the thickness,  $f \in L^2(\Omega)$  is the load, and  $y \in H^2(\Omega) \cap H_0^1(\Omega)$  (the weak solution of (2.2), (2.3)) represents the deflection.

Clearly, (2.2), (2.3) may be rewritten as

$$\Delta z = f \quad \text{in } \Omega, \quad (2.6)$$

$$z = 0 \quad \text{on } \partial\Omega, \quad (2.7)$$

$$\Delta y = z \ell \quad \text{in } \Omega, \quad (2.8)$$

$$y = 0 \quad \text{on } \partial\Omega, \quad (2.9)$$

where  $z \in H^2(\Omega) \cap H_0^1(\Omega)$  is completely determined by  $f$  and  $\ell = u^{-3} \in L^\infty(\Omega)_+$ .

The system (2.6)–(2.9) looks like the optimality conditions of some optimal control problem (i.e., state equation plus adjoint equation) from the point of view of the differential operators.

We formulate such a distributed control problem:

$$\text{Min} \left\{ \frac{1}{2} \int_{\Omega} \ell(x) h^2(x) dx \right\} \quad (2.10)$$

subject to

$$\Delta y = \ell z + \ell h \quad \text{in } \Omega, \quad (2.11)$$

$$y = 0 \quad \text{on } \partial\Omega. \quad (2.12)$$

The control  $h$  belongs to  $L^2(\Omega)$ , and no constraints are imposed;  $z$  is defined by (2.6), (2.7),  $\ell = u^{-3}$ .

The optimal control problem (2.10)–(2.12) admits the trivial solution  $h^* \equiv 0$  in  $\Omega$  (which is unique), and obviously the optimal state  $y^*$  (and  $z$ ) satisfy (2.6)–(2.9) and, consequently, (2.2), (2.3). Moreover, one can see directly that (2.10)–(2.12) is equivalent to the minimization of the usual energy functional associated with (2.2), (2.3). By (2.11),  $h = \ell^{-1} \Delta y - z$ , and we can rewrite (2.10) as

$$\begin{aligned} & \underset{y \in H^2(\Omega) \cap H_0^1(\Omega)}{\text{Min}} \left\{ \frac{1}{2} \int_{\Omega} \frac{1}{\ell(x)} (\Delta y(x) - \ell(x) z(x))^2 dx \right\} \\ &= \underset{y \in H^2(\Omega) \cap H_0^1(\Omega)}{\text{Min}} \left\{ \frac{1}{2} \int_{\Omega} u^3(x) (\Delta y(x))^2 dx - \int_{\Omega} y(x) f(x) dx \right\} \\ & \quad + \frac{1}{2} \int_{\Omega} \ell(x) z(x)^2 dx, \end{aligned}$$

where the last integral does not depend on  $y$ .

*Remark.* This is probably the simplest example that shows that the classical variational method for differential equations may be reformulated as a control problem. If state constraints are added in (2.10)–(2.12), we get a variational inequality for (2.2), (2.3), and the control problem is no longer trivial. In other situations to be mentioned later, such reformulations have a major impact both at the theoretical and the numerical levels.

Coming back to the thickness optimization problem (2.1)–(2.5), the above transformations allow us to reformulate it as follows:

$$\underset{\Omega}{\text{Min}} \left\{ \int_{\Omega} \ell^{-\frac{1}{3}}(x) dx \right\}, \quad (2.13)$$

$$\Delta y = z \ell \quad \text{in } \Omega, \quad (2.14)$$

$$y = 0 \quad \text{on } \partial\Omega, \quad (2.15)$$

$$0 < M^{-3} \leq \ell(x) \leq m^{-3} \quad \text{a.e. in } \Omega, \quad (2.16)$$

$$y \in C. \quad (2.17)$$

The integrand in (2.13) is strictly convex in the interval  $[M^{-3}, m^{-3}]$ . We get the following result.

**Theorem 2.1.** *If  $C$  is convex and if the admissible set is nonvoid, then the problem (2.13)–(2.17) and, consequently, the problem (2.1)–(2.5), has a unique global optimal pair  $[y^*, \ell^*]$ , respectively  $[y^*, u^*]$ , in  $H^2(\Omega) \times L^\infty(\Omega)$ .*

*Remark.* Existence is standard, and uniqueness is a consequence of the strict convexity of (2.13)–(2.17), since the relation  $y \leftrightarrow \ell$  defined by (2.14), (2.15) is linear. However, the original problem (2.1)–(2.5) is nonconvex, since the relation  $y \leftrightarrow u$  in (2.2), (2.3) is strongly nonlinear. Hence, (2.1)–(2.5) may have infinitely many

local minimum points (pairs), but the global minimum is unique. Such results may be extended to clamped plates, i.e., to the case that (2.3) is replaced by

$$y = \frac{\partial y}{\partial n} = 0 \quad \text{on } \partial\Omega,$$

and to other models of plates and beams.

*Remark.* If  $M$  is “large”, and  $C = \{y \in H^2(\Omega); y(x) \geq -\delta \text{ a.e. in } \Omega\}$ ,  $\delta > 0$  “small”, then the set of admissible pairs is nonvoid. Indeed, for  $\ell = M^{-3}$  the corresponding solution  $y_M$  of (2.14) satisfies (for  $p$  “large”):

$$\sup_{x \in \Omega} |y_M(x)| \leq |y_M|_{W^{2,p}(\Omega)} \leq C|z|_{L^p(\Omega)} M^{-3} \rightarrow 0 \quad \text{for } M \rightarrow \infty.$$

Then, Theorem 2.1 may be applied. Existence and uniqueness results are essential stability conditions for numerical methods.

If the state constraint in (2.17) has the above form, then a usual penalization approach approximates (2.13)–(2.17) via

$$\text{Min} \left\{ \int_{\Omega} \ell^{\frac{1}{3}}(x) dx + \frac{1}{2\varepsilon} \int_{\Omega} [y(x) + \delta]_-^2 dx \right\} \quad (2.18)$$

subject to (2.14)–(2.16), with  $\varepsilon > 0$  “small”.

**Proposition 2.2.** *The optimality conditions for the problem (2.18) are given by (2.14), (2.15), the adjoint equation*

$$\Delta p_{\varepsilon} = -\frac{1}{\varepsilon}(y_{\varepsilon} + \delta)_- \quad \text{in } \Omega, \quad (2.19)$$

$$p_{\varepsilon} = 0 \quad \text{on } \partial\Omega, \quad (2.20)$$

and the maximum principle

$$0 \leq \int_{\Omega} \left( -\frac{1}{3}(\ell_{\varepsilon}(x))^{-\frac{4}{3}} + p_{\varepsilon}(x) z(x) \right) (\ell(x) - \ell_{\varepsilon}(x)) dx \quad (2.21)$$

for any  $\ell \in L^{\infty}$  such that  $M^{-3} \leq \ell(x) \leq m^{-3}$  a.e. in  $\Omega$ .

Here,  $[y_{\varepsilon}, \ell_{\varepsilon}]$  is the unique optimal pair of the penalized control problem (2.18), (2.14)–(2.16).

*Remark.* One can prove bang-bang properties for the optimal thickness, [12]. Here, we just show this in a numerical example computed by gradient methods with projection. Notice that the gradient is given by  $-\frac{1}{3}\ell^{-\frac{4}{3}} + p z$ .

*Example 2.3.* Let  $\Omega \subset \mathbb{R}^2$  be an ellipse with semiaxes 0.2 and 1.0. The initial “thickness” is  $\ell_1 \equiv 0.3$  in  $\Omega$ , and the constraint on  $\ell$  is

$$0.1 \leq \ell \leq 19.9 \quad \text{in } \Omega,$$

while  $\delta = 0.1$ . The Figures 2.1, 2.2 a), b) represent an optimal  $\ell$  (which is almost bang-bang in Figure 2.2) together with its section along  $x_2 = 0$  for  $f \equiv -1000$  in  $\Omega$ , and, respectively, for

$$f(x_1, x_2) = \begin{cases} 200 & \text{if } x_1 \geq 0, \\ -2500 & \text{if } x_1 < 0. \end{cases}$$

*Remark.* Relevant for this section are the papers Sprekels and Tiba [12], [13], [14], and Arnăutu, Langmach, Sprekels, and Tiba [1].

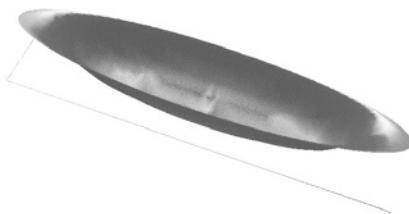


FIGURE 2.1. a)

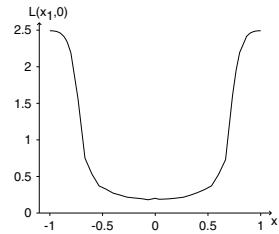


FIGURE 2.1. b)

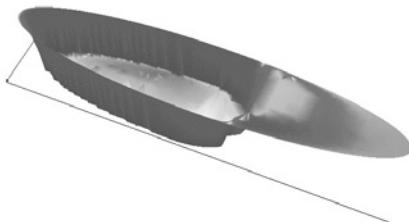


FIGURE 2.2. a)

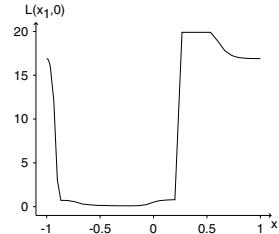


FIGURE 2.2. b)

### 3. Kirchhoff–Love arches

We recall the classical formulation of the model:

$$\begin{aligned} & \int_0^1 \left[ \frac{1}{\varepsilon} (v'_1 - c v_2)(u'_1 - c u_2)(s) + (v'_2 + c v_1)'(u'_2 + c u_1)'(s) \right] ds \\ &= \int_0^1 (f_1 u_1 + f_2 u_2)(s) ds, \quad \forall u_1 \in H_0^1(0, 1), \quad \forall u_2 \in H_0^2(0, 1). \quad (3.1) \end{aligned}$$

The arch is given by a two-dimensional Jordan curve parametrized with respect to the arc length by  $[\varphi_1, \varphi_2]$ , and  $\sqrt{\varepsilon}$  is proportional to the thickness of the arch (assumed constant),  $[f_1, f_2] \in L^2(0, 1)^2$  are, respectively, the tangential and the normal component of the load (assumed to act in the arch plane), while  $[v_1, v_2]$  is a similar representation of the deformation. The coefficient  $c = \varphi_2'' \varphi_1' - \varphi_1'' \varphi_2'$  is the curvature.

A thorough investigation using Dirichlet's principle, Korn's inequality, and the Lax–Milgram lemma, for the solvability of (3.1), may be found in Ciarlet [3], p. 432.

We shall analyze shape optimization problems associated with (3.1), while the thickness is assumed constant. An essential tool is the reformulation of (3.1) via the control variational method.

Let  $\theta(s) = \arctan\left(\frac{\varphi_2'(s)}{\varphi_1'(s)}\right)$  (then  $\theta' = c$ ) denote the angle between the horizontal coordinate axis and the tangent vector  $\varphi'(s) = (\varphi_1'(s), \varphi_2'(s))$ , and let

$$W(s) = \begin{pmatrix} \cos(\theta(s)) & \sin(\theta(s)) \\ -\sin(\theta(s)) & \cos(\theta(s)) \end{pmatrix} \quad (3.2)$$

denote the fundamental matrix of the differential system

$$q_1'(s) = c(s) q_2(s), \quad q_2'(s) = -c(s) q_1(s), \quad s \in [0, 1].$$

The control variational method associates with (3.1), (3.2) the optimal control problem

$$\text{Min} \left\{ L(u, z) = \frac{1}{2\varepsilon} \int_0^1 u^2(s) ds + \frac{1}{2} \int_0^1 z'(s)^2 ds \right\}, \quad (3.3)$$

subject to  $u \in L^2(0, 1)$ ,  $z \in H_0^1(0, 1)$ , to the state equation

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}(t) = \int_0^t W(t) W^{-1}(s) \begin{bmatrix} u(s) + g_1(s) \\ z(s) + g_2(s) \end{bmatrix} ds, \quad t \in [0, 1], \quad (3.4)$$

and to the constraint

$$\int_0^1 W^{-1}(s) \begin{bmatrix} u(s) + g_1(s) \\ z(s) + g_2(s) \end{bmatrix} ds = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \quad (3.5)$$

The formulation (3.3)–(3.5) is meaningful for  $\theta \in L^\infty(0, 1)$ , i.e., for  $\varphi = [\varphi_1, \varphi_2]$  just Lipschitzian. If  $\theta \in W^{1,1}(0, 1)$ , then (3.4) may be rewritten in differential form, i.e.,

$$\begin{aligned} v_1' - cv_2 &= u + g_1 && \text{a.e. in } (0, 1), \\ v_2' + cv_1 &= z + g_2 && \text{a.e. in } (0, 1). \end{aligned} \quad (3.6)$$

The affine terms  $g_1, g_2$  are defined by  $g_1 = \varepsilon_2 \ell$ ,  $-g_2'' = h$ ,  $g_2(0) = g_2(1) = 0$ , where

$$\begin{bmatrix} \ell \\ h \end{bmatrix}(t) = - \int_0^t W(t)W^{-1}(s) \begin{bmatrix} f_1(s) \\ f_2(s) \end{bmatrix}(s) ds. \quad (3.7)$$

A simple computation, involving (3.6), (3.7), and some partial integration, shows that the cost functional (3.3) represents exactly the energy of the original system (3.1). From this point of view, the control variational method (applied to arches by Sprekels and Tiba [15]) is very similar to the approach of Ciarlet and Ciarlet [5], [6], Gratie [8], Ciarlet and Gratie [7]. Using optimal control methods in the minimization of the energy allows the application of powerful tools like the adjoint equation and Pontryagin's maximum principle.

Restriction (3.5) expresses that  $v_1(1) = v_2(1) = 0$  (multiplying by  $W(1)$ ) and completes the initial condition  $v_1(0) = v_2(0) = 0$ , which is a consequence of (3.4). The constraint (3.5) makes the control problem (3.3)–(3.5) nontrivial, in contrast to (2.10)–(2.12) (which has the solution  $h = 0$ ).

As we have already mentioned, the following result reduces the solution to (3.1) to the variational problem (3.3)–(3.5).

**Theorem 3.1.** *If  $\varphi \in W^{3,\infty}(0,1)^2$ , then the unique optimal state  $[v_1^*, v_2^*]$  of the control problem (3.3)–(3.5) is the solution of (3.1).*

*Remark.* The regularity assumption  $\varphi \in W^{3,\infty}(0,1)^2$  is standard in the mathematical literature on the subject. We study and solve (3.3)–(3.5) under the much weaker hypothesis  $\varphi \in W^{1,\infty}(0,1)^2$ , which is one of the gains of the control variational method.

**Theorem 3.2.** *Suppose that  $\theta \in L^\infty(0,1)$ . The couple  $\{[u^*, z^*], [v_1^*, v_2^*]\}$  is optimal for (3.3)–(3.5) if and only if there are  $\lambda_1^*, \lambda_2^* \in \mathbb{R}$  and  $p^*, q^* \in L^\infty(0,1)$  such that:*

$$\begin{aligned} \begin{bmatrix} v_1^* \\ v_2^* \end{bmatrix}(t) &= \int_0^t W(t)W^{-1}(s) \begin{bmatrix} u^*(s) + g_1(s) \\ z^*(s) + g_2(s) \end{bmatrix} ds, \quad \text{for a.e. } t \in (0,1), \\ \int_0^1 W^{-1}(s) \begin{bmatrix} u^*(s) + g_1(s) \\ z^*(s) + g_2(s) \end{bmatrix} ds &= \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \\ \begin{bmatrix} p^* \\ q^* \end{bmatrix}(t) &= W(t) \begin{bmatrix} \lambda_1^* \\ \lambda_2^* \end{bmatrix}, \quad \text{for a.e. } t \in (0,1), \\ u^* &= \varepsilon p^*, \quad (z^*)'' = -q^*, \quad z^*(0) = z^*(1) = 0. \end{aligned}$$

*Remark.* Theorem 3.2 gives the characterization of the solution to (3.3)–(3.5) via the first-order optimality conditions, which are a variant of the Pontryagin maximum principle. They allow, via further arguments, the explicit solution of

(3.3)–(3.5). In particular,  $[\lambda_1^*, \lambda_2^*] \in \mathbb{R}^2$  is the unique minimizer of the dual problem

$$\begin{aligned} \text{Min}_{[\lambda_1, \lambda_2] \in \mathbb{R}^2} & \left\{ \frac{1}{2\varepsilon} \int_0^1 (\lambda_1 \varepsilon \cos(\theta(s)) + \lambda_2 \varepsilon \sin(\theta(s)) + \varepsilon \ell(s))^2 dx \right. \\ & \left. + \frac{1}{2} \int_0^1 [(\lambda_1 w_1 + \lambda_2 w_2 + g_2)'(s)]^2 ds \right\}, \end{aligned} \quad (3.8)$$

where  $w_1, w_2 \in H^2(0, 1) \cap H_0^1(0, 1)$  are some auxiliary mappings defined by

$$w_1''(s) = \sin(\theta(s)), \quad w_2''(s) = -\cos(\theta(s)). \quad (3.9)$$

*Remark.* This is possible since the constraint (3.5) has finite-dimensional range. The dual problem (3.8) is unconstrained and equivalent to its first-order optimality system (a  $(2 \times 2)$  linear algebraic system). This gives the explicit solution for  $\lambda_1^*, \lambda_2^*$  and subsequently, one can compute  $p^*, q^*$ , and  $u^*, z^*$  by the formulas in Theorem 3.2. The state equation (3.4) gives the explicit computation of the deformation (only some integrals have to be approximated by quadrature formulas).

Some examples of deformation computations using (3.8), (3.9), and Theorem 3.2 are given in Figures 3.1–3.3. Notice that in the case of Gothic arches such as in Figure 3.1,  $\theta$  has indeed jumps, that is, the hypothesis  $\theta \in L^\infty(0, 1)$  is essential.

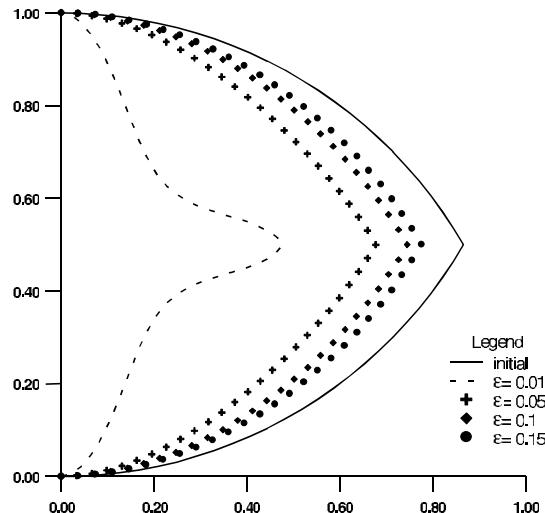


FIGURE 3.1.  $\theta(t) = t$ ,  $t \in [0, \frac{\pi}{3}]$ ,  $\theta(t) = t + \frac{\pi}{3}$ ,  $t \in [\frac{\pi}{3}, \frac{2\pi}{3}]$ ,  
 $f_1(t) = 0$ ,  $f_2(t) = \frac{1}{SE}$ ,  $E = 10$ , © SIAM.

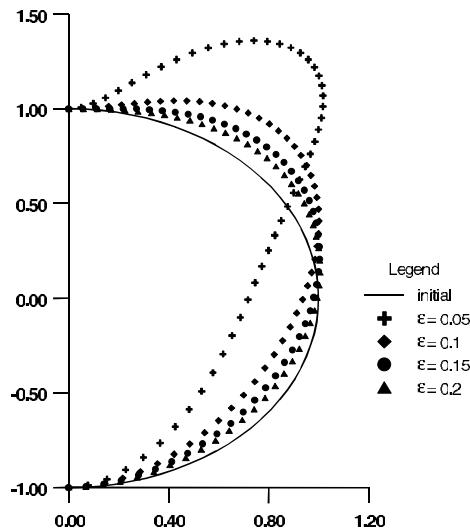


FIGURE 3.2.  $\theta(t) = t$ ,  $t \in [0, \pi]$ ,  $f_1(t) = \frac{\sin(t)}{S}$ ,  $f_2(t) = \frac{\cos(t)}{S}$ ,  
 © SIAM.

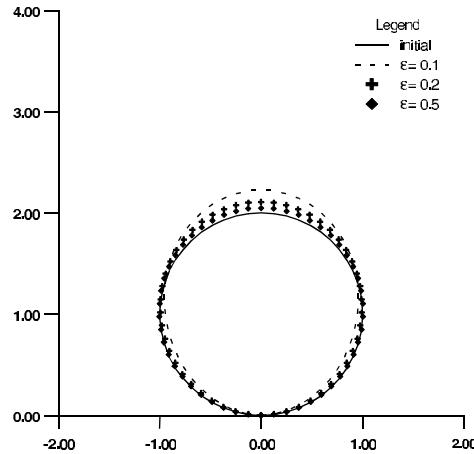


FIGURE 3.3.  $\theta(t) = t$ ,  $f_1(t) = \frac{\sin(t)}{SE}$ ,  $f_2(t) = \frac{\cos(t)}{SE}$ ,  $t \in [0, 2\pi]$ ,  
 $E = 100$ , © SIAM.

Based on this efficient solution method of (3.1), we consider associated shape optimization problems. We study the model problem of the minimization of the normal component of the deflection:

$$\underset{\theta \in U_{\text{ad}}}{\text{Min}} \left\{ \frac{1}{2} \int_0^1 [v_2(s)]^2 ds \right\} \quad (3.10)$$

subject to the optimality system defined in Theorem 3.2.

Notice that this optimality system becomes the state equation for the optimization problem (3.10), and all the unknowns appearing there,  $v_1, v_2, u, z, p, q, \lambda_1, \lambda_2$ , are the new state parameters. The new control unknown is just  $\theta \in U_{\text{ad}} \subset L^\infty(0, 1)$  (control constraints), which completely characterizes the geometry of the arch (its shape).

*Remark.* The significance of the problem (3.10) is to minimize the normal deflection (safety requirement) by choosing an advantageous shape of the structure. The field of forces  $[f_1, f_2] \in L^2(0, 1)^2$  is assumed to be given (for instance, the maximal load to which the arch may be subjected). General cost functionals may be studied in a similar way.

**Theorem 3.3.** *If  $U_{\text{ad}}$  is compact in  $L^\infty(0, 1)$ , then the problem (3.10) has at least one solution  $\theta^* \in U_{\text{ad}}$ .*

Figures 3.4–3.6 are related to the problem (3.10) for three examples of prescribed forces and of constraints on  $\theta$ . The values of the cost functional obtained in various iterations (corresponding to the two-dimensional curves represented by different graphical symbols) are written in the legend of each figure.

*Remark.* Problem (3.10) is a control-by-the-coefficients optimization problem. In the classical setting, the assumption that  $c' = \theta''$  (the curvature) is bounded in  $L^r(0, 1)$ ,  $r > 1$ , ensures existence. Theorem 3.3 uses a much weaker assumption, due to the application of the control variational method. It is possible to use gradient methods for the solution of (3.10) under the same assumption  $\theta \in L^\infty(0, 1)$ . This is a rather technical subject, and we refer to Ignat, Sprekels, and Tiba [9], Neittaanmäki, Sprekels, and Tiba [11], Ch. 6.1.2, for a complete treatment. Figures 3.4–3.6 give some examples of optimal shapes computed by the above approach together with several intermediate iterations. One should notice that Figures 3.4 and 3.5 admit a physical interpretation perfectly matching our numerical experiments. This is a hint that the model and the optimization method are well founded from the viewpoint of physics and have good stability properties.

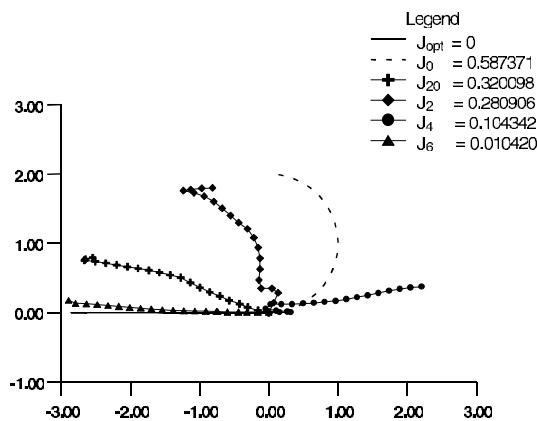


FIGURE 3.4.  $\theta(t) \in [0, \pi]$ ,  $f_1(t) = \frac{1}{S}$ ,  $f_2(t) = 0$ ,  $\theta_0(t) = t$ ,  
 $t \in [0, \pi]$ ,  $\odot$  SIAM.

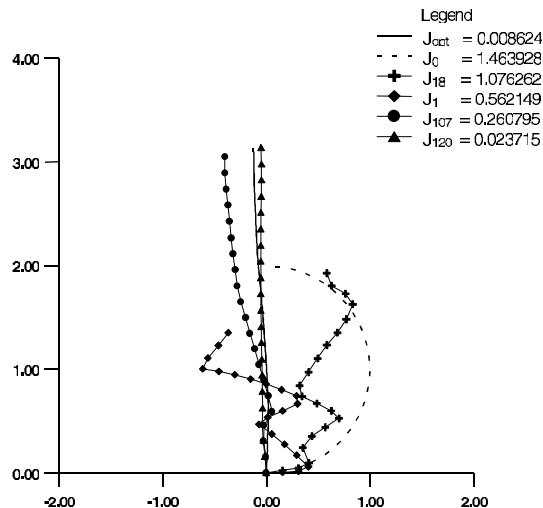


FIGURE 3.5.  $\theta(t) \in [0, \pi]$ ,  $f_1(t) = \frac{\sin(\theta(t))}{S}$ ,  $f_2(t) = \frac{\cos(\theta(t))}{S}$ ,  $\theta_0(t) = t$ ,  
 $t \in [0, \pi]$ ,  $\odot$  SIAM.

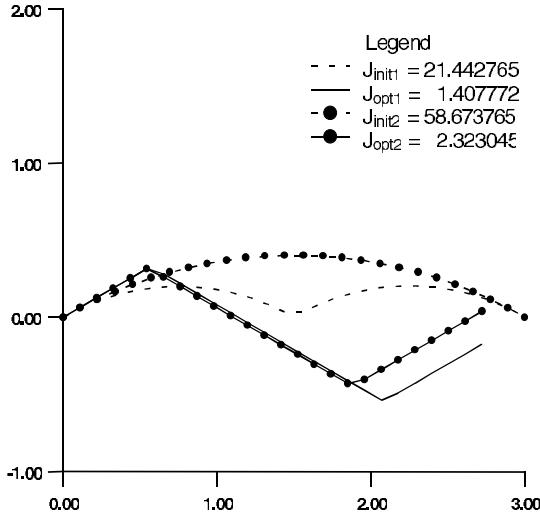


FIGURE 3.6.  $\theta(t) \in [\frac{\pi}{3}, \frac{2\pi}{3}]$ ,  $f_1(t) = \frac{\cos(\theta(t))}{S}$ ,  $f_2(t) = \frac{\sin(\theta(t))}{S}$ ,  
 $t \in [0, \pi]$ ,  $\theta_{01}(t) = \frac{2t+\pi}{3}$ ,  $t \in [0, \frac{\pi}{2}]$ ,  $\theta_{01}(t) = \frac{2t}{3}$ ,  
 $t \in [\frac{\pi}{2}, \pi]$ ,  $\theta_{02}(t) = \frac{t+\pi}{3}$ ,  $t \in [0, \pi]$ ,  $\circledcirc$  SIAM.

#### 4. Curved rods in dimension three

Let  $\bar{\theta} \in W^{2,\infty}(0,1)^3$  be the parametrization of a three-dimensional Jordan curve, and let  $\omega \in \mathbb{R}^2$  be some bounded Lipschitzian domain, not necessarily simply connected. If  $\{\bar{t}, \bar{n}, \bar{b}\}$  denotes some local orthonormal frame associated with the curve  $\bar{\theta}$ , we define the geometric transformation

$$F : \Omega = \omega \times ]0, 1[ \rightarrow F(\Omega) = \widehat{\Omega} \subset \mathbb{R}^3,$$

$$\begin{aligned} F(\bar{x}) &= F(x_1, x_2, x_3) = \bar{\theta}(x_3) + x_1 \bar{n}(x_3) + x_2 \bar{b}(x_3), \\ \forall (x_1, x_2) \in \omega, \quad \forall x_3 \in ]0, 1[. \end{aligned} \tag{4.1}$$

Denote by  $(h_{ij}(\bar{x}))_{i,j=1,3} = J(\bar{x})^{-1}$  the Jacobian of  $F$ ,  $J = \nabla F$ . If  $\text{diam } (\Omega)$  is “small”, then  $\det J(\bar{x}) \geq c_0 > 0$ ,  $\forall \bar{x} \in \Omega$ , and  $F : \Omega \rightarrow \widehat{\Omega}$  is a one-to-one transformation, Ciarlet [4], Thm. 3.1-1. A three-dimensional curved rod around the line  $\theta$  of centroids is the domain  $\widehat{\Omega}$  defined by (4.1).

Starting from the linear elasticity system, and using the supplementary assumption that the displacement at  $\hat{x} \in \widehat{\Omega}$  has the form

$$\bar{y}(\hat{x}) = \bar{\tau}(x_3) + x_1 \bar{N}(x_3) + x_2 \bar{B}(x_3) \tag{4.2}$$

with  $\bar{x} = (x_1, x_2, x_3) = F^{-1}(\hat{x}) \in \Omega$ , we obtain the following model for the curved rod:

$$\begin{aligned}
\mathcal{B}(\bar{y}, \bar{v}) &= \tilde{\lambda} \int_{\Omega} \sum_{i,j=1}^3 \left[ N_i(x_3) h_{1i}(\bar{x}) + B_i(x_3) h_{2i}(\bar{x}) + (\tau'_i(x_3) + x_1 N'_i(x_3) \right. \\
&\quad \left. + x_2 B'_i(x_3)) h_{3i}(\bar{x}) \right] \cdot \left[ M_j(x_3) h_{1j}(\bar{x}) + D_j(x_3) h_{2j}(\bar{x}) \right. \\
&\quad \left. + (\mu'_j(x_3) + x_1 M'_j(x_3) + x_2 D'_j(x_3)) h_{3j}(\bar{x}) \right] |\det J(\bar{x})| d\bar{x} \\
&\quad + \tilde{\mu} \int_{\Omega} \sum_{i < j} \left[ N_i(x_3) h_{1j}(\bar{x}) + B_i(x_3) h_{2j}(\bar{x}) + (\tau'_i(x_3) + x_1 N'_i(x_3) \right. \\
&\quad \left. + x_2 B'_i(x_3)) h_{3j}(\bar{x}) + N_j(x_3) h_{1i}(\bar{x}) + B_j(x_3) h_{2i}(\bar{x}) \right. \\
&\quad \left. + (\tau'_j(x_3) + x_1 N'_j(x_3) + x_2 B'_j(x_3)) h_{3i}(\bar{x}) \right] \\
&\quad \cdot \left[ M_i(x_3) h_{1j}(\bar{x}) + D_i(x_3) h_{2j}(\bar{x}) + (\mu'_i(x_3) + x_1 M'_i(x_3) \right. \\
&\quad \left. + x_2 D'_i(x_3)) h_{3j}(\bar{x}) + M_j(x_3) h_{1i}(\bar{x}) + D_j(x_3) h_{2i}(\bar{x}) \right. \\
&\quad \left. + (\mu'_j(x_3) + x_1 M'_j(x_3) + x_2 D'_j(x_3)) h_{3i}(\bar{x}) \right] |\det J(\bar{x})| d\bar{x} \\
&\quad + 2 \tilde{\mu} \int_{\Omega} \sum_{i=1}^3 \left[ N_i(x_3) h_{1i}(\bar{x}) + B_i(x_3) h_{2i}(\bar{x}) + (\tau'_i(x_3) + x_1 N'_i(x_3) \right. \\
&\quad \left. + x_2 B'_i(x_3)) h_{3i}(\bar{x}) \right] \cdot \left[ M_i(x_3) h_{1i}(\bar{x}) + D_i(x_3) h_{2i}(\bar{x}) \right. \\
&\quad \left. + (\mu'_i(x_3) + x_1 M'_i(x_3) + x_2 D'_i(x_3)) h_{3i}(\bar{x}) \right] |\det J(\bar{x})| d\bar{x} \\
&= \sum_{\ell=1}^3 \int_{\Omega} f_{\ell}(\bar{x})(\mu_{\ell}(x_3) + x_1 M_{\ell}(x_3) + x_2 D_{\ell}(x_3)) |\det J(\bar{x})| d\bar{x}, \tag{4.3}
\end{aligned}$$

for any test functions  $\bar{\mu} = (\mu_1, \mu_2, \mu_3)$ ,  $\bar{M} = (M_1, M_2, M_3)$ ,  $\bar{D} = (D_1, D_2, D_3) \in H_0^1(0, L)^3$ .

We have denoted  $\bar{v} = (\bar{\mu}, \bar{M}, \bar{D}) \in H_0^1(0, 1)^9$ , and  $\bar{y} = (\tau_1, \tau_2, \tau_3, N_1, N_2, N_3, B_1, B_2, B_3) \in H_0^1(0, 1)^9$ , the vector of the nine unknowns. The bilateral null conditions, given by the choice of the space  $H_0^1(0, 1)$ , correspond to clamped curved rods. The vector  $\bar{f} = (f_1, f_2, f_3) \in L^2(0, 1)^3$  represents the body forces acting on the rod, and  $\tilde{\lambda} \geq 0$ ,  $\tilde{\mu} > 0$  are the Lamé coefficients.

The condition (4.2) is very similar to (5.2) in the next section, and the choice of the test functions is, too. That is why we call the model (4.3) to be of a generalized Naghdi type, Ignat, Sprekels, and Tiba [10], by using the name for similar shell models (see next section). Notice that the regularity assumption  $\bar{\theta} \in W^{2,\infty}(0, 1)^3$  is one degree less than the usual assumptions in the literature, Trabucho and Viaño [20]. In the work of Tiba and Vodák [19], an asymptotic model is introduced for

curved rods under mere Lipschitz hypotheses for the parametrization. Shape optimization problems associated with three-dimensional curved rods are analyzed in Arnăutu, Sprekels, and Tiba [2], Neittaanmäki, Sprekels, and Tiba [11], Ch. 6.2.3, including numerical experiments with a fairly complete mathematical justification.

We associate with (4.3) the following optimal control problem:

$$\begin{aligned} \text{Min } & \left\{ \tilde{\lambda} \int_{\Omega} \sum_{i,j=1}^3 U_{ii}(\bar{x}) U_{jj}(\bar{x}) |\det J(\bar{x})| d\bar{x} \right. \\ & + \tilde{\mu} \int_{\Omega} \sum_{i < j} [U_{ij}(\bar{x}) + U_{ji}(\bar{x})]^2 |\det J(\bar{x})| d\bar{x} \\ & + 2 \tilde{\mu} \int_{\Omega} \sum_{i=1}^3 U_{ii}^2(\bar{x}) |\det J(\bar{x})| d\bar{x} - 2 \sum_{i=1}^3 \int_{\Omega} f_i(\bar{x}) [\tau_i(x_3) + x_1 N_i(x_3) \right. \\ & \left. \left. + x_2 B_i(x_3)] |\det J(\bar{x})| d\bar{x} \right\}, \end{aligned} \quad (4.4)$$

subject to the state system

$$\begin{aligned} & N_i(x_3) h_{1j}(\bar{x}) + B_i(x_2) h_{2j}(\bar{x}) + [\tau'_i(x_3) + x_1 N'_i(x_3) + x_2 B'_i(x_3)] h_{3j}(\bar{x}) \\ = & U_{ij}(\bar{x}) \text{ in } \Omega, \end{aligned} \quad (4.5)$$

$$N_i(0) = B_i(0) = \tau_i(0) = 0, \quad i = \overline{1,3}, \quad (4.6)$$

and to the control constraints

$$U = \{U_{ij}\}_{i,j=\overline{1,3}} \in \mathcal{V} \subset L^2(\Omega)^9. \quad (4.7)$$

Here,  $\mathcal{V} \neq \emptyset$  is the closed linear subspace in  $L^2(\Omega)^9$  generated by all functions in  $L_0^2(0,1)$  (with zero mean) used on the “position” of  $\tau'_i, N'_i, B'_i$ ,  $i = \overline{1,3}$ , in (4.5). Clearly,  $N_i, B_i$  can be immediately obtained by simple integration and (4.6), which gives the form of  $U_{ij}$  in the right-hand side of (4.6) and spans  $\mathcal{V}$ .

Notice that the condition  $\tau'_i, N'_i, B'_i \in L_0^2(0,1)$ ,  $i = \overline{1,3}$ , is the same thing as

$$\tau_i(1) = N_i(1) = B_i(1) = 0, \quad i = \overline{1,3},$$

which could be imposed instead of (4.7). We prefer to impose the control constraint (4.7) in this form, since it is explicit. In (4.6),  $(x_1, x_2) \in \omega$  appears as parameter, and the definition of  $\mathcal{V}$  ensures that (4.6) has a unique solution for any  $U = (U_{ij})_{i,j=\overline{1,3}} \in \mathcal{V}$ . If (4.7) is not fulfilled, then (4.6) may have no solution.

**Theorem 4.1.** *The optimal control problem (4.4)–(4.7) has a unique optimal couple  $U^* = \{U_{ij}^*\} \in \mathcal{V}$ ,  $[\tau_i^*, B_i^*, N_i^*]_{i=\overline{1,3}} \in H_0^1(0,1)^9$ , and the optimal state is the unique solution to the system (4.3) that governs the Naghdi generalized model for curved rods.*

This shows that the problem (4.4)–(4.7) is well posed and may be solved instead of (4.3). The next results underline the simplicity of (4.4)–(4.7).

**Proposition 4.2.** *If  $\{U_{ij}^*\}$  is known, then  $[\tau_i^*, N_i^*, B_i^*]_{i=1,3}$  may be computed explicitly.*

This is, of course, valid for any  $\{U_{ij}\}_{i,j=1,3}$  and any  $[\tau_i, N_i, B_i]_{i=1,3}$  that satisfy (4.5), (4.6), that is, the state system may be solved explicitly. There are certain orthogonality relations between the coefficients  $(h_{ij})_{i,j=1,3} = J^{-1}$  that may be obtained by their explicit computation, and this yields

$$B_i^* = \sum_{j=1}^3 U_{ij}^* b_j, \quad i = \overline{1,3}, \quad (4.8)$$

$$N_i^* = \sum_{j=1}^3 U_{ij}^* n_j, \quad i = \overline{1,3}, \quad (4.9)$$

$$\begin{aligned} & (\tau_i^*)' + x_1(N_i^*)' + x_2(B_i^*)' \\ &= \sum_{j=1}^3 U_{ij}^* t_j \det J(\bar{x}) + \sum_{j=1}^3 U_{ij}^* n_j c x^2 - \sum_{j=1}^3 U_{ij}^* b_j c x_1, \quad i = \overline{1,3}, \end{aligned} \quad (4.10)$$

where  $c$  together with  $a, \beta$  are  $L^\infty(0, 1)$  coefficients of curvature and torsion type that may be computed from the “equations of motion”

$$\begin{aligned} \bar{t}'(x_3) &= a(x_3) \bar{b}(x_3) + \beta(x_3) \bar{n}(x_3), \\ \bar{b}'(x_3) &= -a(x_3) \bar{t}(x_3) + c(x_3) \bar{n}(x_3), \\ \bar{n}'(x_3) &= -\beta(x_3) \bar{t}(x_3) - c(x_3) \bar{b}(x_3). \end{aligned}$$

*Remark.* Let us denote by  $\Lambda_i(U_{ij})$  the right-hand side in (4.10). Then, we can perform the following substitution in (4.4):

$$\begin{aligned} & \sum_{i=1}^3 \int_{\Omega} f_i [\tau_i + x_1 N_i + x_2 B_i] \det J d\bar{x} \\ &= - \sum_{i=1}^3 \int_{\Omega} [\tau'_i + x_1 N'_i + x_2 B'_i] \int_0^{x_3} f_i(x_1, x_2, \rho) \det J(x_1, x_2, \rho) d\rho d\bar{x} \\ &= - \sum_{i=1}^3 \int_{\Omega} \Lambda_i(U_{ij}) \int_0^{x_3} f_i(x_1, x_2, \rho) \det J(x_1, x_2, \rho) d\rho d\bar{x}. \end{aligned}$$

In this simple way, the optimal control (4.4)–(4.7) is transformed into a mathematical programming problem defined on  $\mathcal{V} \subset L^2(\Omega)^9$ , since the state disappears. One can also compare the approach from this section to the solution of the Kirchhoff–Love model in Section 3, in dimension 2.

## 5. Generalized Naghdi shells

Let  $\omega \subset \mathbb{R}^2$  be a bounded Lipschitzian domain,  $\varepsilon > 0$  “small”, and  $\Omega = \omega \times ]-\varepsilon, \varepsilon[$ . We assume that

$$\partial\omega = \gamma_0 \cup \gamma_1, \quad \gamma_0 \cap \gamma_1 = \emptyset,$$

and we denote

$$\Gamma_0 = \gamma_0 \times ]-\varepsilon, \varepsilon[,$$

$$\Gamma_1 = \partial\Omega \setminus \Gamma_0,$$

$$\mathcal{V}(\omega) = \left\{ \bar{v} = (v_1, v_2, v_3) \in H^1(\omega)^3; \bar{v}|_{\gamma_0} = 0 \right\}.$$

Let  $p : \omega \rightarrow \mathbb{R}$  be piecewise in  $C^2(\bar{\omega})$ , and let  $\bar{n} : \omega \rightarrow \mathbb{R}^3$ ,  $\bar{n} = (n_1, n_2, n_3)$ , be the unit normal vector to the graph of  $p$  in  $\mathbb{R}^3$ . We denote by  $\bar{\pi} = (\pi_1, \pi_2, \pi_3) = (x_1, x_2, p(x_1, x_2))$  this graph, which will represent the midsurface of the shell. We define the transformation  $F : \Omega \rightarrow F(\Omega) = \hat{\Omega} \subset \mathbb{R}^3$  by

$$F(\bar{x}) = F(x_1, x_2, x_3) = \bar{\pi}(x_1, x_2) + x_3 \bar{n}(x_1, x_2). \quad (5.1)$$

If  $\varepsilon > 0$  is small enough, then (5.1) is a one-to-one transformation (Ciarlet [4], Thm. 3.1-1), which justifies the definition of the shell’s geometry  $\hat{\Omega} = F(\Omega)$ . Denote by  $J = \nabla F$  the Jacobian of  $F$ , and let  $(h_{ij}(\bar{x}))_{i,j=1,3} = J(\bar{x})^{-1}$ . Starting from the linear elasticity system, and using the assumption that the displacement has the form

$$\hat{y}(\hat{x}) = \bar{u}(x_1, x_2) + x_3 \bar{r}(x_1, x_2), \quad \forall \hat{x} \in \hat{\Omega}, \quad \bar{x} = F^{-1}(\hat{x}), \quad (5.2)$$

the following generalized Naghdi shell model of the deformation was deduced in Sprekels and Tiba [16]:

$$\begin{aligned} & \mathcal{B}([\bar{u}, \bar{r}], [\bar{\mu}, \bar{\rho}]) \\ &= \lambda \int_{\Omega} \left\{ \sum_{i=1}^3 \left[ \left( \frac{\partial u_i}{\partial x_1} + x_3 \frac{\partial r_i}{\partial x_1} \right) h_{1i} + \left( \frac{\partial u_i}{\partial x_2} + x_3 \frac{\partial r_i}{\partial x_2} \right) h_{2i} + r_i h_{3i} \right] \right. \\ & \quad \cdot \left. \left\{ \sum_{j=1}^3 \left[ \left( \frac{\partial \mu_j}{\partial x_1} + x_3 \frac{\partial \rho_j}{\partial x_1} \right) h_{1j} + \left( \frac{\partial \mu_j}{\partial x_2} + x_3 \frac{\partial \rho_j}{\partial x_2} \right) h_{2j} + \rho_j h_{3j} \right] \right\} \right. \\ & \quad \cdot |\det J(\bar{x})| d\bar{x} \\ &+ 2\mu \int_{\Omega} \sum_{i=1}^3 \left[ \left( \frac{\partial u_i}{\partial x_1} + x_3 \frac{\partial r_i}{\partial x_1} \right) h_{1i} + \left( \frac{\partial u_i}{\partial x_2} + x_3 \frac{\partial r_i}{\partial x_2} \right) h_{2i} + r_i h_{3i} \right] \\ & \quad \cdot \left[ \left( \frac{\partial \mu_i}{\partial x_1} + x_3 \frac{\partial \rho_i}{\partial x_1} \right) h_{1i} + \left( \frac{\partial \mu_i}{\partial x_2} + x_3 \frac{\partial \rho_i}{\partial x_2} \right) h_{2i} + \rho_i h_{3i} \right] |\det J(\bar{x})| d\bar{x} \end{aligned}$$

$$\begin{aligned}
& + \mu \int_{\Omega} \sum_{i < j} \left\{ \left[ \left( \frac{\partial u_i}{\partial x_1} + x_3 \frac{\partial r_i}{\partial x_1} \right) h_{1j} + \left( \frac{\partial u_i}{\partial x_2} + x_3 \frac{\partial r_i}{\partial x_2} \right) h_{2j} + r_i h_{3j} \right. \right. \\
& \quad \left. \left. + \left( \frac{\partial u_j}{\partial x_i} + x_3 \frac{\partial r_j}{\partial x_1} \right) h_{1i} + \left( \frac{\partial u_j}{\partial x_2} + x_3 \frac{\partial r_j}{\partial x_2} \right) h_{2i} + r_j h_{3i} \right] \right. \\
& \quad \cdot \left[ \left( \frac{\partial \mu_i}{\partial x_1} + x_3 \frac{\partial \rho_i}{\partial x_1} \right) h_{1j} + \left( \frac{\partial \mu_i}{\partial x_2} + x_3 \frac{\partial \rho_i}{\partial x_2} \right) h_{2j} + \rho_i h_{3j} \right. \\
& \quad \left. \left. + \left( \frac{\partial \mu_j}{\partial x_1} + x_3 \frac{\partial \rho_j}{\partial x_1} \right) h_{1i} + \left( \frac{\partial \mu_j}{\partial x_2} + x_3 \frac{\partial \rho_j}{\partial x_2} \right) h_{2i} + \rho_j h_{3i} \right] \right. \\
& \quad \cdot |\det J(\bar{x})| d\bar{x} \\
& = \sum_{i=1}^3 \left\{ \int_{\Omega} f_i(\mu_i + x_3 \rho_i) d\bar{x} + \int_{\Gamma_1} h_i(\mu_i + x_3 \rho_i) d\sigma \right\}, \\
& \quad \forall \bar{\mu} = (\mu_1, \mu_2, \mu_3), \quad \forall \bar{\rho} = (\rho_1, \rho_2, \rho_3) \in \mathcal{V}(\omega). \quad (5.3)
\end{aligned}$$

In (5.3),  $\bar{f} = (f_1, f_2, f_3)$  represents the body forces,  $\bar{h} = (h_1, h_2, h_3)$  the surface tractions,  $\lambda \geq 0$ ,  $\mu > 0$  are the Lamé coefficients, and the shell is partially clamped along  $\Gamma_0$ . The unknown  $\bar{u} \in \mathcal{V}(\omega)$  may be interpreted as the deformation of the midsurface of the shell, and  $\bar{r} \in \mathcal{V}(\omega)$  is the deformation of the normal  $\bar{n}$  (which may also modify its length, in this sense generalizing the classical Naghdi model, Ciarlet [4]).

We show that it is possible to solve directly the generalized Naghdi shell model (5.3) via a control problem governed by a finite number of independent Poisson equations. This choice is motivated by its simplicity, and other choices are also possible (which is a general characteristic of the control variational method):

$$\begin{aligned}
\min_{w \in [L^2(\omega)]^{12}} \left\{ L(w) = \frac{1}{2} \mathcal{B}([\bar{u}, \bar{r}], [\bar{u}, \bar{r}]) + \frac{1}{2} \sum_{i=1}^3 \int_{\omega} [|w_1^i|_{\mathbb{R}^2}^2 + |w_2^i|_{\mathbb{R}^2}^2] dx_1 dx_2 \right. \\
\left. - \frac{1}{2} \sum_{i=1}^3 \int_{\omega} [|\nabla u_i|_{\mathbb{R}^2}^2 + |\nabla r_i|_{\mathbb{R}^2}^2] dx_1 dx_2 \right\}, \quad (5.4)
\end{aligned}$$

subject to

$$\begin{aligned}
\sum_{i=1}^3 \int_{\omega} [\nabla u_i \cdot \nabla \phi_i + \nabla r_i \cdot \nabla \psi_i] dx_1 dx_2 = \sum_{i=1}^3 \int_{\omega} [w_1^i \cdot \nabla \phi_i + w_2^i \cdot \nabla \psi_i] dx_1 dx_2 \\
+ \sum_{i=1}^3 \left\{ \int_{\Omega} f_i(\phi_i + x_3 \psi_i) d\bar{x} + \int_{\Gamma_1} h_i(\phi_i + x_3 \psi_i) d\tau \right\}, \quad \forall \phi, \psi \in \mathcal{V}(\omega). \quad (5.5)
\end{aligned}$$

This is an unconstrained control problem with  $w = [w_1, w_2]$  as control parameter,  $w_\ell = [w_\ell^1, w_\ell^2, w_\ell^3] \in L^2(\omega)^6$ ,  $\ell = 1, 2$ .

**Theorem 5.1.** *The optimal control problem (5.4), (5.5) has a unique optimal couple  $[u^*, r^*] \in \mathcal{V}(\omega)^2$ ,  $[w_1^*, w_2^*] \in L^2(\omega)^{12}$ , and  $[u^*, r^*]$  is the unique solution of the Naghdi generalized model (5.3).*

*Remark.* One can compute the gradient of the cost (5.4) and use gradient methods for the solution of (5.4), (5.5) and, implicitly, of (5.3). In each iteration of the algorithm a finite number of independent Poisson equations (for the state and the adjoint system) have to be solved. If convex state constraints are added to (5.4), (5.5), then one obtains a variational inequality for (5.3). Notice that (5.4) may be interpreted as the energy of the generalized Naghdi model as it is usual in the control variational method. We also quote the work of Sprekels and Tiba [18], including full details of the results introduced in Sections 4 and 5. In the paper Sprekels and Tiba [17], the control variational method is applied to the full elasticity system.

## References

- [1] V. Arnăutu, H. Langmach, J. Sprekels, D. Tiba, *On the approximation and the optimization of plates*. Numer. Funct. Anal. Optim. **21:3–4** (2000), 337–354.
- [2] V. Arnăutu, J. Sprekels, D. Tiba, *Optimization of curved mechanical structures*. SIAM J. Control Optim. **44:2** (2005), 743–775.
- [3] Ph. Ciarlet, *The Finite Element Method for Elliptic Problems*. North-Holland, Amsterdam, 1978.
- [4] Ph. Ciarlet, *Mathematical Elasticity. Vol. III. Theory of Shells*. North-Holland, Amsterdam, 2000.
- [5] Ph. Ciarlet, P. Ciarlet, *Another approach to linearized elasticity and Korn's inequality*. C. R. Acad. Sci. Paris, Ser. I **339** (2004), 307–312.
- [6] Ph. Ciarlet, P. Ciarlet, *Another approach to linearized elasticity and a new proof of Korn's inequality*. Math. Models Methods Appl. Sci. **15:2** (2005), 259–271.
- [7] Ph. Ciarlet, L. Gratie, *A new approach to linear shell theory*. C. R. Acad. Sci. Paris, Ser. I **340** (2005), 471–478.
- [8] L. Gratie, *New unknowns on the midsurface of a Naghdi's shell model*. An. St. Univ. Ovidius **12:2** (2004), 115–126.
- [9] A. Ignat, J. Sprekels, D. Tiba, *Analysis and optimization of nonsmooth arches*. SIAM J. Control Optim. **40** (2001), 1107–1135.
- [10] A. Ignat, J. Sprekels, D. Tiba, *A model of a general elastic curved rod*. Math. Methods Appl. Sci. **25:10** (2002), 835–854.
- [11] P. Neittaanmäki, J. Sprekels, D. Tiba, *Optimization of Elliptic Systems. Theory and Applications*. Springer, New York, 2006.
- [12] J. Sprekels, D. Tiba, *Propriétés de bang-bang généralisées dans l'optimisation des plaques*. C. R. Acad. Sci. Paris, Ser. I **327** (1998), 705–710.

- [13] J. Sprekels, D. Tiba, *A duality approach in the optimization of beams and plates.* SIAM J. Control Optim. **37:2** (1998/1999), 486–501.
- [14] J. Sprekels, D. Tiba, *A duality-type method for the design of beams.* Adv. Math. Sci. Appl. **9:1** (1999), 84–102.
- [15] J. Sprekels, D. Tiba, *Sur les arches lipschitziennes.* C. R. Acad. Sci. Paris, Sér. I **331:2** (2000), 179–184.
- [16] J. Sprekels, D. Tiba, *An analytic approach to a generalized Naghdi shell model.* Adv. Math. Sci. Appl. **12:1** (2002), 175–190.
- [17] J. Sprekels, D. Tiba, *Optimal design of mechanical structures.* In: *Control Theory of Partial Differential Equations*, LN Pure Appl. Math. **242**, Chapman & Hall / CRC, Boca Raton, Florida, 2005, pp. 259–271.
- [18] J. Sprekels, D. Tiba, *The control variational method for differential systems.* submitted to SIAM J. Control Optim, 2007.
- [19] D. Tiba, R. Vodák, *A general asymptotic model for Lipschitzian curved rods.* Adv. Math. Sci. Appl. **15:1** (2005), 137–198.
- [20] L. Trabucho, J.M. Viaño, *Mathematical modelling of rods.* in: Handbook of Numerical Analysis, vol. IV (Ph. Ciarlet, J.L. Lions, eds.), Elsevier, Amsterdam, 1996.

Jürgen Sprekels  
Weierstrass Institute  
for Applied Analysis  
and Stochastics  
Mohrenstrasse 39  
D-10117 Berlin, Germany  
e-mail: sprekels@wias-berlin.de

Dan Tiba  
Institute of Mathematics  
Romanian Academy  
P. O. Box 1-764  
014700 Bucharest, Romania  
e-mail: dan.tiba@imar.ro

# A New Non-linear Semidefinite Programming Algorithm with an Application to Multidisciplinary Free Material Optimization

M. Stingl, M. Kočvara and G. Leugering

**Abstract.** A new method and algorithm for the efficient solution of a class of nonlinear semidefinite programming problems is introduced. The new method extends a concept proposed recently for the solution of convex semidefinite programs based on the sequential convex programming (SCP) idea. In the core of the method, a generally non-convex semidefinite program is replaced by a sequence of subproblems, in which nonlinear constraint and objective functions defined in matrix variables are approximated by block separable convex models. Global convergence is proved under reasonable assumptions. The article is concluded by numerical experiments with challenging Free Material Optimization problems subject to displacement constraints.

**Mathematics Subject Classification (2000).** 74B05, 74P05, 90C90, 90C25,  
90C22.

**Keywords.** Structural optimization, material optimization, semidefinite programming, sequential convex programming.

## 1. Introduction

In the last two decades semidefinite programming problems (SDP) have received more and more attention. One of the main reasons is the large variety of applications leading to semidefinite programs; see [3, 15], for example. As a consequence, various algorithms for solving semidefinite programs have been developed – most of them specialized in linear SDPs. Famous examples are interior point or dual scaling methods (see, for example, [15, 31, 33] and the references therein). Only recently, some of the algorithmic concepts have been generalized to nonlinear semidefinite programs. Nowadays, there are several approaches: For example, in [9] and [11] the

---

This work has been partially supported by the EU Commission in the Sixth Framework Program, Project No. 30717 PLATO-N.

sequential quadratic programming idea is generalized to semidefinite programs. A similar approach combining interior point ideas with sequential semidefinite programming is presented in [16]. As alternative to this class of generalized interior point methods, some Lagrangian type methods have been explored, see, [28, 23], for instance. A related concept based on the modified barrier idea [24] lead to an efficient implementation in the code PENNON [19, 22].

However the situation is still such that many classes of semidefinite programs remain unsolved. The main difficulty is the high computational complexity. On the one hand, for some classes of SDP instances exploitation of sparse data structures is very well understood. For instance, the techniques described in [14] are implemented in many linear SDP solvers [8, 18, 32]. In the last five years, efficient algorithms involving Krylov type methods have been adopted to semidefinite programming and have been successfully applied to relaxations of combinatorial optimization problems; see [36, 25, 22]. Only very recently the authors in [10] found that algebraic properties of certain SDP instances can lead to amazing computational complexity improvements. On the other hand there are still SDP instances, for which all these techniques and concepts fail to work. One prominent example is the so-called free material optimization problem.

Free material optimization (FMO) is a branch of structural optimization. It represents a generalization of so-called topology optimization (see [4]) that, nowadays, is being routinely used in the industry. FMO has been successfully used for conceptual design of aircraft components; the most prominent example is the design of ribs in the leading edge of Airbus A380 [17]. The underlying FMO model was introduced in [5] and [26] and has been studied in several further articles such as [2, 39]. The optimization variable is the (positive definite) material tensor which is allowed to vary from point to point. The method is supported by powerful optimization and numerical techniques, which are based on dualization of the original convex problem and lead to large scale semidefinite programming problems [2]. The dualization approach has however two major disadvantages. First of all, the computational complexity of the method depends cubically on the number of load cases [20], which makes the approach impractical for 3D problems with more than a few (typically 3–5) load cases. This was the main motivation of the authors to develop the method described in [27]. Moreover, it is almost impossible to apply the dual approach to extended (multi-disciplinary) FMO problems, as these problems are typically non-convex. This is a serious drawback, as many design constraints, such as displacement-based constraints are typical requirements arising in many real-world applications; compare [17, 21].

In this article we propose an algorithmic concept – a generalization of the method described in [27] – which is able to cope with the latter class of problems. Like the original approach, the method is based on a class of sequential convex programming algorithms, of which the most prominent representatives are CONLIN [12], the method of moving asymptotes (MMA) [29, 30] and SCPIP [37, 38]. In contrast to the results described in [27], where the focus was on convex, potentially non-smooth semidefinite programs, our main interest here is in

non-convex problems. We follow the ideas presented in [30] in order to establish a global convergence result for our method. Moreover we discuss numerical details of our algorithm and demonstrate by numerical experiments that the new method is a viable alternative and supplement to existing methods in the field of material optimization.

The structure of this article is as follows: In Section 2 we define the basic problem statement. In the third section, we recall the definition of convex, separable hyperbolic approximations of functions defined on matrix spaces. In Section 4, these approximations are used to construct a globally convergent algorithm. Then, in Section 5, we briefly describe the free material optimization (FMO) model including displacement constraints. Finally, in Section 6, we present results of numerical studies with 2D- and 3D-FMO problems, of which the latter ones have not been solved by any existing method yet.

Throughout this article we use the following notation: We denote by  $\mathbb{S}^d$  the space of symmetric  $d \times d$ -matrices equipped with the standard inner product  $\langle \cdot, \cdot \rangle_{\mathbb{S}^d}$  defined by  $\langle A, B \rangle_{\mathbb{S}^d} := \text{Tr}(AB)$  for any pair of matrices  $A, B \in \mathbb{S}^d$ . We denote by  $\mathbb{S}_+^d$  the cone of all positive semidefinite matrices in  $\mathbb{S}^d$  and use the abbreviation  $A \succ_{\mathbb{S}^d} 0$  for matrices  $A \in \mathbb{S}_+^d$ . Moreover, for  $A, B \in \mathbb{S}^d$ , we say that  $A \succ_{\mathbb{S}^d} B$  if and only if  $A - B \succ_{\mathbb{S}^d} 0$ , and similarly for  $A \preccurlyeq_{\mathbb{S}^d} B$ .

## 2. Basic problem statement

Our aim is to solve the following generic semidefinite program:

$$\min_{Y \in \mathbb{S}} f_0(Y) \quad (P)$$

subject to

$$f_\ell(Y) \leq 0, \quad \ell = 1, 2, \dots, L,$$

$$g_k(Y) \leq 0, \quad k = 1, 2, \dots, K,$$

$$\underline{Y}_i \preccurlyeq_{\mathbb{S}^{d_i}} Y_i \preccurlyeq_{\mathbb{S}^{d_i}} \bar{Y}_i, \quad i = 1, 2, \dots, m$$

with

$$\mathbb{S} = \mathbb{S}^{d_1} \times \mathbb{S}^{d_2} \times \dots \times \mathbb{S}^{d_m} \text{ and } (d_1, d_2, \dots, d_m) \in \mathbb{N}^m.$$

We assume that, in general,  $m$  is large ( $10^3 - 10^5$ ) and  $d_i$  are small (2–10). That is, we have many small-size matrix variables and matrix constraints.

In what follows  $F$  denotes the feasible domain of problem  $(P)$ . Throughout the paper we make the following assumptions:

- (A1) The functions  $f_\ell : \mathbb{S} \rightarrow \mathbb{R}$ , ( $\ell = 0, 1, \dots, L$ ) are continuously differentiable.
- (A2) The functions  $g_k : \mathbb{S} \rightarrow \mathbb{R}$  ( $k = 1, 2, \dots, K$ ) are continuously differentiable, convex and separable with respect to the matrix variable  $Y$ .
- (A3) Problem  $(P)$  admits at least one solution.
- (A4) The non-degeneracy constraint qualification (see [15, 7]) holds for  $P$ .

Problems of type  $(P)$  arise in various applications. Our main motivation is to solve the free material optimization problems described in detail in Section 5. However, other applications can be found, e.g., in spline approximation [1] and sparse SDP relaxation of polynomial optimization problems [34].

*Remark 2.1.* Assumption (A1) differs from the assumptions made in [27]. There,  $f$  and all constraint functions are required to be convex. On the other hand, we have a stronger continuity condition in this article. The non-degeneracy constraint qualification used in assumption (A4) is a straightforward generalization of the well-known linear independency constraint qualification.

*Remark 2.2.* For a short note on moderately successful experiences with ‘standard SDP solvers’ applied to a sub-class of  $(P)$ -type problems as well as a brief motivation for the choice of the sequential convex programming framework, the interested reader is referred to [27]. Here we only remark that the concept described throughout this article can not be seen as a universal cure for large-scale (nonlinear) SDP problems, but, as we will see in Section 6, it turns out to be efficient when solving generalized FMO problems.

### 3. A block-separable convex approximation scheme

In this section we briefly outline the concept of block-separable convex approximations (see [27]) of continuously differentiable functions

$$f : \mathbb{S} \rightarrow \mathbb{R}, \text{ where } \mathbb{S} = \mathbb{S}^{d_1} \times \mathbb{S}^{d_2} \times \cdots \times \mathbb{S}^{d_m} \text{ and } (d_1, d_2, \dots, d_m) \in \mathbb{N}^m. \quad (3.1)$$

We introduce the following convenient notation: Let  $I = \{1, 2, \dots, m\}$ . On  $\mathbb{S}$  we define the inner product  $\langle \cdot, \cdot \rangle_{\mathbb{S}} := \sum_{i \in I} \langle \cdot, \cdot \rangle_{\mathbb{S}^{d_i}}$ , where  $\langle \cdot, \cdot \rangle_{\mathbb{S}^{d_i}}$  is the standard inner product in  $\mathbb{S}^{d_i}$  ( $i \in I$ ). Moreover, we denote by  $\|\cdot\|_{\mathbb{S}}$  the norm induced by  $\langle \cdot, \cdot \rangle_{\mathbb{S}}$ . Finally, we denote the directional derivatives of  $f$  of first and second order in directions  $V, W \in \mathbb{S}$  by  $\frac{\partial}{\partial Y} f(Y; V)$  and  $\frac{\partial^2}{\partial Y \partial Y} f(Y; V, W)$ , respectively.

**Definition 3.1.** We call an approximation  $g : \mathbb{S} \rightarrow \mathbb{R}$  of a function  $f$  of type (3.1) a *convex first-order approximation* at  $\bar{Y} = (\bar{Y}_1, \dots, \bar{Y}_m) \in \mathbb{S}$ , if the following assumptions are satisfied:

- (AP1)  $g(\bar{Y}) = f(\bar{Y})$ ,
- (AP2)  $\frac{\partial}{\partial Y_i} g(\bar{Y}) = \frac{\partial}{\partial Y_i} f(\bar{Y})$  for all  $i \in I$ ,
- (AP3)  $g$  is convex.

In the following, we recall the local block separable *convex first-order approximation* scheme for functions of type  $f$ , proposed in [27]. We start with the following definitions:

**Definition 3.2.** Let  $f : \mathbb{S} \rightarrow \mathbb{R}$  be continuously differentiable on a subset  $B \subset \mathbb{S}$ . For all  $i \in I$  we define differential operators entry-wise by

$$(\nabla^i f)_{\ell,j} := \left( \frac{\partial f}{\partial Y_i} \right)_{\ell,j}, \quad 1 \leq \ell, j \leq d_i$$

and denote by  $\nabla_+^i f(\bar{Y})$  and  $\nabla_-^i f(\bar{Y})$  the projections of  $\nabla^i f(\bar{Y})$  onto  $\mathbb{S}_+^{d_i}$  and  $\mathbb{S}_-^{d_i}$ , respectively.

**Definition 3.3.** Let  $f : \mathbb{S} \rightarrow \mathbb{R}$  be continuously differentiable on a subset  $B \subset \mathbb{S}$  and  $\bar{Y} = (\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_m) \in B$ . Moreover let *asymptotes*  $L = (L_1, L_2, \dots, L_m)^\top$ ,  $U = (U_1, U_2, \dots, U_m)^\top$  be given such that

$$L_i \prec_{\mathbb{S}_+^{d_i}} \bar{Y}_i \prec_{\mathbb{S}_+^{d_i}} U_i \quad \text{for all } i \in I$$

and  $\tau := \{\tau_1, \tau_2, \dots, \tau_m\}$  be a set of non-negative real parameters. Then we define the *hyperbolic approximation*  $f_{\bar{Y}}^{L, U, \tau}$  of  $f$  at  $\bar{Y}$  as

$$\begin{aligned} f_{\bar{Y}}^{L, U, \tau}(Y) &:= f(\bar{Y}) \\ &+ \sum_{i=1}^m \langle \nabla_+^i f(\bar{Y}), (U_i - \bar{Y}_i)(U_i - Y_i)^{-1}(U_i - \bar{Y}_i) - (U_i - \bar{Y}_i) \rangle_{\mathbb{S}^{d_i}} \\ &- \sum_{i=1}^m \langle \nabla_-^i f(\bar{Y}), (\bar{Y}_i - L_i)(Y_i - L_i)^{-1}(\bar{Y}_i - L_i) - (\bar{Y}_i - L_i) \rangle_{\mathbb{S}^{d_i}} \\ &+ \sum_{i=1}^m \tau_i \langle (Y_i - \bar{Y}_i)^2, (U_i - Y_i)^{-1} + (Y_i - L_i)^{-1} \rangle_{\mathbb{S}^{d_i}}. \end{aligned} \quad (3.2)$$

The following Theorem ([27]) says that (3.2) is a convex approximation in the sense of Definition 3.1.

#### Theorem 3.4.

- a)  $f_{\bar{Y}}^{L, U, \tau}$  satisfies assumptions (AP1) to (AP3).
- b)  $f_{\bar{Y}}^{L, U, \tau}$  is separable w.r.t. the matrix variables  $Y_1, Y_2, \dots, Y_m$ .
- c) Let  $B$  be a compact subset of  $\mathbb{S}$ ,  $\bar{\tau} \geq \tau_i \geq \underline{\tau} > 0$  for all  $i \in I$ , and asymptotes  $L$  and  $U$  in the sense of definition 3.3 be given. Then  $f_{\bar{Y}}^{L, U, \tau}$  is strongly convex on  $B$ . Moreover the second-order derivative of  $f_{\bar{Y}}^{L, U, \tau}$  is uniformly bounded for all  $\bar{Y} \in B$ .

*Remark 3.5.* Theorem 3.4 differs from the original theorem in [27] in the choice of the asymptotes. Here we restrict ourselves to only one (fixed) choice of asymptotes. The reason for this simplification is twofold. First it helps to unburden the notation. Second, and more important there is no efficient dynamic choice of asymptotes known in the semidefinite programming case. This is in sharp contrast to the standard nonlinear programming situation; see [6, 13, 29, 38].

#### 4. A globally convergent algorithm based on hyperbolic approximations

In the framework of this section we use the local hyperbolic approximations defined in Section 3 in order to establish a solution scheme for the generic optimization problem  $(P)$ :

Given an iteration index  $j$  and an associated feasible point  $Y^j$  of problem  $(P)$ , we define local hyperbolic approximations of  $f_\ell$  ( $\ell = 0, 1, \dots, L$ ) as

$$f_\ell^j(Y) := (f_\ell)_{Y^j}^{\tau^j}(Y) := (f_\ell)_{Y^j}^{L, U, \tau^j}(Y),$$

and local approximations of  $(P)$  close to  $Y^j$  as follows:

$$\begin{aligned} & \min_{Y \in \mathbb{S}} f_0^j(Y) \\ & \text{subject to} \end{aligned} \tag{P<sup>j</sup>}$$

$$\begin{aligned} f_\ell^j(Y) &\leq 0, \quad \ell = 1, 2, \dots, L, \\ g_k(Y) &\leq 0, \quad k = 1, 2, \dots, K, \\ \underline{Y}_i &\preceq_{\mathbb{S}^{d_i}} Y_i \preceq_{\mathbb{S}^{d_i}} \overline{Y}_i, \quad i = 1, 2, \dots, m. \end{aligned}$$

We denote the feasible domain of problem  $(P^j)$  by  $F^j$ . By construction the following corollary is an immediate consequence of Theorem 3.4:

**Corollary 4.1.** *For all  $\ell = 0, 1, \dots, L$ :*

- a)  $f_\ell^j(Y^j) = f_\ell(Y^j)$ .
- b) *The gradients of  $f_\ell^j$  and  $f_\ell$  coincide at  $Y^j$ .*
- c) *The functions  $f^\ell$  are convex.*
- d) *The functions  $f^\ell$  are separable w.r.t.  $Y_1, Y_2, \dots, Y_m$ .*

*Remark 4.2.* Note that the function  $F_\ell(Y, Y', \tau) := (f_\ell)_{Y'}^\tau(Y)$  depends continuously on all its arguments (not only on  $Y$ ). This fact will play an important role in the convergence theory later.

The following proposition states a basic property of  $(P^j)$ .

**Proposition 4.3.** *Each subproblem  $(P^j)$  has a unique solution  $\hat{Y}^j$ .*

*Proof.* The existence and the uniqueness of a solution follows from the strong convexity of the objective function  $f_0^j$  on the compact set  $F^j$ . Furthermore, assumption (A4) (non-degeneracy constraint qualification) and Corollary 4.1 a & b imply that we can find a strictly feasible point of problem  $(P^j)$  and the assertion follows.  $\square$

Now we are able to present the basic algorithm for the solution of  $(P)$ :

*Algorithm 1.* Let asymptotes  $L$  and  $U$  feasible with Definition 3.3 and a constant  $\vartheta > 1$  be given.

- (0) Find  $Y^1 \in F$ .
- (1) Put  $j = 1$ .
- (2) Choose  $\bar{\tau} \geq \tau_1^j, \tau_2^j, \dots, \tau_m^j \geq \underline{\tau} > 0$ .
- (3) Solve problem  $(P^j)$ . Denote the solution by  $Y^+$ .
- (4) If  $f_\ell^j(Y^+) \geq f_\ell(Y^+)$  for all  $\ell = 0, 1, \dots, L$ , GOTO (6).

- (5) Put  $\tau_i^j \leftarrow \vartheta \tau_i^j$  for all  $i \in \{\ell \in \{0, 1, \dots, L\} \mid f_\ell^j(Y^+) < f_\ell(Y^+)\}$  and GOTO (3).
- (6)  $Y^{j+1} = Y^+$ .
- (7) If  $Y^{j+1}$  is stationary for problem (P), STOP; otherwise put  $j = j + 1$  and GOTO (2).

An appropriate update scheme for the parameters  $\tau_1^j, \tau_2^j, \dots, \tau_m^j$  (Step 2) will be proposed in Section 6, where we will also discuss algorithmic details as, for instance, a practical stopping criterion in Step 7. There we will further point out, how we carry out Step 0 above. For a detailed description of the algorithm applied to the solution of the subproblems arising from Step 3, we refer again to [27] and the references therein.

*Remark 4.4.* Algorithm 1 consists of *outer iterations* (Steps 2–7) and *inner iterations* (Step 3–5). The inner iterations replace the line search used in the original algorithm stated in [27]. An interpretation of the inner iterations is as follows: Whenever the condition in Step 4 fails to hold, we increase the influence of the strong convexity term. This results in a more conservative model. In a sense this is related to the trust region idea, which is a popular alternative to line search methods.

We state now the central convergence result for Algorithm 1:

**Theorem 4.5.** *Suppose that assumptions (A1)–(A3) are satisfied. Then, either Algorithm 1 stops at a stationary point of (P), or the sequence  $\{Y^j\}_j$  generated by Algorithm 1 has at least one accumulation point and each accumulation point is a stationary point of (P).*

In order to be able to prove the convergence theorem, we essentially follow the lines of the convergence proof in [30]. Many arguments carry directly over from [30]. Nevertheless we present all auxiliary results here in the semidefinite context for the sake of completeness. We start with the following lemma.

**Lemma 4.6.** *In each outer iteration only a finite number of inner iterations is required until the condition*

$$f_\ell^j(Y^+) \geq f_\ell(Y^+) \quad \text{for all } \ell = 0, 1, \dots, L \quad (4.1)$$

*is satisfied. Moreover the parameters  $\tau_i$  ( $i \in I$ ) remain bounded throughout all outer iterations.*

*Proof.* Using Theorem 3.4 c the assertion of the lemma follows directly from [30, Lemma 7.2 and Lemma 7.3].  $\square$

As a consequence of this lemma, inner iterations can be neglected in the remainder of the convergence proof.

Next we define the following auxiliary problem  $(P_{\tilde{Y}}^\tau)$ :

$$\begin{aligned} \min_{Y \in \mathbb{S}} f_0^{\tau}(Y) \\ \text{subject to} \\ f_\ell^{\tau}(Y) \leq 0, \quad \ell = 1, 2, \dots, L, \\ g_k(Y) \leq 0, \quad k = 1, 2, \dots, K, \\ \underline{Y}_i \preceq_{\mathbb{S}^{d_i}} Y_i \preceq_{\mathbb{S}^{d_i}} \bar{Y}_i, \quad i = 1, 2, \dots, m. \end{aligned} \tag{P}_{\tilde{Y}}^\tau$$

**Lemma 4.7.** *For each  $\tau \in \Theta := [\underline{\tau}, \bar{\tau}]^m$  the following equivalence holds true: A given point  $\tilde{Y}$  is stationary for problem  $(P)$  if and only if it is stationary for  $(P_{\tilde{Y}}^\tau)$ .*

*Proof.* The assertion of Lemma 4.7 follows directly from the first-order approximation properties of the model functions  $(f_\ell)_{\tilde{Y}}^\tau$  stated in Theorem 3.4 a & b.  $\square$

An immediate consequence of Lemma 4.7 is the following: If Algorithm 1 has a fixed point, i.e.,  $Y^{j+1} = Y^j$  then  $Y^j$  is stationary for problem  $(P)$ . Hence we can assume subsequently without loss of generality that

$$(A5) \quad Y^{j+1} \neq Y^j.$$

**Lemma 4.8.** *All iterates  $Y^j$ ,  $j > 1$  remain feasible. Moreover the sequence  $\{f_0(Y^j)\}_j$  is strictly decreasing.*

*Proof.* The first assertion is implied by condition (4.1) and Lemma 4.6. The strict monotonicity of the objective function follows from assumption (A5).  $\square$

Taking into consideration that all iterates are in the compact set defined by the matrix bounds, we conclude from Lemma 4.8 that the sequence  $\{Y^j\}_j$  generated by Algorithm 1 has at least one accumulation point. We denote this point by  $Y^*$ . Now we argue exactly as in [30]: We can find an infinite index set  $\tilde{K}$  such that

$$(C1) \quad Y^k \rightarrow Y^* \text{ as } k \in \tilde{K} \text{ and } k \rightarrow \infty.$$

By the same reasoning there exists a set  $K \subset \tilde{K}$  and a point  $\hat{Y}$  such that

$$(C2) \quad Y^{k+1} \rightarrow \hat{Y} \text{ as } k \in K \text{ and } k \rightarrow \infty.$$

**Lemma 4.9.** *The sequence of function values  $\{f_0(Y^j)\}_j$  converges to  $f_0(Y^*)$ .*

*Proof.* The sequence of function values is monotonically decreasing due to Lemma 4.8. Moreover it is bounded by the global minimum of problem  $(P)$  – cf. assumption (A4). Thus the sequence of function values converges to some  $f^*$ . Finally we see from assertion (C1) that  $f^* = f_0(Y^*)$ .  $\square$

Using the lemma above and the continuity of  $f_0$  we immediately find

$$f_0(Y^*) = f_0(\hat{Y}) \tag{4.2}$$

and conclude:

**Lemma 4.10.**  $\widehat{Y}$  is the unique solution of  $(P_{Y^*}^{\tau^*})$ , where  $\tau^*$  is an accumulation point of the bounded sequence  $\{\tau\}_j$ .

*Proof.* Without restriction of generality we assume that the sequence  $\{\tau\}_{j \in K}$  converges.  $Y^{j+1}$  is as optimal solution a feasible point of problem  $(P^j)$ . Now, letting  $k \in K$  and  $k \rightarrow \infty$  we see that  $\widehat{Y}$  is in the feasible domain of  $(P_{Y^*}^{\tau^*})$ . It remains to prove that  $f_{0Y^*}(\widehat{Y}) \leq f_{0Y^*}(\widetilde{Y})$  for an arbitrary feasible  $\widetilde{Y}$ . We argue by contradiction. Assume that there exists a  $\widetilde{Y}$  feasible for  $(P_{Y^*}^{\tau^*})$  such that  $f_{0Y^*}(\widetilde{Y}) < f_{0Y^*}(\widehat{Y})$ . Then by a continuity argument we can find a point  $\widetilde{Y}'$  in the interior of the feasible domain of  $(P_{Y^*}^{\tau^*})$ , which satisfies the inequality

$$f_{0Y^*}(\widetilde{Y}') < f_{0Y^*}(\widehat{Y}). \quad (4.3)$$

The fact that  $\widetilde{Y}'$  is strictly feasible guarantees the existence of a  $\bar{k} > 0$  such that  $\widetilde{Y}'$  is a feasible point of all  $(P^k)$  with  $k > \bar{k}$ . Now we conclude from Remark 4.2, from  $Y^{k+1} \xrightarrow{K} \widehat{Y}$ ,  $Y^k \xrightarrow{K} Y^*$ ,  $\tau^k \xrightarrow{K} \tau^*$  and (4.3) that we find an index  $k' \geq \bar{k}$  large enough such that  $f_{0Y^k}(\widetilde{Y}') < f_{0Y^k}(Y^{k+1})$ . But this contradicts the optimality of  $Y^{k+1}$  for  $(P^k)$ .  $\square$

**Lemma 4.11.**  $\widehat{Y} = Y^*$ .

*Proof.* By construction  $Y^*$  is in the feasible domain of  $(P_{Y^*}^{\tau^*})$ . From  $f_0(Y^{k+1}) \leq f_0^k(Y^{k+1})$  it follows by letting  $k \in K$  and  $k \rightarrow \infty$  that

$$f_0(\widehat{Y}) < f_{0Y^*}(\widehat{Y}). \quad (4.4)$$

By Lemma 4.7 we have  $f_0(Y^*) = f_{0Y^*}(Y^*)$ , but then (4.2) together with (4.4) shows that  $f_{0Y^*}(Y^*) \leq f_{0Y^*}(\widehat{Y})$ . Now the assertion follows directly from Lemma 4.10.  $\square$

Finally the assertion from Theorem 4.5 follows from Lemma 4.7 and Lemma 4.11.

## 5. Free material optimization

We briefly introduce the free material optimization problem:

Let  $\Omega \subset \mathbb{R}^2$  be a two-dimensional bounded domain<sup>1</sup> with a Lipschitz boundary. By  $u(x) = (u_1(x), u_2(x))$  we denote the displacement vector at a point  $x$  of the body under an external load, and by

$$e_{ij}(u(x)) = \frac{1}{2} \left( \frac{\partial u_i(x)}{\partial x_j} + \frac{\partial u_j(x)}{\partial x_i} \right) \quad \text{for } i, j = 1, 2$$

the associated (small-)strain tensor. We assume that our system is governed by linear Hooke's law, i.e., the stress is a linear function of the strain

$$\sigma_{ij}(x) = E_{ijkl}(x)e_{kl}(u(x)) \quad (\text{in tensor notation}),$$

---

<sup>1</sup>The entire presentation is given for two-dimensional bodies, to keep the notation simple. Analogously, all this can be done for three-dimensional solids.

where  $E$  is the elastic stiffness tensor. The symmetries of  $E$  allow us to write the 2<sup>nd</sup> order tensors  $e$  and  $\sigma$  as vectors

$$e = (e_{11}, e_{22}, \sqrt{2}e_{12})^T \in \mathbb{R}^3, \quad \sigma = (\sigma_{11}, \sigma_{22}, \sqrt{2}\sigma_{12})^T \in \mathbb{R}^3.$$

Correspondingly, the 4<sup>th</sup> order tensor  $E$  can be written as a symmetric  $3 \times 3$  matrix

$$E = \begin{pmatrix} E_{1111} & E_{1122} & \sqrt{2}E_{1112} \\ E_{2222} & \sqrt{2}E_{2212} & \\ \text{sym.} & & 2E_{1212} \end{pmatrix}. \quad (5.1)$$

In this notation, Hooke's law reads as  $\sigma(x) = E(x)e(u(x))$ .

Given an external load function  $f \in [L_2(\Gamma)]^2$ , where  $\Gamma$  is a part of  $\partial\Omega$  that is not fixed by Dirichlet boundary conditions, we are able to state a basic boundary value problem of the type:

$$\text{Find } u \in [H^1(\Omega)]^2, \text{ such that} \quad (5.2)$$

$$\begin{aligned} -\text{div}(\sigma) &= 0 && \text{in } \Omega \\ \sigma \cdot n &= f && \text{on } \Gamma \\ u &= 0 && \text{on } \Gamma_0 \\ \sigma &= E \cdot e(u) && \text{in } \Omega. \end{aligned}$$

Here  $\Gamma$  and  $\Gamma_0$  are open disjunctive subsets of  $\partial\Omega$ . Applying Green's formula, we obtain the weak equilibrium equation

$$\text{Find } u \in V, \text{ such that} \quad (5.3)$$

$$\int_{\Omega} \langle E(x)e(u(x)), e(v(x)) \rangle dx = \int_{\Gamma} f(x) \cdot v(x) dx, \quad \forall v \in V,$$

where  $V = \{u \in [H^1(\Omega)]^2 \mid u = 0 \text{ on } \Gamma_0\} \supset [H_0^1(\Omega)]^2$  reflects the Dirichlet boundary conditions.

In *free material optimization* (FMO), the design variable is the elastic stiffness tensor  $E$  which is a function of the space variable  $x$  (see [5, 26]). The only constraints on  $E$  are that it is physically reasonable, i.e., that  $E$  is symmetric and positive semidefinite. This gives rise to the following definition

$$E_0 := \{E \in L^{\infty}(\Omega)^{3 \times 3} \mid E = E^{\top}, E \succeq \underline{\rho}I \text{ a.e. in } \Omega\},$$

where  $\underline{\rho} \in \mathbb{R}^+$  is a suitable non-negative number and  $I$  denotes the identity matrix. The choice of  $L^{\infty}$  is due to the fact that we allow for maximal-material/minimal-material situations. A frequently used measure for the stiffness of the material tensor is its trace. In order to avoid arbitrarily stiff material, we add pointwise stiffness restrictions of the form  $\text{Tr}(E) \leq \bar{\rho}$ , where  $\bar{\rho}$  is a finite real number. Accordingly, we define the *set of admissible materials* as

$$E := \{E \in L^{\infty}(\Omega)^{3 \times 3} \mid E = E^{\top}, E \succeq \underline{\rho}I, \text{Tr}(E) \leq \bar{\rho} \text{ a.e. in } \Omega\}.$$

Now we are able to present the basic FMO problem:

$$\inf_{\substack{u \in V, \\ E \in \bar{E}}} \int_{\Gamma} f(x) \cdot u(x) dx \quad (5.4)$$

subject to

$u$  solves the equilibrium equation (5.3),

$$v(E) \leq \bar{v}.$$

Here the volume  $v(E)$  is defined as  $\int_{\Omega} \text{Tr}(E) dx$  and  $\bar{v} \in \mathbb{R}$  is an upper bound on overall resources. Moreover, the objective, the so-called compliance functional, measures how well the structure can carry the load  $f$ . More details about the infinite-dimensional problems are given in [2, 35].

Next we show, how we can incorporate so-called displacement constraints in the problem formulation: Given a set of functions  $d^k \in [L_2(\Gamma)]^2$  ( $k = 1, 2, \dots, n_d$ ), we define *linear displacement constraints* as

$$\int_{\Gamma} d^k(x) \cdot u(x) dx \leq r^k, \quad k = 1, 2, \dots, n_d,$$

where  $r_1, r_2, \dots, r^{n_d}$  are real numbers. Then we arrive at the following multidisciplinary FMO problem:

$$\inf_{\substack{u \in V, \\ E \in \bar{E}}} \int_{\Gamma} f(x) \cdot u(x) dx \quad (5.5)$$

subject to

$u$  solves the equilibrium equation (5.3),

$$\int_{\Gamma} d^k(x) \cdot u(x) dx \leq r^k, \quad k = 1, 2, \dots, n_d,$$

$$v(E) \leq \bar{v}.$$

Important questions like existence of an optimal solution of this generalized FMO problem are discussed in a forthcoming paper by the authors. In order to solve problem (5.5) numerically, we use the finite element scheme described in [27], which is based on the discretization schemes used in [2, 35]. After discretization, problem (5.5) becomes

$$\min_{E \in \tilde{E}} f^{\top} A^{-1}(E) f \quad (5.6)$$

subject to

$$(d^k)^{\top} A^{-1}(E) f \leq r^k, \quad k = 1, 2, \dots, n_d, \quad (5.7)$$

$$\sum_{i=1}^m \text{Tr}(E_i) \leq V,$$

where  $\tilde{E}$  is given as

$$\tilde{E} = \left\{ E \in (\mathbb{S}^3)^m \mid E_i \succeq \underline{\rho} I, \text{Tr}(E_i) \leq \bar{\rho}, i = 1, \dots, m \right\}, \quad (5.8)$$

$V$  is a discrete upper bound on resources and  $f, d^k$  ( $k = 1, 2, \dots, n_d$ ) are coefficient vectors in  $\mathbb{R}^N$  associated with a finite element basis  $\vartheta_1, \vartheta_2, \dots, \vartheta_N$  (note that the discrete vectors are denoted by the same symbols as their infinite-dimensional counterparts). Moreover  $A$  is the so-called global stiffness matrix associated with the material  $E$ .

*Remark 5.1.* In [27] the authors have shown that the objective function given in problem (5.6) is convex, smooth and has a dense Hessian. While the displacement constraints (5.7) are smooth and lead to dense Hessians as well, they are in general not convex.

*Remark 5.2.* Note that it is possible to work with several independently acting load cases in all FMO problems above. For simplicity and because the multiple load case is already extensively studied in [27] we restrict ourselves to single load case scenarios in this article.

## 6. Numerical experiments

Before we report on results of our numerical experiments with FMO problems described in the previous section we want to present some algorithmic details:

### 6.1. Algorithmic details

**The choice of the asymptotes.** As already mentioned in Section 3 we use fixed asymptotes. The following choice turned out to be robust:

$$L_i = 0, \quad U_i = 1.1\bar{\rho}\text{Id},$$

where  $\text{Id}$  is the identity matrix in  $\mathbb{S}^3$  and  $\mathbb{S}^6$  for 2D- and 3D-problems, respectively.

**The subproblems.** During all iterations, we solve the subproblems approximately. We use the following strategy: we start with a moderate accuracy of  $\varepsilon = 10^{-3}$  for the KKT error of the subproblem. During the outer iterations we adjust the tolerance according to the current KKT error of the master problem.

**The choice of  $\tau$ .** The parameters  $\tau_i^j$  ( $i \in I$ ) in the  $j$ th outer iteration are initialized such that the following condition is valid:

$$-\nabla^i f_\ell(E^j) + \tau_i^j I \succeq \delta I \quad (i \in I)$$

for all  $i \in I$  and all  $\ell = 0, 1, \dots, L$ . A typical choice for  $\delta$  is  $10^{-4}$ . The constant update factor  $\vartheta$  used in step 4 of Algorithm 1 is typically chosen from the interval  $[2, 10]$ . For a more sophisticated update scheme we refer to [30].

**A practical stopping criterion.** We use two stopping criteria for Algorithm 1. The first one is based on the relative difference of two successive objective function values. We consider this stopping criterion as achieved if the relative difference falls below some given threshold  $\epsilon_1$  (typically  $\epsilon_1 = 10^{-8}$ ). The second stopping criterion is based on the following KKT-related error measures:

$$\begin{aligned}\text{err}_1 &= \left\| \nabla L(Y^l, y^l, u^l, \underline{U}^l, \bar{U}^l) \right\|, \\ \text{err}_2 &= \max\{f_\ell(Y^l), g_k(Y^l) \mid \ell = 1, 2, \dots, L, k = 1, 2, \dots, K\}, \\ \text{err}_3 &= \max \left\{ |y_\ell^l f_\ell(Y^l)|, |u_k^l g_k(Y^l)|, |\langle \underline{U}_j^l, Y_j^l - \underline{Y}_j \rangle|, |\langle \bar{U}_j^l, \bar{Y}_j - Y_j^l \rangle| \right. \\ &\quad \left. \ell = 1, \dots, L, k = 1, \dots, K, j = 1, \dots, m \right\},\end{aligned}$$

where  $Y^l$  is the approximate solution at iterate  $l$ ,  $L$  is the Lagrangian associated with problem  $(P)$  and  $y^l, u^l, \underline{U}^l$  and  $\bar{U}^l$  are the corresponding vectors of Lagrangian (matrix) multipliers associated with the constraint functions  $f_\ell, g_k$  and the lower and upper matrix bound constraints, respectively. Recall that the feasibility of  $Y^l$  w.r.t. the matrix bound constraints is maintained throughout all iterations. Now we define our second stopping criterion as

$$\frac{1}{3} \sum_{i=1}^3 \text{err}_i \leq \epsilon_2, \quad (6.1)$$

where a typical value for  $\epsilon_2$  is  $5 \cdot 10^{-5}$ . Note that we only stop when both stopping criteria are satisfied simultaneously.

**How to find a feasible point?** In general it is known that finding a feasible point of a non-convex optimization problem may be as hard as solving the optimization problem itself. It turns out however that the following strategy inspired by [38] works well for our purposes: If no feasible point is known, we start by solving the following auxiliary problem:

$$\begin{aligned}\min_{Y \in \mathbb{S}} f_0^j(Y) + \sum_{\ell=1, \dots, L} \eta_\ell f_\ell^j(Y) \\ \text{subject to} \\ g_k(Y) \leq 0, \quad k = 1, 2, \dots, K, \\ \underline{Y}_i \preceq_{\mathbb{S}^{d_i}} Y_i \preceq_{\mathbb{S}^{d_i}} \bar{Y}_i, \quad i = 1, 2, \dots, m.\end{aligned}$$

Here the parameters  $\eta_\ell$ ,  $\ell = 1, 2, \dots, L$  are penalty parameters, which are increased until a feasible solution is identified.

**The code.** We have implemented the new algorithm in the C programming language. In what follows we refer to the resulting code as PENSCP. All FMO and finite element computations have been carried out by a prototype version of the software platform PLATO-N.

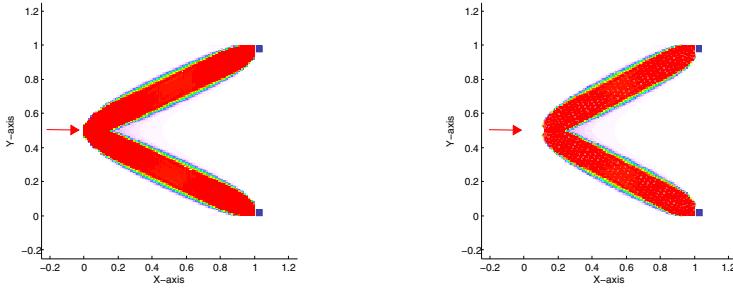


FIGURE 1. Test case 1: no displacement constraint applied; density plot (left) / deformed density plot (right)

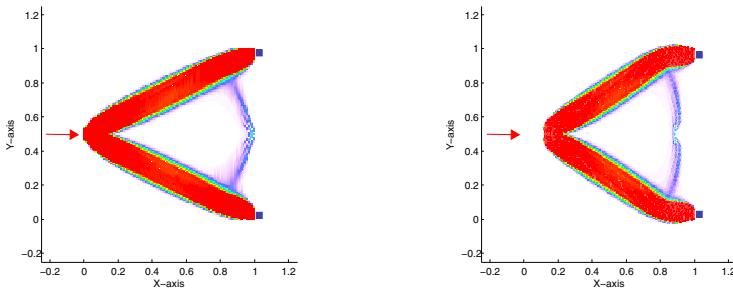


FIGURE 2. Test case 1: including displacement constraint; density plot (left) / deformed density plot (right)

## 6.2. Numerical studies with displacement constrained FMO problems

**2D examples.** The first test case we consider consists of a rectangular domain which is clamped on the corner elements of the right boundary and loaded in the center of the left boundary edge. The design space is discretized by 14.400 quadrilateral elements. We first solved this problem without displacement constraints. PENSCLP needed 256 iterations in order to solve the problem. The optimal density results are depicted in Figure 1.

Next we solve the same problem including a displacement constraint, which forces the center of the right boundary edge to move to the left. The additionally constrained problem was solved in 273 iterations. The resulting density distribution is shown in Figure 2.

The compliance of the modified problem is only larger by approximately 0.5 per cent compared to the original problem. This is because not much additional

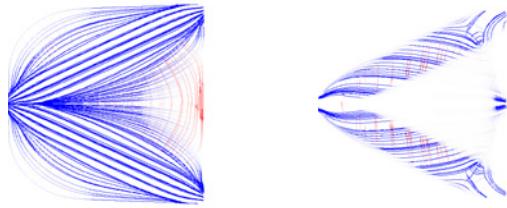


FIGURE 3. Test case 1: principal strain directions; no displacement constraint (left) / with displacement constraint (right)

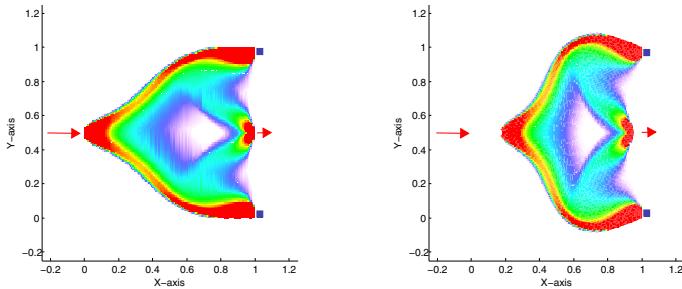


FIGURE 4. Test case 1: additional load & displacement constraint; density plot (left) / deformed density plot (right)

material is needed to move a point, where originally the material is on the lower bound. As a consequence the resulting density plots differ only slightly. In Figure 3 we plot the principal strain directions for both problems. here the difference is already much more significant.

A well-known trick in design of mechanisms which avoids the difficulties described above leading to more significant changes in the design is to prescribe some material in the area which should be moved by the displacement constraint. Alternatively one can apply a small force. Using the latter idea, we obtain the results displayed in Figure 4. This time the solution was found in 361 iterations. The compliance is (of course) much higher now – by about 50 per cent.

In our second test case we load a horizontal bar by vertical forces in the left and right upper corner of the design domain. At the same time the bar is supported from below as indicated in Figure 5. The bar is discretized by 18.000 elements. Solving the problem without displacement constraint (310 iterations) one observes that the complete bar bends over the support. In particular, the central part of the upper boundary moves up.

This time the displacement constraint should invert this effect, i.e., the central part of the upper boundary should move down. The computed density results

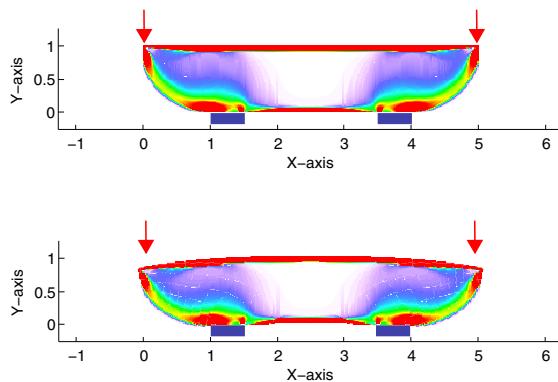


FIGURE 5. Test case 2: no displacement constraint; density plot (top) / deformed density plot (bottom)

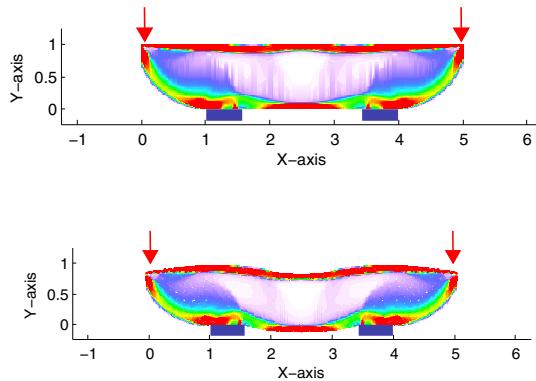


FIGURE 6. Test case 2: displacement constraint applied; density plot (top) / deformed density plot (bottom)

(achieved after 327 iterations) can be seen in Figure 6. Figure 7 provides a comparison of the principal strain directions. The increase of the compliance functional in this example was about 16 per cent.

**3D examples.** In our third test case we want to design a cube, which is loaded in the center of its top surface and clamped in the neighborhood of the corners of the bottom surface. We discretize the cube by approximately 10.000 Hexa elements. The algorithm stops after 815 iterations yielding the density result shown in Figure 8.

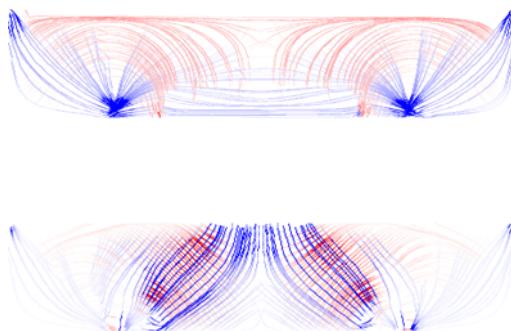


FIGURE 7. Test case 2: principal strain directions; no displacement constraint (top) / with displacement constraint (bottom)

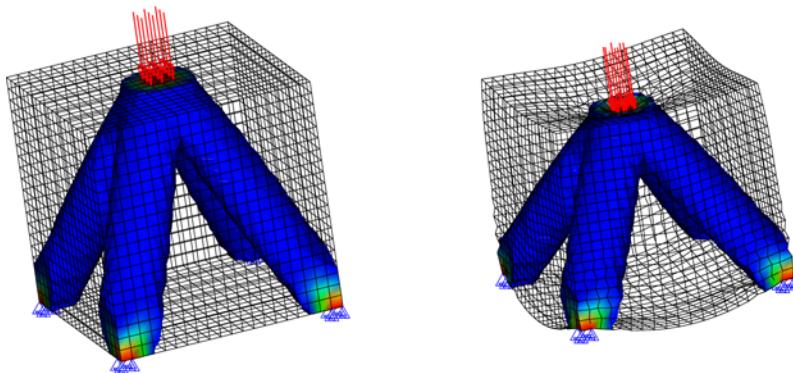


FIGURE 8. Test case 3: no displacement constraint; density plot (left) / deformed density plot (right)

Similar as in test case 1, we now apply a displacement constraint, which should move the area around the center of the bottom face upwards. We stopped the algorithm close to the required precision after 1000 iterations. Figure 9 demonstrates the effect of the displacement constraint. Note that similar as in our first experiment, we added a small force to the center node in the bottom. We suppose that the large number of iterations is related to a poor scaling of the problem.

Our last test case is the 3D counterpart of test case 2. We used a discretization of approximately 8.000 finite elements in this case. Here the original problem converges after 463 iteration, while after adding the displacement constraint 527

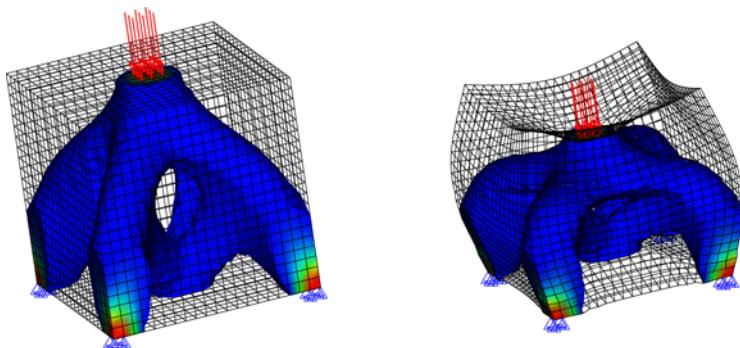


FIGURE 9. Test case 3: displacement constraint applied; density plot (left) / deformed density plot (right)

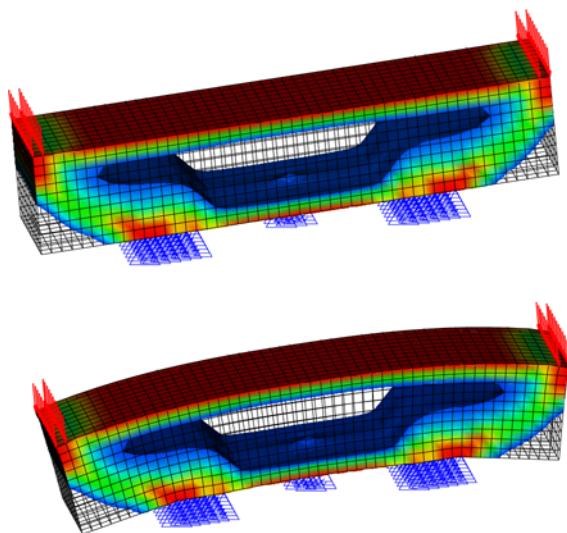


FIGURE 10. Test case 4: no displacement constraint; density plot (top) / deformed density plot (bottom)

iterations are required. Figures 10 and 11 compare the results obtained in both cases. In our last example the compliance increased by 8 per cent.

*Remark 6.1.* We note that in all examples a feasible point was detected after at most five outer iterations. The associated penalty parameters took values between 1 and 100.

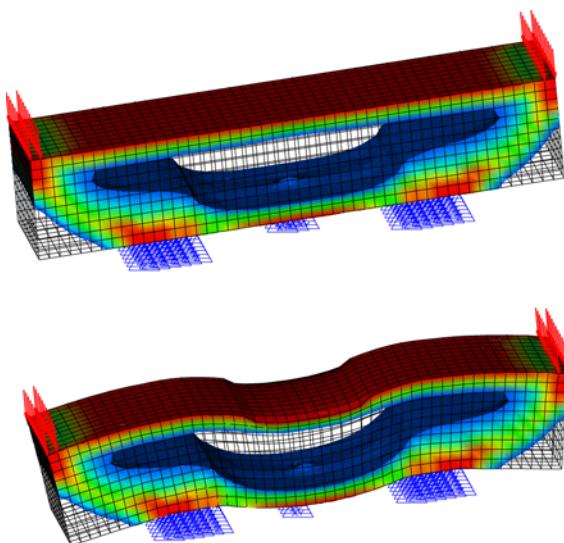


FIGURE 11. Test case 4: displacement constraint applied; density plot (top) / deformed density plot (bottom)

*Remark 6.2.* We want to remark that as alternative to problem (5.5) one could minimize the volume functional subject to displacement and compliance constraints. We tested this formulation on a few examples. The results were comparable with the only difference that the computation time was significantly higher due to more iterations required by the sub solver.

## References

- [1] F. Alizadeh, J. Eckstein, N. Noyan, and G. Rudolf. Arrival rate approximation by nonnegative cubic splines. *Operations Research*, 56:140–156, 2008.
- [2] A. Ben-Tal, M. Kočvara, A. Nemirovski, and J. Zowe. Free material design via semi-definite programming. The multi-load case with contact conditions. *SIAM J. Optimization*, 9:813–832, 1997.
- [3] A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization*. MPS-SIAM Series on Optimization. SIAM Philadelphia, 2001.
- [4] M. Bendsøe and O. Sigmund. *Topology Optimization. Theory, Methods and Applications*. Springer-Verlag, Heidelberg, 2002.
- [5] M.P. Bendsøe, J. M. Guades, R.B. Haber, P. Pedersen, and J.E. Taylor. An analytical model to predict optimal material properties in the context of optimal structural design. *J. Applied Mechanics*, 61:930–937, 1994.
- [6] K.U. Bletzinger. Extended method of moving asymptotes based on second-order information. *Struct. Multidiscip. Optim.*, 5 (3):175–183, 1993.

- [7] F.J. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*. Springer-Verlag New York, 2000.
- [8] B. Borchers. A c library for semidefinite programming. *Optimization Methods and Software*, 11(1):613–623, 1999.
- [9] R. Correa and C.H. Ramirez. A global algorithm for nonlinear semidefinite programming. *SIAM J. Optim.*, 15:1791–1820, 2004.
- [10] E. de Klerk, D.V. Pasechnik, and A. Schrijver. Reduction of symmetric semidefinite programs using the regular\*-representation. *Mathematical Programming*, 109(2/3), 2007.
- [11] B. Fares, D. Noll, and P. Apkarian. Robust control via sequential semidefinite programming. *SIAM Journal on Control and Optimization*, 40(6):1791–1820, 2002.
- [12] C. Fleury. Conlin: An efficient dual optimizer based on convex approximation concepts. *Struct. Optim.*, 1:81–89, 1989.
- [13] C. Fleury. Efficient approximation concepts using second order information. *Int. J. Num. Meth. Engrg.*, 28:2041–2058, 1989.
- [14] K. Fujisawa, M. Kojima, and K. Nakata. Exploiting sparsity in primal-dual interior-point method for semidefinite programming. *Mathematical Programming*, 79:235–253, 1997.
- [15] L. Vanderberghe H. Wolkowicz, R. Saigal. *Handbook of Semidefinite Programming – Theory, Algorithms and Applications*. Kluwer Academic Publishers, 2000.
- [16] F. Jarre. An interior method for nonconvex semidefinite programs. *Optimization and Engineering*, 1:347–372, 2000.
- [17] M. Kočvara, M. Stingl, and J. Zowe. Free material optimization: recent progress. *Optimization*, 57:79–100, 2008.
- [18] M. Kojima, S. Shindoh, and S. Hara. Interior-point methods for the monotone semi-definite linear complementarity problem in symmetric matrices. *SIAM Journal on Optimization*, 7:86–125, 1997.
- [19] M. Kočvara and M. Stingl. PENNON – a code for convex nonlinear and semidefinite programming. *Optimization Methods and Software*, 18(3):317–333, 2003.
- [20] M. Kočvara and M. Stingl. The worst-case multiple-load fmo problem revisited. In N. Olhoff M.P. Bendsoe and O. Sigmund, editors, *IUTAM-Symposium on Topological Design Optimization of Structures, Machines and Material: Status and Perspectives*, pages 403–411. Springer, 2006.
- [21] M. Kočvara and M. Stingl. Free material optimization: Towards the stress constraints. *Structural and Multidisciplinary Optimization*, 33(4-5):323–335, 2007.
- [22] M. Kočvara and M. Stingl. On the solution of large-scale sdp problems by the modified barrier method using iterative solvers. *Mathematical Programming (Series B)*, 109(2-3):413–444, 2007.
- [23] D. Noll. Local convergence of an augmented lagrangian method for matrix inequality constrained programming. *Optimization Methods and Software*, 22(5):777–802, 2007.
- [24] R. Polyak. Modified barrier functions: Theory and methods. *Math. Prog.*, 54:177–222, 1992.
- [25] J. Povh, F. Rendl, and A. Wiegele. A boundary point method to solve semidefinite programs. *Computing*, 78(3):277–286, 2006.

- [26] U. Ringertz. On finding the optimal distribution of material properties. *Structural Optimization*, 5:265–267, 1993.
- [27] M. Stingl and M. Kočvara G. Leugering. A sequential convex semidefinite programming algorithm with an application to multiple-load free material optimization. *SIAM J. Optimization*, accepted, 2008.
- [28] D. Sun, J. Sun, and L. Zhang. The rate of convergence of the augmented lagrangian method for nonlinear semidefinite programming. *Math. Program.*, 114(2):349–391, 2008.
- [29] K. Svanberg. The method of moving asymptotes – a new method for structural optimization. *International Journal for numerical methods in Engineering*, 24:359–373, 1987.
- [30] K. Svanberg. A class of globally convergent optimization methods based on conservative separable approximations. *SIOPT*, 12:555–573, 2002.
- [31] M.J. Todd. Semidefinite optimization. *Acta Numerica*, 10:515–560, 2001.
- [32] R.H. Tutuncu, K.C. Toh, and M.J. Todd. Solving semidefinite-quadratic-linear programs using sdpt3. *Mathematical Programming Ser. B*, 95:189–217, 2003.
- [33] L. Vanderberghe and S. Boyd. Semidefinite programming. *SIAM Rev.*, 38:49–95, 1996.
- [34] H. Waki, S. Kim, M. Kojima, and M. Muramatsu. Sums of squares and semidefinite programming relaxation for polynomial optimization problems with structured sparsity. *SIAM J. Optim.*, 17:218–242, 2006.
- [35] R. Werner. *Free Material Optimization*. PhD thesis, Institute of Applied Mathematics II, Friedrich-Alexander University of Erlangen-Nuremberg, 2000.
- [36] X. Zhao, D. Sun, and K.C. Toh. A Newton-CG augmented Lagrangian method for semidefinite programming. Optimization online,  
[http://www.optimization-online.org/db\\_html/2008/03/1930.html](http://www.optimization-online.org/db_html/2008/03/1930.html), 2008.
- [37] C. Zillober. *Eine global konvergente Methode zur Lösung von Problemen aus der Strukturoptimierung*. PhD thesis, Technische Universität München, 1992.
- [38] C. Zillober. Global convergence of a nonlinear programming method using convex approximations. *Numerical Algorithms*, 27(3):256–289, 2001.
- [39] J. Zowe, M. Kočvara, and M. Bendsøe. Free material optimization via mathematical programming. *Math. Prog., Series B*, 79:445–466, 1997.

M. Stingl and G. Leugering  
 Applied Mathematics II, Department of Mathematics  
 University of Erlangen  
 Martensstr. 3  
 D-91058 Erlangen, Germany  
 e-mail: [stingl@am.uni-erlangen.de](mailto:stingl@am.uni-erlangen.de)  
[leugering@am.uni-erlangen.de](mailto:leugering@am.uni-erlangen.de)

M. Kočvara  
 School of Mathematics, University of Birmingham  
 Birmingham B15 2TT, UK  
 e-mail: [kocvara@maths.bham.ac.uk](mailto:kocvara@maths.bham.ac.uk)

# How to Check Numerically the Sufficient Optimality Conditions for Infinite-dimensional Optimization Problems

Daniel Wachsmuth and Arnd Rösch

**Abstract.** We consider general non-convex optimal control problems. Many results for such problems rely on second-order sufficient optimality conditions. We propose a method to verify whether the second-order sufficient optimality conditions hold in a neighborhood of a numerical solution. This method is then applied to abstract optimal control problems. Finally, we consider an optimal control problem subject to a semi-linear elliptic equation that appears to have multiple local minima.

**Mathematics Subject Classification (2000).** Primary 49J20; Secondary 49K20, 65N15.

**Keywords.** Optimal control, sufficient optimality conditions, verification of optimality conditions, problems with multiple minima.

## 1. Introduction

Let us consider the following model problem. Let  $U$  be a Hilbert space. Denote by  $U_{\text{ad}}$  the space of admissible controls, where  $U_{\text{ad}}$  is a closed, convex and non-empty subset of  $U$ . In addition, let  $f : U \rightarrow \mathbb{R}$  be a twice continuously Fréchet-differentiable function. Then we are considering the problem

$$\min_{u \in U_{\text{ad}}} f(u). \quad (1.1)$$

The first-order necessary optimality condition for (1.1) reads as follows. Let  $\bar{u}$  be a local solution of (1.1). Then the variational inequality

$$f'(\bar{u})(u - \bar{u}) \geq 0 \quad \forall u \in U_{\text{ad}} \quad (1.2)$$

is satisfied. An equivalent characterization is given by the inclusion

$$-f'(\bar{u}) \in N_{U_{\text{ad}}}(\bar{u}),$$

where  $N_{U_{\text{ad}}}(\bar{u})$  denotes the normal cone of  $U_{\text{ad}}$  at  $\bar{u}$ . We will not consider second-order necessary conditions here, instead we refer to [3, 4].

A strong second-order sufficient optimality condition is satisfied at  $\bar{u}$  if the condition (1.2) as well as the coercivity property

$$f''(\bar{u})[v, v] \geq \alpha \|v\|_U^2 \quad \forall v \in U \quad (1.3)$$

hold for some  $\alpha > 0$ . This condition ensures that  $\bar{u}$  is locally optimal, and moreover, the quadratic growth condition holds: there are constants  $r, \delta > 0$  such that

$$f(u) \geq f(\bar{u}) + \delta \|u - \bar{u}\|_U^2 \quad \forall u \in U_{\text{ad}} : \|u - \bar{u}\|_U \leq r.$$

Using the well-known concept of strongly active sets, see, e.g., [3, 4, 8], the subspace, where  $f''$  has to be positive definite, can be confined.

If the second-order sufficient conditions hold at the local minimum  $\bar{u}$  one can prove several properties of the original optimization problem. At first, such a local solution is stable with respect to perturbations. That is, a small perturbation of the optimization problem leads only to a small perturbation in the solution. This stability is a major ingredient for convergence results, since one can interpret approximated problems as perturbations of the original one. This allows to prove local fast convergence of optimization methods (SQP, semi-smooth Newton) as well as convergence rates for finite-element discretizations of optimal control problems.

The importance of sufficient optimality conditions makes it desirable to verify whether these conditions are satisfied for a given problem. However, in condition (1.3) coercivity is assumed for the *unknown* solution. For finite-dimensional problems, one can compute eigenvalues of the Hessian matrix at some approximation of the solution since it is possible to compute this Hessian exactly. In infinite-dimensional problems, the computation of the second derivative is also prone to discretization errors. Hence, it is difficult to check whether the condition is fulfilled. This was the starting point for our investigations. We will propose a different condition, which is in fact a condition at a given approximation  $\bar{u}_h$ . Since only known quantities are involved, there is a chance to check this condition. For the details, we refer to Section 2. We have to admit that we can only deal with problems without two-norm discrepancy. The two-norm discrepancy occurs, whenever the ingredients of the problem are differentiable with respect to a smaller space (say  $L^\infty$ ) and stronger norm, and coercivity of the second derivative only holds with respect to weaker norms (say  $L^2$ ), see for instance [5].

The numerical solution of optimization problems in function spaces is often done by discretization. Let  $U_h$  be a finite-dimensional subspace of  $U$  with basis  $\phi_h^1 \dots \phi_h^{N_h}$ . Then an example discrete problem, which hopefully can be solved on a computer, reads as

$$\min_{u_h \in U_{\text{ad}} \cap U_h} f(u_h).$$

Given a discrete solution  $\bar{u}_h$ , one can introduce the discrete Hessian matrix associated with the discrete problem by

$$H = (h_{ij})_{i,j=1}^{N_h}, \quad h_{i,j} = f''(\bar{u}_h)[\phi_h^i, \phi_h^j].$$

Then one can compute the eigenvalues of  $H$  and check positive definiteness of  $H$ . If  $H$  is positive definite at  $\bar{u}_h$  then  $\bar{u}_h$  is a local minimum of the *discrete* problem. However, it may happen that  $\bar{u}_h$  is not even close to a local solution of the original problem. Hence, the information in  $H$  is almost worthless in this case. We will present an example in Section 3, where exactly this situation occurs.

In Section 4, we will extend our approach to optimal control problems with partial differential equation. Here, we have in mind the following optimization problem

$$\min g(y) + j(u)$$

subject to

$$\begin{aligned} Ay + d(y) &= Bu, \\ u &\in U_{\text{ad}}. \end{aligned}$$

Here, a large class of semilinear elliptic state equation are covered by the analysis. In particular, steady-state Navier-Stokes equation are included. However, the differentiability requirements and the coercivity assumption are formulated with respect to the same spaces and norms. That is, problems with two-norm discrepancy are not covered.

In the article [12], the authors already suggested conditions for the numerical verification of optimality conditions. However, the analysis relied heavily on  $H^2$ -regularity of the solutions. We will overcome this restriction using a different approach for the treatment of the discretization errors.

The plan of the article is as follows. The verifiable condition is developed in Section 2. In Section 3, we introduce an example that shows that the computation of eigenvalues of the discrete Hessian cannot be taken as substitute for the condition of Section 2. The analysis concerned with optimal control problems for a semilinear elliptic equation is done in Section 4. We end the article with a report about an optimal control problem that admits two local solutions, see Section 5.

## 2. Coercivity condition for nonlinear programming

Let  $u_h \in U_{\text{ad}} \cap U_h$  be an arbitrary, admissible point. Ideally,  $u_h$  would be the solution of a discretized problem or an approximation of it as the outcome of some iterative method. But we will not rely on this property, which is a major improvement over [12].

Now, let us present the coercivity condition. At first, we assume that we can find bounds of certain characteristics of  $f'$  and  $f''$ .

**Assumption 2.1.** There are constants  $\epsilon, \alpha, M, R$  such that the following three inequalities hold:

$$f'(u_h)(u - u_h) \geq -\epsilon \|u - u_h\|_U \quad \forall u \in U_{\text{ad}}, \quad (2.1)$$

$$f''(u_h)[v, v] \geq \alpha \|v\|_U^2 \quad \forall v \in U, \quad (2.2)$$

$$|(f''(u) - f''(u_h))[v_1, v_2]| \leq M \|u - u_h\|_U \|v_1\|_U \|v_2\|_U \quad \forall u \in U_{\text{ad}}, \quad (2.3)$$

$$\begin{aligned} \|u - u_h\|_U &\leq R, \\ v_1, v_2 &\in U. \end{aligned}$$

Let us comment on the three inequalities involved in the assumption. The first one (2.1) measures in some sense the residuum in the variational inequality (1.2). The second inequality is a coercivity assumption on  $f''$  at  $u_h$ . The essential difference to (1.3) is that the point, where we have to check for coercivity of  $f''$ , is known.

Moreover, these conditions are analogous to the pre-requisites of convergence theorems of Newton's method: smallness of initial residual, bounded invertibility, and local Lipschitz estimates. See also the comments below.

Let us now take another admissible point  $u \in U_{\text{ad}}$ . With the help of Assumption 2.1, we can estimate the difference between  $f(u)$  and  $f(u_h)$  using Taylor expansion as

$$\begin{aligned} f(u) - f(u_h) &\geq f'(u_h)(u - u_h) + \frac{1}{2}f''(u_h)(u - u_h)^2 \\ &\quad + \int_0^1 \int_0^s (f''(u_h + t(u - u_h)) - f''(u_h))(u - u_h)^2 dt ds \quad (2.4) \\ &\geq -\epsilon \|u - u_h\|_U + \frac{\alpha}{2} \|u - u_h\|_U^2 - \frac{M}{6} \|u - u_h\|_U^3. \end{aligned}$$

In addition to Assumption 2.1, we need a further qualification, which relates the constants appearing there to each other.

**Assumption 2.2.** There exists a real number  $r_+$  with  $R > r_+ > 0$  such that

$$-\epsilon r_+ + \frac{\alpha}{2} r_+^2 - \frac{M}{6} r_+^3 > 0, \quad (2.5)$$

$$\alpha - Mr_+ > 0 \quad (2.6)$$

is satisfied.

Assumptions 2.1 and 2.2 allow us to prove the main result of this section.

**Theorem 2.3.** Let Assumptions 2.1 and 2.2 be satisfied. Then there exists a local solution  $\bar{u}$  of the original problem (1.1) with

$$\|\bar{u} - u_h\|_U < r_+.$$

Furthermore, the second-order sufficient optimality condition (1.3) is satisfied at  $\bar{u}$ .

*Proof.* Let us consider the optimization problem (1.1) but restricted to the closed ball centered at  $u_h$  with radius  $r_+$ ,

$$\min_{u \in \bar{B}(u_h, r_+) \cap U_{\text{ad}}} f(u).$$

Due to (2.6), the function  $f$  is convex on  $\bar{B}(u_h, r_+)$ . Hence, the auxiliary problem admits a global minimum  $\bar{u}$  with  $\|\bar{u} - u_h\|_U \leq r_+$ . Moreover, the second derivative of  $f$  is positive definite at  $\bar{u}$  by (2.2) and (2.6).

By (2.5),  $\bar{u}$  cannot lie on the boundary of  $B(u_h, r_+)$ , since the value of  $f$  is there larger than in  $u_h$ . That is,  $\bar{u}$  is also the global minimum of  $f$  over the intersection of  $U_{\text{ad}}$  with the open ball at  $u_h$  with radius  $r_+$ . Hence,  $\bar{u}$  is a local solution of the original problem (1.1).  $\square$

The consequences of this result are threefold: at first we obtain existence of a solution of the original problem in the specified neighborhood. Secondly, we can estimate the distance to the solution. And third, we can prove that this yet unknown solution fulfills the second-order optimality condition.

The inequality (2.5) is an assumption on the objective functional. We can replace it by an assumption on the first derivative  $f'$ , and can prove an result analogous to Theorem 2.3.

**Assumption 2.4.** There exists a real number  $\tilde{r}_+$  with  $R > \tilde{r}_+ > 0$  such that

$$-\epsilon + \alpha \tilde{r}_+ - \frac{M}{2} \tilde{r}_+^2 > 0 \quad (2.7)$$

is satisfied.

**Theorem 2.5.** *Let the assumptions 2.1 and 2.4 be satisfied. Then there exists a local solution  $\bar{u}$  of the original problem (1.1) with*

$$\|\bar{u} - u_h\|_U < \tilde{r}_+.$$

*Furthermore, the second-order sufficient optimality condition (1.3) is satisfied.*

*Proof.* At first, we have to show that Assumption 2.4 implies the convexity of  $f$  is a neighborhood of  $u_h$ . Let us define a polynomial  $p$  by  $p(r) = -\epsilon + \alpha \tilde{r} - \frac{M}{2} \tilde{r}^2$ . We already know  $p(0) < 0$  and  $p(\tilde{r}_+) > 0$ . Hence, there is a  $\tilde{r}_0 \in (0, \tilde{r}_+)$  such that  $p(\tilde{r}_0) = 0$ . The root  $\tilde{r}_0$  is given by  $\tilde{r}_0 = \frac{\alpha}{M} \left( 1 - \sqrt{1 - \frac{2M\epsilon}{\alpha^2}} \right)$ . Moreover, it holds  $\alpha - M\tilde{r}_0 = \alpha \sqrt{1 - \frac{2M\epsilon}{\alpha^2}} > 0$ . Hence, there is a  $\tilde{r}_1 \in (\tilde{r}_0, \tilde{r}_+)$  such that (2.6) and (2.7) are satisfied for  $\tilde{r}_1$ . This implies the convexity of  $f$  on the ball centered at  $u_h$  with radius  $\tilde{r}_1$ .

As in the proof of the previous Theorem 2.3, we obtain then the existence of a global solution  $\bar{u}$  of the problem

$$\min_{u \in \bar{B}(u_h, \tilde{r}_1) \cap U_{\text{ad}}} f(u).$$

It remains to show that  $\bar{u}$  is not on the boundary of  $\bar{B}(u_h, \tilde{r}_1)$ . Let us take an arbitrary  $u \in U_{\text{ad}}$  with  $\|u - u_h\|_U = \tilde{r}_1$ . Using (2.7), we obtain

$$\begin{aligned} f'(u)(u_h - u) &= f'(u_h)(u_h - u) + f''(u_h)(u_h - u, u - u_h) \\ &\quad + \int_0^1 (f''(u_h + t(u - u_h)) - f''(u_h))(u_h - u, u - u_h) dt \\ &\leq \left( \epsilon - \alpha \|u - u_h\|_U + \frac{M}{2} \|u - u_h\|_U^2 \right) \|u - u_h\|_U < 0. \end{aligned}$$

Hence, the necessary optimality conditions of (1.2) are not fulfilled for any control on the boundary of  $\bar{B}(u_h, \tilde{r}_1)$ . Thus, the solution  $\bar{u}$  satisfies  $\|\bar{u} - u_h\|_U < \tilde{r}_1 \leq \tilde{r}_+$ . Furthermore, it is a local solution of the original problem (1.1).  $\square$

A close inspection of the proof reveals that we can show an improved error estimate:

**Corollary 2.6.** *Let the assumptions of the previous theorem be fulfilled. Then it holds*

$$\|\bar{u} - u_h\|_U \leq \frac{\alpha}{M} \left( 1 - \sqrt{1 - \frac{2M\epsilon}{\alpha^2}} \right).$$

*Proof.* If Assumption 2.4 is satisfied with some  $\tilde{r}_+$ , then it will be satisfied for all  $r$  between  $\frac{\alpha}{M} \left( 1 - \sqrt{1 - \frac{2M\epsilon}{\alpha^2}} \right)$ , which is the first root of the polynomial in (2.7), and  $\tilde{r}_+$ . Then Theorem 2.5 yields the claim.  $\square$

Theorems 2.3 and 2.5 state that the yet unknown solution  $\bar{u}$  satisfies the second-order sufficient optimality condition. This implies that it is possible apply deeper results, which rely on this conditions. For instance, we can apply results for the fast local convergence of optimization methods. That is, if the initial guess is close enough to the solution then the iterates will converge with a high convergence rate towards the solution  $\bar{u}$ .

Let us show exemplarily the fast convergence of Newton's method for generalized equations in the sense of [1, 7] if started at  $u_h$ . The key idea here is to write the variational inequality (1.2) as the inclusion

$$-f'(\bar{u}) \in N_{U_{\text{ad}}}(u).$$

Then the generalized Newton method solves for  $u_{k+1}$  the problem

$$-(f''(u_k)(u - u_k) + f'(u_k)) \in N_{U_{\text{ad}}}(u), \quad (2.8)$$

which is the first-order necessary optimality condition of

$$\min_{u \in U_{\text{ad}}} \frac{1}{2} f''(u_k)(u - u_k)^2 + f'(u_k)(u - u_k).$$

That is, only the objective function is linearized but not the constraint  $u \in U_{\text{ad}}$ . It turns out, that the conditions of Theorem 2.5 are sufficient to ensure local quadratic convergence of the simple iteration (2.8) for the initial choice  $u_0 = u_h$ .

**Theorem 2.7.** *Let the assumptions of Theorem 2.5 be satisfied. Set  $u_0 := u_h$ . Then the sequence of iterates generated by the procedure (2.8) converges quadratically to  $\bar{u}$ .*

*Proof.* Let us assume first, that the equation (2.8) is solvable for some  $k$ . The iterate  $u_{k+1}$  satisfies the variational inequality

$$(f''(u_k)(u_{k+1} - u_k) + f'(u_k), u - u_{k+1}) \geq 0 \quad \forall u \in U_{\text{ad}}. \quad (2.9)$$

Setting  $k = 0$ ,  $u_0 = u = u_h$ , and using (2.1), we obtain the following estimate of the initial step

$$\|u_1 - u_0\|_U \leq \frac{\epsilon}{\alpha}.$$

Let us denote by  $\alpha_k$  the smallest eigenvalue of  $f''(u_k)$ . Then it holds  $\alpha_0 = \alpha$  and  $\alpha_{k+1} \geq \alpha_k - M\|u_k - u_{k+1}\|_U$ . Setting  $u := u_k$  in (2.9), we find applying (2.9) for  $u_k$

$$\begin{aligned} f''(u_k)(u_k - u_{k+1})^2 &\leq f'(u_k)(u_k - u_{k+1}) \\ &= (f'(u_{k-1}) + f''(u_{k-1})(u_k - u_{k-1}), u_k - u_{k+1}) \\ &\quad + \int_0^1 (f''(u_{k-1} + t(u_k - u_{k-1})) - f''(u_{k-1}))(u_k - u_{k-1}, u_k - u_{k+1}) dt. \end{aligned}$$

Setting  $k - 1$  for  $k$  in (2.9), we find the optimality relation for  $u_k$ , which implies that the first part of the right-hand side is non-positive. Applying Assumption 2.1, we obtain

$$f''(u_k)(u_k - u_{k+1})^2 \leq \frac{M}{2}\|u_k - u_{k-1}\|_U^2\|u_k - u_{k+1}\|_U,$$

which gives

$$\|u_k - u_{k+1}\|_U \leq \frac{M}{2\alpha_k}\|u_k - u_{k-1}\|_U^2 \leq \frac{M}{2(\alpha_{k-1} - M\|u_k - u_{k-1}\|_U)}\|u_k - u_{k-1}\|_U^2.$$

Now, we can proceed as in Ortega's proof of the Newton-Kantorovich theorem [10], see also [6]. The technique applied there delivers (a) existence of solutions of (2.9) for all  $k$ , and (b) quadratic convergence. Moreover, the convergence region of Newton's method given by [10] is the ball at  $u_h$  with the radius given by Corollary 2.6.  $\square$

The similarities to the convergence proof of Newton-Kantorovich type is obvious. That is, the assumptions above can be interpreted as assumptions in the context of Newton's method and vice-versa. This observation allows also to apply heuristic techniques to estimate the constants appearing in Assumption 2.1 during the procedure of Newton's method. For a detailed explanation of these techniques we refer to the monograph of Deuflhard [6].

### 3. On coercivity of $f''$ and positive definiteness of the discrete Hessian

The computation of the constants appearing in the Assumption 2.1 above is a difficult task, it is especially hard to find a lower bound for the smallest eigenvalue  $\alpha$  of  $f''$ . Here, it would be advantageous if one could compute  $\alpha$  as an eigenvalue of the discrete problem, which fulfills

$$f''(\tilde{u}_h)(v, v) \geq \alpha \|v\|_U^2 \quad \forall v \in U_h$$

for a finite-dimensional subspace  $U_h \subset U$ . This method is widely employed in numerical experiments to indicate optimality of computed solution. However, despite being attractive from a computational point of view, this method is not save in general and may lead to wrong conclusions.

We will now construct an example with numerical solution  $\tilde{u}_h$  that has the following properties:

- $f'(\tilde{u}_h) = 0$ ,
- $f''(\tilde{u}_h)(v_h, v_h) \geq \alpha \|v_h\|_U^2 \quad \forall v_h \in U_h$  with  $\alpha > 0$ ,
- $\tilde{u}_h$  is not close to a local minimum of the original problem.

That means in particular, that all eigenvalues of the discrete Hessian matrix are positive. Hence,  $\tilde{u}_h$  is a local minimum of the discretized problem. Unfortunately, it appears that  $\tilde{u}_h$  is not even in the neighborhood of a local minimum of the original problem. Thus, the positive definiteness of the discrete Hessian is misleading.

#### Minimizing a fourth-order polynomial

We will consider now a special objective function. Let be given  $u_1 \neq u_2$  from the Hilbert space  $U$ . Then we want to minimize

$$f(u) = \frac{1}{2} \|u - u_1\|_U^2 \|u - u_2\|_U^2. \quad (3.1)$$

Of course, both  $u_1$  and  $u_2$  are global minima of this problem. Now, let us have a look on the derivatives of  $f$ . The first derivative is given by

$$f'(u) = (u - u_1)\|u - u_2\|_U^2 + (u - u_2)\|u - u_1\|_U^2. \quad (3.2)$$

And it turns out that  $\tilde{u} := \frac{1}{2}(u_1 + u_2)$  is a stationary point. If  $U$  is one-dimensional then  $\tilde{u}$  is a local maximum of  $f$ . For higher-dimensional  $U$ ,  $\tilde{u}$  is actually a saddle point as we will see. Hence, let us assume in the sequel that the dimension of  $U$  is greater than one.

The second derivative of  $f$  is given as bilinear form by

$$\begin{aligned} f''(u)(v_1, v_2) &= (\|u - u_1\|_U^2 + \|u - u_2\|_U^2)(v_1, v_2) \\ &\quad + 2(u - u_2, v_1)(u - u_1, v_2) + 2(u - u_1, v_1)(u - u_2, v_2). \end{aligned} \quad (3.3)$$

Formally, one can decompose  $f''$  into  $D + 2VV^T$ , where  $D$  is a positive multiple of the identity and  $VV^T$  is a two-rank perturbation. This simplifies the computation

of eigenvalues of  $f''$ . Let us set  $u = \tilde{u}$  and  $v_1 = v_2 = v$  in (3.3). We obtain

$$f''(\tilde{u})(v, v) = \frac{1}{2} \|u_1 - u_2\|_U^2 \|v\|_U^2 - (v, u_1 - u_2)^2. \quad (3.4)$$

Let us decompose the space  $U$  as the direct sum:  $\text{span}\{u_1 - u_2\} \oplus \{u_1 - u_2\}^\perp$ . Then we can write  $v = v_1 + v_2$  with  $(v_2, u_1 - u_2) = 0$ , which gives

$$f''(\tilde{u})(v, v) = \frac{1}{2} \|u_1 - u_2\|_U^2 (\|v_1\|_U^2 + \|v_2\|_U^2) - (v_1, u_1 - u_2)^2. \quad (3.5)$$

For  $v_1 \in \text{span}\{u_1 - u_2\}$  it holds  $(v_1, u_1 - u_2)^2 = \|u_1 - u_2\|_U^2 \|v_1\|_U^2$ , which implies

$$f''(\tilde{u})(v, v) = \frac{1}{2} \|u_1 - u_2\|_U^2 (\|v_2\|_U^2 - \|v_1\|_U^2).$$

Thus, for the direction of negative curvature  $v = s(u_1 - u_2)$  we have

$$f''(\tilde{u})(v, v) = -\frac{1}{2} \|u_1 - u_2\|_U^2 \|v\|_U^2.$$

With similar arguments, one finds the inequality

$$f''(u)(v, v) \geq \lambda_1(u) \|v\|_U^2$$

with

$$\lambda_1(u) = \|u - u_1\|_U^2 + \|u - u_2\|_U^2 - 2\|u - u_1\|_U \|u - u_2\|_U + 2(u - u_1, u - u_2). \quad (3.6)$$

Let us denote by  $U_h$  a finite-dimensional subspace of  $U$ . The orthogonal projection from  $U$  onto  $U_h$  is denoted by  $\Pi_h$ . Let us recall the expression for  $f''(\tilde{u})$ , cf. 3.4,

$$f''(\tilde{u})(v, v) = \frac{1}{2} \|u_1 - u_2\|_U^2 \|v\|_U^2 - (v, u_1 - u_2)^2.$$

We will now derive conditions such that  $f''(\tilde{u})[v_h, v_h] > 0$  is fulfilled for all  $v_h \neq 0$  from the finite-dimensional space  $U_h$ . Let us consider for a moment directions  $v_h$  with  $\|v_h\|_U = 1$ . The supremum of  $(v_h, u_1 - u_2)$  over all such  $v_h$  is attained at  $v_h = \frac{\Pi_h(u_1 - u_2)}{\|\Pi_h(u_1 - u_2)\|_U}$ , which implies for  $\|v_h\|_U = 1$

$$\begin{aligned} f''(\tilde{u})(v_h, v_h) &= \frac{1}{2} \|u_1 - u_2\|_U^2 - (v_h, u_1 - u_2)^2 \\ &\geq \frac{1}{2} \|u_1 - u_2\|_U^2 - \|\Pi_h(u_1 - u_2)\|_U^2. \end{aligned}$$

Using  $\|u_1 - u_2\|_U^2 = \|(I - \Pi_h)(u_1 - u_2)\|_U^2 + \|\Pi_h(u_1 - u_2)\|_U^2$ , we find that  $f''(\tilde{u})$  is positive definite on  $U_h$  if

$$\|(I - \Pi_h)(u_1 - u_2)\|_U \geq \|\Pi_h(u_1 - u_2)\|_U$$

holds. That is, if the  $L^2$ -norm of the projection  $\Pi_h(u_1 - u_2)$  captures less than one half of the  $L^2$ -norm of  $u_1 - u_2$ , then the bilinear form  $f''(\tilde{u})$  is positive definite on  $U_h$  despite being indefinite on whole  $U$ . Or in other words, if the discretization is too coarse to approximate the direction of negative curvature  $\tilde{v} = u_1 - u_2$  the bilinear form is positive definite on  $U_h$ .

Let us demonstrate that such a situation may occur for a concrete optimization problem. Let us define  $U$  to be the set of square integrable functions on  $I = (0, 1)$ ,  $U := L^2(0, 1)$ . Take an integer number  $N$ , set  $h := 1/N$ . The interval  $I$  is subdivided into  $N$  subintervals  $I_j$  of equal length  $h$ ,  $j = 1 \dots N$ . The discrete subspace  $U_h$  is chosen as the space of piecewise constant function on the intervals  $I_j$ . Let us denote by  $\Pi_h$  the  $L^2$ -projector onto  $U_h$ ,  $\Pi_h : L^2(I) \rightarrow U_h$ . The functions  $u_1, u_2$  we will choose such that

1.  $\tilde{u} = (u_1 + u_2)/2$  is in  $U_h$  for all  $h$ ,
2. the direction of negative curvature  $\tilde{v}$  is ‘hard to approximate’ with functions from  $U_h$ .

Let  $\epsilon > 0$  be a small number. We define

$$u_1(x) = x^{-1/2+\epsilon}, \quad u_2(x) = -u_1(x).$$

Then obviously we have  $u_1 + u_2 = 0 \in U_h$ . Moreover, it holds  $u_1 \in L^2(I)$  and  $u_1 \notin H^1(I)$ . The latter property is the reason, why  $u_1$  can only be approximated with low convergence rates with respect to  $h$ .

**Lemma 3.1.** *For the above choice of  $U_h$  and  $u_1$  it holds*

$$\|u_1 - \Pi_h u_1\|_U \geq g(\epsilon)h^\epsilon$$

with  $g(\epsilon) = \frac{|2\epsilon-1|}{\sqrt{2\epsilon}|2\epsilon+1|}$ .

*Proof.* Let us only consider the approximation of  $u_1$  by a function  $w$  that is constant on the first subinterval  $(0, h)$  and equal to  $u$  on  $(h, 1)$ ,

$$w(x) = \begin{cases} w_0 & \text{if } x \in (0, h) \\ u_1(x) & \text{if } x \in [h, 1) \end{cases}$$

with  $w_0$  in  $\mathbb{R}$ . It is clear that  $\|u_1 - \Pi_h u_1\|_U \geq \|u_1 - w\|_U$  holds. A short computation yields that  $w_0 = \frac{h^{-1/2+\epsilon}}{1/2+\epsilon}$  minimizes  $\|u_1 - w\|_U^2$  over all choices of  $w_0$ , which in turn gives

$$\int_0^h (u_1(x) - w_0)^2 dx = h^{2\epsilon} \frac{(\epsilon - 1/2)^2}{2\epsilon(\epsilon + 1/2)^2},$$

and the claim is proven.  $\square$

Let us remark, that the previous lemma not only gives an arbitrary small convergence rate for  $\epsilon \rightarrow 0+$  but also states that the constant explodes for  $\epsilon$  tending to zero.

In Table 1 we computed the “critical values” of  $h^0$ . If the mesh size  $h$  is larger than  $h^0$ , then one gets a wrong indication by the eigenvalues of the discrete Hessian: The smallest eigenvalue of the Hessian of the discretized problem is positive, but the computed solution is only a saddle point of the original problem. Consequently, the usual strategy to look at the smallest eigenvalue of the Hessian fails for this simple problem. The last line of Table 1 shows that this wrong indication can occur even for very small discretization parameters.

$\epsilon$	$h_0$	
0.05	$1/18$	$= 0.056$
0.04	$1/106$	$= 0.0094$
0.03	$1/1917$	$= 5.2 \cdot 10^{-4}$
0.02	$1/619660$	$= 1.6 \cdot 10^{-6}$

TABLE 1. Critical mesh sizes

#### 4. Application to an abstract optimal control problem

Let us consider now a more complicated optimization problem. We will introduce an additional constraint, which will mimic a partial differential equation. We investigate the minimization of the functional

$$J(y, u) := g(y) + j(u) \quad (4.1)$$

subject to

$$Ay + d(y) = Bu, \quad (4.2)$$

$$u \in U_{\text{ad}}. \quad (4.3)$$

Here,  $y$  denotes the state of the system,  $u$  the control. Let  $U_{\text{ad}}$  be a closed, convex, non-empty subset of the Hilbert space  $U$ .

**Assumption 4.1.** Let  $Y$  be a Banach space. Let  $A : Y \rightarrow Y'$  and  $B : U \rightarrow Y'$  be linear operators. Moreover, we assume  $A$  to be coercive, i.e., it holds  $\langle Ay, y \rangle_{Y', Y} \geq \delta \|y\|_Y^2$  for some  $\delta > 0$  and all  $y \in Y$ .

The functions  $d, g, j$  are twice Fréchet-differentiable as functions from  $Y$  to  $Y'$ ,  $Y$  to  $\mathbb{R}$ , and  $U$  to  $\mathbb{R}$ , respectively. Moreover, we assume for simplicity that  $d$  is monotone with  $d(0) = 0$ .

Thus, the state equation (4.2) has to hold in  $Y'$ . Under the assumptions above, this equation is uniquely solvable for each control  $u$ . Let us denote the solution mapping by  $S$ , i.e.,  $y = S(u)$  is the solution of (4.2). Since  $d$  is monotone, the linearized equation

$$Ay + d'(\tilde{y})y = Bu$$

is solvable, where we set  $\tilde{y} = S(\tilde{u})$ . In addition, there exists an upper bound of the norm of its solution operator  $S'(\tilde{u})$ ,

$$\|S'(u)\|_{L(Y', Y)} = \|(A + d'(y))^{-1}\|_{L(Y', Y)} \leq c_A \quad \forall u \in U, y = S(u).$$

In view of this estimate, we can directly give a Lipschitz estimate for solutions of (4.2)

$$\|S(u_1) - S(u_2)\|_Y \leq c_A \|B\| \cdot \|u_1 - u_2\|_U. \quad (4.4)$$

**Assumption 4.2.** Let us take  $R > 0$  and  $u_h \in U_{\text{ad}}$ . Then we assume that there exist positive constants  $c_{g'}, c_{d'}, c_{d''}, c_{g''}, c_{j''}$  depending on  $R$  such that the local Lipschitz estimates

$$\begin{aligned} \|d'(y) - d'(y_h)\|_{L(Y, Y')} &\leq c_{d'} \|y - y_h\|_Y \\ \|g'(y) - g'(y_h)\|_{Y'} &\leq c_{g'} \|y - y_h\|_Y \\ \|d''(y) - d''(y_h)\|_{L(Y \times Y, Y')} &\leq c_{d''} \|y - y_h\|_Y \\ \|g''(y) - g''(y_h)\|_{(Y \times Y)' } &\leq c_{g''} \|y - y_h\|_Y \\ \|j''(u) - j''(u_h)\|_{(U \times U)' } &\leq c_{j''} \|u - u_h\|_Y \end{aligned}$$

hold for all  $u \in U_{\text{ad}}$  with  $\|u - u_h\|_U < R$  and  $y = S(u)$ .

The problem class covered by Assumptions 4.1 and 4.2 is wide enough to cover distributed or boundary control problems for semilinear elliptic equations. Moreover, the case of the steady-state Navier-Stokes equations fits also in the assumption. However, we have to admit that optimal control problems with two-norm discrepancy are not included.

Let us define the reduced cost functional  $\phi : U \rightarrow \mathbb{R}$  by

$$\phi(u) := g(S(u)) + j(u).$$

The conditions in Section 2, i.e., Assumptions 2.1, 2.2, and 2.4, have now to be interpreted as conditions on the reduced cost functional. The reduced functional of course inherits the structure of the optimal control problem (4.1)–(4.3). So we will express the conditions on  $\phi$  in terms of the original problem.

Let  $(\bar{u}, \bar{y})$  be an admissible pair for (4.1)–(4.3). If  $\bar{u}$  is locally optimal, then there exists an adjoint state  $\bar{p} \in Y$  such that it holds

$$A^* \bar{p} + d'(\bar{y})^* \bar{p} = g'(\bar{y})$$

and

$$(j'(\bar{u}) + B^* \bar{p}, u - \bar{u}) \geq 0 \quad \forall u \in U_{\text{ad}}.$$

Let now  $(u_h, y_h, p_h)$  be some triple consisting of approximations of a locally optimal control, state, and adjoint. Suppose  $u_h$  is an admissible control. Let us assume that we can control the residuals of the optimality system.

**Assumption 4.3.** There are positive constants  $\epsilon_u, \epsilon_y, \epsilon_p$  such that it holds

$$(j'(u_h) + B^* p_h, u - u_h) \geq -\epsilon_u \|u - u_h\|_U \quad \forall u \in U_{\text{ad}}, \quad (4.5)$$

$$\|Ay_h + d(y_h) - Bu_h\|_{Y'} \leq \epsilon_y, \quad (4.6)$$

$$\|A^* p_h + d'(y_h)^* p_h - g'(y_h)\|_{Y'} \leq \epsilon_p. \quad (4.7)$$

This assumption corresponds to (2.1) in Assumption 2.1 of Section 2. We will now investigate the error in the variational inequality (2.1), i.e., we want to estimate  $\epsilon$  in

$$\phi'(u_h)(u - u_h) \geq -\epsilon \|u - u_h\|_U \quad \forall u \in U_{\text{ad}}.$$

To characterize the derivative  $\phi'$  in terms of the original problem let us introduce two auxiliary functions  $y^h$  and  $p^h$  as the solutions of

$$\begin{aligned} Ay^h + d(y^h) &= u_h, \\ A^*p^h + d'(y^h)^*p^h &= g'(y^h). \end{aligned} \quad (4.8)$$

We have the following error estimates for these states and adjoints:

**Lemma 4.4.** *Let  $y^h$  and  $p^h$  be given by (4.8). Then it holds*

$$\|y^h - y_h\|_Y \leq c_A \epsilon_y, \quad (4.9)$$

$$\|p^h - p_h\|_Y \leq c_A ((c_{g'} + c_{d'} \|p_h\|_Y) \|y^h - y_h\|_Y + \epsilon_p). \quad (4.10)$$

*Proof.* The difference  $p^h - p_h$  solves the equation

$$\begin{aligned} A^*(p^h - p_h) + d'(y^h)^*(p^h - p_h) &= g'(y^h) - g'(y_h) \\ &\quad - (A^*p_h + d'(y_h)^*p_h - g'(y_h)) + (d'(y_h)^* - d'(y^h)^*)p_h, \end{aligned}$$

which immediately gives (4.10) using the notations of Assumption 4.2. The difference  $y^h - y_h$  can be treated similarly, and one obtains  $\|y^h - y_h\|_Y \leq c_A \epsilon_y$ .  $\square$

Observe, that  $y^h$  and  $p^h$  can be written as  $y^h = S(u_h)$  and  $p^h = S'(u_h)^*g'(y^h)$ . Hence, we can rewrite the first derivative  $\phi'$  as

$$\begin{aligned} \phi'(u_h)(u - u_h) &= (j'(u_h) + B^*S'(u_h)^*g'(y^h), u - u_h) \\ &= (j'(u_h) + B^*p^h, u - u_h). \end{aligned}$$

**Lemma 4.5.** *The following inequality is satisfied for all admissible controls  $u \in U_{\text{ad}}$*

$$\phi'(u_h)(u - u_h) \geq -\epsilon \|u - u_h\|_U,$$

where  $\epsilon$  is given by

$$\epsilon := \epsilon_u + \|B\| \cdot \|p^h - p_h\|_Y$$

*Proof.* The claim follows immediately from

$$\begin{aligned} \phi''(u_h)(u - u_h) &= (j'(u_h) + B^*p^h, u - u_h) \\ &= (j'(u_h) + B^*p_h + B^*(p^h - p_h), u - u_h) \\ &\geq -(\epsilon_u + \|B\| \cdot \|p^h - p_h\|_Y) \|u - u_h\|_U. \end{aligned} \quad \square$$

Next, we need a coercivity condition on the second derivative of the Lagrangian involving known quantities only.

**Assumption 4.6.** There is a constant  $\delta > 0$  such that

$$j''(u_h)(v, v) + g''(y_h)(z, z) + d''(y_h)(z, z)p_h \geq \delta \|v\|^2 \quad (4.11)$$

holds for all  $v = u - u_h$ ,  $u \in U_{\text{ad}}$  with  $z$  being the solution of the linearized equation.

$$Az + d'(y_h)z = Bv. \quad (4.12)$$

This condition is especially fulfilled for convex  $j$  and  $g$  and under the sign condition  $d''(y_h)p_h > 0$ . We will now derive a lower bound for the eigenvalues of  $\phi''$  analogously to (2.2).

**Lemma 4.7.** *Let  $v = u - u_h$ ,  $u \in U_{\text{ad}}$  be given. Then it holds*

$$\phi''(u_h)(v, v) \geq \alpha \|v\|_U^2$$

with  $\alpha$  given by

$$\begin{aligned} \alpha = & \delta - (c_{g''} + \|p^h - p_h\|_Y(c_{d''} + \|d''(y_h)\|) + c_{d''}\|p_h\|_Y) \|y^h - y_h\|_Y(c_A\|B\|)^2 \\ & - (\|g''(y_h)\|_{(Y \times Y)'} + \|p_h\|_Y\|d''(y_h)\|_{L(Y \times Y, Y')}) (2c_{d'}\|y_h - y^h\|_Y c_A^3 \|B\|^2). \end{aligned}$$

*Proof.* We can write the second derivative of  $\phi$  as

$$\phi''(u_h)(v, v) = j''(u_h)(v, v) + g''(y^h)(z^h, z^h) + d''(y^h)(z^h, z^h)p^h, \quad (4.13)$$

where  $z^h$  solves

$$Az^h + d'(y^h)z^h = Bv. \quad (4.14)$$

Let us denote by  $z$  the solution of (4.12). A priori bounds of  $z$  and  $z^h$  can be calculated as above, and we obtain

$$\|z\|_Y, \|z^h\|_Y \leq c_A\|B\| \cdot \|v\|_U.$$

The difference  $z^h - z$  solves  $A(z^h - z) + d'(y^h)(z^h - z) = (d'(y_h) - d'(y^h))z$ , hence it holds

$$\|z^h - z\|_Y \leq c_A c_{d'} \|y_h - y^h\|_Y c_A \|B\| \|v\|_U.$$

Let us introduce the abbreviations  $s^h := g''(y^h) + d''(y^h)(\cdot, \cdot)p^h$ ,  $s_h := g''(y_h) + d''(y_h)(\cdot, \cdot)p_h$ ;  $s^h, s_h : Y \times Y \rightarrow \mathbb{R}$ . Then we write

$$s^h(z^h, z^h) = s_h(z, z) + (s^h - s_h)(z^h, z^h) + s_h((z^h, z^h) - (z, z)).$$

Here, the first addend appears in the coercivity assumption (4.11). The second addend is estimated as

$$\begin{aligned} \|s^h - s_h\|_{(Y \times Y)'} & \leq \left( c_{g''} + \|p^h - p_h\|_Y \|d''(y_h)\|_{L(Y \times Y, Y')} \right. \\ & \quad \left. + (\|p_h\|_Y + \|p^h - p_h\|_Y) c_{d''} \right) \|y^h - y_h\|_Y. \end{aligned}$$

For the third one we obtain

$$\begin{aligned} \|s_h((z^h, z^h) - (z, z))\|_{(Y \times Y)'} & \leq (\|g''(y_h)\|_{(Y \times Y)} + \|p_h\|_Y \|d''(y_h)\|_{L(Y \times Y, Y')}) \|z^h + z\|_Y \|z^h - z\|_Y. \end{aligned}$$

Putting everything together we finally find

$$\phi''(u_h)(v, v) \geq \alpha \|v\|_U^2$$

with  $\alpha$  equal to

$$\begin{aligned} \delta - & (c_{g''} + \|p^h - p_h\|_Y(c_{d''} + \|d''(y_h)\|) + c_{d''}\|p_h\|_Y) \|y^h - y_h\|_Y(c_A\|B\|)^2 \\ & - (\|g''(y_h)\|_{(Y \times Y)'} + \|p_h\|_Y \|d''(y_h)\|_{L(Y \times Y, Y')}) (2c_{d'}\|y_h - y^h\|_Y c_A^3 \|B\|^2). \quad \square \end{aligned}$$

According to Lemma 4.4, that the negative terms in the estimate of  $\alpha$  are of the order of  $\epsilon_y, \epsilon_p$ . That is, there is hope that  $\alpha$  is positive for small residuals in the optimality system. That will be true in particular, if  $(u_h, y_h, p_h)$  solves a very fine discretized problem and a second-order sufficient optimality condition is fulfilled for the original problem (4.1)–(4.3).

**Corollary 4.8.** *If Assumption 4.6 holds with the linearized equation (4.14) instead of (4.12), then the statement of Lemma 4.7 is valid with*

$$\alpha = \delta - (c_{g''} + \|p^h - p_h\|_Y (c_{d''} + \|d''(y_h)\|) + c_{d''} \|p_h\|_Y) \|y^h - y_h\|_Y (c_A \|B\|)^2.$$

Finally, we have to compute the Lipschitz constant of  $\phi''$  as equivalent to inequality (2.3) in Section 2.

**Lemma 4.9.** *There is a constant  $M > 0$  such that it holds for all  $u \in U_{ad}$  with  $\|u - u_h\|_U < R$*

$$|(\phi''(u) - \phi''(u_h))(v_1, v_2)| \leq M \|u - u_h\|_U \|v_1\|_U \|v_2\|_U \quad \forall v_1, v_2 \in U.$$

An upper bound of  $M$  is given in the course of the proof.

*Proof.* Let  $y$  and  $p$  be the solutions of the state and adjoint equations associated with  $u$ , i.e., they satisfy

$$Ay + d(y) = u, \quad A^*p + d'(y)^*p = g'(y).$$

Then it holds

$$\begin{aligned} (\phi''(u) - \phi''(u_h))(v_1, v_2) &= (j''(u) - j''(u_h))(v_1, v_2) + (g''(y) - g''(y^h))(z_1, z_2) \\ &\quad + (pd''(y) - p^h d''(y^h))(z_1, z_2), \end{aligned}$$

where the  $z_i$ ,  $i = 1, 2$ , are the solutions of the linearized equations  $Az_i + d'(y^h)z_i = Bv_i$ . Using Lipschitz continuity of the solution mapping, cf. (4.4), we obtain

$$\begin{aligned} \|y - y^h\|_Y &\leq c_A \|B\| \cdot \|u - u_h\| \leq c_A \|B\| R, \\ \|y - y_h\|_Y &\leq c_A (\|B\| \cdot \|u - u_h\| + \epsilon_y) \leq c_A (\|B\| R + \epsilon_y). \end{aligned} \tag{4.15}$$

Similarly to (4.10), the difference of the adjoint states is estimated by

$$\begin{aligned} \|p - p^h\|_Y &\leq c_A (c_{g'} + c_{d'} (\|p^h - p_h\|_Y + \|p_h\|_Y)) \|y - y^h\|_Y \\ &\leq c_A (c_{g'} + c_{d'} (\|p^h - p_h\|_Y + \|p_h\|_Y)) c_A \|B\| R. \end{aligned} \tag{4.16}$$

Employing the splitting

$$\begin{aligned} pd''(y) - p^h d''(y^h) &= (p - p^h)d''(y) + p^h(d''(p) - d''(y^h)) \\ &= (p - p^h)(d''(y_h) + d''(y) - d''(y_h)) \\ &\quad + (p_h + p^h - p_h)(d''(y) - d''(y^h)) \end{aligned}$$

we can estimate

$$\begin{aligned} \|d''(y)(\cdot, \cdot)p - d''(y^h)(\cdot, \cdot)p^h\|_{(Y \times Y)'}) &\leq \|p - p^h\|_Y \|d''(y_h)\|_{L(Y \times Y, Y')} \\ &\quad + c_{d''} (\|p - p^h\|_Y \|y - y_h\|_Y + (\|p_h\|_Y + \|p^h - p_h\|_Y) \|y - y^h\|_Y). \end{aligned}$$

And the claim of the Lemma holds with

$$\begin{aligned} M \geq c_{j''} + (c_A \|B\|)^2 & \left( c_{g''} \|y - y^h\|_Y + \|p - p^h\|_Y \|d''(y_h)\|_{L(Y \times Y, Y')} \right. \\ & \left. + c_{d''} (\|p - p^h\|_Y \|y - y_h\|_Y + (\|p_h\|_Y + \|p^h - p_h\|_Y) \|y - y^h\|_Y) \right), \end{aligned}$$

where for  $\|y - y^h\|_Y, \|y - y_h\|_Y, \|p - p^h\|_Y$  the corresponding upper bounds (4.15)–(4.16) has to be used.  $\square$

Lemmata 4.5, 4.7, and 4.9 give the possibility to estimate the constants  $\epsilon, \alpha, M$  that are needed to proceed with the results of Section 2.

**Theorem 4.10.** *Let the constants given by Lemmata 4.5, 4.7, and 4.9 fulfill the Assumption 2.4. Then there exists a local solution  $\bar{u}$  of the optimal control problem (4.1)–(4.3), which satisfies*

$$\|\bar{u} - u_h\|_U \leq \alpha \sqrt{1 - \frac{2M\epsilon}{\alpha^2}}.$$

Moreover, the second-order sufficient condition holds at  $\bar{u}$ .

## 5. An optimal control problem with two local minima

Now, let us apply the technique described above to the following optimal control problem: Minimize

$$\frac{1}{2} \|y - y_d\|_{L^2(\Omega)}^2 + \frac{1}{2} \|u - u_1\|_{L^2(\Omega)}^2 \|u - u_2\|_{L^2(\Omega)}^2 \quad (5.1)$$

subject to the semilinear state equation

$$\begin{aligned} -\Delta y(x) + y(x)^3 &= u(x) && \text{in } \Omega \\ y(x) &= 0 && \text{on } \Gamma \end{aligned} \quad (5.2)$$

and the control constraints

$$u_a(x) \leq u(x) \leq u_b(x) \text{ a.e. on } \Omega. \quad (5.3)$$

Let  $\Omega \subset \mathbb{R}^d$ ,  $d = 2, 3$ , be a bounded domain with Lipschitz boundary  $\Gamma$ . Furthermore, functions  $y_d, u_a, u_b \in L^2(\Omega)$ ,  $u_a(x) \leq u_b(x)$  a.e. on  $\Omega$ , are given.

At first, let us choose the function spaces. We set  $Y := H_0^1(\Omega)$  with  $\|y\|_Y := \|\nabla y\|_{L^2(\Omega)}$  and  $U = L^2(\Omega)$ ,  $\|u\|_U = \|u\|_{L^2(\Omega)}$ . The operator  $A$  is given by  $A = -\Delta$ . The right-hand side operator  $B$  is the embedding operator  $L^2(\Omega) \rightarrow H^{-1}(\Omega)$ . Its norm is bounded as  $\|B\|_{L(U, Y')} \leq I_2$ . Using the notation of the previous section, we define

$$\begin{aligned} d(y) &:= y^3, \\ g(y) &:= \frac{1}{2} \|y - y_d\|_{L^2(\Omega)}^2, \\ j(u) &:= \frac{\nu}{2} \|u - u_1\|_{L^2(\Omega)}^2 \|u - u_2\|_{L^2(\Omega)}^2 \end{aligned}$$

Due to the embedding  $H^1(\Omega) \rightarrow L^4(\Omega)$ , the function  $d$  is differentiable from  $Y = H_0^1(\Omega)$  to  $Y'$ . Let us denote upper bounds of the embedding constants  $H_0^1(\Omega) \rightarrow L^p(\Omega)$  by  $I_p$ ,  $p < \infty$ . They can be computed by eigenvalue estimates for the Laplacian. Furthermore, formulas for  $I_p$  are given in [11].

We computed  $(u_h, y_h, p_h)$  as the solution of the discretized optimal control problem. The discretization was carried out using  $P2$ -elements for the state and  $P1$ -elements for the control.

In Section 4, many constants have to be computed. Let us report, how we computed them for the particular example.

**Solution estimates.** By monotonicity of the semilinearity  $d(y)$  we have,

$$\|S(u_1) - S(u_2)\|_Y \leq I_2 \|u_1 - u_2\|_U.$$

Since  $d'(y)$  is non-negative, it holds,

$$\|(A + d'(y))^{-1}\|_{Y', Y} \leq 1 =: c_A.$$

**Lipschitz estimates.** Let  $u \in U$  be taken with  $\|u - u_h\|_U \leq R$ . Then we have  $\|y - y^h\|_Y := \|S(u) - S(u_h)\|_Y \leq I_2 R$ . Some of the Lipschitz constants, will depend on  $R$ . After an easy computation, one finds

$$\begin{aligned} \|(d'(y) - d'(y_h))z\|_{Y'} &\leq I_4^4 (\|y_h\|_Y + I_2 R) \|y - y_h\|_Y \|z\|_Y \\ &=: c_{d'}(R) \|y - y_h\|_Y \|z\|_Y \\ \|g'(y) - g'(y_h)\|_{Y'} &\leq I_2^2 \|y - y_h\|_Y \\ \|d''(y) - d''(y_h)\|_{L(Y \times Y, Y')} &\leq I_4^4 \|y - y_h\|_Y \\ \|g''(y) - g''(y_h)\| &= 0. \end{aligned}$$

The function  $j$  coincides with the function  $f$  analyzed in Section 3, its derivative was derived in (3.3). Let us now write

$$\begin{aligned} (j''(u) - j''(u_h))(v_1, v_2) \\ = & (\|u - u_1\|_U^2 + \|u - u_2\|_U^2 - \|u_h - u_1\|_U^2 - \|u_h - u_2\|_U^2) (v_1, v_2) \\ & + 2(u - u_2, v_1)(u - u_1, v_2) + 2(u - u_1, v_1)(u - u_2, v_2) \\ & - 2(u_h - u_2, v_1)(u_h - u_1, v_2) - 2(u_h - u_1, v_1)(u_h - u_2, v_2). \end{aligned}$$

Using the identity

$$\|u - u_1\|_U^2 - \|u_h - u_1\|_U^2 = (u - u_h, 2(u_h - u_1) + (u - u_h)),$$

we find for  $\|u - u_h\|_U \leq R$

$$|\|u - u_1\|_U^2 - \|u_h - u_1\|_U^2| \leq (2\|u_h - u_1\|_U + R) \|u - u_h\|_U.$$

Analogously, we get

$$\begin{aligned} & (u - u_2, v_i)(u - u_1, v_j) - (u_h - u_2, v_i)(u_h - u_1, v_j) \\ &= \left( (u - u_2, v_i) - (u_h - u_2, v_i) \right) (u_h - u_1, v_j) + (u - u_2, v_i)(u - u_h, v_j) \\ &= (u - u_h, v_i)(u_h - u_1, v_j) + (u - u_h + u_h - u_2, v_i)(u - u_h, v_j), \end{aligned}$$

which implies the estimate

$$\begin{aligned} & \|(u - u_2, v_i)(u - u_1, v_j) - (u_h - u_2, v_i)(u_h - u_1, v_j)\| \\ & \leq (\|u_h - u_1\|_U + \|u_h - u_2\|_U + R) \|u - u_h\|_U \|v_i\|_U \|v_j\|_U. \end{aligned}$$

Hence, we obtain the following value of  $c_{j''}$ :

$$c_{j''}(R) := 6(\|u_h - u_1\|_U + \|u_h - u_2\|_U + R).$$

**Residual estimates.** Now, let us explain how we obtained bounds for the residuals in Assumption 4.3. If one could compute a function  $q$  such that the inequality

$$(j'(u_h) + B^* p_h + q, u - u_h) \geq 0$$

holds for *all* admissible controls  $u \in U_{\text{ad}}$ , then the lower bound in (4.5) is realized by  $\epsilon_u = \|q\|_U$ . The computation of such a function  $q$  is described for instance in [9].

There are quite a few possibilities to estimate the residuals in the state and the adjoint equation. For instance, one can apply standard a posteriori error estimators of residual type. We used another possibility, as described in [11].

Let  $\sigma \in H(\text{div})$  be given, i.e.,  $\sigma \in L^2(\Omega)^d$  with  $\text{div}(\sigma) \in L^2(\Omega)$ . Then we can estimate

$$\begin{aligned} \|-\Delta y_h + d(y_h) - Bu_h\|_{H^{-1}} &\leq \|-\Delta y_h - \text{div}(\sigma)\|_{H^{-1}} + \|\text{div}(\sigma) + d(y_h) - Bu_h\|_{H^{-1}} \\ &\leq \|\nabla y_h - \sigma\|_{L^2} + I_2 \|\text{div}(\sigma) + d(y_h) - Bu_h\|_{L^2}. \end{aligned}$$

In our computations, we used the Raviart-Thomas elements  $RT_1$  to discretize the space  $H(\text{div})$ . In a post-processing step, we computed  $\sigma_h$  as minimizer of

$$\|\nabla y_h - \sigma\|_{L^2}^2 + I_2^2 \|\text{div}(\sigma) + d(y_h) - Bu_h\|_{L^2}^2.$$

A similar technique was applied to compute the adjoint residual (4.7).

**Coercivity check.** The lower coercivity bound  $\delta$  as in Assumption 4.6 was computed as  $\delta = \lambda_1(u_h)$ , where  $\lambda_1(u)$  is defined by (3.6). Since it holds

$$g''(y_h)(z, z) + d''(y_h)(z, z)p_h = \int_{\Omega} (g''(y_h(x)) + d''(y_h(x))p_h(x))(z(x))^2 dx,$$

we checked the sign of  $g''(y_h) + d''(y_h)p_h$ . If the sign was positive,  $\delta$  was chosen as above, and we could use the estimate given by Corollary 4.8. Otherwise, the computation were repeated on a finer mesh.

**Computation of  $r_+$ .** As one can see above, some of the constants depend on the safety radius  $R$ . That implies that the constants  $\epsilon, \alpha, M$  depend on  $R$  as well. Let us report, how we computed the value  $\tilde{r}_+$ , cf. Assumption 2.4. By a bisection method, we computed an interval  $[r_1, r_2]$  that contains the smallest positive root

of the polynomial

$$p(r) = -\epsilon(r + \rho) + \alpha(r + \rho) - \frac{1}{2}M(r + \rho)r^2$$

with small  $\rho = 10^{-5}$ . Then  $\tilde{r}_+$  was chosen as the right border of the interval,  $\tilde{r}_+ = r_2$ . That is, all the assumptions are fulfilled for  $\tilde{r}_+$  and  $R := \tilde{r}_+ + \rho$ . If the bisection method was not able to find  $r$  such that  $p(r)$  was positive, the whole computation was restarted on a finer mesh.

**Data.** The domain  $\Omega$  was chosen as  $\Omega = (0, 3)^2 \setminus [1, 2]^2$ . Hence,  $\Omega$  is not convex. This implies that the solution of the elliptic equation does not belong to  $H^2(\Omega)$  in general. Thus, the theory as developed in [12] cannot be applied. Furthermore, we took

$$y_d(x_1, x_2) = 0.02 \cdot \sin(\pi x_1) \sin(\pi x_2)$$

and

$$u_1(x) = 0.1, \quad u_2(x) = 0.4,$$

$$u_a(x) = 0.394, \quad u_b(x) = 0.099.$$

**Solution method and results.** The mesh was chosen as a uniform triangulation of the domain with 25.600 triangles, which yields a mesh size of about  $h = 0.035$ . We solved the discretized optimal control problem by the SQP-method with semi-smooth Newton's method for the inner problems. As initial guesses we used  $y_h^0 = 0$  and  $p_h^0 = 0$  for state and adjoint. Starting the SQP-method at  $u_h^0 = 0$  yields the solution depicted in Figure 1.

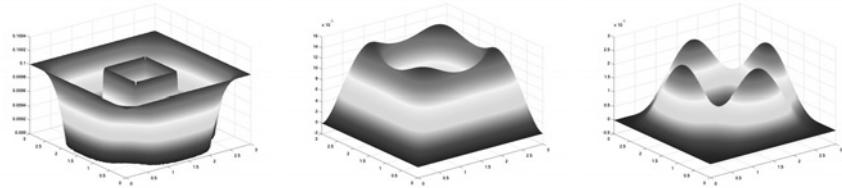


FIGURE 1. First solution: control  $\bar{u}_h^1$ , state  $\bar{y}_h^1$ , adjoint state  $\bar{p}_h^1$

For a different starting point, we obtained a different solution. Choosing  $u_0 = 0.5$  yields the solution triple shown in Figure 2.

The results of the numerical verification technique are as follows. The radius  $\tilde{r}_+^1 = 5.773 \cdot 10^{-4}$  satisfies Assumption 2.4 and thus the requirements of Theorem 4.10. Hence, there exists a local solution  $\bar{u}^1$  of (5.1)–(5.3) in the neighborhood of  $u_h^1$  with the error estimate

$$\|\bar{u}^1 - \bar{u}_h^1\|_U \leq 5.773 \cdot 10^{-4}.$$

Moreover, the second derivative of the reduced cost functional is positive definite and it holds

$$\phi''(\bar{u}^1)(v, v) \geq 0.7202\|v\|_U^2 \quad \forall v \in U.$$

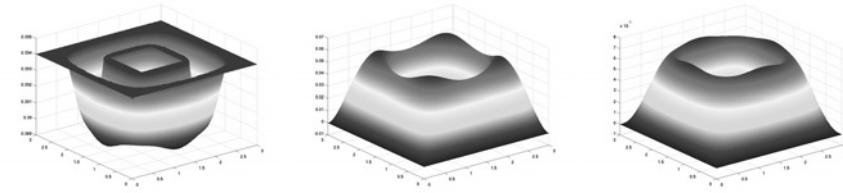


FIGURE 2. Second solution: control  $\bar{u}_h^2$ , state  $\bar{y}_h^2$ , adjoint state  $\bar{p}_h^2$

Similarly, we found the radius  $\tilde{r}_+^2 = 2.727 \cdot 10^{-3}$  for the second discrete solution, which gives the existence of a locally optimal control  $\bar{u}^2$  with

$$\|\bar{u}^2 - \bar{u}_h^2\|_U \leq 2.727 \cdot 10^{-3}$$

and

$$\phi''(\bar{u}^2)(v, v) \geq 0.5991 \|v\|_U^2 \quad \forall v \in U.$$

Since  $\|u_h^1 - u_h^2\|_U$  is much larger than  $\tilde{r}_+^1 + \tilde{r}_+^2$ , the controls  $\bar{u}^1$  and  $\bar{u}^2$  are clearly separated. Consequently, the optimal control problem (5.1)–(5.3) with the data as given above has at least two locally optimal controls.

**Convergence rates.** We computed solutions for different mesh sizes. The coarsest mesh was obtained by a uniform triangulation with 400 triangles. The meshes were then refined using a grading strategy [2] to cope with the re-entrant corners.

The convergence behaviour of the SQP-method did not change: depending on the initial guess the obtained solutions were either near  $\bar{u}^1$  or  $\bar{u}^2$ . In Table 2 we listed the error bounds  $r_+^1$  and  $r_+^2$ .

$h$	$r_+^1$	$r_+^2$
0.28284	$2.7311 \cdot 10^{-3}$	$1.2382 \cdot 10^{-2}$
0.18284	$1.2450 \cdot 10^{-3}$	$5.7057 \cdot 10^{-3}$
0.09913	$5.8972 \cdot 10^{-4}$	$2.7202 \cdot 10^{-3}$
0.05134	$2.8629 \cdot 10^{-4}$	$1.3268 \cdot 10^{-3}$
0.02609	$1.4096 \cdot 10^{-4}$	$6.6074 \cdot 10^{-4}$
0.01315	$6.9948 \cdot 10^{-5}$	$3.3381 \cdot 10^{-4}$

TABLE 2. Error bounds for different meshes

As one can see, the error bounds  $r_+^1$  and  $r_+^2$  decrease like  $h$ . For a uniform discretization of the non-convex domain  $\Omega$  one would expect lower convergence rates. The optimal convergence rate is then recovered using mesh-grading in the vicinity of the re-entrant corners.

## References

- [1] W. Alt. The Lagrange-Newton method for infinite dimensional optimization problems. *Numerical Functional Analysis and Optimization*, 11:201–224, 1990.
- [2] T. Apel, A. Sändig, and J. Whiteman. Graded mesh refinement and error estimates for finite element solutions of elliptic boundary value problems in non-smooth domains. *Math. Methods Appl. Sci.*, 19(1):63–85, 1996.
- [3] J.F. Bonnans and A. Shapiro. *Perturbation analysis of optimization problems*. Springer, New York, 2000.
- [4] E. Casas and F. Tröltzsch. Second-order necessary and sufficient optimality conditions for optimization problems and applications to control theory. *SIAM J. Optim.*, 13:406–431, 2002.
- [5] E. Casas, F. Tröltzsch, and A. Unger. Second-order sufficient optimality conditions for a nonlinear elliptic control problem. *J. for Analysis and its Applications*, 15:687–707, 1996.
- [6] P. Deuflhard. *Newton methods for nonlinear problems*, volume 35 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2004.
- [7] A.L. Dontchev. Local analysis of a Newton-type method based on partial linearization. *AMS Lectures in Appl. Math.*, pages 295–306, 1996.
- [8] A.L. Dontchev, W.W. Hager, A.B. Poore, and B. Yang. Optimality, stability, and convergence in optimal control. *Appl. Math. Optim.*, 31:297–326, 1995.
- [9] K. Krumbiegel and A. Rösch. A new stopping criterion for iterative solvers for control constrained optimal control problems. *Archives of Control Sciences*, 18(1):17–42, 2008.
- [10] J.M. Ortega. The Newton-Kantorovich theorem. *Amer. Math. Monthly*, 75:658–660, 1968.
- [11] M. Plum. Computer-assisted enclosure methods for elliptic differential equations. *Linear Algebra Appl.*, 324(1-3):147–187, 2001.
- [12] A. Rösch and D. Wachsmuth. Numerical verification of optimality conditions. *SIAM J. Control Optim.*, 47(5):2557–2581, 2008.

Daniel Wachsmuth

Johann Radon Institute for Computational and Applied Mathematics (RICAM)  
Austrian Academy of Sciences  
Altenbergerstraße 69  
A-4040 Linz, Austria  
e-mail: daniel.wachsmuth@ricam.oeaw.ac.at

Arnd Rösch

Universität Duisburg-Essen  
Fachbereich Mathematik  
Forsthausweg 2  
D-47057 Duisburg, Germany  
e-mail: arnd.roesch@uni-due.de

# Hidden Boundary Shape Derivative for the Solution to Maxwell Equations and Non Cylindrical Wave Equations

Jean-Paul Zolésio

**Abstract.** This paper deals with the shape derivative of boundary shape functionals governed by electromagnetic 3D time-depending Maxwell equations solution  $E, H$  (or non cylindrical wave equation to begin with) involving derivatives terms at the boundary whose existence are related to hidden-like regularity results. Under weak regularity of data we compute the shape derivative of  $L^2$  norms at the boundary of the electrical field  $E$  and magnetic field  $H$  solution to the 3D Maxwell system. This analysis is obtained via MinMax parameter derivative under saddle point existence, so that no shape sensitivity analysis of the solution  $E, H$  is needed, see [1]. The results are completely new concerning the non-cylindrical wave equation as well as for Maxwell equations. Of course this technic is true for all classical shape derivative of quasi convex functionals governed by classical linear problems such as elliptic or parabolic problems (including elasticity). For these situations the result as been developed in many former papers after [12] and later the book [11]. So that the strong material and shape derivative of the “state” is still necessary only for non linear problems for which the Langrangian is definitively not convex-concave, then we use the weak form of the Implicit function theorem (in order to deal with the minimum regularity), see [14] R. Dziri and J.P. Zolesio, but there is still hope to extend this analysis to a larger class of “local saddle points”.

**Mathematics Subject Classification (2000).** 49K20, 35L05, 35L55, 35L99.

**Keywords.** Shape derivative, sharp regularity, saddle point, moving domain, maxwell equations.

## 1. Introduction

We obtain shape derivative results associated with Neumann boundary-like conditions under minimal regularity of the data, extending some older results derived for the Dirichlet boundary condition for wave equations. In fact we extend here the results in two directions: the context of the more difficult Neumann conditions and the full sharp regularity for the Maxwell equation in presence of metallic obstacle, leading to mixed Dirichlet-Neumann condition at the boundary for the Electrical vector field  $E$ . For example ([7],[4]) this situation includes the very simple example for the classical wave equation with homogeneous Dirichlet boundary conditions:

$$y_{tt} - \Delta y = f \text{ in } \Omega, \quad y = 0 \text{ on } \partial\Omega.$$

With  $(y(0), y_t(0)) \in H_0^1(\Omega) \times L^2(\Omega)$ ,  $f \in L^1(I, L^2(\Omega))$ , this situation leads to a solution  $y$  which fails to be in  $L^2(I, H^2(\Omega))$ , nevertheless, by hidden regularity, the normal derivative satisfies  $\frac{\partial}{\partial n} y \in L^2(I \times \partial\Omega)$ ; then we consider shape functionals in the following form:

$$J(\Omega_s) = \int_0^\tau \int_{\partial\Omega_s} \left( \frac{\partial}{\partial n} y_s \right)^2 dt d\Gamma$$

and we show the shape differentiability for any perturbation  $\Omega_s$  in the form  $\Omega_s = T_s(Z)(\Omega)$  where the vector field  $Z$  is assumed to be Lipschitz continuous. The proof makes use of the *extractor technique* introduced at ICIAM 1995 ([6]) and several papers ([7],[4],[3]);

The paper is based on two “main ingredients”, the so-called “extractor identity” (whose proof will be derived in the next section):

$$\forall \phi \in H^2(\Omega_s), \int_{\Gamma_s} \left( - \left( \frac{\partial}{\partial n_s} \phi \right)^2 + |\nabla_{\Gamma_s} \phi|^2 \right) d\Gamma_s = j_{\Omega_s}(\phi),$$

where

$$j_{\Omega_s}(\phi) = \int_{\Omega_s} [2\langle D^2 b_{\Omega_s}^h \cdot \nabla \phi, \nabla \phi \rangle + |\nabla \phi|^2 \Delta b_{\Omega_s}^h - \Delta \phi \nabla \phi \cdot \nabla b_{\Omega_s}^h] dx \quad (1.1)$$

and the Min Max differentiability (for convex-concave Lagrangian  $L(s, \cdot, \cdot)$  with saddle points...)

$$J(\Omega_s) = \text{Min}_{\phi \in K(\Omega)} \text{Max}_{\psi \in K(\Omega)} L(s, \phi, \psi)$$

where  $K(\Omega) = H_0^1(\Omega)$ , for the homogeneous Dirichlet boundary value problem,  $K(\Omega) = H^1(\Omega)$  for the Neumann boundary condition and

$$L(s, \phi, \psi) = j_{\Omega_s}(\phi o T_s^{-1}) + \int_{\Omega_s} [\nabla(\phi o T_s^{-1}) \cdot \nabla(\psi o T_s^{-1}) - F \psi o T_s^{-1}] dx$$

which, by the Min Max parameter derivative principle (1987) that we recall in the next section, insure the existence of the derivative

$$dJ(\Omega, Z) := \frac{\partial}{\partial s} J(\Omega_s)|_{s=0} = \frac{\partial}{\partial s} L(0, y, p)$$

where  $(y, p)$  is the saddle point. Here we must assume  $F := \Delta y \in L^2(D)$ ,  $\Omega_s \subset D$ ,  $0 \leq s$ . By considering the calculus of this derivative by change of variable  $T_s$  or as moving domain derivative we immediately derive the distributed expression and the shape gradient supported by the boundary (and associated with a new regularity result), see also ([7]).

To illustrate this technique with a *new result*, we begin with the scalar wave equation in a moving domain (or non cylindrical evolution), then we adapt this technique to the for the time depending Electric field on a metallic obstacle. The Maxwell case is much more delicate but uses exactly the same steps. The expression of the two Lagrangian functionals under consideration will be more technical as we shall be obliged to treat separately the vector wave equation solution  $E$ , the tangential part  $E^\tau$  then the normal component  $e$ .

We consider a domain  $\Omega$  with boundary  $\Gamma$  on which the boundary condition  $E_\Gamma = 0$  applies <sup>(1)</sup>. Assuming free divergence initial data  $E_i \in H^i(\Omega, R^N)$ ,  $i = 0, 1$  and free divergence current  $J \in L^2(0, \tau, L^2(\Omega, R^N))$ ; we derive that, at the boundary, the magnetic field verifies  $H \in H^{1/2}(0, \tau, L^2(\Gamma, R^3))$  while  $\operatorname{curl} E \in H^{-1/2}(0, \tau, L^2(\Gamma, R^3))$ . Associated with hidden regularity associated respectively with the “tangent electric field”  $E^\tau$  and the normal one  $e$  we introduce a new singular boundary functional  $J(\Omega_s)$  which, through hidden regularity and a new convex-concave Lagrangian formulation turns out to be shape differentiable (i.e., differentiable with respect to  $s$  with a shape derivative  $dJ(\Omega, Z)$  linear continuous with respect to the direction field  $Z$ ). We also investigate the expression of the shape gradient density on the boundary. The place in this paper does not allow for complete analysis of this gradient but we give here the complete keys in order to reach it in finite number of pages... This analysis is critical in electromagnetic as far as the effect of shape singularities (for example discontinuity lines for the normal to the surface) play a central role. Namely when dealing with harmonic regimes one usually prefers the integral representation of the solution. This theory based on the Corton-Kreiss isomorphism assumes the boundary to be of regularity  $C^2$ . We first derive that  $(DE.n)_\Gamma \in L^2([0, \tau] \times \Gamma, R^3)$ ,  $E.n \in H^{1/2}(I, L^2(\Gamma))$  and  $\nabla_\Gamma E.n \in H^{-1/2}(I, L^2(\Gamma))$ . This last regularity proof follows a pseudo-differential extractor technique. To derive the differentiability under weak regularity we make use of the derivative (with respect to a parameter  $s$ ) of a Min Max formulation under continuous saddle point. We then introduce the Lagrangian  $L(s, y, p)$  and we recall the basic semi-derivative result we use in several papers since 1987.

## 2. The derivative of a MinMax with saddle point

For convenience we recall here a result from ([1], 1987). Let  $E, F$  be two Banach spaces and  $\mathbf{L}(s, e, f)$  be a function defined from  $[0, 1] \times K_E \times K_F$  into  $R$ , where  $K_E, K_F$  are convex sets, respectively in  $E$  and  $F$ . The *Lagrangian* functional  $\mathbf{L}$  is convex l.s.c. with respect to  $e$ , concave u.s.c. with respect to  $f$  and continuously

---

<sup>1</sup>For  $N = 3$  this condition can be written  $E \times n = 0$

differentiable with respect to the parameter  $s$ . Assume there exists a non empty set  $S(s)$  of saddle points. Then it always takes the following form:

$$S(s) = A(s) \times B(s), \quad A(s) \subset K_E, \quad B(s) \subset K_F,$$

such that:

$$\begin{aligned} \forall a(s) \in A(s), \quad & \forall b(s) \in B(s), \quad \forall \alpha \in K_A, \quad \forall \beta \in K_B, \\ \mathbf{L}(s, a(s), \beta) \leq \mathbf{L}(s, a(s), b(s)) \leq \mathbf{L}(t, \alpha, b(s)), \end{aligned}$$

so that  $\forall \alpha' \in K_E, \forall \beta' \in K_F$  we have

$$-\mathbf{L}(0, \alpha', b(0)) \leq -\mathbf{L}(0, a(0), b(0)) \leq -\mathbf{L}(0, a(0), \beta')$$

by choosing  $\alpha = a(0), \beta = b(0), \alpha' = a(s), \beta' = b(s)$ , and adding the two previous inequalities we get: for any  $s > 0$ :

$$\begin{aligned} \frac{\mathbf{L}(s, a(s), b(0)) - \mathbf{L}(0, a(s), b(0))}{s} &\leq \frac{\mathbf{L}(s, a(s), b(s)) - \mathbf{L}(0, a(0), b(0))}{s} \\ &\leq \frac{\mathbf{L}(s, a(0), b(s)) - \mathbf{L}(0, a(0), b(s))}{s}. \end{aligned}$$

Under reasonable smoothness assumptions on  $\mathbf{L}$  and Kuratowski continuity of the sets  $A(s)$  and  $B(s)$  we get the semi-derivative of

$$\begin{aligned} l(s) &= \text{Min}_{a \in K_E} \text{Max}_{b \in K_F} \mathbf{L}(s, a, b) \\ l'(0) &= \text{Min}_{a \in A(0)} \text{Max}_{b \in B(0)} \frac{\partial}{\partial s} \mathbf{L}(0, a, b). \end{aligned} \tag{2.1}$$

In the following section we shall make use of that semi-derivative in the specific situation in which the set  $S(0)$  is reduced to a unique pair,  $A(0) = \{y\}, B(0) = \{p\}$ , where  $Y$  and  $p$  will be the “state” and “adjoint-state” solution associated with the wave (or Maxwell) equation under consideration. In this situation the function  $l$  is differentiable at  $s = 0$  and the derivative (2.1) takes the following form:

$$l'(0) = \frac{\partial}{\partial s} \mathbf{L}(0, y, p).$$

### 3. Shape derivative for singular boundary term in non-cylindrical wave equation

In order to illustrate on a simple example (a non-vectorial one) the method, while giving a very new result, we consider the non-cylindrical wave equation under homogeneous Dirichlet boundary condition on a moving domain  $\Omega_t$ . We assume that this moving domains evolves under the action of a smooth vector field  $V$  whose flow vector  $T_t(V)$  is a smooth one to-one mapping from the bounded domain  $D \subset R^N$  into itself and letting invariant the boundary  $\partial D$  which is also assumed to be smooth.

$$\Omega_t = T_t(V)(\Omega_0) \subset D, \quad Q_V = \cup_{0 < t < \tau} \{t\} \times \Omega_t$$

For any  $F \in L^2(0, \tau, L^2(D))$  we consider the solution  $y \in H := C^1([0, \tau], H_0^1(\Omega_t)) \cap C^0([0, \tau], L^2(\Omega_t))$  as solution to

$$y_{tt} - \Delta y = F \text{ in } Q_V, \quad y(0) = y_0, \quad y_t(0) = y_1.$$

For any given elements  $y_0 \in H_0^1(D)$ ,  $y_1 \in L^2(D)$ .

### 3.1. Perturbation field $Z$

Consider a smooth vector field  $Z(s, t, x)$  where  $s$  is the small perturbation parameter and consider the perturbed tube  $\Omega_t^s = T_s(Z(t))(\Omega_t)$  and denote

$$Q^s = \cup_{0 < t < \tau} \{t\} \times \Omega_t^s$$

and consider the element

$$y^s \in H^s := C^1([0, \tau], H_0^1(\Omega_t^s)) \cap C^0([0, \tau], L^2(\Omega_t^s))$$

solution to

$$y_{tt}^s - \Delta y^s = F \text{ in } Q^s, \quad y^s(0) = y_0, \quad y_t^s(0) = y_1.$$

(We assume  $Z(s, 0, x) = 0$ , so that  $\Omega_0^s = \Omega_0$ .)

### 3.2. Extractor identity

Let

$$\mathbf{E}(s) = \int_0^\tau \int_{\Omega_t^s} \{ -(y_t oT_s(Z(t))^{-1})^2 + |\nabla(y oT_s(Z(t))^{-1})|^2 \} dx dt$$

and

$$e = \frac{\partial}{\partial s} \mathbf{E}(s)|_{s=0}.$$

The calculus by moving domains leads to:

$$\begin{aligned} e_1 &= \int_0^\tau \int_{\partial\Omega_t} \{ -(y_t)^2 + |\nabla y|^2 \} \langle Z(0, t, x), n_t(x) \rangle d\Gamma_t(x) dt \\ &\quad + 2 \int_0^\tau \int_{\Omega_t} \{ y_t(t, x) \nabla y_t(t, x) \cdot Z(0, t, x) \\ &\quad - \langle \nabla y(t, x), \nabla(\langle \nabla y, (Z(0, t, x)) \rangle) \rangle \} dx dt. \end{aligned}$$

The calculus by change of variable leads to:

$$\begin{aligned} e_2 &= \frac{\partial}{\partial s} \left( \int_0^\tau \int_{\Omega_t} \{ -(y_t oT_s(Z))^{-2} oT_s(Z) \right. \\ &\quad \left. + |\nabla(y oT_s(Z))^{-1}|^2 oT_s(Z) \} \det DT_s(Z) dx dt \right) \\ &= \frac{\partial}{\partial s} \left( \int_0^\tau \int_{\Omega_t} \{ -(y_t)^2 + |DT_s(Z)^{-*} \cdot \nabla(y)|^2 \} \det DT_s(Z) dx dt \right). \end{aligned}$$

That is

$$\begin{aligned} e_2 &= \int_0^\tau \int_{\Omega_t} \{ -(y_t)^2 + |\nabla(y)|^2 \} \operatorname{div}(Z(0, t, x) dx dt) \\ &\quad - \int_0^\tau \int_{\Omega_t} \{ \langle \nabla y, DZ(0, t, x)^*. \nabla y \rangle dx dt \}. \end{aligned}$$

We pay attention to the second term of  $e_1$ :

$$\begin{aligned} &2 \int_0^\tau \int_{\Omega_t} \{ y_t(t, x) \nabla y_t(t, x).Z(0, t, x) - \langle \nabla y(t, x), \nabla(\langle \nabla y, Z(0, t, x) \rangle) \} dx dt \\ &= \int_{\Omega_\tau} y_t(\tau, x) \nabla y(\tau, x).Z(0, \tau, x) dx - \int_{\Omega_0} y_t(0, x) \nabla y(0, x).Z(0, 0, x) dx \\ &\quad - 2 \int_0^\tau \int_{\Omega_t} \{ y_{tt}(t, x) \nabla y(t, x).Z(0, t, x) \\ &\quad - 2 \int_0^\tau \int_{\Omega_t} \{ y_t(t, x) \nabla y(t, x).Z_t(0, t, x) \\ &\quad - \int_0^\tau \int_{\partial\Omega_t} y_t(t, x) \langle \nabla y(t, x), Z(0, t, x) \rangle v(t, x) d\Gamma_t(x) dt \\ &\quad + 2 \int_0^\tau \int_{\Omega_t} \Delta y \langle \nabla y, Z(0, t, x) \rangle dx dt \\ &\quad - 2 \int_0^\tau \int_{\partial\Omega_t} \langle \nabla y(t, x), n_t(x) \rangle \langle \nabla y, Z(0, t, x) \rangle d\Gamma_t(x) dt. \end{aligned}$$

We choose the transverse perturbation vector field  $Z$  such that  $Z(0, \tau, x) = 0$  so that the second term cancels with (for example)  $Z$  in the  $s$ -autonomous form

$$Z(s, t, x) = (\tau - t)W(t, x),$$

with  $W = \nabla b_{\Omega_t}^h$  we have  $\langle \nabla y, Z(0, t, x) \rangle = (\tau - t)(\frac{\partial}{\partial n_t} y)^2$  and  $y_t(t, x) + \frac{\partial}{\partial n_t} y v(t, x) = 0$  on  $\partial\Omega_t$ . We obtain the (non-cylindrical) extractor identity  $e_1 = e_2$  which leads to:

$$\begin{aligned} &\int_0^\tau \int_{\partial\Omega_t} \{ (\tau - t)(\frac{\partial}{\partial n_t} y)^2 (1 - v(t)^2) d\Gamma_t(x) dt \\ &\quad - \int_0^\tau \int_{\partial\Omega_t} y_t(t, x) \langle \nabla y(t, x), Z(0, t, x) \rangle v(t, x) d\Gamma_t(x) dt \\ &\quad - 2 \int_0^\tau \int_{\partial\Omega_t} \langle \nabla y(t, x), n_t(x) \rangle \langle \nabla y, Z(0, t, x) \rangle d\Gamma_t(x) dt \\ &= \tau \int_{\Omega_0} y_1(x) \langle \nabla y_0, \nabla b_{\Omega_0}^h \rangle dx + 2 \int_0^\tau (\tau - t) \int_{\Omega_t} F(t, x) \nabla y(t, x). \nabla b_{\Omega_t}^h \\ &\quad + 2 \int_0^\tau \int_{\Omega_t} \{ y_t(t, x) \nabla y(t, x). \nabla b_{\Omega_t}^h \end{aligned}$$

$$\begin{aligned}
& + \int_0^\tau (\tau - t) \int_{\Omega_t} \{ -(y_t)^2 + |\nabla(y)|^2 \} \Delta b_{\Omega_t}^h(0, t, x) dx dt \\
& - \int_0^\tau (\tau - t) \int_{\Omega_t} \{ \langle \nabla y, D^2 b_{\Omega_t}^h(0, t, x) \cdot \nabla y \rangle \} dx dt
\end{aligned}$$

and we obtain:

$$\begin{aligned}
& \int_0^\tau \int_{\partial\Omega_t} \{ (\tau - t) (\frac{\partial}{\partial n_t} y)^2 \} d\Gamma_t(x) dt \\
& = -\tau \int_{\Omega_0} y_1(x) \langle \nabla y_0, \nabla b_{\Omega_0}^h \rangle dx - 2 \int_0^\tau \int_{\Omega_t} \{ (\tau - t) F + y_t \} \nabla y \cdot \nabla b_{\Omega_t}^h \\
& + \int_0^\tau (\tau - t) \int_{\Omega_t} \{ \langle D^2 b_{\Omega_t}^h \cdot \nabla y, \nabla y \rangle + (y_t^2 - |\nabla y|^2) \Delta b_{\Omega_t}^h \} dx dt.
\end{aligned} \tag{3.1}$$

## 4. Derivatives

Let  $\mathbf{Z}(s, t, x)$  be any smooth vector field with  $\mathbf{Z}(s, 0, x) = 0$  and  $\Omega_t^s := T_s(\mathbf{Z})(\Omega_t)$ . Let  $y^s \in H(Q^s)$  and

$$y_s := y^s o T_s(\mathbf{Z}) \in H(Q).$$

We have  $\phi \in H(Q)$  iff  $\phi o T_s(\mathbf{Z})^{-1} \in H(Q^s)$ , so that we introduce the following Lagrangian:

$$\begin{aligned}
\mathbf{L}(s, \phi, \psi) & = -\tau \int_{\Omega_0} y_1(x) \langle \nabla y_0, \nabla b_{\Omega_0}^h \rangle dx \\
& - 2 \int_0^\tau \int_{\Omega_t^s} \{ (\tau - t) F + (\phi o T_s(\mathbf{Z})^{-1})_t \} \nabla(\phi o T_s(\mathbf{Z})^{-1}) \cdot \nabla b_{\Omega_t^s}^h \\
& + \int_0^\tau (\tau - t) \int_{\Omega_t^s} \{ \langle D^2 b_{\Omega_t^s}^h \cdot \nabla(\phi o T_s(\mathbf{Z})^{-1}), \nabla(\phi o T_s(\mathbf{Z})^{-1}) \rangle \\
& + ((\phi o T_s(\mathbf{Z})^{-1})_t)^2 - |\nabla(\phi o T_s(\mathbf{Z})^{-1})|^2 \} \Delta b_{\Omega_t^s}^h \} dx dt \\
& + \int_0^\tau \int_{\Omega_t^s} ((\phi o T_s(\mathbf{Z})^{-1})_t (\psi o T_s(\mathbf{Z})^{-1})_t \\
& - \langle \nabla(\phi o T_s(\mathbf{Z})^{-1}), \nabla(\psi o T_s(\mathbf{Z})^{-1}) \rangle - F \psi) dx dt + \int_{\Omega_0} y_1 \psi(0) dx.
\end{aligned} \tag{4.1}$$

### 4.1. Adjoint equation

We characterise the second component  $p = b(0)$  of the saddle point  $S(0)$ ; this is done by writing the necessary condition of optimality with respect to  $\phi$ , at  $\psi = p$ .

We get the following variational problem:

$$\begin{aligned} \forall \phi, \quad & -2 \int_0^\tau \int_{\Omega_t} \{ \phi_t \langle \nabla y, \nabla b_{\Omega_t}^h \rangle + y_t \langle \nabla \phi, \nabla b_{\Omega_0}^h \rangle \} dx dt \\ & + 2 \int_0^\tau (\tau - t) \int_{\Omega_t} \{ \langle D^2 b_{\Omega_t}^h \cdot \nabla y, \nabla \phi \rangle + 2(y_t \phi_t - \langle \nabla y, \nabla \phi \rangle) \Delta b_{\Omega_t}^h \} dx dt \\ & + \int_0^\tau \int_{\Omega_t} (\phi_t p_t - \langle \nabla \phi, \nabla p \rangle) dx dt = 0. \end{aligned} \quad (4.2)$$

#### 4.2. The shape functional $J$

We set:

$$\begin{aligned} J(Q^s) &:= \int_0^\tau \int_{\partial \Omega_t^s} \{ (\tau - t) \left( \frac{\partial}{\partial n_t} y^s \right)^2 d\Gamma_t(x) dt \\ &= \text{Min}_{\{\phi \in H(Q), \phi(0) = y_0\}} \text{Max}_{\{\psi \in H(Q), \psi(\tau) = 0\}} \mathbf{L}(s, \phi, \psi), \end{aligned} \quad (4.3)$$

such that

$$\frac{\partial}{\partial s} J(Q^s)|_{s=0} = \frac{\partial}{\partial s} \mathbf{L}(0, y, p)$$

where  $p$  turns out to be the (unique) backward adjoint state which will be made more precise.

#### 4.3. $s$ -derivative of the Lagrangian

By obvious change of variable we have:

$$\begin{aligned} \mathbf{L}(s, \phi, \psi) &= -\tau \int_{\Omega_0} y_1(x) \langle \nabla y_0, \nabla b_{\Omega_0}^h \rangle dx - 2 \int_0^\tau \int_{\Omega_t} \{ (\tau - t) F o T_s \\ &+ (\phi o T_s^{-1})_t o T_s \} \nabla(\phi o T_s^{-1}) o T_s \cdot (\nabla b_{\Omega_0}^h) o T_s J(s) \\ &+ \int_0^\tau (\tau - t) \int_{\Omega_t} \{ \langle D^2 b_{\Omega_t}^h o T_s \cdot (\nabla \phi o T_s^{-1}) o T_s, (\nabla \phi o T_s^{-1}) o T_s \rangle \\ &+ ((\phi o T_s^{-1})_t o T_s)^2 - |(\nabla(\phi o T_s^{-1}) o T_s|^2) \Delta b_{\Omega_t}^h o T_s \} dx dt \\ &+ \int_0^\tau \int_{\Omega_t} ((\phi o T_s^{-1})_t o T_s (\psi o T_s^{-1})_t o T_s \\ &- \langle \nabla(\phi o T_s^{-1}) o T_s, \nabla(\psi o T_s^{-1}) o T_s \rangle - F o T_s \psi) J(s) dx dt - \int_{\Omega_0} y_1 \psi(0) dx, \end{aligned}$$

where  $J(s) = \det(DT_s)$ . Now we assume that the vector field  $\mathbf{Z}$  does not depends on  $t$  so that  $(\phi o T_s^{-1})_t o T_s = \phi_t$ . Then it simplifies for

$$\begin{aligned} \mathbf{L}(s, \phi, \psi) &= -\tau \int_{\Omega_0} y_1(x) \langle \nabla y_0, \nabla b_{\Omega_0}^h \rangle dx - 2 \int_0^\tau \int_{\Omega_t} \{ (\tau - t) F o T_s \\ &+ \phi_t \langle D T_s^{-*} \cdot \nabla \phi, (\nabla b_{\Omega_t}^h) o T_s \rangle \} J(s) dx dt \\ &+ \int_0^\tau (\tau - t) \int_{\Omega_t} \{ \langle D^2 b_{\Omega_t}^h o T_s \cdot D T_s^{-*} \cdot \nabla \phi, D T_s^{-*} \cdot \nabla \phi \rangle \} \end{aligned} \quad (4.4)$$

$$\begin{aligned}
& + (\phi_t)^2 - |(\nabla(\phi)|^2) \Delta b_{\Omega_t}^h oT_s \} dx dt \\
& + \int_0^\tau \int_{\Omega_t} (\phi_t \psi_t - \langle DT_s^{-*} \cdot \nabla \phi, DT_s^{-*} \cdot \nabla \psi \rangle - F oT_s \psi) J(s) dx dt \\
& - \int_{\Omega_0} y_1 \psi(0) dx.
\end{aligned}$$

**4.3.1. The  $s$  derivative.** This derivative is now very simple, as the perturbation parameter  $s$  just appears through the geometrical terms, then we get:

$$\begin{aligned}
\frac{\partial}{\partial s} \mathbf{L}(s, \phi, \psi)|_{s=0} = & -2 \int_0^\tau \int_{\Omega_t} \{ (\tau - t) \nabla F \cdot \mathbf{Z}(0) - \phi_t \langle \nabla \phi, D\mathbf{Z}(0) \cdot \nabla b_{\Omega_0}^h \rangle \} dx dt \\
& + \phi_t \langle \nabla \phi, D^2 b_{\Omega_0}^h \cdot \mathbf{Z}(0) \rangle + \phi_t \langle \nabla \phi, \nabla b_{\Omega_0}^h \rangle \operatorname{div} \mathbf{Z}(0) \\
& + \int_0^\tau (\tau - t) \int_{\Omega_t} \{ \langle [D^3 b_{\Omega_t}^h \cdot \mathbf{Z}(0) - D^2 b_{\Omega_t}^h \cdot D\mathbf{Z}^*(0)]. \nabla \phi, \nabla \phi \rangle \\
& - \int_0^\tau (\tau - t) \int_{\Omega_t} \{ \langle D^2 b_{\Omega_t}^h \cdot \nabla \phi, D\mathbf{Z}^*(0) \cdot \nabla \phi \rangle \\
& + (\phi_t)^2 - |(\nabla(\phi)|^2) \nabla(\Delta b_{\Omega_t}^h) \cdot \mathbf{Z}(0) \} dx dt \\
& + \int_0^\tau \int_{\Omega_t} (\langle D\mathbf{Z}(0) \cdot \nabla \phi, \nabla \psi \rangle \\
& + \langle \nabla \phi, D\mathbf{Z}(0) \cdot \nabla \psi \rangle - \nabla F \cdot \mathbf{Z}(0) \psi) dx dt \\
& + \int_0^\tau \int_{\Omega_t} (\phi_t \psi_t - \langle \nabla \phi, \nabla \psi \rangle - F \psi) \operatorname{div} \mathbf{Z}(0) dx dt.
\end{aligned}$$

This expression completely proves that the Lagrangian is  $s$ -differentiable for any  $\phi$  and  $\psi$  in  $H(Q)$  so that the derivative of the Lagrangian fully applies as soon as we will show that the adjoint problem has a unique solution  $p$  in  $H(Q)$  which turns out to be the solution of the classical so-called “adjoint problem” in optimal control theory. The previous expression of the derivative is rather heavy and from the structure theorem (of the derivative) we know that this expression should be obtained as a boundary one, that is as a measure acting at the lateral boundary  $\Sigma = \cup_{0 < t < \tau} \{t\} \times \partial \Omega_t$  on the normal component term  $z(t) := \langle \mathbf{Z}(0, t, .), n_t(.) \rangle$ . To obtain this expression we could make intensive use of the Stoke's by part integration in the previous expression of the Lagrangian derivative but we do prefer to do moving boundary derivative from the expression of the Lagrangian itself (4.1) and not from (4.4):

$$\begin{aligned}
dJ(Q; \mathbf{Z}) = & -2 \int_0^\tau \int_{\partial \Omega_t} \{ (\tau - t) F + y_t \} \nabla y \cdot \nabla b_{\Omega_t}^h z d\Gamma_t dt \\
& + \int_0^\tau (\tau - t) \int_{\partial \Omega_t} \{ \langle D^2 b_{\Omega_t}^h \cdot \nabla y, \nabla y \rangle \\
& + ((y_t)^2 - |\nabla y|^2) \Delta b_{\Omega_t}^h \} z d\Gamma_t(x) dt
\end{aligned} \tag{4.5}$$

$$+ \int_0^\tau \int_{\partial\Omega_t} ( y_t p_t - \langle \nabla y, \nabla p \rangle - F p ) z(t, x) d\Gamma_t(x) dt.$$

#### 4.4. The shape gradient

It is a measure concentrated on  $\Sigma$  given by:

$$\begin{aligned} G = & \{(\tau - t) F + y_t\} \nabla y \cdot \nabla b_{\Omega_t}^h + \{ \langle D^2 b_{\Omega_t}^h \cdot \nabla y, \nabla y \rangle \\ & + ((y_t)^2 - |\nabla y|^2) \Delta b_{\Omega_t}^h \} + (y_t p_t - \langle \nabla y, \nabla p \rangle - F p). \end{aligned}$$

This expression simplifies a lot as  $D^2 b \nabla y = 0$  (Dirichlet condition),

$$\begin{aligned} y_t &= -v \frac{\partial}{\partial n_t} y, \quad \text{etc. } \dots \\ G &= \left\{ (\tau - t) F - v \frac{\partial}{\partial n_t} y \right\} \frac{\partial}{\partial n_t} y \\ &+ (\tau - t) (v^2 - 1) \left( \frac{\partial}{\partial n_t} y \right)^2 \Delta b_{\Omega_t}^h + (v^2 - 1) \frac{\partial}{\partial n_t} y \frac{\partial}{\partial n_t} p. \end{aligned}$$

**4.4.1. Distribution on the boundary.** The shape gradient associated with the cylindrical evolution problem would be obtain simply by making  $v = 0$  in this expression. The existence of the shape derivative  $dJ(\Omega; \mathbf{Z})$  implies the existence, as an element in  $W^{1,\infty}(\partial\Omega)'$  of the *non-linear* weak term

$$g = (\tau - t) \left\{ F \frac{\partial}{\partial n} y - \left( \frac{\partial}{\partial n} y \right)^2 \Delta b_\Omega^h \right\} - \frac{\partial}{\partial n} y \frac{\partial}{\partial n} p.$$

### 5. Maxwell equations

#### 5.1. Free divergence solutions to Maxwell and wave equations

As  $E$  is the electrical field we deal with vector functions, say

$$E \in C^0([0, \tau], H^1(\Omega, R^N))$$

where the time interval is  $I = ]0, \tau[$  while  $\Omega$  is a bounded smooth domain with boundary  $\Gamma$ . Throughout this paper we shall be concerned with divergence free initial conditions  $E_0, E_1$  and right-hand side  $F$  for the classical wave equation formulated in the cylindrical evolution domain  $Q = I \times \Omega$ . We shall discuss the boundary conditions on the lateral boundary  $\Sigma = I \times \Gamma$ .

#### 5.2. Wave deriving from Maxwell equation

Assuming perfect media ( $\epsilon = \mu = 1$ ) the Ampère law is

$$\operatorname{curl} \mathbf{H} = \frac{\partial}{\partial t} E + J, \tag{5.1}$$

where  $J$  is the electric current density. The Faraday's law is

$$\operatorname{curl} E = - \frac{\partial}{\partial t} \mathbf{H}. \tag{5.2}$$

The conservations laws are:

$$\operatorname{div} E = \rho, \quad \operatorname{div} \mathbf{H} = 0, \quad (5.3)$$

where  $\rho$  is the volume charge density. From (5.1,5.2), as  $\operatorname{div} \operatorname{curl} = 0$ , we obtain

$$\operatorname{div} J = -\operatorname{div}(E_t) = -\rho_t.$$

We assume  $\rho = 0$ , which implies  $\operatorname{div} J = 0$ . Under this assumption any  $E$  solving (5.1) is divergence free as soon as the initial condition  $E_0$  is. We shall also assume  $\operatorname{div} E_0 = 0$  so that (5.5) will be a consequence of (5.1). With  $F = -J_t$ , we similarly get  $\operatorname{div} F = 0$  and  $E$  solves the usual Maxwell equation:

$$E_{tt} + \operatorname{curl} \operatorname{curl} E = F, \quad E(0) = E_0, \quad E_t(0) = E_1. \quad (5.4)$$

**Lemma 5.1.** *We assume  $\operatorname{div} F = \operatorname{div} E_0 = \operatorname{div} E_1 = 0$  then any solution  $E$  to Maxwell equation (5.4) verifies the conservation condition (5.3) (with  $\rho = 0$ ):*

$$\operatorname{div} E = 0. \quad (5.5)$$

We have the classical identity

$$\operatorname{curl} \operatorname{curl} E = -\Delta E + \nabla(\operatorname{div} E)$$

so that  $E$  also solves the following wave problem

$$E_{tt} - \Delta E = F, \quad E(0) = E_0, \quad E_t(0) = E_1. \quad (5.6)$$

### 5.3. Boundary conditions

The physical boundary condition for metallic boundary is  $E \times n = 0$  which can be written as the homogeneous Dirichlet condition on the tangential component of the field:

$$E_\Gamma = 0 \text{ on } \Gamma \quad (5.7)$$

We introduce the following Fourier-like boundary condition involving the mean curvature  $\Delta b_\Omega = \lambda_1 + \lambda_2$  of the surface  $\Gamma$ .

$$\Delta b_\Omega E \cdot n + \langle D E \cdot n, n \rangle = 0 \text{ on } \Gamma. \quad (5.8)$$

In flat pieces of the boundary this condition simplifies to the usual Neumann condition.

**Proposition 5.2.** *Let  $E$  be a smooth element ( $E \in \mathbf{H}^2$ , see below) and the 3 free divergence elements ( $E_0, E_1, F$ ) be given in*

$$H^2(\Omega, R^3) \times H^1(\Omega, R^3) \times L^2(0, \tau, H^1(\Omega, R^3)).$$

*Then:*

- i) *Let  $E$  be solution to Maxwell-metallic system (5.4), (5.7). Then  $E$  solves the mixed wave problem (5.6), (5.7), (5.8) and, from Lemma 5.1,  $E$  solves also the free divergence condition (5.5).*
- ii) *Let  $E$  be solution to the wave equation (5.6) with “metallic” b.c. (5.7), then  $E$  verifies the Fourier-like condition (5.8) if and only if  $E$  verifies the free divergence condition (5.5).*

- iii) Let  $E$  be a free divergence solution to the “metallic” wave problem (5.6), (5.5), (5.7), then  $E$  solves the Maxwell problem (5.4), (5.7), (5.8).

*Proof.* We consider  $e = \operatorname{div}E$ ; if  $E$  is solution to Maxwell problem (5.4) then  $e$  solves the scalar wave equation with initial conditions  $e_i = \operatorname{div}E_i = 0$ ,  $i = 0, 1$  and right-hand side  $f = \operatorname{div}F = 0$ . If  $E$  solves (5.8) then we get  $e = 0$  as from the following result we get  $e = 0$  on the boundary:

**Lemma 5.3.** *Let  $E \in H^2(\Omega)$  solving the tangential Dirichlet condition (5.7), then we have the following expression for the trace of  $\operatorname{div}E$ :*

$$\operatorname{div}E|_{\Gamma} = \Delta b_{\Omega} \langle E, n \rangle + \langle DE.n, n \rangle \text{ on } \Gamma. \quad (5.9)$$

*Proof of the lemma.* The divergence successively decomposes as follows at the boundary (see ([12],[11])) as follows:

$$\begin{aligned} \operatorname{div}E|_{\Gamma} &= \operatorname{div}_{\Gamma}(E) + \langle DE.n, n \rangle = \operatorname{div}_{\Gamma}(E.n \vec{n}) + \operatorname{div}_{\Gamma}(E_{\Gamma}) + \langle DE.n, n \rangle \\ &= \langle \nabla_{\Gamma}(E.n), n \rangle + E.n \operatorname{div}_{\Gamma}(\vec{n}) + \operatorname{div}_{\Gamma}(E_{\Gamma}) + \langle DE.n, n \rangle. \end{aligned}$$

Obviously  $\langle \nabla_{\Gamma}(E.n), n \rangle = 0$ , the mean curvature of the surface  $\Gamma$  is  $\Delta b_{\Omega} = \operatorname{div}_{\Gamma}(\vec{n})$  and if the field  $E$  solves the tangential Dirichlet condition (5.7) we get the following simple expression for the restriction to the boundary of the divergence:

$$\operatorname{div}(E)|_{\Gamma} = \Delta b_{\Omega} \langle E, n \rangle + \langle DE.n, n \rangle.$$

Then if  $E$  solves the extra condition “Fourier-like” (5.8) we get  $e = 0$  on  $\Gamma$  so that  $e = 0$ .

#### 5.4. The Wave-Maxwell mixed problem

From the previous considerations it derives that under the free divergence assumption for the 3 data  $E_0, E_1, F$  the 3 following problems are equivalents (in the sense that any smooth solution of one of them is solution to the two others): Maxwell problem (5.4, 5.7), Free-Wave pb (5.6, 5.5, 5.7), Mixed-Wave pb (5.6, 5.7, 5.8). We emphasize that any solution to Maxwell pb solves the free divergence condition (5.5) and the Fourier-like condition (5.8). Any solution to the Mixed-Wave pb solves (for free) the free div condition (5.5). Any solution to Free-Wave pb solves (for free) the Fourier-like condition (5.8). The purpose of this paper is to develop the proof of the following regularity result:

**Proposition 5.4.** *Let  $(E_0, E_1, J)$  be divergence free vectors fields in*

$$H^1(\Omega, R^3)^2 \times L^2(\Omega, R^3) \times H^1(I, L^2(\Omega, R^3)).$$

*with zero tangential components:  $(E_0)_{\Gamma} = 0$ . Assume also  $\operatorname{curl}E_1 = 0$ . The Maxwell problem (5.4, 5.7) has a unique solution*

$$E \in C^0(\bar{I}, H^1(\Omega, R^3)) \cap C^1(\bar{I}, L^2(\Omega, R^3))$$

*verifying the boundary regularity:*

$$\operatorname{curl}E|_{\Gamma} \in H^{-1/2}(I \times \Gamma, R^3). \quad (5.10)$$

so that the magnetic field  $\mathbf{H}$  at the boundary verifies

$$\mathbf{H}|_{\Gamma} \in H^{1/2}(I, L^2(\Gamma, R^3)). \quad (5.11)$$

Moreover we have

$$E|_{\Gamma} \in H^{1/2}(I, L^2(\Gamma, R^3)). \quad (5.12)$$

Moreover, if  $J|_{\Gamma} \in L^2(I, L^2(\Gamma))$ , from Ampère's law (5.1) we obtain

$$\operatorname{curl} \mathbf{H}|_{\Gamma} \in H^{-1/2}(I, L^2(\Gamma, R^3)). \quad (5.13)$$

**5.4.1. Tangential decomposition.** For any vector field  $G \in H^1(\Omega, R^N)$  we designate by  $G_{\Gamma}$  the tangential part  $G_{\Gamma} = G|_{\Gamma} - \langle G, n \rangle \vec{n}$  and (see ([11], [9], [5], [10])) we consider its tangential Jacobian matrix  $D_{\Gamma}G = D(G_{\Gamma})|_{\Gamma}$  and its transposed  $D_{\Gamma}^*$ . To derive the regularity result we shall be concerned the three following terms at the boundary:

$$(DE.n)_{\Gamma}, \quad \nabla_{\Gamma}(E.n), \quad E_t.$$

**Lemma 5.5.**  $\forall E \in H^2(\Omega, R^N)$ , we have by direct calculus:

$$DE|_{\Gamma} = DE.n \otimes n + D_{\Gamma}E. \quad (5.14)$$

Obviously, as  $E = E_{\Gamma} + \langle E, n \rangle n$ , we have:

$$D_{\Gamma}E = D_{\Gamma}E_{\Gamma} + D_{\Gamma}(E.n n)$$

such that

$$\text{when } E_{\Gamma} = 0, \quad D_{\Gamma}E = D_{\Gamma}(\langle E, n \rangle n). \quad (5.15)$$

Now, as  $D_{\Gamma}(\langle E, n \rangle n) = \langle E, n \rangle D_{\Gamma}(n) + n \otimes \nabla_{\Gamma}(\langle E, n \rangle)$ , and as  $D_{\Gamma}(n) = D^2 b_{\Omega}$ , we get the

**Lemma 5.6.** Assume  $E_{\Gamma} = 0$ , then we have

$$D_{\Gamma}E = \langle E, n \rangle D^2 b_{\Omega}|_{\Gamma} + n \otimes \nabla_{\Gamma}(\langle E, n \rangle).$$

Moreover as

$$\operatorname{div}_{\Gamma} E := \operatorname{div} E|_{\Gamma} - \langle DE.n, n \rangle.$$

When  $\operatorname{div} E = 0$  we get  $\langle DE.n, n \rangle = -\operatorname{div}_{\Gamma} E$ , and if also  $E_{\Gamma} = 0$  we have  $\langle DE.n, n \rangle = -\operatorname{div}_{\Gamma}(\langle E, n \rangle n)$ , that is:

**Lemma 5.7.** We consider the mean curvature  $H = \Delta b_{\Omega}$ , then:

$$E_{\Gamma} = 0, \quad \operatorname{div} E = 0 \quad (5.16)$$

imply

$$\text{i)} \quad \langle DE.n, n \rangle = -H E.n \quad (5.17)$$

$$\text{ii)} \quad DE.n = \langle DE.n, n \rangle n + (DE.n)_{\Gamma} = -H E.n n + (DE.n)_{\Gamma} \quad (5.17)$$

and

$$\text{iii)} \quad |DE.n|^2 = H^2 |E.n|^2 + |(DE.n)_{\Gamma}|^2. \quad (5.18)$$

$$DE = -H E.n n \otimes n + (DE.n)_{\Gamma} \otimes n + E.n D^2 b + n \otimes \nabla_{\Gamma}(E.n) \quad (5.19)$$

$$\text{iv)} \quad DE..DE = H^2 |E.n|^2 + |(DE.n)_{\Gamma}|^2 + |E.n|^2 D^2 b .. D^2 b + |\nabla_{\Gamma}(E.n)|^2. \quad (5.20)$$

**Proposition 5.8.** Let  $E \in H^2(\Omega, R^N)$ ,  $\operatorname{div} E = 0$ ,  $E_{\Gamma=0}$ , then:

$$DE..DE|_{\Gamma} = (H^2 + D^2 b..D^2 b) |E.n|^2 + |(DE.n)_{\Gamma}|^2 + |\nabla_{\Gamma}(E.n)|^2$$

that is

$$DE..DE|_{\Gamma} = |DE.n|^2 + |E.n|^2 D^2 b..D^2 b + |\nabla_{\Gamma}(E.n)|^2. \quad (5.21)$$

### 5.5. DE boundary estimate

We have  $2\epsilon = DE + D^* E$ ,  $2\sigma = DE - D^* E$ , so that

$$DE = \epsilon(E) + \sigma(E).$$

And

$$\|\operatorname{curl} E\|_{L^2(\Gamma, R^3)}^2 \leq 4 \|DE\|_{L^2(\Gamma), R^{N^2}}^2. \quad (5.22)$$

From the decomposition (5.14) we have:

$$\|DE\|_{L^2(I, L^2(\Gamma, R^3))} \leq \|DE.n \otimes n\| + \|D^2 b E.n\|.$$

But

$$\|DE.n \otimes n\|^2 = \int_0^\tau \int_{\Gamma} (DE.n \otimes n) .. (DE.n \otimes n) dt d\Gamma.$$

That is

$$\begin{aligned} \|DE.n \otimes n\|_{L^2(I, L^2(\Gamma, R^3))}^2 &\leq \int_0^\tau \int_{\Gamma} |DE.n|^2 dt d\Gamma \\ &= \int_0^\tau \int_{\Gamma} \{ |(DE.n)_{\Gamma}|^2 + |\langle DE.n, n \rangle|^2 \} dt d\Gamma. \end{aligned}$$

But as  $\langle DE.n, n \rangle = -\langle E, n \rangle D^2 b_{\Omega}$  we get the desired estimate.

### 5.6. Extractor identity

Let  $I = ]0, \tau[$  be the time interval and for  $k \geq 1$  we consider:

$$H^k = C^0(\bar{I}, H^k(\Omega, R^3)) \cap C^1(\bar{I}, H^{k-1}(\Omega, R^3)), \quad (5.23)$$

$$\mathbf{H}^k = \{E \in H^k, \operatorname{div} E = 0, E_{\Gamma} = 0 \text{ on } \Gamma\}. \quad (5.24)$$

Let  $F \in L^2(I, L^2(\Omega, R^3))$ ,  $E_0 \in H^1(\Omega, R^3)$ ,  $E_1 \in L^2(\Omega, R^3)$  with  $\operatorname{div} E_0 = \operatorname{div} E_1 = 0$ .

We consider  $E \in \mathbf{H}^1$  solution to

$$A.E := E_{tt} - \Delta E = F \in L^2(I, L^2(\Omega, R^3)), \quad (5.25)$$

$$E(0) = E_0, \quad E_t(0) = E_1. \quad (5.26)$$

### 5.7. The extractor $e(V)$

Let  $E \in \mathbf{H}^2$ , and  $V \in C^0([0, \tau], C^2(D, R^N))$ ,  $\langle V(t, .), n \rangle = 0$  on  $\partial D$ ; we consider its flow mapping  $T_s = T_s(V)$  and the derivative:

$$\mathbf{e}(V) = \frac{\partial}{\partial s} \{ \mathbf{E}(V, s) \}|_{s=0}, \quad (5.27)$$

where

$$\mathbf{E}(V, s) := \int_0^1 \int_{\Omega_s} ( |E_t oT_s^{-1}|^2 - D(EoT_s^{-1} .. D(EoT_s^{-1})) ) dx dt. \quad (5.28)$$

By change of variable

$$D(EoT_s^{-1})oT_s = DE.DT_s^{-1}. \quad (5.29)$$

We get the second expression

$$\mathbf{E}(V, s) = \int_0^1 \int_{\Omega} ( |E_t|^2 - (DE.DT_s^{-1}) .. (DE.DT_s^{-1}) ) J(t) dx dt. \quad (5.30)$$

We have two expressions (5.28) and (5.30) for the same term  $\mathbf{E}(V, s)$ . The first one is an integral on a mobile domain  $\Omega_s(V)$  while the second one is an integral over the fixed domain  $\Omega$ . Therefore, taking the derivative with respect to the parameter  $s$  we shall obtain two different expressions for  $\mathbf{e}$  that we shall respectively denote by  $\mathbf{e}_1$  and  $\mathbf{e}_2$ .

**5.7.1. Expression for  $\mathbf{e}_1$ .** As the element  $E$  is smooth,  $\mathbf{H}^2$  we can directly apply the classical results from ([11]) and setting For shortness assume  $\operatorname{div} V = 0$  so that  $J(t) = 1$ , in this specific case we get

$$\mathbf{e} = \frac{\partial}{\partial s} \mathbf{E}|_{s=0}$$

we get

$$\begin{aligned} \mathbf{e}_1 &= 2 \int_0^1 \int_{\Omega} \{ E_t \cdot (-DE_t.V) - DE .. D(-DE.V) \} dx dt \\ &\quad + \int_0^1 \int_{\Gamma} \{ |E_t|^2 - DE .. DE \} v d\Gamma dt. \end{aligned} \quad (5.31)$$

**5.7.2. Green-Stokes Theorem.** The by part integration formula applies as:

$$\begin{aligned} \int_0^1 \int_{\Omega} \{ DE .. D(DE.V) \} dx dt &= \int_0^1 \int_{\Omega} \langle -\Delta E, DE.V \rangle dx dt \\ &\quad + \int_0^1 \int_{\Gamma} \langle DE.n, DE.V \rangle d\Gamma(x) dt. \end{aligned} \quad (5.32)$$

**5.7.3. By part time integration.** Then

$$\begin{aligned} \int_0^1 \int_{\Omega} E_t.(DE_t.V) dx dt &= \int_0^1 \int_{\Omega} (-E_{tt}.(DE.V) + E_t.(DE.W)) dx dt \\ &\quad - \int_{\Omega} E_t(0).(DE(0).W) dx. \end{aligned} \quad (5.33)$$

Then assuming the initial condition  $E_0 \in H^1(\Omega, R^3)$ ,  $E_1 \in L^2(\Omega, R^3)$ ,

$$\begin{aligned} e_1 &= 2 \int_0^1 \int_{\Omega} \{ (E_{tt}.(DE.V) - E_t.(DE.W)) - \langle \Delta E, DE.V \rangle \} dx dt \\ &\quad + 2 \int_{\Omega} E_1.(DE_0.W) dx \\ &\quad + \int_0^1 \int_{\Gamma} \{ (|E_t|^2 - DE..DE) \langle V, n \rangle + 2 \langle DE.n, DE.V \rangle \} d\Gamma(x) dt. \end{aligned} \quad (5.34)$$

The discussion is now on the last boundary integral.

**5.7.4. Specific choice for  $V$  at the boundary.** As the boundary  $\Gamma = \partial\Omega \in C^2$  we can apply all intrinsic geometry material introduced in ([5]) and  $p = p_{\Gamma}$  denoting the projection mapping onto the manifold  $\Gamma$  (which is smoothly defined in a tubular neighborhood of  $\Gamma$ ) we consider the oriented distance function  $b = b_{\Omega} = d_{\Omega^c} - d_{\Omega}$  where  $\Omega^c = R^n \setminus \bar{\Omega}$ , and its “localized version” defined as follows (see([8])):

Let  $h > 0$  be “a small” positive number and  $\rho_h(\cdot) \geq 0$  be a cutting scalar smooth function such that  $\rho_h(z) = 0$  when  $|z| > h$ ,  $\rho(z) = 1$  when  $|z| < h/2$ ; then we set

$$b_{\Omega}^h = \rho_h o b_{\Omega}$$

and the associate localized projection mapping

$$p_h = I_d - b_{\Omega}^h \nabla b_{\Omega}^h$$

smoothly defined in the tubular neighborhood  $\mathbf{U}_h(\Gamma) = \{x \in D \text{ s.t. } |b_{\Omega}(x)| < h\}$ . Let be given any smooth element  $v \in C^0(\Gamma)$  we consider the vector field  $V$  in the following form

$$V(t, x) = W(x)(1-t), \quad W(x) = v o p_h \nabla b_{\Omega}^h. \quad (5.35)$$

Then the last term (boundary integral) in (5.34) takes the following form:

$$\int_0^1 \int_{\Gamma} \{ (|E_t|^2 - DE..DE) + 2 \langle DE.n, DE.n \rangle \} v(1-t) d\Gamma(x) dt$$

we get:

$$\begin{aligned} \mathbf{e}_1 &= \int_0^1 \int_{\Gamma} (|E_t|^2 - DE..DE + 2 |DE.n|^2) v d\Gamma dt \\ &\quad + 2 \int_Q (E_{tt}.DE.V - \langle \Delta E, DE.V \rangle) dx dt - \int_{\Omega} \langle E_t(0), DE(0).W \rangle dx. \end{aligned}$$

As from (5.21) we have

$$DE..DE = |DE.n|^2 + D^2 b_\Omega .. D^2 b_\Omega |E.n|^2 + |\nabla_\Gamma E.n)|^2$$

and as

$$|DE.n|^2 = |(DE.n)_\Gamma|^2 + (\Delta b_\Omega)^2 |E.n|^2$$

we obtain:

**Proposition 5.9.**

$$\begin{aligned} \mathbf{e}_1 &= \int_0^1 \int_\Gamma (\tau - t) \left\{ |E_t|^2 + |(DE.n)_\Gamma|^2 \right. \\ &\quad \left. - |\nabla_\Gamma(E.n)|^2 + |E.n|^2(H^2 - D^2 b .. D^2 b) \right\} v d\Gamma dt \\ &\quad + 2 \int_Q \langle A.E, DE.V \rangle dQ - 2 \int_\Omega \langle E_1, D(E_0).W \rangle dx \end{aligned} \quad (5.36)$$

**5.7.5. Second expression for  $\mathbf{e}$ .** From (5.30) we obtain the  $s$  derivative as a distributed integral term as follows

$$\mathbf{e}_2 = \int_Q \{ (|E_t|^2 - DE..DE) \operatorname{div} V(0) - 2 DE..(-DE.DV) \} dx dt$$

**5.7.6. Extractor identity.** As  $\mathbf{e} = \mathbf{e}_1 = \mathbf{e}_2$  we get

$$\begin{aligned} &\int_\Sigma (\tau - t) \{ (|E_t|^2 - |\nabla_\Gamma(E.n)|^2 + |(DE.n)_\Gamma|^2 + |E.n|^2 (H^2 - D^2 b .. D^2 b)) \} v d\Sigma \\ &= \int_Q \{ (|E_t|^2 - DE..DE) \operatorname{div} V - 2 DE..(-DE.DV) \} dx dt \\ &\quad - \int_Q 2(E_{tt} - \Delta E).DE.V dQ + \int_\Omega 2 \langle E_1, DE_0.W \rangle dx; \end{aligned} \quad (5.37)$$

that is

$$\begin{aligned} &\int_\Sigma (\tau - t) \{ (|E_t|^2 - |\nabla_\Gamma(E.n)|^2 + |(DE.n)_\Gamma|^2) \} v d\Sigma \\ &= \int_Q \{ (|E_t|^2 - DE..DE) \operatorname{div} V - 2 DE..(-DE.DV) \} dx dt \\ &\quad - 2 \int_Q 2 \langle A.E, DE.V \rangle dQ + \int_\Omega 2 \langle E_1, DE_0.W \rangle dx \\ &\quad + \int_\Sigma (1-t)|E.n|^2(D^2 b .. D^2 b - H^2)v d\Sigma. \end{aligned} \quad (5.38)$$

Notice that the curvature terms

$$D^2 b .. D^2 b - H^2 = \lambda_1^2 + \lambda_2^2 - (\lambda_1 + \lambda_2)^2 = -2\kappa.$$

where  $\kappa = \lambda_1 \lambda_2$  is the Gauss curvature of the boundary  $\Gamma$ .

## 6. Regularity at the boundary

We shall apply twice this last identity.

### 6.1. Tangential field $E^\tau$

In a first step we consider the “tangential vector field” obtained as

$$E^\tau := E - E \cdot \nabla b_\Omega^h \cdot \nabla b_\Omega^h.$$

We get

$$E_{tt}^\tau - \Delta E^\tau = (E_{tt} - \Delta E) - (E_{tt} - \Delta E) \cdot \nabla b_\Omega^h \cdot \nabla b_\Omega^h + C.E,$$

that is  $A.E^\tau = (A.E)^\tau + C.E.$ , where the commutator is  $C.E \in L^2(0,T;L^2(\Omega,R^3))$  is given by:

$$\begin{aligned} C.E &= -E \cdot \nabla (\Delta b_\Omega^h) \cdot \nabla b_\Omega^h - 2D^2 b_\Omega^h \cdot D E \cdot \nabla b_\Omega^h \\ &\quad - E \cdot \nabla b_\Omega^h \cdot \nabla (\Delta b_\Omega^h) - 2D^2 b_\Omega^h \cdot \nabla (E \cdot \nabla b_\Omega^h). \end{aligned}$$

The conclusion formally derives as follows: as  $E^\tau \in L^2(I, H^1(\Omega, R^3))$  we get the traces terms

$$E^\tau \cdot n = E_t^\tau = 0 \in L^2(I, H^{1/2}(\Gamma)),$$

then as  $e_1 = e_2$  we conclude by taking the vector field in the form

$$V(t, x) = (\tau - t) \nabla b_\Omega^h = (\tau - t) \rho'_h \partial b_\Omega \nabla b_\Omega.$$

That is  $v = 1$ , and we get:

$$\begin{aligned} &\int_0^\tau (\tau - t) \int_\Gamma \{|(DE^\tau \cdot n)_\Gamma|^2\} d\Gamma dt \\ &= \int_Q (\tau - t) \{ (-DE^\tau \cdot DE^\tau) \operatorname{div}(\nabla b_\Omega^h) + 2DE^\tau \cdot (DE^\tau \cdot D(\nabla b_\Omega^h)) \} dx dt \\ &\quad - 2 \int_Q (\tau - t) \langle A.E^\tau, DE^\tau \cdot (\nabla b_\Omega^h) \rangle dQ + \int_\Omega 2 \langle E_1^\tau, DE_0^\tau \cdot (\nabla b_\Omega^h) \rangle dx. \end{aligned}$$

#### 6.1.1. Regularity result for $E^\tau$ .

**Proposition 6.1.** *Let  $\Omega$  be a bounded domain in  $R^3$  with boundary  $\Gamma$  being a  $C^2$  manifold. Let  $h$  verifying the condition (6.3). There exists a constant  $M > 0$  such that for any data  $(E_0, E_1, F) \in L^2(\Omega, R^3) \times H^1(\Omega, R^3) \times L^2(\Omega, R^3)$ , the vector  $E^\tau \in \mathbf{H}^1(0, 2\tau) := C^0([0, 2\tau], H^1(\Omega, R^3)) \cap C^1([0, 2\tau], L^2(\Omega, R^3))$  verifies*

$$(DE^\tau \cdot n)_\Gamma \in L^2(0, \tau, L^2(\Gamma, R^3)) \tag{6.1}$$

and

$$\begin{aligned} &\int_0^T \int_\Gamma \{|(DE^\tau \cdot n)_\Gamma|^2\} d\Gamma dt \\ &\leq M \|\nabla b_\Omega^h\|_{W^{1,\infty}(\Omega, R^N)} \{ \|E\|_{\mathbf{H}^1(0, 2T)}^2 + |F|_{L^2([0, 2T] \times \Omega, R^3)} \\ &\quad + 1/T |E_0^\tau|_{H^1(\Omega, R^3)}^2 + 1/T |E_1^\tau|_{L^2(\Omega, R^3)}^2 \}. \end{aligned} \tag{6.2}$$

Notice that

$$\nabla b_\Omega^h = \rho'_h ob_\Omega \nabla b_\Omega$$

such that

$$\|\nabla b_\Omega^h\|_{L^\infty(R^N, R^N)} \leq \text{Max}_{\{0 \leq s \leq h\}} |\rho'_h(s)|$$

while

$$\begin{aligned} D^2 b_\Omega^h &= D(\rho'_h ob_\Omega \nabla b_\Omega) \\ &= \rho''_h ob_\Omega \nabla b_\Omega \times \nabla b_\Omega + \rho'_h ob_\Omega D^2 b_\Omega \end{aligned}$$

such that

$$\begin{aligned} \|D^2 b_\Omega^h\|_{L^\infty(R^n, R^{N^2})} &\leq \{\text{Max}_{\{0 \leq s \leq h\}} |\rho''_h(s)| \\ &\quad + \text{Max}_{\{0 \leq s \leq h\}} |\rho'_h(s)| \ \|D^2 b_\Omega\|_{L^\infty(\mathbf{U}_h(\Gamma), R^{N^2})} \}. \end{aligned}$$

By the choice of  $\rho_h$  in the form  $\rho_h(s) = f(2s/h - 1)$  when  $h/2 < s < h$  and  $F(x) = 2x^3 - 3x^2 + 1$  we obtain

$$\|\rho_h\|_{C^2([0,h])} \leq \frac{8}{h^2}.$$

Thus the previous estimate is in the form

$$\|D^2 b_\Omega^h\|_{L^\infty(R^n, R^{N^2})} \leq C_0 \frac{1}{h^2} \|D^2 b_\Omega\|_{L^\infty(\mathbf{U}_h(\Gamma), R^{N^2})} \}.$$

For the *larger*  $h$  such that the condition holds

$$D^2 b_\Omega \in L^\infty(\mathbf{U}_h(\Gamma), R^{N^2}). \quad (6.3)$$

## 6.2. Boundary functional

It can be verified that

$$D(E^\tau).n = (DE.n)_\Gamma,$$

such that

$$\begin{aligned} J(\Omega) &:= \int_0^\tau (\tau - t) \int_\Gamma \{|(DE.n)_\Gamma|^2\} d\Gamma dt \\ &= \int_0^\tau (\tau - t) \int_\Gamma \{|(DE^\tau.n)_\Gamma|^2\} d\Gamma dt \\ &= \int_Q (\tau - t) \{ (-DE^\tau .. DE^\tau) \text{div}(\nabla b_\Omega^h) + 2 DE^\tau .. (DE^\tau . D(\nabla b_\Omega^h)) \} dx dt \\ &\quad - 2 \int_Q (\tau - t) \langle F^\tau + C, DE^\tau . (\nabla b_\Omega^h) \rangle dQ + \int_\Omega 2 \langle E_1^\tau, DE_0^\tau . (\nabla b_\Omega^h) \rangle dx. \end{aligned}$$

**6.2.1. The Lagrangian.** Let  $\Omega_s = T_s(\mathbf{Z})(\Omega)$  and we consider  $E^s$  the solution of the previous Maxwell system in the cylinder  $Q^s = \cup_{0 < t < \tau} \{t\} \times \Omega_s$  then

$$\begin{aligned} J(\Omega_s) &= \int_0^\tau (\tau - t) \int_{\Gamma_s} \{|(DE^s \cdot n_s)_{\Gamma_s}|^2\} d\Gamma_s dt \\ &= \text{Min}_{\{(R, \phi) \in K_0^\tau \times K_0\}} \text{Max}_{\{(S, \psi) \in (K_\tau)^2\}} \mathbf{L}(s; R, \phi, S, \psi) \end{aligned}$$

where

$$\begin{aligned} K_0^\tau &= \{R \in \mathbf{H}(0, \tau), R(0) = E_0^\tau\} \\ K_0 &= \{R \in \mathbf{H}(0, \tau), R(0) = E_0\} \\ K_1 &= \{S \in \mathbf{H}(0, \tau), S(\tau) = 0\} \end{aligned}$$

$$\begin{aligned} &\mathbf{L}(s, R, \phi, S, \psi) \\ &= \int_{Q^s} (\tau - t) \{ (-D(RoT_s^{-1})..D(RoT_s^{-1})) \text{div}(\nabla b_\Omega^h) \\ &\quad + 2 D(RoT_s^{-1})..(D(RoT_s^{-1}).D(\nabla b_\Omega^h)) \} dx dt \\ &\quad - 2 \int_{Q^s} (\tau - t) \langle F^\tau + C, D(RoT_s^{-1}).(\nabla b_\Omega^h) \rangle dQ + \int_\Omega 2 \langle E_1^\tau, DE_0^\tau.(\nabla b_\Omega^h) \rangle dx \\ &\quad + \int_{Q^s} ((\phi oT_s^{-1})_t (\psi oT_s^{-1})_t - \langle \nabla((\phi oT_s^{-1})_t), \nabla((\psi oT_s^{-1})_t) \rangle - F(\psi oT_s^{-1})_t) \\ &\quad + \int_{Q^s} (\tau - t) \{ (-D(RoT_s^{-1})..D(RoT_s^{-1})) \text{div}(\nabla b_\Omega^h) \\ &\quad + 2 D(RoT_s^{-1})..(D(RoT_s^{-1}).D(\nabla b_\Omega^h)) \} dx dt \\ &\quad + \int_{\Omega^{s)} (E_1^\tau (SoT_s^{-1})_t + E_1 (\psi oT_s^{-1})_t) dx. \end{aligned} \tag{6.4}$$

where

$$\begin{aligned} C.R &= -R \cdot \nabla(\Delta b_{\Omega_s}^h) \cdot \nabla b_{\Omega_s}^h - 2D^2 b_{\Omega_s}^h .. DR \cdot \nabla b_{\Omega_s}^h \\ &\quad - R \cdot \nabla b_{\Omega_s}^h \cdot \nabla(\Delta b_{\Omega_s}^h) - 2D^2 b_{\Omega_s}^h \cdot \nabla(R \cdot \nabla b_{\Omega_s}^h). \end{aligned}$$

### 6.3. $s$ -derivative

By change of variable  $T_s$  we obtain the expression of the Lagrangian as an integral over the non perturbed domain  $\Omega$  and concerning the element  $R$  and  $S$  (rather than  $RoT_s^{-1}$  and  $SoT_s^{-1}$ ). Notice that

$$\begin{aligned} (C.(RoT_s^{-1}))oT_s &= -\langle R, DT_s^{-*} \cdot \nabla((\Delta b_{\Omega_s}^h)oT_s) \rangle DT_s^{-*} \cdot \nabla(b_{\Omega_s}^h oT_s) \\ &\quad - 2(D^2 b_{\Omega_s}^h)oT_s .. (DR)oDT_s^{-1} DT_s^{-*} \cdot \nabla(b_{\Omega_s}^h oT_s) \\ &\quad - \langle R, DT_s^{-*} \cdot \nabla(b_{\Omega_s}^h oT_s) \rangle DT_s^{-*} \cdot \nabla(\Delta b_{\Omega_s}^h oT_s) \\ &\quad - 2D^2 b_{\Omega_s}^h oT_s \cdot \{ DT_s^{-*} \cdot \nabla(\langle DT_s^{-1}.R, DT_s^{-*} \cdot \nabla(b_{\Omega_s}^h oT_s) \rangle) \} \end{aligned}$$

Obviously as soon as  $E$  and  $P$  are in the space energy  $H(Q)$  the functional  $J(Q^s)$  is differentiable as soon as all the geometrical terms are differentiable in  $L^\infty(D)$ , which is true for a  $C^3$  boundary.

#### 6.4. Gradient calculus

As in the previous wave example, the moving domain derivative of (6.4) is easier. We get

$$\begin{aligned}
 & dJ(Q; \mathbf{Z}) \\
 &= \int_0^\tau \int_{\partial\Omega} (\tau - t) \{ (-DE^\tau .. DE^\tau) \operatorname{div}(\nabla b_\Omega^h) \\
 &\quad + \int_0^\tau \int_{\partial\Omega} 2DE^\tau .. (DE^\tau . D^2 b_\Omega^h) \} z(t, x) d\Gamma(x) dt \\
 &\quad - 2 \int_0^\tau \int_{\partial\Omega} (\tau - t) \{ \langle DE^\tau . n, [2D^2 b_\Omega^h - \Delta b_\Omega^h I_d] . DE^\tau . n \rangle z(t, x) d\Gamma(x) dt \\
 &\quad - 2 \int_0^\tau \int_{\partial\Omega} (\tau - t) \langle F^\tau + \{ E^\tau . \nabla \Delta b . n - 2(D^2 b .. DE^\tau) n \\
 &\quad - E . n \nabla \Delta b - 2D^2 b \nabla_\Gamma(E, n) \}, DE^\tau . n \rangle z(t, x) d\Gamma(x) dt \\
 &\quad + 2 \int_0^\tau \int_{\partial\Omega} (\tau - t) \langle F^\tau + C.E, DE^\tau . n \rangle z(t, x) d\Gamma(x) dt \\
 &\quad + \int_0^\tau \int_{\partial\Omega} (E_t P_t - DE^\tau .. DP - \langle C.E, P \rangle) z(t, x) d\Gamma(x) dt \\
 &\quad + 2 \int_0^\tau \int_{\partial\Omega} DE^\tau . n DP . n z(t, x) d\Gamma(x) dt + \int_{\partial\Omega} E_1 P_t(0) z(t, x) d\Gamma(x) dt.
 \end{aligned} \tag{6.5}$$

## 7. The normal vector field $e$

We set

$$e = E . \nabla b_\Omega^h$$

### Lemma 7.1.

$$e_{tt} - \Delta e = (E_{tt} - \Delta E) . \nabla b_\Omega^h + \theta . E, \tag{7.1}$$

where

$$\theta . E = D^2 b_\Omega^h .. DE + \operatorname{div}(D^2 b_\Omega^h . E) \tag{7.2}$$

$$\frac{\partial}{\partial n} e = \langle DE . n, n \rangle = -\Delta b_\Omega e \text{ on } \Gamma. \tag{7.3}$$

Then  $e$  solves the wave problem with *curvature Fourier-like* boundary condition:

$$e_{tt} - \Delta e = F . \nabla b_\Omega^h + D^2 b_\Omega^h .. DE + \operatorname{div}(D^2 b_\Omega^h . E) = F_n + \theta \tag{7.4}$$

$$\frac{\partial}{\partial n} e + \Delta b_\Omega e = 0, \quad e(0) = E_0 . \nabla b_\Omega^h, \quad e_t(0) = E_1 . \nabla b_\Omega^h,$$

where  $F_n = F \cdot \nabla b_\Omega^h$ , we shall use the notation

$$\Theta = F \cdot \nabla b_\Omega^h + \theta$$

and  $\Delta b_\Omega$  is the mean curvature  $\frac{1}{R_1} + \frac{1}{R_2}$  of the surface  $\Gamma$ .

### 7.1. Extension to $R$

Let

$$\rho \in C^2(R), \quad \rho \geq 0, \quad \text{supp } \rho \subset [-2\tau, +2\tau], \quad \rho = 1 \text{ on } [-\tau, +\tau].$$

We set

$$\tilde{e} = \rho(t) e(t), \quad t \geq 0, \quad = \rho(t)(e_0 + t e_1), \quad t \leq 0,$$

which turns out to be solution on  $R$  to the wave problem

$$\tilde{e}_{tt} - \Delta \tilde{e} = H, \quad \frac{\partial}{\partial n} \tilde{e} = g,$$

where

$$g = -\Delta b_\Omega \tilde{e} \text{ on } \Gamma \text{ for } t > 0.$$

$$g = \rho(t) \left( \frac{\partial}{\partial n} e_0 + t \frac{\partial}{\partial n} e_1 \right) \text{ on } \Gamma \text{ for } t < 0,$$

where  $H \in L^2(R, L^2(\Omega))$  verifies

$$\begin{aligned} H &= \rho(t)\Theta + \rho'' e + 2\rho' \frac{\partial}{\partial t} e \quad \text{if } t > 0 \\ H &= \rho''(e_0 + t e_1) + 2\rho' e_1 - \rho(\Delta e_0 + t \Delta e_1) \quad \text{if } t < 0 \end{aligned}$$

## 8. Fourier transform

We consider

$$z(\zeta)(x) = \int_{-\infty}^{+\infty} \exp(-i\zeta t) \tilde{e}(t, x) dt \tag{8.1}$$

which turns out to be a solution to

$$\frac{\partial}{\partial n} z = \mathbf{F}.g \text{ on } \Gamma.$$

We consider the perturbed domain  $\Omega_s = T_s(V)(\Omega)$  with boundary  $\Gamma_s = T_s(V)(\Gamma)$ , and

$$\mathbf{E}(s, V) = \int_{-\infty}^{+\infty} d\zeta \int_{\Omega_s(V)} |\zeta| |zoT_s(V)^{-1}|^2 + \frac{1}{1+|\zeta|} |\nabla(zoT_s(V)^{-1})|^2 dx.$$

Moreover

$$e = \left( \frac{d}{ds} \mathbf{E}(s, V) \right)_{s=0}.$$

We compute this by two different ways:

**8.0.1. By moving domain derivative.** Let

$$\begin{aligned} e_1 &= \int_{-\infty}^{+\infty} d\zeta \int_{\Omega} |\zeta| 2\operatorname{Re}\{\langle z, \nabla \bar{z}.(-V) \rangle\} + \frac{1}{1+|\zeta|} 2\operatorname{Re}\{\langle \nabla z, \nabla(\nabla \bar{z}.(-V)) \rangle\} ) dx \\ &\quad + \int_{-\infty}^{+\infty} d\zeta \left( \int_{\Gamma} \{ |\zeta| |z|^2 + \frac{1}{1+|\zeta|} |\nabla z|^2 \} \langle V, n \rangle d\Gamma(x) \right). \end{aligned}$$

By Stokes Theorem we get,

$$\begin{aligned} &\int_{-\infty}^{+\infty} d\zeta \int_{\Omega} \frac{1}{1+|\zeta|} 2\operatorname{Re}\{\nabla(z).\nabla(\nabla \bar{z}.(-V))\} ) dx \\ &= \int_{-\infty}^{+\infty} d\zeta \int_{\Omega} \frac{1}{1+|\zeta|} 2\operatorname{Re}\{\Delta(z), (\nabla \bar{z}.V)\} ) dx \\ &\quad - \int_{-\infty}^{+\infty} \int_{\Gamma} \frac{1}{1+|\zeta|} 2\operatorname{Re}\{\langle \nabla z.n, \nabla \bar{z}.V \rangle\} ). \end{aligned}$$

As  $V = vn$  on  $\Gamma$  we get for the last term:

$$- \int_{-\infty}^{+\infty} \int_{\Gamma} \frac{1}{1+|\zeta|} 2\operatorname{Re}\{\langle \nabla z.n, \nabla \bar{z}.n \rangle\} ) v.$$

But on  $\Gamma$  we have

$$\langle \nabla z.n, \nabla \bar{z}.n \rangle = |\mathbf{F}.g|^2$$

Finally we get

$$\begin{aligned} e_1 &= \int_{-\infty}^{+\infty} \int_{\Gamma} \{ |\zeta| |z|^2 + \frac{1}{1+|\zeta|} \operatorname{Re}\{\langle \nabla z, \nabla \bar{z} \rangle\} - 2|\mathbf{F}.g|^2 \} v \\ &\quad + \int_{-\infty}^{+\infty} d\zeta \int_{\Omega} (|\zeta| 2\operatorname{Re}\{\langle z, \nabla \bar{z}.(-V) \rangle\} + \frac{1}{1+|\zeta|} 2\operatorname{Re}\{\Delta z, (\nabla \bar{z}.V)\}) ) dx. \end{aligned}$$

Then

$$\begin{aligned} &\int_{-\infty}^{+\infty} \int_{\Gamma} \{ |\zeta| |z|^2 + \frac{1}{1+|\zeta|} |\nabla_{\Gamma} z|^2 \} ) v \\ &= \int_{-\infty}^{+\infty} \int_{\Gamma} \frac{1}{1+|\zeta|} |\mathbf{F}.g|^2 d\Gamma dt \\ &\quad - \int_{-\infty}^{+\infty} d\zeta \int_{\Omega} \frac{1}{1+|\zeta|} 2\operatorname{Re}\{|\zeta|^2 z + \mathbf{F}.H)(\nabla \bar{z}.V)\} ) dx \\ &\quad + \int_{-\infty}^{+\infty} d\zeta \int_{\Omega} (|\zeta| 2\operatorname{Re}\{\langle z, \nabla \bar{z}.V \rangle\} + e_2). \end{aligned}$$

From which there exists  $M > 0$  such that

$$\int_{-\infty}^{+\infty} \int_{\Gamma} \{ |\zeta| |z|^2 + \frac{1}{1+|\zeta|} |\nabla_{\Gamma} z|^2 \} ) v \leq M \{ \|z\|_{L^2(R, H^1(\Omega))}^2 + \|z\|_{L^2(R, L^2(\Gamma))}^2 \}.$$

We have  $\sqrt{|\zeta|}z \in L^2(R_\zeta, L^2(\Gamma))$  and  $\frac{1}{\sqrt{|\zeta|}}\nabla_\Gamma z \in L^2(R_\zeta, L^2(\Gamma, R^N))$ , we conclude that

$$\begin{aligned} E.n &\in H^{1/2}(I, L^2(\Gamma)) \cap L^2(I, H^{1/2}(\Gamma)) \\ \nabla_\Gamma(E.n) &\in H^{-1/2}(I, L^2(\Gamma, R^N)). \end{aligned}$$

### 8.1. Expression by change of variable

We obtain

$$\begin{aligned} \mathbf{E}(s, V) &= \int_{-\infty}^{+\infty} d\zeta \int_{\Omega} ( |\zeta| |z|^2 J(s) + \frac{1}{1+|\zeta|} |DT_s^{-*} \cdot \nabla z|^2 |^2 J(s) ) dx \\ e_2 &= \int_{-\infty}^{+\infty} d\zeta \left( \int_{\Omega} \{ |\zeta| |z|^2 + \frac{1}{1+|\zeta|} |\nabla z|^2 \} \operatorname{div} V(0) - 2 \frac{1}{1+|\zeta|} \langle DV(0).z, \nabla z \rangle \right) dx \end{aligned}$$

#### 8.1.1. The shape functional.

$$J(\Omega) := \int_{-\infty}^{+\infty} \int_{\Gamma} \{ |\zeta| |z|^2 + \frac{1}{1+|\zeta|} |\nabla_\Gamma z|^2 \} d\Gamma d\zeta$$

Then, taking  $V = \nabla b_\Omega^h$ ,  $v = \langle V(0), n \rangle = 1$  on  $\partial\Omega$ , we get

$$\begin{aligned} J(\Omega) &= \int_{-\infty}^{+\infty} \int_{\Gamma} \frac{1}{1+|\zeta|} |\mathbf{F}.g|^2 d\Gamma dt \\ &\quad - \int_{-\infty}^{+\infty} d\zeta \int_{\Omega} \frac{1}{1+|\zeta|} 2\operatorname{Re}\{ |\zeta|^2 z + \mathbf{F}.H)(\nabla \bar{z} \cdot \nabla b_\Omega^h) \} dx \\ &\quad + \int_{-\infty}^{+\infty} d\zeta \int_{\Omega} ( |\zeta| 2\operatorname{Re}\{ \langle z, \nabla \bar{z} \cdot \nabla b_\Omega^h \rangle \} \\ &\quad + \int_{-\infty}^{+\infty} d\zeta \left( \int_{\Omega} \{ |\zeta| |z|^2 + \frac{1}{1+|\zeta|} |\nabla z|^2 \} \operatorname{div} V(0) \right. \\ &\quad \left. - 2 \frac{1}{1+|\zeta|} \langle D\nabla b_\Omega^h.z, \nabla z \rangle \right) dx \end{aligned}$$

with

$$\Omega_s = T_s(\Omega), \quad T_s = T_s(Z).$$

We get:

$$J(\Omega) = \operatorname{Min}_{\{(\phi, \Phi) \in K_0\}} \operatorname{Max}_{\{(\psi, \Psi) \in K_\tau\}} \mathbf{L}(s, (\phi, \Phi), (\psi, \Psi)),$$

where, with the notation

$$\begin{aligned} H_0^i &= C^i([0, \tau], H_0^1(\Omega)) \cap C^{i+1}([0, \tau], L^2(\Omega)) \\ H^i &= \{ \Phi \in C^i([0, \tau], H^1(\Omega)) \cap C^{i+1}([0, \tau], L^2(\Omega)), \Phi_\Gamma = 0 \} \\ K_0 &= \{ (\phi, \Phi) \in H^1 \times H_0^3, \phi(0) = E_0.n, \Phi(0) = E_0 \} \\ K_\tau &= \{ (\psi, \Psi) \in H^1 \times H_0^3, \psi(\tau) = 0, \Psi(\tau) = 0 \} \end{aligned}$$

$$\begin{aligned}
& \mathbf{L}(s, (\phi, \Phi), (\psi, \Psi)) \\
&= \int_{-\infty}^{+\infty} \int_{\Gamma_s(Z)} \frac{1}{1+|\zeta|} |\mathbf{F}.g|^2 d\Gamma_s dt \\
&\quad - \int_{-\infty}^{+\infty} d\zeta \int_{\Omega_s(Z)} \frac{1}{1+|\zeta|} 2\mathbf{Re}\{|\zeta|^2 zoT_s^{-1} + \mathbf{F}.H)(\nabla zo\bar{T}_s^{-1}.\nabla b_{\Omega_s(Z)}^h)\} dx \\
&\quad + \int_{-\infty}^{+\infty} d\zeta \int_{\Omega_s(Z)} (|\zeta| 2\mathbf{Re}\{\langle zoT_s^{-1}, \nabla zo\bar{T}_s^{-1}.\nabla b_{\Omega_s(Z)}^h \rangle\}) \\
&\quad + \int_{-\infty}^{+\infty} d\zeta \left( \int_{\Omega_s(Z)} \{|\zeta| |zoT_s^{-1}|^2 + \frac{1}{1+|\zeta|} |\nabla zoT_s^{-1}|^2\} \operatorname{div} V(0) \right. \\
&\quad \left. - 2 \frac{1}{1+|\zeta|} \langle D^2 b_{\Omega_s(Z)}^h . zoT_s^{-1}, \nabla zoT_s^{-1} \rangle \right) dx \\
&\quad + \int_0^\tau \int_{\Omega_s(Z)} ((eoT_s^{-1})_t(\theta.[\psi oT_s^{-1}])_t - \langle \nabla(\theta.[eoT_s^{-1}]), \nabla(\theta[\psi oT_s^{-1}]) \rangle) dxdt \\
&\quad + \int_0^\tau \int_{\partial\Omega_s(Z)} \Delta b_{\Omega_s(Z)}^h \langle \Phi oT_s^{-1}, n_s \rangle \phi oT_s^{-1} d\Gamma(x) dt \\
&\quad + \int_0^\tau \int_{\Omega_s(Z)} \theta_s.[\Phi oT_s^{-1}] \psi oT_s^{-1} d\Gamma(x) dt \\
&\quad + \int_0^\tau \int_{\Omega_s(Z)} ((\Phi oT_s^{-1})_t.(\Psi oT_s^{-1})_t - D(\Phi oT_s^{-1})..D(\Psi oT_s^{-1}) - F.\Psi oT_s^{-1}) dxdt.
\end{aligned}$$

Notice that

$$\theta_s.[\Phi oT_s^{-1}] = D^2 b_{\Omega_s}^h .. D(\Phi oT_s^{-1}) + \operatorname{div}(D^2 b_{\Omega_s}^h . \Phi oT_s^{-1})$$

such that

$$\theta_s.[\Phi oT_s^{-1}]oT_s = D^2 b_{\Omega_s}^h oT_s .. D(\Phi).(DT_s)^{-1} + \operatorname{div}(D^2 b_{\Omega_s}^h . \Phi oT_s^{-1}).$$

Thus, with (8.1), and the change of variable  $T_s$  we get:

$$\begin{aligned}
& J(\Omega_s) \\
&= \int_{-\infty}^{+\infty} d\zeta \int_{\Omega} \frac{1}{1+|\zeta|} 2\mathbf{Re}\{|\zeta|^2 z + \mathbf{F}.H)(DT_s^{-*}.\nabla \bar{z}, DT_s^{-1}.\nabla(b_{\Omega_s oT_s}^h)\} J(s) dx \\
&\quad + \int_{-\infty}^{+\infty} d\zeta \int_{\Omega} (|\zeta| 2\mathbf{Re}\{\langle z(DT_s^{-1}.\nabla \bar{z}), DT_s^{-1}.\nabla b_{\Omega_s}^h oT_s \rangle\}) J(s) \\
&\quad + \int_{-\infty}^{+\infty} d\zeta \left( \int_{\Omega} \{|\zeta| |z|^2 + \frac{1}{1+|\zeta|} |DT_s^{-1}.\nabla z|^2\} \operatorname{div} V(0) J(s) \right. \\
&\quad \left. - 2 \frac{1}{1+|\zeta|} \langle D^2 b_{\Omega_s}^h oT_s, DT_s^{-1}.\nabla z \rangle J(s) \right) dx \\
&\quad + \int_0^\tau \int_{\Omega} (\phi_t(D^2 b_{\Omega_s}^h oT_s .. D(\Phi).(DT_s)^{-1} + \operatorname{div}(D^2 b_{\Omega_s}^h . \Phi))_t
\end{aligned}$$

$$\begin{aligned}
& - \langle \nabla(D^2 b_{\Omega_s}^h o T_s .. D(\Phi).(DT_s)^{-1} + \operatorname{div}(D^2 b_{\Omega_s}^h . \Phi o T_s^{-1})), \nabla(\theta[\psi o T_s^{-1}]) \rangle J(s) dx dt \\
& + \int_0^\tau \int_\Omega (\Phi_t . \Psi_t - (DT_s^{-1} . D\Phi) .. (DT_s^{-1} . D\Psi)) J(s) dx dt \\
& + \dots .
\end{aligned}$$

And we would get by similar calculus the expression of

$$dJ(\Omega; Z) = \frac{\partial}{\partial s} J(\Omega_s)|_{s=0}.$$

## References

- [1] M. Cuer, J.-P. Zolésio Control of singular problem via differentiation of a min-max, *Systems & Control Letters*, vol. 11, no. 2, pp. 151–158, 1988.
- [2] M.C. Delfour, J.-P. Zolésio Pseudodifferential extractor and hidden regularity, in preparation.
- [3] John Cagnol and Jean-Paul Zolésio. Hidden shape derivative in the wave equation. In *Systems modeling and optimization (Detroit, MI, 1997)*, vol. 396 of *Chapman & Hall/CRC Res. Notes Math.*, pp. 42–52. Chapman & Hall/CRC, Boca Raton, FL, 1999.
- [4] John Cagnol and Jean-Paul Zolésio. Shape control in hyperbolic problems. In *Optimal control of partial differential equations (Chemnitz, 1998)*, vol. 133 of *Internat. Ser. Numer. Math.*, pp. 77–88. Birkhäuser, Basel, 1999.
- [5] M.C. Delfour and J.-P. Zolésio. *Shapes and geometries*. Advances in Design and Control. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2001. Analysis, differential calculus, and optimization.
- [6] Michel C. Delfour and Jean-Paul Zolésio. Curvatures and skeletons in shape optimization. *Z. Angew. Math. Mech.*, vol. 76, no. 3, pp. 199–202, 1996.
- [7] Michel C. Delfour and Jean-Paul Zolésio. Hidden boundary smoothness for some classes of differential equations on submanifolds. In *Optimization methods in partial differential equations (South Hadley, MA, 1996)*, vol. 209 of *Contemp. Math.*, pp. 59–73. Amer. Math. Soc., Providence, RI, 1997.
- [8] Michel C. Delfour and Jean-Paul Zolésio. *Oriented Distance Function and Its Evolution Equation for Initial Sets with Thin Boundary*. In *SIAM J. Optim.*, vol. 42, no. 6, pp. 2286–2304, 2003.
- [9] B. Kawohl, O. Pironneau, L. Tartar, and J.-P. Zolésio. *Optimal shape design*, vol. 1740 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2000.
- [10] M. Moubachir and J.-P. Zolésio. *Moving Shape Analysis and Control, Application to fluid structure interaction*. Pure and Applied Mathematics, 277. Chapman and Hall/CRC, Taylor and Francis Group, Boca Raton, FL, 2006. A Serie of Monograph and Textbooks.
- [11] Jan Sokołowski and Jean-Paul Zolésio. *Introduction to shape optimization*, vol. 16 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1992. Shape sensitivity analysis.

- [12] Jean-Paul Zolésio. Material Derivative. in Shape Analysis In *Optimization of distributed parameter structures, Vol. II (Iowa City, Iowa, 1980)*, vol. 50 of *NATO Adv. Study Inst. Ser. E: Appl. Sci.*, pp. 1457–1473. Nijhoff, The Hague, 1981.
- [13] John E. Lagnese. Exact Boundary Controllability of Maxwell's Equations in a General Region In *SIAM J. Control and Optimiz.*, vol. 27, no. 2, pp. 374–388, 1989.

Jean-Paul Zolésio  
CNRS-INLN and INRIA  
Sophia Antipolis, France.  
e-mail: [jean-paul.zolesio@inln.cnrs.fr](mailto:jean-paul.zolesio@inln.cnrs.fr)