LECTURE NOTES IN COMPUTATIONAL SCIENCE AND ENGINEERING 101

Ronald Hoppe *Editor*

Optimization with PDE Constraints

ESF Networking Program 'OPTPDE'

Springer

Lecture Notes
in Computational Science
and Engineering

Editors:

More information about this series at
http://www.springer.com/series/3527

Ronald Hoppe
Editor

Optimization with PDE Constraints

ESF Networking Program 'OPTPDE'

Springer

Editor
Ronald Hoppe
Institut für Mathematik
Universität Augsburg
Augsburg
Germany

ISSN 1439-7358 ISSN 2197-7100 (electronic)
ISBN 978-3-319-08024-6 ISBN 978-3-319-08025-3 (eBook)
DOI 10.1007/978-3-319-08025-3
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014948961

Mathematics Subject Classification (2010): 49J20, 49J52, 49K20, 49M05, 65K10

Front cover: Ariane 5 flight VA208 (detail), ©ESA/CNES/ARIANESPACE-Optique Video du CSG. By kind authorization of the European Space Agency (ESA).

Printed on acid-free paper

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

Preface

This volume contains contributions on the history, mathematical analysis, and numerical solution of constrained optimal control and optimization problems where a partial differential equation (PDE) or a system of PDEs appears as an essential part of the constraints. The appropriate treatment of such problems requires a fundamental understanding of the subtle interplay between optimization in function spaces and numerical discretization techniques and relies on advanced methodologies from the theory of PDEs and numerical analysis as well as scientific computing. The contributions reflect part of the work that has been done within the European Science Foundation (ESF) Networking Programme optimization with PDEs (OPTPDE). The OPTPDE programme has been launched in October 2008 for a 5-year period and has been supported by seventeen national science foundations and research institutions from 12 European countries:

Fonds zur Förderung der wissenschaftlichen Forschung in Österreich
Austrian Science Research Fund, Austria
Fonds National de la Recherche Scientifique (FNRS)
National Fund for Scientific Research, Belgium
Akademie ved Ceske republiky (GACR)
Academy of Sciences of the Czech Republic, Czech Republic
Suomen Akatemia/Finlands Akademi
Academy of Finland, Finland
Deutsche Forschungsgemeinschaft (DFG)
German Research Foundation, Germany
Istituto Nazionale di Alta Matematica (INdAM)
National Institute for Advanced Mathematics, Italy,
Scuola Internazionale dei Studi Avanzati (SISSA)
Universita di Roma Tor Vegata (Dip. di Matematica),
Dip di Matematica F. Brioschi di Politecnico di Milano,
Universita di Padova,

Universita di Roma La Sapienza (Dip. di Matematica), Italy
Fonds National de la Recherche (FNR)
National Research Fund, Luxembourg
Polska Akademia Nauk (PAN)
Polish Academy of Sciences, Poland
Ministerio de Educacion y Ciencia (MEC)
Ministry of Education and Science, Spain
Vetenskapsradet (VR)
Swedish Research Council, Sweden
Schweizerischer Nationalfonds (SNF)
Swiss National Science Foundation, Switzerland
Engineering and Physical Sciences Research Council (EPSRC), United Kingdom

A primary goal was to bring together experts from the optimization/optimal control and the numerical PDE communities and to use the emerging synergies to make significant progress in the mathematical treatment of challenging problems in PDE constrained optimization. To this end, a series of conferences on general PDE constrained optimization and workshops on specific topics have been organized that have both deepened existing cooperations and triggered new scientific relations between the participants. The contributions in this volume are original, peer-reviewed research articles by participants of the ESF OPTPDE Programme and their coworkers:

The contribution by Petr Beremlijski, Jaroslav Haslinger, Jiři L. Outrata, and Róbert Pathó deals with the numerical solution of shape optimization in frictional contact mechanics. The essential idea is to transform the discretized problem to a nonsmooth minimization problem which is then solved by a bundle trust method using the generalized differential calculus of Mordukhovich. Phase field methods for the recovery of a binary function from blurred and noisy data are a significant task in image processing and optimal control of PDEs. The article by Charles Brett, Charles M. Elliott, and Andreas S. Dedner presents an approach to the numerical solution of this inverse problem based on a combination of the Mumford-Shah model and a phase field approximation to the perimeter regularization.

Multigrid methods and adaptive sequential quadratic programming are two novel techniques for the numerical solution of shape optimization and optimal control problems associated with evolutionary partial differential equations that are applied in the contribution by Maurizio Falcone and Marco Verani. The approximation relies on the coupling between a proper orthogonal decomposition and the classical dynamic programming approach.

Adaptive finite elements for PDE constrained optimization represent a subject that emerged from the cooperation between the optimization and numerical analysis communities. The article by Alexandra Gaevskaya, Michael Hintermüller, Ronald H.W. Hoppe, and Caroline Löbhard is concerned with such techniques for the numerical solution of optimally controlled elliptic variational inequalities. Based on the equivalence with Mathematical Problems with Complementarity Constraints

(MPCCs), the convergence of discrete stationary points to a stationary point in function space is shown. Moreover, a residual-type a posteriori error estimator is developed which up to data oscillations provides both an upper and a lower bound for the global discretization error.

The history of constrained optimization is the subject of the contribution by Martin Gander, Felix Kwok, and Gerhard Wanner. Referring to a lot of original sources, the authors forge a bridge from the origins in the eighteenth century (Varignon, Johann Bernoulli, Lagrange) to Pontryagin's celebrated maximum principle and the modern theory of PDE constrained optimization.

Topology optimization using the concept of topological derivatives is a powerful tool for the optimal design in solid mechanics. The article by S.M. Giusti, Jan Sokolowski, and Jan Stebel deals with the application of this concept to frictionless contact problems by minimizing the structural compliance for a given amount of material. Several numerical examples demonstrate the robustness of the suggested approach.

The numerical solution of three-dimensional contact problems with orthotropic friction by using the Coulomb friction cone without any approximation is the main goal of the contribution by Jaroslav Haslinger, Radek Kučera, and Tomáš Kozubek. The suggested algorithm relies on an appropriate discretization of the dual variational formulation involving the Lagrange multipliers on the contact boundary. Its performance is illustrated by the documentation of numerical results for several model examples.

The article by Günter Leugering, Jan Sokolowski, and Antoni Żochowski is concerned with shape-topological differentiability of energy-type objective functionals for unilateral problems in domains with cracks. Using tools from nonsmooth analysis, the authors study the dependence of the Griffith shape functional on domain perturbations far from the cracks and obtain its directional shape and topological derivatives with respect to boundary variations of an inclusion.

The boundary stabilization for finite difference semidiscretizations of the one-dimensional wave equation with variable density and diffusion coefficients is the central theme of the article by Aurora Marica and Enrique Zuazua. Adding a suitable artificial viscosity to the finite difference approximation, by an application of the classical multiplier technique at the discrete level it is shown that the discrete decay rate is uniform as the mesh size tends to zero.

The theory of a posteriori error estimates of functional type is known to provide guaranteed upper and lower bounds for the global discretization error. In the contribution by Pekka Neittaanmäki and Sergey Repin, the theory is applied to finite element discretizations of distributed optimal control problems for second order elliptic boundary value problems. In this way, guaranteed bounds for the cost functional as well as computable error estimates for the state and the control functions are obtained.

A shape sensitivity analysis of the work functional for the compressible Navier-Stokes equations in a bounded domain with an obstacle is the subject of the contribution by Pavel I. Plotnikov and Jan Sokolowski. The main tool in the analysis is the Kuratowski-Mosco convergence of sequences of compact sets.

Moreover, establishing the continuity of typical cost functionals with respect to the Kuratowski-Mosco convergence, it is shown that the problem of minimizing the work of hydrodynamic forces in a class of obstacles with fixed volume admits a solution.

The contribution by Jean-Pierre Puel provides new results on local exact controllability and null controllability for the incompressible Navier-Stokes equations. The presented approach is based on new global Carleman estimates for the Stokes problem associated with the linearized equations and some fine interpolation results.

The editor would like to express his sincere thanks to those who have made it possible to produce this book. Particular thanks go to the European Science Foundation for including OPTPDE as one of the ESF PESC Networking Programmes and to the European national science foundations and research institutions listed above for their financial support. I am also indebted to the editors of the Lecture Notes in Computational Science and Engineering for considering this book as a volume within this series and to Claus Ascheron of Springer-Verlag for his continuous advice and support during the preparation and production of this volume.

Augsburg, Germany Ronald H.W. Hoppe
January 2014

Contents

Contributors

P. Beremlijski Centre of Excellence, IT4 Innovations and Department of Applied Mathematics, VŠB-Technical University of Ostrava, Ostrava-Poruba, Czech Republic

C. Brett Mathematics Institute, University of Warwick, Coventry, UK

A.S. Dedner Mathematics Institute, University of Warwick, Coventry, UK

C.M. Elliott Mathematics Institute, University of Warwick, Coventry, UK

M. Falcone Dipartimento di Matematica, SAPIENZA-Università di Roma, Roma, Italy

A. Gaevskaya Institute of Mathematics, University of Augsburg, Augsburg, Germany

M. Gander Section de Mathématiques, Université de Genève, Genève, Switzerland

S.M. Giusti Facultad Regional Córdoba, Departamento de Ingenieria Civil, Universidad Tecnológica Nacional (UTN/FRC-CONICET), Córdoba, Argentina

J. Haslinger Department of Numerical Mathematics, Charles University in Prague, Praha 8, Czech Republic

Centre of Excellence, IT4Innovations and Department of Applied Mathematics, VŠB-Technical University of Ostrava, Ostrava-Poruba, Czech Republic

M. Hintermüller Department of Mathematics, Humboldt-Universität zu Berlin, Berlin, Germany

R.H.W. Hoppe Institute of Mathematics, University of Augsburg, Augsburg, Germany

Department of Mathematics, University of Houston, Houston, TX, USA

T. Kozubek Centre of Excellence IT4I, VŠB-TUO, Ostrava, Czech Republic

R. Kučera Centre of Excellence IT4I, VŠB-TUO, Ostrava, Czech Republic

F. Kwok Section de Mathématiques, Université de Genève, Genève, Switzerland

G. Leugering Department Mathematik, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

C. Löbhard Department of Mathematics, Humboldt-Universität zu Berlin, Berlin, Germany

A. Marica Institute for Mathematics and Scientific Computing, University of Graz, Graz, Austria

P. Neittaanmäki Department Mathematical Information Technology, University of Jyväskylä, Jyväskylä, Finland

J. Outrata Institute of Information Theory and Automation, Czech Academy of Sciences, Prague 8, Czech Republic

R. Pathó Department of Numerical Mathematics, Charles University in Prague, Praha 8, Czech Republic

P.I. Plotnikov Lavrentyev Institute of Hydrodynamics, Novosibirsk, Russia

J.-P. Puel Laboratoire de Mathématiques de Versailles, Université de Versailles, Versailles Cedex, France

S. Repin Department Mathematical Information Technology, University of Jyväskylä, Jyväskylä, Finland

J. Sokolowski Laboratoire de Mathématiques, Institut Élie Cartan, UMR7502 (Université Lorraine, CNRS, INRIA), Université de Lorraine, Vandoeuvre-lès-Nancy Cedex, France

J. Stebel Institute of Mathematics of the Academy of Sciences of the Czech Republic, Praha 1, Czech Republic

M. Verani MOX - Modelling and Scientific Computing - Dipartimento di Matematica, Politecnico di Milano, Milano, Italy

G. Wanner Section de Mathématiques, Université de Genève, Genève, Switzerland

A. Żochowski Systems Research Institute of the Polish Academy of Sciences, Warszawa, Poland

E. Zuazua BCAM-Basque Center for Applied Mathematics, Bilbao, Basque Country, Spain

Ikerbasque - Basque Foundation for Science, Bilbao, Basque Country, Spain

Numerical Solution of 2D Contact Shape Optimization Problems Involving a Solution-Dependent Coefficient of Friction

Petr Beremlijski, Jaroslav Haslinger, Jiří Outrata, and Róbert Pathó

Abstract This contribution deals with numerical solution of shape optimization problems in frictional contact mechanics. The state problem in our case is given by 2D static Signorini problems with Tresca friction and a solution-dependent coefficient of friction. A suitable Lipschitz continuity assumption on the coefficient of friction is made, ensuring unique solvability of the discretized state problems and Lipschitz continuity of the corresponding control-to-state mapping. The discrete shape optimization problem can be transformed into a nonsmooth minimization problem and handled by the bundle trust method. In each step of the method, the state problem is solved by the method of successive approximations and necessary subgradient information is computed using the generalized differential calculus of B. Mordukhovich.

Keywords Frictional contact • Nonsmooth analysis • Shape optimization

P. Beremlijski (✉)
Centre of Excellence, IT4 Innovations and Department of Applied Mathematics, VŠB-Technical University of Ostrava, 17. listopadu 15/2172, CZ-708 33 Ostrava-Poruba, Czech Republic
e-mail: petr.beremlijski@vsb.cz

J. Haslinger
Department of Numerical Mathematics, Charles University in Prague, Sokolovská 83, CZ-186 75 Praha 8, Czech Republic

Centre of Excellence, IT4 Innovations, VŠB-Technical University of Ostrava, 17. listopadu 15/2172, CZ-708 33 Ostrava-Poruba, Czech Republic
e-mail: hasling@karlin.mff.cuni.cz

J. Outrata
Institute of Theory of Information and Automation of the AS CR, Pod Vodárenskou věží 4, CZ-182 08 Praha 8, Czech Republic
e-mail: outrata@utia.cas.cz

Grad. School of Information Technology and Math. Sciences, University of Ballarat, Australia

R. Pathó
Department of Numerical Mathematics, Charles University in Prague, Sokolovská 83, CZ-186 75 Praha 8, Czech Republic
e-mail: patho@karlin.mff.cuni.cz

R. Hoppe (ed.), *Optimization with PDE Constraints*, Lecture Notes in Computational Science and Engineering 101,
DOI 10.1007/978-3-319-08025-3_1

Mathematics Subject Classification (2010). Primary 49M25; Secondary 35J86, 74P10

1 Introduction

Shape optimization is a branch of optimal control theory in which control variables are related to the geometry of optimized structures (size, shape or topology). By an appropriate change of the geometry one tries to get a structure with some desired properties. Usually, its behavior is modeled by partial differential equations. In practice, however, one can meet situations when physical systems are governed by variational inequalities. A common feature of such optimization problems is the fact that the control-to-state mapping might be nonsmooth and, consequently, the whole optimization problem is generally nonsmooth, as well. If it is so, then special tools of nonsmooth analysis have to be used to perform sensitivity analysis which provides necessary gradient-like information for nonsmooth minimization methods. Contact problems represent one of typical applications of variational inequalities in mechanics of solids: one tries to find an equilibrium state of a system of a finite number of loaded deformable bodies which are possibly in mutual contact taking into account effects of friction on common parts. Just the presence of friction complicates the analysis. If the friction obeys the Coulomb law [5], then the respective mathematical model leads to an implicit variational inequality. Shape optimization with contact problems involving Coulomb friction in 2 and 3D has been theoretically studied in [1], and [2], respectively, including numerical experiments. Another type of friction was considered in [8], namely contact problems with given friction and a solution-dependent coefficient of friction. Shape optimization with this type of the state problems has been theoretically analyzed in [7]. The goal of the present chapter is to illustrate applicability of theoretical results concerning sensitivity analysis for numerical realization of model examples.

The paper is organized as follows: after introducing the notation, we recall some basic notions from the theory of generalized differential calculus that will be used later in Sect. 3. In Sect. 2 we present the state problem, the shape optimization problem and also quote some results concerning their solvability. Next, we introduce a suitable discretization and review conditions under which discrete optimal shapes exist and converge to an optimal one as the discretization parameter tends to zero. Assuming unique solvability of the discrete state problems in Sect. 3, we compute shape sensitivities of the cost functional and the discrete state variable employing modern methods of variational analysis [12]. Using these results, numerical examples in Sect. 4 illustrate the feasibility of this approach in solving shape optimization problems involving complicated boundary conditions.

Throughout the paper we use the following notation: the symbol $H^k(\Omega)$ ($k \geq 0$ integer) stands for the Sobolev space of functions which are together with their derivatives up to order k square integrable in Ω, i.e. elements of $L^2(\Omega)$ (we set

$H^0(\Omega) := L^2(\Omega)$). The norm in $H^k(\Omega)$ will be denoted by $\|\cdot\|_{k,\Omega}$. Vector-valued functions and the respective spaces of vector-valued functions will be denoted by bold characters. Bold characters will also be used for vectors in $\mathbb{R}^n$, with the Euclidean scalar product $\langle\cdot,\cdot\rangle_n$ and corresponding norm $\|\cdot\|_n$. For a set $A \subset X$, $\overline{A}$ stands for the *closure* of A with respect to the topology of X. For $X = \mathbb{R}^n$ and $\bar{\mathbf{x}} \in A$ we denote by $\hat{N}_A(\vec{x})$ the *Fréchet (regular) normal cone* to A at $\bar{\mathbf{x}}$:

$$\hat{N}_A(\bar{\mathbf{x}}) := \left\{\mathbf{x}^* \in \mathbb{R}^n \;\middle|\; \limsup_{\mathbf{x} \xrightarrow{A} \bar{\mathbf{x}}} \frac{\langle \mathbf{x}^*, \mathbf{x} - \bar{\mathbf{x}}\rangle_n}{\|\mathbf{x} - \bar{\mathbf{x}}\|_n} \leq 0\right\},$$

whereas the *limiting (Mordukhovich) normal cone* to A at $\bar{\mathbf{x}}$ will be denoted by $N_A(\bar{\mathbf{x}})$:

$$N_A(\bar{\mathbf{x}}) := \operatorname{limsup}_{\mathbf{x} \xrightarrow{A} \bar{\mathbf{x}}} \hat{N}_A(\mathbf{x}).$$

Here the symbol "Limsup" stands for the Kuratowski-Painlevé outer limit of sets (cf. [16]). Given a multifunction $Q : \mathbb{R}^n \to \mathbb{R}^m$, we denote its graph by $\operatorname{Gr} Q := \{(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^m \mid \mathbf{y} \in Q(\mathbf{x})\}$. The *regular coderivative* of Q at a reference point $(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \in \operatorname{Gr} Q$ is given by the multifunction $\hat{D}^*Q(\bar{\mathbf{x}}, \bar{\mathbf{y}}) : \mathbb{R}^m \to \mathbb{R}^n$, which is defined as follows:

$$\hat{D}^*Q(\bar{\mathbf{x}}, \bar{\mathbf{y}})(\mathbf{y}^*) := \{\mathbf{x}^* \in \mathbb{R}^n \mid (\mathbf{x}^*, -\mathbf{y}^*) \in \hat{N}_{\operatorname{Gr} Q}(\bar{\mathbf{x}}, \bar{\mathbf{y}})\}.$$

Analogously, the multifunction $D^*Q(\bar{\mathbf{x}}, \bar{\mathbf{y}}) : \mathbb{R}^m \to \mathbb{R}^n$, defined by

$$D^*Q(\bar{\mathbf{x}}, \bar{\mathbf{y}})(\mathbf{y}^*) := \{\mathbf{x}^* \in \mathbb{R}^n \mid (\mathbf{x}^*, -\mathbf{y}^*) \in N_{\operatorname{Gr} Q}(\bar{\mathbf{x}}, \bar{\mathbf{y}})\}$$

is called the *limiting (Mordukhovich) coderivative* of Q at $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$. Further, we will employ another important notion from the theory of generalized differentiation, namely that of calmness: a multifunction Q is said to be *calm* at $(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \in \operatorname{Gr} Q$ provided $\exists L > 0$ and $\exists$ neighbourhoods U, V of $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$, respectively, such that:

$$Q(\mathbf{x}) \cap V \subset Q(\bar{\mathbf{x}}) + L\|\mathbf{x} - \bar{\mathbf{x}}\|_n \mathbb{B}_m \quad \forall \mathbf{x} \in U,$$

where $\mathbb{B}_m$ stands for the closed unit ball in $\mathbb{R}^m$, centered at the origin.

2 Problem Formulation and Discretization

Throughout the chapter we assume that the positive real parameters a, b and $0 < C_0 < b$ are fixed.

Let us consider an *elastic body*, represented by the domain $\Omega := \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 \in (0, a),\ \alpha(x_1) < x_2 < b\}$, where $\alpha \in C^{0,1}([0, a])$, $0 \leq \alpha \leq C_0$. Suppose

that the boundary $\partial\Omega$ is decomposed according to the boundary conditions into three pairwise disjoint, relatively open subsets: along $\Gamma_u \subset \partial\Omega$, $\mathrm{meas}_1\,\Gamma_u > 0$ the body is clamped, on $\Gamma_P \subset \partial\Omega$ surface tractions of density $\mathbf{P} = (P_1, P_2) \in \mathbf{L}^2(\Gamma_P)$ act and along $\Gamma_c := \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 \in (0, a),\ x_2 = \alpha(x_1)\} = \mathrm{Gr}\,\alpha$, the body may come in contact with the rigid foundation $M = \{(x_1, x_2) \in \mathbb{R}^2 \mid x_2 \le 0\}$. Due to the special geometry, the *non-penetration conditions* on the contact boundary Γ_c can be expressed exactly and take the following form:

$$\left.\begin{aligned} &u_2(x_1, \alpha(x_1)) \ge -\alpha(x_1), \quad T_2(\mathbf{u})(x_1, \alpha(x_1)) \ge 0, \\ &\big(u_2(x_1, \alpha(x_1)) + \alpha(x_1)\big) T_2(\mathbf{u})(x_1, \alpha(x_1)) = 0 \end{aligned}\right\} \quad \text{for } x_1 \in (0, a). \tag{2.1}$$

Here $\mathbf{u} = (u_1, u_2) : \Omega \to \mathbb{R}^2$ is a displacement vector and $\mathbf{T}(\mathbf{u}) = (T_1(\mathbf{u}), T_2(\mathbf{u})) : \partial\Omega \to \mathbb{R}^2$ is the stress vector associated with $\mathbf{u}$. In addition to (2.1) we shall consider effects of friction between Ω and M. We use the friction law of Tresca type, i.e. with an a-priori given slip bound $g : \Gamma_c \to \mathbb{R}_+$, but with a coefficient of friction $\mathscr{F} : R_+ \to R_+$ which depends on the solution. Thus the *friction conditions* on Γ_c read as follows:

$$\left.\begin{aligned} &u_1 = 0 \Longrightarrow |T_1(\mathbf{u})| \le \mathscr{F}(0) g \\ &u_1 \ne 0 \Longrightarrow T_1(\mathbf{u}) = -\mathrm{sgn}(u_1)\mathscr{F}(|u_1|) g \end{aligned}\right\} \quad \text{on } \Gamma_c. \tag{2.2}$$

Finally, Ω will be subject to body forces of density $\mathbf{F} = (F_1, F_2) \in \mathbf{L}^2(\Omega)$. The equilibrium state of Ω is characterized by a displacement vector $\mathbf{u}$ which satisfies the system of the linear equilibrium equations in Ω, the classical boundary conditions on Γ_P, Γ_u and the unilateral and friction conditions (2.1) and (2.2), resp., on Γ_c.

In order to give the weak form of the Signorini problem with given friction and a solution-dependent coefficient of friction, we denote the space of virtual displacements by $\mathbf{V} := \{\mathbf{v} = (v_1, v_2) \in \mathbf{H}^1(\Omega) \mid \mathbf{v} = \mathbf{0}$ a.e. on $\Gamma_u\}$ and the closed, convex cone of kinematically admissible displacements by $\mathbf{K} := \{\mathbf{v} \in \mathbf{V} \mid v_2(x_1, \alpha(x_1)) \ge -\alpha(x_1)$ a.e. in $(0, a)\}$. Further, let $a : \mathbf{V} \times \mathbf{V} \to \mathbb{R}$ and $L : \mathbf{V} \to \mathbb{R}$ be defined by:

$$a(\mathbf{u}, \mathbf{v}) := \int_\Omega \sigma(\mathbf{u}) : \varepsilon(\mathbf{v})\, d\mathbf{x}, \qquad L(\mathbf{v}) := \int_\Omega \mathbf{F} \cdot \mathbf{v}\, d\mathbf{x} + \int_{\Gamma_P} \mathbf{P} \cdot \mathbf{v}\, ds,$$

where the stress tensor $\sigma(\mathbf{u})$ is linked to the linearized strain tensor $\varepsilon(\mathbf{u}) := \frac{1}{2}(\nabla\mathbf{u} + (\nabla\mathbf{u})^T)$ by a linear Hooke's law: $\sigma(\mathbf{u}) = \mathscr{C}\varepsilon(\mathbf{u})$. We assume that the fourth order stiffness tensor $\mathscr{C} \in L^\infty(\Omega)$ satisfies the usual symmetry and ellipticity conditions.

Definition 2.1. By a *weak solution* to the Signorini problem with Tresca friction and a solution-dependent coefficient of friction $\mathscr{F}$ we mean any $\mathbf{u} \in \mathbf{K}$ satisfying:

$$a(\mathbf{u}, \mathbf{v} - \mathbf{u}) + \int_{\Gamma_c} \mathscr{F}(|u_1|) g(|v_1| - |u_1|)\, ds \ge L(\mathbf{v} - \mathbf{u}) \quad \forall \mathbf{v} \in \mathbf{K}. \tag{$\mathscr{P}$}$$

Note that ($\mathscr{P}$) is an implicit variational inequality of the second kind. Its solvability was addressed in [8] and is summarized in the following theorem.

Theorem 2.2. *(i) For any nonnegative $g \in L^2(\Gamma_c)$ and nonnegative, bounded, continuous $\mathscr{F}$ there exists a solution to ($\mathscr{P}$).*

(ii) There exists a constant $C_{max} > 0$ such that the solution to ($\mathscr{P}$) is unique, provided that $g \in L^\infty(\Gamma_c)$ and $\mathscr{F}$ is bounded, Lipschitz continuous with modulus $C_L > 0$ such that: $C_L \|g\|_{L^\infty(\Gamma_c)} < C_{max}$.

Up to this point we used one fixed domain Ω and solved the corresponding problem ($\mathscr{P}$) on it. When optimizing the contact boundary we consider α to be a parameter, by means of which one can change the shape of Ω. Our aim is to find α^* from an *admissible set* U_{ad} such that the pair $(\alpha^*, \mathbf{u}^*)$, where $\mathbf{u}^*$ solves the corresponding problem ($\mathscr{P}$) on $\Omega(\alpha^*)$, minimizes a given cost functional J on U_{ad}. To emphasize the fact that Ω is parametrized by α, we will write α as the argument. In agreement with this convention, notation $\Omega(\alpha)$, $\Gamma_c(\alpha)$, $\mathbf{V}(\alpha)$, $\mathbf{K}(\alpha)$, $\mathscr{P}(\alpha)$, etc. will be used instead of Ω, Γ_c, $\mathbf{V}$, $\mathbf{K}$, ($\mathscr{P}$), etc.

In what follows we shall restrict ourselves to α belonging to the following admissible set U_{ad}:

$$U_{ad} := \{\alpha \in C^{1,1}([0,a]) \mid 0 \le \alpha \le C_0,\ |\alpha'| \le C_1 \text{ in } [0,a],\\ |\alpha''| \le C_2 \text{ a.e. in } (0,a), \text{meas}\,\Omega(\alpha) = C_3\}, \tag{2.3}$$

i.e. U_{ad} contains all functions which are together with their first derivatives Lipschitz equi-continuous in $[0,a]$ and preserve the constant area of $\Omega(\alpha)$. We assume that the positive constants C_0, C_1, C_2 and C_3 are chosen in such a way that $U_{ad} \neq \emptyset$. Further, we need to clarify the meaning of all functions appearing in the definition of ($\mathscr{P}$) for various $\alpha \in U_{ad}$. To this end, let $\hat{\Omega} := (0,a) \times (0,b)$ and assume that the functions $\mathscr{C}$, $\mathbf{F}$, $\mathbf{P}$ and g are restrictions of some $\hat{\mathscr{C}} \in L^\infty(\hat{\Omega})$, $\hat{\mathbf{F}} \in \mathbf{L}^2(\hat{\Omega})$, $\hat{\mathbf{P}} \in \mathbf{L}^2(\partial\hat{\Omega})$ and $\hat{g} \in H^1(\hat{\Omega})$, $\hat{g} \ge 0$ onto $\Omega(\alpha)$, $\Gamma_p(\alpha)$ and $\Gamma_c(\alpha)$, respectively.

Let $S : U_{ad} \ni \alpha \mapsto \{\mathbf{u} \in \mathbf{K}(\alpha) \mid \mathbf{u} \text{ solves } \mathscr{P}(\alpha)\}$ denote the *control-to-state* mapping and let $J : \text{Gr}\, S \to \mathbb{R}$ be a given *cost functional*. Note that S is a *multivalued* mapping, in general.

Definition 2.3. A domain $\Omega(\alpha^*)$ is said to be *optimal* iff there exists $\mathbf{u}^* \in S(\alpha^*)$ satisfying:

$$J(\alpha^*, \mathbf{u}^*) \le J(\alpha, \mathbf{u}) \quad \forall (\alpha, \mathbf{u}) \in \text{Gr}\, S. \tag{$\mathbb{P}$}$$

Below we recall the result from [7] stating, that there exists an optimal shape in U_{ad}, defined by (2.3), for a large class of cost functionals.

Theorem 2.4. *Let the assumptions of Theorem 2.2(i) hold and suppose that J is lower semicontinuous in the following sense:*

$$\left.\begin{array}{l}\alpha_n \to \alpha \ \text{ in } C^1([0,a]),\ \alpha_n, \alpha \in U_{ad},\\ \mathbf{y}_n \rightharpoonup \mathbf{y} \ \text{ in } \mathbf{H}^1(\hat{\Omega}),\ \mathbf{y}_n, \mathbf{y} \in \mathbf{H}^1(\hat{\Omega})\end{array}\right\} \Longrightarrow \liminf_{n\to\infty} J(\alpha_n, \mathbf{y}_n|_{\Omega(\alpha_n)}) \geq J(\alpha, \mathbf{y}|_{\Omega(\alpha)}).$$

Then ($\mathbb{P}$) *has a solution.*

Proof. It is sufficient to prove that Gr S is compact in the above defined topology—see [7, Lemma 1]. □

In the second part of this section we shortly describe a discrete version of ($\mathbb{P}$) and provide sufficient conditions ensuring unique solvability of the discretized state problems and convergence of discrete optimal shapes to an optimal one in the sense of Definition 2.3.

Every discretization of ($\mathbb{P}$) is twofold: (i) one has to approximate the admissible set U_{ad} and (ii) to discretize the state problem. In order to make the forthcoming presentation more straightforward, we shall use continuous, piecewise linear functions α_h as design variables. However, they are not practical from the engineering point of view and therefore will be replaced by Bézier functions in numerical experiments. For the approximation of ($\mathscr{P}$) we shall use standard piecewise linear triangular finite elements.

Let $d \geq 1$ be a given integer and set $h := a/d$. By δ_h we denote the equidistant partition of $[0, a]$:

$$\delta_h: \quad 0 \equiv a_0 < a_1 < \cdots < a_{d(h)} \equiv a, \qquad a_j = a + jh,\ j = 0, \ldots, d. \tag{2.4}$$

With any δ_h we associate the set U_{ad}^h defined by

$$\begin{aligned} U_{ad}^h := \{\alpha_h \in C([0,a]) \mid \alpha_h|_{[a_{i-1},a_i]} \in P_1([a_{i-1}, a_i]) \quad & \forall i = 1, \ldots, d,\\ 0 \leq \alpha_h(a_i) \leq C_0 \quad & \forall i = 0, \ldots, d,\\ |\alpha_h(a_i) - \alpha_h(a_{i-1})| \leq C_1 h \quad & \forall i = 1, \ldots, d,\\ |\alpha_h(a_{i+1}) - 2\alpha_h(a_i) + \alpha_h(a_{i-1})| \leq C_2 h^2, \quad & \forall i = 1, \ldots, d-1,\\ \text{meas}\, \Omega(\alpha_h) = C_3\}, & \end{aligned}$$

where $C_0, \ldots, C_3$ are the same as in (2.3). Notice that $U_{ad}^h \not\subset U_{ad}$, i.e. U_{ad}^h is an *external approximation* of U_{ad}.

Since for each $\alpha_h \in U_{ad}^h$ the domain $\Omega(\alpha_h)$ is polygonal, one can construct its triangulation $\mathscr{T}(h, \alpha_h)$ whose nodes lie on the lines $\{a_i\} \times \mathbb{R}$, $i = 0, 1, \ldots, d$.

Moreover, we shall assume that for each $h > 0$ the family $\{\mathscr{T}(h,\alpha_h) \mid \alpha_h \in U_{ad}^h\}$ consists of *topologically equivalent* triangulations (cf. [6, p. 32]) and that $\{\mathscr{T}(h,\alpha_h)\}$ are *uniformly regular* with respect to $(h,\alpha_h) \in (0,\infty) \times U_{ad}^h$. The domain $\Omega(\alpha_h)$ with the triangulation $\mathscr{T}(h,\alpha_h)$ will be denoted by $\Omega_h(\alpha_h)$ or just shortly Ω_h.

On $\Omega_h(\alpha_h)$ we construct the following piecewise linear approximations of $\mathbf{V}(\alpha_h)$ and $\mathbf{K}(\alpha_h)$:

$$\mathbf{V}_h(\alpha_h) := \{\mathbf{v}_h \in \mathbf{C}(\overline{\Omega}_h) \mid \mathbf{v}_h|_T \in (P_1(T))^2 \ \forall T \in \mathscr{T}(h,\alpha_h),\ \mathbf{v}_h = \mathbf{0} \text{ on } \overline{\Gamma}_u(\alpha_h)\},$$

and

$$\mathbf{K}_h(\alpha_h) := \{\mathbf{v}_h = (v_{h1}, v_{h2}) \in \mathbf{V}_h(\alpha_h) \mid v_{h2}(a_i, \alpha_h(a_i)) \geq -\alpha_h(a_i)\ \forall a_i \in \mathscr{N}_h\},$$

respectively, where $\mathscr{N}_h$ is the set of all *contact nodes*, i.e. $a_i \in \mathscr{N}_h$ iff $(a_i, \alpha_h(a_i)) \in \overline{\Gamma}_c(\alpha_h) \setminus \overline{\Gamma}_u(\alpha_h)$. Observe, that $\mathbf{K}_h(\alpha_h) \subset \mathbf{K}(\alpha_h)\ \ \forall h > 0\ \forall \alpha_h \in U_{ad}^h$.

Definition 2.5. By a solution to the discretized Signorini problem with given friction and a solution-dependent coefficient of friction we mean any function $\mathbf{u}_h := \mathbf{u}_h(\alpha_h) \in \mathbf{K}_h(\alpha_h)$ satisfying:

$$\left.\begin{aligned} &a_{\alpha_h}(\mathbf{u}_h, \mathbf{v}_h - \mathbf{u}_h) + \int_0^a \mathscr{F}(r_h|u_{h1} \circ \alpha_h|) g \circ \alpha_h (|v_{h1} \circ \alpha_h| - \\ &|u_{h1} \circ \alpha_h|) \sqrt{1 + (\alpha_h')^2}\, dx_1 \geq L_{\alpha_h}(\mathbf{v}_h - \mathbf{u}_h) \quad \forall \mathbf{v}_h \in \mathbf{K}_h(\alpha_h), \end{aligned}\right\} \qquad (\mathscr{P}_h(\alpha_h))$$

where $r_h : C([0,a]) \to C([0,a])$ stands for the piecewise linear Lagrange interpolation operator on δ_h and for any $w \in H^1(\Omega(\alpha_h))$ the symbol $w \circ \alpha_h$ denotes the function $x \mapsto w(x, \alpha_h(x))$, $x \in (0,a)$.

Theorem 2.6. *(i) Let the assumptions of Theorem 2.2(i) be satisfied. Then $(\mathscr{P}_h(\alpha_h))$ has a solution for any $h > 0$ and $\alpha_h \in U_{ad}^h$.*

(ii) There exists a constant $C_{max}^h > 0$ such that the solution to $(\mathscr{P}_h(\alpha_h))$ is unique, provided that the following conditions hold: $g \in C(\overline{\hat{\Omega}})$ and $\mathscr{F}$ is nonnegative, bounded, Lipschitz continuous with modulus $C_L > 0$ such that $C_L \|g\|_{C(\overline{\hat{\Omega}})} < C_{max}^h$.

Proof. It can be found in [8] and [15]. □

Note, that by looking at the explicit form of the constants $C_{\max}$ and $C_{\max}^h$ appearing in Theorems 2.2 and 2.6 we find that: (1) $C_{\max}^h$ can be chosen *independently* of h and (2) $C_{\max}^h < C_{\max}$. Hence, Theorem 2.6(*ii*) implies Theorem 2.2(*ii*).

Analogously to S and ($\mathbb{P}$) we define the discrete control-to-state mapping S_h and the discrete shape optimization problem ($\mathbb{P}_h$):

$$\left.\begin{array}{l}\text{minimize } J(\alpha_h, \mathbf{u}_h)\\ \text{subj. to } \quad (\alpha_h, \mathbf{u}_h) \in \operatorname{Gr} S_h\end{array}\right\} \tag{$\mathbb{P}_h$}$$

In the next theorem the symbol "˜" above a function $\mathbf{v} \in \mathbf{H}^1(\Omega(\alpha_h))$ denotes its extension to $\hat{\Omega}$ satisfying: $\|\tilde{\mathbf{v}}\|_{1,\hat{\Omega}} \leq \tilde{C}\,\|\mathbf{v}\|_{1,\Omega(\alpha_h)}\ \forall \mathbf{v} \in \mathbf{H}^1(\Omega(\alpha_h))$, such that the constant $\tilde{C}$ is independent of α_h and $h > 0$. Since $\{\Omega(\alpha_h) \mid \alpha_h \in U_{ad}^h,\ h > 0\}$ is a system satisfying the *uniform cone property*, such an extension exists (see [3]). For details we refer to [7].

Theorem 2.7. *Suppose that J is continuous in the following sense:*

$$\left.\begin{array}{l}\alpha_h \to \alpha \ \text{ in } C([0,a]),\ \alpha_h \in U_{ad}^h\\ \mathbf{u}_h \rightharpoonup \mathbf{u} \ \text{ in } \mathbf{H}^1(\hat{\Omega}),\ \mathbf{u}_h, \mathbf{u} \in \mathbf{H}^1(\hat{\Omega})\end{array}\right\} \Longrightarrow \lim_{h\to 0_+} I(\alpha_h, \mathbf{u}_h|_{\Omega(\alpha_h)}) = I(\alpha, \mathbf{u}|_{\Omega(\alpha)})$$

and let the assumptions of Theorem 2.6(ii) hold. Then:

(j) there exists at least one solution $(\alpha_h^*, \mathbf{u}_h^*)$ to ($\mathbb{P}_h$) $\forall h > 0$;

(jj) for every sequence of discrete optimal pairs $\{(\alpha_h^*, \mathbf{u}_h^*)\}$, $h \to 0_+$ there exists a subsequence $\{(\alpha_{h_j}^*, \mathbf{u}_{h_j}^*)\}$, $j \to \infty$ and functions $\alpha^* \in U_{ad}$, $\mathbf{u}^* \in \mathbf{H}^1(\hat{\Omega})$ such that:

$$\alpha_{h_j}^* \to \alpha^* \ \text{ in } \ C([0,a]) \quad \text{and} \quad \tilde{\mathbf{u}}_{h_j}^* \rightharpoonup \mathbf{u}^* \ \text{ in } \ \mathbf{H}^1(\hat{\Omega}),\ j \to \infty, \tag{2.5}$$

where $(\alpha^*, \mathbf{u}^*|_{\Omega(\alpha^*)})$ solves ($\mathbb{P}$). In addition, every accumulation point of $\{(\alpha_h^*, \mathbf{u}_h^*)\}$ in the sense of (2.5) has this property.

Proof. It follows from Theorems 6, 7 and Lemma 7 in [7]. □

We conclude this section with the algebraic form of the discrete state problem ($\mathscr{P}_h(\alpha_h)$), in particular with its reduced version involving only state variables defined on the contact boundary Γ_c. In the rest of the paper $h > 0$ shall be *fixed*.

Let us set $n := \dim \mathbf{V}_h(\alpha_h)$ and $p := \operatorname{card} \mathscr{N}_h$, i.e. p is the number of the contact nodes. For the sake of simplicity let us further assume that $p = d(h) + 1$ (cf. (2.4)). Then U_{ad}^h is isomorphic to a convex, compact set $\mathscr{U}_{ad} \subset \mathbb{R}_+^p$ by means of the mapping $\alpha_h \mapsto \boldsymbol{\alpha} = (\alpha_h(a_0), \ldots, \alpha_h(a_{d(h)}))$. Further, the set $\mathbf{K}_h(\alpha_h)$ may be identified with the closed, convex set:

$$\mathscr{K}(\boldsymbol{\alpha}) := \{\mathbf{v} \in \mathbb{R}^n \mid \mathbf{v}_v \geq -\boldsymbol{\alpha}\}, \quad \boldsymbol{\alpha} \in \mathscr{U}_{ad},$$

where $\mathbf{v}_\nu \in \mathbb{R}^p$ stands for the subvector of $\mathbf{v} \in \mathbb{R}^n$ consisting of the second components of the displacement vector $\mathbf{v}$ at all contact nodes, i.e. $(\mathbf{v}_\nu)_i = v_{h2}(a_{i-1}, \alpha_h(a_{i-1}))$ for each $i = 1, \ldots, p$. Analogously, $\mathbf{v}_\tau \in \mathbb{R}^p$ consists of the first components of $\mathbf{v}$ at the contact nodes. The frictional term in $(\mathscr{P}_h(\alpha_h))$ should be *approximated* by a quadrature formula whose integration nodes coincide with the contact nodes. Hence, the algebraic formulation of the discrete Signorini problem with a solution-dependent coefficient of friction reads as:

$$\left.\begin{array}{l}\text{Find } \mathbf{u} \in \mathscr{K}(\boldsymbol{\alpha}) \text{ such that for every } \mathbf{v} \in \mathscr{K}(\alpha): \\ \langle \mathbb{A}(\boldsymbol{\alpha})\mathbf{u}, \mathbf{v}-\mathbf{u}\rangle_n + \sum_{i=1}^{p} \omega_i(\boldsymbol{\alpha}) \mathscr{F}(|(\mathbf{u}_\tau)_i|)\big(|(\mathbf{v}_\tau)_i| - |(\mathbf{u}_\tau)_i|\big) \geq \langle \mathbf{L}(\boldsymbol{\alpha}), \mathbf{v}-\mathbf{u}\rangle_n\end{array}\right\} \quad (\mathscr{P}'(\boldsymbol{\alpha}))$$

where $\mathbb{A} \in C^1(\mathscr{U}_{ad}; \mathbb{R}^{n\times n})$ and $\mathbf{L} \in C^1(\mathscr{U}_{ad}; \mathbb{R}^n)$ denote the matrix- and vector-valued function, resp., associating with any $\boldsymbol{\alpha} \in \mathscr{U}_{ad}$ the stiffness matrix $\mathbb{A}(\boldsymbol{\alpha})$ and the load vector $\mathbf{L}(\boldsymbol{\alpha})$. Finally, let us assume that $\omega_i \in C^1(\mathscr{U}_{ad}; (0, \infty))$ $\forall i = 1, \ldots, p$.

Instead of dealing with $(\mathscr{P}'(\boldsymbol{\alpha}))$ directly, we shall introduce Lagrange multipliers $\lambda \in \mathbb{R}^p_+$ to release the constraint $\mathbf{v} \in \mathscr{K}(\boldsymbol{\alpha})$, and employ the Schur complement technique to eliminate all internal variables and reduce the state problem to the contact boundary. Since it will be more convenient for sensitivity analysis, the resulting variational inequality is formulated as a *generalized equation* (GE) (for details the reader is kindly referred to [7] and also [1, 2]):

$$\left.\begin{array}{l}\mathbf{0} \in \mathbb{A}_{\tau\tau}(\boldsymbol{\alpha})\mathbf{u}_\tau + \mathbb{A}_{\tau\nu}(\boldsymbol{\alpha})\mathbf{u}_\nu - \mathbf{L}_\tau(\boldsymbol{\alpha}) + Q_1(\boldsymbol{\alpha}, \mathbf{u}_\tau) \\ \mathbf{0} = \mathbb{A}_{\nu\tau}(\boldsymbol{\alpha})\mathbf{u}_\tau + \mathbb{A}_{\nu\nu}(\boldsymbol{\alpha})\mathbf{u}_\nu - \mathbf{L}_\nu(\boldsymbol{\alpha}) - \boldsymbol{\lambda} \\ \mathbf{0} \in \mathbf{u}_\nu + \boldsymbol{\alpha} + N_{\mathbb{R}^p_+}(\boldsymbol{\lambda}).\end{array}\right\} \tag{2.6}$$

In our case the multifunction $Q_1 : \mathscr{U}_{ad} \times \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ is defined as:

$$\big(Q_1(\boldsymbol{\alpha}, \mathbf{u}_\tau)\big)_i := \omega_i(\boldsymbol{\alpha}) \mathscr{F}(|(\mathbf{u}_\tau)_i|) \partial |(\mathbf{u}_\tau)_i| \quad \forall i = 1, \ldots, p,$$

where "∂" denotes the subdifferential of convex functions, $N_{\mathbb{R}^p_+}(\cdot)$ is the normal cone in the sense of convex analysis and submatrices $\mathbb{A}_{\tau\tau}$, $\mathbb{A}_{\tau\nu}$, $\mathbb{A}_{\nu\nu} \in C^1(\mathscr{U}_{ad}; \mathbb{R}^{p\times p})$ are parts of the Schur complement to the stiffness matrix with $\mathbb{A}_{\nu\tau} = \mathbb{A}^T_{\tau\nu}$. In addition, note that $\mathbb{A}_{\tau\tau}$ and $\mathbb{A}_{\nu\nu}$ are positive definite uniformly with respect to $\boldsymbol{\alpha} \in \mathscr{U}_{ad}$.

The next theorem states that GE (2.6) is uniquely solvable and its solution depends Lipschitz continuously on the shape variable α.

Theorem 2.8. *There exists a constant $C_L > 0$, independent of h and $\alpha \in \mathscr{U}_{ad}$ such that if $\mathscr{F}$ is Lipschitz continuous with modulus C_L, then the corresponding control-to-state mapping $S : \mathscr{U}_{ad} \to \mathbb{R}^{3p}$, $\alpha \mapsto \{(\mathbf{u}_\tau, \mathbf{u}_\nu, \lambda) \mid (\mathbf{u}_\tau, \mathbf{u}_\nu, \lambda)$ solves (2.6)$\}$ is single-valued and Lipschitz continuous.*

3 Discrete Sensitivity Analysis

Introducing the state variable $\mathbf{y} := (\mathbf{u}_\tau, \mathbf{u}_\nu, \boldsymbol{\lambda}) \in \mathbb{R}^{3p}$, the GE (2.6) may be written in the following compact form:

$$\mathbf{0} \in F(\boldsymbol{\alpha}, \mathbf{y}) + Q(\boldsymbol{\alpha}, \mathbf{y}), \tag{3.1}$$

with $\boldsymbol{\alpha} \in \mathscr{U}_{ad}$ being the control variable and

$$F(\boldsymbol{\alpha}, \mathbf{y}) := \begin{pmatrix} \mathbb{A}_{\tau\tau}(\boldsymbol{\alpha}) & \mathbb{A}_{\tau\nu}(\boldsymbol{\alpha}) & 0 \\ \mathbb{A}_{\nu\tau}(\boldsymbol{\alpha}) & \mathbb{A}_{\nu\nu}(\boldsymbol{\alpha}) & -\mathbb{I} \\ 0 & \mathbb{I} & 0 \end{pmatrix} \mathbf{y} - \begin{pmatrix} \mathbf{L}_\tau(\boldsymbol{\alpha}) \\ \mathbf{L}_\nu(\boldsymbol{\alpha}) \\ -\boldsymbol{\alpha} \end{pmatrix}, \quad Q(\boldsymbol{\alpha}, \mathbf{y}) := \begin{pmatrix} Q_1(\boldsymbol{\alpha}, \mathbf{y}_1) \\ 0 \\ N_{\mathbb{R}^p_+}(\mathbf{y}_3) \end{pmatrix}.$$

Note that F is single-valued, continuously differentiable in its domain of definition and Q is a closed-graph multifunction. The algebraic *shape optimization problem* reads as the following Mathematical Program with Equilibrium Constraints (MPEC):

$$\left.\begin{array}{ll} \text{minimize} & J(\boldsymbol{\alpha}, \mathbf{y}) \\ \text{subj. to} & \mathbf{0} \in F(\boldsymbol{\alpha}, \mathbf{y}) + Q(\boldsymbol{\alpha}, \mathbf{y}), \\ & \boldsymbol{\alpha} \in \mathscr{U}_{ad}, \end{array}\right\} \tag{$\mathbb{P}$}$$

where J is a given continuously differentiable cost functional. In what follows we shall assume that the assumptions of Theorem 2.8 are satisfied. Then ($\mathbb{P}$) may be equivalently reformulated as a nonlinear optimization problem:

$$\left.\begin{array}{ll} \text{minimize} & \mathscr{J}(\boldsymbol{\alpha}) := J(\boldsymbol{\alpha}, S(\boldsymbol{\alpha})) \\ \text{subj. to} & \boldsymbol{\alpha} \in \mathscr{U}_{ad}. \end{array}\right\} \tag{$\tilde{\mathbb{P}}$}$$

Since $\mathscr{J}$ is locally Lipschitz continuous, ($\tilde{\mathbb{P}}$) can be solved by standard methods of nonsmooth optimization. Such algorithms, however, require typically knowledge of some subgradient information, usually in the form of one (arbitrary) subgradient from the Clarke subdifferential $\overline{\partial}\mathscr{J}$ (cf. [4, Theorem 2.5.1]) in each iteration step. This can be conducted by using the chain rule from [4, Theorem 2.6.6]:

$$\overline{\partial}\mathscr{J}(\overline{\boldsymbol{\alpha}}) = \nabla_\alpha J(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{y}}) + \left(\overline{\partial} S(\overline{\boldsymbol{\alpha}})\right)^T \nabla_y J(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{y}}), \tag{3.2}$$

valid at any reference point $\overline{\boldsymbol{\alpha}} \in \mathscr{U}_{ad}$, $\bar{\mathbf{y}} := S(\overline{\boldsymbol{\alpha}})$. Thus, for the required subgradient information it is sufficient to compute an element from $\left(\overline{\partial} S(\overline{\boldsymbol{\alpha}})\right)^T \nabla_y J(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{y}})$, where $\overline{\partial} S(\overline{\boldsymbol{\alpha}})$ stands for the *generalized Jacobian* of Clarke, defined in [4, Definition 2.6.1]. The rest of the section is devoted to this task.

First, observe that Lipschitz continuity of S and formula (2.23) in [11] yield:

$$\big(\overline{\partial} S(\overline{\boldsymbol{\alpha}})\big)^T \mathbf{y}^* = \operatorname{conv} D^* S(\overline{\boldsymbol{\alpha}}(\mathbf{y}^*)) \quad \forall \mathbf{y}^* \in \mathbb{R}^{3p}.$$

Comparing with (3.2), we see that it is sufficient to determine one element from the set $D^* S(\overline{\boldsymbol{\alpha}})(\nabla_y J(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{y}}))$ and we are done. The latter task will be accomplished using the following theorem.

Theorem 3.1. *Let $(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{y}}) \in Gr\ S$ be fixed and introduce the mapping: $\Phi : \mathbb{R}^p \times \mathbb{R}^{3p} \to \mathbb{R}^p \times \mathbb{R}^{3p} \times \mathbb{R}^{3p}$, $(\boldsymbol{\alpha}, \mathbf{y}) \mapsto \big(\boldsymbol{\alpha}, \mathbf{y}, -F(\boldsymbol{\alpha}, \mathbf{y})\big)^T$. Then the following hold:*

(i) The multifunction $M : \mathbb{R}^p \times \mathbb{R}^{3p} \times \mathbb{R}^{3p} \to \mathbb{R}^p \times \mathbb{R}^{3p}$, $\mathbf{p} \mapsto \{\, (\boldsymbol{\alpha}, \mathbf{y}) \mid \mathbf{p} + \Phi(\boldsymbol{\alpha}, \mathbf{y}) \in Gr\ Q \,\}$ is calm at $(\mathbf{0}, \mathbf{0}, \mathbf{0}, \overline{\boldsymbol{\alpha}}, \bar{\mathbf{y}})^T$.

(ii) For every $\mathbf{p}^* \in D^* S(\overline{\boldsymbol{\alpha}})(\nabla_y J(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{y}}))$ there exists a vector $\mathbf{v}^* \in \mathbb{R}^{3p}$ such that $(\mathbf{p}^*, \mathbf{v}^*)$ is a solution of the (limiting) adjoint GE:

$$\begin{pmatrix} \mathbf{p}^* \\ -\nabla_y J(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{y}}) \end{pmatrix} \in \nabla F(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{y}})^T \mathbf{v}^* + D^* Q(\Phi(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{y}}))(\mathbf{v}^*). \tag{AGE}$$

Proof. Part (i) was proved in [7, Lemma 8], whereas part (ii) follows from (i) and [9, Theorem 4.1]. For details see [7]. □

Note that due to Lipschitz continuity of S, (AGE) attains at least one solution $\mathbf{p}^*$ and at points $(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{y}})$, where Q is normally regular, i.e. $\hat{N}_{\mathrm{Gr}\ Q}(\Phi(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{y}})) = N_{\mathrm{Gr}\ Q}(\Phi(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{y}}))$, every solution $\mathbf{p}^*$ of (AGE) belongs to $D^* S(\overline{\boldsymbol{\alpha}})(\nabla_y J(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{y}}))$. In the nonregular case, however, the set of solutions of (AGE) is in general larger than $D^* S(\overline{\boldsymbol{\alpha}})(\nabla_y J(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{y}}))$ and so this procedure may lead to a subgradient, which lies outside of $(\overline{\partial} S(\overline{\boldsymbol{\alpha}}))^T \nabla_y J(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{y}})$. Nevertheless, numerical experience shows that this phenomenon occurs very rarely and typically does not negatively influence the behavior of bundle methods, which we use for the solution of ($\tilde{\mathbb{P}}$) (cf. [17]).

In the rest of this section we will devote our attention to the solution of (AGE), in particular to expression of the coderivative $D^* Q$ in terms of the problem data. To begin with, note that the components of Q are *decoupled* (this is a consequence of the assumed model of given friction), hence its coderivative can be computed componentwise:

$$D^* Q(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{y}}, \bar{\mathbf{q}})(\mathbf{q}^*) = \begin{pmatrix} D^* Q_1(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{y}}_1, \bar{\mathbf{q}}_1)(\mathbf{q}_1^*) \\ 0 \\ D^* N_{\mathbb{R}^p_+}(\bar{\mathbf{y}}_3, \bar{\mathbf{q}}_3)(\mathbf{q}_3^*) \end{pmatrix} \quad \forall \mathbf{q}^* \in \mathbb{R}^{3p}, \tag{3.3}$$

at any given point $(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{y}}, \bar{\mathbf{q}}) \in \mathrm{Gr}\, Q$. The third component is standard and the exact formula for it may be found e.g. in [13, Lemma 2.2]. In order to deal with the first component, let us write the multifunction $Q_1 : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^p$ as a composition of an outer multifunction Z_1 and an inner single-valued, smooth mapping Ψ:

$$Q_1(\boldsymbol{\alpha}, \mathbf{u}) = \begin{pmatrix} \omega_1(\boldsymbol{\alpha}) \mathscr{F}(|u_1|) \partial |u_1| \\ \omega_2(\boldsymbol{\alpha}) \mathscr{F}(|u_2|) \partial |u_2| \\ \vdots \\ \omega_p(\boldsymbol{\alpha}) \mathscr{F}(|u_p|) \partial |u_p| \end{pmatrix} = (Z_1 \circ \Psi)(\boldsymbol{\alpha}, \mathbf{u}), \tag{3.4}$$

where

$$\Psi = (\Psi_1, \ldots, \Psi_p) : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^{3p}, \qquad \Psi_j(\boldsymbol{\alpha}, \mathbf{u}) := (\omega_j(\boldsymbol{\alpha}), u_j)^T,$$

and

$$Z_1 : \mathbb{R}^{3p} \to \mathbb{R}^p, \quad \mathbf{y} \mapsto (Z(\mathbf{y}_1), \ldots, Z(\mathbf{y}_p))^T,$$

with

$$Z : (0, \infty) \times \mathbb{R} \to \mathbb{R}, \quad (x_1, x_2) \mapsto x_1 \mathscr{F}(|x_2|) \partial |x_2|.$$

Now the chain rule from [16, Theorem 10.40] allows us to compute the coderivative of the composite multifunction (3.4) as follows:

Theorem 3.2. *Let $(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{u}}, \bar{\mathbf{q}}) \in Gr\, Q_1$ be such that the following condition holds:*

$$Ker\, \nabla \Psi(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{u}})^T \cap D^* Z_1(\Psi(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{u}}), \bar{\mathbf{q}})(\mathbf{0}) = \{\mathbf{0}\}. \tag{3.5}$$

Then:

$$\begin{aligned} \forall \mathbf{q}^* \in \mathbb{R}^p : \quad D^* Q_1(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{u}}, \bar{\mathbf{q}})(\mathbf{q}^*) &\subset \nabla \Psi(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{u}})^T D^* Z_1(\Psi(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{u}}), \bar{\mathbf{q}})(\mathbf{q}^*) \\ &= \nabla \Psi(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{u}})^T \begin{pmatrix} D^* Z(\Psi_1(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{u}}), \bar{q}_1)(\bar{q}_1^*) \\ D^* Z(\Psi_2(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{u}}), \bar{q}_2)(\bar{q}_2^*) \\ \vdots \\ D^* Z(\Psi_p(\overline{\boldsymbol{\alpha}}, \bar{\mathbf{u}}), \bar{q}_p)(\bar{q}_p^*) \end{pmatrix}. \end{aligned} \tag{3.6}$$

By means of (3.3) and (3.6) we have reduced the computation of $D^* Q$ to that of $D^* Z$. Due to the particularly simple structure of Z, this can be done relatively easily and has been investigated in detail in [7, Section 6.2]. We summarize these results in the next theorem.

Theorem 3.3. *Let $(\bar{x}_1, \bar{x}_2, \bar{z}) \in Gr\ Z$ be a given point and $z^* \in \mathbb{R}$ arbitrary. Then exactly one from the following cases holds true:*

(1) $\bar{x}_2 > 0$, *then:* $D^*Z(\bar{x}_1, \bar{x}_2, \bar{z})(z^*) = \{z^*\mathscr{F}(\bar{x}_2)\} \times D^*\mathscr{F}(\bar{x}_2)(\bar{x}_1 z^*)$;
(2) $\bar{x}_2 < 0$, *then:* $D^*Z(\bar{x}_1, \bar{x}_2, \bar{z})(z^*) = \{-z^*\mathscr{F}(-\bar{x}_2)\} \times \big(-D^*\mathscr{F}(-\bar{x}_2)(-\bar{x}_1 z^*)\big)$;
(3) $\bar{x}_2 = 0$, $|\bar{z}| < \bar{x}_1\mathscr{F}(0)$, *then:*

$$D^*Z(\bar{x}_1, 0, \bar{z})(z^*) = \begin{cases} \{0\} \times \mathbb{R}\ , & \text{if } z^* = 0, \\ \emptyset\ , & \text{otherwise}; \end{cases}$$

(4) $\bar{x}_2 = 0$, $\bar{z} = \bar{x}_1\mathscr{F}(0)$, *then:*

$$D^*Z(\bar{x}_1, 0, \bar{x}_1\mathscr{F}(0))(z^*) \begin{cases} \subset \{z^*\mathscr{F}(0)\} \times D^*\mathscr{F}(0)(\bar{x}_1 z^*), & \text{if } z^* > 0, \\ = \{z^*\mathscr{F}(0)\} \times (-\infty,\ \bar{x}_1 z^* D^+\mathscr{F}(0)], & \text{if } z^* < 0, \\ = \{0\} \times \mathbb{R}, & \text{if } z^* = 0, \end{cases}$$

where the symbol $D^+\mathscr{F}(0) := \limsup_{\eta \to 0_+} \frac{\mathscr{F}(\eta) - \mathscr{F}(0)}{\eta}$ *stands for the upper Dini derivative of* $\mathscr{F}$ *at* 0;
(5) $\bar{x}_2 = 0$, $\bar{z} = -\bar{x}_1\mathscr{F}(0)$, *then:*

$$D^*Z(\bar{x}_1, 0, -\bar{x}_1\mathscr{F}(0))(z^*) \begin{cases} = \{-z^*\mathscr{F}(0)\} \times [\bar{x}_1 z^* D^+\mathscr{F}(0),\ +\infty), & \text{if } z^* > 0, \\ \subset \{-z^*\mathscr{F}(0)\} \times \big(-D^*\mathscr{F}(0)(-\bar{x}_1 z^*)\big), & \text{if } z^* < 0, \\ = \{0\} \times \mathbb{R}, & \text{if } z^* = 0. \end{cases}$$

Using this result one may construct and solve the (AGE). Moreover, one has:

Corollary 3.4. *The condition* (3.5) *holds at each* $(\bar{\boldsymbol{\alpha}}, \bar{\mathbf{u}}, \bar{\mathbf{q}}) \in Gr\ Q_1$.

Proof. See [7, Corollary 2]. □

4 Numerical Results

The theoretical results of the previous sections will now be used for computation of model examples. We assume that the friction coefficient $\mathscr{F}$ is defined by

$$\mathscr{F}(t) = 0.25 \cdot \frac{1}{t^2 + 1} \quad \forall t \in \mathbb{R}_+, \tag{4.1}$$

and the a-priori given slip bound is $g = 150$. Further we assume that the cost functional J is continuously differentiable so that the composite map $\mathscr{J}$ is locally Lipschitzian. Therefore one can use the implicit programming approach [14] to solve the shape optimization problem ($\mathbb{P}$). For the minimization of $\mathscr{J}$ we used the Matlab implementation of the bundle trust method BT (see [17]). This method is very robust and was designed just for the minimization of nonsmooth functions. It requires in every step the value of the objective function and its arbitrary Clarke subgradient (for more details see [4]), i.e., for each admissible $\boldsymbol{\alpha}$ we have to be able to find a solution of the state problem $(\mathbf{u}, \lambda) = S(\boldsymbol{\alpha})$ and to compute one arbitrary Clarke subgradient of $\mathscr{J}$ at $\boldsymbol{\alpha}$. This issue was discussed in the preceding section.

Since the Signorini problem with given friction and a solution-dependent coefficient of friction can be equivalently formulated as a fixed-point problem (see [8]), the method of successive approximations will be used for its numerical solution. Each iterative step is represented by the Signorini problem with given friction and the given coefficient friction computed from the previous iteration.

These techniques were implemented and the following experiments were solved by MatSol library [10] developed in the Matlab environment.

For the solution of model examples, we slightly modify the set U_{ad}^h. The purpose of this modification is to decrease the number of control variables and, at the same time, to get a smooth shape of the contact boundary. Therefore, the boundary Γ_c will be modelled by Bézier functions of order d. The system of points $\{A_i\}_{i=0}^d$, where $A_i = (ih, \alpha_i), \alpha_i \in \mathbb{R}, i = 0, 1, \ldots, d, h = a/d$ defines the so-called control points of the Bézier function F_α of order d on $[0, a]$:

$$F_\alpha(x) = \sum_{i=0}^{d} \alpha^i \beta_d^i(x), \quad \beta_d^i(x) = \frac{1}{a^d} \binom{d}{i} x^i (a - x)^{d-i}, \ \ x \in [0, a].$$

Discretized shapes are determined by the vector $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_d)$, where α_i is the second component of A_i, $i = 0, \ldots, d$. The end points of F_α coincide with the first and last control point. The graph of F_α itself lies in the convex envelope of the control points. This means that any upper and lower bounds imposed on the components of $\boldsymbol{\alpha}$ are automatically satisfied for F_α, too.

The new shape optimization problem using this type of the design variables is defined as follows:

$$\left.\begin{array}{ll} \text{minimize} & J(\boldsymbol{\alpha}, S(\boldsymbol{\alpha})) \\ \text{subj. to} & \boldsymbol{\alpha} \in \mathscr{U}, \end{array}\right\} \qquad (\mathbb{P}_B)$$

where

$$\mathscr{U} = \{\boldsymbol{\alpha} \in \mathbb{R}^{d+1} \mid 0 \le \alpha^i \le C_0,\ i = 0, 1, \ldots, d;$$
$$|\alpha^{i+1} - \alpha^i| \le C_1 h,\ i = 0, 1, \ldots, d-1;$$
$$|\alpha^{i+1} - 2\alpha^i + \alpha^{i-1}| \le C_2 h^2,\ i = 1, 2, \ldots, d-1;$$
$$C_{31} \le \text{meas}\ \Omega(\boldsymbol{\alpha}) \le C_{32}\}$$

and $C_0, C_1, C_2, C_{31}, C_{32}$ are given positive constants. The first $d+1$ box constraints guarantee that $|F_\alpha(x)| \le C_0\ \forall x \in [0, a]$. The second and the third set of the constraints take care of the smoothness of the optimal shape. It is well known that if the control points satisfy these two conditions, then $|F'_\alpha(x)| \le C_1$ and $|F''_\alpha(x)| \le C_2\ \forall x \in [0, a]$. The last constraint is added to control the volume of the domain. Unlike the constant volume constraint considered in the theoretical part of this paper, this time we use the inequality constraints for the volume of $\Omega(\boldsymbol{\alpha})$. The last constraint has a physical meaning of preserving the weight of the structure in prescribed limits.

We will present results of two examples solved by the mentioned implicit programming technique combined with the BT code. In both examples we use the same data and change only the cost function J. The shape of the elastic body $\Omega(\boldsymbol{\alpha}), \boldsymbol{\alpha} \in \mathscr{U}$, is defined through a Bézier function F_α as follows (cf. Fig. 1):

$$\Omega(\boldsymbol{\alpha}) = \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 \in (0, a),\ F_\alpha(x_1) < x_2 < b\}.$$

From Fig. 1 one also sees the distribution of external pressures on the boundary Γ_P, given as $\mathbf{P^1} = (0; -60\ \text{MPa})$ on $(0, 1.8) \times \{1\}$ and zero on $(1.8, 2) \times \{1\}$, while $\mathbf{P^2} = (50\ \text{MPa}; 30\ \text{MPa})$ on $\{2\} \times (0, 1)$. Further, Γ_u is the part of the boundary where the zero displacements are prescribed.

The set of the admissible designs $\mathscr{U}$ is specified as follows: $a = 2, b = 1$ and $C_0 = 0.75$, $C_1 = 0.85$, $C_2 = 10$, $C_{31} = 1.88$, $C_{32} = 1.95$. In both examples

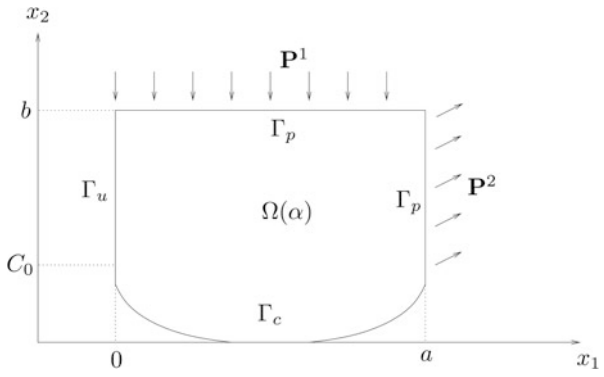


Fig. 1 The elastic body and applied loads

the Young modulus $E = 1$ GPa and the Poisson constant $\sigma = 0.3$ are used. The state problems on $\Omega(\boldsymbol{\alpha})$ are discretized by isoparametric quadrilateral elements of Lagrange type. The total number of nodes (vertices of quadrilaterals) is 1,800 for any $\boldsymbol{\alpha} \in \mathscr{U}$. The dimension of the control vector $\boldsymbol{\alpha}$, generating the Bézier function and defining $\Omega(\boldsymbol{\alpha})$, is 20.

Example 1. In the first example we try to smooth down peaks of the normal contact stress distribution. To this aim, one should minimize the max norm of the discrete normal contact stress $\boldsymbol{\lambda}$. The objective function $\mathscr{J}$, however, must be continuously differentiable, so we will use (p power of) p-norm of vectors with $p = 4$. The shape optimization problem then reads as follows:

$$\begin{aligned} &\text{minimize} \quad \|\lambda\|_4^4 \\ &\text{subj. to} \quad \boldsymbol{\alpha} \in \mathscr{U}. \end{aligned}$$

In Fig. 2 we depict the initial shape and the distribution of the von Mises stress in the loaded body. Figure 3 shows the optimal shape and the von Mises stress in

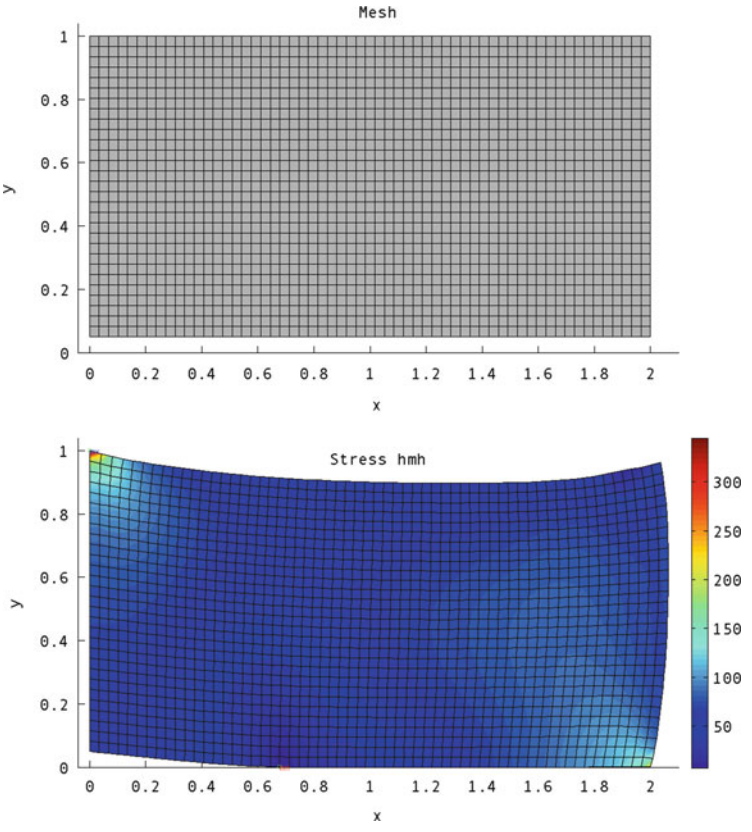


Fig. 2 Example 1, initial design

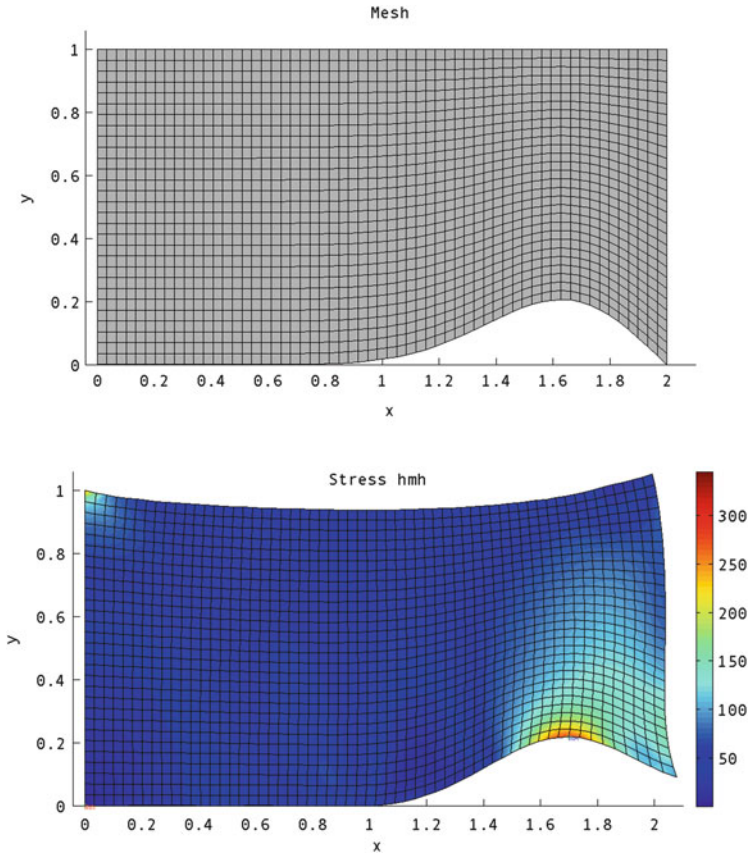


Fig. 3 Example 1, optimal design

the deformed optimal body. Finally, Fig. 4 compares the contact normal stresses for the initial $\Omega(\boldsymbol{\alpha}_0)$ (left) and optimal $\Omega(\boldsymbol{\alpha}_{\text{opt}})$ (right) shape, respectively. The obtained optimal value of the cost functional $\mathscr{J}(\boldsymbol{\alpha}_{\text{opt}}) = 1.9623 \cdot 10^8$ compared to $\mathscr{J}(\boldsymbol{\alpha}_0) = 6.0151 \cdot 10^8$ represents a decrease by 67 %. The decrease of the peak stress is also quite significant.

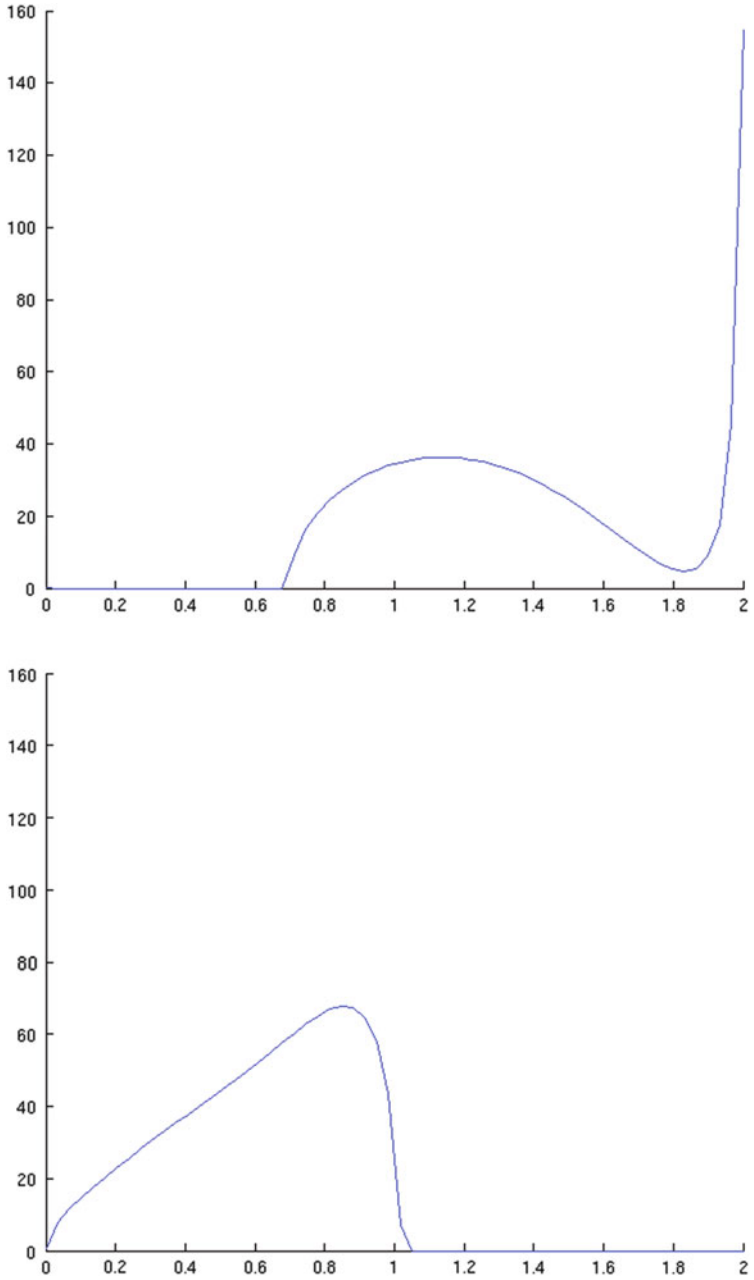


Fig. 4 Example 1, normal stress for initial (*left*) and optimal (*right*) design

Example 2. Here we try to identify the contact normal stress $\boldsymbol{\lambda}$ with a prescribed value $\bar{\boldsymbol{\lambda}}$. The shape optimization problem can be written as

$$\begin{aligned} &\text{minimize} \quad \|\bar{\lambda} - \lambda\|_2^2 \\ &\text{subj. to} \quad \boldsymbol{\alpha} \in \mathscr{U}. \end{aligned}$$

This vector was chosen to model a function, depicted in Fig. 7 by the dotted line.

The initial design and its deformation with the distribution of the von Mises stress is presented in Fig. 5, while Fig. 6 shows the optimal design before and

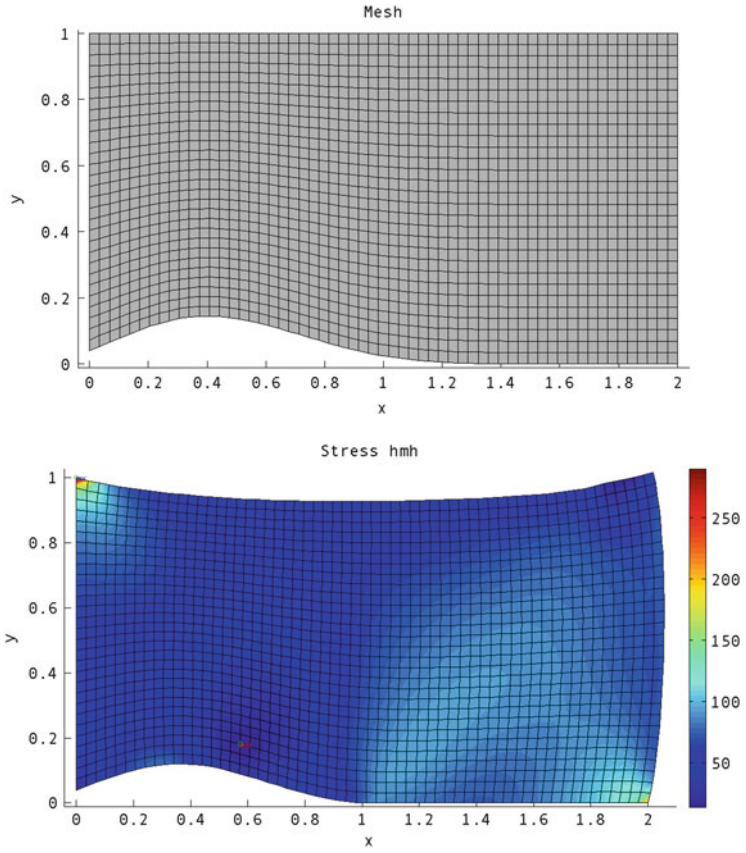


Fig. 5 Example 2, initial design

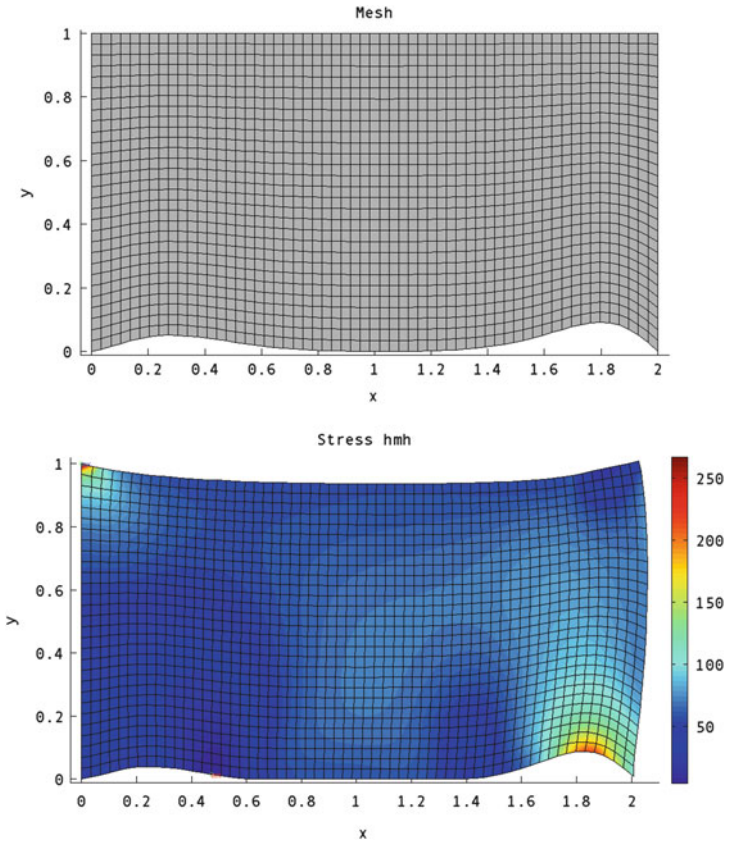


Fig. 6 Example 2, optimal design

after deformation. Finally, Fig. 7 compares the contact normal stresses with the prescribed values. While the initial contact stresses are far from the prescribed values, the stresses for the optimal shape follow very closely $\overline{\boldsymbol{\lambda}}$. Note that during the optimization process the initial value $\mathscr{J}(\boldsymbol{\alpha}_0) = 5.910 \cdot 10^4$ of the cost functional dropped by two orders of magnitude to $\mathscr{J}(\boldsymbol{\alpha}_{\text{opt}}) = 9.1457 \cdot 10^2$.

In order to emphasize the importance of proper modelling of contact problems, let us compute the same example, now with a coefficient of friction which *does not depend* on the solution. In particular, we set $\mathscr{F}(t) = 0.25$ for every $t \geq 0$, but keep all other parameters of Example 2 unchanged. Starting from the same

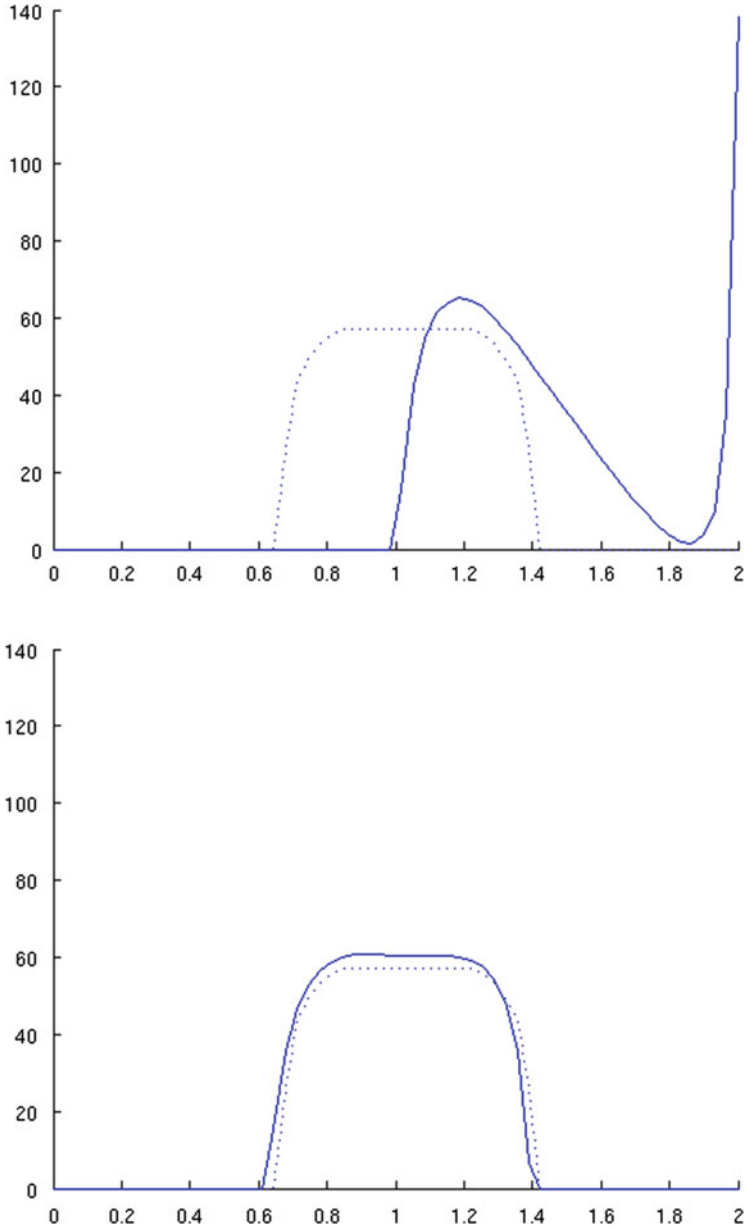


Fig. 7 Example 2, normal stress for initial (*left*) and optimal (*right*) design

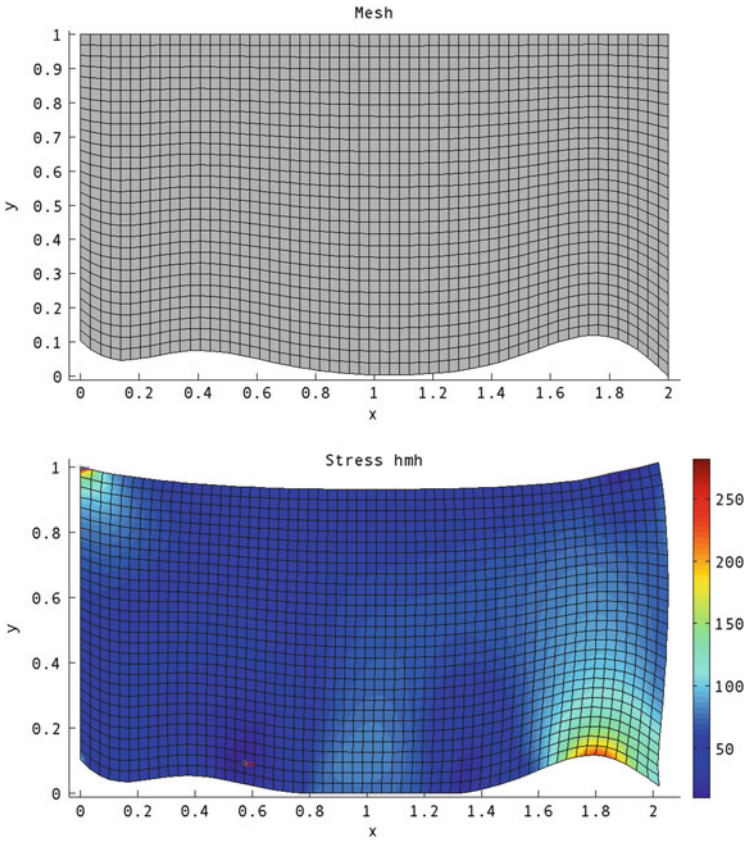


Fig. 8 Example 2 with $\mathscr{F} = \text{const}$; optimal design $\Omega(\overline{\boldsymbol{\alpha}}_{\text{opt}})$

initial configuration $\Omega(\boldsymbol{\alpha}_0)$ as in Example 2, the algorithm converges to a solution $\Omega(\overline{\boldsymbol{\alpha}}_{\text{opt}})$—see Fig. 8.

Then we solve the original contact problem with the coefficient of friction given by (4.1) on $\Omega(\overline{\boldsymbol{\alpha}}_{\text{opt}})$ and show the distribution of the normal stress along $\Gamma_c(\overline{\boldsymbol{\alpha}}_{\text{opt}})$ in Fig. 9. Comparing Fig. 9 with Fig. 7, one can see that the "approximate optimal design" $\Omega(\overline{\boldsymbol{\alpha}}_{\text{opt}})$, obtained by replacing the original state problem with a simpler one, is not optimal at all.

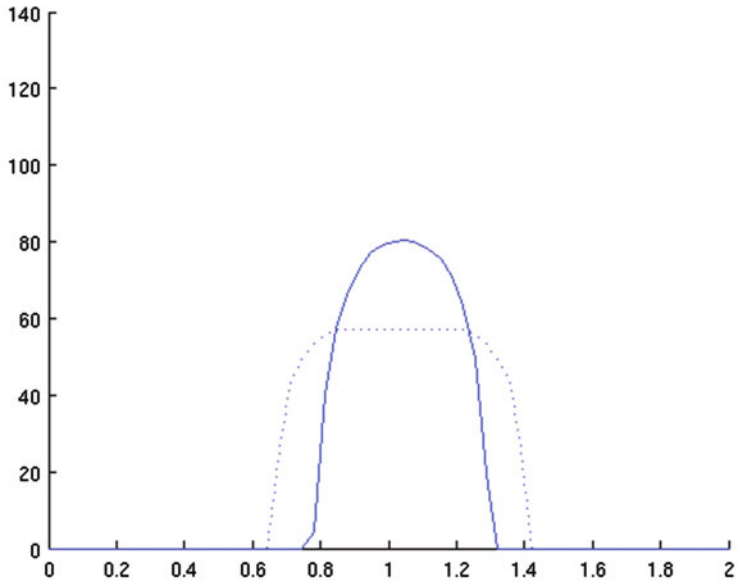


Fig. 9 Example 2; normal stress distribution on $\Omega(\overline{\boldsymbol{\alpha}}_{\text{opt}})$

Acknowledgements The work was supported by the ESF OPTPDE Research Programme. The first and the second author acknowledge also the support of the European Regional Development Fund in the IT4 Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070) and support of the project SPOMECH—Creating a multidisciplinary R&D team for reliable solution of mechanical problems, reg. no. CZ.1.07/2.3.00/20.0070 within Operational Programme 'Education for competitiveness' funded by Structural Funds of the European Union and state budget of the Czech Republic. The second, the third and the fourth author acknowledge the support of the Grant GAČR P201/12/0671. In addition, the third author expresses his gratitude to the ARC project DP110102011. The fourth author acknowledges also the support of GAUK no. 719912.

References

[1] P. Beremlijski, J. Haslinger, M. Kočvara, J.V. Outrata, Shape optimization in contact problems with Coulomb friction. SIAM J. Opt. **13**, 561–587 (2002)
[2] P. Beremlijski, J. Haslinger, M. Kočvara, R. Kučera, J.V. Outrata, Shape optimization in three-dimensional contact problems with Coulomb friction. SIAM J. Opt. **20**, 416–444 (2009)
[3] D. Chenais, On the existence of a solution in a domain identification problem. J. Math. Anal. App. **52**, 189–219 (1975)
[4] F. Clarke, *Optimization and Nonsmooth Analysis* (Wiley, New York, 1983)
[5] J. Haslinger, I. Hlaváček, J. Nečas, *Numerical Methods for Unilateral Problems in Solid Mechanics.* In: ed. by P.G. Ciarlet, J.-L. Lions, Handbook of Numerical Analysis, Vol. IV, Part 2, pp. 313–491 (North Holland, Amsterdam, 1996)
[6] J. Haslinger, R.A.E. Mäkinen, *Introduction to Shape Optimization: Theory, Approximation and Computation. Advances in Design and Control* (SIAM, Philadelphia, 2003)

[7] J. Haslinger, J.V. Outrata, R. Pathó, Shape optimization in 2D contact problems with given friction and a solution-dependent coefficient of friction. Set Valued Var. Anal. **20**, 31–59 (2012)
[8] J. Haslinger, O. Vlach, Signorini problem with a solution dependent coefficient of friction (model with given friction): approximation and numerical realization. Appl. Math. **50**, 153–171 (2005)
[9] R. Henrion, A. Jourani, J. Outrata, On the calmness of a class of multifunctions. SIAM J. Opt. **13**, 603–618 (2002)
[10] T. Kozubek, A. Markopoulos, T. Brzobohatý, R. Kučera, V. Vondrák, Z. Dostál, MatSol - MATLAB efficient solvers for problems in engineering. http://industry.it4i.cz/en/products/matsol
[11] B.S. Mordukhovich, Generalized differential calculus for nonsmooth and set-valued mappings. J. Math. Anal. Appl. **183**, 250–288 (1994)
[12] B.S. Mordukhovich, *Variational Analysis and Generalized Differentiation, I: Basic Theory, II: Applications.* Grundlehren Series (Fundamental Principles of Mathematical Sciences), Vols. 330 and 331 (Springer, Berlin-Heidelberg-New York, 2006)
[13] J.V. Outrata, Optimality conditions for a class of mathematical programs with equilibrium constraints. Math. Oper. Res. **24**, 627–644 (1999)
[14] J.V. Outrata, M. Kočvara, J. Zowe, *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints: Theory, Applications and Numerical Results* (Kluwer Academic, Dordrecht-Boston-London, 1998)
[15] R. Pathó, Shape optimization in contact problems with friction. Diploma Thesis, Charles University in Prague, 2009
[16] R.T. Rockafellar, R. Wets, *Variational Analysis* (Springer, Berlin-Heidelberg-New York, 1998)
[17] H. Schramm, J. Zowe, A version of the bundle idea for minimizing a nonsmooth function: conceptual idea, convergence analysis, numerical results. SIAM J. Opt. **2**, 121–152 (1992)

Phase Field Methods for Binary Recovery

Charles Brett, Charles M. Elliott, and Andreas S. Dedner

Abstract We consider the inverse problem of recovering a binary function from blurred and noisy data. Such problems arise in many applications, for example image processing and optimal control of PDEs. Our formulation is based on the Mumford-Shah model, but with a phase field approximation to the perimeter regularisation. We use a double obstacle potential as well as a smooth double well potential. We introduce an iterative method for solving the problem, develop a suitable discretisation of this iterative method, and prove some convergence results. Numerical simulations are presented which illustrate the usefulness of the approach and the relative merits of the phase field models.

Keywords Binary recovery • Image processing • Mumford-Shah model • Optimal control • Phase field models

Mathematics Subject Classification (2010). Primary 49N45; Secondary 65K10, 68U10

1 Introduction

A fundamental problem in the field of image processing is the following. Suppose we have a function $\bar{u}$ defined on a bounded and piecewise smooth domain $\Omega \subset \mathbb{R}^N$ for $N \leq 3$, which has been transformed by a linear operator S, and then corrupted by additive noise ζ, such that we have data

$$y_d := S\bar{u} + \zeta.$$

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) Grant EP/H023364/1 and by the ESF within the Programme OPTPDE.

C. Brett (✉) • C.M. Elliott • A.S. Dedner
Mathematics Institute, University of Warwick, Coventry, CV4 7AL, UK
e-mail: c.brett@warwick.ac.uk; c.m.elliott@warwick.ac.uk; a.s.dedner@warwick.ac.uk

R. Hoppe (ed.), *Optimization with PDE Constraints*, Lecture Notes in Computational Science and Engineering 101,
DOI 10.1007/978-3-319-08025-3_2

The problem is to recover $\bar{u}$ given y_d. Two immediate issues are that (a) ζ is unknown, so we will not be able to find $\bar{u}$ even with a good model for the space in which it lies (b) inverting S may be ill-posed, so it will be difficult to find an approximation to $\bar{u}$ even if $\zeta = 0$.

We investigate this problem in the case that $\bar{u}$ is a binary function. We develop the theory with S an abstract operator, but in examples we take S to be the solution operator of an elliptic PDE. In this case the problem becomes one in PDE constrained optimal control.

Our approach to modelling the problem is to minimise an energy functional consisting of an L^2 fidelity term plus a phase field approximation to minimal perimeter regularisation. This can be thought of as a relaxation of the Mumford-Shah segmentation model. In our phase field approximation we use the Ginzburg-Landau functional with both the smooth double well and double obstacle potentials.

1.1 Motivating Examples

First we give examples from both image processing and optimal control of PDEs motivating the study of this problem:

- **Image segmentation**—We can represent a barcode by a 1D function which takes the value -1 when the barcode is white and 1 when it is black. When a barcode is scanned by a barcode reader this function becomes blurred (due to scattering in the air) and noisy (due to measurement error and imperfections in the barcode). So the machine only sees a corrupted signal, but from this it needs to determine the scanned barcode.
- **Elliptic source recovery**—Suppose we have noisy data of a quantity y, which is related to another quantity $\bar{u}$ by some physical law. For example, let $\bar{u}$ represent a heat source, then the long term temperature distribution y may be related to $\bar{u}$ by the solution of an elliptic PDE. Our goal could be to find the heat source that produces a particular temperature distribution.

1.2 Background Material

For the above problems to be tractable we naturally require some knowledge of the form of the operator S and the noise ζ. We also usually assume a specific form of $\bar{u}$, as this influences the best model to use. For example, in the barcode problem we could assume that the function we are trying to recover is a binary function taking the values -1 and 1, and that the bars have a minimum width. Some sets of assumptions on S, ζ and $\bar{u}$ that are made in the literature are the following:

1. *Denoising and deblurring*—S is a blurring operator (maybe the identity), ζ is Gaussian noise, and $\bar{u}$ is a piecewise smooth function [11, 13, 31].

2. *Segmentation*—S is a blurring operator (maybe the identity), ζ is Gaussian noise, and $\bar{u}$ is binary function [16, 21, 29]. These are the assumptions we make in this work.
3. *Binary image restoration*—S is the identity, we have 'salt and pepper' noise, and $\bar{u}$ is a binary function [14]. This kind of noise gives each point of a binary function a probability of switching to the other value, so the data y_d is also binary.

Note that the above sets of assumptions have been named using terminology from image processing. Although our problem can be thought of as either an image processing or PDE constrained optimal control problem depending on the choice of S, we found most of the relevant literature to be from the image processing community. This is unsurprising since image processing is one of the main applications of binary recovery. We end up taking S to be the solution operator of an elliptic PDE, but try to use neutral language which reflects that our problem arises in these two fields.

For segmentation, which we focus on in this work, a large proportion of the literature modifies one of the following two models when formulating the problem of Sect. 1 mathematically. We now introduce these models so the reader can see how our approach fits with the existing literature.

- **Model 1 (Mumford-Shah).** This model, which was introduced in [29], looks for piecewise smooth functions that minimise an energy functional.

  Let Ω_i be disjoint open subsets with piecewise smooth boundaries such that the closure of $\bigcup \Omega_i$ is Ω. Let u be a function that is differentiable on $\bigcup \Omega_i$, but which is allowed to be discontinuous across $\Gamma := \bigcup \partial\Omega_i \setminus \partial\Omega$. Then the Mumford-Shah model involves minimising

$$E_1(u, \Gamma) = \frac{1}{2} \int_\Omega (u - y_d)^2 + \mu \int_{\Omega \setminus \Gamma} |\nabla u|^2 + \sigma \, |\Gamma| \,, \tag{1.1}$$

  where $|\Gamma|$ denotes the $N-1$ dimensional Hausdorff measure of Γ. The $|\Gamma|$ term encourages minimising the length of the interface over which u is discontinuous.

  If we restrict to minimising over binary functions that take the unknown value a_i on Ω_i $(i = 0, 1)$, then this energy functional becomes

$$E_2(\{a_i\}, \Gamma) = \frac{1}{2} \sum_i \int_{\Omega_i} (a_i - y_d)^2 + \sigma \, |\Gamma| \,.$$

  For fixed Γ note that E_2 is minimised with respect to $\{a_i\}$ by setting

$$a_i = \frac{1}{|\Omega_i|} \int_{\Omega_i} y_d \,.$$

  So the problem reduces to just finding Γ, the locations of the discontinuities.

Due to the spaces of functions we are minimising over, both of the above variants of the Mumford-Shah model are nonconvex problems. In our work will use a relaxation of (1.1) based on a phase field approximation.

- **Model 2 (ROF).** The ROF (Rudin-Osher-Fatemi) model of [31] involves solving the following constrained minimisation problem over a suitable space of functions:

$$\begin{aligned} &\text{Minimise } |u|_{TV} \\ &\text{with } \int_\Omega u = \int_\Omega y_d \text{ and } \int_\Omega (u - y_d)^2 = s^2. \end{aligned} \tag{1.2}$$

The term $|u|_{TV}$ represents the total variation of u, and it can be defined even if u is not continuous; the total variation of a function $u \in L^1(\Omega)$ is

$$|u|_{TV} := \sup\{-\int_\Omega u \text{div}(\phi)\, \mathrm{d}x : \phi \in C_c^1(\Omega, \mathbb{R}^N), \|\phi\|_{L^\infty(\Omega)} \leq 1\}.$$

Sometimes the notation $\int_\Omega |\nabla u|$ is used instead of $|u|_{TV}$ to highlight that the total variation of u is equal to this quantity when it is well defined. The first constraint in (1.2) says that the noise has zero mean and the second that it has standard deviation s.

$BV(\Omega, \mathbb{R})$ is the subspace of functions in $L^1(\Omega)$ which have finite total variation. Minimising this model over $u \in BV(\Omega, \mathbb{R})$ can be related to the following problem for some value of σ:

$$\text{Minimise} \quad \frac{1}{2}\|u - y_d\|^2_{L^2(\Omega)} + \sigma\, |u|_{TV} \text{ over } BV(\Omega, \mathbb{R}). \tag{1.3}$$

Note that (1.3) can be thought of as a relaxation of (1.1) with $\mu = 0$; we minimise over a larger space of functions in order to get a convex problem.

If we restrict to minimising over binary functions then (1.3) becomes similar to the Mumford-Shah model. Suppose u only takes the known values $a_0 < a_1$ (i.e. $u \in BV(\Omega, \{a_0, a_1\})$), then

$$|u|_{TV} = (a_1 - a_0)\text{Per}(\{u = a_1\}) = (a_1 - a_0)\,|\Gamma|\,,$$

where the perimeter function $\text{Per}(\Sigma) := \int_\Omega |\nabla \chi_\Sigma|$ and Γ is the set over which u is discontinuous. So for binary functions, total variation regularisation is equivalent to both perimeter regularisation and the interfacial length regularisation in the Mumford-Shah model. In fact (1.1) and (1.3) become equivalent.

Suppose that in addition to $u \in BV(\Omega, \{a_0, a_1\})$ we have salt and pepper noise. Then the data is binary and both models reduce to the geometric problem

$$\min_{\Sigma_u \subset \Omega} |\Sigma_u \Delta \Sigma_d| + \sigma(a_1 - a_0)\text{Per}(\Sigma_u).$$

Here Σ_u and Σ_d denote respectively the sets where the unknown u and data y_d take the value a_1, $|\cdot|$ is now the N dimensional Hausdorff measure, and $\Sigma_u \Delta \Sigma_d$ is the symmetric difference between the sets.

1.3 Phase Field Model

We base our model on the Mumford-Shah model, but minimise over the space of functions $BV(\Omega, \{a_0, a_1\})$, and generalise it to include the blurring operator S, which we suppose is known a priori. So we have the following nonconvex model with a parameter σ, which we will shortly relax in a different way to (1.3):

$$\underset{u \in BV(\Omega, \{a_0, a_1\})}{\arg\min} \quad F(u) := \frac{1}{2} \|Su - y_d\|^2_{L^2(\Omega)} + \sigma \mathrm{Per}(\{u = a_1\}). \tag{1.4}$$

We require $S : L^2(\Omega) \to L^2(\Omega)$ to be continuous, linear, and have the mean preservation property i.e. $S1 = 1$ and hence $Sc = c$ for any constant function c. Continuity is helpful for proving existence of minimisers. Linearity and the mean preservation property allow us to recover a function $\bar{u} : \Omega \to \{a_0, a_1\}$ from data y_d by recovering a function $\bar{u} : \Omega \to \{-1, 1\}$ from a scaled and shifted copy of y_d, so long as a_0 and a_1 are known. We assume this to be the case and will therefore restrict our attention to $a_0 = -1$ and $a_1 = 1$ from now onwards.

Some examples of forms S could take are:

1. *Solution operator of elliptic PDE*—Let $Su := y$, where y solves the elliptic boundary value problem

$$\begin{aligned} -\alpha \Delta y + y &= u \quad \text{in } \Omega \\ \frac{\partial y}{\partial \nu} &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{1.5}$$

   For any $u \in L^2(\Omega)$ this equation has a unique weak solution $y \in H^1(\Omega)$ which satisfies the stability estimate

$$\|y\|_{L^2(\Omega)} = \|Su\|_{L^2(\Omega)} \leq C_s(\alpha) \|u\|_{L^2(\Omega)}, \tag{1.6}$$

   where $C_s(\alpha) := \frac{1}{1+\alpha/C_p}$ and C_p is the Poincaré constant. So S has all the required properties. We also observe that evaluating S is well-posed, but inverting S is ill-posed, which motivates the need for our model. This is the operator we use for our numerics.
2. *Convolution operator*—Let

$$Su := \phi_\alpha * u,$$

where ϕ_α is a suitable probability distribution of 'size' α, for example the Gaussian distribution

$$\phi_\alpha(x) = \frac{1}{\alpha\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\alpha^2}\right)$$

of mean zero and variance α, and $*$ is the convolution operation. Such an operator is used in the barcode problem of [16,21] and [17].

In both of these examples we have a parameter α which controls the extent of the blurring effect. Large α corresponds to heavy blurring and small α corresponds to light blurring. In our work the value of α is known a priori since we assume complete knowledge of S. However there are applications where we may want to relax this assumption, for example the barcode problem of [21]. In this application we do not know a-priori the distance of the barcode from the scanner, which means the level of blurring is unknown. This can be dealt with by fixing α to be some reasonable guess, or optimising for α at the same time as u.

We relax the model (1.4) by replacing the perimeter functional by the Ginzburg-Landau functional $G_\varepsilon : L^1(\Omega) \to [0, \infty]$ defined by

$$G_\varepsilon(u) := \begin{cases} \int_\Omega \frac{\varepsilon}{2} |\nabla u|^2 + \frac{1}{\varepsilon}\Psi(u) & u \in H^1(\Omega) \\ \infty & \text{otherwise} \end{cases}$$

for some suitable $\Psi : \mathbb{R} \to \mathbb{R}$, and then minimising over $H^1(\Omega)$ instead of $BV(\Omega, \{-1, 1\})$. So we consider

$$\underset{u \in H^1(\Omega)}{\arg\min}\, F_\varepsilon(u) := \frac{1}{2}\|Su - y_d\|^2_{L^2(\Omega)} + \frac{\sigma}{c(\Psi)}\left(\int_\Omega \frac{\varepsilon}{2}|\nabla u|^2 + \frac{1}{\varepsilon}\Psi(u)\right). \tag{1.7}$$

We will focus on two different forms for the potential Ψ; the smooth double well potential

$$\Psi_1(u) := \frac{1}{4}(1 - u^2)^2,$$

and the double obstacle potential

$$\begin{aligned} \Psi_2(u) :=& \frac{1}{2}(1 - u^2) + I_{[-1,1]}(u) \\ =& \begin{cases} \frac{1}{2}(1 - u^2) & |u| \le 1 \\ \infty & |u| > 1 \end{cases}. \end{aligned}$$

This approach, which is called a phase field approximation, results in a diffuse interface with minimisers no longer just taking the values $\{-1, 1\}$, but values in

the interval $[-1, 1]$. It is still a nonconvex problem, but it has the advantage of allowing us to minimise over a smoother space of functions for which there is better developed theory. We are able to justify this approach with the following result.

Theorem 1.1. *Let Ψ be the smooth double well potential Ψ_1. Then $G_\varepsilon(u)$ Γ-converges in $L^1(\Omega)$ as $\varepsilon \to 0$ to*

$$\begin{cases} c(\Psi_1)\mathrm{Per}(\{u = 1\}) & u \in BV(\Omega, \{-1, 1\}) \\ \infty & \textit{otherwise} \end{cases},$$

where $c(\Psi_1) = 2\int_{-1}^{1} \sqrt{2\Psi_1(s)}\mathrm{d}s = \frac{4\sqrt{2}}{3}$.

Proof. See [28]. □

A similar result holds for the double obstacle potential, and performing a calculation we get that $c(\Psi_2) = \frac{\pi}{2}$ (see [10]). To simplify notation we let $\sigma_i = \sigma/c(\Psi_i)$. This ensures that the weighting given to the regularisation is asymptotically σ for both potentials.

The different potentials lead to different formulations and we need to use different approaches to solve them. In particular, Ψ_1 leads to nonlinearity in the zeroth order terms, where as Ψ_2 causes nonlinearity by imposing constraints on the solution.

1.4 Literature Review

We now mention other parts of the literature which overlap with aspects of this work.

Barcode Problem. The 1D version of our problem is related to the barcode problem of Esedoglu in [21]. This work was later extended by Choksi and Gennip in [16]. Choksi et al. [17] uses similar ideas on QR barcodes. References for more general image processing literature can be found in Sect. 1.1.

PDE Constrained Inverse Problems. A survey of the literature from the optimal control perspective can be found in [30]. In addition, [33] describes a number of applications where we want to recover piecewise constant functions, such as magnetic resonance imaging (MRI). The thesis [26] discusses a wide range of techniques for geometric inverse problems. Tai and Li [34] recovers a piecewise constant diffusion coefficient from an elliptic PDE in 2D using the level set method.

Phase Field. In [26] there is a brief discussion of using a phase field approximation with the smooth double well potential for binary recovery. References [21] and [16] use this idea for numerical simulations, though they do not justify the approach analytically. Theory for the phase field approximation with the double obstacle potential can be found in papers by Blowey and Elliott, including [8, 9] and [10]. In [32] the double obstacle potential is used in the context of image processing, but without deblurring.

Level Set Method. This is an alternative way of recovering the discontinuities in our problem. It is discussed in [33] and [34].

Approximation of Mumford-Shah. Chambolle and Del Maso [12] and related papers prove Γ-convergence results for finite element approximations of the Mumford-Shah functional. These results have some relation to the convergence results that we obtain using a different approach.

Our work differs from existing work, and hence offers a new contribution, in the following respects:

- We introduce the phase field approximation to the model right from the start (rather than at the last minute in order to allow numerical simulations). We therefore prove rigorous analytical results for this approximate model, which puts our approach on a much firmer footing than in existing work.
- Not only the smooth double well potential, but also the double obstacle potential is used for the phase field approximation. Results are proved for both simultaneously using an abstract framework.
- We thoroughly investigate the dependency of the model on the parameters and perform a systematic comparison of the smooth double well potential and the double obstacle potential on a 1D problem. This highlights some advantages and attractive features of the latter in this setting.

1.5 Layout

In Sect. 2 we introduce an abstract optimisation problem, an iterative method for finding critical points of this problem, and prove a convergence result for the iterative method. In Sect. 3 we show that (1.7) fits into this framework with both the smooth double well and double obstacle potentials. In Sect. 4 we discuss a gradient flow formulation of (1.7) and its link to the iterative method. In Sect. 5 we discretise the iterative method and prove another convergence result. We also look at a finite element discretisation for a particular choice of S. In Sect. 6 we demonstrate that implementations of the iterative method work well in 1 and 2 dimensions. In Sect. 7 the performance of using both potentials is compared in detail for a 1D problem. In Appendix A we describe how we choose the parameters in our model for the numerics.

2 Abstract Framework

Rather than developing separate theory for solving (1.7) with the smooth double well and obstacle potentials, it is advantageous to introduce an abstract framework that both problems fit into.

To this end let V and H be real Hilbert spaces with V compactly embedded in H, and let W be a closed convex nonempty subset of V. Let $b : V \times V \to \mathbb{R}$ and $c : H \times H \to \mathbb{R}$ be symmetric continuous bilinear forms with the properties

$$\begin{aligned} \exists\, \beta \text{ s.t. } b(\eta, \eta) \geq \beta \|\eta\|_V^2 \quad & \forall \eta \in V \\ c(\eta, \eta) \geq 0 \quad & \forall \eta \in H. \end{aligned}$$

Let $l : V \to \mathbb{R}$ be a bounded linear functional and $J : V \to \mathbb{R}$ a continuous convex functional. With these objects we can define the energy functional $I : V \to \mathbb{R}$ by

$$I(\eta) := \frac{1}{2} b(\eta, \eta) + J(\eta) - \frac{1}{2} c(\eta, \eta) - l(\eta),$$

which for positive constants α_0 and C_0 we assume satisfies

$$I(\eta) \geq \alpha_0 \|\eta\|_V^2 - C_0 \quad \forall \eta \in W. \tag{2.1}$$

Remark 2.1. The functional I can be decomposed in different ways into b, J, c and l.

2.1 Optimisation Formulation

Consider the following optimisation problem: Find $u \in W$ such that

$$I(u) = \inf_{\eta \in W} I(\eta). \tag{2.2}$$

We can show existence of a solution to (2.2) with the following general result.

Proposition 2.2. *Let $A_1(\cdot) : V \to \mathbb{R}$ be weakly lower semicontinuous and let $A_2(\cdot) : H \to \mathbb{R}$ be continuous. If $A(\eta) := A_1(\eta) + A_2(\eta)$ is bounded below, then the following optimisation problem has a solution: Find $u \in W$ such that*

$$A(u) = \inf_{\eta \in W} A(\eta).$$

Proof. This follows from standard theory; we construct an infimising sequence which we know is bounded in V, so have a subsequence which weakly converges to an element of W, and this element is a minimiser of A by the properties of A_1 and A_2. □

Corollary 2.3. *(2.2) has a solution.*

Proof. Take $A_1(\eta) := \frac{1}{2}b(\eta,\eta) + J(\eta) - l(\eta)$ and $A_2(\eta) := -\frac{1}{2}c(\eta,\eta)$. Recall that continuous convex functionals are weakly lower semicontinuous, so A_1 and A_2 satisfy the requirements of Theorem 2.2. □

Note that in general there is not a unique solution to (2.2).

2.2 Variational Inequality Formulation

By standard theory, solutions to (2.2) must satisfy the following: Find $u \in W$ such that

$$b(u, \eta - u) + J(\eta) - J(u) \geq c(u, \eta - u) + l(\eta - u) \quad \forall \eta \in W. \tag{2.3}$$

Here we have used that J is a convex function, so it has a subdifferential ∂J, which by definition satisfies

$$J(\eta) - J(u) \geq \langle v, \eta - u \rangle \quad \forall v \in \partial J(u),$$

where $\langle \cdot, \cdot \rangle$ denotes the duality pairing between V^* and V. If J is in addition Gâteaux differentiable then (2.3) is equivalent to the following variational inequality: Find $u \in W$ such that

$$b(u, \eta - u) + \langle J'(u), \eta - u \rangle \geq c(u, \eta - u) + l(\eta - u) \quad \forall \eta \in W. \tag{2.4}$$

We often call solutions of (2.3) critical points of (2.2).

Remark 2.4. If $c(\eta,\eta) \leq \kappa b(\eta,\eta)$ for all $\eta \in V$ with $\kappa < 1$, then (2.3) has a unique solution. When we fit (1.7) into this framework, we find that this would require ε to be large. We intend to take ε small so that (1.7) approximates (1.4), which means we will not necessarily have uniqueness.

Note that solutions of (2.2) solve (2.3), but the converse is not necessarily true. We nevertheless aim to solve (2.3), as this is much easier in practice. Once a solution has been found, additional tests would have to be used to verify that the solution is a local minimiser of I.

2.3 Iterative Method

We apply to (2.3) the following generalisation of the iterative method of Barrett and Elliott [1]: Given $u^0 \in W$, for $n = 1, 2, \ldots$ find $u^n \in W$ such that

$$b(u^n, \eta - u^n) + J(\eta) - J(u^n) \geq c(u^{n-1}, \eta - u^n) + l(\eta - u^n) \quad \forall \eta \in W. \tag{2.5}$$

If J is in addition Gâteaux differentiable then this is equivalent to the following iterative method: Given $u^0 \in W$, for $n = 1, 2, \ldots$ find $u^n \in W$ such that

$$b(u^n, \eta - u^n) + \langle J'(u^n), \eta - u^n \rangle \geq c(u^{n-1}, \eta - u^n) + l(\eta - u^n) \quad \forall \eta \in W. \tag{2.6}$$

Note that $b(\eta, \eta) + J(\eta)$ is convex and $-c(\eta, \eta) - l(\eta)$ is concave.

Equations (2.5) and (2.6) have unique solutions as they are equivalent to minimising a convex functional over W. Moreover we can prove the following convergence result.

Theorem 2.5. *Every sequence $\{u^n\}$ generated by (2.5) satisfies*

$$I(u^n) + c(u^n - u^{n-1}, u^n - u^{n-1}) + \beta \|u^n - u^{n-1}\|_V^2 \leq I(u^{n-1}) \tag{2.7}$$

and has a subsequence which converges in V to a critical point of (2.2) i.e. a solution of (2.3). Also, the limit of any subsequence of $\{u^n\}$ that converges weakly in V, and hence strongly in H, is a critical point of (2.2).

Proof. The proof is an extension to that of Theorem 6.1 in [1]. To deduce (2.7) we test (2.5) with $\eta = u^{n-1}$ and use the coercivity of b. Because of the assumptions on I, $\{u^n\}$ is uniformly bounded in V, so we can extract a subsequence, which we also denote by $\{u^n\}$, that converges weakly in V and strongly in H to some element $u \in W$. The assumptions on b, c, l and J allow us to pass to the limit in (2.5) and deduce that u satisfies (2.3). The same argument applies to any subsequence, which proves the second part of the theorem.

To see why the convergence in the first part of the theorem is strong in V, note that now we know u satisfies (2.3), we can combine this inequality with (2.5) to get

$$b(u - u^n, u - u^n) \leq c(u - u^{n-1}, u - u^n).$$

The result then follows using the coercivity of b and the strong convergence of u^n in H. □

3 Binary Recovery Application

We now show that (1.7) with both the smooth double well and double obstacle potentials can be fitted into the framework of the previous section.

3.1 Smooth Double Well Potential

Set $V, W := H^1(\Omega), H := L^2(\Omega)$, let $S : H \to H$ satisfy the assumptions in Sect. 1.3, and take

$$\begin{aligned} b(u,\eta) &:= (Su, S\eta) + \sigma_1 \varepsilon (\nabla u, \nabla \eta) \\ c(u,\eta) &:= \frac{\sigma_1}{\varepsilon}(u,\eta) \\ l(u) &:= (S^* y_d, u) \\ J(u) &:= \frac{\sigma_1}{4\varepsilon}\int_\Omega u^4. \end{aligned}$$

Here and throughout this document $(\cdot,\cdot)$ denotes the $L^2(\Omega)$ inner product. S^* denotes the adjoint operator of S, which is defined as follows: For real Hilbert spaces U, V the adjoint operator of a continuous linear operator $A : U \to V$ is the operator $A^* : V \to U$ such that

$$(Au, v)_V = (u, A^* v)_U \quad \forall u \in U, v \in V.$$

The above objects have the properties required in Sect. 2. Coercivity of b can be shown using a contradiction argument and that $S0 = 0$. J is well defined and continuous since $H^1(\Omega)$ is continuously embedded in $L^6(\Omega)$ for $\Omega \subset \mathbb{R}^N$ with $N \leq 3$. I satisfies assumption (2.1) since

$$\int_\Omega \frac{u^4}{4} - \frac{u^2}{2} \geq \int_\Omega \frac{u^2}{2} - 1 = \frac{1}{2}\|u\|^2_{L^2(\Omega)} - |\Omega|,$$

and so

$$\begin{aligned} I(u) \geq \frac{\sigma_1 \varepsilon}{2}\|\nabla u\|^2_{L^2(\Omega)} + \frac{\sigma_1}{\varepsilon}\|u\|^2_{L^2(\Omega)} - \frac{\sigma_1}{\varepsilon}|\Omega| \geq \\ \sigma_1 \min\left\{\frac{\varepsilon}{2}, \frac{1}{\varepsilon}\right\}\|u\|^2_V - \frac{\sigma_1}{\varepsilon}|\Omega| \quad \forall u \in W. \end{aligned}$$

Moreover I equals F_ε from (1.7) with the smooth double well potential (up to an additive constant), so (2.2) becomes: Given $y_d \in L^2(\Omega)$ find

$$\arg\min_{u \in H^1(\Omega)} F_1(u) := \frac{1}{2}\|Su - y_d\|^2_{L^2(\Omega)} + \sigma_1\left(\int_\Omega \frac{\varepsilon}{2}|\nabla u|^2 + \frac{1}{\varepsilon}\Psi_1(u)\right). \tag{3.1}$$

J is Gâteaux differentiable, so solutions to (3.1) satisfy (2.4), which becomes: Given $y_d \in L^2(\Omega)$, find $u \in H^1(\Omega)$ such that

$$(S^*(Su - y_d), \eta) + \sigma_1 \varepsilon (\nabla u, \nabla \eta) + \frac{\sigma_1}{\varepsilon}(u^3 - u, \eta) = 0 \quad \forall \eta \in H^1(\Omega).$$

In this example we have an equality instead of a variational inequality because W is the full space V.

Equation (2.6) gives the following iterative method for solving the above variational inequality, and it converges by Theorem 2.5: Given $y_d \in L^2(\Omega)$ and $u^0 \in H^1(\Omega)$, for $n = 1, 2, \ldots$ find $u = u^n \in H^1(\Omega)$ such that

$$(S^*(Su - y_d), \eta) + \sigma_1 \varepsilon(\nabla u, \nabla \eta) + \frac{\sigma_1}{\varepsilon}(u^3 - u^{n-1}, \eta) = 0 \quad \forall \eta \in H^1(\Omega). \tag{3.2}$$

3.2 Double Obstacle Potential

Define $K := \{u \in H^1(\Omega) : |u| \leq 1 \text{ a.e. in } \Omega\}$. Set $V := H^1(\Omega), W := K, H := L^2(\Omega)$, let $S : H \to H$ satisfy the assumptions in Sect. 1.3, and take

$$\begin{aligned} b(u, \eta) &:= (Su, S\eta) + \sigma_2 \varepsilon(\nabla u, \nabla \eta) \\ c(u, \eta) &:= \frac{\sigma_2}{\varepsilon}(u, \eta) \\ l(u) &:= (S^* y_d, u) \\ J(u) &:= 0. \end{aligned}$$

The above objects have the properties required in Sect. 2. As with the smooth double well potential, I satisfies assumption (2.1) since for $u \in W$ we have

$$-\int_\Omega \frac{u^2}{2} \geq \int_\Omega \frac{u^2}{2} - 1 = \frac{1}{2}\|u\|^2_{L^2(\Omega)} - |\Omega|\,.$$

Moreover I equals F_ε from (1.7) with the double obstacle potential (up to an additive constant), so (2.2) becomes: Given $y_d \in L^2(\Omega)$ find

$$\arg\min_{u \in K} F_2(u) := \frac{1}{2}\|Su - y_d\|^2_{L^2(\Omega)} + \sigma_2 \left(\int_\Omega \frac{\varepsilon}{2} |\nabla u|^2 + \frac{1}{2\varepsilon}(1 - u^2) \right). \tag{3.3}$$

Solutions to (3.3) satisfy (2.4), which becomes: Given $y_d \in L^2(\Omega)$, find $u \in K$ such that

$$(S^*(Su - y_d), \eta - u) + \sigma_2 \varepsilon(\nabla u, \nabla \eta - \nabla u) - \frac{\sigma_2}{\varepsilon}(u, \eta - u) \geq 0 \quad \forall \eta \in K.$$

Equation (2.6) gives the following iterative method for solving the above variational inequality, which converges by Theorem 2.5: Given $y_d \in L^2(\Omega)$ and $u^0 \in K$, for $n = 1, 2, \ldots$ find $u = u^n \in K$ such that

$$(S^*(Su - y_d), \eta - u) + \sigma_2 \varepsilon(\nabla u, \nabla \eta - \nabla u) - \frac{\sigma_2}{\varepsilon}(u^{n-1}, \eta - u) \geq 0 \quad \forall \eta \in K. \tag{3.4}$$

3.3 Alternative Iterative Methods

In (3.2) and (3.4) the S^*S term is taken implicitly, so we need to be able to invert the operator $S^*S - \sigma_i \varepsilon \Delta$ efficiently, otherwise these iterative methods will be too computationally expensive. In some cases this may be possible, for example if S is the identity, but in general this is not the case.

As we remarked earlier, the definitions of b and c that make I correspond to (3.1) and (3.3) are not unique. For example we can set $b(u,\eta) = B(u,\eta) + \rho(u,\eta)$ and $c(u,\eta) = C(u,\eta) + \rho(u,\eta)$ for some $\rho \geq 0$. The $\rho(u,\eta)$ terms cancel out in I, so defining B and C the same way b and c were defined earlier in this section gives the same optimisation problems (3.1) and (3.3). But the corresponding iterative methods are different. The point of this is that the $\rho(u,\eta)$ term is convex (when $\eta = u$), so it gives us more flexibility in how we define B and C while still having b and c satisfy the coercivity and positivity assumptions.

In particular, for suitably large ρ we can take the S^*S term explicitly (which in our framework corresponds to moving it from b to c), and also take the $\frac{\sigma_i}{\varepsilon}(u,\eta)$ term implicitly (i.e. move it from c to b). So for our examples this corresponds to taking

$$b(u,\eta) := \rho(u,\eta) + \sigma_i \varepsilon(\nabla u, \nabla \eta) - \frac{\sigma_i}{\varepsilon}(u,\eta),$$

$$c(u,\eta) := \rho(u,\eta) - (S^*Su,\eta).$$

A restriction such as $\rho > \max\{\frac{\sigma_i}{\varepsilon}, C_s^2\}$, where C_s is the stability constant from (1.6), is then sufficient for both b to be coercive and c to be nonnegative. So we have the following iterative methods, which are in general easier to solve computationally than (3.2) and (3.4).

Example 3.1 (Smooth Double Well). Given $y_d \in L^2(\Omega)$ and $u^0 \in H^1(\Omega)$, for $n = 1, 2, \ldots$ find $u = u^n \in H^1(\Omega)$ such that

$$\rho(u - u^{n-1}, \eta) + (S^*(Su^{n-1} - y_d), \eta) + \sigma_1 \varepsilon(\nabla u, \nabla \eta) + \frac{\sigma_1}{\varepsilon}(u^3 - u, \eta) = 0 \tag{3.5}$$

for all $\eta \in H^1(\Omega)$.

Example 3.2 (Double Obstacle). Given $y_d \in L^2(\Omega)$ and $u^0 \in K$, for $n = 1, 2, \ldots$ find $u = u^n \in K$ such that

$$\rho(u - u^{n-1}, \eta - u) + (S^*(Su^{n-1} - y_d), \eta - u) + \tag{3.6}$$

$$\sigma_2 \varepsilon(\nabla u, \nabla \eta - \nabla u) - \frac{\sigma_2}{\varepsilon}(u, \eta - u) \geq 0$$

for all $\eta \in K$.

When solving Example 3.1 in practice, it is more convenient for us to solve a linear equation. Therefore we linearise the $J'(u)$ term in (3.5) and consider the following iterative method.

Example 3.3 (Smooth Double Well). Given $y_d \in L^2(\Omega)$ and $u^0 \in H^1(\Omega)$, for $n = 1, 2, \ldots$ find $u = u^n \in H^1(\Omega)$ such that

$$\begin{aligned}\rho(u - u^{n-1}, \eta) + (S^*(Su^{n-1} - y_d), \eta) \\ + \sigma_1 \varepsilon(\nabla u, \nabla \eta) + \frac{\sigma_1}{\varepsilon}((u^{n-1})^2 u - u, \eta) = 0\end{aligned} \tag{3.7}$$

for all $\eta \in H^1(\Omega)$.

This iterative method lies outside of our framework, so the convergence theory does not necessarily hold. However it works well in practice.

To finish this section we show how we can reformulate the iterative methods to remove $S^*(Su^{n-1} - y_d)$ when S is defined as in (1.5). For example, (3.6) becomes the following.

Example 3.4 (Double Obstacle). Given $y_d \in L^2(\Omega)$ and $u^0 \in K$, for $n = 1, 2, \ldots$ find $u = u^n \in K$ such that

$$\rho(u - u^{n-1}, \eta - u) + (p^{n-1}, \eta - u) + \sigma_1 \varepsilon(\nabla u, \nabla \eta - \nabla u) - \frac{\sigma_1}{\varepsilon}(u, \eta - u) \geq 0$$

for all $\eta \in K$, where $p^{n-1} \in H^1(\Omega)$ solves

$$\alpha(\nabla p^{n-1}, \nabla \eta) + (p^{n-1}, \eta) = (y^{n-1} - y_d, \eta) \quad \forall \eta \in H^1(\Omega),$$

and y^{n-1} solves the weak form of (1.5) with $u = u^{n-1}$.

4 Gradient Flow

In this section we investigate the gradient flow method for finding critical points of (3.1) and (3.3) from an initial guess u_0. We prove that this method has some desirable properties, and note the link the to iterative method of the previous sections.

4.1 Smooth Double Well Potential

Let u_0 denote the initial guess of the solution and consider the L^2 gradient flow of F_1 in (3.1).

Problem 4.1. Given $y_d \in L^2(\Omega)$ and $u_0 \in H^1(\Omega)$, find $u \in L^2(0, T; H^1(\Omega))$ with weak time derivative $\partial_t u \in L^2(0, T; L^2(\Omega))$ such that $u(0) = u_0$ and

$$(\partial_t u(t), \eta) + (S^*(Su(t) - y_d), \eta) + \sigma_1 \varepsilon(\nabla u(t), \nabla \eta) + \frac{\sigma_1}{\varepsilon}(\Psi_1'(u(t)), \eta) = 0 \tag{4.1}$$

for all $\eta \in H^1(\Omega)$ and almost all $t \in (0, T)$.

Theorem 4.2. *Problem 4.1 has a unique solution.*

Proof. Note that Problem 4.1 is very similar to the Allen-Cahn equation with the smooth double well potential, and the proof follows using standard techniques. See for example the references in Theorem 4.5, where existence and uniqueness is proved for smooth potentials in order to show existence and uniqueness for the double obstacle potential in the limit. □

Theorem 4.3. *If u is a sufficiently smooth solution of Problem 4.1 then the energy $F_1(u(t))$ decreases over time.*

Proof. For some $t \in (0, T)$ we can test (4.1) with $\eta = \partial_t u(t)$ to get

$$\begin{aligned} \|\partial_t u(t)\|^2_{L^2(\Omega)} &+ (S^*(Su(t) - y_d), \partial_t u(t)) \\ &+ \sigma_1 \varepsilon (\nabla u(t), \nabla \partial_t u(t)) + \frac{\sigma_1}{\varepsilon} (\Psi_1'(u(t)), \partial_t u(t)) = 0. \end{aligned} \tag{4.2}$$

Note that

$$\begin{aligned} (S^*(Su(t) - y_d), \partial_t u(t)) &= \frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t} \|Su(t) - y_d\|^2_{L^2(\Omega)}, \\ (\nabla u(t), \nabla \partial_t u(t)) &= \frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t} \|\nabla u(t)\|^2_{L^2(\Omega)}, \\ (\Psi_1'(u(t)), \partial_t u(t)) &= \frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega \Psi_1(u(t)), \end{aligned}$$

so Eq. (4.2) is equivalent to

$$\|\partial_t u(t)\|^2_{L^2(\Omega)} + \frac{\mathrm{d}}{\mathrm{d}t} \left(\frac{1}{2} \|Su(t) - y_d\|^2_{L^2(\Omega)} + \frac{\sigma_1 \varepsilon}{2} \|\nabla u(t)\|^2_{L^2(\Omega)} + \frac{\sigma_1}{\varepsilon} \int_\Omega \Psi_1(u(t)) \right) = 0.$$

Therefore as long as $\partial_t u(t)$ is not zero almost everywhere we have

$$0 > -\|\partial_t u(t)\|^2_{L^2(\Omega)} \geq \frac{\mathrm{d}}{\mathrm{d}t} F_1(u(t)),$$

and hence the energy decreases. □

4.2 Double Obstacle Potential

We can formulate a gradient flow for F_2 from (3.3) in a similar way.

Problem 4.4. Given $y_d \in L^2(\Omega)$ and $u_0 \in H^1(\Omega)$, find $u \in K_T$ with $\partial_t u \in L^2(0,T;L^2(\Omega))$ such that $u(0) = u_0$ and

$$\begin{aligned}(\partial_t u(t), \eta - u(t)) + (S^*(Su(t) - y_d), \eta - u(t)) \\ + \sigma_2 \varepsilon(\nabla u(t), \nabla\eta - \nabla u(t)) - \frac{\sigma_2}{\varepsilon}(u(t), \eta - u(t)) \geq 0\end{aligned} \tag{4.3}$$

for all $\eta \in K$ and almost all $t \in (0,T)$. Here

$$K_T := \{u \in L^2(0,T;H^1(\Omega)) : |u| \leq 1 \text{ a.e. in } (0,T)\times\Omega\}.$$

Theorem 4.5. *Problem 4.4 has a unique solution. Moreover, if u is a sufficiently smooth solution then the energy $F_2(u(t))$ decreases over time.*

Proof. This follows from a slight modification to the arguments for the double obstacle Allen-Cahn inequality in [6, 7, 9, 10, 15] to allow for the S^*Su term. □

For both potentials it is important to consider whether $u(t)$ converges to a steady state as $t \to \infty$. These types of issues are investigated in [27], and in [15] for the 1D double obstacle potential. We do not discuss this as the focus of this work is on iterative methods.

4.3 Link to Iterative Methods

Particular first order discretisations in time of the gradient flow formulations are equivalent to the iterative methods of the previous section with $\rho = \frac{1}{\Delta t}$. But we only want to solve the optimisation problems (3.1) and (3.3); we are not interested in the accuracy of solutions to (4.1) and (4.3) at each point in time, but rather how well they approximate minimisers of F_1 and F_2 for large t. For this reason our method for solving (3.1) and (3.3) should focus on decreasing the energy. The iterative methods of the previous sections are designed to have this property, where as discretisations in time of the gradient flows may not.

The scheme denoted by (2.5) of Barrett and Elliott motivated the convexity splitting implicit/explicit Euler scheme used in [20]. See also [22].

5 Discretisation

In this section we discretise the abstract iterative method of Sect. 2 in space and analyse convergence of the discretisation. We then apply this theory to a finite element discretisation of (3.5) and (3.6) for S defined by (1.5).

5.1 Discrete Abstract Framework

Suppose we have a family of subspaces $V_h \subset V$ and closed convex nonempty subsets $W_h \subset V_h$ which approximate functions in W increasingly well as some parameter $h \to 0$. In particular we suppose we have an approximation operator $P_h : W \to W_h$ such that

$$\|\eta - P_h\eta\|_V \to 0 \text{ as } h \to 0 \quad \forall \eta \in W, \tag{5.1}$$

and that every sequence $\{\eta_h\} \subset W_h$ satisfies

$$\eta_h \rightharpoonup \eta \text{ in } V \text{ as } h \to 0 \implies \eta \in W. \tag{5.2}$$

Remark 5.1. Note that we do not require $W_h \subset W$. If this holds then (5.2) follows automatically because W is a closed convex subset of a Banach space, and hence is weakly sequentially closed.

We now assume there exist objects b_h, c_h and l_h which satisfy the same assumptions as b, c and l, with the boundedness and coercivity constants independent of h. We define

$$I_h(\eta) := \frac{1}{2}b_h(\eta,\eta) + J(\eta) - \frac{1}{2}c_h(\eta,\eta) - l_h(\eta),$$

and as in (2.1) we assume that there exist positive constants α_1 and C_1 independent of h such that

$$I_h(\eta_h) \geq \alpha_1\|\eta_h\|_V^2 - C_1 \quad \forall \eta_h \in W_h. \tag{5.3}$$

So minimisers of I_h over W_h (which exist, since I_h satisfies the same assumptions as I) satisfy the following discrete problem: Find $u_h \in W_h$ such that

$$b_h(u_h, \eta_h - u_h) + J(\eta_h) - J(u_h) \geq c_h(u_h, \eta_h - u_h) + l_h(\eta_h - u_h) \quad \forall \eta_h \in W_h. \tag{5.4}$$

If J is in addition Gâteaux differentiable then this is equivalent to the following discrete variational inequality: Find $u_h \in W_h$ such that

$$b_h(u_h, \eta_h - u_h) + \langle J'(u_h), \eta_h - u_h\rangle \geq c_h(u_h, \eta_h - u_h) + l_h(\eta_h - u_h) \quad \forall \eta_h \in W_h.$$

We need b_h, c_h and l_h to approximate their continuous counterparts as $h \to 0$. So we make the additional assumptions that for any bounded sequence $\{v_h\} \subset W$ we have

$$\|(b - b_h)(v_h, \cdot)\|_{V^*} = \sup_{\eta_h \in V_h \setminus \{0\}} \frac{|b(v_h, \eta_h) - b_h(v_h, \eta_h)|}{\|\eta_h\|_V} \to 0, \tag{5.5}$$

$$\|(c - c_h)(v_h, \cdot)\|_{V^*} = \sup_{\eta_h \in V_h \setminus \{0\}} \frac{|c(v_h, \eta_h) - c_h(v_h, \eta_h)|}{\|\eta_h\|_V} \to 0,$$

$$\|l - l_h\|_{V^*} = \sup_{\eta_h \in V_h \setminus \{0\}} \frac{|l(\eta_h) - l_h(\eta_h)|}{\|\eta_h\|_V} \to 0$$

as $h \to 0$. With these assumptions solutions of the discrete variational inequality (5.4) approximate solutions of the continuous variational inequality (2.3) as $h \to 0$, as the following theorem shows.

Theorem 5.2. *For any sequence $h_n \to 0$ the sequence $\{u_{h_n}\}$ of solutions to (5.4) has a subsequence which converges weakly in V, and hence strongly in H, to a critical point of (2.2) i.e. a solution of (2.3). Moreover, the limit of any subsequence of $\{u_{h_n}\}$ that converges weakly in V, and hence strongly in H, is a critical point of (2.2).*

Proof. For a given h we can find $u_h = \arg\min_{\eta_h \in W_h} I_h(\eta_h)$, then for any $\eta_h \in W_h$,

$$I_h(u_h) \le I_h(\eta_h) = \frac{1}{2} b_h(\eta_h, \eta_h) + J(\eta_h) - \frac{1}{2} c_h(\eta_h, \eta_h) - l_h(\eta_h).$$

Fix $\eta \in W$ and set $\eta_h = P_h \eta \in W_h$. So $\{\eta_h\}$ is bounded in V by (5.1), which means $|b_h(\eta_h, \eta_h) - b(\eta_h, \eta_h)| \le \|(b_h - b)(\eta_h, \cdot)\|_{V^*} \|\eta_h\|_V \le C$. Here and throughout this section C denotes a generic constant independent of h which may vary from line to line. A similar result holds for l_h, and c_h is nonnegative, so

$$I_h(u_h) \le \frac{1}{2} b(\eta_h, \eta_h) + J(\eta_h) + |l(\eta_h)| + C.$$

By the boundedness of b and l,

$$I_h(u_h) \le C(\|\eta_h\|_V^2 + J(\eta_h) + \|\eta_h\|_V).$$

Combining this with (5.3) we get

$$\|u_h\|_V \le C(\|\eta_h\|_V + J(\eta_h) + 1).$$

Now (5.1) and the continuity of J give that $J(\eta_h) \le C$. In addition (5.1) implies that for h less than some h_0, $\|\eta_h\|_V \le \|\eta\|_V + C$, and therefore $\|u_h\|_V \le C$.

From the above it follows that for any sequence $h_n \to 0$, $\{u_{h_n}\}$ is bounded in V. So we can find a subsequence, which we also denote by $\{u_{h_n}\}$, that converges weakly in V and strongly in H to some $u \in V$. In fact $u \in W$ by (5.2). We now show that u is a solution of (2.3).

Note that for all $\eta \in W$ we have

$$\begin{aligned}
&\liminf_{n\to\infty} b_{h_n}(u_{h_n}, P_{h_n}\eta - u_{h_n}) \\
&= \liminf_{n\to\infty} \Big(b_{h_n}(u_{h_n}, P_{h_n}\eta - u_{h_n}) \pm b(u_{h_n}, P_{h_n}\eta - u_{h_n}) \pm b(u_{h_n}, \eta - u_{h_n}) \Big) \\
&= \liminf_{n\to\infty} \Big((b_{h_n} - b)(u_{h_n}, P_{h_n}\eta - u_{h_n}) + b(u_{h_n}, P_{h_n}\eta - \eta) + b(u_{h_n}, \eta - u_{h_n}) \Big) \\
&= \liminf_{n\to\infty} b(u_{h_n}, \eta - u_{h_n}) \\
&\le b(u, \eta - u).
\end{aligned}$$

The final equality follows because $\lim_{n\to\infty}(b_{h_n} - b)(u_{h_n}, P_{h_n}\eta - u_{h_n}) = 0$ by (5.5) and $\lim_{n\to\infty} b(u_{h_n}, P_{h_n}\eta - \eta) = 0$ by (5.1). The inequality follows from the lower semicontinuity of $b(\cdot,\cdot)$ and the continuity of $b(\cdot, \eta)$. Similar results hold for the c_h and l_h terms. This and the continuity and weak lower semicontinuity of J gives

$$\begin{aligned}
b(u, \eta - u) + J(\eta) - J(u) &\ge \liminf_{n\to\infty} \Big(b_{h_n}(u_{h_n}, P_{h_n}\eta - u_{h_n}) + J(P_{h_n}\eta) - J(u_{h_n}) \Big) \\
&\ge \liminf_{n\to\infty} \Big(c_{h_n}(u_{h_n}, P_{h_n}\eta - u_{h_n}) + l_{h_n}(P_{h_n}\eta - u_{h_n}) \Big) \\
&\ge c(u, \eta - u) + l(\eta - u) \quad \forall \eta \in W.
\end{aligned}$$

Hence u is indeed a solution of (2.3).

The same argument applies to any weakly convergent subsequence, which proves the second part of the theorem. □

Remark 5.3. We could also assume we have functionals J_h satisfying the same assumptions as J, with the continuity independent of h, plus the additional property that $v_{h_n} \rightharpoonup v$ in W for $h_n \to 0$ implies $\liminf_{n\to\infty} J_{h_n}(v_{h_n}) \ge J(v)$. Then a proof almost identical to the above gives convergence for (5.4) with J replaced by J_h. This allows numerical integration to be used on the J term.

As with (2.3) in Sect. 2, we can consider an iterative method for solving (5.4): Given $u_h^0 \in W_h$, for $n = 1, 2, \ldots$ find $u_h^n \in W_h$ such that

$$b_h(u_h^n, \eta_h - u_h^n) + J(\eta_h) - J(u_h^n) \ge c_h(u_h^{n-1}, \eta_h - u_h^n) + l_h(\eta_h - u_h^n) \quad \forall \eta_h \in W_h.$$

If J is in addition Gâteaux differentiable then this is equivalent to the following iterative method: Given $u_h^0 \in W_h$, for $n = 1, 2, \ldots$ find $u_h^n \in W_h$ such that

$$b_h(u_h^n, \eta_h - u_h^n) + (J'(u_h^n), \eta_h - u_h^n) \ge c_h(u_h^{n-1}, \eta_h - u_h^n) + l_h(\eta_h - u_h^n) \quad \forall \eta_h \in W_h.$$

Since b_h, c_h and l_h satisfy the same assumptions as b, c, and l, the above iterative method still has the energy decreasing property, and we get convergence of iterates

to a solution of (5.4). Then as $h \to 0$ the solutions of (5.4) converge to critical points of (2.2) by Theorem 5.2.

5.2 Finite Element Discretisation of (3.5) and (3.6)

Assume that Ω is polyhedral and let $\{T_h\}$ be a family of uniform regular triangulations of Ω into disjoint open simplices with a maximal element size h. Associated with each T_h we have the piecewise linear finite element space

$$V_h := \{v \in C^0(\bar{\Omega}) : v|_T \in P_1(T) \text{ for all } T \in T_h\} \subset H^1(\Omega),$$

where $P_1(T)$ is the set of all linear affine functions on T. Also define

$$K_h := \{v_h \in V_h : |v_h| \leq 1 \text{ in } \Omega\}$$

so that we have a finite element space analogous to K. Note that $K_h \subset K$ so Remark 5.1 applies. Take P_h to be the operator that maps $u \in W$ to the unique $P_h u \in W_h$ such that

$$(P_h u, \eta_h - u)_{H^1(\Omega)} \geq (u, \eta_h - u)_{H^1(\Omega)} \quad \forall \eta_h \in W_h.$$

This operator satisfies Eq. (5.1), see e.g. Chapter 2 in [24].

Let S be the solution operator of (1.5), and denote by S_h the discrete blurring operator. We intend this to approximate S, so we define S_h to map $u \in L^2(\Omega)$ to the unique $y_h \in V_h$ satisfying

$$\alpha(\nabla y_h, \nabla \eta_h) + (y_h, \eta_h) = (u, \eta_h) \quad \forall \eta_h \in V_h. \tag{5.6}$$

A stability estimate the same as (1.6) holds, so

$$\|y_h\|_{L^2(\Omega)} = \|S_h u\|_{L^2(\Omega)} \leq C_s(\alpha)\|u\|_{L^2(\Omega)}, \tag{5.7}$$

where as before $C_s(\alpha) = \frac{1}{1+\alpha/C_p}$. Also standard error analysis for elliptic PDEs says

$$\|y - y_h\|_{L^2(\Omega)} \leq Ch\|y\|_{H^1(\Omega)},$$

which combined with (5.7) gives that

$$\|(S - S_h)u\|_{L^2(\Omega)} \leq Ch\|Su\|_{H^1(\Omega)} \leq Ch\|u\|_{L^2(\Omega)}. \tag{5.8}$$

Example 5.4 (Smooth Double Well). Take the same definitions as in Example 3.1. In addition take V_h as above, $W_h := V_h$, and define

$$\begin{aligned} b_h(u_h, \eta_h) &:= \rho(u_h, \eta_h) + \sigma_1 \varepsilon(\nabla u_h, \nabla \eta_h) - \frac{\sigma_1}{\varepsilon}(u_h, \eta_h) \\ c_h(u_h, \eta_h) &:= \rho(u_h, \eta_h) - (S_h u_h, S_h \eta_h) \\ l_h(u_h) &:= (S_h^* y_{d,h}, u_h) \\ J(u_h) &:= \frac{\sigma_1}{4\varepsilon} \int_\Omega u_h^4, \end{aligned}$$

where S_h is the discrete elliptic operator defined by (5.6), and $y_{d,h}$ is the L^2-projection of y_d onto V_h.

For $\rho > \max\{\frac{\sigma_1}{\varepsilon}, C_s^2\}$, where C_s is the stability constant from (5.7), all the assumptions of Theorem 2.5 are satisfied, so we get the decreasing energy property and convergence of iterates for the following discrete iterative method: Given $y_{d,h}$, $u_h^0 \in V_h$, for $n = 1, 2, \ldots$ find $u_h = u_h^n \in V_h$ such that

$$\begin{aligned} \rho(u_h - u_h^{n-1}, \eta_h) + (p_h^{n-1}, \eta_h) + \sigma_1 \varepsilon(\nabla u_h, \nabla \eta_h) \\ + \frac{\sigma_1}{\varepsilon}(u_h^3 - u_h, \eta_h) = 0 \quad \forall \eta_h \in V_h, \end{aligned}$$

where y_h^{n-1}, $p_h^{n-1} \in V_h$ satisfy

$$\begin{aligned} \alpha(\nabla y_h^{n-1}, \nabla \eta_h) + (y_h^{n-1}, \eta_h) &= (u_h^{n-1}, \eta_h) \\ \alpha(\nabla p_h^{n-1}, \nabla \eta_h) + (p_h^{n-1}, \eta_h) &= (y_h^{n-1} - y_{d,h}, \eta_h) \end{aligned}$$

for all $\eta_h \in V_h$.

The assumptions of Theorem 5.2 are also satisfied, since for a weakly convergent sequence $\{v_h\} \in V$ we have

$$\begin{aligned} |(b_h - b)(v_h, \eta_h)| &= |(S_h v_h, S_h \eta_h) - (S v_h, S \eta_h)| \\ &\le |(S_h v_h, (S_h - S)\eta_h)| + |((S_h - S)v_h, S\eta_h)| \\ &\le \|S_h v_h\|_{L^2(\Omega)} \|(S_h - S)\eta_h\|_{L^2(\Omega)} + \|(S_h - S)v_h\|_{L^2(\Omega)} \|S\eta_h\|_{L^2(\Omega)}. \end{aligned}$$

Now using (5.7) and (5.8) we get

$$\|(b_h - b)(v_h, \cdot)\|_{H^1(\Omega)^*} \le C h \|v_h\|_{H^1(\Omega)},$$

and so $\|(b_h - b)(v_h, \cdot)\|_{H^1(\Omega)^*} \to 0$ as $h \to 0$ by the boundedness of $\|v_h\|_V$. Similar results hold for c_h and l_h. Therefore we have convergence of limit points of the above discrete iterative method to critical points of (3.1) as $h \to 0$.

Remark 5.5. As mentioned before Example 3.3, when solving the smooth double well problem in practice, we solve a finite element discretisation of the linearised iterative method (3.7): Given $y_{d,h}$, $u_h^0 \in V_h$, for $n = 1, 2, \ldots$ find $u_h = u_h^n \in V_h$ such that

$$\begin{aligned} \rho(u_h - u_h^{n-1}, \eta_h) + (p_h^{n-1}, \eta_h) + \sigma_1 \varepsilon(\nabla u_h, \nabla \eta_h) \\ + \frac{\sigma_1}{\varepsilon}((u_h^{n-1})^2 u_h - u_h, \eta_h) = 0 \quad \forall \eta_h \in V_h, \end{aligned} \tag{5.9}$$

where y_h^{n-1}, $p_h^{n-1} \in V_h$ satisfy

$$\begin{aligned} \alpha(\nabla y_h^{n-1}, \nabla \eta_h) + (y_h^{n-1}, \eta_h) &= (u_h^{n-1}, \eta_h) \\ \alpha(\nabla p_h^{n-1}, \nabla \eta_h) + (p_h^{n-1}, \eta_h) &= (y_h^{n-1} - y_{d,h}, \eta_h) \end{aligned} \tag{5.10}$$

for all $\eta_h \in V_h$.

We use numerical integration on the linearised term. Note that the theorems do not necessarily hold for this iterative method, but it performs well in practice.

Example 5.6 (Double Obstacle). Take the same definitions as in Example 3.2. In addition take V_h as above, $W_h := K_h$, and define

$$\begin{aligned} b_h(u_h, \eta_h) &:= \rho(u_h, \eta_h) + \sigma_2 \varepsilon(\nabla u_h, \nabla \eta_h) - \frac{\sigma_2}{\varepsilon}(u_h, \eta_h) \\ c_h(u_h, \eta_h) &:= \rho(u_h, \eta_h) - (S_h u_h, S_h \eta_h) \\ l_h(u_h) &:= (S_h^* y_{d,h}, u_h) \\ J(u_h) &:= 0, \end{aligned}$$

where S_h is the discrete elliptic operator defined by (5.6), and $y_{d,h}$ is the L^2-projection of y_d onto V_h.

For $\rho > \max\{\frac{\sigma_2}{\varepsilon}, C_s^2\}$ all the assumptions of Theorem 2.5 are satisfied, so we get the decreasing energy property and convergence of iterates for the following discrete iterative method: Given $y_{d,h} \in V_h$ and $u_h^0 \in K_h$, for $n = 1, 2, \ldots$ find $u_h = u_h^n \in K_h$ such that

$$\begin{aligned} \rho(u_h - u_h^{n-1}, \eta_h - u_h) + (p_h^{n-1}, \eta_h - u_h) + \sigma_2 \varepsilon(\nabla u_h, \nabla \eta_h - \nabla u_h) \\ - \frac{\sigma_2}{\varepsilon}(u_h, \eta_h - u_h) \geq 0 \quad \forall \eta_h \in K_h \end{aligned} \tag{5.11}$$

where y_h^{n-1}, $p_h^{n-1} \in V_h$ satisfy

$$\begin{aligned} \alpha(\nabla y_h^{n-1}, \nabla \eta_h) + (y_h^{n-1}, \eta_h) &= (u_h^{n-1}, \eta_h) \\ \alpha(\nabla p_h^{n-1}, \nabla \eta_h) + (p_h^{n-1}, \eta_h) &= (y_h - y_{d,h}, \eta_h) \end{aligned}$$

for all $\eta_h \in V_h$.

Theorem 5.2 gives convergence of limit points of the above discrete iterative method to critical points of (3.3) as $h \to 0$.

5.3 Algorithms

The discrete iterative methods in Examples 5.4 and 5.6 lead to the following algorithms for binary image recovery, which we implement and test in the next section.

5.3.1 Smooth Double Well Potential

Given $y_{d,h} \in V_h$ and an initial guess $u_h^0 \in V_h$, set $n = 1$ then:

1. Solve (5.10) for y_h^{n-1} then p_h^{n-1};
2. Solve (5.9) for u_h^n;
3. If $\|u_h^n - u_h^{n-1}\|_{L^2(\Omega)} < \text{TOL}$ terminate the algorithm. Else set $n = n + 1$ and go to step 1;

An alternative stopping criterion would be to wait until the change in energy $\left|F_1(u_h^n) - F_1(u_h^{n-1})\right|$ is sufficiently small. This has the advantage that the energy decreasing result then guarantees our algorithm terminates. However the stopping criterion in the above algorithm also gives a strong indication of a steady state, and it seems to work better in practice.

Note that despite the blurring and noise, $y_{d,h}$ still contains a lot of information about the solution. Therefore it makes sense to scale and threshold $y_{d,h}$ in order to get a good initial guess for u_h^0.

5.3.2 Double Obstacle Potential

The algorithm for this potential is the same as for the smooth double well potential, but we instead solve the variational inequality (5.11) in step 2.

One method for solving the variational inequalities at each iteration is the primal-dual active set (PDAS) method. It is applied to solving the variational inequalities arising in the Allen-Cahn inequality in [6]. We implemented this method and found it to work well. However, for the numerics in the next sections we use an alternative method known as the Truncated Nonsmooth Newton Multigrid (TNNMG) method (see [23, 25]), which performs very well.

6 Numerics

In this section we show some numerical examples of binary recovery in 1 and 2 dimensions. The data is blurred by the solution operator of the elliptic PDE (1.5), with the parameter α controlling the level of blurring. It also has additive Gaussian noise of mean zero and variance γ.

We do the recovery using the discrete iterative methods of Remark 5.5 (based on the smooth double well potential) and (5.11) (based on the double obstacle potential). In practice we observe convergence of the full sequence of iterates to steady states, which are discrete critical points of (1.7). As we take ε and h small, we believe that these critical points closely approximate a global minimiser of the model (1.4). This is because the iterative methods give us discrete critical points of the approximate model (1.7), which seem to be at least discrete local minimisers of (1.7), as different initial iterates and (valid) values of ρ do not lead to different steady states. In addition, for small ε (and appropriate h) the critical points are close to being binary i.e. feasible minimisers of the model (1.4). We cannot be certain how close they really are to the global minimisers of (1.4) due to the lack of explicitly known global minimisers for interesting problems. Regardless, by artificially generating data from a known binary function, the numerical results show that for small ε (and appropriate h) our iterative methods are effective at recovering something close to the binary function.

The weighting given to the regularisation (the parameter σ), which defines the nonconvex model (1.4), is an important but challenging issue. If we take σ too small then recovered functions still have artifacts of the noise. If σ is too large then we loose some features we actually want to keep. We show some figures and discuss some results on the choice of σ for related problems in Appendix A.1, however the theory does not apply to our particular problem. In this section we just take values of σ that we have experimentally determined to work well for the problem at hand.

For the implementation we use the Distributed and Unified Numerics Environment (DUNE), see [2–5, 18, 19]. DUNE provides interfaces for grids, solvers and finite element spaces. Therefore once the algorithms are implemented, it takes minimal effort to change features of the implementation that would usually be fixed, such as the grid type, the dimension of the problem, and the type of finite elements used.

6.1 1D Numerics

The test problem in 1D is inspired by the barcode problem of [21], which was mentioned as a motivating example in Sect. 1.1. We try to recover a binary function taking the values $\{-1, 1\}$, which one can imagine represents a cross section of a barcode (with values of -1 corresponding to black parts of the barcode and values of 1 corresponding to white parts). We suppose this binary function is corrupted,

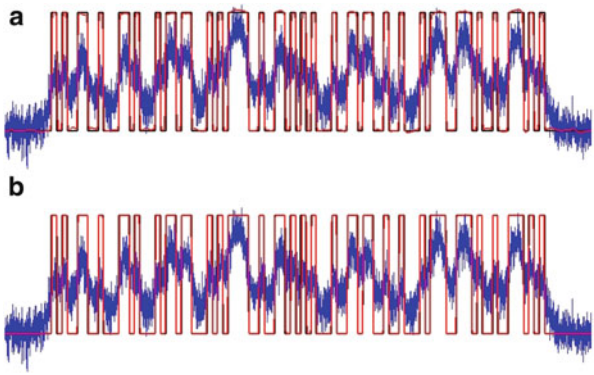


Fig. 1 $\alpha = 1e-4, \gamma = 0.4, \sigma = 1e-4, \varepsilon = 5.31e-4$ and $h = 1.67e-4$. (**a**) Smooth double well potential. (**b**) Double obstacle potential

giving blurred and noisy data that we want to decode. The main difference between our test problem and the barcode problem in [21] is that we have chosen blurring caused by the solution operator of an elliptic PDE instead of a convolution. Although this is not a realistic blurring operator specified by this application, if our approach is effective for this blurring operator then it is likely to be effective for other blurring operators.

The recovery using both the smooth double well and double obstacle potentials can be found in Fig. 1. The black lines represent the binary function that we want to recover, the blue lines are the artificial data we generate by adding blurring and noise, and the red lines are the recovered functions for each potential. Even by eye it is not clear exactly how many 'bars' are in the binary functions, or the correct widths of the bars. But the recovered functions closely match the binary function we started with (which is why the black lines are almost hidden by the red lines), showing that our approach is effective. The figure also makes apparent one of the advantages of the double obstacle potential, which is that recovered functions take a form closer to what we actually want; binary functions.

6.2 2D Numerics

The test problems in 2D involve recovering binary functions with discontinuities of various shapes. In this dimension the problems have a natural interpretation as deblurring and denoising of images, but we also view them as binary source recovery problems for elliptic PDEs.

Figure 2 shows the recovery of a binary function using (1.7) with the smooth double well potential. The discontinuity is a 'blob' shape and is marked by a black line. The blurred and noisy data for this function is shown in Fig. 2a,b. Figure 2c shows the recovered function, with a yellow line marking the zero level set. We can see that the yellow line closely matches the black line, except for a slight mismatch at the concave parts of the discontinuity. Note that we cannot make the interface

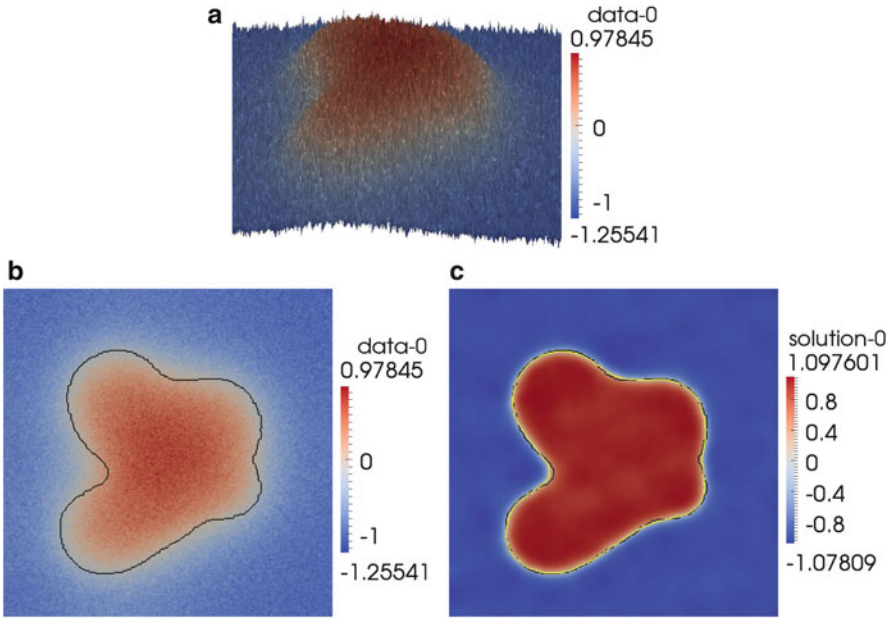


Fig. 2 $\alpha = 0.01$, $\gamma = 0.2$, $\sigma = 1e-4$, $\varepsilon = 0.00879$ and $h = 0.00345$ using the smooth double well potential

as small as for the 1D problem as the resolution of the grid needed to resolve it makes this computationally expensive. Our implementation is capable of adaptivity, which lessens this cost somewhat, but we will not demonstrate this functionality in this work. With this simple visualisation the recovered function using the double obstacle potential looks very similar, so we do not include a figure of it.

Figure 3 (which can be interpreted in the same way as Fig. 2) shows the recovery of a binary function with a letter 'A' shaped discontinuity. This time we use the double obstacle potential in (1.7), though the recovered function using the smooth double well potential looks similar. This example shows that the model can also recover discontinuities with corners reasonably accurately, but there is some rounding of these corners due to the regularisation.

To finish this section we show an example which relates to an application of binary image recovery in 2D. Figure 4 shows the recovery of a binary function representing a QR code with 25×25 blocks (the size typically used to encode a URL). The yellow lines mark the discontinuity of the binary function. Figure 4a shows the data with a red line marking the zero level set, and Fig. 4b shows the recovered function. We see that features which are blurred below the zero level set (and which therefore would not be recovered by a simple projection) are nevertheless recovered by the model.

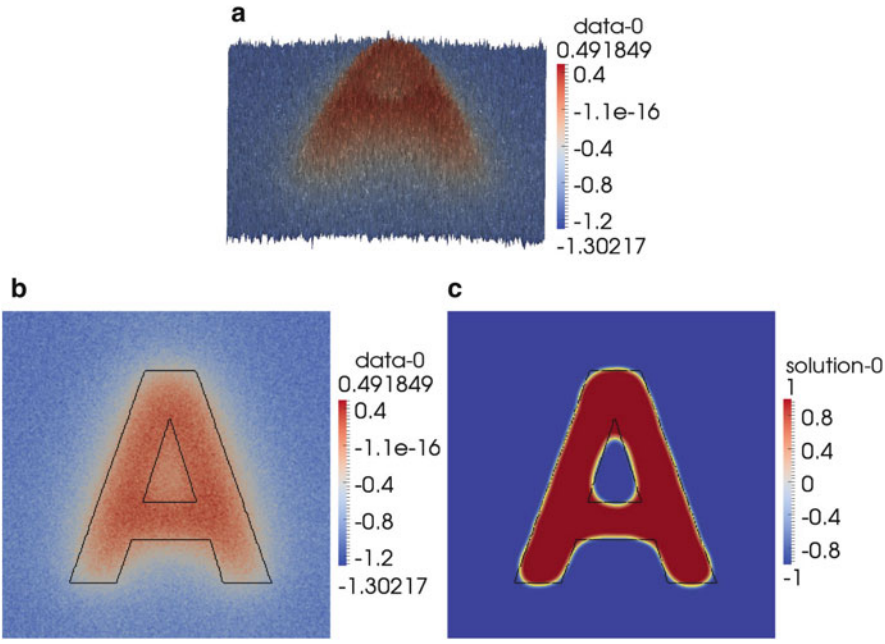


Fig. 3 $\alpha = 0.01$, $\gamma = 0.2$, $\sigma = 1e-4$, $\varepsilon = 0.00879$ and $h = 0.00345$ using the double obstacle potential

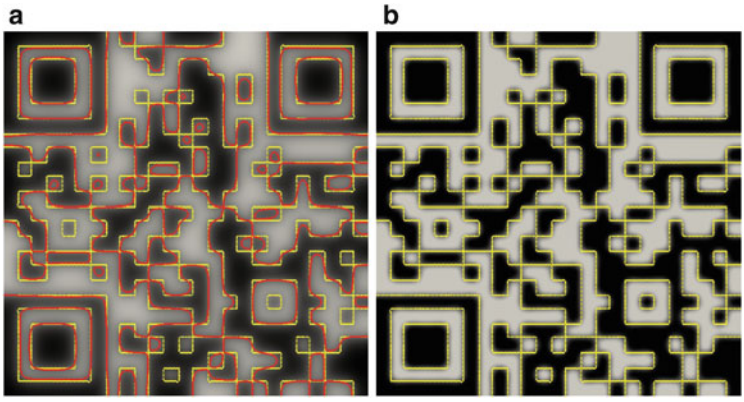


Fig. 4 $\alpha = 5e-4$, $\gamma = 0$, $\sigma = 1e-5$, $\varepsilon = 0.00373$ and $h = 0.00146$ using the double obstacle potential

7 Comparison of Potentials in 1D

Due to the Γ-convergence result of Theorem 1.1, we expect that critical points of (1.7) for a given value of σ using either the smooth double well or double obstacle potential will converge to critical point of (1.4) in the limit of small ε. Of course the

critical points they converge to are not guaranteed to be the same, but agreement of the limits is observed in practice, and for very small ε the recovered functions for both potentials are almost indistinguishable. However it is well known that for phase field type problems, the interface should be well resolved in order for an accurate spatial approximation. This means that the smaller ε, the more grid points needed, and the higher the computational cost of the iterative methods. For many applications we only want to recover the location of the discontinuities in a binary function, which we suppose are given by the zero level set of the recovered function.

This motivates us to consider in this section how well we can recover the locations of the discontinuities with ε of moderate size (rather than as small as possible), which is computationally cheaper. In this case the choice of potential does not just affect the implementation and speed of the iterative method; the recovered functions will in general look quite different, and there may be differences in how accurately or reliably the locations of the discontinuities are recovered.

As in Sect. 6 we consider a problem with blurring caused by the solution operator of the elliptic PDE (1.5) and additive Gaussian noise of mean zero and variance γ. We use the discrete iterative method of Remark 5.5 for the smooth double well potential and (5.11) for the double obstacle potential.

At this stage it is helpful to recall the parameters we have introduced so far, as well as introduce a new parameter ω, the width of the smallest bar in the binary function. The parameters are contained in Table 1, and have been classified as follows:

- Problem parameters—Define the problem we are trying to solve. In applications we have no control over these, though we suppose they are known a priori.
- Model parameters—Specify the model we will use to solve the problem. Different values can lead to the recovery of quite different functions, so they need to be chosen carefully.
- Approximation parameters—We do not work with the model, but rather an approximation of it. These parameters control how good the approximation is.

Table 1 Parameter types

Parameter	Description	Type of parameter	Optimal value
ω	Width of smallest bar in binary function	Problem	–
α	Level of blurring	Problem	–
γ	Level of noise	Problem	–
σ	Weighting given to perimeter regularisation	Model	$\omega/80$
ε	Order of width of interface	Approximation	$\omega/4\pi$
h	Grid width	Discretisation	$\omega/32$
u^0	Initial iterate	Iteration	–
ρ	Parameter in iterative method	Iteration	DW: 0.833, DO: 0.588
TOL	Stopping criterion	Implementation	DW: $3e-4$, DO: $3.5e-4$

- Discretisation parameters—Affect the accuracy of the spatial discretisation in the iterative method.
- Iteration parameters—Determine the behavior of the iterative method.
- Implementation parameters—Control the finer details of the implementation.

We also have a number of less significant implementation parameters that handle the imprecision of computer arithmetic. These will be set to sensible values and ignored in our discussion.

Motivated by the above discussion we now investigate differences between the smooth double well and double obstacle potentials in accuracy, reliability, speed, and implementational complexity.

7.1 Accuracy

Denote the binary function we want to recover by $\bar{u}$ and the recovered function by $u_{\varepsilon,h}$. We measure the accuracy of the recovery by calculating the error quantity

$$E(u_{\varepsilon,h}) := \frac{1}{4}\big|\,|P(u_{\varepsilon,h})|_{TV} - |\bar{u}|_{TV}\,\big| + \frac{1}{2}\|P(u_{\varepsilon,h}) - \bar{u}\|_{L^1(\Omega)},$$

where P is the L^2 projection onto the space $BV(\Omega, \{-1, 1\})$ (i.e. $P(u_{\varepsilon,h}) = 1$ when $u_{\varepsilon,h} \geq 0$ and -1 when $u_{\varepsilon,h} < 0$). $|u|_{TV}$ is the total variation of u, as defined in Sect. 1.2. The integer part of $E(u_{\varepsilon,h})$ tells us the absolute difference between the number of bars in the projected recovered function and $\bar{u}$. The decimal part tells us whether the discontinuities in the projected recovered function are in the correct locations. So E measures the accuracy of the recovery in a sense that matters in applications.

We project because our best guess of $\bar{u}$ should lie in $BV(\Omega, \{-1, 1\})$. The downside of this is that $P(u_{\varepsilon,h})$ is not a minimiser of (1.7). It is important to note that the recovery using the double obstacle potential is naturally much closer to being binary than with the smooth double well potential, so projection is less necessary. This is a big advantage of using the double obstacle potential, which must be remembered when values of $E(u_{\varepsilon,h})$ seem comparable.

The test problems we use for our comparison use the same binary function as in Sect. 6.1 (which has $\omega = \frac{1}{113}$), and different levels of blurring and noise i.e. a range of values of α and γ. We first fix σ based on the size of ω (as described in Appendix A.1) then choose good values of the approximation and discretisation parameters (as described in Appendices A.2 and A.3). So we have $\sigma = 1e-4$, $\varepsilon = 7.06e-4$ and $h = 2.77e-4$ for both potentials. Each realisation of the noise will be different, so we calculate an average E over multiple realisations of the noise. As we observed earlier, we get the same steady state of (1.7) regardless of the choice of iteration parameters. The same is true for implementation parameters. So we ignore both these types of parameters in our discussion of accuracy.

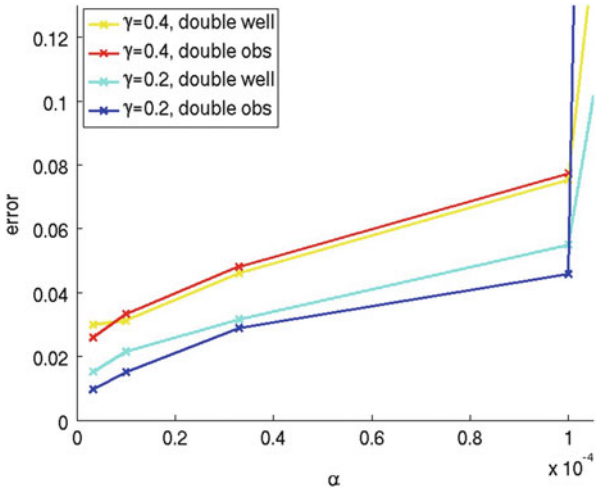


Fig. 5 The error (averaged over many realisations of the noise) for both potentials at different levels of blurring and noise

We see in Fig. 5 that neither potential is the most accurate in all circumstances. For moderate levels of noise ($\gamma = 0.2$), the double obstacle potential leads to a slightly more accurate recovery. However for high levels of noise ($\gamma = 0.4$), the smooth double well potential seems to perform slightly better. Without projection the double obstacle potential always leads to a recovery which is significantly more accurate than the smooth double well potential.

7.2 Reliability

By reliability we refer to the range of problems (i.e. the levels of blurring and noise) over which a binary function can be recovered with reasonable accuracy; as the amount of blurring and noise are increased, eventually the recovered function does not resemble the binary function we wanted. Note that this range will depend on σ. We do not do a detailed comparison of reliability, but feel that it is comparable for both potentials. For example, we can see in Fig. 5 that $\alpha = 1e-4$ and $\gamma = 0.4$ is roughly the limit at which the correct number of bars can be recovered using either potential.

7.3 Speed

The time it takes to recover a function which resembles the binary function is an important practical consideration. Where as accuracy is independent of the implementation, this is certainly not the case for speed. All but the inner workings of each iterative method in our implementation are identical, so we will do our best to make a fair comparison of speed.

Table 2 Average runtimes for $\alpha = 1e-4$ and $\gamma = 0.2$

	Time for rough recovery (s)	Time for accurate recovery (s)
Smooth double well		
Average time/it	0.0359	0.181
# iterations	11	170
Runtime	0.41	29.9
Double obstacle		
Average time/it	0.0639	0.255
# iterations	9	170
Runtime	0.58	42.6

We perform this comparison for the binary function of Sect. 6.1, one choice of blurring and noise ($\alpha = 1e-4$ and $\gamma = 0.2$), and σ as in Sects. 7.1 and 7.2. Choices of ε and h as well as iteration and implementation parameters have a big impact on speed, so we will test two different combinations of these parameters. Our timings can be found in Table 2.

The runtimes for 'accurate recovery' use ε and h as in Sects. 7.1 and 7.2, and TOL as described in Appendix A.5. These values have been chosen to ensure robustness. The table also contains timings for 'rough recovery', where less conservative parameter values are used ($\varepsilon = \frac{\omega}{2\pi}$, $h = \frac{\omega}{20}$, and TOL as described in Appendix A.5). For many problems we can still get a reasonable recovery with these parameter values, and it lowers the computation time significantly.

The recovery times are comparable for each potential for both rough and accurate recovery, though the smooth double well potential has a slight advantage for this size of problem. However we remark that the recovery time of the double obstacle potential scales better as the number of degrees of freedom in the discretisation increases, so it has better performance in 2D.

7.4 Implementational Complexity

Implementing the iterative method for the double obstacle potential is less standard as we are solving variational inequality rather than a PDE. But it is no more complicated than implementing adaptivity, which is needed for the computational cost of the iterative method for the smooth double well potential to scale well to dimensions 2 and higher.

7.5 Summary of Comparison

Both potentials can accurately recover binary functions over the same range of blurring and noise. If no projection is used, the double obstacle potential produces significantly more accurate results. Even with projection it is more accurate for

moderate levels of blurring and noise. Our implementation using the smooth double well potential is slightly quicker for both accurate and rough binary recovery on our 1D test problem. However our implementation using the double obstacle potential, which is overall no more complicated, scales better to many degrees of freedom and so tends to be quicker in higher dimensions.

Appendix A Parameter Choices

In this appendix we describe our methodology for choosing parameter values for the numerical tests and comparisons in Sects. 6 and 7.

A.1 Choice of Model Parameter σ

We recover different functions for different values of σ, so it is important to choose the 'right' value. This is illustrated in Fig. 6, where we show the recovered functions for the same problem as in Fig. 1a for different values of σ. We see that $\sigma = 5e-3$ leads to too few bars being recovered. The recovered function for $\sigma = 1e-6$ follows the noise too much and does not resemble a binary function. With $\sigma = 1e-4$ we recover something close to the binary function that generated the data, so we consider this to be a good value.

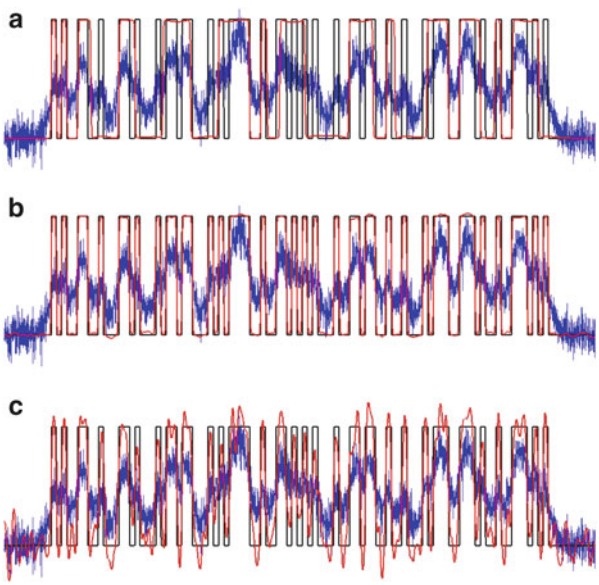


Fig. 6 The problem of Fig. 1a with different values of σ. (**a**) $\sigma = 5e-3$. (**b**) $\sigma = 1e-4$. (**c**) $\sigma = 1e-6$

It is known that the choice of σ in (1.4) should be related to the variance of the noise. Noise with a large variance requires a large σ in order for good recovery. We could try to figure out the variance of the noise from the data and use this to choose σ, however there is not an explicit form for the relationship. Instead we choose σ based on the length scale of the features that we want to recover (i.e. the parameter ω), and use the same σ for all levels of noise. In applications this is generally known a priori e.g. for barcode recovery. This approach works well because we take σ be as large as possible while not removing the features we want to recover, and hence perform the maximum amount of denoising. We do not seem to pay a significant price for this large σ in cases where the noise is small, and this approach leads to a simple rule for choosing σ. The literature that gives us a heuristic way of choosing such a σ is introduced below.

The following result shows that it is unwise to take σ too large.

Proposition A.1. *There exists a $\sigma^* > 0$ such that the minimiser of (1.4) is 0 iff $\sigma > \sigma^*$.*

Proof. Proposition 5.7 in [13]. □

But we also need to be careful not to take σ too small. In fact, since S is known we have the following result in the 1D case.

Theorem A.2. *In the absence of noise there exists a $\sigma_* > 0$ such that the minimiser of (1.4) is $\bar{u}$ whenever $\sigma \leq \sigma_*$.*

Proof. Proposition 5 in [21]. □

Another interesting result is Theorem 1.1 part 2 in [16], which proves more explicit conditions on σ to ensure exact recovery in the case that S is a convolution with a hat function in 1D. Due to our complicated form for S we are forced to use a more heuristic argument to choose a good value for σ.

Chan et al. [14] shows that for the 1D case in the absence of blurring and noise (i.e. binary data), local and global minimisers of (1.4) can be calculated explicitly for a given value of σ. These considerations suggest we should take σ to be smaller than a quarter of the size of the smallest object we want to recover. In particular, $\sigma = \frac{\omega}{8}$ seems like a sensible choice. But this assumes binary data. We have blurring, which means the differences between the functions in the $\|Su - y_d\|^2_{L^2(\Omega)}$ term can be much smaller. Hence we take σ an order of magnitude smaller i.e. $\sigma = \frac{\omega}{80}$. This σ is still larger than the length scale of the noise (which is of order h), so the results in [14] say it will be removed. Numerical experiments confirm that this choice of σ works well in practice.

A.2 Choice of ε

The phase field approximation in (1.7) results in solutions with interfaces of width $o(\varepsilon)$. In order for an accurate spatial approximation we need a reasonable number

of grid points across the interfaces. So a smaller ε requires more grid points and a higher computational cost. With this is mind we want to take ε as large as we can while still resolving the finest features of the binary function. So the choice of ε should be related to the value of ω.

We assume that there is a linear relationship between the optimal choice of ε and ω and deduce the constant of proportionality c_1 such that we get a good recovery with $\pi\varepsilon = c_1\omega$. Note that $\pi\varepsilon$ is the asymptotic width of the interface for minimisers of the Ginzburg-Landau functional with the double obstacle potential, and a good approximation with the smooth double well potential. The width of interfaces in minimisers of (1.7), a perturbed Ginzburg-Landau functional, are approximately the same size. So c_1 can be thought of as the relative width of the interface compared to the width of the smallest bar.

To determine c_1 we recover a simple binary function which can be seen in Fig. 7. We take $\omega_1 = \omega_2 = \omega_3 = 0.2$ (i.e. bars of equal widths), as we found the case where all bars are at the finest length scale to be the hardest for accurate recovery. We consider different levels of blurring and noise and compute the error E of the recovered functions. We take σ to be the optimal value of $\frac{\omega}{80}$ that we decided upon in Appendix A.1, and take $\pi\varepsilon = 50h$ to ensure that effects of the spatial discretisation do not distort our results.

We observe that for a high signal to noise ratio we can take c_1 very large and still get accurate recovery ($\alpha = 0.01$ in Fig. 8), even though the bars do not separate properly (see Fig. 9a). For low signal to noise ratios ($\alpha = 0.1$ in Fig. 8) we need

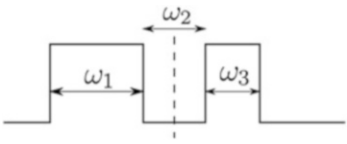


Fig. 7 A simple binary function

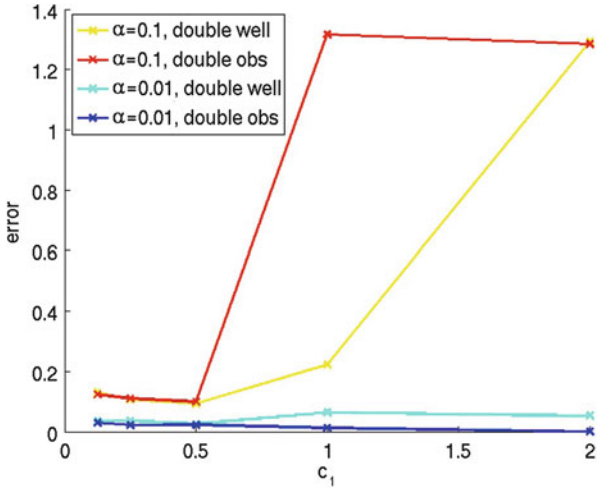


Fig. 8 Errors (averaged over many realisations of the noise) for both potentials at different levels of blurring and $\gamma = 0.2$

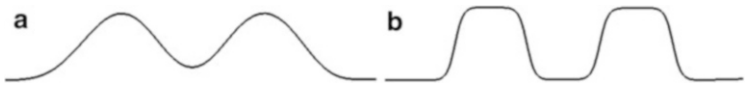


Fig. 9 The interfaces using the smooth double well potential with different values of c_1. Figure 9b shows the interfaces for $c_1 = 0.25$, which we decide is the optimal parameter value. (**a**) $c_1 = 2.0$. (**b**) $c_1 = 0.25$

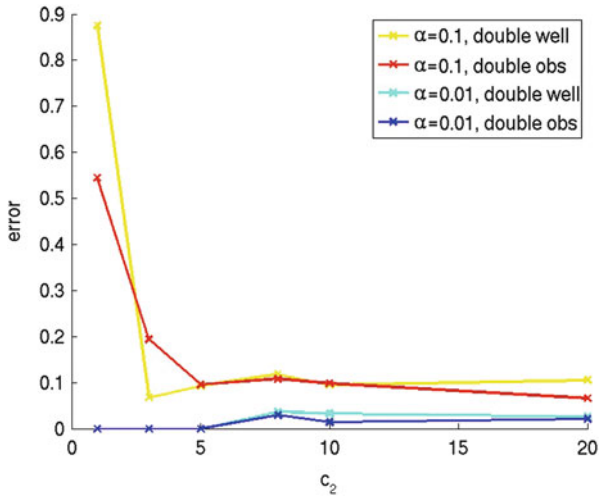


Fig. 10 Errors (averaged over many realisations of the noise) for both potentials at different levels of blurring and $\gamma = 0.2$ with different values of c_2

to take $c_1 \leq 0.5$ for accurate recovery, though it is not until $c \leq 0.25$ that the interfaces start to look reasonably sharp (see Fig. 9b). As expected there is not an accuracy penalty for taking c_1 too small, however it increases computation time by forcing us to take smaller h in order to resolve the interfaces. This motivates us to take $c_1 = 0.25$ i.e. $\pi\varepsilon = \frac{\omega}{4}$.

A.3 Choice of h

We use the same test problems as in Appendix A.2 to deduce a constant factor c_2 such that we get a good recovery with $\pi\varepsilon = c_2 h$. Hence c_2 can be thought of as the number of grid elements across each interface.

With a high signal to noise ratio ($\alpha = 0.01$ in Fig. 10) it can actually be advantageous to have few grid points across the interface. In this case the recovered function would have to deviate a long way from the binary function in order for the projection to take an incorrect value on even a single grid point, and the data does not force sufficient deviation. As a result we can actually get perfect recovery on coarse grids. However, if we have a poorly resolved interface we are not well

approximating our model and we may get a bad recovery for low signal to noise ratios ($\alpha = 0.1$ in Fig. 10).

We do not want to adjust the relationship between ε and h for different levels of blurring and noise; we want a relationship for each potential that always works. This means we must properly resolve the interfaces. Figure 10 suggests that we can take $c_2 = 5$ for both potentials, however this leads to slightly jagged interfaces. Therefore we will again favour robustness and choose $c_2 = 8$ i.e. $\pi\varepsilon = 8h$.

A.4 Choice of Iterative Parameter

The discrete iterative methods of Sect. 5.2 have values $\bar{\rho}$ independent of h such that for all $\rho > \bar{\rho}$ the iterates decrease in energy and converge in some sense. For example, a possible $\bar{\rho}$ for the iterative method of Example 5.6 applied to the problem in Sect. 7.3 is $\max\{\frac{\sigma_2}{\varepsilon}, C_s^2\} = 0.999$, where we use the Poincaré constant $1/\pi$. However in practice we observe that the iterates of this method decrease in energy and converge for $\rho \geq 0.833$. It is advantageous to take ρ small, as this results in fewer iterations and uses less total computational effort. So to maximise speed we experimentally determine a value of ρ which is as small as possible while still reliably giving a decrease in energy and convergence of iterates. This approach also works for the iterative method of Remark 5.5 for the double well potential, which lies outside of our framework. So for the speed comparison in Sect. 7.3 we use $\rho = 0.833$ for the smooth double well potential and $\rho = 0.588$ for the double obstacle potential. In the rest of the numerics, where speed is less of a concern, ρ is taken large (and larger than $\bar{\rho}$ if it is known) to ensure we get the expected behaviour of the iterative methods.

A.5 Choice of Stopping Criterion

We will never quite reach the steady state of the iterative method, so a decision needs to be made about when we are sufficiently close. For this purpose we use the stopping criterion introduced in Sect. 5.3 which terminates the algorithms when the L^2 norm of the difference between consecutive iterations is less than TOL.

Mostly we take TOL small so that we are effectively finding the exact steady state, but for the comparison of speed in Sect. 7.3 we need to avoid unnecessary iterations. Figure 11 suggests about 170 iterations will take us quite close to the steady state for the problem under consideration. This corresponds to taking TOL$= 3e-4$ for the smooth double well and TOL$= 3.5e-4$ for the double obstacle, and we use these values for the 'accurate recovery'.

In practice we just want a sufficiently accurate recovery as quickly as possible. Our feeling is that the binary function is usually sufficiently accurately recovered once the error is below 0.1. At this stage the correct number of bars have formed and the locations are probably known well enough (e.g. for a different algorithm to interpret the binary function as a barcode). We see in Fig. 11 that the smooth

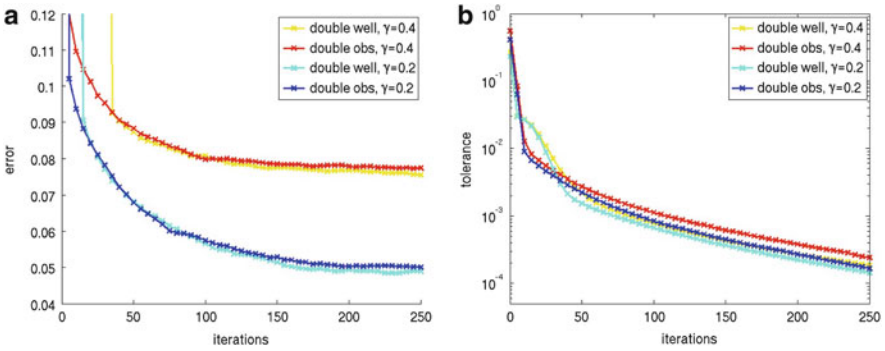


Fig. 11 The error (averaged over many realisations of the noise) after a given number of iterations for both potentials for the problem of Sect. 7.3. (**a**) Error. (**b**) TOL

double well potential achieves this in around 11 iteration, which corresponds to TOL$= 1.5e-2$. The double obstacle potential achieves this in around 9 iterations, which corresponds to TOL$= 4e-2$. We take these values for the 'rough recovery'.

Acknowledgements We are grateful to Carsten Gräser for sharing his Dune-Solvers code for the TNNMG method.

References

[1] J.W. Barrett, C.M. Elliott, *Finite element approximation of a free boundary problem arising in the theory of liquid drops and plasma physics.* Math. Modell. Numer. Anal. **25**, 213–252 (1991)

[2] P. Bastian, M. Blatt, A. Dedner, C. Engwer, R. Klöfkorn, M. Ohlberger, O. Sander, A generic grid interface for parallel and adaptive scientific computing. Part I: Abstract framework. Computing **82**, 103–119 (2008)

[3] P. Bastian, M. Blatt, A. Dedner, C. Engwer, R. Klöfkorn, R. Kornhuber, M. Ohlberger, O. Sander, A generic grid interface for parallel and adaptive scientific computing. Part II: Implementation and tests in DUNE. Computing **82**, 121–138 (2008)

[4] P. Bastian, M. Blatt, A. Dedner, C. Engwer, J. Fahlke, C. Gräser, R. Klöfkorn, M. Nolte, M. Ohlberger, O. Sander, DUNE web page, 2011. http://www.dune-project.org

[5] M. Blatt, P. Bastian, The iterative solver template library. In: Applied Parallel Computing. State of the Art in Scientific Computing, ed. by B. Kagström, E. Elmroth, J. Dongarra, J. Wasniewski, pp. 666–675, Vol. 4699 of Lecture Notes in Computer Science (Springer, Berlin-Heidelberg-New York, 2007)

[6] L. Blank, M. Butz, H. Garcke, Solving the Cahn-Hilliard variational inequality with a semi-smooth Newton method. ESAIM Contr Optim. Calculus Variat. **17**, 931–954 (2011)

[7] L. Blank, H. Garcke, L. Sarbu, V. Styles, Primal-dual active set methods for Allen-Cahn variational inequalities with nonlocal constraints. Numer. Meth. Part. Differ. Equat. **29**(3), 999–1030 (2013) http://onlinelibrary.wiley.com/doi/10.1002/num.v29.3/issuetoc

[8] J.F. Blowey, C.M. Elliott, The Cahn-Hilliard gradient theory for phase separation with non-smooth free energy. Part I: Mathematical analysis. Eur. J. Appl. Math. **2**, 233–279 (1991)

[9] J.F. Blowey, C.M. Elliott, The Cahn-Hilliard gradient theory for phase separation with non-smooth free energy. Part II: Numerical analysis. Eur. J. Appl. Math. **3**, 147–179 (1992)
[10] J.E. Blowey, C.M. Elliott, Curvature dependent phase boundary motion and parabolic double obstacle problems. In: Degenerate Diffusions, ed. by W.M. Ni, L.A. Peletier, L. Vazquez (Springer, Berlin-Heidelberg-New York, 1993), pp. 19–60
[11] A. Chambolle, P. Lions, Image recovery via total variation minimization and related problems. Numer. Math. **76**, 167–188 (1997)
[12] A. Chambolle, G. Del Maso, Discrete approximation of the Mumford-Shah functional in dimension two. ESAIM Math. Modell. Numer. Anal. **33**, 651–672 (1999)
[13] T.F. Chan, S. Esedoglu, Aspects of total variation regularized L1 function approximation. SIAM J. Appl. Math. **65**, 1817–1837 (2005)
[14] T. Chan, S. Esedoglu, M. Nikolova, Algorithms for finding global minimizers of image segmentation and denoising models. SIAM J. Appl. Math. **66**, 1632–1648 (2006)
[15] X. Chen, C.M. Elliott, Asymptotics for a parabolic double obstacle problem. Proc. R. Soc. Lond. A. **444**(1922), 429–445 (1994)
[16] R. Choksi, Y. Van Gennip, Deblurring of one dimensional bar codes via total variation energy minimisation. SIAM J. Imag. Sci. **3**, 735–764 (2010)
[17] R. Choksi, Y. Van Gennip, A. Oberman, Anisotropic total variation regularized L1-approximation and denoising/deblurring of 2D bar codes. Preprint, arXiv:1007.1035 (2010)
[18] A. Dedner, R. Klöfkorn, M. Nolte, M. Ohlberger, A generic interface for parallel and adaptive scientific computing: Abstraction principles and the DUNE-FEM module. Computing **90**, 165–196 (2010)
[19] A. Dedner, R. Klöfkorn, M. Nolte, M. Ohlberger, DUNE-FEM web page. 2011. http://dune.mathematik.uni-freiburg.de
[20] C.M. Elliott, A.M. Stuart, The global dynamics of discrete semilinear parabolic equations. SIAM J. Numer. Anal. **30**, 1622–1663 (1993)
[21] S. Esedoglu, Blind deconvolution of bar code signals. Inverse Probl. **20**, 121–135 (2004)
[22] D.J. Eyre, An unconditionally stable one-step scheme for gradient systems. Unpublished article (1998)
[23] C. Gräser, R. Kornhuber, Multigrid methods for obstacle problems. J. Comput. Math. **27**, 1–44 (2009)
[24] R. Glowinski, *Numerical Methods for Nonlinear Variational Problems* (Springer, Berlin-Heidelberg-New York, 1984)
[25] C. Gräser, Convex minimization and phase field models. PhD thesis, Freie Universität Berlin, 2011
[26] B. Hackl, Geometry variations, level set and phase-field methods for perimeter regularized geometric inverse problems. PhD thesis, Johannes Kepler Universität Linz, 2006
[27] J.K. Hale, *Asymptotic Behavior of Dissipative Systems* (American Mathematical Society, Providence, 1988)
[28] L. Modica, S. Mortola, Un esempio di Γ-convergenza. Bollettino dell'Unione Matematica Italiana **14-B**, 285–299 (1977)
[29] D.B. Mumford, J. Shah, Optimal approximations by piecewise smooth functions and associated variational problems. Comm. Pure Appl. Math. **42**, 577–685 (1989)
[30] N. Petra, G. Stadler, Model variational inverse problems governed by partial differential equations. Technical Report ADA555315, University of Texas at Austin, Institute for Computational Engineering and Sciences, 2011
[31] L.I. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms. Phys. D **60**, 259–268 (1992)
[32] L. Sarbu, Primal-dual active set methods for Allen-Cahn variational inequalities. PhD thesis, University of Sussex, 2010
[33] X. Tai, T.F. Chan, A survey on multiple level set methods with applications for identifying piecewise constant functions. Int. J. Numer. Anal. Modell. **1**, 25–47 (2004)
[34] X. Tai, H. Li, A piecewise constant level set method for elliptic inverse problems. Appl. Numer. Math. **57**, 686–696 (2007)

Recent Results in Shape Optimization and Optimal Control for PDEs

Maurizio Falcone and Marco Verani

Abstract In this paper we will present some recent advances in the numerical approximation of two classical problems: shape optimization and optimal control for evolutive partial differential equations. For shape optimization we present two novel techniques which have shown to be rather efficient on some applications. The first technique is based on multigrid methods whereas the second relies on an adaptive sequential quadratic programming. With respect to the optimal control of evolutive problems, the approximation is based on the coupling between a POD representation of the dynamical system and the classical Dynamic Programming approach. We look for an approximation of the value function characterized as the weak solution (in the viscosity sense) of the corresponding Hamilton-Jacobi equation. Several tests illustrate the main features of the above methods.

Keywords Dynamic programming • Evolutionary partial differential equations • Multigrid methods • Optimal control • Proper orthogonal decomposition • Shape optimization

Mathematics Subject Classification (2010). Primary 49L20; Secondary 35K51, 65M55

The authors acknowledge support by the European Science Foundation within the Programme 'Optimization with PDE Constraints' (OPTPDE).

M. Falcone (✉)
Dipartimento di Matematica, SAPIENZA - Università di Roma, P.za Aldo Moro, 2, 00185 Roma, Italy
e-mail: falcone@mat.uniroma1.it

M. Verani
MOX - Modelling and Scientific Computing - Dipartimento di Matematica, Politecnico di Milano, P.zza L. Da Vinci, 32 20133 Milano, Italy
e-mail: marco.verani@polimi.it

R. Hoppe (ed.), *Optimization with PDE Constraints*, Lecture Notes in Computational Science and Engineering 101,
DOI 10.1007/978-3-319-08025-3_3

1 Introduction

In this survey we will present some recent advances in the numerical approximation of two classical problems: shape optimization and optimal control for evolutive partial differential equations. These results have been achieved with the contributions of the researcher working in the teams at Milano Politecnico and Roma "La Sapienza" within the ESF OPTPDE project.

Shape optimization problems are ubiquitous in science, engineering and industrial applications. Indeed, starting with the foundation of PDE-based optimization [30], shape design has became one of the most frequent application in technologies and it is nowadays one main focus of aerodynamics simulation (see, e.g., [31, 42]).

A central role in the formulation and development of computational frameworks for shape optimization has been played by elliptic shape optimization problems [36] that correspond to cases of potential flow allowing simpler investigation. Nevertheless, these problems arise in many important applications as nozzle and airfoil design, and in the design of beams and plates. Along this development, one of the most remarkable advances in shape design has been to replace the approach of parametric optimization with the concept of continuous shape design (see, e.g., the books [17,22,28,31,36,38]). In fact, in the former approach the control variable (i.e., the shape) is restricted to belong to a finite dimensional space spanned by suitable basis functions, while in the latter case it is an element of an infinite-dimensional space. This second approach opens enormous perspective in the formulation of more accurate and sophisticated shape optimization problems.

The possibility of formulating the shape optimization problems at the infinite-dimensional level poses new challenges to the design and implementation of numerical optimization schemes that properly accommodate the infinite dimensionality of the control function. In particular, a successful and effective algorithm must allow the control function to be *adaptively* approximated and optimized to any desired degree of accuracy.

With respect to shape optimization, the purpose of this paper is twofold. We first formulate and analyze a multigrid shape optimization framework that extends principles and techniques of the multigrid strategy for PDE solvers and accommodates the infinite-dimensionality of the control variables; then we introduce an adaptive strategy able to automatically deal with the approximation of the optimal geometry combined with the approximation of the underlying PDE. As we said, our second problem will be the approximation of a finite horizon optimal control problem for an evolutive partial differential equation, e.g. the advection–diffusion equation. The basic ingredient of the method is the coupling between an adaptive reduced basis representation of the solution and a Dynamic Programming scheme for the evolutive Hamilton-Jacobi equation characterizing the value function. Since the theory of weak solutions for Hamilton-Jacobi equation is rather complete in any dimension, the method can in principle solve a rather general class of optimal

control problems. The approach described here is clearly different from the more classical approach based on the solution of the system of necessary conditions obtained via the Pontryagin maximum principle. The main advantage is that we naturally obtain optimal control in feedback form but the price we pay is related to the well known curse of dimensionality of Dynamic Programming. We try to circumvent this problem using new tools which have emerged in recent years to deal with optimal control problems in infinite dimension. In particular, we will use new techniques to reduce the number of dimensions in the description of the dynamical system or, more in general, of the solution of the problem that one is trying to optimize. These methods are generally called *reduced-order methods* and include for example the POD (Proper Orthogonal Decomposition) method and reduced basis approximation (see [35]). In some particular case, as for the heat equation, even 5 basis functions will suffice to have a rather accurate POD representation of the solution. Having this in mind, it is reasonable to start thinking to a different approach based on Dynamic Programming (DP) and Hamilton-Jacobi-Bellman equations (HJB). In this new approach we will first develop a reduced basis representation of the solution along a reference trajectory and then use this basis to set-up a control problem in the new space of coordinates. Then, the corresponding Hamilton-Jacobi equation will just need 3–5 variables to represent the state of the system. It is well known that the solution of HJB equation is not an easy task from the numerical point of view since viscosity solutions of the HJB equation are typically non regular (just Lipschitz continuous). Optimal control problems for ODEs were solved by Dynamic Programming, both analytically and numerically (see [4] for a general presentation of this theory). From the numerical point of view, this approach has been developed for many classical control problems obtaining convergence results and a-priori error estimates ([19,21] and the book [20]). We should mention that a first tentative in this direction has been made by Kunisch and co-authors in a series of papers [23,24,27] for diffusion dominated equations. In particular, in the paper by Kunisch, Volkwein and Xie [26] one can see a feedback control approach based on a coupling between POD basis approximation and HJB equations for the viscous Burgers equation.

Note that restricting the dimension to a rather low number of basis functions (typically 4) naturally affects the accuracy of the POD approximation. In fact, under this restriction, the POD method does not always have enough informations to follow correctly the solution of the evolutive problem. We circumvent this problem updating our POD basis during the evolution and splitting the problem into subproblems. Every sub-problem is set in an interval $I_j = [t_j, t_{j+1}]$ where we recompute the POD basis. Behind the adaptive method and the choice of the t_j there are two important *a-posteriori* estimators: the first is related to the computation of the POD basis function whereas the second takes into account the residual of the dynamics.

2 Two Approaches for Shape Optimization: Multigrid and Adaptivity

Shape optimization problems governed by partial differential equations (PDE) can be formulated as constrained minimization problems with respect to the shape of a domain Ω in $\mathbb{R}^d$. If $u = u(\Omega)$ is the solution of a PDE in Ω, the state equation,

$$\mathcal{A}u(\Omega) = f, \tag{2.1}$$

and $J(\Omega, u(\Omega))$ is a cost functional, then we consider the minimization problem

$$\Omega^* \in \mathcal{U}_{ad} : \qquad J(\Omega^*, u(\Omega^*)) = \inf_{\Omega \in \mathcal{U}_{ad}} J(\Omega, u(\Omega)), \tag{2.2}$$

where $\mathcal{U}_{ad}$ is a set of admissible domains in $\mathbb{R}^d$. This is a constrained minimization problem for J.

In this section we review two different shape optimization algorithms, namely the Multigrid Sequential Quadratic Programming (**MSQP**) presented in [3] and the Adaptive Sequential Quadratic Programming algorithm (**ASQP**) introduced in [32]. Such algorithms build a sequence of domains $\{\Omega^{(\ell)}\}_{\ell \geq 0}$ converging to a local minimizer of the shape optimization problem (2.1)–(2.2). To motivate and briefly describe the ideas underlying **MSQP** and **ASQP**, we need the concept of shape derivative $\nabla J(\Omega; \mathrm{w})$ of $J(\Omega)$ in the direction of a normal velocity w. By resorting to the celebrated Hadamard-Zolésio structure theorem (see, e.g., [17, 38]), it is well known that the shape derivative $\nabla J(\Omega; \mathrm{w})$ can be always written as

$$\nabla J(\Omega; \mathrm{w}) = \int_\Gamma G(\Omega)\mathrm{w}, \tag{2.3}$$

for a proper choice of the function $G(\Omega)$, named the *Riesz representation* of the shape derivative, that in general depends on the solution $u(\Omega)$ of the state equation (2.1). To review **MSQP** and **ASQP**, we preliminary introduce an infinite dimensional Sequential Quadratic Programming (∞-**ASQP**) algorithm. Let $\Omega^{(\ell)}$ be the current iterate and $\Omega^{(\ell+1)}$ be the new one. We let $\Gamma^{(\ell)} := \partial\Omega^{(\ell)}$ and $\mathbb{V}(\Gamma^{(\ell)})$ be a Hilbert space defined on $\Gamma^{(\ell)}$, with norm $\|\cdot\|_{\mathbb{V}(\Gamma^{(\ell)})}$. We further let $b_{\Gamma^{(\ell)}}(\cdot,\cdot) : \mathbb{V}(\Gamma^{(\ell)}) \times \mathbb{V}(\Gamma^{(\ell)}) \to \mathbb{R}$ be a continuous and coercive bilinear form with respect to the norm $\|\cdot\|_{\mathbb{V}(\Gamma^{(\ell)})}$, which gives rise to the elliptic self-adjoint operator $\mathcal{B}^{(\ell)}$ on $\Gamma^{(\ell)}$ defined by $\langle \mathcal{B}^{(\ell)}\mathrm{v}, \mathrm{w}\rangle_{\Gamma^{(\ell)}} = b_{\Gamma^{(\ell)}}(\mathrm{v}, \mathrm{w})$. We then consider the following *quadratic model* $Q^{(\ell)} : \mathbb{V}(\Gamma^{(\ell)}) \to \mathbb{R}$ of J at $\Omega^{(\ell)}$

$$Q^{(\ell)}(\mathrm{w}) := J(\Omega^{(\ell)}) + \nabla J(\Omega^{(\ell)}; \mathrm{w}) + \frac{1}{2}\langle \mathcal{B}^{(\ell)}\mathrm{w}, \mathrm{w}\rangle. \tag{2.4}$$

We denote by $\mathrm{v}^{(\ell)}$ the minimizer of $Q^{(\ell)}(\mathrm{w})$, namely $\mathrm{v}^{(\ell)}$ satisfies

$$\mathrm{v}^{(\ell)} \in \mathbb{V}(\Gamma^{(\ell)}): \qquad b_{\Gamma^{(\ell)}}(\mathrm{v}^{(\ell)}, \mathrm{w}) = -\langle G^{(\ell)}, \mathrm{w}\rangle_{\Gamma^{(\ell)}} \qquad \forall \mathrm{w} \in \mathbb{V}(\Gamma^{(\ell)}), \tag{2.5}$$

with $g^{(\ell)} := g(\Omega^{(\ell)})$. It is easy to check that $\mathrm{v}^{(\ell)}$ is the unique minimizer of $Q^{(\ell)}(\mathrm{w})$ and that the coercivity of the form $b_{\Gamma^{(\ell)}}(\cdot,\cdot)$ implies that $\mathrm{v}^{(\ell)}$ is an admissible descent direction; i.e. $\nabla J(\Omega^{(\ell)}; \mathrm{v}^{(\ell)}) < 0$.

Once $\mathrm{v}^{(\ell)}$ has been found, we need to determine a stepsize that is not too small and guarantees sufficient decrease of the functional J. To accomplish this goal we identify a range of admissible stepsizes by adapting the classical Armijo-Wolfe conditions in $\mathbb{R}^n$: given constants $0 < \alpha < \beta < 1$, we seek a stepsize $\mu \in \mathbb{R}^+$ satisfying

$$J(\Omega^{(\ell)} + \mu \mathbf{v}^{(\ell)}) \le J(\Omega^{(\ell)}) + \alpha\mu\, \nabla J(\Omega^{(\ell)}; \mathrm{v}^{(\ell)}), \tag{2.6}$$

$$\nabla J(\Omega^{(\ell)} + \mu \mathbf{v}^{(\ell)}; \mathrm{v}^{(\ell)}) \ge \beta\, \nabla J(\Omega^{(\ell)}; \mathrm{v}^{(\ell)}), \tag{2.7}$$

where $\Omega^{(\ell)} + \mu\mathbf{v}^{(\ell)} := \{\mathbf{y} \in \mathbb{R}^d : \mathbf{y} = \mathbf{x} + \mu\mathbf{v}^{(\ell)}(\mathbf{x}),\ \mathbf{x} \in \Omega^{(\ell)}\}$ is the updated domain and $\mathbf{v}^{(\ell)} = \mathrm{v}^{(\ell)}\boldsymbol{\nu}^{(\ell)}$ is a normal vector field.

We are now ready to introduce the infinite dimensional Sequential Quadratic Programming algorithm (∞-**ASQP**) for solving the constrained optimization problem (2.1)–(2.2):

∞-SQP Algorithm

Given the initial domain $\Omega^{(0)}$, set $\ell = 0$ and iterate:

(*a*) Compute $u^{(\ell)} = u(\Omega^{(\ell)})$ by solving (2.1)

(*b*) Compute the Riesz representation $G^{(\ell)} = G(\Omega^{(\ell)})$ of (2.3)

(*c*) Compute the search direction $\mathrm{v}^{(\ell)}$ by solving (2.5)

(*d*) Determine an admissible stepsize $\mu^{(\ell)}$ satisfying (2.6)-(2.7)

(*e*) Update: $\Omega^{(\ell+1)} = \Omega^{(\ell)} + \mu^{(\ell)}\mathbf{v}^{(\ell)}$; $\ell := \ell + 1$

It is important to note that, the ∞-SQP algorithm is not feasible as it stands, because it requires the exact computation of the following quantities at each iteration: the solution $u^{(\ell)}$ to the state equation (2.1); the solution $\mathrm{v}^{(\ell)}$ to the linear subproblem (2.5); the values of the functional J and of its derivative dJ in the line search routine. In the following, we review the Adaptive Sequential Quadratic Programming algorithm (**ASQP**) (see Sect. 2.1) and the Multigrid Sequential Quadratic Programming (**MSQP**) (see Sect. 2.2) as possible feasible variants of the ∞-SQP algorithm.

2.1 MSQP: A Multigrid Shape Optimization Algorithm

In this section we sketch the ideas underlying the construction of **MSQP**, we present the algorithm and we report some enlightening numerical results (see [3] for more details). Generally speaking, in **MSQP** the boundary of the domain, i.e., the control variable, is represented at various levels k of discretization and the resulting multigrid shape optimization scheme acts directly on the geometry of the domain combining a single-grid shape gradient optimizer with a coarse-grid correction (minimization) step, recursively within a hierarchy of levels.

As we focus on multigrid concepts, we need to define an iterative optimization process that can be applied at every level of discretization with the aim of improving the shape towards the optimum. In our case, this is a shape-gradient optimizer, denoted by $\mathtt{SQP}_k$, that acts similarly to a Jacobi smoother in a classical multigrid scheme. In practice, $\mathtt{SQP}_k$ is a feasible variant of the ∞-**SQP** stated at the level k of discretization (see e.g. [16, 18]).

In addition to the iterative scheme mentioned above, the formulation of a multigrid scheme requires to define a coarse-grid correction step that complements the action of the single-grid optimization procedure. To construct this step, suitable intergrid transfer operators are required together with the formulation of a coarse optimization problem that correctly approximates the fine-level shape optimization problem. On the other hand, to define the coarse shape optimization problem, the multigrid optimization framework introduced in [29, 33] is extended to the present case where the optimization variable is a geometrical object.

The approach presented in [3] is in contrast to previous attempts [7, 11–15] to define a consistent multigrid framework for shape optimization where the computational domain is discretized by finite elements and the control boundary is represented through parameterized shape functions. Therefore, within the hierarchy of levels defined by the multigrid strategy, the approach of **MSQP** allows to construct a coarse-grid correction step that can be understood from the geometrical [40] and optimization [5,6,29,33,41] point of views, whereas the idea in [7,12] of coarsening by taking a subset of shape parameters appears based on heuristic consideration.

To prepare the description of the **MSQP** algorithm, we first introduce the hierarchy of spaces $\mathcal{U}_{ad}^k$ of *discrete* admissible configurations. According to this, we denote by Ω_k an element of $\mathcal{U}_{ad}^k$ and by Γ_k the corresponding boundary (see Fig. 1 for an example where the deformable part of the domain is the graph of a function). Then we introduce the finite element space to approximate the solution of the PDE on Ω_k: let $\mathcal{T}_k(\Omega_k)$ be a conforming and shape-regular triangulation of Ω_k and $\mathbb{V}(\Omega_k)$ denote the associated space of finite elements. If we define the discrete reduced functional $\hat{J}_k(\Gamma_k)$ at k-level as

$$\hat{J}_k(\Gamma_k) := J(\Omega_k, y_k(\Omega_k)) \tag{2.8}$$

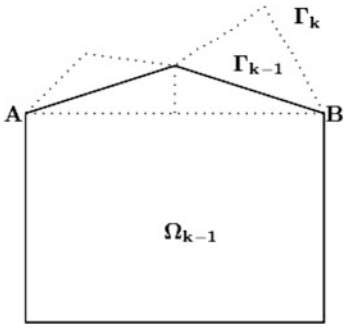


Fig. 1 An example of discrete control boundary represented at different levels of discretization: Γ_k (*dotted*) and $\Gamma_{k-1} = I_k^{k-1}\Gamma_k$ (*solid*)

then the reduced discrete shape optimization problem at level k reads as follows:

$$\Gamma_k^* = \text{argmin}_{\Gamma_k \in \mathcal{U}_{ad}^k} \hat{J}_k(\Gamma_k) \,. \tag{2.9}$$

Finally, for multigrid purpose, we need to define intergrid transfer operators acting on: functions in $\mathcal{U}_{ad}^k$; geometric boundaries Γ_k; and functions defined on geometric boundaries (i.e., shape gradients). To simplify the exposition (see [3] for more rigorous definitions) we will not use different symbols to distinguish among the above operators: we will always denote by I_k^{k-1} the restriction operators and by I_{k-1}^k the corresponding prolongation operators, the difference being clear from the context (see Fig. 1).

Finally, we introduce a hierarchy of nested shape optimization problems that will be solved at different levels of discretization. At k-level of discretization, we consider a function g_k to be defined iteratively in terms of g_{k+1}, where we set $g_K = 0$, being K the finest level of discretization (see below, Step 4 of the **MSQP** algorithm, for the precise recursive definition of g_k). The corresponding shape optimization problem at k-level reads as follows:

$$\min_{\Gamma_k \in \mathcal{U}_{ad}^k} F_k(\Gamma_k) := \hat{J}_k(\Gamma_k) - \int_{\Omega_k} g_k \, d\Omega \,. \tag{2.10}$$

It is clear that at the finest level K, the problem (2.10) corresponds to the original discrete shape optimization problem. Our aim is to formulate a multigrid shape optimization scheme for solving the minimization problem (2.10) for all levels k.

Let $\Gamma_k^{(0)}$ be the initial optimization boundary at level k and g_k be given. The following steps define one multigrid V-cycle that will be denoted by $\Gamma_k^{(new)} = \texttt{MSQP}(\Gamma_k^{(old)}, k, g_k)$.

MSQP Algorithm

If $k = 1$ (coarsest resolution) then the minimization problem (2.10) is solved exactly. Else if $k > 1$:

(1) Apply one-grid optimization

$$\Gamma_k^{(\ell+1)} = \mathtt{SQP}_k(\Gamma_k^{(\ell)}), \qquad \ell = 0, 1, \ldots, m_1 - 1.$$

(2) Compute the gradient residual

$$r_k = g_k - \nabla \hat{J}_k(\Gamma_k^{(m_1)}).$$

(3) Restrict the residual and the approximate solution to coarse levels

$$r_{k-1} = I_k^{k-1} r_k, \qquad \hat{\Gamma}_{k-1} = I_k^{k-1} \Gamma_k^{(m_1)}.$$

(4) Setup the coarse-grid problem

$$g_{k-1} = \nabla \hat{J}_{k-1}(\hat{\Gamma}_{k-1}) + r_{k-1}.$$

(5) Call the **MSQP** scheme to compute the coarse-grid minimizer for $\min F_{k-1}(\Gamma_{k-1})$: $\tilde{\Gamma}_{k-1} = \mathtt{MSQP}(\hat{\Gamma}_{k-1}, k-1, g_{k-1})$ such that

$$\tilde{\Gamma}_{k-1} \approx \operatorname{argmin} F_{k-1}(\Gamma_{k-1}).$$

(6) Construct the multigrid coarse-to-fine descent direction

$$\gamma_k = I_{k-1}^k \left(\tilde{\Gamma}_{k-1} - \hat{\Gamma}_{k-1} \right).$$

(7) Optimize along γ_k with α-linesearch

$$\Gamma_k^{m_1+1} = \Gamma_k^{(m_1)} + \alpha\, \gamma_k$$

(8) Apply one-grid optimization

$$\Gamma_k^{(\ell+1)} = \mathtt{SQP}_k(\Gamma_k^{(\ell)}), \qquad \ell = m_1 + 1, \ldots, m_1 + m_2.$$

(9) End.

In [3] it is proved that the multigrid coarse-to-fine direction γ_k built in Step 6 is indeed a descent direction. Moreover, it should be clear that the **MSQP** scheme given above will be applied iteratively, thus resulting in a sequence of V-cycles with finest level K and $g_K = 0$. Therefore, we also refer to the following algorithm as the **MSQP** scheme.

MSQP Algorithm

Input finest level K, initial Γ_K^0, $g_K = 0$, Tolerance ϵ, iteration counter $\ell = 0$, max number iterations ℓ_{max} and iterate:

(1) Compute $\Gamma_K^{\ell+1} = \mathrm{MSQP}(\Gamma_K^\ell, K, g_K)$

(2) Check convergence: if $\|\nabla \hat{J}(\Gamma_K^\ell)\| > \epsilon$ and $\ell < \ell_{max}$ then $\ell := \ell + 1$ and go to Step 1.

(3) End

In the following, we report some numerical results, originally presented in [3], where a shape optimization problem governed by an elliptic PDE has to be solved. In particular, let $y = y(\Omega)$ be the unique solution to the following elliptic partial differential equation

$$-\Delta y = f \qquad \text{in } \Omega \tag{2.11}$$

$$y = y_b \qquad \text{on } \partial\Omega\,, \tag{2.12}$$

where y_b is a given function defined in $\mathbb{R}^2$. Let r be a given function and $\lambda_1, \lambda_2, A, P > 0$ be given positive parameters. We consider the following cost functional

$$J(y, \Omega) := \int_\Omega r(y)\, d\Omega + \frac{\lambda_1}{2}\left(\int_{\partial\Omega} d\Gamma - P\right)^2 + \frac{\lambda_2}{2}\left(\int_\Omega d\Omega - A\right)^2, \tag{2.13}$$

which depends on the solution y of the problem (2.11)–(2.12), on the difference between the perimeter of $\partial\Omega$ and a given target value P and on the difference between the area of Ω and a given target value A.

The set $\mathcal{U}_{ad}^k$ of the admissible configurations is obtained by deforming the upper part of the domain, which is described by the graph of a piecewise linear function defined on a one dimensional grid. Increasing k amounts to decrease the mesh-size of the grid. As shown in Fig. 2, the **MSQP** algorithm, for different values of the finest level K of discretization, is able to efficiently approximate the optimal domain (in this case represented by the unitary square).

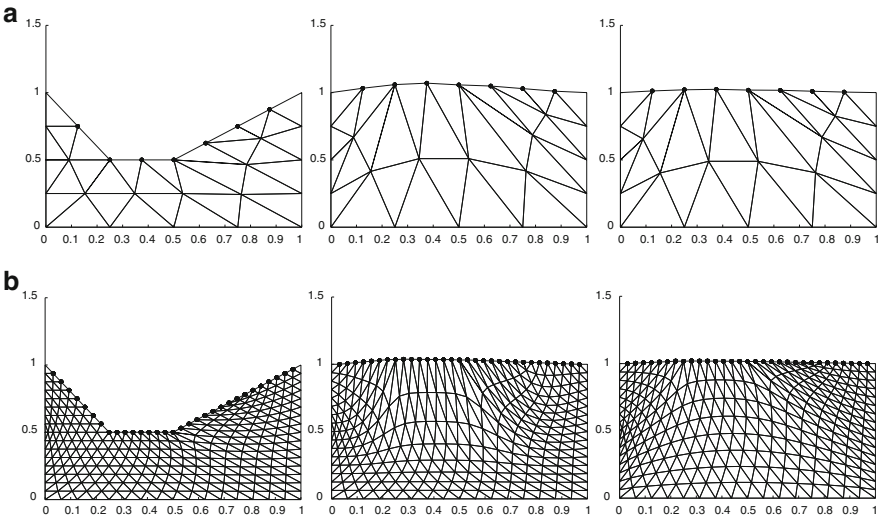


Fig. 2 Performance of the **MSQP** scheme for different values of the finest level K of discretization. The algorithm converges towards the optimal shape represented by the unitary square. (**a**) Finest level of discretization $K = 1$. Initial configuration (*left*), after 1 iteration (*middle*) and 12 iterations (*right*). (**b**) Finest level of discretization $K = 3$. Initial configuration (*left*), after 1 iteration (*middle*) and 9 iterations (*right*)

2.2 ASQP: An Adaptive Shape Optimization Algorithm

An alternative feasible variant of the ∞-**SQP** is represented by the Adaptive Sequential Quadratic Programming (**ASQP**) algorithm originally introduced in [32]. The **ASQP** scheme replaces all the non-computable operations of ∞-**SQP** (the solution to the state equation (2.1), the solution to the linear subproblem (2.5), the values of the functional J and of its derivative dJ in the line search routine) by adaptive finite dimensional approximations, whose accuracies are adjusted relative to the energy decrease for each iteration. It is worth noticing that the adaptive procedure driving **ASQP** has to deal with two distinct sources of error:

- *PDE Error*: this hinges on the approximation of (2.1), the values of the functional J and its derivative (2.3);
- *Geometric Error*: this relates to the approximation of (2.5) which yields the new domain.

Since it is wasteful to impose a PDE error finer than the expected geometric error, we have a natural mechanism to balance the computational effort.

In the following, we briefly describe the **ASQP** algorithm (see [32] for more details). Recall that $\ell \geq 1$ stands for the adaptive counter and $\Omega^{(\ell)}$ is the current domain produced by ASQP with deformable boundary $\Gamma^{(\ell)}$. Let $\mathbb{S}^{(\ell)} = \mathbb{S}_{\mathcal{T}^{(\ell)}}(\Omega^{(\ell)})$ and $\mathbb{V}^{(\ell)} = \mathbb{V}_{\mathcal{T}^{(\ell)}}(\Gamma^{(\ell)})$ be the finite element spaces on the bulk and boundary, which

are compatible and fully determined by one underlying mesh $\mathcal{T}^{(\ell)}$ of $\Omega^{(\ell)}$. We define **ASQP** as follows:

ADAPTIVE SEQUENTIAL QUADRATIC PROGRAMMING ALGORITHM (**ASQP**)

Given the initial domain $\Omega^{(0)}$, a triangulation $\mathcal{T}^{(0)}$ of $\Omega^{(0)}$, and the parameters $0 < \theta \leq \frac{1}{5}$, set $\gamma = \frac{1}{2} - \theta(1+\theta)$, $k = 0$, $\varepsilon^{(0)} = +\infty$, $\mu^{(0)} = 1$, repeat the following steps:

(1) $[\mathcal{T}^{(\ell)}, U^{(\ell)}, Z^{(\ell)}, J^{(\ell)}, G^{(\ell)}] = \texttt{APPROXJ}(\Omega^{(\ell)}, \mathcal{T}^{(\ell)}, \varepsilon^{(\ell)})$
(2) $[\mathbf{V}^{(\ell)}, \mathcal{T}^{(\ell)}] = \texttt{DIRECTION}(\Omega^{(\ell)}, \mathcal{T}^{(\ell)}, G^{(\ell)}, \theta)$
(3) $[\Omega^{(\ell+1)}, \mathcal{T}^{(\ell+1)}, \mu^{(\ell+1)}] = \texttt{LINESEARCH}(\Omega^{(\ell)}, \mathcal{T}^{(\ell)}, \mathbf{V}^{(\ell)}, J^{(\ell)}, \mu^{(\ell)})$
(4) $\varepsilon^{(\ell+1)} := \gamma\mu^{(\ell+1)}\|\mathbf{V}^{(\ell+1)}\|^2_{\Gamma^{(\ell)}}$; $\ell \leftarrow \ell + 1$.

In theory this algorithm is an infinite loop giving a more accurate approximation as the iterations progress, but in practice we implement a stopping criteria in LINESEARCH.

The modules **APPROXJ** and **DIRECTION** are driven by different adaptive strategies and corresponding different tolerances, say a PDE tolerance γ and a geometric tolerance θ. Their relative values allow for different distributions of the computational effort in dealing with the PDE and the geometry.

The routine DIRECTION enriches/coarsens the space $\mathbb{V}^{(\ell)}$ to control the quality of the descent direction, guaranteeing a geometric error proportional to $\mu^{(\ell)}\|V^{(\ell)}\|^2_{\Gamma^{(\ell)}}$, namely

$$\left|J(\Omega^{(\ell)} + \mu^{(\ell)}\mathbf{V}^{(\ell)}) - J(\Omega^{(\ell)} + \mu^{(\ell)}\mathbf{v}^{(\ell)})\right| \leq \delta\mu^{(\ell)}\|V^{(\ell)}\|^2_{\Gamma^{(\ell)}}, \tag{2.14}$$

with $\delta := \theta(1+\theta) \leq \frac{3}{2}\theta$. On the other hand, the module **APPROXJ** enriches/coarsens the space $\mathbb{S}^{(\ell)}$ to control the error in the approximate functional value $J^{(\ell)}(\Omega^{(\ell)} + \mu^{(\ell)}\mathbf{V}^{(\ell)})$ to the prescribed tolerance $\gamma\mu^{(\ell)}\|V^{(\ell)}\|^2_{\Gamma^{(\ell)}}$,

$$\left|J(\Omega^{(\ell)} + \mu^{(\ell)}\mathbf{V}^{(\ell)}) - J^{(\ell)}(\Omega^{(\ell)} + \mu^{(\ell)}\mathbf{V}^{(\ell)}))\right| \leq \gamma\mu^{(\ell)}\|V^{(\ell)}\|^2_{\Gamma^{(\ell)}}, \tag{2.15}$$

where $\gamma = \frac{1}{2} - \delta \geq \delta$ prevents excessive numerical resolution relative to the geometric one. This is achieved within the module APPROXJ via the *Dual Weighted Residual* method (DWR) [8], taylored to the approximation of the functional value J. The remaining modules perform the following tasks. The module **SOLVE** finds approximate solutions $U^{(\ell)} \in \mathbb{S}^{(\ell)}$ of (2.1) and $Z^{(\ell)} \in \mathbb{S}^{(\ell)}$ of an adjoint equation (necessary for the computation of the shape derivative) while **RIESZ** builds on $\mathbb{S}^{(\ell)}$ an approximation $G^{(\ell)}$ to the shape derivative. Finally, the module **LINESEARCH** enforces an inexact version of Wolfe's conditions.

We observe that the test (2.15) is not very demanding for DWR. So we expect coarse meshes at the beginning, and a combination of refinement and coarsening

later as DWR detects geometric singularities, such as corners, and sorts out whether they are genuine to the problem or just due to lack of numerical resolution. This aspect of **ASQP** is a novel paradigm in adaptivity and is detailed in [32].

In the following, we report some numerical examples originally presented in [32] to highlight the main features of the adaptive algorithm. In particular, we consider the drag reduction shape optimization problem governed by Stokes equation (see e.g. [37]). Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with its boundary subdivided into an *inflow* part Γ_{in}, an *outflow* part Γ_{out}, a part considered as *walls* Γ_w, and an obstacle Γ_s which is the deformable part to be optimized. The velocity $\mathbf{u} := \mathbf{u}(\Omega)$ and the pressure $p := p(\Omega)$ solve the following problem:

$$\begin{aligned} -\mathrm{div}(\mathbf{T}(\mathbf{u}, p)) &= 0 && \text{in } \Omega \\ \mathrm{div}\mathbf{u} &= 0 && \text{in } \Omega \\ \mathbf{u} &= \mathbf{u}_d && \text{on } \Gamma_{in} \cup \Gamma_s \cup \Gamma_w \\ \mathbf{T}(\mathbf{u}, p)\boldsymbol{\nu} &= 0 && \text{on } \Gamma_{out} \end{aligned} \tag{2.16}$$

where $\mathbf{T}(\mathbf{u}, p) := 2\mu\boldsymbol{\epsilon}(\mathbf{u}) - p\mathbf{I}$ is the Cauchy tensor with $\boldsymbol{\epsilon}(\mathbf{u}) = \frac{\nabla\mathbf{u}+\nabla\mathbf{u}^T}{2}$, $\mu > 0$ is the viscosity, and

$$\mathbf{u_d} = \begin{cases} \mathbf{v}_\infty & \text{on } \Gamma_{in} \\ \mathbf{0} & \text{on } \Gamma_w \cup \Gamma_s, \end{cases}$$

with $\mathbf{v}_\infty = V_\infty \hat{\mathbf{v}}_\infty$, $\hat{\mathbf{v}}_\infty$ being the unit vector pointing in the direction of the incoming flow and V_∞ a scalar function.

We let the cost functional measuring the obstacle *drag* be

$$J[\Omega, (\mathbf{u}, p)] := -\int_{\Gamma_s} (\mathbf{T}(\mathbf{u}, p)\boldsymbol{\nu}) \cdot \hat{\mathbf{v}}_\infty \, \mathrm{d}S, \tag{2.17}$$

where $(\mathbf{u}, p)$ solves (2.16). We would like to minimize the linear boundary functional J subject to the state constraint (2.16) among all admissible configurations with *fixed volume* that can be obtained by piecewise smooth perturbations of the obstacle boundary Γ_s.

In Fig. 3 we report the initial and final optimal configuration. As an effect of the DWR error indicator, the mesh refinement takes place mostly around the deformable shape, whereas in the rest of the domain Ω the mesh is rather coarse.

In Figs. 4, 5, 6, we show the efficacy of the adaptive **ASQP** method to sort out whether a geometric singularity is genuine to the problem or just due to the lack of numerical resolution. In the first case (genuine singularities) the method preserves the singularities and further refine them, whereas in the latter case (non-genuine singularities) the algorithm coarsens the (unnecessarily) over refined regions.

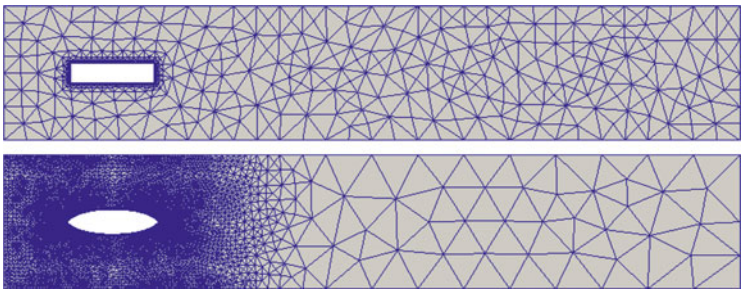


Fig. 3 Initial (*top*) and final (*bottom*) configuration. The **ASQP** algorithm obtains the optimal "rugby ball" shape [37]. The mesh refinement takes place mostly around the deformable shape, whereas in the rest of Ω the mesh is rather coarse: this is related to DWR mesh refinement (and coarsening)

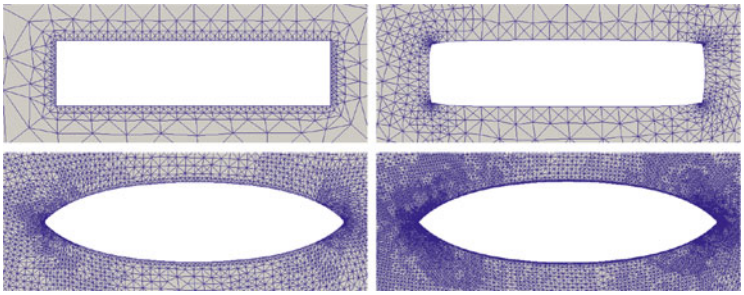


Fig. 4 Zoom of the evolution of the deformable shape. The initially refined corners (*top*) are subsequently smoothed out and coarsened (see Fig. 5). The new corners of the rugby ball, instead, are genuine singularities and are preserved and further refined by **ASQP** (*bottom*)

Fig. 5 Detection of genuine geometric singularities. Evolution of the initial upper-left corner of the deformable shape (see top of Figs. 3 and 4). The adaptive **ASQP** method is able to sort out whether geometric singularities are genuine to the problem or just due to lack of numerical resolution and to coarsen overrefined regions of the computational grid

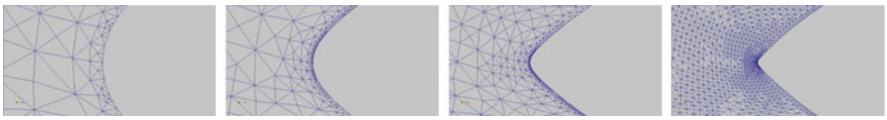


Fig. 6 Detection of genuine geometric singularities. Zoom on the evolution of the left-hand part of the deformable shape (see top of Fig. 3 and bottom of Fig. 4). The adaptive **ASQP** method is able to recognize the corner of the rugby ball as genuine singularity of the problem and to refine the mesh to improve both the PDE and the geometric approximation

3 Optimal Control for Evolutive PDEs

Let us now turn our attention to optimal control problems for evolutive partial differential equations. The classical approach is based on open-loop controls and on the Pontryagin maximum principle which leads to a backward-forward system characterizing the optimal couple state-control (see e.g. [34,39]). We have followed a different idea, trying to apply the Dynamic Programming approach. The results presented here are illustrated in [1,2].

3.1 The POD Approximation Method for Evolutive PDEs

We briefly describe some important features of the POD approximation, more details as well as precise results can be found in the notes by Volkwein [43]. Let us consider a matrix $Y \in \mathbb{R}^{m\times n}$, with rank $d \leq \min\{m,n\}$. We will call y_j the j-th column of the matrix Y. We are looking for an orthonormal basis $\{\psi_i\}_{i=1}^{\ell} \in \mathbb{R}^m$ with $\ell \leq n$ such that the minimum of the following functional is reached:

$$J(\psi_1,\dots,\psi_\ell) = \sum_{j=1}^{n} \left\| y_j - \sum_{i=1}^{\ell} \langle y_j, \psi_i\rangle \psi_i \right\|^2 . \tag{3.1}$$

The solution of this minimization problem is given in the following theorem

Theorem 3.1. *Let $Y = [y_1,\dots,y_n] \in \mathbb{R}^{m\times n}$ be a given matrix with rank $d \leq \min\{m,n\}$. Further, let $Y = \Psi\Sigma V^T$ be the Singular Value Decomposition (SVD) of Y, where $\Psi = [\psi_1,\dots,\psi_m] \in \mathbb{R}^{m\times m}$, $V = [v_1,\dots,v_n] \in \mathbb{R}^{n\times n}$ are orthogonal matrices and the matrix $\Sigma \in \mathbb{R}^{m\times n}$ is diagonal, $\Sigma = diag\{\sigma_1,\dots,\sigma_m\}$. Then, for any $\ell \in \{1,\dots,d\}$ the solution to (3.1) is given by the left singular vectors $\{\psi_i\}_{i=1}^{\ell}$, i.e, by the first ℓ columns of Ψ.*

The vectors $\{\psi_i\}_{i=1}^{\ell}$ will be indicated as the *POD basis of rank ℓ*.. This idea is really useful, in fact we get a representation of a solution for the original dynamics solving an equation of lower dimension. Whenever it is possible to compute a POD basis of rank ℓ, we get a problem of lower dimension ℓ which will be of manageable size provided ℓ is very small.

Let us consider the following ODEs system

$$\begin{cases} \dot{y}(s) = Ay(s) + f(s,y(s)), \ s \in (0,T] \\ y(0) = y_0 \end{cases} \tag{3.2}$$

where $y_0 \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times m}$ and $f : [0,T]\times\mathbb{R}^m \to \mathbb{R}^m$ is continuous and locally Lipschitz to ensure uniqueness.

The system (3.2) can be also interpreted as a semi-discrete problem, where the matrix A represents the discretization in space of an elliptic operator, e.g. the Laplace operator. To compute the POD basis functions, first of all we have to construct a time grid $0 \leq t_1 \leq \ldots \leq t_n = T$ and we assume to know the solution of (3.2) at given time t_j, $j = 1, \ldots, N$. We call *snapshots* the solution at those fixed times, they will be used to find a proper POD basis. For the moment, let us skip the problem of selecting the snapshots sequence to obtain an efficient POD basis since this is a rather difficult problem, we refer the interested reader to [25]). Given a snapshots sequence, Theorem 3.1 allows to compute our POD basis, namely, $\{\psi_j\}_{j=1}^{\ell}$.

Assume we can write the solution in reduced form as

$$y^{\ell}(s) = \sum_{j=1}^{\ell} y_j^{\ell}(s)\psi_j = \sum_{j=1}^{\ell} \langle y^{\ell}(s), \psi_j \rangle \psi_j, \qquad \forall s \in [0, T]$$

substituting this formula into (3.2) we obtain the equivalent dynamics in the reduced coordinate space

$$\begin{cases} \sum_{j=1}^{\ell} \dot{y}_j^{\ell}(s)\psi_j = \sum_{j=1}^{\ell} y_j^{\ell}(s) A\psi_j + f(s, y^{\ell}(s)), & s \in (0, T] \\ \sum_{j=1}^{\ell} y_j^{\ell}(0)\psi_j = y_0. \end{cases} \tag{3.3}$$

Our new problem (3.3) has $\ell \leq m$ unknown coefficient functions which are indicated by $y_j^{\ell}(s)$, $j = 1, \ldots, \ell$. The problem is now in low dimension, using a compact notation we get:

$$\begin{cases} \dot{y}^{\ell}(s) = A^{\ell} y^{\ell}(s) + F(s, y^{\ell}(s)) \\ y^{\ell}(0) = y_0^{\ell} \end{cases}$$

where

$$A^{\ell} \in \mathbb{R}^{\ell \times \ell} \qquad \text{with } (A^{\ell})_{ij} = \langle A\psi_i, \psi_j \rangle,$$

$$y^{\ell} = \begin{pmatrix} y_1^{\ell} \\ \vdots \\ y_{\ell}^{\ell} \end{pmatrix} : [0, T] \to \mathbb{R}^{\ell}$$

$F = (F_1, \ldots, F_{\ell})^T : [0, T] \times \mathbb{R}^{\ell} \to \mathbb{R}^{\ell}$,

$$F_i(s, y) = \left\langle f\left(s, \sum_{j=1}^{\ell} y_j \psi_j\right), \psi_i \right\rangle \quad \text{for } s \in [0, T] \;\; y = (y_1, \ldots y_{\ell}) \in \mathbb{R}^{\ell},$$

finally obtaining the representation of y_0 in $\mathbb{R}^\ell$

$$y_0^\ell = \begin{pmatrix} \langle y_0, \psi_1 \rangle \\ \vdots \\ \langle y_0, \psi_\ell \rangle \end{pmatrix} \in \mathbb{R}^\ell.$$

In order to apply the POD method to our optimal control problem, the number ℓ of POD basis functions plays a crucial role. In fact, we would like to keep ℓ as low as possible still capturing the behavior of the original dynamics. Then, the main question is: how can we measure the accuracy of our POD approximation? We need to define an accuracy parameter and a good choice is given by the following ratio

$$\mathcal{E}(\ell) = \frac{\sum_{i=1}^{\ell} \sigma_i}{\sum_{i=1}^{d} \sigma_i}, \tag{3.4}$$

where the σ_i are the singular value obtained by the SVD. Clearly, when $\mathcal{E}(\ell)$ is close to one this means that the approximation is rather accurate because it keeps the main features of the original dynamics. This is also strictly related to the truncation error due to the projection of y_j onto the space generated by the orthonormal basis $\{\psi\}_{i=1}^{\ell}$, in fact:

$$J(\psi_1, \ldots, \psi_\ell) = \sum_{j=1}^{n} \left\| y_j - \sum_{i=1}^{\ell} \langle y_j, \psi_i \rangle \psi_i \right\|^2 = \sum_{i=\ell+1}^{d} \sigma_i^2$$

3.2 An Optimal Control Problem via POD Approximation

Following [1] we present this approach for the finite horizon control problem. Consider the controlled system

$$\begin{cases} \dot{y}(s) = f(y(s), u(s), s), \quad s \in (t, T] \\ y(t) = x \in \mathbb{R}^n, \end{cases} \tag{3.5}$$

with $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, we will denote by $y : [t, T] \to \mathbb{R}^n$ its solution, by u the control $u : [t, T] \to \mathbb{R}^m$, and by

$$\mathcal{U} = \{u : [0, T] \to U\}$$

the set of admissible controls where $U \subset \mathbb{R}^m$ is a compact set. Whenever we want to emphasize the dependency of the solution on the control u we will write $y(t;u)$. Assume that there exists a unique solution trajectory for (3.5) provided the controls are measurable (a precise statement can be found in [4]). For the finite horizon optimal control problem the cost functional will be given by

$$\min_{u\in\mathcal{U}} J_{x,t}(u) := \int_t^T L(y(s,u),u(s),s)e^{-\lambda s}\,ds + g(y(T)) \tag{3.6}$$

where $L : \mathbb{R}^n \times \mathbb{R}^m \to \mathcal{R}$ is the running cost, (x,t) is the initial condition and $\lambda \geq 0$ is the discount factor.

The goal is to find a state-feedback control law $u(t) = \Phi(y(t),t)$, in terms of the state equation $y(t)$, where Φ is the feedback map. To derive optimality conditions we use the well-known *dynamic programming principle* due to Bellman (see [4]). We first define the value function:

$$v(x,t) := \inf_{u\in\mathcal{U}} J_{x,t}(u) \tag{3.7}$$

Proposition 3.2 (DPP). *For all $x \in \mathbb{R}^n$ and $t \leq \tau \leq T$ then:*

$$v(x,t) = \min_{u\in\mathcal{U}} \left\{ \int_t^\tau L(y(s),u(s),s)e^{-\lambda s}\,ds + v(y(\tau),T-\tau) \right\}. \tag{3.8}$$

Due to (3.8) we can derive the *Hamilton-Jacobi-Bellman* equations (HJB):

$$-\frac{\partial v}{\partial t}(y,t) = \min_{u\in U} \{L(y,u,t) + \nabla v(y,t)\cdot f(y,u,t)\}, \tag{3.9}$$

complemented by the terminal condition $v(x,T) = g(x)$. This is a nonlinear partial differential equation of the first order which is hard to solve analytically although a general theory of weak solutions is available [4]. Rather we can solve it numerically by means of a finite differences or semi-Lagrangian schemes (see the book [20] for a comprehensive analysis of approximation schemes for Hamilton-Jacobi equations). For a semi-Lagrangian discretization one starts by a discrete version of (HJB) by discretizing the underlined control problem and then project the semi-discrete scheme on a grid obtaining the fully discrete scheme

$$\begin{cases} v_i^{n+1} = \min\limits_{u\in U}[\Delta t\, L(x_i, n\Delta t, u) + I[v^n](x_i + \Delta t\, F(x_i,t_n,u))] \\ v_i^0 = g(x_i), \end{cases}$$

with $x_i = i\,\Delta x$, $t_n = n\Delta t$, $v_i^n := v(x_i,t_n)$ and $I[\cdot]$ is an interpolation operator which is necessary to compute the value of v^n at the point $x_i + \Delta t\, F(x_i,t_n,u)$ (in general, this point will not be a node of the grid). The interested reader will find in [21] a detailed presentation of the scheme and a priori error estimates for its numerical approximation.

It is also important to note that we need to compute the minimum in order to get the value v_i^{n+1}. Since, in general, v^n is not a smooth function, we compute the minimum by means of a minimization method which does not use derivatives (this can be done by the Brent algorithm as in [10]).

The main advantage of this approach is that it allows to compute the optimal feedback via the value function. However, there are two major difficulties: our weak solutions (in the viscosity sense) are in general non-smooth and the approximation in high dimension is not feasible due to the huge amount of data required. The request to solve an HJB in high dimension comes up naturally whenever we want to control evolutive PDEs. Just to give an idea, if we build a grid in $[0, 1] \times [0, 1]$ with a discrete step $\Delta x = 0.01$ we have 10^4 nodes: to solve an HJB in that dimension is simply impossible. The POD method allows us to obtain reduced models even for rather complicated dynamics opening the way to a feasible solution of the HJB equation.

Consider the following abstract problem:

$$\begin{cases} \dfrac{d}{ds}\langle y(s), \varphi\rangle_H + a(y(s), \varphi) = \langle B(u(s), \varphi\rangle_{V',V} \quad \forall \varphi \in V \\ y(t) = y_0 \quad \in H, \end{cases} \tag{3.10}$$

where $B : U \to V'$ is a linear and continuous operator. We assume that a space of admissible controls $\mathcal{U}_{ad}$ is given in such a way that for each $u \in \mathcal{U}_{ad}$ and $y_0 \in H$ there exists a unique solution y of (3.10). V and H are two Hilbert spaces, with $\langle \cdot, \cdot \rangle_H$ we denote the scalar product in H; $a : V \times V \to \mathcal{R}$: is symmetric coercive and bilinear. Then, we introduce the cost functional of the finite horizon problem

$$\mathcal{J}_{y_0,t}(u) := \int_t^T L(y(s), u(s), s)e^{-\lambda s}\, ds + g(y(T)),$$

where $L : V \times U \times [0, T] \to \mathcal{R}$. The optimal control problem is

$$\min_{u \in \mathcal{U}} \mathcal{J}_{y_0,t}(u) \tag{3.11}$$

$$\text{subject to the constraint:} \quad y \in W_{loc}(0, T; V) \times \mathcal{U} \text{ solves (3.10)}$$

with $W_{loc}(0, T) = \bigcap_{T>0} W(0, T)$, where $W(0, T)$ is the standard Sobolev space:

$$W(0, T) = \{\varphi \in L^2(0, T; V), \varphi_t \in L^2(0, T; V')\}.$$

The model reduction approach for an optimal control problem (3.11) is based on the Galerkin approximation of dynamic with some informations on the controlled dynamic (snapshots). To compute a POD solution for (3.11) we make the following ansatz

$$y^\ell(x, s) = \sum_{i=1}^{\ell} w_i(s)\psi_i(x), \tag{3.12}$$

where $\{\psi\}_{i=1}^{\ell}$ is the POD basis. The computation of the POD basis functions follows three easy steps:

1. Computation of the snapshots for the solution at given times, $y(s_j)$.
2. Collect the snapshots into a matrix Y and compute the singular value decomposition of $Y = U\Sigma V^T$.
3. Take the first ℓ columns of U, they will be the POD basis of rank ℓ.

Now let us introduce mass and stiffness matrix:

$$M = ((m_{ij})) \in \mathbb{R}^{\ell\times\ell} \text{ with } m_{ij} = \langle \psi_j, \psi_i \rangle_H,$$
$$S = ((s_{ij})) \in \mathbb{R}^{\ell\times\ell} \text{ with } m_{ij} = a(\psi_j, \psi_i),$$

and the control map $b : U \to \mathbb{R}^{\ell}$ is defined by:

$$u \to b(u) = (b(u)_i) \in \mathbb{R}^{\ell} \text{ with } b(u)_i = \langle Bu, \psi_i \rangle_H.$$

The coefficients of the initial condition $y^{\ell}(0) \in \mathbb{R}^{\ell}$ are determined by $w_i(0) = (w_0)_i = \langle y_0, \psi \rangle_X$, $1 \le i \le \ell$, and the solution of the reduced dynamic problem is denoted by $w^{\ell}(s) \in \mathbb{R}^{\ell}$. Then, the Galerkin approximation is given by

$$\min J^{\ell}_{w_0^{\ell},t}(u) \tag{3.13}$$

with $u \in \mathcal{U}$ and w solves the following equation:

$$\begin{cases} \dot{w}^{\ell}(s) = F(w^{\ell}(s), u(s), s) \quad s > 0, \\ w^{\ell}(0) = w_0^{\ell}. \end{cases} \tag{3.14}$$

The cost functional is defined as:

$$J^{\ell}_{w_0^{\ell},t}(u) = \int_0^T L(w^{\ell}(s), u(s), s)e^{-\lambda s}\, dt + g(w^{\ell}(T)),$$

with w^{ℓ} and y^{ℓ} linked to (3.12) and the nonlinear map $F : \mathbb{R}^{\ell} \times U \to \mathbb{R}^{\ell}$ is given by

$$F(w^{\ell}, u, s) = M^{-1}(-Sw^{\ell}(s) + b(u(s))).$$

The value function v^{ℓ}, defined for the initial state $w_0 \in \mathbb{R}^{\ell}$, reads as

$$v^{\ell}(w_0^{\ell}, t) = \inf_{u\in\mathcal{U}} J^{\ell}_{w_0^{\ell},t}(u)$$

and w^{ℓ} solves (3.13) with the control u and initial condition w_0.

To complete the scenario, let us explain how we have computed the intervals defining the domain where we are going to solve the HJB equation in reduced

coordinate. Clearly we need to restrict the computation to a bounded domain Υ_h in $\mathbb{R}^\ell$. We would like to find an invariant domain for the discrete dynamics, i.e. a domain Υ_h such that $y + \Delta t F(y,u) \in \Upsilon_h$ for each $y \in \Upsilon_h$ and $u \in \mathcal{U}$. We can choose $\Upsilon_h = [a_1, b_1] \times [a_2, b_2] \times \ldots [a_\ell, b_\ell]$ with $a_1 \geq a_2 \geq \ldots \geq a_\ell$. How should we compute the intervals $[a_i, b_i]$?

Ideally the intervals should be chosen so that the dynamics contains all the components of the controlled trajectory. Moreover, they should be encapsulated because we expect that their importance should decrease monotonically with their index and that our interval lengths decrease quickly.

Let us suppose to discretize the space control $U = \{u_1, \ldots, u_M\}$ where U is symmetric with respect to the origin, i.e. $\bar{u} \in U$ implies $-\bar{u} \in U$.

Hence, if $y^\ell(s) = \sum_{i=1}^{\ell} \langle y(s), \psi_i \rangle \psi_i = \sum_{i=1}^{\ell} w_i(s)\psi_i$, as a consequence, the coefficients $w_i(s) \in [a_i, b_i]$. We consider the trajectories solution $y(s, u_j)$ such that the control is constant $u(s) \equiv u_j$ for each t_j, $j = 1, \ldots, M$. Then, we have

$$y^\ell(s, u_j) = \sum_{i=1}^{\ell} \langle y(s, u_j), \psi_i \rangle \psi_i .$$

We write $y^\ell(s, u_j)$ to stress the dependence on the constant control u_j. Each trajectory $y^\ell(s, u_j)$ corresponds to a set of coefficients $w_i^{(j)}(t)$ for $i = 1, \ldots, \ell$, $j = 1, \ldots, M$. Every coefficient $w_i^{(j)}(s)$ belongs to an interval $[\underline{w}_i^{(j)}, \overline{w}_i^{(j)}]$ so, for $i = 1, \ldots, \ell,$, we define:

$$a_i \equiv \min\{\underline{w}_i^{(1)}, \ldots, \underline{w}_i^{(M)}\}$$
$$b_i \equiv \max\{\overline{w}_i^{(1)}, \ldots, \overline{w}_i^{(M)}\}.$$

Note that when we do not find an invariant domain to set up our computation we must introduce appropriate boundary conditions to manage the trajectories leaving the domain (see [19, 20] for more details on this technical problem).

3.3 Numerical Experiments

In this section we present some numerical tests for the controlled heat equation and for the advection-diffusion equation with a quadratic cost functional. Consider the following advection-diffusion equation:

$$\begin{cases} y_s(x,s) - \varepsilon y_{xx}(x,s) + c y_x(x,s) = u(s) \\ y(x,0) = y_0(x), \end{cases} \tag{3.15}$$

with $x \in [a, b]$, $s \in [0, T]$, $\varepsilon \in \mathbb{R}_+$ and $c \in \mathbb{R}$.

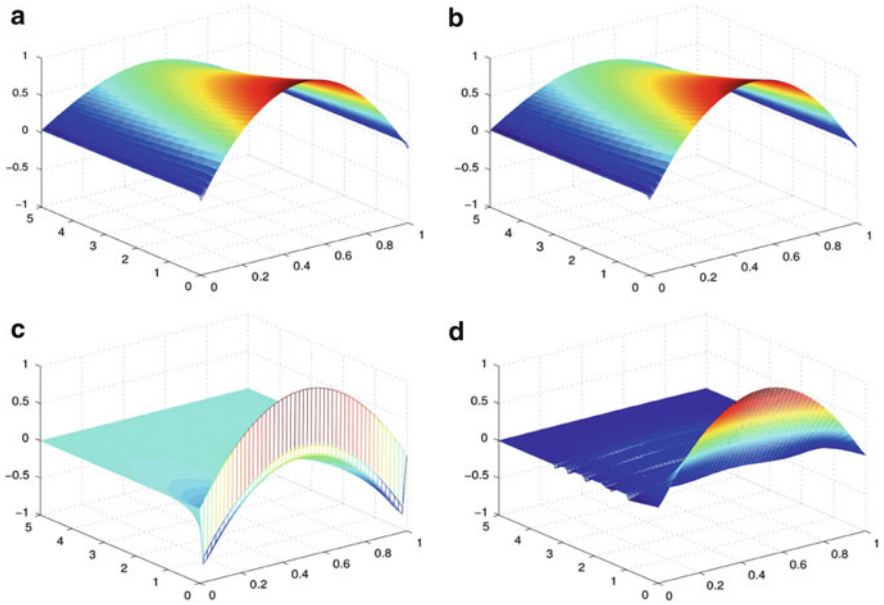


Fig. 7 Test 1: (**a**) Heat Equation without control; (**b**) Heat Equation for $\hat{u}(t) = 0$, 3 POD basis; (**c**) Controlled solution with LQR-MATLAB; (**d**) Approximate solution POD (3 basis functions) coupled with HJB

Note that changing the parameters c and ε we can obtain the heat equation ($c = 0$) and the advection equation ($\varepsilon = 0$).

The functional to be minimized is

$$J_{y_0,t}(u(\cdot)) = \int_0^T ||y(x,s) - \hat{y}(x,s)||^2 + R||u(s)||^2 \, ds, \tag{3.16}$$

i.e. we want to stay close to a reference trajectory $\hat{y}$ while minimizing the norm of u. Note that we dropped the discount factor setting $\lambda = 0$. Typically in our test problems $\hat{y}$ is obtained by applying a particular control $\hat{u}$ to the dynamics. The numerical simulations reported here have been made on a server SUPERMICRO 8045C-3RB with 2 cpu INTEL Xeon Quad-Core 2.4 Ghz and 32 GB RAM under SLURM (https://computing.llnl.gov/linux/slurm/).

Test 1: Heat Equation with Smooth Initial Data. We compute the snapshots with a centered/forward Euler scheme with space step $\Delta x = 0.02$, and time step $\Delta t = 0.012$, $\varepsilon = 1/60, c = 0, R = 0.01$ and $T = 5$. The initial condition is $y_0(x) = 5x - 5x^2$, and $\hat{y}(x,s) = 0$. In Fig. 7 we compare four different approximations concerning the heat equation: (a) is the solution for $\hat{u}(t) = 0$, (b) is its approximation via POD (non-adaptive), (c) is the direct LQR solution computed by MATLAB without POD and, finally, the approximate optimal solution obtained

coupling POD and HJB. The approximate value function is computed for $\Delta t = 0.1$ $\Delta x = 0.1$ whereas the optimal trajectory as been obtained with $\Delta t = 0.01$. Test 1, and even Test 2, have been solved in about half an hour of CPU time.

Note that in this example the approximate solution is rather accurate because the regularity of the solution is high due to the diffusion term. Since in the limit the solution tends to the average value, the choice of the snapshots will not affect too much the solution, i.e. even a rough choice of the snapshots will give us a good approximation. The difference between Fig. 7c and 7d is due to the fact that the control space is continuous for 7c and discrete for 7d.

Test 2: Heat Equation with No-Smooth Initial Data. In this section we change the initial condition with a function which is only Lipschitz continuous: $y_0(x) = 1 - |x|$. According to Test 1, we consider the same parameters. (see Fig. 8).
Riccati's equation has been solved by a MATLAB LQR routine. Thus, we have used the solution given by this routine as the correct solution in order to compare the errors in L^1 and L^2 norm between the reduced Riccati's equation and our approach based on the reduced HJB equation. Since we do not have any information, the snapshots are computed for $\hat{u} = 0$. This is only a guess, but in the parabolic case it fits well due to the diffusion term.

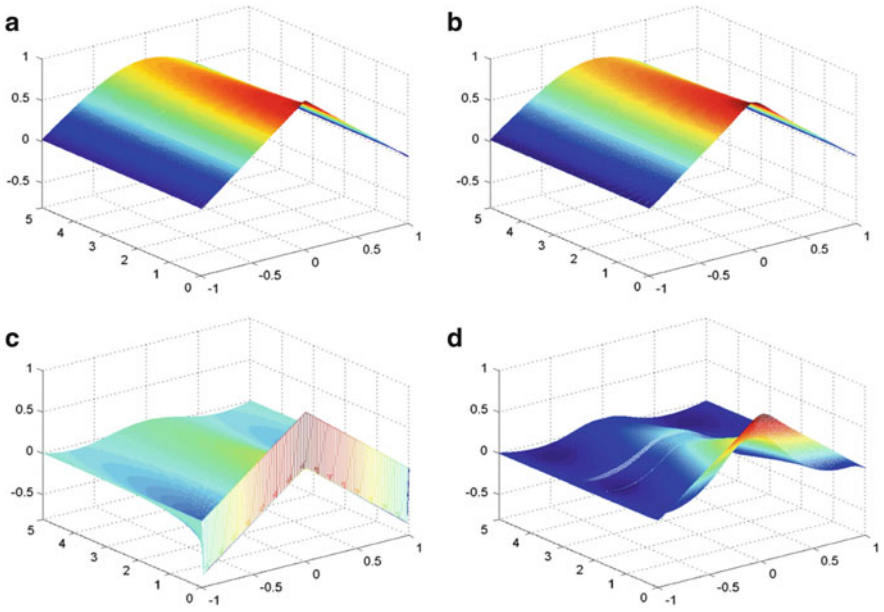


Fig. 8 Test 2: (**a**) exact solution for $\hat{u} = 0$; (**b**) Exact solution for $\hat{u} = 0$ POD (3 basis functions); (**c**) Approximate optimal solution for LQR-MATLAB; (**d**) Approximate solution POD (3 basis functions) coupled with HJB

Table 1 Test 2: L^1 and L^2 errors at time T for the optimal approximate solution

	L^1	L^2
$y^{LQR} - y^{POD+LQR}$	0.0221	0.0172
$y^{LQR} - y^{POD+HJB}$	0.0204	0.0171

As in Test 1, the choice of the snapshots does not affect strongly the approximation due to the asymptotic behavior of the solution. The presence of a Lipschitz continuous initial condition has almost no influence on the global error (see Table 1).

4 The Adaptive POD Approximation Method

We now present an adaptive method to compute POD basis. As we have seen in Sect. 3 we have a big constraint on the number of variables in the state space for numerical solution of an HJB.

For a parabolic equation, one can try to solve the problem with only three/four POD basis functions; they are enough to describe the solution in a rather accurate way. In fact the singular values decay pretty soon and it is rather easy to work with a really low-rank dimensional problem.

On the contrary, hyperbolic equations do not have this nice property and they will need more POD basis functions to get accurate results. Then, it is quite natural to split the problem into subproblems having different POD basis functions. The crucial point is to decide the splitting in order to have the same number of basis functions in each subdomain with a guaranteed accuracy in the approximation.

4.1 Numerical Experiments for the Adaptive POD Approximation Method

Let us first give an illustrative example for the parabolic case, considering a 1D advection-diffusion equation:

$$\begin{cases} y_s(x,s) - \varepsilon y_{xx}(x,s) + c y_x(x,s) = 0 \\ y(x,0) = y_0(x), \end{cases} \tag{4.17}$$

with $x \in [a,b], s \in [0,T], \varepsilon, c \in \mathbb{R}$.

We use a finite difference approximation for this equation based on an explicit Euler method in time combined with the standard centered approximation of the second order term and with an up-wind correction for the advection term. The snapshots will be taken from the sequence generated by the finite difference method.

The final time is $T = 5$, moreover $a = -1$, $b = 4$. The initial condition is $y_0(x) = 5x - 5x^2$, when $0 \leq x \leq 1$, 0 otherwise.

For $\varepsilon = 0.05$ and $c = 1$ with only 3 POD basis functions, the approximation fails (see Fig. 9). Note that in this case the advection is dominating the diffusion, a low number of POD basis functions will not suffice to get an accurate approximation (Fig. 9b). However, the adaptive method which only uses 3 POD basis functions will give accurate results (Fig. 9d).

The idea which is behind the adaptive method is rather simple and easy to implement. Instead of taking into account the whole interval $[0, T]$, we prefer to split it in sub-intervals

$$[0, T] = \cup_{k=0}^{K} [T_k, T_{k+1}]$$

where K is a-priori unknown, $T_0 = 0$, $T_K = T$ and $T_k = t_i$ for some i. In this way, choosing properly the length of the k-th interval $[T_k, T_{k+1}]$, we consider only the snapshots falling in that sub-interval, typically there will be at least three snapshots in every sub-interval. In this way we will have enough informations in every sub-interval and we can apply the standard routines (explained in Sect. 3) to get a "local" POD basis.

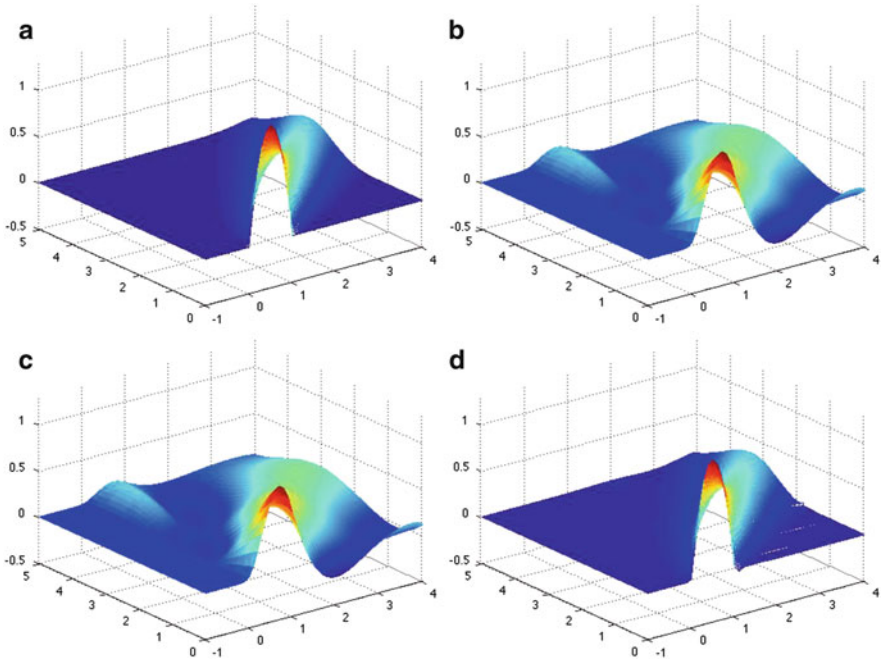


Fig. 9 Equation (4.17): (**a**) solved with finite difference; (**b**) POD-Galerkin approximation with 3 basis functions; (**c**) solved via POD-Galerkin approximation with 5 basis functions; (**d**) Adaptive POD 3 basis functions

Now let us explain how to divide our time interval $[0, T]$. We will choose a parameter to check the accuracy of the POD approximation and define a threshold. Above that threshold we loose in accuracy and we need to compute a new POD basis. A good parameter to check the accuracy is $\mathcal{E}(\ell)$ (see (3.4)), as it was suggested by several authors. The method to define the splitting of $[0, T]$ and the size of every sub-interval works as follows. We start computing the SVD of the matrix Y that gives us informations about our dynamics in the whole time interval. We check the accuracy at every t_i, $i = 1, \ldots N$, and if at t_k the indicator is above the tolerance we set $T_1 = t_k$ and we divide the interval in two parts, $[0, T_1)$ and $(T_1, T]$. Now we just consider the snapshots related the solution up to the time T_1. We iterate this idea until the indicator is below the threshold. When the first interval is found, we restart the procedure in the interval $[T_1, T]$ and we stop when we reach the final time T. Note that the extrema of every interval coincide by construction with one of our discrete times $t_i = i\Delta t$ so that the global solution is easily obtained linking all the sub-problems which always have a snapshot as initial condition. A low value for the threshold will also guarantee that we will not have big jumps passing from one sub-interval to the next. Once we know we got nice POD basis functions we compute the solution of the problem in each sub-intervals. Moreover, in each intervals $[T_k, T_{k+1}]$ we check the residual of the solution previously computed. If the residual is not below a given threshold, we split again the problem into two subproblems. This two subproblems need to update their own basis functions that will satisfy, of course, the error estimator applied to the POD method, since we are considering only a subset of the snapshots.

This idea can be applied also when we have a controlled dynamic (see [2]). First of all we have to decide how to collect the snapshots, since the control $u(t)$ is completely unknown. One can make a guess and use the dynamics and the functional corresponding to that guess, by these informations we can compute the POD basis. Once the POD basis is obtained we will get the optimal feedback law after having solved a reduced HJB equation as we already explained. Let us summarize the POD adaptive method in the following step-by-step presentation.

ALGORITHM

Start: Inizialization
Step 1: collect the snapshots in [0,T]
Step 2: divide $[0, T]$ according to $\mathcal{E}(\ell)$
For i=0 to N-1
Do
  Step 3: apply SVD to get the POD basis in each sub-interval $[t_i, t_{i+1}]$
  Step 4: discretize the space of controls
  Step 5: project the dynamics onto the (reduced) POD space
  Step 6: select the intervals for the POD reduced variables
  Step 7: solve the corresponding HJB in the reduced space

```
      for the interval [t_i, t_{i+1}]
  Step 8: go back to the original coordinate space
End
```

Test 3: Controlled Advection-Diffusion Equation. The advection-diffusion equation needs a different method. We can not use the same $\hat{y}$ we had in the parabolic case, mainly because in Riccati's equation the control is free and is not bounded, on the contrary when we solve an HJB we have to discretize the space of controls. We modified the problem in order to deal with bang-bang controls. We get $\hat{y}$ in (3.16) just plugging in the control $\hat{u} \equiv 0$. We have considered the control space corresponding only to three values in $[-1, 1]$, then $U = \{-1, 0, 1\}$. We first have tried to get a controlled solution, without any adaptive method and, as expected, we obtained a bad approximation (see Fig. 10). From Fig. 10 it is clear that POD with four basis functions is not able to catch the behavior of the dynamics, so we have applied our adaptive method.

We have consider: $T = 3, \Delta x = 0.1, \Delta t = 0.008$, $a = -1$, $b = 4$, $R = 0.01$. According to our algorithm, the time interval $[0, 3]$ was divided into $[0, 0.744] \cup [0.744, 1.496] \cup [1.496, 3]$. As we can see our last interval is bigger than the others, this is due to the diffusion term (see Fig. 11). The L^2-error is 0.0761, and the computation of the optimal solution via HJB has required about six hours of CPU time. In Fig. 4 we compare the exact solution with the numerical solution based on a POD representation. Note that, in this case, the choice of only 4 basis functions for the whole interval $[0, T]$ gives a very poor result due to the presence of the advection term. Looking at Fig. 5 one can see the improvement of our adaptive technique which takes always 4 basis functions in each sub-interval.

In order to check the quality of our approximation we have computed the numerical residual, defined as:

$$\mathcal{R}(y) = \|y_s(x, s) - \varepsilon y_{xx}(x, s) + c y_x(x, s) - u(s)\|.$$

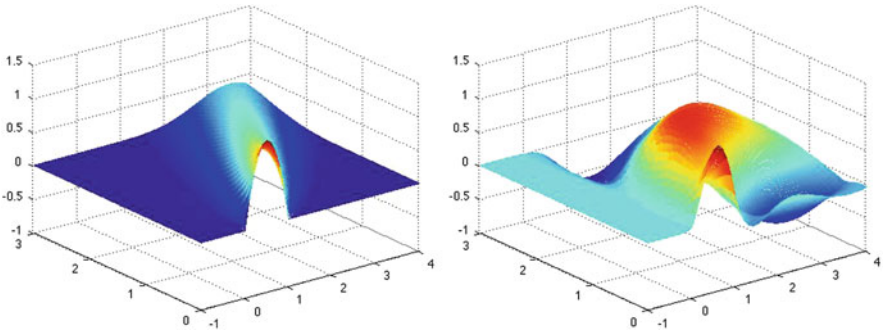


Fig. 10 Test 3: Solution $\hat{y}$ (*left*), approximate solution with POD (4 basis functions) (*right*)

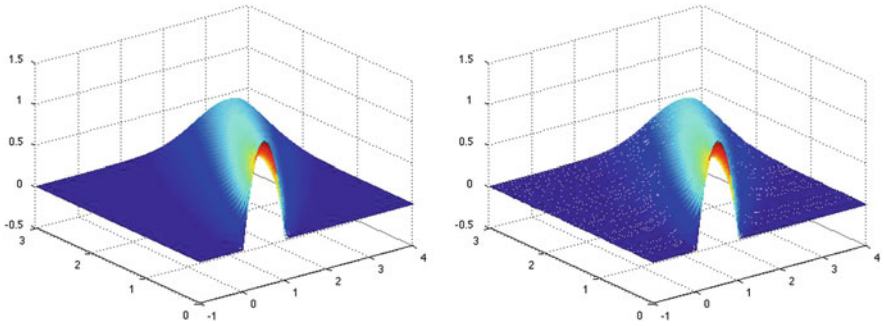


Fig. 11 Test 3: Solution for $\hat{u} \equiv 0$ (*left*), approximate optimal solution (*right*)

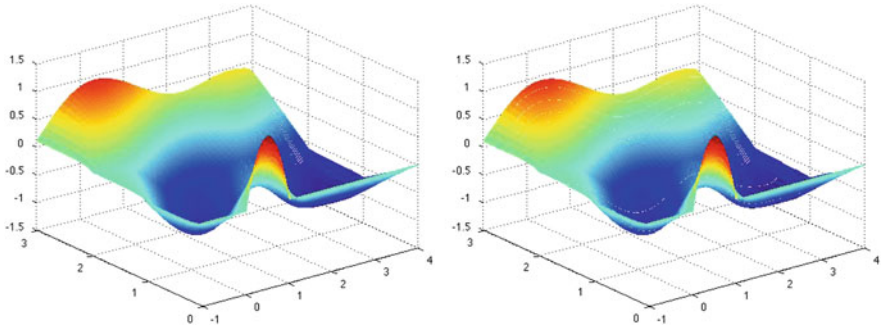


Fig. 12 Test 4: Solution for $\hat{u}$ (*left*), approximate optimal solution (*right*)

The residual for the solution of the control problem computed without our adaptive technique is 1.1, whereas the residual for the adaptive method is $2 * 10^{-2}$. As expected from the pictures, there is a big difference between these two value.

Test 4: Controlled Advection-Diffusion Equation. In this test we take a different $\hat{y}$, namely the solution of (3.15) corresponding to the control

$$\hat{u}(t) = \begin{cases} -1 & 0 \leq t < 1 \\ 0 & 1 \leq t < 2 \\ 1 & 2 \leq t \leq 3. \end{cases}$$

We want to emphasize we can obtain nice results when the space of controls has few element. The parameters were the same used in Test 3. The L^2-error is 0.09, and the time was the same we had in Test 3. In Fig. 12 we can see our approximation. In Fig. 6 one can see that the adaptive technique can also deal with discontinuous controls.

In this test, the residual for the solution of the control problem without our adaptive technique is 2, whereas the residual for the adaptive method is $3 * 10^{-2}$. Again, the residual shows the higher accuracy of the adaptive routine.

5 Conclusions

We presented some recent results concerning the numerical approximation of shape optimization problems and optimal control problems governed by evolutive partial differential equations. In particular, with respect to shape optimization problems, we introduced and discussed two novel techniques, namely a fully geometric multigrid approach and an adaptive sequential quadratic programming algorithm. Several numerical experiments assessed the efficacy of the proposed strategies. Concerning the optimal control of evolutive problems, we detailed how a reasonable coupling between POD and HJB equation can produce feedback controls for infinite dimensional problem. For advection dominated equations that simple idea has to be implemented in a clever way to be successful. It particular, the application of an adaptive technique is crucial to obtain accurate approximations with a low number of POD basis functions. This is still an essential requirement when dealing with the Dynamic Programming approach, which suffers from the curse-of-dimensionality although recent developments in the methods used for HJB equations will allow to increase this bound in the next future (for example by applying patchy techniques, see [9]).

Another important point is the discretization of the control space. In our examples, the number of optimal control is rather limited and this will be enough for problems which have a bang-bang structure for optimal controls. In general, we will need also an approximation of the control space via reduced basis methods. This point as well as a more detailed analysis of the procedure outlined in this paper will be addressed in our future work.

References

[1] A. Alla, M. Falcone, An adaptive POD approximation method for the control of advection-diffusion equations. In: *Control and Optimization with PDE Constraints*, ed. by K. Kunisch, K. Bredies, C. Clason, G. Von Winckel, International Series of Numerical Mathematics (Birkhäuser, Basel, 2013)

[2] A. Alla, M. Falcone, A time adaptive POD method for optimal control problems. In: Proceedings of the 1st IFAC CPDE Workshop, Paris, September 2013

[3] P.F. Antonietti, A. Borzì, M. Verani, Multigrid shape optimization governed by elliptic PDEs. SIAM J. Contr Optim. **51**(2), 1417–1440 (2013)

[4] M. Bardi, I. Capuzzo Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations* (Birkhäuser, Basel, 1997)

[5] A. Borzì, On the convergence of the MG/OPT method. PAMM **5**, 735–736 (2005)

[6] A. Borzì, V. Schulz, *Computational Optimization of Systems Governed by Partial Differential Equations* (SIAM, Philadelphia, 2011)
[7] F. Beux, A. Dervieux, A hierarchical approach for shape optimization. Eng. Comput. **11**, 25–48 (1994)
[8] R. Becker, R. Rannacher, An optimal control approach to a posteriori error estimation in finite element methods. Acta Numer. **10**, 1–102 (2001)
[9] S. Cacace, E. Cristiani, M. Falcone, A. Picarelli, A patchy dynamic programming scheme for a class of Hamilton-Jacobi-Bellman equations. SIAM J. Sci. Comp. **34**, 2625–2649 (2012)
[10] E. Carlini, M. Falcone, R. Ferretti, An efficient algorithm for Hamilton-Jacobi equations in high dimension. Comput. Visual. Sci. **7**, 15–29 (2004)
[11] J.-A. Désidéri, B.A. El Majd, A. Janka, Nested and self-adaptive Bézier parameterizations for shape optimization. J. Comput. Phys. **224**, 117–131 (2007)
[12] J.-A. Désidéri, Hierarchical optimum-shape algorithms using embedded Bezier parameterizations. In: *Numerical Methods for Scientific Computing, Variational Problems and Applications*, ed. by Y. Kuznetsov et al. (CIMNE, Barcelona, 2003)
[13] J.-A. Désidéri, A. Janka, Multilevel shape parameterization for aerodynamic optimization - application to drag and noise reduction of transonic/supersonic business jet. In: *European Congress on Computational Methods in Applied Sciences and Engineering*, ed. by E. Heikkola et al. (2003)
[14] J.-A. Désidéri, Two-level ideal algorithm for parametric shape optimization. In: *Advances in Numerical Mathematics*, ed. by W. Fitzgibbon, R. Hoppe, J. Periaux, O. Pironneau, Y. Vassilevski (Proceedings of International Conferences, Moscow, 2005)
[15] J.-A. Désidéri, A. Dervieux, Hierarchical methods for shape optimization in aerodynamics - I: multilevel algorithms for parametric shape optimization. In: *Introduction to Optimization and Multidisciplinary Design*, ed. by J. Periaux, H. Deconinck, Lecture Series 2006-3 (Von Karman Institute for Fluid Dynamics Publish., Belgium, 2006)
[16] F. de Gournay, Velocity extension for the level-set method and multiple eigenvalues in shape optimization. SIAM J. Contr Optim. **45**, 343–367 (2006)
[17] M.C. Delfour, J.-P. Zolesio, *Shapes and Geometries Analysis, Differential Calculus, and Optimization* (SIAM, Philadelphia, 2011)
[18] G. Dogan, P. Morin, R.H. Nochetto, M. Verani, Discrete gradient flows for shape optimization and applications. Comput. Meth. Appl. Mech. Eng. **196**, 3898–3914 (2007)
[19] M. Falcone, Numerical solution of dynamic programming equations. Appendix of the book by M. Bardi and I. Capuzzo Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations* (Birkhüser, Basel-Boston, 1997), pp. 471–504
[20] M. Falcone, R. Ferretti, *Semi-Lagrangian Approximation Schemes for Linear and Hamilton-Jacobi Equations* (SIAM, Philadelphia, 2014)
[21] M. Falcone, T. Giorgi, An approximation scheme for evolutive Hamilton-Jacobi equations. In: *Stochastic Analysis, Control, Optimization and Applications: A Volume in Honor of* ed. by W.H. Fleming, W.M. McEneaney, G. Yin, Q. Zhang (Birkhäuser, Basel, 1999), pp. 289–303
[22] A. Henrot, M. Pierre, *Variation et Optimisation de Formes* (Springer, Berlin-Heidelberg-New York, 2005)
[23] K. Kunisch, S. Volkwein, Control of burgers' equation by a reduced order approach using proper orthogonal decomposition. J. Optim. Theory Appl. **102**, 345–371 (1999)
[24] K. Kunisch, S. Volkwein, Galerkin proper orthogonal decomposition methods for parabolic problems. Numer. Math. **90**, 117–148 (2001)
[25] K. Kunisch, S. Volkwein, Optimal snapshot location for computing POD basis functions. ESAIM: M2AN **44**, 509–529 (2010)
[26] K. Kunisch, S. Volkwein, L. Xie, HJB-POD based feedback design for the optimal control of evolution problems. SIAM J. Appl. Dyn. Syst. **4**, 701–722 (2004)
[27] K. Kunisch, L. Xie, POD-based feedback control of Burgers equation by solving the evolutionary HJB equation. Comput. Math. Appl. **49**, 1113–1126 (2005)
[28] K. Ito, K. Kunisch, *Lagrange Multiplier Approach to Variational Problems and Applications* (SIAM, Philadelphia, 2008)

[29] R.M. Lewis, S.G. Nash, Model problems for the multigrid optimization of systems governed by differential equations. SIAM J. Sci. Comput. **26**, 1811–1837 (2005)
[30] J.L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations* (Springer, Berlin-Heidelberg-New York, 1971)
[31] B. Mohammadi, O. Pironneau, *Applied Shape Optimization for Fluids* (Oxford University Press, Oxford, 2001)
[32] P. Morin, R.H. Nochetto, S. Pauletti, M. Verani, Adaptive finite element method for shape optimization. ESAIM: Contr Optim. Calculus Variat. **18**, 1122–1149 (2012)
[33] S.G. Nash, A multigrid approach to discretized optimization problems. Optim. Meth. Software **14**, 99–116 (2000)
[34] P. Neittanmaki, D. Tiba, *Optimal Control of Nonlinear Parabolic Systems: Theory, Algorithms and Applications* (Marcel Dekker, New York, 1994)
[35] A.T. Patera, G. Rozza, *Reduced Basis Approximation and A Posteriori Error Estimation for Parametrized Partial Differential Equations* (MIT Pappalardo Graduate Monographs in Mechanical Engineering, Boston, 2006)
[36] O. Pironneau, *Optimal Shape Design for Elliptic Systems* (Springer, Berlin-Heidelberg-New York, 1984)
[37] O. Pironneau, On optimum profiles in Stokes flow. J. Fluid Mech. **59**, 117–128 (1973)
[38] J. Sokolowski, J.-P. Zolesio, *Introduction to Shape Optimization, Shape Sensitivity Analysis* (Springer, Berlin-Heidelberg-New York, 1992)
[39] F. Tröltzsch, *Optimal Control of Partial Differential Equations: Theory, Methods and Applications* (AMS, Providence, 2010)
[40] U. Trottenberg, C. Oosterlee, A. Schüller, *Multigrid* (Academic Press, London, 2001)
[41] M. Vallejos, A. Borzì, Multigrid optimization methods for linear and bilinear elliptic optimal control problems. Computing **82**, 31–52 (2008)
[42] J.C. Vassberg, A. Jameson, *Aerodynamic Shape Optimization, Part I & II* (Von Karman Institute, Brussels, 2006)
[43] S. Volkwein, Model reduction using proper orthogonal decomposition. Prepint, Fachbereich Mathematik, Universität Konstanz, 2011

Adaptive Finite Elements for Optimally Controlled Elliptic Variational Inequalities of Obstacle Type

A. Gaevskaya, M. Hintermüller, R.H.W. Hoppe, and C. Löbhard

Abstract We are concerned with the numerical solution of distributed optimal control problems for second order elliptic variational inequalities by adaptive finite element methods. Both the continuous problem as well as its finite element approximations represent subclasses of Mathematical Programs with Equilibrium Constraints (MPECs) for which the optimality conditions are stated by means of stationarity concepts in function space (Hintermüller and Kopacka, SIAM J. Optim. 20:868–902, 2009) and in a discrete, finite dimensional setting (Scheel and Scholtes, Math. Oper. Res. 25:1–22, 2000) such as (ε-almost, almost) C- and S-stationarity. With regard to adaptive mesh refinement, in contrast to the work in (Hintermüller, ESAIM Control Optim. Calc. Var., 2012, submitted) which adopts a goal oriented dual weighted approach, we consider standard residual-type a posteriori error estimators.

The first main result states that for a sequence of discrete C-stationary points there exists a subsequence converging to an almost C-stationary point, provided the associated sequence of nested finite element spaces is limit dense in its continuous counterpart. As the second main result, we prove the reliability and efficiency of the residual-type a posteriori error estimators. Particular emphasis is put on the approximation of the reliability and efficiency related consistency errors by

A. Gaevskaya
Institute of Mathematics, University of Augsburg Universitätsstr. 14, D-86159 Augsburg, Germany
e-mail: gaevskaya@math.uni-augsburg.de

M. Hintermüller • C. Löbhard
Department of Mathematics, Humboldt-Universität zu Berlin Unter den Linden 6, D-10099 Berlin, Germany
e-mail: hint@math.hu-berlin.de; loebhard@math.hu-berlin.de

R.H.W. Hoppe (✉)
Institute of Mathematics, University of Augsburg Universitätsstr. 14, D-86159 Augsburg, Germany

Department of Mathematics, University of Houston 659 P.G. Hoffman, Houston, TX 77204-3008, USA
e-mail: hoppe@math.uni-augsburg.de; rohop@math.uh.edu

R. Hoppe (ed.), *Optimization with PDE Constraints*, Lecture Notes in Computational Science and Engineering 101,
DOI 10.1007/978-3-319-08025-3_4

heuristically motivated computable quantities and on the approximation of the continuous active, strongly active, and inactive sets by their discrete counterparts.

A detailed documentation of numerical results for two representative test examples illustrates the performance of the adaptive approach.

Keywords A posteriori error analysis • Elliptic variational inequalities • Finite elements • Optimal control • Stationarity

Mathematics Subject Classification (2000). Primary 65K15; Secondary 49M99; 65K10; 90C56.

1 Introduction

This paper is devoted to the study of adaptive finite element methods for the approximation of optimally controlled elliptic variational inequalities of obstacle type. Such problems can be formulated as Mathematical Programs with Complementarity Constraints (MPCCs) representing a subclass of Mathematical Programs with Equilibrium Constraints (MPECs) which have been investigated both in function space [4, 18, 27, 30–34] as well as in finite dimensions [10, 26, 29, 36–38]. Due to the inherent non-convexity and non-differentiability, MPECs are not amenable to classical approaches from optimal control/optimization theory and thus require tools from non-smooth analysis such as generalized derivatives. In particular, this leads to optimality systems in terms of various stationarity concepts such as C(larke)-stationarity and S(trong)-stationarity (cf., e.g., [18] for MPECs in function space). For the spatial discretization of the problems we use continuous, piecewise linear finite elements with respect to an adaptively generated hierarchy of geometrically conforming simplicial triangulations of the computational domain. Although adaptive mesh refinement relying on various a posteriori error estimators has been extensively studied for elliptic variational inequalities (cf., e.g., [2, 6–8,22,24,35,42,43]) as well as for unconstrained and control and/or state constrained elliptic optimal control problems (cf., e.g., [5,12,14–17,19,20,23,40,45]), the only adaptive approach for optimally controlled elliptic variational inequalities we are aware of is the one in [21] based on goal oriented dual weighted residuals. Instead, here we study standard residual-type a posteriori error estimators in terms of element and edge residuals and prove both reliability and efficiency up to consistency errors and data oscillations.

The paper is organized as follows: After introducing basic notations and some preliminary results, in Sect. 2 we state the distributed optimal control problem for a second order elliptic variational inequality of obstacle type, specify the associated active and inactive sets including a possible set of biactivity in case of a lack of strict complementarity, and introduce the relevant stationarity concepts in function space. Section 3 is devoted to the finite element approximation of the problem under consideration giving rise to a discrete optimally controlled variational inequality,

the specification of the discrete active and inactive sets, and the discrete stationarity concepts. Particular emphasis is put on suitable extensions of the discrete Lagrange multipliers which will play a significant role both in the subsequent convergence analysis and in the a posteriori error analysis. In Sect. 4, we prove the first main result of this paper. Under the assumption that the sequence of nested finite element spaces is limit dense in the function space for the continuous state and adjoint state, we show that for a bounded sequence of discrete C-stationary points there exists a subsequence which converges to an almost C-stationary point (cf. Theorem 4.2). Section 5 is concerned with the a posteriori error analysis based on residual-type a posteriori error estimators.

As the second main result, we establish reliability and efficiency of the error estimator up to consistency errors due to a mismatch in complementarity and data oscillations (cf. Theorem 5.1 and Theorem 5.1). Since in the original formulation the consistency errors are not a posteriori, we provide heuristically motivated fully computable quantities in terms of approximations of the characteristic functions of the continuous active and inactive sets as well as of the continuous states and multipliers (cf. Sect. 5.4). The final Sect. 6 contains a documentation of numerical results for two representative test examples, one with strict complementarity and the other without. The numerical results exhibit experimental convergence rates that asymptotically approach the expected optimal convergence rates. Moreover, it is shown that at least some of the heuristically derived approximations of the consistency errors provide close upper bounds.

2 The Optimal Control Problem and Stationarity Concepts

2.1 Notations and Preliminaries

For a bounded Lipschitz domain $\Omega \subset \mathbb{R}^2$, we denote by $\mathcal{D}(\Omega)$ the space of infinitely often continuously differentiable functions with compact support in Ω, and we refer to $\mathcal{D}(\Omega)'$ as the dual space of distributions. Further, we adopt standard notation from Lebesgue and Sobolev space theory (cf., e.g., [1]). In particular, for $D \subseteq \Omega$, we denote by $L^2(D)$ the Hilbert space of square integrable functions on D with inner product $(\cdot,\cdot)_{0,D}$ and associated norm $\|\cdot\|_{0,D}$. $L^2(D)_+$ refers to the positive cone of $L^2(D)$ with respect to the partial order on $L^2(D)$, i.e., $L^2(D)_+ := \{v \in L^2(D) \mid v \geq 0 \text{ a.e. in } D\}$. For $k \in \mathbb{N}$, we denote by $H^k(D)$ the Sobolev space with inner product $(\cdot,\cdot)_{k,D}$, seminorm $|\cdot|_{k,D}$, and norm $\|\cdot\|_{k,D}$. We define $H_0^k(D)$ as the closure of $\mathcal{D}(D)$ in $H^k(D)$ and refer to $H^{-k}(D)$ as the dual space. In particular, we set $V := H_0^1(\Omega)$ so that $V^* = H^{-1}(\Omega)$, and we refer to $\langle\cdot,\cdot\rangle$ as the dual pairing between V^* and V. We define V_+ as the positive cone of V with respect to the partial ordering inherited from $L^2(\Omega)$, i.e., $V_+ := \{v \in V \mid v \geq 0 \text{ a.e. in } \Omega\}$ and we refer to V_+^* as the positive cone of V^*, i.e., $V_+^* := \{\lambda \in V^* \mid \langle\lambda, v\rangle \geq 0 \text{ for all } v \in V_+\}$.

As far as localizations of functionals $\lambda \in V^*$ are concerned, we note that for a distribution $T \in \mathcal{D}(\Omega)'$ and an open set $\omega \subseteq \Omega$ it is said that $T = 0$ on ω, if $T(v) = 0$ for all $v \in \mathcal{D}(\Omega)$ with $\mathrm{supp}(v) \subseteq \omega$ (cf., e.g., [41]). Further, denoting by $\mathcal{O}_T$ the maximal open set where $T = 0$, the support of T is defined by $\mathrm{supp}(T) := \Omega \setminus \mathcal{O}_T$. We set $V_\omega := \{v \in V \mid \mathrm{supp}(v) \subseteq \bar{\omega}\}$. Since a functional $\lambda \in V^*$ can be viewed as a distribution, we introduce the set

$$V_{\omega,0} := \{v \in V_\omega \mid v|_{\Omega\setminus\omega} = 0 \text{ a.e.}, v|_\omega \in H_0^1(\omega)\} \tag{2.1}$$

of test functions and say that $\lambda = 0$ on ω, if $\langle \lambda, v\rangle = 0$ for all $v \in V_{\omega,0}$ (for alternative definitions see [18]). Further, we say that $\lambda \geq 0$ ($\lambda \leq 0$) on ω, if $\langle \lambda, v\rangle \geq 0$ ($\langle \lambda, v\rangle \leq 0$) for all $v \in V_{\omega,0} \cap V_+$. The support of $\lambda \in V^*$ is defined by

$$\mathrm{supp}(\lambda) := \Omega \setminus \mathcal{O}_\lambda. \tag{2.2}$$

We note that $V_{\omega,0} \subseteq V_\omega$. If ω is Lipschitz, we have $V_{\omega,0} = V_\omega$ (cf., e.g., [27]).

In the sequel, we will need characterizations of functionals $\lambda \in V^*$ with restricted support. To this end, we first consider the question of extension by zero of $v|_\omega$, $v \in V$, for $\omega \subseteq \Omega$. If ω is Lipschitz, we denote by $\partial\omega^0(v)$ that part of the boundary $\partial\omega$ such that $v = 0$ a.e. on $\partial\omega^0(v)$ and $v \neq 0$ a.e. on $\partial\omega \setminus \partial\omega^0(v)$. Then, for $v \in V$ and an open Lipschitz domain $\omega \subseteq \Omega$ there exist an open Lipschitz set $\tilde{\omega}$ such that $\omega \subseteq \tilde{\omega} \subseteq \Omega$ and a function $v_\omega^{ext} \in V_{\tilde{\omega},0}$ with $v_\omega^{ext}|_\omega = v|_\omega$ a.e. in ω. If $\partial\omega^0(v) \neq \emptyset$, $\tilde{\omega}$ can be chosen so that $\partial\tilde{\omega} \cap \partial\omega = \partial\omega^0(v)$. If ω is non-Lipschitz, the previous property remains true, if ω is replaced by $\mathrm{Lip}(\omega)$ which is the minimal open Lipschitz set with $\omega \subseteq \mathrm{Lip}(\omega)$.

The following result allows to make use of the restricted support of functionals in V^* to describe their action on functions from V.

Proposition 2.1. *For $\lambda \in V^*$ set $\Lambda := int(supp(\lambda))$, if $supp(\lambda)$ is Lipschitz, and $\Lambda := Lip(int(supp(\lambda)))$, otherwise. For any $v \in V$ there exist an open Lipschitz set $\tilde{\Lambda}$ with $\Lambda \subseteq \tilde{\Lambda} \subseteq \Omega, \partial\tilde{\Lambda} \cap \partial\Lambda = \partial\Lambda^0(v)$ and a function $v_\Lambda^{ext} \in V_{\tilde{\Lambda},0}$ such that $v_\Lambda^{ext}|_\Lambda = v|_\Lambda$ a.e. in Λ and*

$$\langle \lambda, v\rangle = \langle \lambda, v_\Lambda^{ext}\rangle. \tag{2.3}$$

Proof. Since Λ is an open Lipschitz domain, there exist $\tilde{\Lambda}$ with $\Lambda \subseteq \tilde{\Lambda} \subseteq \Omega, \partial\tilde{\Lambda} \cap \partial\Lambda = \partial\Lambda^0(v)$ and a function $v_\Lambda^{ext} \in V_{\tilde{\Lambda},0}$ such that $v_\Lambda^{ext}|_\Lambda = v|_\Lambda$ a.e. in Λ. Hence, it suffices to prove (2.3). Let $\bar{v} \in V_{\Omega\setminus\Lambda,0}$ be defined according to

$$\bar{v} = \begin{cases} 0 \text{ in } \Lambda, \\ v - v_\Lambda^{ext} \text{ in } \mathrm{int}(\Omega \setminus \Lambda). \end{cases}$$

In view of the construction of Λ it holds $\mathrm{int}(\Omega \setminus \Lambda) \subseteq \mathcal{O}_\lambda$, where $\mathcal{O}_\lambda$ is the maximal open set where λ vanishes, and hence, $\langle \lambda, \bar{v} \rangle = 0$. It follows that $\langle \lambda, v \rangle = \langle \lambda, v_\Lambda^{ext} \rangle + \langle \lambda, \bar{v} \rangle = \langle \lambda, v_\Lambda^{ext} \rangle$. □

Remark 2.2. We note that $\langle \lambda, v \rangle = \langle \lambda, v|_{\mathrm{supp}(\lambda)} \rangle$ only if $v \in V_{\mathrm{supp}(\lambda),0}$. Otherwise, λ 'reaches' the values of v slightly outside of $\mathrm{int}(\mathrm{supp}(\lambda))$.

2.2 The Optimal Control Problem

Given a domain $\Omega \subset \mathbb{R}^2$ with boundary $\Gamma = \partial\Omega$, a bilinear form $a(\cdot,\cdot) : V \times V \to \mathbb{R}$, where $V := H_0^1(\Omega)$, a desired state y^d, a shift control u^d, a force density f, an upper obstacle ψ, and a regularization parameter α such that

$$\Omega \text{ is a bounded, polygonal Lipschitz domain,} \tag{2.4a}$$

$$a(\cdot,\cdot) : V \times V \to \mathbb{R} \text{ is symmetric, bounded and V-elliptic, i.e.,}$$

$$|a(y,v)| \le C\,\|y\|_{1,\Omega}\,\|v\|_{1,\Omega}, \quad \gamma\,\|y\|_{1,\Omega}^2 \le a(y,y),\ \gamma, C > 0, \tag{2.4b}$$

$$y^d \in L^2(\Omega), \quad u^d \in L^2(\Omega), \quad f \in L^2(\Omega), \tag{2.4c}$$

$$\psi \in V, \quad \alpha > 0, \tag{2.4d}$$

we consider the following distributed optimal control problem with a variational inequality constraint:

$$\text{Minimize} \qquad J(y,u) := \frac{1}{2}\,\|y - y^d\|_{0,\Omega}^2 + \frac{\alpha}{2}\,\|u - u^d\|_{0,\Omega}^2 \tag{2.5a}$$

$$\text{over} \qquad (y,u) \in V \times L^2(\Omega),$$

$$\text{subject to} \qquad a(y, y - v) \le (f + u, y - v)_{0,\Omega},\ v \in K, \tag{2.5b}$$

$$K := \{v \in V \mid v \le \psi \text{ a.e. in } \Omega\}.$$

Here, J is referred to as the objective functional, y and u stand for the state and the control, and K denotes the constraint set which makes (2.5b) to a variational inequality of obstacle type. We further denote by $A : V \to V^*$ the bounded linear operator associated with the bilinear form $a(\cdot,\cdot)$. Although the subsequent analysis can be carried out for a general second order elliptic differential operator in divergence form, in the sequel we will restrict ourselves to the case $A = -\Delta$.

The optimal control problem (2.5) can be equivalently written in the so-called control-reduced form by means of the control-to-state map $S : L^2(\Omega) \to V$ which assigns to a control $u \in L^2(\Omega)$ the unique solution of the variational inequality (2.5b):

$$\begin{aligned} &\text{Minimize} && J^{red}(u) := \frac{1}{2}\,\|Su - y^d\|^2_{0,\Omega} + \frac{\alpha}{2}\,\|u - u^d\|^2_{0,\Omega} && (2.6)\\ &\text{over} && u \in L^2(\Omega). \end{aligned}$$

The existence of minimizers for (2.5) is guaranteed by the following result:

Theorem 2.3. *Under the assumptions* (2.4) *on the data, the optimal control problem* (2.5) *admits an optimal solution.*

Proof. We refer to [4,31]. □

By introducing a slack variable $\sigma \in V^*$, the variational inequality constraint (2.5b) can be equivalently reformulated in terms of a complementarity system so that (2.5) reads:

$$\begin{aligned} &\text{Minimize} && J(y,u) := \frac{1}{2}\,\|y - y^d\|^2_{0,\Omega} + \frac{\alpha}{2}\,\|u - u^d\|^2_{0,\Omega} && (2.7a)\\ &\text{over} && (y,\sigma,u) \in V \times V^* \times L^2(\Omega), \\ &\text{subject to} && a(y,v) = (f+u,v)_{0,\Omega} - \langle \sigma, v\rangle,\ v \in V, && (2.7b)\\ & && \psi - y \in V_+,\ \sigma \in V_+^*,\ \langle \sigma, \psi - y\rangle = 0. \end{aligned}$$

The problem (2.7) is commonly referred to as a Mathematical Program with Complementarity Constraints (MPCC).

2.3 Continuous Active and Inactive Sets

For given $u \in L^2(\Omega)$, (2.5b) represents an obstacle problem which, under the assumptions (2.4), admits a unique solution $(y,\sigma) \in V \times V^*$ (cf., e.g., [25]). The complementary behavior of y and σ according to (2.7b) gives rise to the following definitions:

Definition 2.4. We define the active set $\mathcal{A}$ as the maximal open subset $D \subseteq \Omega$ such that $\psi - y = 0$ a.e. in D. We denote by $\mathcal{I} := \bigcup_{\varepsilon>0} B_\varepsilon(\psi - y)$ the inactive set, where $B_\varepsilon(\psi - y)$ is the maximal open set $D \subseteq \Omega$ such that $\psi - y \geq \varepsilon$ a.e. in D. Finally, $\mathcal{F}(y) := \Omega \setminus (\mathcal{A} \cup \mathcal{I})$ is said to be the free boundary with respect to y.

Obviously, the sets $\mathcal{A}, \mathcal{I}$, and $\mathcal{F}(y)$ provide a partition of Ω, i.e., it holds $\Omega = \mathcal{A} \cup \mathcal{I} \cup \mathcal{F}(y)$. An alternative partition can be achieved in terms of properties of the multiplier σ:

Definition 2.5. The zero set $\mathcal{Z}$ is defined as the maximal open set D such that $\langle \sigma, v \rangle = 0$ for all $v \in V_{D,0}$, whereas the set $\mathcal{C} := \text{int}(\text{supp}(\sigma))$ is referred to as the strongly active set (for the definitions of $V_{D,0}$ and $\text{supp}(\sigma)$ see (2.1) and (2.2) in Sect. 2.1). The set $\mathcal{F}(\sigma) := \Omega \setminus (\mathcal{Z} \cup \mathcal{C})$ is called the free boundary with respect to σ.

Remark 2.6. If in addition to the assumptions (2.4) on the data of the problem we suppose

$$\Omega \subset \mathbb{R}^2 \text{ is convex or of class } \mathcal{C}^{1,1}, \tag{2.8a}$$

$$\psi \in V \cap H^2(\Omega), \tag{2.8b}$$

the solution of the obstacle problem satisfies $(y, \sigma) \in V \cap H^2(\Omega) \times L^2(\Omega)$. In this regular case, we define the active and the inactive set according to $\mathcal{A}_{reg} := \text{int}(\{x \in \Omega \mid \psi(x) - y(x) = 0\}), \mathcal{I}_{reg} := \text{int}(\Omega \setminus \mathcal{A}_{reg})$. Moreover, the zero set $\mathcal{Z}_{reg}$ is the maximal open set $D \subseteq \Omega$ such that $\sigma = 0$ a.e. in D, and the strongly active set is given by $\mathcal{C}_{reg} := \text{int}(\Omega \setminus \mathcal{Z}_{reg})$.

The special case where $\psi - y$ and the slack variable σ are simultaneously zero in some subset of Ω is taken care of by the definition of the so-called biactive set:

Definition 2.7. The set $\mathcal{B} := \text{int}(\mathcal{A} \setminus \mathcal{C})$ is called the biactive set. If $\text{meas}(\mathcal{B}) = 0$, the solution of the obstacle problem is said to satisfy the strict complementarity condition. Otherwise, it is said that the solution exhibits a lack of strict complementarity.

The following results which were proven in [11] provide characterizations of the active set, the inactive set, the zero set, and of the slack variable σ. They all refer to the complementarity conditions (2.7b).

Proposition 2.8. *For any $v \in V_+$ let the zero set $\Omega^0(v)$ be the maximal open set $D \subseteq \Omega$ such that $v = 0$ a.e. in D and let $\Omega^+(v) := \bigcup_{\varepsilon>0} B_\varepsilon(v)$ be the positive set, where $B_\varepsilon(v)$ is the maximal open set $D \subseteq \Omega$ such that $v \geq \varepsilon$ a.e. in D. Then, it holds*

$$\mathcal{A} = \Omega^0(\psi - y), \quad \mathcal{I} = \Omega^+(\psi - y). \tag{2.9}$$

Moreover, for any $v \in V_+$ such that $\langle \sigma, v \rangle = 0$ it holds

$$\Omega^+(v) \subseteq \mathcal{Z}. \tag{2.10}$$

Corollary 2.9. *For any $v \in V$ such that $\langle \sigma, v^+ \rangle = 0$ and $\langle \sigma, v^- \rangle = 0$ it holds*

$$v = 0 \text{ in } \mathcal{C} \quad \text{and} \quad \langle \sigma, v \rangle = 0. \tag{2.11}$$

Proposition 2.10. *The slack variable σ satisfies*

$$\sigma = 0 \text{ in } \mathcal{I}, \text{ i.e., } \mathcal{C} \subseteq \mathcal{A}, \tag{2.12a}$$

$$\sigma = f + u - A\psi \text{ in } \mathcal{A}. \tag{2.12b}$$

Corollary 2.11. *A lack of strict complementarity of the solution of the obstacle problem occurs if and only if there exists a set $\mathcal{B} \subseteq \mathcal{A}$ such that $f + u - A\psi = 0$ in $\mathcal{B}$. Hence, there must hold $\langle A\psi, v\rangle = (f + u, v)_{0,\mathcal{B}}$, i.e., $A\psi|_{\mathcal{B}} \in L^2(\mathcal{B})$.*

2.4 Stationarity Concepts

In this subsection, we present various concepts of stationarity associated with the optimal control problem (2.5). We note that for MPCC in function space the concepts of C(larke)-stationarity and S(trong)-stationarity have been introduced in [18].

Definition 2.12. For $(y, \sigma, u) \in V \times V^* \times L^2(\Omega)$ assume that there exists a pair $(p, \mu) \in V \times V^*$ such that the following conditions hold true

$$a(y, v) = (f + u, v)_{0,\Omega} - \langle \sigma, v\rangle, \ v \in V, \tag{2.13a}$$

$$\psi - y \in V_+, \ \sigma \in V_+^*, \ \langle \sigma, \psi - y\rangle = 0, \tag{2.13b}$$

$$a(p, v) = (y^d - y, v)_{0,\Omega} - \langle \mu, v\rangle, \ v \in V, \tag{2.13c}$$

$$p = \alpha\,(u - u^d), \tag{2.13d}$$

$$p = 0 \text{ a.e. in } \mathcal{C}, \tag{2.13e}$$

$$\langle \mu, p\rangle \geq 0, \tag{2.13f}$$

$$\langle \mu, \psi - y\rangle = 0. \tag{2.13g}$$

A triple $(y, \sigma, u) \in V \times V^* \times L^2(\Omega)$ is called

(i) an ε-almost C-stationary point of (2.5), if (2.13a)–(2.13g) hold true and the pair $(p, \mu) \in V \times V^*$ satisfies:
For all $\varepsilon > 0$ there exists $U_\varepsilon \subseteq \mathcal{I}$ with $\text{meas}(\mathcal{I} \setminus U_\varepsilon) \leq \varepsilon$ such that

$$\langle \mu, v\rangle = 0, \ v \in V_{U_\varepsilon}, \tag{2.13h}$$

(ii) an almost C-stationary point of (2.5), if (2.13a)–(2.13g) hold true and the pair $(p, \mu) \in V \times V^*$ fulfills

$$\langle \mu, v\rangle = 0, \ v \in V_{\mathcal{I},0}, \tag{2.13i}$$

(iii) a C-stationary point of (2.5), if (2.13a)–(2.13g) hold true and the pair $(p, \mu) \in V \times V^*$ satisfies

$$\langle \mu, v \rangle = 0, \ v \in V_{\mathcal{I}}. \tag{2.13j}$$

Definition 2.13. Let $(y, \sigma, u) \in V \times V^* \times L^2(\Omega)$ be an ε-almost C-stationary point (almost C-stationary, C-stationary) point of (2.5). Then, the triple (y, σ, u) is said to be an ε-almost S-stationary (almost S-stationary, S-stationary) point of (2.5), if the pair $(p, \mu) \in V \times V^*$ additionally satisfies

$$\langle \mu, v \rangle \geq 0, \ v \in V_{\mathcal{B}} \cap V_+, \tag{2.14a}$$

$$p \geq 0 \text{ a.e. in } \mathcal{B}. \tag{2.14b}$$

Remark 2.14. In the Definitions 2.12 and 2.13, the function $p \in V$ is referred to as the adjoint state and Eq. (2.13c) is called the adjoint state equation. The functional $\mu \in V^*$ is said to be the Lagrange multiplier associated with the adjoint state equation.

Remark 2.15. In the previous Definitions 2.12 and 2.13, S-stationarity is the strongest and ε-almost C-stationarity is the weakest concept. The hierarchy of the above introduced stationarity concepts is displayed in the commuting diagram below:

$$\begin{array}{ccccc} \text{S-stationarity} & \Longrightarrow & \text{almost S-stationarity} & \Longrightarrow & \varepsilon\text{-almost S-stationarity} \\ \Downarrow & & \Downarrow & & \Downarrow \\ \text{C-stationarity} & \Longrightarrow & \text{almost C-stationarity} & \Longrightarrow & \varepsilon\text{-almost C-stationarity} \end{array}$$

The following result reveals local properties of almost C-stationary points with respect to the sets $\mathcal{C}, \mathcal{B}$, and $\mathcal{I}$ defined in Sect. 2.3.

Proposition 2.16. *Let $(y, \sigma, u) \in V \times V^* \times L^2(\Omega)$ be an almost C-stationary point of (2.5) and let $(p, \mu) \in V \times V^*$ be the associated adjoint state and Lagrange multiplier. Then, with regard to the strongly active set $\mathcal{C}$, the biactive set $\mathcal{B}$, and the inactive set $\mathcal{I}$ it holds*

	$\mathcal{C}$	$\mathcal{B}$	$\mathcal{I}$
y	$= \psi$ a.e.	$= \psi$ a.e.	–
p	$= 0$ a.e.	$= -\alpha\ (\Delta\psi + f + u^d)$ a.e.	–
u	$= u^d$ a.e.	$= -\Delta\psi - f$ a.e.	–
σ	$= f + u^d + \Delta\psi$	$= 0$	$= 0$
μ	$= y^d - \psi$	$= y^d - \psi + \alpha\ \Delta(\Delta\psi + f + u^d)$	$= 0$

Proof. In view of the definitions of the sets $\mathcal{A}, \mathcal{C}$, and $\mathcal{B}$, we obviously have $y = \psi$ a.e. in $\mathcal{A} = \mathcal{C} \cup \mathcal{B}$. Taking $V_{\mathcal{B},0} \subseteq V_{\mathcal{Z},0}$ and $V_{\mathcal{I},0} \subseteq V_{\mathcal{Z},0}$ into account, it holds

$$\langle \sigma, v \rangle = 0, \quad v \in V_{\mathcal{B},0}, \qquad \langle \sigma, v \rangle = 0, \quad v \in V_{\mathcal{I},0}.$$

Further, due to (2.13d) and (2.13e)

$$p = 0 \text{ a.e. in } \mathcal{C}, \qquad u = u^d \text{ a.e. in } \mathcal{C}.$$

Hence, (2.13c) implies

$$\langle \mu, v \rangle = (y^d - \psi, v)_{0,\mathcal{C}}, \quad v \in V_{\mathcal{C},0},$$

i.e., $\mu|_{\mathcal{C}} = y^d - \psi \in L^2(\mathcal{C})$. By (2.13a) it holds

$$\langle \sigma, v \rangle = (f + u^d, v)_{0,\mathcal{C}} - a(\psi, v), \quad v \in V_{\mathcal{C},0},$$

whence $\sigma = f + u^d + \Delta\psi$ a.e. in $\mathcal{C}$. Moreover, in $\mathcal{B}$ we have

$$(f + u, v)_{0,\mathcal{B}} = (\nabla\psi, \nabla v)_{0,\mathcal{B}}, \quad v \in V_{\mathcal{B},0}.$$

Consequently, the weak divergence of $\nabla\psi$ in $\mathcal{B}$ exists and equals $-(f + u)|_{\mathcal{B}} \in L^2(\mathcal{B})$. It follows that $-\Delta\psi = f + u$ a.e. in $\mathcal{B}$. Hence,

$$u = -\Delta\psi - f \text{ a.e. in } \mathcal{B},$$

and, due to (2.13d)

$$p = -\alpha\,(\Delta\psi + f + u^d) \text{ a.e. in } \mathcal{B}.$$

The previous equation gives rise to $\Delta\psi + f + u^d \in H^1(\mathcal{B})$. Hence, (2.13c) implies

$$\langle \mu, v \rangle = (y^d - \psi, v)_{0,\mathcal{B}} + \alpha\, a(\Delta\psi + f + u^d, v), \quad v \in V_{\mathcal{B},0}.$$

□

Stationarity in the Regular Case. If in addition to the assumptions (2.4) on the data of the problem we suppose

$$\Omega \text{ is either convex and polygonal or of class } C^{1,1}, \tag{2.15a}$$

$$\psi \in V \cap H^2(\Omega), \tag{2.15b}$$

for fixed $u \in L^2(\Omega)$ the solution (y, σ) of the obstacle problem belongs to $V \cap H^2(\Omega) \times L^2(\Omega)$. In this regular case, the optimal control problem (2.5) can be

rewritten according to:

$$\text{Minimize} \qquad J(y,u) := \frac{1}{2}\,\|y - y^d\|_{0,\Omega}^2 + \frac{\alpha}{2}\,\|u - u^d\|_{0,\Omega}^2 \tag{2.16a}$$

$$\text{over} \qquad (y,\sigma,u) \in V \times L^2(\Omega) \times L^2(\Omega),$$

$$\text{subject to} \qquad a(y, y - v) = (f + u - \sigma, v)_{0,\Omega},\ v \in V, \tag{2.16b}$$

$$\psi - y \geq 0 \text{ a.e. in } \Omega,\ \sigma \geq 0 \text{ a.e. in } \Omega,\ (\sigma, \psi - y)_{0,\Omega} = 0.$$

The stationarity concepts can be formulated as in Definitions 2.12 and 2.13.

3 Finite Element Approximation

For a null sequence $\mathcal{H}$ of positive real numbers we assume $\{\mathcal{T}_h(\Omega)\}_{h\in\mathcal{H}}$ to be a shape regular family of geometrically conforming simplicial triangulations of the computational domain Ω. For $D \subset \bar{\Omega}$, we denote by $\mathcal{N}_h(D)$, $\mathcal{E}_h(D)$, and $\mathcal{T}_h(D)$ the sets of nodal points, edges, and triangles of $\mathcal{T}_h(\Omega)$ in D. For $T \in \mathcal{T}_h(\Omega)$, we refer to h_T and $|T|$ as the diameter and the area of T, whereas for $E \in \mathcal{E}_h(\bar{\Omega})$ we denote by h_E the length of the edge E. We further introduce the following patches of triangles of $\mathcal{T}_h(\Omega)$:

$$\omega_a := \bigcup\{T \in \mathcal{T}_h(\Omega) \mid a \in N_h(T)\}, \tag{3.1a}$$

$$\omega_E := \bigcup\{T \in \mathcal{T}_h(\Omega) \mid E \in \mathcal{E}_h(T)\}, \tag{3.1b}$$

$$\omega_T := \bigcup\{T' \in \mathcal{T}_h(\Omega) \mid \mathcal{N}_h(T') \cap \mathcal{N}_h(T) \neq \emptyset\}, \tag{3.1c}$$

and the following set of edges of $\mathcal{E}_h(\Omega)$:

$$\mathcal{E}_h^a := \bigcup\{E \in \mathcal{E}_h(\Omega) \mid a \in \mathcal{N}_h(E)\}. \tag{3.2}$$

Moreover, for $T \in \mathcal{T}_h(\Omega)$ we refer to $P_k(T), k \in \mathbb{N}_0$, as the linear space of polynomials of degree $\leq k$ on T, and we define

$$S_h^{(1)} := \{v_h \in C(\bar{\Omega}) \mid v_h|_T \in P_1(T),\ T \in \mathcal{T}_h(\Omega)\} \tag{3.3}$$

as the finite element space of continuous piecewise linear functions. We set

$$V_h := \{v_h \in S_h^{(1)} \mid v_h|_\Gamma = 0\} \tag{3.4}$$

and denote by $\varphi_h^{(a)}$ the nodal basis function associated with $a \in \mathcal{N}_h(\Omega)$ such that $V_h = \text{span}(\{\varphi_h^{(a)} \mid a \in \mathcal{N}_h(\Omega)\})$ with $\dim V_h = N_h := \text{card}(\mathcal{N}_h(\Omega))$. As the dual space of V_h we consider linear combinations of the Dirac delta functionals δ_a associated with $a \in \mathcal{N}_h(\Omega)$, i.e.,

$$M_h := \{\lambda_h \in \mathcal{M}(\bar{\Omega}) \mid \lambda_h = \sum_{a \in \mathcal{N}_h(\Omega)} \lambda_h(a)\, \delta_a,\ \lambda_h(a) \in \mathbb{R}\}. \tag{3.5}$$

Here, $\mathcal{M}(\bar{\Omega})$ stands for the space of regular Borel measures.

3.1 The Discrete Optimal Control Problem

For the finite element approximation of the optimal control problem (2.5) we denote by $\psi_h \in V_h$ and $u_h^d \in S_h^{(1)}$ the interpolants of $\psi \in V$ and $u^d \in L^2(\Omega)$ in V_h and $S_h^{(1)}$ and refer to $y_h^d \in S_h^{(1)}$ and $f_h \in S_h^{(1)}$ as the L^2-projections of $y^d \in L^2(\Omega)$ and $f \in L^2(\Omega)$ onto $S_h^{(1)}$. Approximating the state $y \in V$ and the control $u \in L^2(\Omega)$ by finite element functions $y_h \in V_h$ and $u_h \in S_h^{(1)}$, the discrete optimal control problem is given as follows:

$$\begin{aligned}
&\text{Minimize} && J_h(y_h, u_h) := \frac{1}{2}\, \|y_h - y_h^d\|_{0,\Omega}^2 + \frac{\alpha}{2}\, \|u_h - u_h^d\|_{0,\Omega}^2 && \text{(3.6a)}\\
&\text{over} && (y_h, u_h) \in V_h \times S_h^{(1)}, &\\
&\text{subject to} && a(y_h, y_h - v_h) \le (f_h + u_h, y_h - v_h)_{0,\Omega},\ v_h \in K_h, && \text{(3.6b)}\\
& && K_h := \{v_h \in V_h \mid v_h \le \psi_h \text{ in } \Omega\}. &
\end{aligned}$$

We refer to J_h and K_h as the discrete objective functional and the discrete constraint set and to y_h and u_h as the discrete state and the discrete control.

Denoting by $S_h : S_h^{(1)} \to V_h$ the discrete control-to-state map which assigns to a control $u_h \in S_h^{(1)}$ the unique solution $y_h \in V_h$ of the discrete variational inequality (3.6b), the control-reduced form of (3.6) reads:

$$\begin{aligned}
&\text{Minimize} && J_h^{red}(u_h) := \frac{1}{2}\, \|S_h u_h - y_h^d\|_{0,\Omega}^2 + \frac{\alpha}{2}\, \|u_h - u_h^d\|_{0,\Omega}^2 && \text{(3.7)}\\
&\text{over} && u_h \in S_h^{(1)}. &
\end{aligned}$$

Theorem 3.1. *The discrete optimal control problem* (3.6) *admits an optimal solution* $(y_h, u_h) \in V_h \times S_h^{(1)}$.

Proof. The proof can be given in much the same way as that of Theorem 2.3. □

As in the continuous regime, by introducing a slack variable $\sigma_h \in M_h$, the discrete optimal control problem (3.6) can be equivalently reformulated as the discrete complementarity problem:

$$\text{Minimize} \qquad J_h(y_h,u_h) := \frac{1}{2}\,\|y_h - y_h^d\|_{0,\Omega}^2 + \frac{\alpha}{2}\,\|u_h - u_h^d\|_{0,\Omega}^2 \tag{3.8a}$$

$$\text{over} \qquad (y_h,\sigma_h,u_h) \in V_h \times M_h \times S_h^{(1)},$$

$$\text{subject to} \qquad a(y_h,v_h) = (f_h + u_h, v_h)_{0,\Omega} - \langle\langle \sigma_h, v_h\rangle\rangle,\ v_h \in V_h, \tag{3.8b}$$

$$y_h \in K_h,\ \sigma_h \in M_h \cap \mathcal{M}_+(\bar\Omega),\ \langle\langle \sigma_h, \psi_h - y_h\rangle\rangle = 0,$$

where $\langle\langle\cdot,\cdot\rangle\rangle$ refers to the dual pairing between $C(\bar\Omega)$ and $\mathcal{M}(\bar\Omega)$.

3.2 Discrete Active and Inactive Sets

For $v_h \in V_h$ we denote by

$$\mathcal{Z}_h(v_h) := \{a \in \mathcal{N}_h(\bar\Omega) \mid v_h(a) = 0\}, \quad \mathcal{C}_h(v_h) := \mathcal{N}_h(\bar\Omega) \setminus \mathcal{Z}_h(v_h) \tag{3.9}$$

the sets of zero and non-zero nodal points with respect to $v_h \in V_h$, and we partition the triangulation $\mathcal{T}_h(\Omega)$ into the sets of zero, non-zero, and mixed triangles with respect to $v_h \in V_h$ according to

$$\mathcal{T}_h(\Omega) = \mathcal{T}_h^z(v_h) \cup \mathcal{T}_h^c(v_h) \cup \mathcal{T}_h^m(v_h), \tag{3.10}$$

where

$$\mathcal{T}_h^z(v_h) := \{T \in \mathcal{T}_h(\Omega) \mid \mathcal{N}_h(T) \subset \mathcal{Z}_h(v_h)\}, \tag{3.11a}$$

$$\mathcal{T}_h^c(v_h) := \{T \in \mathcal{T}_h(\Omega) \mid \mathcal{N}_h(T) \subset \mathcal{C}_h(v_h)\}, \tag{3.11b}$$

$$\mathcal{T}_h^m(v_h) := \mathcal{T}_h(\Omega) \setminus (\mathcal{T}_h^z(v_h) \cup \mathcal{T}_h^c(v_h)). \tag{3.11c}$$

Definition 3.2. For $y_h \in K_h$ we denote by $A_h := \mathcal{Z}_h(\psi_h - y_h) \cap \mathcal{N}_h(\Omega)$ and $I_h := \mathcal{C}_h(\psi_h - y_h) \cap \mathcal{N}_h(\Omega)$ the sets of active and inactive nodal points. A nodal point is said to be an isolated active (inactive) nodal point, if $\mathcal{N}_h(\omega_a) \setminus \{a\} \subset I_h \cup \mathcal{N}_h(\Gamma)$ ($\mathcal{N}_h(\omega_a) \setminus \{a\} \subset A_h \cup \mathcal{N}_h(\Gamma)$). Moreover, the sets

$$\mathcal{A}_h := \bigcup\{T \in \mathcal{T}_h^z(\psi_h - y_h)\}, \tag{3.12a}$$

$$\mathring{\mathcal{I}}_h := \bigcup\{T \in \mathcal{T}_h^c(\psi_h - y_h)\}, \tag{3.12b}$$

$$\mathcal{F}_h(y_h) := \bigcup\{T \in \mathcal{T}_h^m(\psi_h - y_h)\} \tag{3.12c}$$

are referred to as the discrete active set, the discrete purely inactive set, and the discrete free boundary with respect to y_h. The set

$$\mathcal{I}_h := \overset{\circ}{\mathcal{I}}_h \cup \mathcal{F}_h(y_h) \tag{3.12d}$$

is said to be the discrete inactive set.

An edge $E \in \mathcal{E}_h(\bar{\Omega})$ is called active (purely inactive), if $\mathcal{N}_h(E) \subset \mathcal{A}_h$ ($\mathcal{N}_h(E) \subset \mathcal{I}_h$). The sets of active and purely inactive edges will be denoted by $\mathcal{E}_{\mathcal{A}_h}$ and $\overset{\circ}{\mathcal{E}}_{\mathcal{I}_h}$. We set $\mathcal{E}_{\mathcal{F}_h(y_h)} := \mathcal{E}_h(\bar{\Omega}) \setminus (\mathcal{E}_{\mathcal{A}_h} \cup \overset{\circ}{\mathcal{E}}_{\mathcal{I}_h})$ and $\mathcal{E}_{\mathcal{I}_h} := \overset{\circ}{\mathcal{E}}_{\mathcal{I}_h} \cup \mathcal{E}_{\mathcal{F}_h(y_h)}$. An active edge $E \in \mathcal{E}_{\mathcal{A}_h}$ is called isolated, if $E \in \mathcal{E}_{\mathcal{A}_h} \setminus \mathcal{E}_h(\mathcal{A}_h)$.

Likewise, for $\lambda_h \in M_h$ we denote by

$$\mathcal{Z}_h(\lambda_h) := \{a \in \mathcal{N}_h(\Omega) \mid \lambda_h(a) = 0\}, \quad \mathcal{C}_h(\lambda_h) := \mathcal{N}_h(\Omega) \setminus \mathcal{Z}_h(\lambda_h) \tag{3.13}$$

the sets of zero and non-zero nodal points with respect to λ_h and we partition $\mathcal{T}_h(\Omega)$ as follows

$$\mathcal{T}_h(\Omega) = \mathcal{T}_h^z(\lambda_h) \cup \mathcal{T}_h^c(\lambda) \cup \mathcal{T}_h^m(\lambda_h), \tag{3.14}$$

where

$$\mathcal{T}_h^z(\lambda_h) := \{T \in \mathcal{T}_h(\Omega) \mid \mathcal{N}_h(T) \subset \mathcal{Z}_h(\lambda_h) \cup \mathcal{N}_h(\Gamma)\}, \tag{3.15a}$$

$$\mathcal{T}_h^c(\lambda_h) := \{T \in \mathcal{T}_h(\Omega) \mid T \cap \Gamma = \emptyset \text{ and } \mathcal{N}_h(T) \subset \mathcal{C}_h(\lambda_h)\} \cup \{T \in \mathcal{T}_h(\Omega) \mid T \cap \Gamma \neq \emptyset \wedge \mathcal{N}_h(T) \cap \mathcal{N}_h(\Omega) \subset \mathcal{C}_h(\lambda_h) \wedge T \subset \mathcal{A}_h\}, \tag{3.15b}$$

$$\mathcal{T}_h^m(\lambda_h) := \mathcal{T}_h(\Omega) \setminus (\mathcal{T}_h^z(\lambda_h) \cup \mathcal{T}_h^c(\lambda_h)). \tag{3.15c}$$

Definition 3.3. For $\sigma_h \in M_h \cap \mathcal{M}_+(\bar{\Omega})$ the sets $Z_h := \mathcal{Z}_h(\sigma_h)$ and $C_h := \mathcal{C}_h(\sigma_h)$ are said to be the sets of zero and strongly active nodal points. Isolated zero (strongly active) nodal points are defined analogously to Definition 3.2.

An edge $E \in \mathcal{E}_h(\bar{\Omega})$ is said to be strongly active (purely zero), if $\mathcal{N}_h(E) \subseteq C_h$ ($\mathcal{N}_h(E) \subseteq Z_h$). The sets of strongly active and purely zero edges are denoted by $\mathcal{E}_{\mathcal{C}_h}$ and $\overset{\circ}{\mathcal{E}}_{\mathcal{Z}_h}$. We set $\mathcal{E}_{\mathcal{F}_h(\sigma_h)} := \mathcal{E}_h(\bar{\Omega}) \setminus (\mathcal{E}_{\mathcal{C}_h} \cup \overset{\circ}{\mathcal{E}}_{\mathcal{Z}_h})$ and $\mathcal{E}_{\mathcal{Z}_h} := \overset{\circ}{\mathcal{E}}_{\mathcal{Z}_h} \cup \mathcal{E}_{\mathcal{F}_h(\sigma_h)}$.

Moreover, the sets

$$\overset{\circ}{\mathcal{Z}}_h := \bigcup \{T \in \mathcal{T}_h^z(\sigma_h)\}, \tag{3.16a}$$

$$\mathcal{C}_h := \bigcup \{T \in \mathcal{T}_h^c(\sigma_h)\}, \tag{3.16b}$$

$$\mathcal{F}_h(\sigma_h) := \bigcup \{T \in \mathcal{T}_h^m(\sigma_h)\} \tag{3.16c}$$

are referred to as the discrete purely zero set, the discrete strongly active set, and the discrete free boundary with respect to σ_h. The set

$$\mathcal{Z}_h := \overset{\circ}{\mathcal{Z}}_h \cup \mathcal{F}_h(\sigma_h) \tag{3.16d}$$

is said to be the discrete zero set and the set

$$\mathcal{B}_h := \mathrm{cl}(\mathcal{A}_h \setminus \mathcal{C}_h) \tag{3.16e}$$

is called the discrete biactive set. If $\mathcal{B}_h = \emptyset$, we say that discrete strict complementarity holds true. Otherwise, there is a lack of discrete strict complementarity.

Zero (strongly active) edges and isolated zero (isolated strongly active) edges are defined similarly to Definition 3.2.

3.3 Discrete Stationarity Concepts

The discrete (strongly) active sets $\mathcal{A}_h, \mathcal{C}_h$, the discrete biactive set $\mathcal{B}_h$ and the discrete inactive set $\mathcal{I}_h$ will be used to classify stationary points in the discrete regime.

Definition 3.4. For $(y_h, \sigma_h, u_h) \in V_h \times M_h \times S_h^{(1)}$ assume that there exist $(p_h, \mu_h) \in V_h \times M_h$ such that it holds

$$a(y_h, v_h) = (f + u_h, v_h)_{0,\Omega} - \langle\langle \sigma_h, v_h \rangle\rangle,\ v_h \in V_h, \tag{3.17a}$$

$$\psi_h - y_h \geq 0,\ \sigma_h \in M_h \cap \mathcal{M}_+(\bar{\Omega}),\ \langle\langle \sigma_h, \psi_h - y_h \rangle\rangle = 0, \tag{3.17b}$$

$$a(p_h, v_h) = (y^d - y_h, v_h)_{0,\Omega} - \langle\langle \mu_h, v_h \rangle\rangle,\ v_h \in V_h, \tag{3.17c}$$

$$p_h = \alpha\ (u_h - u_h^d), \tag{3.17d}$$

$$p_h(a) = 0,\ a \in \mathcal{C}_h, \tag{3.17e}$$

$$\mu_h(a) = 0,\ a \in \mathcal{I}_h. \tag{3.17f}$$

The triple $(y_h, \sigma_h, u_h) \in V_h \times M_h \times S_h^{(1)}$ is called

(i) a discrete C-stationary point of (3.6), if the pair $(p_h, \mu_h) \in V_h \times M_h$ satisfies

$$\mu_h(a)\ p_h(a) \geq 0,\ a \in \mathcal{B}_h, \tag{3.17g}$$

(ii) a discrete S-stationary point of (3.6), if the pair $(p_h, \mu_h) \in V_h \times M_h$ fulfills

$$\mu_h(a) \geq 0,\ p_h(a) \geq 0,\ a \in \mathcal{B}_h, \tag{3.17h}$$

(iii) a discrete stationary point of (3.6), if $\mathcal{B}_h = \emptyset$, i.e.,

$$\mathcal{C}_h = \mathcal{A}_h. \tag{3.17i}$$

Remark 3.5. In view of (3.17e) and (3.17f), condition (3.17g) implies

$$\langle\langle \mu_h, p_h \rangle\rangle \geq 0. \tag{3.18}$$

However, the reverse does not hold true. If $\langle\langle \mu_h, p_h \rangle\rangle = \sum_{a \in \mathcal{N}_h(\mathcal{B}_h)} \mu_h(a)\, p_h(a) \geq 0$, this does not imply that every summand is nonnegative. In other words, condition (3.18) is weaker than (3.17g).

3.4 Extensions of the Discrete Lagrange Multipliers

In this subsection, we will first derive an explicit representation of the operation of the discrete Lagrange multipliers σ_h and μ_h on functions $v_h \in V_h$ and then provide two extensions $\hat{\sigma}_h, \hat{\mu}_h$ and $\tilde{\sigma}_h, \tilde{\mu}_h$ to functionals on V. The extensions $\hat{\sigma}_h, \hat{\mu}_h$ will be used in the convergence analysis of the finite element approximations in Sect. 4, whereas the extensions $\tilde{\sigma}_h, \tilde{\mu}_h$ will play an essential role in the a posteriori error analysis in Sect. 5.

For notational convenience, we introduce the operator $I_{D_h} : V_h \to V_h, D_h \subset \mathcal{N}_h(\Omega)$, defined by means of

$$I_{D_h}(v_h)(a) := \begin{cases} v_h(a)\,, a \in D_h \\ 0\,, a \in \mathcal{N}_h(\Omega) \setminus D_h \end{cases}, \quad v_h \in V_h. \tag{3.19}$$

It follows that I_{C_h} is the identity on $\mathcal{C}_h$, vanishes on $\overset{\circ}{\mathcal{Z}}_h$, whereas for $D = T \in \mathcal{T}_h(\mathcal{F}_h(\sigma_h))$ and $D = E \in \mathcal{E}_{\mathcal{F}_h(\sigma_h)}$:

$$I_{C_h}(v_h)|_D = \sum_{a \in \mathcal{N}_h(D) \cap \mathcal{C}_h} v_h(a)\, \varphi_h^{(a)}.$$

Likewise, I_{A_h} is the identity on $\mathcal{A}_h$, vanishes on $\overset{\circ}{\mathcal{I}}_h$, whereas for $D = T \in \mathcal{T}_h(\mathcal{F}_h(y_h))$ and $D = E \in \mathcal{E}_{\mathcal{F}_h(y_h)}$:

$$I_{\mathcal{A}_h}(v_h)|_D = \sum_{a \in \mathcal{N}_h(D) \cap \mathcal{A}_h} v_h(a)\, \varphi_h^{(a)}.$$

Proposition 3.6. *Let σ_h, μ_h be the discrete Lagrange multipliers from* Definition 3.4*, let $\mathcal{F}_h(y_h), \mathcal{F}_h(\sigma_h)$ be the discrete free boundaries with respect to y_h and σ_h according to* (3.12d) *and* (3.16c)*, and let I_{D_h} be given by* (3.19)*. Then, for $v_h \in V_h$ it holds*

$$\langle\langle \sigma_h, v_h \rangle\rangle = \sum_{T \in \mathcal{T}_h(\mathcal{C}_h \cup \mathcal{F}_h(\sigma_h))} \Big((f + u_h, I_{C_h}(v_h))_{0,T} - (\nabla y_h, \nabla I_{C_h}(v_h))_{0,T} \Big) = \tag{3.20a}$$

$$\sum_{T \in \mathcal{T}_h(\mathcal{C}_h \cup \mathcal{F}_h(\sigma_h))} (f + u_h, I_{C_h}(v_h))_{0,T} - \sum_{E \in \mathcal{E}_{\mathcal{C}_h} \cup \mathcal{E}_{\mathcal{F}_h(\sigma_h)}} (v_E \cdot [\nabla y_h]_E, I_{C_h}(v_h))_{0,E},$$

$$\langle\langle \mu_h, v_h \rangle\rangle = \sum_{T \in \mathcal{T}_h(\mathcal{A}_h \cup \mathcal{F}_h(y_h))} \Big((y^d - y_h, I_{A_h}(v_h))_{0,T} - (\nabla p_h, \nabla I_{A_h}(v_h))_{0,T} \Big) = \tag{3.20b}$$

$$\sum_{T \in \mathcal{T}_h(\mathcal{A}_h \cup \mathcal{F}_h(y_h))} (y^d - y_h, I_{A_h}(v_h))_{0,T} - \sum_{E \in \mathcal{E}_{\mathcal{A}_h} \cup \mathcal{E}_{\mathcal{F}_h(y_h)}} (v_E \cdot [\nabla p_h]_E, I_{A_h}(v_h))_{0,E}.$$

Proof. In view of (3.17a) and (3.17c) we have

$$\langle\langle \sigma_h, \varphi_h^{(a)} \rangle\rangle = (f + u_h, \varphi_h^{(a)})_{0,\omega_a} - (\nabla y_h, \nabla \varphi_h^{(a)})_{0,\omega_a}, \quad a \in \mathcal{C}_h,$$
$$\langle\langle \mu_h, \varphi_h^{(a)} \rangle\rangle = (y^d - y_h, \varphi_h^{(a)})_{0,\omega_a} - (\nabla p_h, \nabla \varphi_h^{(a)})_{0,\omega_a}, \quad a \in \mathcal{A}_h.$$

Due to (3.16d) and (3.17f) $\sigma_h(a) = 0, a \in \mathcal{Z}_h$, and $\mu_h(a) = 0, a \in \mathcal{I}_h$, whence

$$\sigma_h(a) = \begin{cases} \sum_{T \in \mathcal{T}_h(\omega_a)} \Big((f + u_h, \varphi_h^{(a)})_{0,T} - (\nabla y_h, \nabla \varphi_h^{(a)})_{0,T} \Big), & a \in \mathcal{C}_h \\ 0, & a \in \mathcal{Z}_h \end{cases}, \tag{3.21}$$

and

$$\mu_h(a) = \begin{cases} \sum_{T \in \mathcal{T}_h(\omega_a)} \Big((y^d - y_h, \varphi_h^{(a)})_{0,T} - (\nabla p_h, \nabla \varphi_h^{(a)})_{0,T} \Big), & a \in \mathcal{A}_h \\ 0, & a \in \mathcal{I}_h \end{cases}. \tag{3.22}$$

Applying Green's formula elementwise to the second terms on the right-hand side in (3.21) and (3.22) yields

$$\sigma_h(a) = \begin{cases} \sum_{T \in \mathcal{T}_h(\omega_a)} (f + u_h, \varphi_h^{(a)})_{0,T} - \sum_{E \in \mathcal{E}_h^a} (v_E \cdot [\nabla y_h]_E, \varphi_h^{(a)})_{0,E}, & a \in \mathcal{C}_h \\ 0, & a \in \mathcal{Z}_h \end{cases}, \tag{3.23}$$

and

$$\mu_h(a) = \begin{cases} \sum_{T\in\mathcal{T}_h(\omega_a)} (y^d - y_h, \varphi_h^{(a)})_{0,T} - \sum_{E\in\mathcal{E}_h^a} (\nu_E \cdot [\nabla p_h]_E, \varphi_h^{(a)})_{0,E} \ , a \in \mathcal{A}_h \\ 0 \ , a \in \mathcal{I}_h \end{cases} . \tag{3.24}$$

Taking $\langle\langle \sigma_h, v_h \rangle\rangle = \sum_{a\in\mathcal{N}_h(\mathcal{C}_h)} \sigma_h(a) v_h(a)$ into account, from (3.21) and (3.23) we deduce

$$\langle\langle \sigma_h, v_h \rangle\rangle = \sum_{a\in\mathcal{N}_h(\mathcal{C}_h)} \Big(\sum_{T\in\mathcal{T}_h(\omega_a)} \Big((f + u_h, v_h(a)\, \varphi_h^{(a)})_{0,T} - (\nabla y_h, v_h(a) \nabla \varphi_h^{(a)})_{0,T} \Big) \Big)$$

and

$$\langle\langle \sigma_h, v_h \rangle\rangle \ = \\ \sum_{a\in\mathcal{N}_h(\mathcal{C}_h)} \Big(\sum_{T\in\mathcal{T}_h(\omega_a)} (f + u_h, v_h(a)\varphi_h^{(a)})_{0,T} - \sum_{E\in\mathcal{E}_h(\mathcal{E}_h^a)} (\nu_E \cdot [\nabla y_h]_E, v_h(a)\varphi_h^{(a)})_{0,E} \Big).$$

Regrouping the summands in the above expressions gives (3.20a). The representation (3.20b) follows similarly. □

The first extensions $\hat{\sigma}_h, \hat{\mu}_h \in V^*$ of the discrete multipliers are defined in a similar way to the finite element analysis of variational inequalities of obstacle type (cf., e.g., [6]), whereas the second extensions $\tilde{\sigma}_h, \tilde{\mu}_h \in V^*$ are defined in view of Proposition 3.6.

Definition 3.7. Let $(y_h, \sigma_h, u_h, p_h, \mu_h) \in V_h \times M_h \times S_h^{(1)} \times V_h \times M_h$ satisfy (3.17a)–(3.17f). We define functionals $\hat{\sigma}_h, \hat{\mu}_h \in V^*$ by means of

$$\langle \hat{\sigma}_h, v \rangle := (f + u_h, v)_{0,\Omega} - a(y_h, v), \quad v \in V, \tag{3.25a}$$

$$\langle \hat{\mu}_h, v \rangle := (y^d - y_h, v)_{0,\Omega} - a(p_h, v), \quad v \in V, \tag{3.25b}$$

and functionals $\tilde{\sigma}_h, \tilde{\mu}_h \in V^*$ according to

$$\langle \tilde{\sigma}_h, v \rangle \ := \tag{3.26a}$$

$$\sum_{T\in\mathcal{T}_h(\mathcal{Z}_h)} (f + u_h, v)_{0,T} - \sum_{E\in\mathcal{E}_{\mathcal{Z}_h}} (\nu_E \cdot [\nabla y_h]_E, v)_{0,E} + F_h^{(\sigma)}(P_h^{SZ} v), \quad v \in V,$$

$$\langle \tilde{\mu}_h, v \rangle \ := \tag{3.26b}$$

$$\sum_{T\in\mathcal{T}_h(\mathcal{A}_h)} (y^d - y_h, v)_{0,T} - \sum_{E\in\mathcal{E}_{\mathcal{A}_h}} (\nu_E \cdot [\nabla p_h]_E, v)_{0,E} + F_h^{(\mu)}(P_h^{SZ} v), \quad v \in V,$$

where P_h^{SZ} stands for the Scott–Zhang interpolation operator (see, e.g., [9, 39]) and

$$F_h^{(\sigma)}(v_h) := \tag{3.26c}$$

$$\sum_{T\in\mathcal{T}_h(\mathcal{F}_h(\sigma_h))} (f + u_h, I_{C_h}(v_h))_{0,T} - \sum_{E\in\mathcal{E}_{\mathcal{F}_h(\sigma_h)}} (v_E \cdot [\nabla y_h]_E, I_{C_h}(v_h))_{0,E}, \tag{3.26d}$$

$$F_h^{(\mu)}(v_h) := \tag{3.26e}$$

$$\sum_{T\in\mathcal{T}_h(\mathcal{F}_h(y_h))} (y^d - y_h, I_{A_h}(v_h))_{0,T} - \sum_{E\in\mathcal{E}_{\mathcal{F}_h(y_h)}} (v_E \cdot [\nabla p_h]_E, I_{A_h}(v_h))_{0,E}. \tag{3.26f}$$

Remark 3.8. For later use in Sect. 5, we recall the definition of the Scott–Zhang interpolation operator: For each $a \in \mathcal{T}_h(\Omega)$ let $T \in \omega_a$ be an arbitrarily but fixed chosen element. Further, let $\{\Phi_T^{(a)} \mid a \in \mathcal{N}_h(T)\}$ be the $L^2(T)$-dual basis of $\{\varphi_h^{(a)} \mid a \in \mathcal{N}_h(T)\}$. Then, $P_h^{SZ} : L^2(\Omega) \to V_h$ is defined by means of

$$P_h^{SZ} v := \sum_{a\in\mathcal{N}_h(\Omega)} (P_h^{SZ} v)(a)\varphi_h^{(a)}, \tag{3.27}$$

where the nodal coefficients $(P_h^{SZ} v)(a)$ are given by

$$(P_h^{SZ} v)(a) := \int_T \Phi_T^{(a)}(x) v(x)\, dx. \tag{3.28}$$

Proposition 3.9. *The functionals $\hat{\sigma}_h, \hat{\mu}_h \in V^*$ and $\tilde{\sigma}_h, \tilde{\mu}_h \in V^*$ are extensions of $\sigma_h, \mu_h \in M_h$, i.e., for $v_h \in V_h$ it holds*

$$\langle \hat{\sigma}_h, v_h\rangle = \langle \tilde{\sigma}_h, v_h\rangle = \langle\langle \sigma_h, v_h\rangle\rangle,$$
$$\langle \hat{\mu}_h, v_h\rangle = \langle \tilde{\mu}_h, v_h\rangle = \langle\langle \mu_h, v_h\rangle\rangle.$$

Proof. The results are immediate consequences of (3.17) and Proposition 3.6. □

Remark 3.10. Fine properties of the extensions $\hat{\sigma}_h, \hat{\mu}_h \in V^*$ in terms of localizations involving the discrete active/inactive sets are difficult to obtain, whereas the extensions $\tilde{\sigma}_h, \tilde{\mu}_h \in V^*$ obviously satisfy

$$\mathcal{C}_h \subseteq \operatorname{supp}(\tilde{\sigma}_h) \subseteq \mathcal{C}_h \cup \mathcal{F}_h(\sigma_h), \tag{3.29a}$$

$$\operatorname{supp}(\tilde{\mu}_h) \subseteq \mathcal{A}_h \cup \mathcal{F}_h(y_h). \tag{3.29b}$$

The precise structure of $\tilde{\sigma}_h \in V^*$ depends on the definition of the Scott–Zhang interpolation operator P_h^{SZ}. In particular, under the condition

$$\text{For all } a \in \mathcal{C}_h \text{ there exists } T^{(a)} \subset \omega_a \text{ such that } T^{(a)} \subset \mathcal{C}_h, \tag{3.30}$$

we obtain $\mathrm{supp}(\tilde{\sigma}_h) = \mathcal{C}_h$, if the triangles satisfying (3.30) are used in the definition of P_h^{SZ}. We note that (3.30) excludes isolated strongly active nodal points and edges. However, utilizing a Scott–Zhang interpolation operator defined by averaging over edges instead of triangles (see [39]), allows to show $\mathrm{supp}(\tilde{\sigma}_h) = \mathcal{C}_h$, if we only exclude isolated strongly active nodal points. Similar remarks apply to $\tilde{\mu}_h$, i.e., it is possible to achieve $\mathrm{supp}(\tilde{\mu}_h) \subseteq \mathcal{A}_h$ instead of (3.29b), if no isolated active nodal points occur and the modified P_h^{SZ} is used.

4 Convergence Analysis of the Finite Element Approximation

In this section, we prove that for a sequence of discrete C-stationary points there exists a subsequence converging to an almost C-stationary point. To this end, we assume:

- **($\mathbf{A_1}$)** $\{(y_h, u_h, \sigma_h)\}_{\mathcal{H}}$ is a sequence of global minima of (3.7) or the sequences $\{y_h\}_{\mathcal{H}}$ and $\{u_h\}_{\mathcal{H}}$ are uniformly bounded in $L^2(\Omega)$.
- **($\mathbf{A_2}$)** The obstacle ψ satisfies $\Delta\psi \in L^2(\Omega)$.

Remark 4.1. Under assumption **($\mathbf{A_2}$)** we may restrict ourselves to the case $\psi = 0$, since otherwise we can replace f by $f + \Delta\psi$ and y^d by $y^d - \psi$.

Theorem 4.2. *Let* $\{(y_h, \sigma_h, u_h)\}_{\mathcal{H}}, (y_h, \sigma_h, u_h) \in V_h \times M_h \times S_h^{(1)}, h \in \mathcal{H}$, *be a sequence of discrete C-stationary points of* (3.6). *Further, let* $\{(p_h, \mu_h)\}_{\mathcal{H}}, (p_h, \mu_h) \in V_h \times M_h, h \in \mathcal{H}$, *be the sequence of associated discrete adjoint states and multipliers computed with respect to a sequence* $\{V_h\}_{\mathcal{H}}$ *of nested finite element spaces. Finally, let* $\hat{\sigma}_h \in V^*$ *and* $\hat{\mu}_h \in V^*$ *be the extensions of the multipliers* σ_h *and* μ_h *as given by* (3.25).

If the assumptions **($\mathbf{A_1}$)** *and* **($\mathbf{A_2}$)** *are satisfied and the sequence* $\{V_h\}_{\mathcal{H}}$ *is limit dense in* V, *then there exist a subsequence* $\mathcal{H}' \subset \mathcal{H}$ *and an almost C-stationary point* $(y^*, \sigma^*, u^*) \in V \times V^* \times L^2(\Omega)$ *of* (2.5) *with associated adjoint state* $p^* \in V$ *and multiplier* $\mu^* \in V^*$ *such that for* $h \in \mathcal{H}', h \to 0$ *it holds*

$$y_h \to y^* \quad \text{in } V, \tag{4.1a}$$

$$y_h \to y^* \quad \text{in } L^2(\Omega), \tag{4.1b}$$

$$\hat{\sigma}_h \to \sigma^* \quad \text{in } V^*, \tag{4.1c}$$

$$u_h \to u^* \quad \text{in } L^2(\Omega), \tag{4.1d}$$

$$p_h \rightharpoonup p^* \quad \text{in } V, \tag{4.1e}$$

$$p_h \to p^* \quad \text{in } L^2(\Omega), \tag{4.1f}$$

$$\hat{\mu}_h \overset{*}{\rightharpoonup} \mu^* \quad \text{in } V^*. \tag{4.1g}$$

Moreover, if $\{S_h^{(1)}\}_{\mathcal{H}}$ *is limit dense in* $H^1(\Omega)$*, we have*

$$\langle \mu^*, y^* v \rangle = 0 \quad \textit{for all } v \in C^1(\bar{\Omega}). \tag{4.1h}$$

Proof. Assume that $\{(y_h, \sigma_h, u_h)\}_{\mathcal{H}}$ is a sequence of global minima. The triple $(y_h, \sigma_h, u_h) = (0, -f_h, 0)$ is a feasible point for (3.6) and hence, $J_h(y_h, u_h) \leq J_h(0, -f_h)$. By the inverse triangle inequality and Young's inequality it follows that the sequences $\{y_h\}_{\mathcal{H}}$ and $\{u_h\}_{\mathcal{H}}$ are bounded in $L^2(\Omega)$.

If $\{(y_h, \sigma_h, u_h)\}_{\mathcal{H}}$ is a sequence of stationary points, the boundedness of $\{y_h\}_{\mathcal{H}}$ and $\{u_h\}_{\mathcal{H}}$ in $L^2(\Omega)$ follows from assumption $(\mathbf{A_1})$.

Choosing $v_h = y_h$ in (3.17a) and $v_h = p_h$ in (3.17c) and taking (2.4b),(3.17b), and (3.18) into account, we obtain

$$\begin{aligned}
\gamma \, \|y_h\|_{1,\Omega}^2 &\leq a(y_h, y_h) = (f + u_h, y_h)_{0,\Omega} \leq \Big(\|f\|_{0,\Omega} + \|u_h\|_{0,\Omega} \Big) \, \|y_h\|_{1,\Omega}, \\
\gamma \, \|p_h\|_{1,\Omega}^2 &\leq a(p_h, p_h) = (y^d - y_h, p_h)_{0,\Omega} - \langle\langle \mu_h, p_h \rangle\rangle \\
&\leq (y^d - y_h, p_h)_{0,\Omega} \leq \Big(\|y^d\|_{0,\Omega} + \|y_h\|_{0,\Omega} \Big) \, \|p_h\|_{1,\Omega}.
\end{aligned}$$

In view of the boundedness of $\{y_h\}_{\mathcal{H}}$ and $\{u_h\}_{\mathcal{H}}$ in $L^2(\Omega)$, the preceding two inequalities imply the boundedness of $\{y_h\}_{\mathcal{H}}$ and $\{p_h\}_{\mathcal{H}}$ in V. Moreover, observing (2.4b), for $v \in V$ we have

$$\begin{aligned}
|\langle \hat{\sigma}_h, v \rangle| &\leq \|f + u_h\|_{0,\Omega} \, \|v\|_{0,\Omega} + C \, \|y_h\|_{1,\Omega} \, \|v\|_{1,\Omega} \\
&\leq (\|f + u_h\|_{0,\Omega} + C \, \|y_h\|_{1,\Omega}) \, \|v\|_{1,\Omega}, \\
|\langle \hat{\mu}_h, v \rangle| &\leq \|y^d - y_h\|_{0,\Omega} \, \|v\|_{0,\Omega} + C \, \|p_h\|_{1,\Omega} \, \|v\|_{1,\Omega} \\
&\leq (\|y^d - y_h\|_{0,\Omega} + C \, \|p_h\|_{1,\Omega}) \, \|v\|_{1,\Omega},
\end{aligned}$$

whence

$$\|\hat{\sigma}_h\|_{V^*} \leq \|f + u_h\|_{0,\Omega} + C \, \|y_h\|_{1,\Omega}), \quad \|\hat{\mu}_h\|_{V^*} \leq \|y^d - y_h\|_{0,\Omega} + C \, \|y_h\|_{1,\Omega}).$$

This implies boundedness of the sequences $\{\hat{\sigma}_h\}_{\mathcal{H}}$ and $\{\hat{\mu}_h\}_{\mathcal{H}}$ in V^*. Consequently, there exist a subsequence $\mathcal{H}' \subset \mathcal{H}$ and a point $(y^*, \sigma^*, u^*, p^*, \mu^*) \in V \times V^* \times L^2(\Omega) \times V \times V^*$ such that for $h \in \mathcal{H}', h \to 0$ it holds

$$y_h \rightharpoonup y^* \quad \text{in } V, \quad p_h \rightharpoonup p^* \quad \text{in } V, \tag{4.2a}$$

$$u_h \rightharpoonup u^* \quad \text{in } L^2(\Omega), \tag{4.2b}$$

$$\hat{\sigma}_h \rightharpoonup^* \sigma^* \quad \text{in } V^*, \quad \hat{\mu}_h \rightharpoonup^* \mu^* \quad \text{in } V^*. \tag{4.2c}$$

Due to the Rellich–Kondrachov theorem V is compactly embedded in $L^2(\Omega)$ and hence, (4.2a) implies (4.1b),(4.1f).

For another subsequence, still denoted by $\mathcal{H}'$, we further deduce that for $h \in \mathcal{H}', h \to 0$ we have $y_h \to y^*$ and $p_h \to p^*$ pointwise almost everywhere. Hence, $y_h \le 0, h \in \mathcal{H}'$, implies $y^* \le 0$ almost everywhere (a.e.) in Ω.

Next, we show that the point $(y^*, \sigma^*, u^*, p^*, \mu^*)$ satisfies the state equation (2.13a), the adjoint state equation (2.13c), and (2.13d). Since $\{V_h\}_{\mathcal{H}}$ is limit dense in V, for any $v \in V$ we find a sequence $\{v_h\}_{\mathcal{H}}, v_h \in V_h, h \in \mathcal{H}$, such that $v_h \to v$ for $h \to 0$. Observing (4.2), for $h \in \mathcal{H}', h \to 0$, we deduce

$$
\begin{aligned}
&a(y_h, v_h) \to a(y^*, v), \quad a(p_h, v_h) \to a(p^*, v),\\
&(f + u_h, v_h)_{0,\Omega} \to (f + u^*, v)_{0,\Omega}, \quad (y^d - y_h, v_h)_{0,\Omega} \to (y^d - y^*, v)_{0,\Omega},\\
&\langle\langle \sigma_h, v_h\rangle\rangle = \langle\langle \hat{\sigma}_h, v_h\rangle\rangle \to \langle \sigma^*, v\rangle, \quad \langle\langle \mu_h, v_h\rangle\rangle = \langle\langle \hat{\mu}_h, v_h\rangle\rangle \to \langle \mu^*, v\rangle.
\end{aligned}
$$

Hence, passing to the limit in (3.17a) and (3.17c), we find that $(y^*, \sigma^*, u^*, p^*, \mu^*)$ satisfies (2.13a) and (2.13c).

The limit density of $\{V_h\}_{\mathcal{H}}$ in V further implies $u_h^d \to u^d, h \to 0$. Consequently, (3.17d) and (4.2) imply that (4.1d) holds true and that the pair (p^*, u^*) fulfills (2.13d).

Next, we verify $\sigma^* \in V_+^*$. Since $\{(V_h)_+\}_{\mathcal{H}}$ is limit dense in V_+, for any $v \in V_+$ there exists a sequence $\{v_h\}_{\mathcal{H}}, v_h \in (V_h)_+, h \in \mathcal{H}$, such that $v_h \to v$ as $h \to 0$. Observing $\sigma_h \in \mathcal{M}_+(\bar{\Omega})$ and (4.2c), we find

$$
0 \le \langle\langle \sigma_h, v_h\rangle\rangle = \langle \hat{\sigma}_h, v_h\rangle \to \langle \sigma^*, v\rangle,
$$

whence $\langle \sigma^*, v\rangle$ for any $v \in V_+$.

In order to establish strong convergence of the states in V, due to (3.6b) we have

$$
a(y_h, y_h) \le a(y_h, v_h) + (f + u_h, y_h - v_h)_{0,\Omega}, \quad v_h \in V_h \cap V_-. \tag{4.3}
$$

Since the sequence $\{V_h \cap V_-\}_{\mathcal{H}}$ is limit dense in V_-, there exists a sequence $\{v_h\}_{\mathcal{H}}, v_h \in V_h \cap V_-, h \in \mathcal{H}$, such that $v_h \to y^* \in V_-$ as $h \to 0$. Taking (2.4b) and (4.3) into account, it holds

$$
\begin{aligned}
\gamma \, \|y_h - y^*\|_{1,\Omega}^2 &\le a(y_h - y^*, y_h - y^*) = a(y_h, y_h) - a(y_h, y^*) - a(y^*, y_h - y^*)\\
&\le a(y_h, v_h) + (f + u_h, v_h)_{0,\Omega} - a(y_h, y^*) - a(y^*, y_h - y^*).
\end{aligned}
$$

Due to the already proven assertions (4.1b),(4.1d) and in view of (4.2a) the right-hand side in the preceding inequality converges to zero which implies (4.1a). Moreover, observing (3.17b),(3.17f), and (4.1a), it follows that

$$
0 = \langle \hat{\sigma}_h, y_h\rangle \to \langle \sigma^*, y^*\rangle, \quad 0 = \langle \hat{\mu}_h, y_h\rangle \to \langle \mu^*, y^*\rangle,
$$

whence $\langle \sigma^*, y^*\rangle = \langle \mu^*, y^*\rangle = 0$.

For the proof of (4.1c), we note that the compact embedding of $L^2(\Omega)$ in V^* implies $u_h \to u^*$ in V^* as $\mathcal{H}' \ni h \to 0$. Since $A \in \mathcal{L}(V, V^*)$ is bounded, we obtain

$$\begin{aligned}\|\hat{\sigma}_h - \sigma^*\|_{V^*} &\le \|Ay_h - Ay^*\|_{V^*} + \|u_h - u^*\|_{V^*} \\ &\le \|A\|_{\mathcal{L}(V,V^*)} \|y_h - y^*\|_V + \|u_h - u^*\|_{V^*} \to 0 \ (h \to 0),\end{aligned}$$

which implies (4.1c). Moreover, due to (3.17e),(4.1c), and (4.1e) we have

$$0 = \langle \hat{\sigma}_h, p_h \rangle \to \langle \sigma^*, p^* \rangle \ (\mathcal{H}' \ni h \to 0),$$

whence $\langle \sigma^*, p^* \rangle = 0$.

Next, we show $\langle \mu^*, p^* \rangle \ge 0$. To this end, setting $v_h = p_h$ in (3.17c) and observing $\langle\langle \mu_h, p_h \rangle\rangle \ge 0$, we find

$$0 \ge a(p_h, p_h) - (y^d - y_h, p_h)_{0,\Omega}. \tag{4.4}$$

Since the functional $v \in V \mapsto a(v, v)$ is lower semicontinuous and convex, it is weakly lower semicontinuous whence due to (4.2a)

$$a(p^*, p^*) \le \liminf a(p_h, p_h).$$

On the other hand, the already proven assertions (4.1b),(4.1f) imply

$$(y^d - y_h, p_h)_{0,\Omega} \to (y^d - y^*, p^*)_{0,\Omega} \ (\mathcal{H}' \ni h \to 0).$$

Consequently, passing to the limit in (4.4) and taking into account that the triple (y^*, σ^*, u^*) satisfies (2.13c), we obtain

$$0 \ge a(p^*, p^*) - (y^d - y^*, p^*)_{0,\Omega} = -\langle \mu^*, p^* \rangle,$$

which proves $\langle \mu^*, p^* \rangle \ge 0$.

In order to verify that p^* satisfies (2.13e), we show

$$\langle \sigma^*, (p^*)^+ \rangle = \langle \sigma^*, (p^*)^- \rangle = 0, \tag{4.5}$$

which implies $p^* = 0$ in $\mathcal{C}^* = \mathrm{int}(\mathrm{supp}(\sigma^*))$ by Corollary 2.9. We note that (4.2a) gives rise to

$$(p_h)^+ \rightharpoonup (p^*)^+, \quad (p_h)^- \rightharpoonup (p^*)^- \text{ in } V \text{ as } \mathcal{H}' \ni h \to 0$$

(cf., e.g., [27]). Together with (3.17e), this leads to

$$0 = \langle\langle \sigma_h, (p_h)^+ \rangle\rangle \to \langle \sigma^*, (p^*)^+ \rangle, \quad 0 = \langle\langle \sigma_h, (p_h)^- \rangle\rangle \to \langle \sigma^*, (p^*)^- \rangle \ (\mathcal{H}' \ni h \to 0),$$

which proves (4.5).

It remains to show that (y^*, σ^*, u^*) is an almost C-stationary point and to prove (4.1h). In order to verify (4.1h), let $v \in C^1(\bar{\Omega})$. We have $y^* v \in V$ (cf., e.g., [13]). Since the sequence $\{S_h^{(1)}\}_{\mathcal{H}}$ is limit dense in $H^1(\Omega)$, there exists a sequence $\{v_h\}_{\mathcal{H}}, v_h \in S_h^{(1)}, h \in \mathcal{H}$, such that $v_h \to v$ $(\mathcal{H} \ni h \to 0)$. Observing $v_h \in C(\bar{\Omega}), y_h \in C_0(\Omega)$, we have $v_h y_h \in C_0(\Omega), h \in \mathcal{H}$, which together with $(v_h y_h)|_T \in H^1(T), T \in \mathcal{T}_h(\Omega)$, implies $v_h y_h \in V, h \in \mathcal{H}$. Taking (4.1a) into account, we deduce $y_h v_h \to y^* v$ in V as $\mathcal{H}' \ni h \to 0$. Since $(y_h v_h)(a) = 0, a \in A_h$, it follows that

$$0 = \langle \hat{\mu}_h, y_h v_h \rangle \to \langle \mu^*, y^* v \rangle \ (\mathcal{H}' \ni h \to 0).$$

Hence, $\langle \mu^*, y^* v \rangle = 0$ which proves (4.1h), since $v \in C^1(\bar{\Omega})$ was chosen arbitrarily.

In order to prove (2.13i), we note that (3.17f) yields

$$\langle \hat{\mu}_h, v_h \rangle = 0, \quad v_h \in V_h \cap V_{\mathcal{I}_h \cup \mathcal{F}_h(y_h)}. \tag{4.6}$$

On the other hand, due to the pointwise a.e. convergence of $\{y_h\}_{\mathcal{H}'}$ to y^*, for sufficiently small $h_1 \in \mathcal{H}'$ we have

$$y_h < 0 \text{ a.e. in } \mathcal{I}^*, \quad \mathcal{H}' \ni h \le h_1, \tag{4.7}$$

which shows $\mathcal{I}^* \subseteq \mathcal{I}_h$ for $h \le h_1$. For $h \le h_1$ we define

$$\tilde{\mathcal{I}}_h := \bigcup \{T \in \mathcal{T}_h(\Omega) \mid \operatorname{int}(T) \subseteq \mathcal{I}^*\},$$

such that $\tilde{\mathcal{I}}_h \subseteq \mathcal{I}^* \subseteq \mathcal{I}_h, \mathcal{H}' \ni h \le h_1$. Since $\tilde{\mathcal{I}}_h$ may be empty, we choose $h_2 \in \mathcal{H}'$ sufficiently small so that $\tilde{\mathcal{I}}_h \neq \emptyset$ for $\mathcal{H}' \ni h \le h_2$. Setting $h_3 := \min(h_1, h_2)$, we thus have

$$\emptyset \neq \tilde{\mathcal{I}}_h \subseteq \mathcal{I}^* \subseteq \mathcal{I}_h, \quad \mathcal{H}' \ni h \le h_3. \tag{4.8}$$

Now, let $v \in C_{\mathcal{I}^*,0} := \{v \in C_0(\Omega) \mid v|_{\mathcal{I}^*} \in C_0^\infty(\mathcal{I}^*), v|_{\Omega \setminus \mathcal{I}^*} = 0\}$ be chosen arbitrarily, but fixed. Since $\operatorname{supp}(v) \subseteq \mathcal{I}^*$, there exists $h(v) \in \mathcal{H}', h(v) \le h_3$, such that

$$\operatorname{supp}(v) \subseteq \tilde{\mathcal{I}}_h \subseteq \mathcal{I}^* \subseteq \mathcal{I}_h, \quad \mathcal{H}' \ni h \le h(v).$$

Obviously, we have $v \in V_{\tilde{\mathcal{I}}_{h(v)}} \subseteq V_{\mathcal{I}^*,0}$ and $\tilde{\mathcal{I}}_{h(v)} \subseteq \mathcal{I}_h, h \le h(v)$, whence

$$V_h \cap V_{\tilde{\mathcal{I}}_{h(v)}} \subseteq V_h \cap V_{\mathcal{I}_h \cup \mathcal{F}_h(y_h)}, \quad h \le h(v).$$

Observing (4.6), it follows that

$$\langle \hat{\mu}_h, v_h \rangle = 0, \quad v_h \in V_h \cap V_{\tilde{\mathcal{I}}_{h(v)}},\ h \le h(v). \tag{4.9}$$

Since the sequence $\{V_h \cap V_{\tilde{\mathcal{I}}_{h(v)}}\}_{h \le h(v)} \subset V_{\tilde{\mathcal{I}}_{h(v)}}$ is limit dense in $V_{\tilde{\mathcal{I}}_{h(v)}}$, there exists a sequence $\{v_h\}_{h \le h(v)}$, $v_h \in V_h \cap V_{\tilde{\mathcal{I}}_{h(v)}}$, $h \le h(v)$, such that $v_h \to v$ as $h(v) \ge h \to 0$. In view of (4.2c) and (4.9), it follows that

$$0 = \langle \hat{\mu}_h, v_h \rangle \to \langle \mu^*, v \rangle \ (h(v) \ge h \to 0),$$

which gives $\langle \mu^*, v \rangle = 0, v \in C_{\mathcal{I}^*,0}$. The density of $C_{\mathcal{I}^*,0}$ in $V_{\mathcal{I}^*,0}$ implies (2.13i). □

5 A Posteriori Error Control

In this section, we want to derive a residual-type a posteriori error estimator for the discretization errors in the state, the adjoint state, and the control

$$e_{h,y} := y - y_h, \quad e_{h,p} := p - p_h, \quad e_{h,u} := u - u_h \tag{5.1}$$

that provides both an upper bound (reliability) and a lower bound (efficiency) up to consistency errors and data oscillations. The total discretization error $e_h := (e_{h,y}, e_{h,p}, e_{h,u})$ will be measured in the norm

$$|||e_h||| := \left(\|e_{h,y}\|^2_{1,\Omega} + \|e_{h,p}\|^2_{1,\Omega} + \|e_{h,u}\|^2_{0,\Omega} \right)^{1/2}, \tag{5.2}$$

and we will show

$$\eta_h^2 - e^c_{h,eff} - \mathrm{osc}^2_{h,eff} \lesssim |||e_h|||^2 \lesssim \eta_h^2 + e^c_{h,rel} + \mathrm{osc}^2_{h,rel}.$$

Here, η_h is the residual a posteriori error estimator, whereas $e^c_{h,rel}, e^c_{h,eff}$ and $\mathrm{osc}_{h,rel}, \mathrm{osc}_{h,eff}$ stand for the consistency errors and data oscillations associated with the reliability and efficiency estimates.

5.1 Components of the Reliability and Efficiency Estimates

In this subsection, we introduce the residual-type a posteriori error estimator consisting of element and edge residuals, discuss the consistency errors due to a mismatch in complementarity between the continuous and the discrete regime, and present the data oscillations.

5.1.1 Residual-Type a Posteriori Error Estimator

The residual-type a posteriori error estimator η_h is given by

$$\eta_h := \left((\eta_h^{(1)})^2 + (\eta_h^{(2)})^2\right)^{1/2}, \tag{5.3}$$

where $\eta_h^{(1)}$ and $\eta_h^{(2)}$ consist of element residuals and edge residuals associated with the state equation (2.13a) and the adjoint state equation (2.13c)

$$\eta_h^{(1)} := \Big(\sum_{T\in\mathcal{T}_h(\mathcal{Z}_h)} (\eta_T^{(1)})^2 + \sum_{E\in\mathcal{E}_{\mathcal{Z}_h}} (\eta_E^{(1)})^2 \Big)^{1/2}, \tag{5.4a}$$

$$\eta_h^{(2)} := \Big(\sum_{T\in\mathcal{T}_h(\mathcal{I}_h)} (\eta_T^{(2)})^2 + \sum_{E\in\mathcal{E}_{\mathcal{I}_h}} (\eta_E^{(2)})^2 \Big)^{1/2}. \tag{5.4b}$$

In particular, the element residuals $\eta_T^{(\nu)}$ and the edge residuals $\eta_E^{(\nu)}$, $1 \le \nu \le 2$, are given by

$$\eta_T^{(1)} := h_T \, \|f + u_h\|_{0,T}, \quad , \eta_T^{(2)} := h_T \, \|y^d - y_h\|_{0,T}, \tag{5.5a}$$

$$\eta_E^{(1)} := h_E^{1/2} \, \|\nu_E \cdot [\nabla y_h]_E\|_{0,E}, \quad , \eta_E^{(2)} := h_E^{1/2} \, \|\nu_E \cdot [\nabla p_h]_E\|_{0,E}. \tag{5.5b}$$

5.1.2 Consistency Error (Mismatch in Complementarity)

We distinguish between reliability and efficiency related consistency errors.
Consistency Error for the Reliability Estimate.

$$e_{h,rel}^c := e_{h,\sigma}^{(1)} + e_{h,\sigma}^{(2)} + e_{h,\mu}^{(1)} + e_{h,\mu}^{(2)}, \tag{5.6}$$

where $e_{h,\sigma}^{(\nu)}, e_{h,\mu}^{(\nu)}$, $1 \le \nu \le 2$, are given by

$$e_{h,\sigma}^{(1)} := \langle \tilde{\sigma}_h - \sigma, y - y_h \rangle, \quad e_{h,\sigma}^{(2)} := -\langle \tilde{\sigma}_h - \sigma, p - p_h \rangle, \tag{5.7a}$$

$$e_{h,\mu}^{(1)} := \langle \tilde{\mu}_h - \mu, y - y_h \rangle, \quad e_{h,\mu}^{(2)} := \langle \tilde{\mu}_h - \mu, p - p_h \rangle. \tag{5.7b}$$

Consistency Error for the Efficiency Estimate.

$$e_{h,eff}^c := \Big(\sum_{T\in\mathcal{T}_h(\mathcal{Z}_h)} e_T^{(\sigma)} + \sum_{T\in\mathcal{T}_h(\mathcal{I}_h)} e_T^{(\mu)} + \sum_{E\in\mathcal{E}_{\mathcal{Z}_h}} e_{\omega_E}^{(\sigma)} + \sum_{E\in\mathcal{E}_{\mathcal{I}_h}} e_{\omega_E}^{(\mu)} \Big), \tag{5.8}$$

where $e_T^{(\sigma)}, e_T^{(\mu)}$, and $e_{\omega_E}^{(\sigma)}, e_{\omega_E}^{(\mu)}$ are given by

$$e_T^{(\sigma)} := |(f_h + u_h)\, b_T|_{1,T}^{-1}\, \langle \sigma, (f_h + u_h)\, b_T \rangle, \tag{5.9a}$$

$$e_T^{(\mu)} := |(y_h^d - y_h)\, b_T|_{1,T}^{-1}\, \langle \mu, (y_h^d - y_h)\, b_T \rangle, \tag{5.9b}$$

$$e_{\omega_E}^{(\sigma)} := |v_E \cdot [\nabla y_h]_E\, b_E|_{1,\omega_E}^{-1}\, \langle \sigma, v_E \cdot [\nabla y_h]_E\, b_E \rangle, \tag{5.9c}$$

$$e_{\omega_E}^{(\mu)} := -|v_E \cdot [\nabla p_h]_E\, b_E|_{1,T}^{-1}\, \langle \mu, v_E \cdot [\nabla p_h]_E\, b_E \rangle, \tag{5.9d}$$

and b_T, b_E stand for the element and edge bubble functions.

5.1.3 Data Oscillations

As in case of the consistency errors, we distinguish between reliability and efficiency related data oscillations.

Data Oscillations for the Reliability Estimate.

$$\mathrm{osc}_{h,rel} := \Big(\sum_{T \in \mathcal{T}_h(\Omega)} \mathrm{osc}_T^2(u^d) \Big)^{1/2}, \tag{5.10}$$

where $\mathrm{osc}_T(u^d)$ is given by

$$\mathrm{osc}_T(u^d) := \|u^d - u_h^d\|_{0,T}. \tag{5.11}$$

Data Oscillations for the Efficiency Estimate.

$$\mathrm{osc}_{h,eff} := \Big(\sum_{T \in \mathcal{T}_h(\mathcal{Z}_h)} \mathrm{osc}_T^2(f) + \sum_{T \in \mathcal{T}_h(\mathcal{I}_h)} \mathrm{osc}_T^2(y^d) \Big)^{1/2}, \tag{5.12}$$

where $\mathrm{osc}_T(f)$ and $\mathrm{osc}_T(y^d)$ are given by

$$\mathrm{osc}_T(f) := h_T\, \|f - f_h\|_{0,T}, \quad \mathrm{osc}_T(y^d) := h_T\, \|y^d - y_h^d\|_{0,T}. \tag{5.13}$$

5.2 *Reliability of the Error Estimator*

Theorem 5.1. *Let* (y, σ, u, p, μ) *and* $(y_h, \sigma_h, u_h, p_h, \mu_h)$ *be solutions of* (2.13a)–(2.13g) *and* (3.17a)–(3.17f) *and let* $\eta_h, e_{h,rel}^c, osc_{h,rel}$ *be the residual-type error estimator, the consistency error, and the data oscillations as given by* (5.3),(5.6), *and* (5.10). *Then, it holds*

$$|||e_h|||^2 \lesssim \eta_h^2 + e_{h,rel}^c + osc_{h,rel}^2. \tag{5.14}$$

The proof of Theorem 5.1 will be given by a series of lemmas.

We note that neither $e_{h,y}$ nor $e_{h,p}$ satisfy Galerkin orthogonality due to the presence of u, u_h in the right-hand sides of the continuous and discrete state equations (2.13a),(3.17a) and of y, y_h in the right-hand sides of the continuous and discrete adjoint state equations (2.13c),(3.17c). As in the case of the a posteriori error analysis of finite element approximations of control and/or state constrained distributed optimal control problems for second order elliptic PDEs, Galerkin orthogonality can be achieved with respect to an auxiliary state $y(u_h) \in V$ and an auxiliary adjoint state $p(y_h) \in V$ which are defined as the unique solutions of the variational equations

$$a(y(u_h), v) = (f + u_h, v)_{0,\Omega} - \langle \tilde{\sigma}_h, v \rangle, \quad v \in V, \tag{5.15a}$$

$$a(p(y_h), v) = (y^d - y_h, v)_{0,\Omega} - \langle \tilde{\mu}_h, v \rangle, \quad v \in V. \tag{5.15b}$$

In fact, it follows easily from (5.15a),(3.17a) and (5.15b),(3.17c) that

$$a(y(u_h) - y_h, v_h) = 0, \quad v_h \in V_h, \tag{5.16a}$$

$$a(p(y_h) - p_h, v_h) = 0, \quad v_h \in V_h. \tag{5.16b}$$

Lemma 5.2. *Under the assumptions of* Theorem 5.1 *let* $y(u_h), p(y_h)$ *be the auxiliary state and the auxiliary adjoint state as given by* (5.15a) *and* (5.15b) *and let* $\eta_h^{(1)}$ *and* $\eta_h^{(2)}$ *be the components of the residual a posteriori error estimator according to* (5.4a) *and* (5.4b)*. Then, it holds*

$$\|y(u_h) - y_h\|_{1,\Omega} \lesssim \eta_h^{(1)}, \tag{5.17a}$$

$$\|p(y_h) - p_h\|_{1,\Omega} \lesssim \eta_h^{(2)}. \tag{5.17b}$$

Proof. Denoting by P_h^C Clément's quasi-interpolation operator (cf., e.g., [44]), due to Proposition 3.9 and (5.16a) for $e := y(u_h) - y_h$ it holds

$$\|e\|_{1,\Omega}^2 \lesssim a(e, e) = r(e - P_h^C e), \tag{5.18}$$

where the residual $r(\cdot)$ is given by

$$r(v) := (f + u_h, v)_{0,\Omega} - \langle \tilde{\sigma}_h, v \rangle - a(y_h, v), \quad v \in V.$$

In view of the representation (3.26a) of the extension $\tilde{\sigma}_h$ of the discrete multiplier σ_h, by straightforward estimation we obtain

$$r(e - P_h^C e) \le |\sum_{T\in\mathcal{T}_h(\mathcal{Z}_h)} (f + u_h, e - P_h^C e)_{0,T}| \tag{5.19}$$
$$+ |\sum_{E\in\mathcal{E}_{\mathcal{Z}_h}} (\nu_E \cdot [\nabla y_h]_E, e - P_h^C e)_{0,E}| + |F_h^{(\sigma)}(P_h^{SZ}(e - P_h^C e))|.$$

Taking advantage of the properties

$$\|e - P_h^C e\|_{0,T} \lesssim h_T\, |e|_{1,\omega_h^T}, \quad \|e - P_h^C e\|_{0,E} \lesssim h_T^{1/2}\, |e|_{1,\omega_h^E}$$

of Clément's quasi-interpolation operator, for the first two terms on the right-hand side of (5.19) it follows that

$$|\sum_{T\in\mathcal{T}_h(\mathcal{Z}_h)} (f + u_h, e - P_h^C e)_{0,T}| \le \tag{5.20a}$$
$$\sum_{T\in\mathcal{T}_h(\mathcal{Z}_h)} \|f + u_h\|_{0,T}\, \|e - P_H^C e\|_{0,T} \lesssim \sum_{T\in\mathcal{T}_h(\mathcal{Z}_h)} \eta_T^{(1)}\, |e|_{1,\omega_h^T},$$
$$|\sum_{E\in\mathcal{E}_{\mathcal{Z}_h}} (\nu_E \cdot [\nabla y_h]_E, e - P_h^C e)_{0,E}| \le \tag{5.20b}$$
$$\sum_{E\in\mathcal{E}_{\mathcal{Z}_h}} \|\nu_E \cdot [\nabla y_h]_E\|_{0,E}\, \|e - P_h^C e\|_{0,E} \lesssim \sum_{E\in\mathcal{E}_{\mathcal{Z}_h}} \eta_E^{(1)}\, |e|_{1,\omega_h^E}.$$

For the third term on the right-hand side in (5.19), in view of (3.26c) and the definition of the Scott–Zhang interpolation operator P_h^{SZ} we obtain

$$|F_h^{(\sigma)}(P_h^{SZ}(e - P_h^C e))| \le \tag{5.21}$$
$$\sum_{T\in\mathcal{T}_h(\mathcal{F}_h(\sigma_h))} \Big(\|f + u_h\|_{0,T} \sum_{a\in\mathcal{N}_h(T)\cap\mathcal{C}_h} \|(P_h^{SZ}(e - P_h^C e))(a)\varphi_h^{(a)}\|_{0,T}\Big)$$
$$+ \sum_{E\in\mathcal{E}_{\mathcal{F}_h(\sigma_h)}} \|\nu_E \cdot [\nabla y_h]_E\|_{0,E}\, \|(P_h^{SZ}(e - P_h^C e))(a'_E)\varphi_h^{(a'_E)}\|_{0,E},$$

where a'_E stands for the single nodal point in $\mathcal{N}_h(E) \cap C_h$, $E \in \mathcal{E}_h(\mathcal{F}_h(\sigma_h))$. Using elementary properties of nodal basis functions

$$\|\varphi_h^{(a)}\|_{0,T} \lesssim h_T,\ a \in \mathcal{N}_h(T), \quad \|\varphi_h^{(a)}\|_{0,E} \lesssim h_E^{1/2},\ a \in \mathcal{N}_h(E), \tag{5.22}$$

as well as the following property of P_h^{SZ} (see, e.g., [39])

$$|(P_h^{SZ}v)(a)| \lesssim h_T^{-1}\, \|v\|_{0,T},\ a \in \mathcal{N}_h(T),\ v \in L^2(\Omega), \tag{5.23}$$

it follows that

$$\sum_{a\in\mathcal{N}_h(T)\cap\mathcal{C}_h} |(P_h^{SZ}(e-P_h^C e))(a)|\,\|\varphi_h^{(a)}\|_{0,T} \lesssim \tag{5.24a}$$

$$h_T \sum_{a\in\mathcal{N}_h(T)\cap\mathcal{C}_h} |(P_h^{SZ}(e-P_h^C e))(a)| \lesssim$$

$$h_T \sum_{a\in\mathcal{N}_h(T)\cap\mathcal{C}_h} h_{T^{(a)}}^{-1}\,\|e-P_h^C e\|_{0,T_a} \lesssim h_T \sum_{a\in\mathcal{N}_h(T)\cap\mathcal{C}_h} |e|_{1,\omega_h^{T^{(a)}}},$$

$$\|(P_h^{SZ}(e-P_h^C e))(a'_E)\varphi_h^{(a'_E)}\|_{0,E} = \tag{5.24b}$$

$$|(P_h^{SZ}(e-P_h^C e))(a'_E)|\,\|\varphi_h^{(a'_E)}\|_{0,E} \lesssim$$

$$h_E^{1/2}\,h_{T^{(a'_E)}}^{-1}\,\|e-P_h^C e\|_{0,T^{(a'_E)}} \lesssim h_E^{1/2}\,|e|_{1,\omega_h^{T^{(a'_E)}}},$$

where $T^{(a)}$ denotes the fixed element in ω_h^a which is used in the computation of the nodal coefficient $(P_h^{SZ}(e-P_h^C e))(a)$ (cf. (3.28)). Using (5.24a),(5.24b) in (5.21) yields

$$|F_h^{(\sigma)}(P_h^{SZ}(e-P_h^C e))| \lesssim \tag{5.25}$$
$$\sum_{T\in\mathcal{T}_h(\mathcal{F}_h(\sigma_h))} \eta_T^{(1)}\,|e|_{1,\tilde{\omega}^T} + \sum_{E\in\mathcal{E}_{\mathcal{F}_h(\sigma_h)}} \eta_E^{(1)}\,|e|_{1,\omega_h^{T^{(a'_E)}}},$$

where

$$\tilde{\omega}^T := \bigcup_{a\in\mathcal{N}_h(T)\cap\mathcal{C}_h} \omega_h^{T^{(a)}}.$$

Combining (5.20a),(5.20b), and (5.25), from (5.19) we deduce

$$|r(e-P_h^C e)| \lesssim \sum_{T\in\mathcal{T}_h(\mathcal{Z}_h)} \eta_T^{(1)}\,|e|_{1,\hat{\omega}^T} + \sum_{E\in\mathcal{E}_{\mathcal{Z}_h}} \eta_E^{(1)}\,|e|_{1,\hat{\omega}^E}, \tag{5.26}$$

where

$$\hat{\omega}^T := \begin{cases} \tilde{\omega}^T \cup \omega_h^T, & T\in\mathcal{T}_h(\mathcal{F}_h(\sigma_h)) \\ \omega_h^T, & \text{otherwise} \end{cases}, \quad \hat{\omega}^E := \begin{cases} \omega_h^{T^{(a'_E)}} \cup \omega_h^E, & E\in\mathcal{E}_{\mathcal{F}_h(\sigma_h)} \\ \omega_h^E, & \text{otherwise} \end{cases}.$$

Applying the Cauchy–Schwarz inequality in (5.26) and taking into account that ω_h^T and ω_h^E have a finite overlap, it follows that

$$|r(e - P_h^C e)| \lesssim \Big(\sum_{T \in \mathcal{T}_h(\mathcal{Z}_h)} (\eta_T^{(1)})^2 \Big)^{1/2} \Big(\sum_{T \in \mathcal{T}_h(\mathcal{Z}_h)} |e|^2_{1,\hat{\omega}^T} \Big)^{1/2}$$
$$+ \Big(\sum_{E \in \mathcal{E}_{\mathcal{Z}_h}} (\eta_E^{(1)})^2 \Big)^{1/2} \Big(\sum_{E \in \mathcal{E}_{\mathcal{Z}_h}} |e|^2_{1,\hat{\omega}^E} \Big)^{1/2} \lesssim$$
$$\Big(\sum_{T \in \mathcal{T}_h(\mathcal{Z}_h)} (\eta_T^{(1)})^2 + \sum_{E \in \mathcal{E}_{\mathcal{Z}_h} \cup \mathcal{E}_{\mathcal{F}_h(\sigma_h)}} (\eta_E^{(1)})^2 \Big)^{1/2} \cdot$$
$$\Big(\sum_{T \in \mathcal{T}_h(\mathcal{Z}_h)} |e|^2_{1,\hat{\omega}^T} + \sum_{E \in \mathcal{E}_{\mathcal{Z}_h}} |e|^2_{1,\hat{\omega}^E} \Big)^{1/2} \lesssim \eta_h^{(1)} \, |e|_{1,\Omega}.$$

Using the preceding inequality in (5.18) gives (5.17a).

For the proof of (5.17b) we set $e := p(y_h) - p_h$ and obtain

$$\|e\|^2_{1,\Omega} \lesssim a(e,e) = r(e - P_h^C e), \tag{5.27}$$

where the residual $r(\cdot)$ is given by

$$r(v) := (y^d - y_h, v)_{0,\Omega} - \langle \tilde{\mu}_h, v \rangle - a(p_h, v), \quad v \in V.$$

The representation (3.26b) of the extension $\tilde{\mu}_h$ yields

$$r(e - P_h^C e) = \sum_{T \in \mathcal{T}_h(\Omega)} (y^d - y_h, e - P_h^C e)_{0,T} - \sum_{E \in \mathcal{E}_h(\Omega)} (\nu_E \cdot [\nabla p_h]_E, e - P_h^C e)_{0,E}$$
$$- \langle \tilde{\mu}_h, e - P_h^C e \rangle = \sum_{T \in \mathcal{T}_h(\mathcal{I}_h)} (y^d - y_h, e - P_h^C e)_{0,T}$$
$$- \sum_{E \in \mathcal{E}_{\mathcal{I}_h}} (\nu_E \cdot [\nabla p_h]_E, e - P_h^C e)_{0,E} - F_h^{(\mu)}(P_h^{SZ}(e - P_h^C e)).$$

The terms on the right-hand side can be estimated from above in much the same way as before resulting in

$$|r(e - P_h^C e)| \lesssim \eta_h^{(2)} \, |e|_{1,\Omega}, \tag{5.28}$$

which together with (5.27) allows to conclude. □

Lemma 5.3. *Under the assumptions of* Theorem 5.1 *let* $y, y(u_h)$ *be the state and the auxiliary state and let* $p, p(y_h)$ *be the adjoint state and the auxiliary adjoint state. Further, let* $\eta_h^{(1)}$ *and* $\eta_h^{(2)}$ *be the components of the residual a posteriori error estimator according to* (5.4a) *and* (5.4b) *and let* $e_{h,\sigma}^{(1)}, e_{h,\mu}^{(2)}$ *be the consistency error terms given by* (5.7a),(5.7b)*. Then, it holds*

$$\|y - y(u_h)\|_{1,\Omega}^2 \lesssim \|e_{h,u}\|_{0,\Omega}^2 + (\eta_h^{(1)})^2 + e_{h,\sigma}^{(1)}, \tag{5.29a}$$

$$\|p - p(y_h)\|_{1,\Omega}^2 \lesssim \|e_{h,y}\|_{0,\Omega}^2 + (\eta_h^{(2)})^2 + e_{h,\mu}^{(2)}. \tag{5.29b}$$

Proof. Subtracting (5.15a) from (2.13a) yields

$$a(y - y(u_h), v) = (e_{h,u}, v)_{0,\Omega} + \langle \tilde{\sigma}_h - \sigma, v\rangle, \quad v \in V. \tag{5.30}$$

Choosing $v = y - y(u_h)$ and observing (2.4b), we get

$$\gamma \, \|y - y(u_h)\|_{1,\Omega}^2 \le a(y - y(u_h), y - y(u_h)) \; = \tag{5.31}$$
$$(e_{h,u}, y - y(u_h))_{0,\Omega} + \langle \tilde{\sigma}_h - \sigma, y_h - y(u_h)\rangle + e_{h,\sigma}^{(1)}.$$

The Cauchy–Schwarz inequality and Young's inequality give

$$|(e_{h,u}, y - y(u_h))_{0,\Omega}| \le \frac{\gamma}{4} \, \|y - y(u_h)\|_{0,\Omega}^2 + \frac{1}{\gamma} \, \|e_{h,u}\|_{0,\Omega}^2. \tag{5.32}$$

Moreover, if we choose $v = y_h - y(u_h)$ in (5.30), we obtain

$$\langle \tilde{\sigma}_h - \sigma, y_h - y(u_h)\rangle = (e_{h,u}, y(u_h) - y_h)_{0,\Omega} + a(y - y(u_h), y_h - y(u_h)).$$

Another application of the Cauchy–Schwarz inequality and Young's inequality yield

$$|\langle \tilde{\sigma}_h - \sigma, y_h - y(u_h)\rangle| \; \le \tag{5.33}$$
$$\frac{\gamma}{4} \, \|y - y(u_h)\|_{1,\Omega}^2 + \frac{2}{\gamma} \, \|y_h - y(u_h)\|_{1,\Omega}^2 + \frac{\gamma}{4} \, \|e_{h,u}\|_{0,\Omega}^2.$$

Using (5.32),(5.33) in (5.31) and setting

$$C_1 := \frac{\gamma^2 + 4}{2\gamma^2}, \quad C_2 := \frac{4}{\gamma^2}, \quad C_3 := \frac{2}{\gamma}, \tag{5.34}$$

it follows that

$$\|y - y(u_h)\|_{1,\Omega}^2 \le C_1 \, \|e_{h,u}\|_{0,\Omega}^2 + C_2 \, \|y_h - y(u_h)\|_{1,\Omega}^2 + C_3 \, e_{h,\sigma}^{(1)}. \tag{5.35}$$

The second term on the right-hand side in (5.35) can be estimated from above by (5.17a) which results in (5.29a).

The estimate (5.29b) can be established by using similar arguments. In fact, subtracting (5.15b) from (2.13c) yields

$$a(p - p(y_h), v) = -(e_{h,y}, v)_{0,\Omega} + \langle \tilde{\mu}_h - \mu, v\rangle, \quad v \in V. \tag{5.36}$$

Choosing $v = p - p(y_h)$ and $v = p_h - p(y_h)$, we obtain

$$\begin{aligned}\gamma \, \|p - p(y_h)\|^2_{1,\Omega} &\le a(p - p(y_h), p - p(y_h)) \\ &= (e_{h,y}, p(y_h) - p)_{0,\Omega} + \langle \tilde{\mu}_h - \mu, p_h - p(y_h)\rangle + e^{(2)}_{h,\mu}, \\ \langle \tilde{\mu}_h - \mu, p_h - p(y_h)\rangle &= (e_{h,y}, p_h - p(y_h))_{0,\Omega} + a(p - p(y_h), p_h - p(y_h)).\end{aligned}$$

An application of the Cauchy–Schwarz inequality and Young's inequality gives

$$\|p - p(y_h)\|^2_{1,\Omega} \le C_1 \, \|e_{h,y}\|^2_{0,\Omega} + C_2 \, \|p_h - p(y_h)\|^2_{1,\Omega} + C_3 \, e^{(2)}_{h,\mu}, \tag{5.37}$$

from which (5.29b) can be deduced in view of (5.17b). □

Lemma 5.4. *Under the assumptions of* Theorem 5.1 *let* $\eta_h, e^c_{h,rel}$, *and* $osc_{h,rel}$ *be the residual-type error estimator* (5.3), *the consistency error term* (5.6), *and the data oscillation* (5.10). *Then, it holds*

$$\|e_{h,u}\|^2_{0,\Omega} \lesssim \eta^2_h + e^c_{h,rel} + osc^2_{h,rel}. \tag{5.38}$$

Proof. Combining (2.13d) and (3.17d) we obtain

$$\begin{aligned}\|e_{h,u}\|^2_{0,\Omega} &= (e_{h,u}, u - u_h)_{0,\Omega} \\ &= (e_{h,u}, u^d - u^d_h)_{0,\Omega} + (e_{h,u}, (u - u^d) - (u_h - u^d_h))_{0,\Omega} \\ &= (e_{h,u}, u^d - u^d_h)_{0,\Omega} + \alpha^{-1} \, (e_{h,u}, p - p_h)_{0,\Omega}.\end{aligned} \tag{5.39}$$

The first term on the right-hand side in (5.39) can be estimated from above by

$$|(e_{h,u}, u^d - u^d_h)_{0,\Omega}| \le \frac{1}{4} \, \|e_{h,u}\|^2_{0,\Omega} + \mathrm{osc}^2_h(u^d). \tag{5.40}$$

The second term can be split according to

$$(e_{h,u}, p - p_h)_{0,\Omega} = (e_{h,u}, p - p(y_h))_{0,\Omega} + (e_{h,u}, p(y_h) - p_h)_{0,\Omega}. \tag{5.41}$$

For the estimation of the first term on the right-hand side in (5.41) we choose $v = p - p(y_h)$ in (5.30) which gives

$$a(y - y(u_h), p - p(y_h)) = (e_{h,u}, p - p(y_h))_{0,\Omega} + \langle \tilde{\sigma}_h - \sigma, p - p(y_h)\rangle. \tag{5.42}$$

On the other hand, choosing $v = y - y(u_h)$ in (5.36) yields

$$a(p - p(y_h), y - y(u_h)) = -(e_{h,y}, y - y(u_h))_{0,\Omega} + \langle \tilde{\mu}_h - \mu, y - y(u_h)\rangle. \tag{5.43}$$

Combining (5.42) and (5.43) and using the symmetry of $(\cdot,\cdot)$, it follows that

$$(e_{h,u}, p - p(y_h))_{0,\Omega} = -(e_{h,y}, y - y(u_h))_{0,\Omega} + \tag{5.44}$$
$$\langle \tilde{\sigma}_h - \sigma, p(y_h) - p_h\rangle + \langle \tilde{\mu}_h - \mu, y_h - y(u_h)\rangle + e^{(2)}_{h,\sigma} + e^{(1)}_{h,\mu}.$$

Now, choosing $v = p(y_h) - p_h$ in (5.30) and $v = y_h - y(u_h)$ in (5.36), for the second and third term on the right-hand side in (5.44) we find

$$\langle \tilde{\sigma}_h - \sigma, p(y_h) - p_h\rangle = -(e_{h,u}, p(y_h) - p_h)_{0,\Omega} + a(y - y(u_h), p(y_h) - p_h),$$
$$\langle \tilde{\mu}_h - \mu, y_h - y(u_h)\rangle = (e_{h,y}, y_h - y(u_h))_{0,\Omega} + a(p - p(y_h), y_h - y(u_h)),$$

and hence,

$$(e_{h,u}, p - p(y_h))_{0,\Omega} = -\|e_{h,y}\|^2_{0,\Omega} - (e_{h,u}, p(y_h) - p_h)_{0,\Omega} + \tag{5.45}$$
$$a(p - p(y_h), y_h - y(u_h)) + a(y - y(u_h), p(y_h) - p_h) + e^{(2)}_{h,\sigma} + e^{(1)}_{h,\mu}.$$

Using (5.45) in (5.41) results in

$$(e_{h,u}, p - p_h)_{0,\Omega} = a(p - p(y_h), y_h - y(u_h)) + \tag{5.46}$$
$$a(y - y(u_h), p(y_h) - p_h) - \|e_{h,y}\|^2_{0,\Omega} + e^{(2)}_{h,\sigma} + e^{(1)}_{h,\mu}.$$

For the first term on the right-hand side in (5.46), Young's inequality gives

$$|a(p - p(y_h), y_h - y(u_h))| \le \frac{\varepsilon}{2}\,\|y_h - y(u_h)\|^2_{1,\Omega} + \frac{1}{2\varepsilon}\,\|p - p(y_h)\|^2_{1,\Omega}.$$

Using (5.37) and choosing $\varepsilon = C_1/2$, we get

$$|a(p - p(y_h), y_h - y(u_h))| \le \tag{5.47}$$
$$\|e_{h,y}\|^2_{0,\Omega} + \frac{C_2}{C_1}\,\|p_h - p(y_h)\|^2_{1,\Omega} + \frac{C_1}{4}\,\|y_h - y(u_h)\|^2_{1,\Omega} + \frac{C_3}{C_1}\,e^{(2)}_{h,\mu}.$$

The second term on the right-hand side in (5.46) can be estimated from above similarly:

$$|a(y - y(u_h), p(y_h) - p_h)| \le \frac{\varepsilon}{2}\,\|p_h - p(y_h)\|^2_{1,\Omega} + \frac{1}{2\varepsilon}\,\|y - y(u_h)\|^2_{1,\Omega}.$$

Observing (5.35), we choose $\varepsilon = 2C_1/\alpha$ and obtain

$$|a(y - y(u_h), p(y_h) - p_h)| \le \frac{\alpha}{4} \|e_{h,u}\|^2_{0,\Omega} + \frac{\alpha C_2}{4C_1} \|y_h - y(u_h)\|^2_{1,\Omega} + \frac{C_1}{\alpha} \|p_h - p(y_h)\|^2_{1,\Omega} + \frac{\alpha C_3}{4C_1} e^{(1)}_{h,\sigma}. \tag{5.48}$$

Using (5.40) and (5.46)–(5.48) in (5.39), it follows that

$$\|e_{h,u}\|^2_{0,\Omega} \lesssim \|p_h - p(y_h)\|^2_{1,\Omega} + \|y_h - y(u_h)\|^2_{1,\Omega} + e^c_{h,rel} + \mathrm{osc}^2_{rel}. \tag{5.49}$$

The assertion (5.38) follows from (5.49) by taking (5.17a),(5.17b) from Lemma 5.2 into account. □

Proof of Theorem 5.1. In view of

$$e_{h,y} = y - y(u_h) + y(u_h) - y_h,$$
$$e_{h,p} = p - p(y_h) + p(y_h) - p_h,$$

the estimate (5.14) follows from the preceding Lemmas 5.2, 5.3, and 5.4. □

5.3 *Efficiency of the Error Estimator*

Theorem 5.5. *Let* (y, σ, u, p, μ) *and* $(y_h, \sigma_h, u_h, p_h, \mu_h)$ *be solutions of* (2.13a)–(2.13g) *and* (3.17a)–(3.17f) *and let* $\eta_h, e^c_{h,eff}, osc_{h,eff}$ *be the residual-type error estimator, the consistency error, and the data oscillations as given by* (5.3),(5.8), *and* (5.12). *Then, it holds*

$$\eta^2_h - e^c_{h,eff} - osc^2_{h,eff} \lesssim |\!|\!| e_h |\!|\!|^2. \tag{5.50}$$

The proof of Theorem 5.5 will be provided by the subsequent two lemmas taking into account the following well-known properties (cf., e.g., [44]) of the element bubble functions

$$\|q_h\|^2_{0,T} \lesssim (q_h, q_h\, b_T)_{0,T}, \quad q_h \in P_1(T), \tag{5.51a}$$
$$\|q_h\, b_T\|_{0,T} \lesssim \|q_h\|_{0,T}, \quad q_h \in P_1(T), \tag{5.51b}$$
$$h_T^{-1} \|q_h\|_{0,T} \lesssim |q_h\, b_T|_{1,T} \lesssim h_T^{-1} \|q_h\|_{0,T}, \quad q_h \in P_1(T), \tag{5.51c}$$

and of the edge bubble functions

$$\|q_h\|_{0,E}^2 \lesssim (q_h, q_h\, b_E)_{0,E}, \quad q_h \in P_1(E), \tag{5.52a}$$

$$\|q_h\, b_E\|_{0,E} \lesssim h_E^{1/2}\, \|q_h\|_{0,E}, \quad q_h \in P_1(E), \tag{5.52b}$$

$$h_E^{-1/2}\, \|q_h\|_{0,E} \lesssim |q_h\, b_E|_{1,\omega_E} \lesssim h_E^{-1/2}\, \|q_h\|_{0,E}, \quad q_h \in P_1(E). \tag{5.52c}$$

Lemma 5.6. *Under the assumptions of* Theorem 5.5 *let* $\eta_T^{(\nu)}, 1 \le \nu \le 2$, e_T^σ, e_T^μ, *and* $osc_T(f), osc_T(y^d)$ *be the element residuals* (5.5a), *the consistency error terms* (5.9a),(5.9b), *and the data oscillations* (5.13). *Then, for all* $T \in \mathcal{T}_h(\mathcal{Z}_h)$ *it holds*

$$\eta_T^{(1)} \lesssim \|e_{h,y}\|_{1,T} + h_T\, \|e_{h,u}\|_{0,T} + e_T^\sigma + osc_T(f), \tag{5.53}$$

whereas for all $T \in \mathcal{T}_h(\mathcal{I}_h)$ *we have*

$$\eta_T^{(2)} \lesssim \|e_{h,p}\|_{1,T} + h_T\, \|e_{h,y}\|_{0,T} + e_T^\mu + osc_T(y^d). \tag{5.54}$$

Proof. Setting $\psi_T^\sigma := (f_h + u_h)\, b_T$, using (5.51a), $\Delta y_h|_T = 0$, Green's formula, and $\psi_T^\sigma|_{\partial T} = 0$, we obtain

$$\begin{aligned} &h_T^2\, \|f_h + u_h\|_{0,T}^2 \lesssim h_T^2\, (f_h + u_h, \psi_T^\sigma)_{0,T} = \\ &h_T^2\, (f_h + u_h + \Delta y_h, \psi_T^\sigma)_{0,T} = h_T^2\, (f_h + u_h, \psi_T^\sigma)_{0,T} - h_T^2\, a(y_h, \psi_T^\sigma). \end{aligned} \tag{5.55}$$

On the other hand, since ψ_T^σ is an admissible test function in (2.13a), we have

$$a(y, \psi_T^\sigma) - (f + u, \psi_T^\sigma)_{0,T} + \langle \sigma, \psi_T^\sigma \rangle = 0. \tag{5.56}$$

Using (5.56) in (5.55), it follows that

$$\begin{aligned} &h_T^2\, \|f_h + u_h\|_{0,T}^2 \lesssim h_T^2\, \Big(a(y, \psi_T^\sigma) - (f + u, \psi_T^\sigma)_{0,T} + \langle \sigma, \psi_T^\sigma \rangle\Big) - \\ &h_T^2\, \Big(a(y_h, \psi_T^\sigma) - (f_h + u_h, \psi_T^\sigma)_{0,T}\Big) = \\ &h_T^2\, \Big(a(y - y_h, \psi_T^\sigma) - (f - f_h, \psi_T^\sigma)_{0,T} - (u - u_h, \psi_T^\sigma)_{0,T} + \langle \sigma, \psi_T^\sigma \rangle\Big) \le \\ &h_T^2\, \Big(|e_{h,y}|_{1,T}\, |\psi_T^\sigma|_{1,T} + \|e_{h,u}\|_{0,T}\, \|\psi_T^\sigma\|_{0,T} + e_T^\sigma\, |\psi_T^\sigma|_{1,T}\Big). \end{aligned} \tag{5.57}$$

In view of (5.51b) and (5.51c), it holds

$$\begin{aligned} &h_T^{-1}\, \|f_h + u_h\|_{0,T} \lesssim |\psi_T^\sigma|_{1,T} = |(f_h + u_h)\, b_T| \lesssim h_T^{-1}\, \|f_h + u_h\|_{0,T}, \\ &\|\psi_T^\sigma\|_{0,T} \lesssim \|f_h + u_h\|_{0,T}. \end{aligned} \tag{5.58}$$

Now, using (5.58) in (5.57), we get

$$h_T^2 \, \|f_h + u_h\|_{0,T}^2 \lesssim h_T \, \|f_h + u_h\|_{0,T} \left(\|e_{h,y}\|_{1,T} + h_T \, \|e_{h,u}\|_{0,T} + e_T^\sigma + \mathrm{osc}_T(f) \right).$$

Combining the preceding estimate with $\eta_T^{(1)} \leq h_T \, \|f_h + u_h\|_{0,T} + \mathrm{osc}_T(f)$ yields (5.53). The assertion (5.54) can be shown by similar arguments. □

Lemma 5.7. *Under the assumptions of* Lemma 5.6 *let* $\eta_E^{(\nu)}, 1 \leq \nu \leq 2$, *and* $e_{\omega_E}^\sigma, e_{\omega_E}^\mu$ *be the edge residuals and consistency error terms as given by* (5.5b) *and* (5.9c),(5.9d). *Further, for* $E = T_+ \cap T_-, T_\pm \in \mathcal{T}_h(\Omega)$ *let*

$$\eta_{\omega_E}^{(1)} := \eta_{T_+}^{(1)} + \eta_{T_-}^{(1)}, \quad osc_{\omega_E}(f) := osc_{T_+}(f) + osc_{T_-}(f), \tag{5.59a}$$

$$\eta_{\omega_E}^{(2)} := \eta_{T_+}^{(2)} + \eta_{T_-}^{(2)}, \quad osc_{\omega_E}(y^d) := osc_{T_+}(y^d) + osc_{T_-}(y^d). \tag{5.59b}$$

Then, for $E \in \mathcal{E}_{\mathcal{Z}_h}$ *we have*

$$\eta_E^{(1)} \lesssim \|e_{h,y}\|_{1,\omega_E} + h_E \, \|e_{h,u}\|_{0,\omega_E} + \eta_{\omega_E}^{(1)} + e_{\omega_E}^\sigma + osc_{\omega_E}(f), \tag{5.60}$$

whereas for all $E \in \mathcal{E}_{\mathcal{I}_h}$ *it holds*

$$\eta_E^{(2)} \lesssim \|e_{h,p}\|_{1,\omega_E} + h_E \, \|e_{h,y}\|_{0,\omega_E} + \eta_{\omega_E}^{(2)} + e_{\omega_E}^\mu + osc_{\omega_E}(y^d). \tag{5.61}$$

Proof. For $E \in \mathcal{E}_{\mathcal{Z}_h}$ we set $\psi_E^\sigma := \nu_E \cdot [\nabla y_h]_E \, b_E$. Then, (5.52a) implies

$$\begin{aligned} (\eta_E^{(1)})^2 &= h_E \, \|\nu_E \cdot [\nabla y_h]_E\|_{0,E}^2 \lesssim h_E \, (\nu_E \cdot [\nabla y_h]_E, \psi_E^\sigma)_{0,E} \\ &= h_E \, (\nu_{\partial T_+} \cdot \nabla y_h|_{\partial T_+}, \psi_E^\sigma)_{0,\partial T_+} + h_E \, (\nu_{\partial T_-} \cdot \nabla y_h|_{\partial T_-}, \psi_E^\sigma)_{0,\partial T_-}, \end{aligned} \tag{5.62}$$

where we have used that $\psi_E^\sigma|_{E'} = 0$, $E' \in \partial T_\pm \setminus \{E\}$. Further, Green's formula and $\Delta y_h|_{T_\pm} = 0$ yield

$$a_{T_\pm}(y_h, \psi_E^\sigma) = (\nabla y_h, \nabla \psi_E^\sigma)_{0,T_\pm} = (\nu_{\partial T_\pm} \cdot \nabla y_h|_{T_\pm}, \psi_E^\sigma)_{0,T_\pm}. \tag{5.63}$$

Using (5.63) in (5.62) gives

$$(\eta_E^{(1)})^2 \lesssim h_E \, a_{\omega_E}(y_h, \psi_E^\sigma). \tag{5.64}$$

Taking into account that ψ_E^σ is an admissible test function in (2.13a), we get

$$a_{\omega_E}(y, \psi_E^\sigma) - (f + u, \psi_E^\sigma)_{0,\omega_E} + \langle \sigma, \psi_E^\sigma \rangle = 0. \tag{5.65}$$

Combining (5.64) and (5.65), we obtain

$$
\begin{aligned}
(\eta_E^{(1)})^2 \lesssim\, & h_E\, a(y_h - y, \psi_E^\sigma) + h_E\, (f_h + u_h, \psi_E^\sigma)_{0,\Omega_E} + \\
& h_E\, (f - f_h, \psi_E^\sigma)_{0,\omega_E} + h_E\, (u - u_h, \psi_E^\sigma)_{0,\omega_E} - h_E\, \langle \sigma, \psi_E^\sigma \rangle \\
\le\, & h_E\, |y - y_h|_{1,\omega_E}\, |\psi_E^\sigma|_{1,\omega_E} + h_E\, \|\psi_E^\sigma\|_{0,\omega_E} \Big(\|f_h + u_h\|_{0,\omega_E} + \\
& \|u - u_h\|_{0,\omega_E} + \|f - f_h\|_{0,\omega_E} \Big) + h_E\, e_{\omega_E}^\sigma\, |\psi_E^\sigma|_{1,\omega_E}.
\end{aligned} \tag{5.66}
$$

Moreover, (5.52b) and (5.52c) imply

$$
\begin{aligned}
h_E^{-1/2}\, \|\nu_E \cdot [\nabla y_h]_E\|_{0,E} &\lesssim |\psi_E^\sigma|_{1,\omega_E} = |\nu_E \cdot [\nabla y_h]_E\, b_E|_{1,\omega_E} \\
&\lesssim h_E^{-1/2}\, \|\nu_E \cdot [\nabla y_h]_E\|_{0,E}, \\
\|\psi_E^\sigma\|_{0,\omega_E} &\lesssim h_E^{1/2}\, \|\nu_E \cdot [\nabla y_h]_E\|_{0,E}.
\end{aligned} \tag{5.67}
$$

Using (5.67) in (5.66) yields

$$
\eta_E^{(1)} \lesssim \|e_{h,y}\|_{1,\omega_E} + h_E\, \|e_{h,u}\|_{0,\omega_E} + h_E\, \|f_h + u_h\|_{0,\omega_E} + e_{\omega_E}^\sigma + \operatorname{osc}_{\omega_E}(f).
$$

Due to the shape regularity of the triangulation, for $E \in \mathcal{E}_h(T)$ we have $h_E \lesssim h_T \lesssim h_E$ and hence,

$$
\begin{aligned}
& h_E\, \|f_h + u_h\|_{0,\omega_E} \le h_E\, \|f_h + u_h\|_{0,T_+} + h_E\, \|f_h + u_h\|_{0,T_-} \lesssim \\
& h_{T_+}\, \|f_h + u_h\|_{0,T_+} + h_{T_-}\, \|f_h + u_h\|_{0,T_-} \lesssim \eta_{\omega_E}^{(1)}.
\end{aligned}
$$

The preceding two estimates result in (5.60). The assertion (5.61) can be verified by similar arguments. □

5.4 *Estimation of the Consistency Error*

In this subsection, we provide fully computable quantities for the approximation of the reliability and efficiency related consistency errors.

5.4.1 Approximation of Characteristic Functions

In this paragraph, following [15,17,28] in case of adaptive finite element approximations of control and/or state constrained optimally controlled second order elliptic boundary value problems, we provide approximations of the characteristic functions $\chi_{\mathcal{A}}$ and $\chi_{\mathcal{Z}}$ of the active set $\mathcal{A}$ and the zero set $\mathcal{Z}$ by means of the available finite

element solutions. Here and in the forthcoming paragraphs we will use realizations $\sigma_h', \mu_h' \in V_h$ of the discrete multipliers σ_h, μ_h with respect to the finite element spaces V_h according to

$$(\sigma_h', v_h)_{0,\Omega} = \langle\langle \sigma_h, v_h \rangle\rangle, \quad (\mu_h', v_h)_{0,\Omega} = \langle\langle \mu_h, v_h \rangle\rangle, \quad v_h \in V_h.$$

Moreover, we introduce a mesh function $\bar{h} \in S_h^{(1)}$ whose nodal values $\bar{h}(a)$ are given by averaging over local patches:

$$\bar{h}(a) := (\mathrm{card}(\omega_a))^{-1} \sum_{T \in \mathcal{T}_h(\omega_a)} h_T, \quad a \in \mathcal{N}_h(\bar{\Omega}).$$

The approximations of the characteristic functions are defined by means of

$$\chi_{h,\mathcal{A}}(a) := 1 - \frac{(\psi_h - y_h)(a)}{\gamma \bar{h}(a)^r + (\psi_h - y_h)(a)}, \quad a \in \mathcal{N}_h(\bar{\Omega}), \tag{5.68a}$$

$$\chi_{h,\mathcal{Z}}(a) := 1 - \frac{\sigma_h'(a)}{\gamma \bar{h}(a)^r + \sigma_h'(a)}, \quad a \in \mathcal{N}_h(\bar{\Omega}), \tag{5.68b}$$

where $0 < \gamma \leq 1$ and $r > 0$ are fixed. In case of uniform meshes with $\bar{h} \approx h = \max_{T \in \mathcal{T}_h(\bar{\Omega})} h_T$, the following result reflects the approximation properties of $\chi_{h,\mathcal{A}}$ and $\chi_{h,\mathcal{Z}}$.

Proposition 5.8. *For $0 \leq \varepsilon < 1$ and γ, r as in (5.68a),(5.68b) consider the partition*

$$\mathcal{I} \cap \mathcal{I}_h = \mathcal{I}_1 \cup \mathcal{I}_2,$$

where the sets $\mathcal{I}_\nu$, $1 \leq \nu \leq 2$, are given by

$$\mathcal{I}_1 := \{x \in \mathcal{I} \mid 0 < \psi_h(x) - y_h(x) \leq \gamma h^{\varepsilon r}\}, \; \mathcal{I}_2 := \{x \in \mathcal{I} \mid \psi_h(x) - y_h(x) > \gamma h^{\varepsilon r}\}.$$

Then, it holds

$$\|\chi_{\mathcal{A}} - \chi_{h,\mathcal{A}}\|_{0,\omega} \begin{cases} = 0 & , \omega \subset \mathcal{A} \cap \mathcal{A}_h \\ < \min(|\omega|^{1/2}, \gamma^{-1} h^{-r} \|\psi_h - y_h\|_{0,\Omega} & , \omega \subset \mathcal{A} \cap \mathcal{I}_h \\ = |\omega|^{1/2} & , \omega \subset \mathcal{I} \cap \mathcal{A}_h \\ < |\omega|^{1/2} & , \omega \subset \mathcal{I}_1 \\ < |\omega|^{1/2} h^{r(1-\varepsilon)} & , \omega \subset \mathcal{I}_2 \end{cases}.$$

Proof. Without loss of generality we may assume $h \leq 1$. For the proof we distinguish several cases.

Case 1 ($\omega \subset \mathcal{A} \cap \mathcal{A}_h$): Obviously, $\chi_{\mathcal{A}}|_\omega = \chi_{h,\mathcal{A}}|_\omega = 1$.
Case 2 ($\omega \subset \mathcal{A} \cap \mathcal{I}_h$): We have $\chi_{\mathcal{A}}|_\omega = 1$ and hence,

$$(\chi_{\mathcal{A}} - \chi_{h,\mathcal{A}})|_\omega = \frac{(\psi_h - y_h)|_\omega}{\gamma h^r + (\psi_h - y_h)|_\omega}.$$

Since $(\psi_h - y_h)|_\omega > 0$ and $\gamma h^r > 0$, it follows that

$$(\chi_{\mathcal{A}} - \chi_{h,\mathcal{A}})|_\omega < \gamma^{-1} h^{-r} (\psi_h - y_h)|_\omega \quad \text{and} \quad (\chi_{\mathcal{A}} - \chi_{h,\mathcal{A}})|_\omega < 1,$$

which allows to conclude.
Case 3 ($\omega \subset \mathcal{I} \cap \mathcal{A}_h$): The assertion follows readily from $\chi_{\mathcal{A}}|_\omega = 0$ and $\chi_{h,\mathcal{A}}|_\omega = 1$.
Case 4 ($\omega \subset \mathcal{I} \cap \mathcal{I}_h$): We have $\chi_{\mathcal{A}}|_\omega = 0$ and

$$(\chi_{\mathcal{A}} - \chi_{h,\mathcal{A}})|_\omega = \frac{\gamma h^r}{\gamma h^r + (\psi_h - y_h)|_\omega}.$$

For $\omega \subset \mathcal{I}_1$ this implies $(\chi_{\mathcal{A}} - \chi_{h,\mathcal{A}})|_\omega < 1$, and we conclude. On the other hand, for $\omega \subset \mathcal{I}_2$, taking $h \leq 1$ into account, we find

$$(\chi_{\mathcal{A}} - \chi_{h,\mathcal{A}})|_\omega < \min(1, h^{r(1-\varepsilon)}) = h^{r(1-\varepsilon)},$$

which proves the assertion. □

5.4.2 Approximation of the Continuous Active/Inactive Sets

Based on the approximations $\chi_{h,\mathcal{A}}, \chi_{h,\mathcal{Z}}$ of the characteristic functions of the continuous sets $\mathcal{A}$ and $\mathcal{Z}$, we derive approximations of the continuous (strongly) active, biactive, inactive, and zero sets. To this end, for $0 < \kappa \leq 1$ and $0 < r' \leq r$ we first define nodal sets $\bar{A}_h, \bar{I}_h, \bar{C}_h, \bar{Z}_h$, and $\bar{B}_h$ as approximations of their continuous counterparts according to

$$\begin{aligned}
\bar{A}_h &:= \{a \in \mathcal{N}_h(\bar{\Omega}) \mid \chi_{h,\mathcal{A}}(a) \geq 1 - \kappa \bar{h}(a)^{r'}\}, \quad \bar{I}_h := \mathcal{N}_h(\bar{\Omega}) \setminus \bar{A}_h, \\
\bar{C}_h &:= \Big(\mathcal{N}_h(\Omega) \setminus \{a \in \mathcal{N}_h(\Omega) \mid \chi_{h,\mathcal{Z}}(a) \geq 1 - \kappa \bar{h}(a)^{r'}\}\Big) \cap \bar{A}_h, \\
\bar{Z}_h &:= \mathcal{N}_h(\Omega) \setminus \bar{C}_h, \quad \bar{B}_h := \bar{A}_h \cap \bar{Z}_h.
\end{aligned}$$

These sets constitute a suitable basis for the specification of approximations $\bar{\mathcal{A}}_h$ of $\mathcal{A}$, $\bar{\mathcal{I}}_h$ of $\mathcal{I}$, $\bar{\mathcal{C}}_h$ of $\mathcal{C}$, and $\bar{\mathcal{Z}}_h$ of $\mathcal{Z}$ by means of

$$\bar{\mathcal{A}}_h := \bigcup \{T \in \mathcal{T}_h(\Omega) \mid T \in \bar{\mathcal{A}}_h^T\}, \ \bar{\mathcal{A}}_h^T := \{T \in \mathcal{T}_h(\Omega) \mid \mathcal{N}_h(T) \subseteq \bar{A}_h\}, \tag{5.69a}$$

$$\bar{\mathcal{I}}_h := \bigcup \{T \in \mathcal{T}_h(\Omega) \mid T \in \bar{\mathcal{I}}_h^T \cup \bar{\mathcal{F}}_{y_h}^T\}, \tag{5.69b}$$

$$\bar{\mathcal{I}}_h^T := \{T \in \mathcal{T}_h(\Omega) \mid \mathcal{N}_h(T) \subseteq \bar{I}_h\},\ \bar{\mathcal{F}}_{y_h}^T := \mathcal{T}_h(\Omega) \setminus (\bar{\mathcal{A}}_h^T \cup \bar{\mathcal{I}}_h^T),$$

$$\bar{\mathcal{C}}_h := \bigcup \{T \in \mathcal{T}_h(\Omega) \mid T \in \bar{\mathcal{C}}_h^T\},\ \bar{\mathcal{C}}_h^T := \{T \in \mathcal{T}_h(\Omega) \mid \mathcal{N}_h(T) \subseteq \bar{C}_h\} \cup \tag{5.69c}$$

$$\{T \in \mathcal{T}_h(\Omega) \mid T \cap \Gamma \neq \emptyset \wedge \mathcal{N}_h(T) \cap \mathcal{N}_h(\Omega) \neq \emptyset \wedge T \subseteq \bar{\mathcal{A}}_h^T\},$$

$$\bar{\mathcal{Z}}_h := \bigcup \{T \in \mathcal{T}_h(\Omega) \mid T \in \bar{\mathcal{Z}}_h^T \cup \bar{\mathcal{F}}_{\sigma_h}^T\}, \tag{5.69d}$$

$$\bar{\mathcal{Z}}_h^T := \{T \in \mathcal{T}_h(\Omega) \mid \mathcal{N}_h(T) \subseteq \bar{Z}_h \cup \mathcal{N}_h(\Gamma)\},\ \bar{\mathcal{F}}_{\sigma_h}^T := \mathcal{T}_h(\Omega) \setminus (\bar{\mathcal{C}}_h^T \cup \bar{\mathcal{Z}}_h^T).$$

The biactive set $\mathcal{B}$ and the free boundaries $\mathcal{F}(y)$ and $\mathcal{F}(\sigma)$ are approximated by

$$\bar{\mathcal{B}}_h := \bigcup \{T \in \mathcal{T}_h(\Omega) \mid T \in \bar{\mathcal{B}}_h^T\},\ \bar{\mathcal{B}}_h^T := \bar{\mathcal{A}}_h^T \setminus \bar{\mathcal{C}}_h^T, \tag{5.69e}$$

$$\bar{\mathcal{F}}_{y_h} := \bigcup \{T \in \mathcal{T}_h(\Omega) \mid T \in \bar{\mathcal{F}}_{y_h}^T\}, \tag{5.69f}$$

$$\bar{\mathcal{F}}_{\sigma_h} := \bigcup \{T \in \mathcal{T}_h(\Omega) \mid T \in \bar{\mathcal{F}}_{\sigma_h}^T\}. \tag{5.69g}$$

In the documentation of the numerical results in the following Sect. 6, we will measure the quality of the approximation of the active set $\mathcal{A}$ and the strongly active set $\mathcal{C}$ by the a posteriori quantities

$$e_{\ell,\mathcal{A}}^{dva} := \|\chi_{\mathcal{A}_\ell} - \chi_{\bar{\mathcal{A}}_\ell}\|_{L^1(\Omega)},\quad e_{\ell,\mathcal{C}}^{dva} := \|\chi_{\mathcal{C}_\ell} - \chi_{\bar{\mathcal{C}}_\ell}\|_{L^1(\Omega)}, \tag{5.70}$$

where the upper index 'dva' stands for 'discrete versus approximate', and compare them with the quantities

$$e_{\ell,\mathcal{A}}^{evd} := \|\chi_{\mathcal{A}} - \chi_{\mathcal{A}_\ell}\|_{L^1(\Omega)},\quad e_{\ell,\mathcal{C}}^{evd} := \|\chi_{\mathcal{C}} - \chi_{\mathcal{C}_\ell}\|_{L^1(\Omega)}, \tag{5.71a}$$

$$e_{\ell,\mathcal{A}}^{eva} := \|\chi_{\mathcal{A}} - \chi_{\bar{\mathcal{A}}_\ell}\|_{L^1(\Omega)},\quad e_{\ell,\mathcal{C}}^{eva} := \|\chi_{\mathcal{C}} - \chi_{\bar{\mathcal{C}}_\ell}\|_{L^1(\Omega)}. \tag{5.71b}$$

Here, the upper indices 'evd' and 'eva' mean 'exact versus discrete' and 'exact versus approximate'. Obviously, these latter quantities are only available, if the exact solution is known.

5.4.3 Approximation of the Continuous States and Multipliers

We derive approximations of the state y and the adjoint state p as well as various approximations of the multipliers σ and μ in terms of the approximations of the continuous active/biactive, strongly active, inactive, zero nodal points (sets) and free boundaries provided in the previous paragraph 5.4.2. Motivated by superconvergence results through local averaging (cf., e.g., [3]), we define approximations

$\bar{y}_h \in V_h$ of y and $\bar{p}_h \in V_h$ of p according to

$$\bar{y}_h(a) := \begin{cases} \mathrm{card}(\mathcal{N}_h(\omega_a))^{-1} \sum\limits_{a' \in \mathcal{N}_h(\omega_a)} y_h(a') \,, & a \in \bar{I}_h \\ \psi_h(a) \,, & a \in \bar{A}_h \end{cases}, \tag{5.72a}$$

$$\bar{p}_h(a) := \begin{cases} \mathrm{card}(\mathcal{N}_h(\omega_a))^{-1} \sum\limits_{a' \in \mathcal{N}_h(\omega_a)} p_h(a') \,, & a \in \bar{Z}_h \\ 0 \,, & a \in \bar{C}_h \end{cases}. \tag{5.72b}$$

Likewise, we define approximations σ_h'' and μ_h'' of σ and μ by means of

$$\sigma_h''(a) := \begin{cases} \mathrm{card}(\mathcal{N}_h(\omega_a))^{-1} \sum\limits_{a' \in \mathcal{N}_h(\omega_a)} \sigma_h'(a') \,, & a \in \bar{C}_h \\ 0 \,, & a \in \bar{Z}_h \end{cases}, \tag{5.73a}$$

$$\mu_h''(a) := \begin{cases} \mathrm{card}(\mathcal{N}_h(\omega_a))^{-1} \sum\limits_{a' \in \mathcal{N}_h(\omega_a)} \mu_h'(a') \,, & a \in \bar{I}_h \\ 0 \,, & a \in \bar{A}_h \end{cases}. \tag{5.73b}$$

Remark 5.9. The functions $\bar{y}_h, \bar{p}_h$ will replace y, p in the approximation of the consistency error $e^c_{h,rel}$, whereas σ_h'', μ_h'' will be used in the approximation of $e^c_{h,eff}$ and in a further form of the approximation of $e^c_{h,rel}$ (cf. paragraph 5.4.4).

For the approximation of the multipliers σ, μ in the consistency error $e^c_{h,rel}$ we will use an alternative approximation which relies on the structural properties of the multipliers. If the sets $\mathcal{C}$ and $\mathcal{A}$ are the union of a finite number of connected pairwise disjoint Lipschitz sets, for any $v \in V$ Proposition 2.1 guarantees the existence of sets $\tilde{\mathcal{C}}, \tilde{\mathcal{A}}$ and functions $v^{ext}_{\mathcal{C}}, v^{ext}_{\mathcal{A}} \in V$ such that $\mathcal{C} \subseteq \tilde{\mathcal{C}} \subseteq \Omega \,, \mathcal{A} \subseteq \tilde{\mathcal{A}} \subseteq \Omega$ and

$$\langle \sigma, v \rangle = \langle \sigma, v^{ext}_{\mathcal{C}} \rangle = (f + u, v^{ext}_{\mathcal{C}})_{0,\tilde{\mathcal{C}}} - (\nabla y, \nabla v^{ext}_{\mathcal{C}})_{0,\tilde{\mathcal{C}}},$$

$$\langle \mu, v \rangle = \langle \mu, v^{ext}_{\mathcal{A}} \rangle = (y^d - y, v^{ext}_{\mathcal{A}})_{0,\tilde{\mathcal{A}}} - (\nabla p, \nabla v^{ext}_{\mathcal{A}})_{0,\tilde{\mathcal{A}}}.$$

Employing the structural information provided in Proposition 2.16, we obtain

$$\langle \sigma, v \rangle \; = \tag{5.74a}$$

$$\Big((f + u^d \,, v)_{0,\mathcal{C}} - (\nabla \psi, \nabla v)_{0,\mathcal{C}} \Big) - \Big((\Delta \psi, v^{ext}_{\mathcal{C}})_{0,(\tilde{\mathcal{C}} \setminus \mathcal{C}) \cap \mathcal{B}} + (\nabla \psi, \nabla v^{ext}_{\mathcal{C}})_{0,(\tilde{\mathcal{C}} \setminus \mathcal{C}) \cap \mathcal{B}} \Big)$$

$$+ \; \Big((f + u^d + \alpha^{-1} p, v^{ext}_{\mathcal{C}})_{0,(\tilde{\mathcal{C}} \setminus \mathcal{C}) \cap \mathcal{I}} - (\nabla y, \nabla v^{ext}_{\mathcal{C}})_{0,(\tilde{\mathcal{C}} \setminus \mathcal{C}) \cap \mathcal{I}} \Big),$$

$$\langle \mu, v \rangle = (y^d - \psi, v)_{0,\mathcal{A}} \tag{5.74b}$$

$$+ \; \alpha \, (\nabla(\Delta \psi + f + u^d), \nabla v)_{0,\mathcal{B}} + \Big((y^d - y, v^{ext}_{\mathcal{A}})_{0,\tilde{\mathcal{A}} \setminus \mathcal{A}} - (\nabla p, \nabla v^{ext}_{\mathcal{A}})_{0,\tilde{\mathcal{A}} \setminus \mathcal{A}} \Big).$$

In order to provide a fully computable approximation, we replace the unknown sets $\mathcal{C}, \mathcal{B}, \mathcal{A}, \mathcal{I}$, and the unknown functions y, p by their previously defined approximations $\bar{\mathcal{C}}_h, \bar{\mathcal{B}}_h, \bar{\mathcal{A}}_h, \bar{\mathcal{I}}_h$, and $\bar{y}_h$, $\bar{p}_h$. Moreover, $\tilde{\mathcal{C}}$, $\tilde{\mathcal{A}}$ are chosen according to

$$\tilde{\mathcal{C}} := \bar{\mathcal{C}}_h \cup \bar{\mathcal{F}}_{\sigma_h}, \quad \tilde{\mathcal{A}} := \bar{\mathcal{A}}_h \cup \bar{\mathcal{F}}_{y_h}, \tag{5.75}$$

whereas $v_{\mathcal{C}}^{ext}, v_{\mathcal{A}}^{ext}$ are approximated by

$$v_{\bar{\mathcal{C}}_h}^{ext} := I_{\bar{\mathcal{C}}_h}(v_h), \quad v_{\bar{\mathcal{A}}_h}^{ext} := I_{\bar{\mathcal{A}}_h}(v_h), \quad v_h \in V_h. \tag{5.76}$$

Here, I_{D_h}, $D_h \subseteq \mathcal{N}_h(\Omega)$, is the operator from (3.19).

Using the previous approximations in (5.74) and assuming sufficient regularity of the data in $\bar{\mathcal{B}}_h$, we obtain the following approximations of the action of σ, μ on functions in V_h:

$$\begin{aligned}
\langle \sigma, v_h \rangle \approx \langle \bar{\sigma}_h^{(1)}, v_h \rangle = & \sum_{T \in \mathcal{T}_h(\bar{\mathcal{C}}_h)} \Big((f + u^d, v_h)_{0,T} - (\nabla \psi, \nabla v_h)_{0,T} \Big) \qquad (5.77a) \\
& - \sum_{T \in \mathcal{T}_h(\bar{\mathcal{F}}_{\sigma_h} \cap \bar{\mathcal{B}}_h)} \Big((\Delta \psi, I_{\bar{\mathcal{C}}_h}(v_h))_{0,T} + (\nabla \psi, \nabla I_{\bar{\mathcal{C}}_h} v_h)_{0,T} \Big) \\
& + \sum_{T \in \mathcal{T}_h(\bar{\mathcal{F}}_{\sigma_h} \cap \bar{\mathcal{I}}_h)} \Big((f + u^d + \alpha^{-1} \bar{p}_h, I_{\bar{\mathcal{C}}_h}(v_h))_{0,T} - (\nabla \bar{y}_h, \nabla I_{\bar{\mathcal{C}}_h}(v_h))_{0,T} \Big), \\
\langle \mu, v_h \rangle \approx \langle \bar{\mu}_h^{(1)}, v_h \rangle = & \sum_{T \in \mathcal{T}_h(\bar{\mathcal{A}}_h)} (y^d - \psi, v_h)_{0,T} \qquad (5.77b) \\
& + \alpha \sum_{T \in \mathcal{T}_h(\bar{\mathcal{B}}_h)} (\nabla(\Delta \psi + f + u^d), \nabla v_h)_{0,T} \\
& + \sum_{T \in \mathcal{T}_h(\bar{\mathcal{F}}_{y_h})} \Big((y^d - \bar{y}_h, I_{\bar{\mathcal{A}}_h}(v_h))_{0,T} - (\nabla \bar{p}_h, \nabla I_{\bar{\mathcal{A}}_h}(v_h))_{0,T} \Big).
\end{aligned}$$

As far as the regularity of the data is concerned, in the proof of Proposition 2.16 we have seen that $\Delta \psi \in L^2(\mathcal{B})$ and $\Delta \psi + f + u^d \in H^1(\mathcal{B})$. If $\bar{\mathcal{B}}_h \subset \mathcal{B}$ or else $\Delta \psi \in L^2(\bar{\mathcal{B}}_h)$ and $\Delta \psi + f + u^d \in L^2(\bar{\mathcal{B}}_h)$ hold true, (5.77a) and (5.77b) are well defined. Otherwise, employing the values of $\bar{y}_h$ and $\bar{p}_h$ in $\bar{\mathcal{B}}_h$, we can use the following simplification of the approximations of the action of σ, μ on functions in V_h:

$$\langle \sigma, v_h \rangle \approx \langle \bar{\sigma}_h^{(2)}, v_h \rangle = \sum_{T \in \mathcal{T}_h(\bar{\mathcal{C}}_h)} \Big((f + u^d, v_h)_{0,T} - (\nabla \psi, \nabla v_h)_{0,T} \Big) \tag{5.78a}$$

$$- \sum_{T \in \mathcal{T}_h(\bar{\mathcal{F}}_{\sigma_h} \cap \bar{\mathcal{B}}_h)} (\nabla \psi, \nabla I_{\bar{C}_h}(v_h))_{0,T}$$

$$+ \sum_{T \in \mathcal{T}_h(\bar{\mathcal{F}}_{\sigma_h} \cap \bar{\mathcal{I}}_h)} \Big((f + u^d + \alpha^{-1} \bar{p}_h, I_{\bar{C}_h}(v_h))_{0,T} - (\nabla \bar{y}_h, \nabla I_{\bar{C}_h}(v_h))_{0,T} \Big),$$

$$\langle \mu, v_h \rangle \approx \langle \bar{\mu}_h^{(2)}, v_h \rangle = \sum_{T \in \mathcal{T}_h(\bar{\mathcal{A}}_h)} (y^d - \psi, v_h)_{0,T} - \sum_{T \in \mathcal{T}_h(\bar{\mathcal{B}}_h)} (\nabla \bar{p}_h, \nabla v_h)_{0,T} \tag{5.78b}$$

$$+ \sum_{T \in \mathcal{T}_h(\bar{\mathcal{F}}_{y_h})} \Big((y^d - \bar{y}_h, I_{\bar{A}_h}(v_h))_{0,T} - (\nabla \bar{p}_h, \nabla I_{\bar{A}_h}(v_h))_{0,T} \Big).$$

5.4.4 Approximation of the Consistency Errors

For the consistency error $e^c_{h,rel}$ we will use three different types of approximations

$$e^c_{h,rel} \approx \bar{e}^{c,(k)}_{h,rel} := \bar{e}^{1,(k)}_{h,\sigma} + \bar{e}^{2,(k)}_{h,\sigma} + \bar{e}^{1,(k)}_{h,\mu} + \bar{e}^{2,(k)}_{h,\mu}, \quad 1 \le k \le 3. \tag{5.79}$$

For the first two approximations $\bar{e}^{c,(k)}_{h,rel}$, $1 \le k \le 2$, we use the approximation of the multipliers by (5.77) and (5.78):

$$\bar{e}^{1,(k)}_{h,\sigma} := \langle \tilde{\sigma}_h - \bar{\sigma}_h^{(k)}, \bar{y}_h - y_h \rangle, \quad \bar{e}^{2,(k)}_{h,\sigma} := \langle \tilde{\sigma}_h - \bar{\sigma}_h^{(k)}, p_h - \bar{p}_h \rangle, \tag{5.80a}$$

$$\bar{e}^{1,(k)}_{h,\mu} := \langle \tilde{\mu}_h - \bar{\mu}_h^{(k)}, \bar{y}_h - y_h \rangle, \quad \bar{e}^{2,(k)}_{h,\mu} := \langle \tilde{\mu}_h - \bar{\mu}_h^{(k)}, \bar{p}_h - p_h \rangle. \tag{5.80b}$$

The third approximation $\bar{e}^{c,(3)}_{h,rel}$ is obtained by using the approximation of the multipliers by local averaging (cf. (5.73)):

$$\bar{e}^{1,(3)}_{h,\sigma} := \langle \tilde{\sigma}_h - \sigma_h'', \bar{y}_h - y_h \rangle, \quad \bar{e}^{2,(3)}_{h,\sigma} := \langle \tilde{\sigma}_h - \sigma_h'', p_h - \bar{p}_h \rangle, \tag{5.81a}$$

$$\bar{e}^{1,(3)}_{h,\mu} := \langle \tilde{\mu}_h - \mu_h'', \bar{y}_h - y_h \rangle, \quad \bar{e}^{2,(3)}_{h,\mu} := \langle \tilde{\mu}_h - \mu_h'', \bar{p}_h - p_h \rangle. \tag{5.81b}$$

Further, we compute upper bounds $\bar{e}^{c,(k)}_{h,rel}$, $1 \le k \le 3$, according to

$$\bar{e}^{c,(k)}_{h,rel} \le \bar{E}^{c,(k)}_{h,rel} := \bar{E}^{1,(k)}_{h,\sigma} + \bar{E}^{2,(k)}_{h,\sigma} + \bar{E}^{1,(k)}_{h,\mu} + \bar{E}^{2,(k)}_{h,\mu}, \quad 1 \le k \le 3, \tag{5.82}$$

where $\bar{E}^{\nu,(k)}_{h,\sigma}$, $\bar{E}^{\nu,(k)}_{h,\mu}$, $1 \le \nu \le 2$, are given by summing up the absolute values of the elementwise contributions of $\bar{e}^{\nu,(k)}_{h,\sigma}$, $\bar{e}^{\nu,(k)}_{h,\mu}$, $1 \le \nu \le 2$.

For the approximation of the consistency error $e^c_{h,eff}$ we use the approximation of the multipliers by local averaging as given by (5.73):

$$\bar{e}^c_{h,eff} \lesssim \bar{E}^c_{h,eff} := \sum_{T\in\mathcal{T}_h(\mathcal{Z}_h)} h_T^2 \, \|\sigma_h''\|^2_{0,T} + \sum_{T\in\mathcal{T}_h(\mathcal{I}_h)} h_T^2 \, \|\mu_h''\|^2_{0,T}. \tag{5.83}$$

6 Numerical Results

In this section, we present numerical results for problems with and without strict complementarity illustrating the performance of the suggested finite element approximation. We note that for adaptively refined meshes it is appropriate to measure the decay in the error *err* in terms of the degrees of freedom (DOF) provided by the finite element mesh. In particular, if there exists a real number $\tau > 0$ such that $err = O(DOF^{-\tau})$, then τ is said to be the convergence rate of the error with respect to the degrees of freedom. In the numerical experiments, we are dealing with a hierarchy $\{\mathcal{T}_{h_n}(\Omega)\}_{n\in\mathbb{N}}$ of nested simplicial meshes with associated degrees of freedom $DOF(n)$. Denoting by $err(n), n \in \mathbb{N}$, the error with respect to the mesh $\mathcal{T}_{h_n}(\Omega)$, we refer to

$$\tau_n := \frac{\log(err(n-1)/err(n))}{\log(DOF(n)/DOF(n-1))}, \quad n \in \mathbb{N}, \tag{6.1}$$

as the experimental convergence rate in terms of the degrees of freedom. On a double logarithmic scale, the numbers τ_n correspond to the negative slopes of the lines connecting $\log(err(n-1))$ and $\log(err(n))$. In the subsequent numerical examples, we will compare these lines both for uniform refinement and adaptive refinement. In the regular case, we expect the slopes to be approximately the same, whereas for less regular solutions the slope for adaptive refinement is expected to be larger than in case of uniform refinement.

Example 6.1. We consider $A = -\Delta$ on the L-shaped domain $\Omega = (-2,2)^2 \setminus ([0,2]\times[-2,0])$. In polar coordinates, given

$$y^*(r,\varphi) = -\gamma(r)\, r^{2/3} \sin(\frac{2}{3}\varphi),$$

$$\sigma^*(r) = \begin{cases} 1\,, & r \geq \bar{r} := 0.5 \\ 0\,, & \text{otherwise} \end{cases}, \quad u^*(r,\varphi) = y^*(r,\varphi),$$

where

$$\gamma(r) = \begin{cases} 0\,, & r \geq \bar{r} \\ 16r^3 - 12r^2 + 1\,, & \text{otherwise} \end{cases},$$

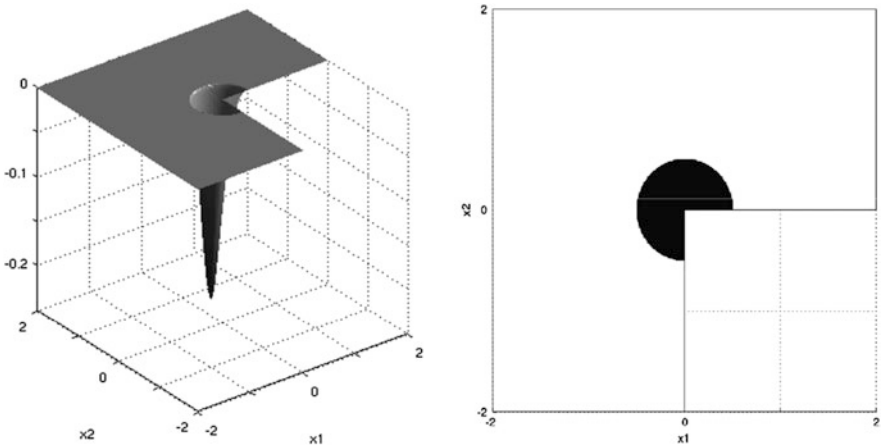


Fig. 1 Example 6.1. Optimal state y^* (*left*) and inactive set $\mathcal{I}^*$, marked in *black* (*right*)

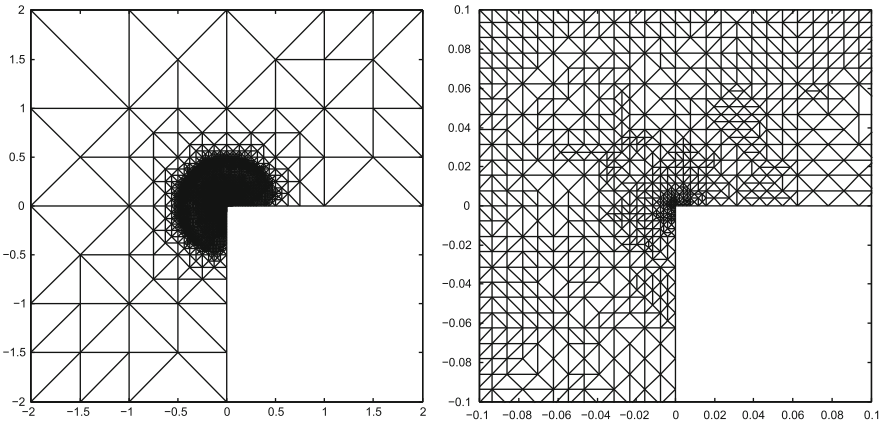


Fig. 2 Example 6.1. Final mesh (*left*) and zoom into the vicinity of the singularity at the origin (*right*)

it can be easily verified that the triple (y^*, σ^*, u^*) with the adjoint state $p^* = y^*$ and the multiplier $\mu^* = \sigma^*$ is an S-stationary point of (2.5) with respect to the data

$$y^d = \mu^* - \Delta p^* + y^*, \quad u^d = 0,$$
$$f = \sigma^* - \Delta y^* - p^*, \quad \alpha = 1, \quad \psi = 0.$$

Further, we have $\mathcal{I}^* = \{(r, \varphi) \mid r \in (0, \bar{r}), \varphi \in (0, 3\pi/2)\}$, $\mathcal{Z}^* = \mathcal{I}^*$, and hence, $\mathcal{B}^* = \emptyset$. The state y^* and the inactive set $\mathcal{I}^*$ are displayed in Fig. 1. The adaptively generated final mesh with 33468 DOFs and a zoom into the vicinity of the singularity of the state at the origin are shown in Fig. 2.

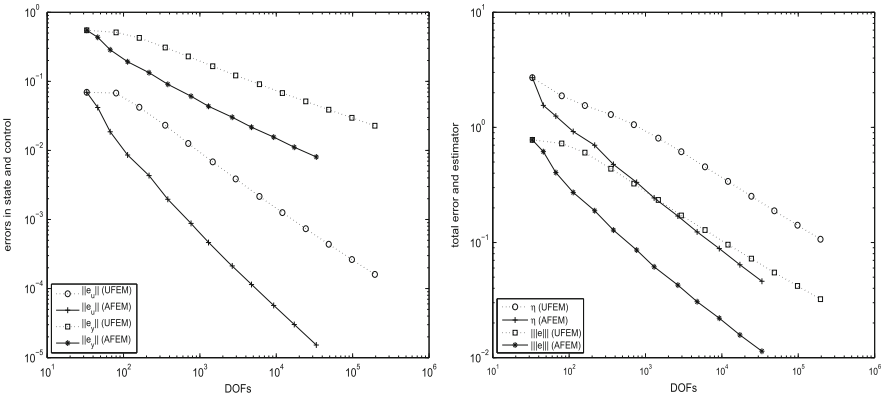


Fig. 3 Example 6.1. Convergence history: Decrease of the errors in the state $\|e_{h,y}\|_{1,\Omega}$ and in the control $\|e_{h,u}\|_{0,\Omega}$ as a function of the DOFs on a logarithmic scale (for uniform (UFEM) and adaptive (AFEM) refinement (*left*). Decrease of the estimator η_h and the total error $|||e_h|||$ as a function of the DOFs on a logarithmic scale (for uniform (UFEM) and adaptive (AFEM) refinement (*right*)

Table 1 Example 6.1: Experimental convergence rates (uniform and adaptive refinement)

	$\|e_{h,y}\|_{1,\Omega}$		$\|e_{h,p}\|_{1,\Omega}$		$\|e_{h,u}\|_{0,\Omega}$		$\|\|\|e_h\|\|\|$	
n	Unif.	Adapt.	Unif.	Adapt.	Unif.	Adapt.	Unif.	Adapt.
3	0.26	1.11	0.26	1.11	0.68	2.15	0.28	1.15
4	0.41	0.76	0.41	0.76	0.76	1.48	0.42	0.78
5	0.43	0.56	0.43	0.56	0.88	1.06	0.44	0.57
6	0.44	0.68	0.44	0.68	0.83	1.40	0.45	0.69
7	0.45	0.57	0.45	0.57	0.82	1.15	0.45	0.57
8	0.41	0.64	0.41	0.64	0.82	1.21	0.41	0.64
9	0.42	0.51	0.42	0.51	0.78	1.09	0.43	0.51
10	0.40	0.57	0.40	0.57	0.76	1.07	0.40	0.57
11	0.40	0.49	0.40	0.49	0.75	1.04	0.40	0.50
12	0.39	0.54	0.39	0.54	0.73	1.02	0.39	0.54
13	0.38	0.49	0.38	0.49	0.72	1.03	0.38	0.49

The convergence history is documented in Fig. 3 (left) which shows the decrease of the errors in the state $\|e_{h,y}\|_{1,\Omega}$ and in the control $\|e_{h,u}\|_{0,\Omega}$ as a function of the DOFs on a logarithmic scale both for uniform refinement (UFEM) and for adaptive refinement (AFEM). Likewise, Fig. 3 (right) shows the decrease of the total error $|||e_h|||$ and of the estimator η_h as a function of the DOFs on a logarithmic scale, again both for uniform refinement (UFEM) and for adaptive refinement (AFEM).

Table 1 contains the computed experimental convergence rates (cf. 6.1) for the approximation of the state, the adjoint state, the control, and the total error in case

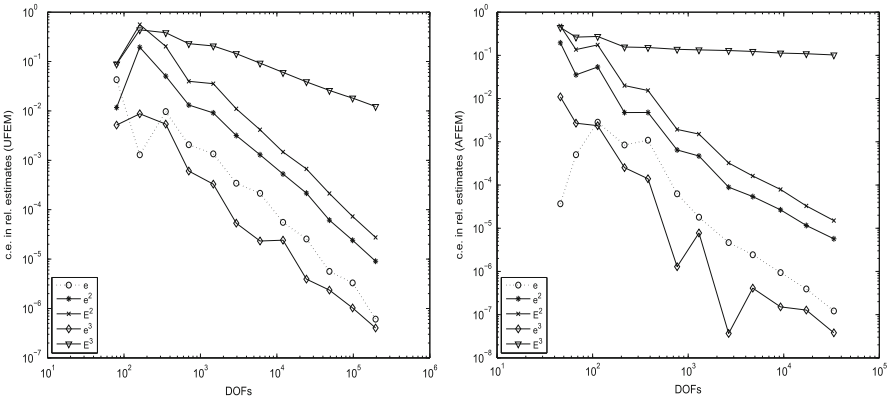


Fig. 4 Example 6.1. Decrease of the reliability related consistency error $e = |e^c_{h,rel}|$ (*dotted line*) and its estimates $e^k = |\bar{e}^{c,(k)}_{h,rel}|$, $E^k = \bar{E}^{c,(k)}_{h,rel}$, $2 \le k \le 3$, (*solid lines*) as functions of the DOFs on a logarithmic scale for uniform refinement (*left*) and adaptive refinement (*right*)

of uniform and adaptive refinement. We see that asymptotically the expected optimal convergence rates are achieved.

As far as the consistency errors and their estimates are concerned, we have to distinguish between the reliability related consistency errors $e^c_{h,rel}$ (cf. (5.6)) and the efficiency related consistency errors $e^c_{h,eff}$ (cf. (5.8)). Figure 4 displays the decay of $|e^c_{h,rel}|$ and its estimates $|\bar{e}^{c,(k)}_{h,rel}|$, $\bar{E}^{c,(k)}_{h,rel}$, $2 \le k \le 3$, as a function of the DOFs on a logarithmic scale for uniform refinement (left) and for adaptive refinement (right) (we note that $\bar{e}^{c,(1)}_{h,rel}$, $\bar{E}^{c,(1)}_{h,rel}$ and $\bar{e}^{c,(2)}_{h,rel}$, $\bar{E}^{c,(2)}_{h,rel}$ coincide for problems featuring strict complementarity which is the case in Example 1).

We observe that $|\bar{e}^{c,(2)}_{h,rel}|$ and $\bar{E}^{c,(2)}_{h,rel}$ provide upper bounds for $|e^c_{h,rel}|$ with approximately the same decay rates. On the other hand, $|\bar{e}^{c,(3)}_{h,rel}|$ slightly underestimates $|e^c_{h,rel}|$, whereas $\bar{E}^{c,(3)}_{h,rel}$ grossly overestimates $|e^c_{h,rel}|$ with an insufficient decay rate in particular for adaptive refinement.

Similarly, in Fig. 5 the decay of the efficiency related consistency errors $e^c_{h,eff}$ and their estimates $\bar{E}^{c,1}_{h,eff}$ are shown as functions of the DOFs on a logarithmic scale for uniform refinement (left) and adaptive refinement (right). After a pre-asymptotic phase, the estimates $\bar{E}^{c,1}_{h,eff}$ represent close upper bounds of $e^c_{h,eff}$ featuring essentially the same decay rates.

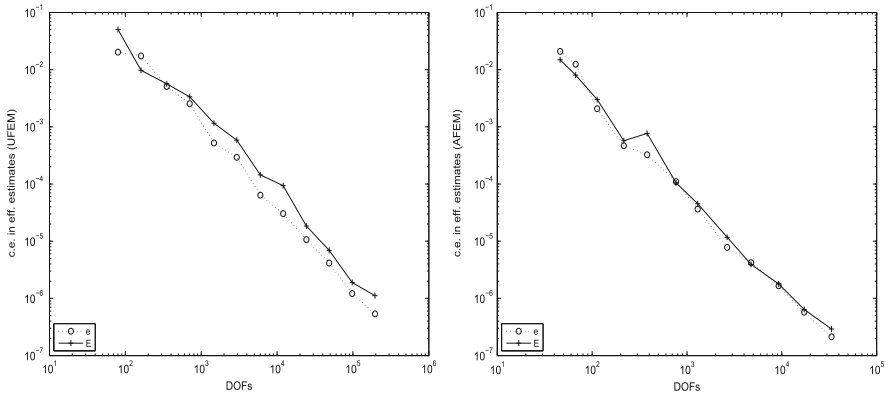


Fig. 5 Example 6.1. Decrease of the efficiency related consistency error $e = e^{c}_{h,eff}$ (*dotted line*) and its estimate $E = \bar{E}^{c,1}_{h,eff}$ (*solid line*) as functions of the DOFs on a logarithmic scale for uniform refinement (*left*) and adaptive refinement (*right*)

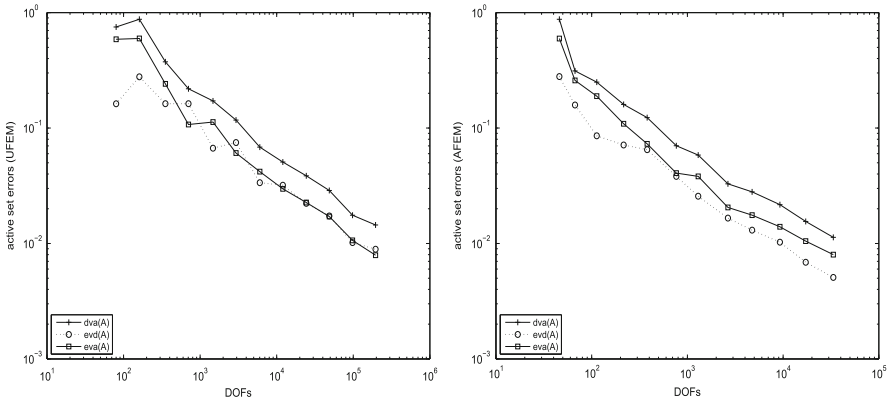


Fig. 6 Example 6.1. Approximation of the active set $\mathcal{A}$: quantities $e^{evd}_{h,\mathcal{A}}$ (*dotted line*) and $e^{dva}_{h,\mathcal{A}}$, $e^{eva}_{h,\mathcal{A}}$ (*solid lines*) as functions of the DOFs on a logarithmic scale for uniform refinement (*left*) and adaptive refinement (*right*)

Finally, Fig. 6 displays the decay of the errors with regard to the approximation of the active set $\mathcal{A}$ in terms of the quantities $e^{evd}_{h,\mathcal{A}}$, $e^{eva}_{h,\mathcal{A}}$, and $e^{dva}_{h,\mathcal{A}}$ (cf. (5.70),(5.71)). Recalling that the quantities $e^{evd}_{h,\mathcal{A}}$ and $e^{eva}_{h,\mathcal{A}}$ are the L^1-norms of the difference between the characteristic function of the continuous active set $\mathcal{A}$ on one hand and the characteristic function of the discrete active set $\mathcal{A}_h$ resp. the characteristic function of the approximate active set $\bar{\mathcal{A}}_h$ on the other hand, we see that the a posteriori quantity $e^{dva}_{h,\mathcal{A}}$ yields a close upper bound with approximately the same decay rates.

Example 6.2. The second example which has been considered in [18,21] features a problem with lack of strict complementarity. We consider $A = -\Delta$ on $\Omega = (0,1)^2$. Given

$$y^*(x_1,x_2) = \begin{cases} -z_1(x_1)z_2(x_2)\ , & (x_1,x_2) \in (0,0.5)\times(0,0.8) \\ 0\ , & \text{else} \end{cases},$$
$$\sigma^*(x_1,x_2) = 2\max(0, -|x_1-0.8| - |(x_2-0.2)x_1 - 0.3| + 0.35),$$
$$u^*(x_1,x_2) = y^*(x_1,x_2),$$

where

$$z_1(x_1) := -4096\,x_1^6 + 6144\,x_1^5 - 3072\,x_1^4 + 512\,x_1^3,$$
$$z_2(x_2) := -244.140625\,x_2^6 + 585.9375\,x_2^5 - 468.75\,x_2^4 + 125\,x_2^3,$$

it can be easily verified that the triple (y^*,σ^*,u^*) with the adjoint state $p^* = y^*$ and the multiplier $\mu^* = \sigma^*$ is an S-stationary point of (2.5) with respect to the data

$$y^d = \mu^* - \Delta p^* + y^*, \quad u^d = 0,$$
$$f = \sigma^* - \Delta y^* - p^*, \quad \alpha = 1, \quad \psi = 0.$$

Further, we have $\mathcal{I}^* = \{(x_1,x_2) \mid (x_1,x_2) \in (0,0.5)\times(0,0.8)\}$, $\mathcal{C}^* = \{(x_1,x_2) \mid |x_1-0.8| + |(x_2-0.2)x_1 - 0.3| \le 0.35\}$, and hence, $\mathcal{B}^* = \Omega \setminus (\mathcal{I}^* \cup \mathcal{C}^*) \neq \varnothing$. The optimal state y^* and the optimal multiplier σ^* are shown in Fig. 7, whereas the inactive set $\mathcal{I}^*$ and the strongly active set $\mathcal{C}^*$ are displayed in Fig. 8. Figure 9 shows the adaptively generated mesh at level $n = 7$ with 2439 DOFs and the final mesh (level $n = 11$) with 34159 DOFs.

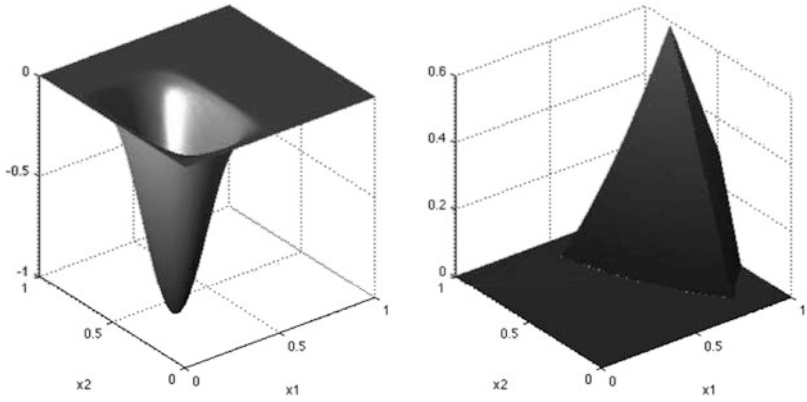


Fig. 7 Example 6.2. Optimal state y^* (*left*) and optimal multiplier σ^* (*right*)

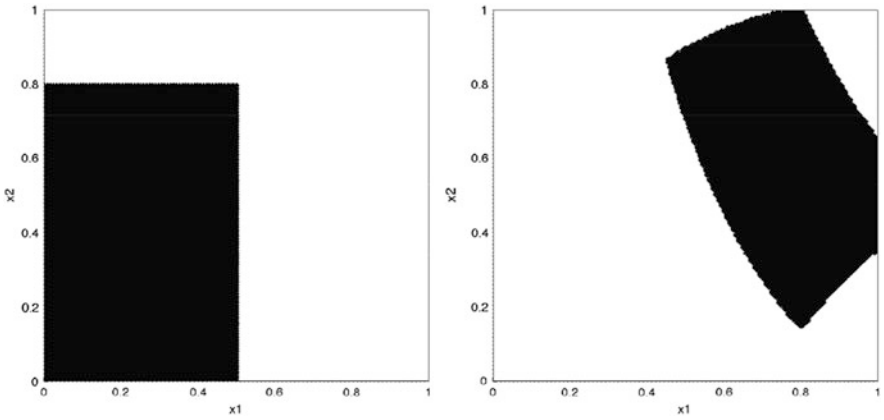


Fig. 8 Example 6.2. The inactive set $\mathcal{I}^*$ (*left*) and the strongly active set $\mathcal{C}^*$, both marked in *black* (*right*)

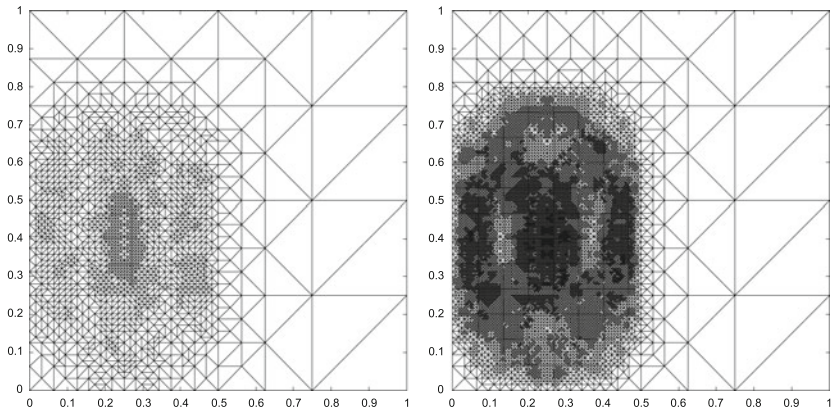


Fig. 9 Example 6.2. Mesh at refinement level $n = 7$ (*left*) and final mesh (*right*)

As in the first example, Fig. 10 displays the decrease of the errors in the state, in the control, in the total error, and in the estimator as functions of the DOFs on a logarithmic scale, whereas Table 2 contains the associated computed experimental convergence rates. Since the solution is smooth, uniform refinement is already optimal, i.e., in Table 2 we observe almost the same rates for uniform and adaptive refinement. However, as can be seen in Fig. 10, the error reductions are slightly less for adaptive refinement.

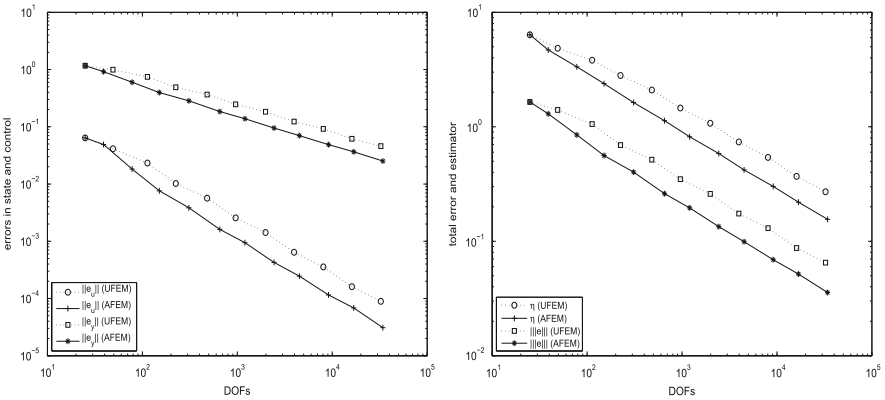


Fig. 10 Example 6.2. Convergence history: Decrease of the errors in the state $\|e_{h,y}\|_{1,\Omega}$ and in the control $\|e_{h,u}\|_{0,\Omega}$ as functions of the DOFs on a logarithmic scale (for uniform (UFEM) and adaptive (AFEM) refinement (*left*). Decrease of the estimator η_h and the total error $|||e_h|||$ as a function of the DOFs on a logarithmic scale (for uniform (UFEM) and adaptive (AFEM) refinement (*right*)

Table 2 Example 6.2: Experimental convergence rates (uniform and adaptive refinement)

	$\|e_{h,y}\|_{1,\Omega}$		$\|e_{h,p}\|_{1,\Omega}$		$\|e_{h,u}\|_{0,\Omega}$		$\|\|\|e_h\|\|\|$	
n	Unif.	Adapt.	Unif.	Adapt.	Unif.	Adapt.	Unif.	Adapt.
2	0.24	0.61	0.24	0.61	0.65	1.42	0.25	0.63
3	0.34	0.63	0.34	0.63	0.69	1.33	0.35	0.64
4	0.61	0.47	0.61	0.47	1.20	0.95	0.62	0.47
5	0.39	0.58	0.39	0.58	0.78	1.16	0.39	0.58
6	0.57	0.47	0.57	0.47	1.14	0.88	0.57	0.47
7	0.41	0.53	0.41	0.53	0.81	1.12	0.41	0.54
8	0.57	0.49	0.57	0.49	1.15	0.90	0.57	0.49
9	0.42	0.52	0.42	0.52	0.83	1.07	0.42	0.52
10	0.57	0.47	0.57	0.47	1.15	0.85	0.58	0.47
11	0.44	0.53	0.44	0.53	0.84	1.12	0.42	0.53

Figure 11 shows the decrease of the reliability related consistency error $e = |e^c_{h,rel}|$ (dotted line) and its estimates $e^k = |\bar{e}^{c,(k)}_{h,rel}|, E^k = \bar{E}^{c,(k)}_{h,rel}, 1 \le k \le 3$, as functions of the DOFs on a logarithmic scale both for uniform refinement (left) and for adaptive refinement (right). We see a very similar behavior as in Example 1, i.e., for $1 \le k \le 2$, the quantities $e^k = |\bar{e}^{c,(k)}_{h,rel}|$ and $E^k = \bar{E}^{c,(k)}_{h,rel}$ provide close upper bounds, whereas $|\bar{e}^{c,(3)}_{h,rel}|$ underestimates and $\bar{E}^{c,(3)}_{h,rel}$ grossly overestimates the consistency error $|e^c_{h,rel}|$.

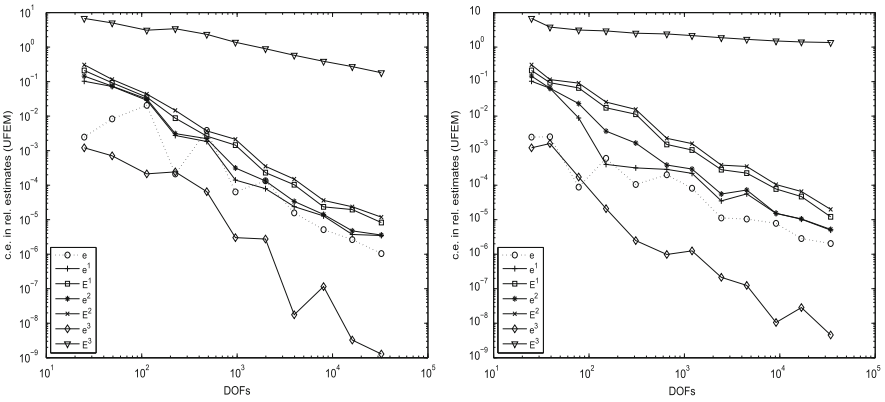


Fig. 11 Example 6.2. Decrease of the reliability related consistency error $e = |e^c_{h,rel}|$ (*dotted line*) and its estimates $e^k = |\tilde{e}^{c,(k)}_{h,rel}|$, $E^k = \tilde{E}^{c,(k)}_{h,rel}$, $1 \le k \le 3$, (*solid lines*) as functions of the DOFs on a logarithmic scale for uniform refinement (*left*) and adaptive refinement (*right*)

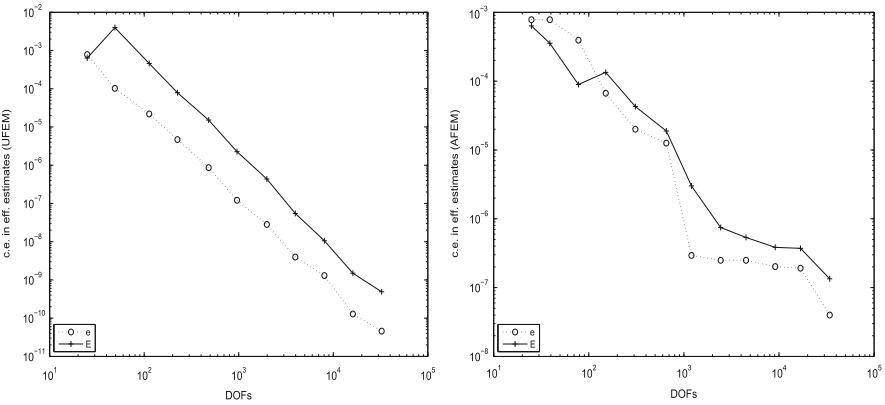


Fig. 12 Example 6.2. Decrease of the efficiency related consistency error $e = |e^c_{h,eff}|$ (*dotted line*) and its estimate $E = \tilde{E}^{c,(1)}_{h,eff}$ (*solid line*) as functions of the DOFs on a logarithmic scale for uniform refinement (*left*) and adaptive refinement (*right*)

Figure 12 displays the decrease of the efficiency related consistency error $e^c_{h,eff}$ and its estimate $\tilde{E}^{c,1}_{h,eff}$ as functions of the DOFs on a logarithmic scale for uniform refinement (left) and adaptive refinement (right). As in Example 1, after some pre-asymptotic phase in the adaptive regime, the estimates provide upper bounds of the consistency error.

Example 2 features the occurrence of a strongly active set $\mathcal{C}^*$ and hence, we are interested in how well the a posteriori quantities $e^{dva}_{h,\mathcal{A}}$ and $e^{dva}_{h,\mathcal{C}}$ coincide with $e^{eva}_{h,\mathcal{A}}$, $e^{evd}_{h,\mathcal{A}}$ and $e^{eva}_{h,\mathcal{C}}$, $e^{evd}_{h,\mathcal{C}}$, respectively. This is reflected in Figs. 13 and 14.

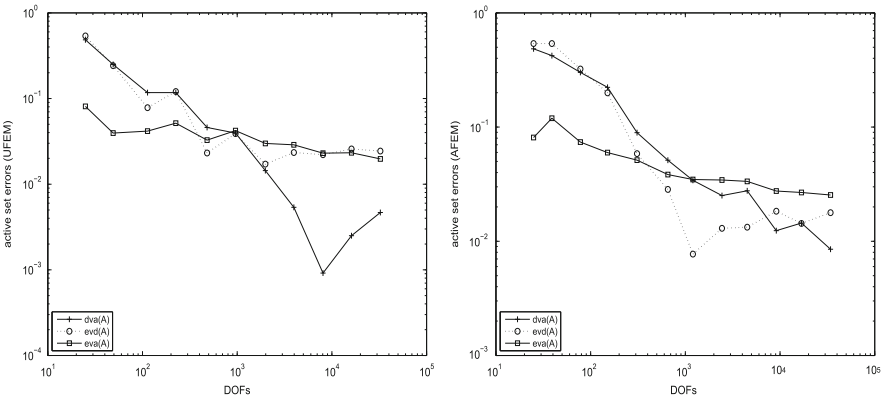


Fig. 13 Example 6.2. Approximation of the active set $\mathcal{A}$: quantities $e^{evd}_{h,\mathcal{A}}$ (*dotted line*) and $e^{dva}_{h,\mathcal{A}}, e^{eva}_{h,\mathcal{A}}$ (*solid lines*) as functions of the DOFs on a logarithmic scale for uniform refinement (*left*) and adaptive refinement (*right*)

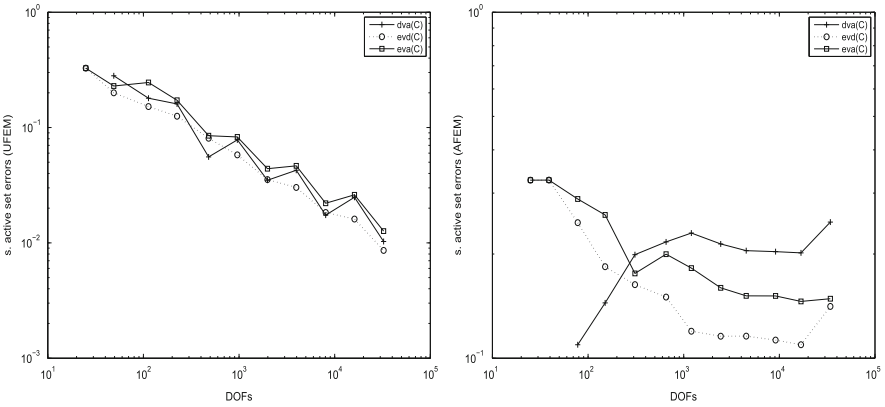


Fig. 14 Example 6.2. Approximation of the strongly active set $\mathcal{C}$: quantities $e^{evd}_{h,\mathcal{C}}$ (*dotted line*) and $e^{dva}_{h,\mathcal{C}}, e^{eva}_{h,\mathcal{C}}$ (*solid lines*) as functions of the DOFs on a logarithmic scale for uniform refinement (*left*) and adaptive refinement (*right*)

Acknowledgements A.G. has been partially supported by a grant from the European Science Foundation within the Networking Programme 'OPTPDE'. M.H. acknowledges support by the German Research Fund (DFG) through the Research Center MATHEON Project C28 and C31 and the SPP 1253 "Optimization with Partial Differential Equations", and the Austrian Science Fund (FWF) through the START Project Y 305-N18 Interfaces and Free Boundaries and the SFB Project F32 04-N18 "Mathematical Optimization and Its Applications in Biomedical Sciences". R.H.W.H. has been supported by the DFG Priority Programs SPP 1253 and SPP 1506, by the NSF grants DMS-0914788, DMS-1115658, by the Federal Ministry for Education and Research (BMBF) within the projects 'FROPT' and 'MeFreSim', and by the European Science Foundation within the Networking Programme 'OPTPDE'.

References

[1] R.A. Adams, J.J.F. Fournier, *Sobolev Spaces*, 2nd edn. (Academic, New York, 2003)
[2] M. Ainsworth, J.T. Oden, C.Y. Lee, Local a posteriori error estimators for variational inequalities. Numer. Math. Partial Differ. Equ. **9**, 23–33 (1993)
[3] I. Babuska, J. Whiteman, T. Strouboulis, *Finite Elements: An Introduction to the Method and Error Estimation*. (Oxford University Press, Oxford, 2011)
[4] V. Barbu, *Optimal Control of Variational Inequalities*. (Pitman, Boston/London/Melbourne, 1984)
[5] R. Becker, H. Kapp, R. Rannacher, Adaptive finite element methods for optimal control of partial differential equations: basic concept. SIAM J. Control Optim. **39**, 113–132 (2000)
[6] D. Braess, A posteriori error estimators for obstacle problems: another look. Numer. Math. **101**, 415–421 (2005)
[7] D. Braess, C. Carstensen, R.H.W. Hoppe, Convergence analysis of a conforming adaptive finite element method for an obstacle problem. Numer. Math. **107**, 455–471 (2007)
[8] D. Braess, C. Carstensen, R.H.W. Hoppe, Error reduction in adaptive finite element approximations of elliptic obstacle problems. J. Comp. Math. **27**, 148–169 (2009)
[9] S.C. Brenner, L.R. Scott, *The Mathematical Theory of Finite Element Methods*, 3rd edn. (Springer, New York, 2008)
[10] F. Facchinei, J.S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*. (Springer, Berlin/Heidelberg/New York, 2003)
[11] A. Gaevskaya, Adaptive finite element methods for optimally controlled elliptic variational inequalities, Ph.D. thesis, Institute for Mathematics, University of Augsburg, 2013
[12] A. Gaevskaya, R.H.W. Hoppe, S. Repin, Functional approach to a posteriori error estimation for elliptic optimal control problems with distributed control. J. Math. Sci. **144**, 4535–4547 (2007)
[13] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*. (Pitman, Boston/London/Melbourne, 1985)
[14] A. Günther, M. Hinze, A posteriori error control of a state constrained elliptic control problem. J. Numer. Math. **16**, 307–322 (2008)
[15] M. Hintermüller, R.H.W. Hoppe, Goal-oriented adaptivity in control constrained optimal control of partial differential equations. SIAM J. Control Optim. **47**, 1721–1743 (2008)
[16] M. Hintermüller, R.H.W. Hoppe, Goal oriented mesh adaptivity for mixed control-state elliptic optimal control problems, in *Applied and Numerical Partial Differential Equations. Scientific Computing in Simulation, Optimization and Control in a Multidisciplinary Context*, ed. by W. Fitzgibbon, Y. Kuznetsov, P. Neittaanmäki, J. Périaux, O. Pironneau. Computational Methods in Applied Sciences, vol. 15. (Springer, Berlin/Heidelberg/New York, 2009)
[17] M. Hintermüller, R.H.W. Hoppe, Goal-oriented adaptivity in pointwise state constrained optimal control of partial differential equations. SIAM J. Control Optim. **48**, 5468–5487 (2010)
[18] M. Hintermüller, I. Kopacka, Mathematical programs with complementarity constraints in function space: C- and strong stationarity and a path-following algorithm. SIAM J. Optim. **20**, 868–902 (2009)
[19] M. Hintermüller, R.H.W. Hoppe, Y. Iliash, M. Kieweg, An a posteriori error analysis of adaptive finite element methods for distributed elliptic control problems with control constraints. ESAIM Control Optim. Calc. Var. **14**, 540–560 (2008)
[20] M. Hintermüller, M. Hinze, R.H.W. Hoppe, Weak-duality based adaptive finite element methods for PDE-constrained optimization with pointwise gradient state-constraints. J. Comp. Math. **30**, 101–123 (2012)
[21] M. Hintermüller, R.H.W. Hoppe, C. Löbhard, Dual-weighted goal-oriented adaptive finite elements for optimal control of elliptic variational inequalities. ESAIM Control Optim. Calc. Var. **20**, 524–546 (2014)

[22] R.H.W. Hoppe, R. Kornhuber, Adaptive multilevel methods for obstacle problems. SIAM J. Numer. Anal. **31**, 301–323 (1994)
[23] R.H.W. Hoppe, Y. Iliash, C. Iyyunni, N. Sweilam, A posteriori error estimates for adaptive finite element discretizations of boundary control problems. J. Numer. Anal. **14**, 57–82 (2006)
[24] C. Johnson, Adaptive finite element methods for the obstacle problem. Math. Models Methods Appl. Sci. **2**, 483–487 (1992)
[25] D. Kinderlehrer, G. Stampacchia, *An Introduction to Variational Inequalities and Its Applications*. (SIAM, Philadelphia, 2000)
[26] D. Klatte, B. Kummer, *Nonsmooth Equations in Optimization: Regularity, Calculus, Methods, and Applications*. (Kluwer, Dordrecht, 2002)
[27] I. Kopacka, Mpecs/mpccs in functional space: first order optimality concepts, path-following and multilevel algorithms, Ph.D. thesis, Institute of Applied Mathematics, Karl-Franzens University at Graz, 2009
[28] R. Li, W. Liu, H. Ma, T. Tang, Adaptive finite element approximation for distributed elliptic optimal control problems. SIAM J. Control Optim. **41**, 1321–1349 (2002)
[29] Z.Q. Luo, J.S. Pang, D. Ralph, *Mathematical Programs with Equilibrium Constraints*. (Cambridge University Press, Cambridge, 1996)
[30] F. Mignot, Contrôle dans les inèquations variationelles elliptiques. J. Funct. Anal. **22**, 130–185 (1976)
[31] F. Mignot, J.P. Puel, Optimal control in some variational inequalities. SIAM J. Control Optim. **22**, 466–476 (1984)
[32] B.S. Mordukhovich, *Variational Analysis and Generalized Differentiation I: Basic Theory*. (Springer, Berlin/Heidelberg/New York, 2006)
[33] B.S. Mordukhovich, *Variational Analysis and Generalized Differentiation II: Applications*. (Springer, Berlin/Heidelberg/New York, 2006)
[34] P. Neittaanmäki, J. Sprekels, D. Tiba, *Optimization of Elliptic Systems: Theory and Applications*. (Springer, Berlin/Heidelberg/New York, 2006)
[35] R. Nochetto, K.G. Siebert, A. Veeser, Pointwise a posteriori error control for elliptic obstacle problems. Numer. Math. **95**, 163–195 (2003)
[36] J. Outrata, M. Kocvara, J. Zowe, *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints*. (Kluwer, Dordrecht, 1998)
[37] J. Outrata, J. Zowe, A numerical approach to optimization problems with variational inquality constraints. Math. Program. **68**, 105–130 (1995)
[38] H. Scheel, S. Scholtes, Mathematical programs with complementarity constraints: stationarity, optimality, and sensitivity. Math. Oper. Res. **25**, 1–22 (2000)
[39] L.R. Scott, S. Zhang, Finite element interpolation of nonsmooth functions satisfying boundary conditions. Math. Comput. **54**, 483–493 (1990)
[40] K.G. Siebert, A. Veeser, A unilaterally constrained quadratic minimization with adaptive finite elements. SIAM J. Optim. **18**, 260–289 (2007)
[41] R.S. Strichartz, *A Guide to Distribution Theory and Fourier Transforms*. (World Scientific, River Edge, 2003)
[42] F.T. Suttmeier, On a direct approach to adaptive FE-discretizations for elliptic variational inequalities. J. Numer. Math. **13**, 73–80 (2005)
[43] A. Veeser, Efficient and reliable a posteriori error estimators for elliptic obstacle problems. SIAM J. Numer. Anal. **39**, 146–167 (2001)
[44] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*. (Teubner & Wiley, Stuttgart, 1996)
[45] B. Vexler, W. Wollner, Adaptive finite elements for elliptic optimization problems with control constraints. SIAM J. Optim. **47**, 509–534 (2008)

Constrained Optimization: From Lagrangian Mechanics to Optimal Control and PDE Constraints

Martin J. Gander, Felix Kwok, and Gerhard Wanner

Abstract The history of constrained optimization spans nearly three centuries. The principal warhorse, Lagrange multipliers, was discovered by Lagrange in the Statics section of his famous book on Mechanics from 1788, by applying the idea of virtual velocities to problems in statics with constraints. The idea of virtual velocities, in turn, goes back to a letter of Johann Bernoulli from 1715 to Varignon, in which he announced a very simple rule for solving hundreds of Varignon's problems in the blink of an eye. Varignon then explains this rule in his book published in 1725. Half a century later, Bernoulli's rule was chosen by Lagrange as the general principle for the foundation of his mechanics, with the multipliers as the main tool for treating mechanical constraints. In the second edition of his mechanics, published in 1811, Lagrange stressed the importance of his multipliers also for constrained optimization. In particular, they provide spectacular simplifications of entire chapters of Euler's treatise on Variational Calculus from 1744. Lagrange multipliers is however a much farther reaching concept; we show how one can discover the important primal and dual equations in optimal control and the famous maximum principle of Pontryagin using only Lagrange multipliers. Pontryagin and his group, however, did not discover the maximum principle this way, since they were coming from a completely different area of mathematics. We finally give the complete formulation of PDE constrained optimization based on duality introduced by Lions, and conclude with an outlook on more recent applications.

Keywords Constrained optimization • Optimal control • PDE constrained optimization • Variational methods

Our intention is not to write a full historical paper, but to highlight the parts of the historical development we find interesting as mathematicians. For full details on the history of constrained optimization with complete references, see [45] and [46].

M.J. Gander (✉) • F. Kwok • G. Wanner
Faculté des Sciences, Section Mathématiques, Université de Genève, CH-1211 Genève 4, Suisse, Geneva, Switzerland
e-mail: martin.gander@unige.ch; felix.kwok@unige.ch; gerhard.wanner@unige.ch

R. Hoppe (ed.), *Optimization with PDE Constraints*, Lecture Notes in Computational Science and Engineering 101,
DOI 10.1007/978-3-319-08025-3_5

Mathematics Subject Classification (2010). Primary 01-02; Secondary 49-03, 65K10

1 Lagrange Multipliers Originating from Mechanics

> "Le *Traité de Dynamique* de M. d'Alembert, ... parut en 1743, ... Cette méthode réduit toutes les loix du mouvement des corps à celles de leur équilibre, & ramene ainsi la Dynamique à la Statique" [33, Seconde Partie, p. 179]

Lagrange's method of multipliers originates from Lagrange's research in mechanics, more precisely his *Mécanique analytique* [33], first published in 1788, with a second, improved edition [34] in 1811/1815. In his long introductions, Lagrange traces the following history for his work:

1. *Archimedes, Pappus, Varignon*: For nearly 2,000 years, research in mechanics concerned mainly *Statics*, beginning with the discovery of the law of the *lever* by Archimedes. Then, mainly by researchers as Pappus, Stevin, Roberval and Descartes, theories for the equilibria of ever more complicated "machines" were developed, culminating in the *Nouvelle Mécanique* by Varignon.
2. *Galilei, Newton, Leibniz, the Bernoulli brothers, Euler*: The next period then concentrated on the *Dynamics* of increasingly complex mechanical systems (mass points, liquids, rigid bodies) with more and more analytical methods (differential equations).
3. *Lagrange*: Finally, the "principle of d'Alembert" from 1743 reduces problems in dynamics back to problems in statics (see quotation), so that Lagrange's *Mécanique analytique* again started with an extensive "première partie" on statics, comprising nearly 200 pages, as a foundation for the now-called *Lagrangian mechanics* in the second part. The main idea there was the *Principle of Virtual Velocities*, which first appeared in a letter of Joh. Bernoulli from 1715 to Varignon. The extension of this idea to *constrained* mechanical problems then led to the invention of *Lagrange multipliers*.

1.1 Archimedes' Proof for the Lever

The very first great discovery in Statics was made by Archimedes with the law of the lever: *two bodies are in equilibrium if their weights are inversely proportional to their arm lengths* (see Fig. 1 and [1]).

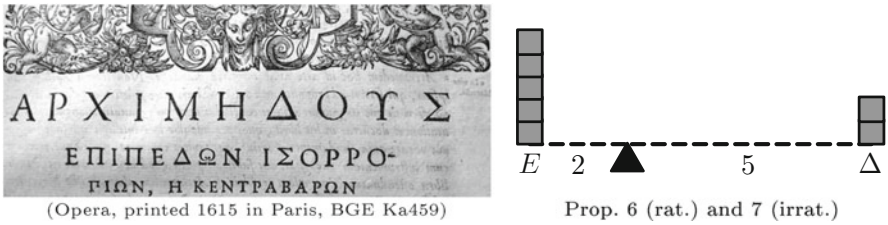


Fig. 1 Archimedes' law for the lever

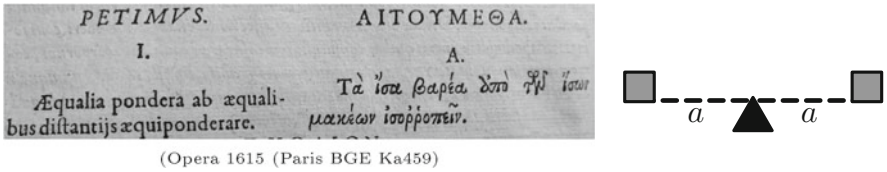


Fig. 2 Archimedes' hypothesis

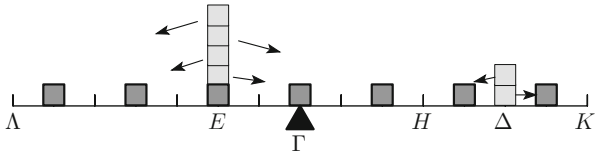
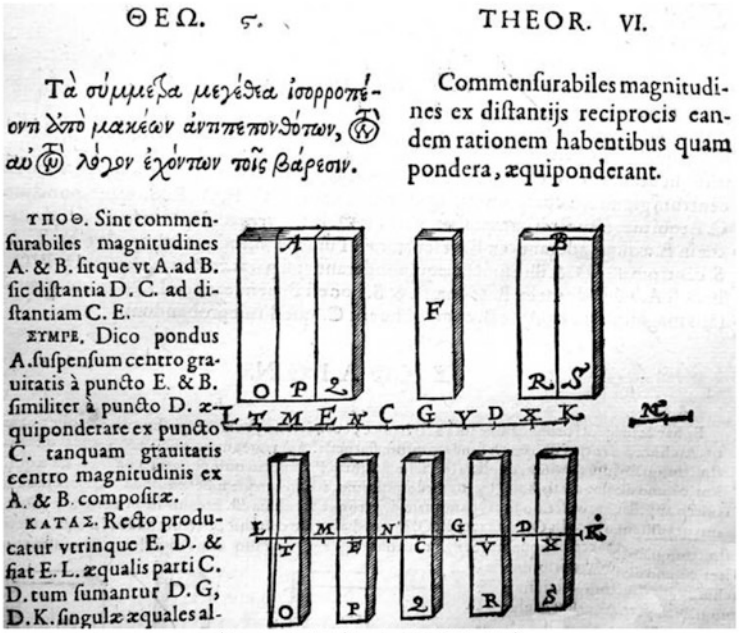


Fig. 3 Archimedes' proof of his Prop. 6

The proof of Archimedes is very beautiful: He started from the axiom that equal weights at equal distances are in equilibrium (see Fig. 2).

Then, after more axioms, several preliminary propositions and corollaries, he proved his Proposition 6, valid for rational ratios of weights, in two pages of Greek text. His idea was to distribute the weight units left and right in a symmetric way to obtain an overall symmetric configuration (see Fig. 3 for an illustration in the case of a $5 : 2$ lever). Figure 4 shows the corresponding proposition and figure for the ratio $3 : 2$, which appear in the 1615 edition of Archimedes' *Opera* (observe that the letters L, E, C, G, D, K of the Latinized version correspond to Archimedes' $\Lambda, E, \Gamma, H, \Delta, K$).



(Opera 1615 (Paris BGE Ka459)

Fig. 4 Archimedes' Prop. 6 with figure from the 1615 edition

1.2 Virtual Velocities and Joh. Bernoulli's "Regle"

> ... il n'y a pas un seul cas d'equilibre dans toute la mechanique tant des fluides que des solides, qui ne puisse etre expliqué par cette regle ... J'ay donc raison d'appeller le grand et le premier principe de statique sur lequel j'ay fondé ma regle ... (Joh. Bernoulli in his letter to Varignon, 1715)

> ... je crois pouvoir avancer que tous les principes généraux qu'on pourrait peut-être encore découvrir dans la science de l'équilibre ne seront que le même principe des vitesses virtuelles, envisagé différemment, et dont ils ne différeront que dans l'expression [34, Sect. I, §17].

All the efforts during the centuries after Archimedes in generalizing this result to more and more complicated situations culminated in the work of Pierre Varignon, who elaborated during many decades his *Nouvelle Mécanique* [51], consisting of two heavy volumes published posthumously in 1725,[1] with hundreds of results illustrated on 64 plates of figures (see Fig. 5).

When this work was nearly completed, Joh. Bernoulli explained in a letter to *Mr. le Chev. Renau*, with a copy to Varignon, his "regle" based on the *Virtual Velocities*,

[1] On the frontispiece is written "Dont le projet fut donné en M.DC.LXXXVII".

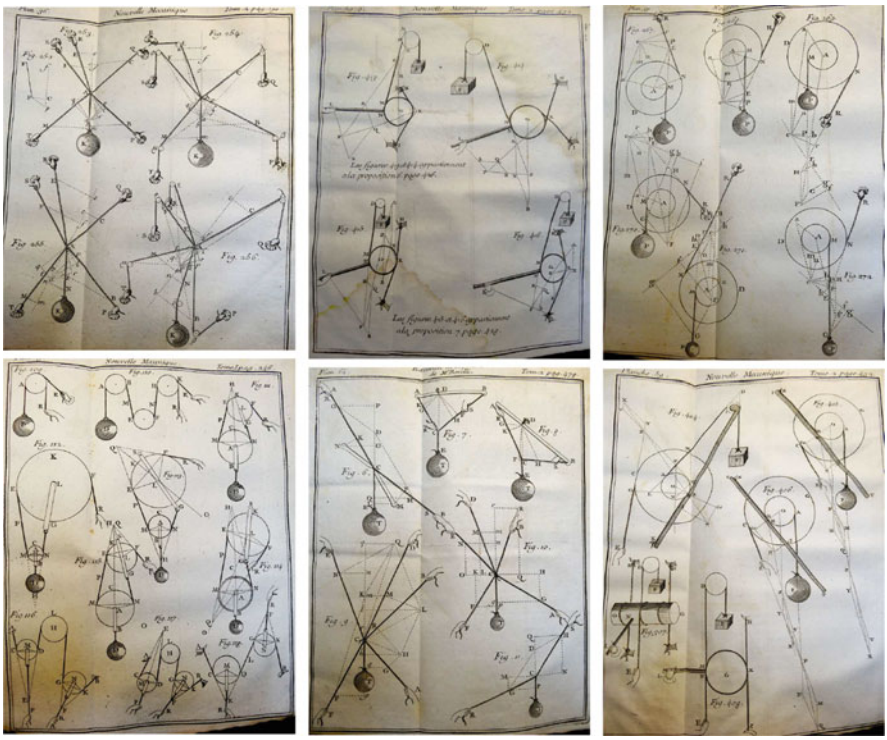


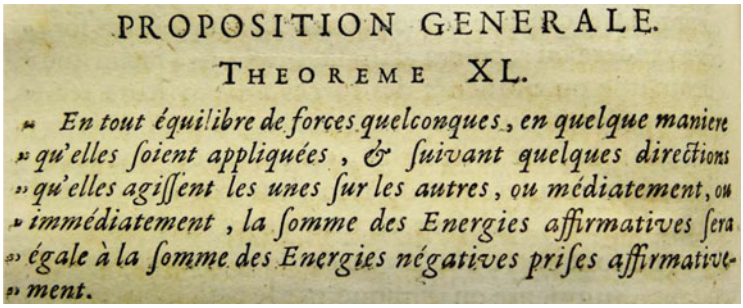
Fig. 5 Six out of the 64 figure plates from Varignon [51]; (the *upper left figure* of the *upper left plate* explains the principle of virtual velocities as in Fig. 9 below)

which allowed one to replace *all such figures* by *one general equation*. Varignon had some difficulty in admitting that all his work over decades was declared to be an "easy game"[2] and contested the general truth of this rule. Bernoulli then got angry[3] and explained his ideas in more detail, written in a second letter, dated Feb. 26, 1715.[4] Varignon then included Bernoulli's "regle" as "Theoreme XL" in "Section IX" ("Corollaire general de la Théorie précedente") of his book, by saying that, unfortunately, it was too late to rewrite all the rest of the book (see Fig. 6).

[2] "Votre *projet d'une nouvelle mechanique* fourmille d'un grand nombre d'exemples, dont quelques uns à en juger par les figures paroissent assez compliqués; mais je vous deffie de m'en proposer un à votre choix, que je ne resolve sur le champ et comme en jouant par ma dite regle."

[3] "... cependant permettez moy que je vous reproche ici une nonchalance qui vous est arrivé assez souvent en ce que vous portez quelques fois votre jugement un peu à la legere, sans examiner, si ce que vous croyez etre une objection en est veritablement une ; ... c'est donc pour une autre fois que je vous donne cet avertissement à fin que vous soyez à l'avenir sur vos gardes, quand il s'agit de juger..."

[4] Varignon gave in his book the wrong date 1717, which was also copied by Lagrange.



PROPOSITION GENERALE.

THEOREME XL.

« En tout équilibre de forces quelconques, en quelque maniere » qu'elles ſoient appliquées, & ſuivant quelques directions » qu'elles agiſſent les unes ſur les autres, ou médiatement, ou » immédiatement, la ſomme des Energies affirmatives ſera » égale à la ſomme des Energies négatives priſes affirmative- » ment.

Fig. 6 Bernoulli's "regle" as published by Varignon [51, Vol. II, p. 176]

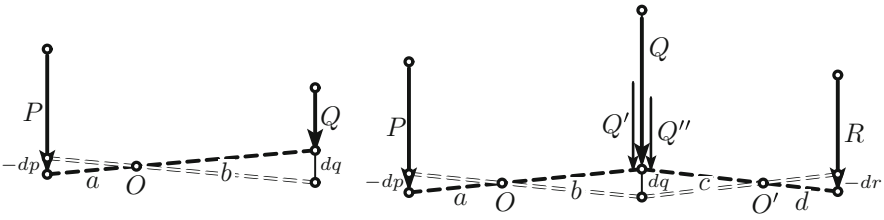


Fig. 7 The Lever (*left*); Composed levers (*right*)

We now describe the derivation of Benoulli's "regle" following the text of Lagrange [33, Prem. Partie, Sect. II]. However we do not follow the style of Lagrange, who proudly avoided the use of any figures.

We start with a system containing *two* forces P and Q, illustrated here by a lever (see Fig. 7, left) attached at O with arm lengths a and b. We then suppose that the system receives a virtual velocity during an infinitely small interval of time, such that the lever arms receive infinitely small displacements dp and dq proportional to a and b. Archimedes' law then tells us that for equilibrium to occur, the virtual velocities and the forces must be inversely proportional. Thus, if we pay attention to the signs of the displacements, we obtain

$$\frac{P}{Q} = -\frac{dq}{dp} \qquad \text{or} \qquad Pdp + Qdq = 0.$$

Let us now make the system more complicated by considering *three* forces P, Q and R instead of two (Fig. 7, right). We decompose the force Q as sum $Q = Q' + Q''$ in such a way that both subsystems to the left and right are in equilibrium, i.e., such that

$$Pdp + Q'dq = 0 \quad \text{and} \quad Q''dq + Rdr = 0,$$

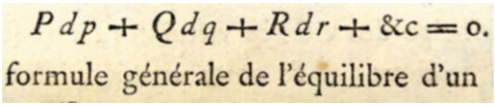


Fig. 8 Bernoulli's rule as published by Lagrange [33]

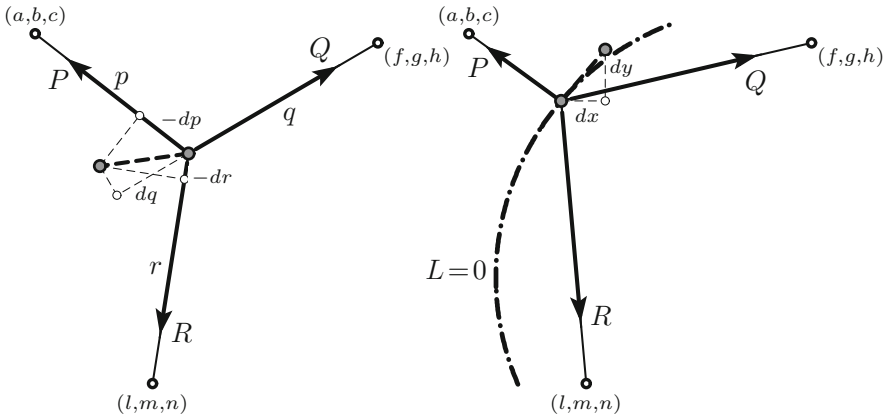


Fig. 9 A point attached by three forces (*left*); as constrained problem (*right*)

so we get $Pdp + Qdq + Rdr = 0$ as condition for an equilibrium. By adding more and more forces to the system, we obtain

$$\boxed{Pdp + Qdq + Rdr + \ldots = 0} \tag{1.1}$$

for an equilibrium. This equation, expressed in words and not in formulas, was precisely Joh. Bernoulli's "regle" of Fig. 6. The terms $Pdp, Qdq, \ldots$ were called "Energies" by Bernoulli. Lagrange calls them "moments" of the forces and calls (1.1) "la formule générale de l'équilibre" (see Fig. 8).

Example. The first example Lagrange considers in detail (in Sect. V) is a point mass attached by several forces P, Q, R to fixed points with Cartesian coordinates $(a, b, c), (f, g, h), (l, m, n)$ (see Fig. 9, left). Inserting

$$p = \sqrt{(x-a)^2 + (y-b)^2 + (z-c)^2}, \quad dp = \frac{1}{p} \cdot ((x-a)dx + (y-b)dy + (z-c)dz)\,,$$

and similarly for dq, dr, formula (1.1) becomes

$$Xdx + Ydy + Zdz = 0 \tag{1.2}$$

where $X = P\frac{x-a}{p} + Q\frac{x-f}{q} + R\frac{x-l}{r}$, $Y = P\frac{y-b}{p} + Q\frac{y-g}{q} + R\frac{y-m}{r}$ and $Z = P\frac{z-c}{p} + Q\frac{z-h}{q} + R\frac{z-n}{r}$. Since, at the moment, our point mass is completely free, dx, dy and dz are independent,[5] and the condition for equilibrium is

$$X = 0\,, \qquad Y = 0 \qquad \text{and} \qquad Z = 0\,. \tag{1.3}$$

In the case where the forces P, Q, R are equal (or proportional) to the distances p, q, r, this formula simplifies considerably and the equilibrium position becomes the barycenter of the triangle spanned by the three fixed points (or of a pyramid in the case of four forces, a result which Lagrange attributes to Leibniz).

1.3 The Discovery of the Multiplier Method

Suppose now (see Fig. 9, right) that the point mass is restricted to a surface $L = 0$, so that in (1.2) the displacements dx, dy, dz are *not* independent, but are restricted to the tangent space of $L = 0$, i.e. they must satisfy

$$dL = \frac{\partial L}{\partial x}dx + \frac{\partial L}{\partial y}dy + \frac{\partial L}{\partial z}dz = 0\,. \tag{1.4}$$

This means geometrically that, whenever (1.4) holds, i.e. the vector (dx, dy, dz) is orthogonal to $(\frac{\partial L}{\partial x}, \frac{\partial L}{\partial y}, \frac{\partial L}{\partial z})$, we must satisfy (1.2) as well, i.e. the vector (dx, dy, dz) must also be orthogonal to (X, Y, Z). As a consequence, both vectors must be parallel so that there exists a constant λ such that

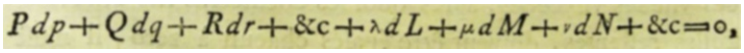
$$X + \lambda\frac{\partial L}{\partial x} = 0\,, \qquad Y + \lambda\frac{\partial L}{\partial y} = 0 \quad \text{and} \quad Z + \lambda\frac{\partial L}{\partial z} = 0\,. \tag{1.5}$$

However, vectors and scalar products were not yet familiar concepts to Lagrange, so he argued differently ("Il n'est pas difficile de prouver par la théorie de l'élimination des équations linéaires..."): we eliminate one of the unknowns, say dz, by multiplying (1.4) with a suitable constant, which is $\lambda = -Z/\frac{\partial L}{\partial z}$, and add it to (1.2), which gives

$$\left(X + \lambda\frac{\partial L}{\partial x}\right)\cdot dx + \left(Y + \lambda\frac{\partial L}{\partial y}\right)\cdot dy = 0\,, \qquad Z + \lambda\frac{\partial L}{\partial z} = 0\,.$$

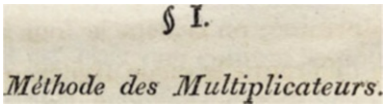
Here, dx and dy are independent and equation (1.5) must be satisfied, the last one being the formula for λ.

[5] dp, dq, dr are not independent at the equilibrium point.



$$P\,dp + Q\,dq + R\,dr + \&c + \lambda\,dL + \mu\,dM + \nu\,dN + \&c = 0,$$

Fig. 10 Lagrange's "équation générale" for ALL problems of equilibria



§ I.

Méthode des Multiplicateurs.

Fig. 11 Heading of §1 in Sect. IV of Lagrange [34]

Condition (1.5) just means that we have applied the virtual velocity argument, *without constraints*, to the system

$$X dx + Y dy + Z dz + \lambda dL = 0\,. \tag{1.6}$$

Lagrange realizes that this "multiplier" λ, whose invention originated from the theory of linear equations, also has a physical meaning: it represents the constant which, when multiplied with the vector $(\frac{\partial L}{\partial x}, \frac{\partial L}{\partial y}, \frac{\partial L}{\partial z})$, yields the force that holds the particle onto the surface $L = 0$.

To include an additional constraint $M = 0$, we see from linear algebra that we can simply add another term μdM, and so on. Finally, one can generalize (1.1) to any system with any number of constraints by writing

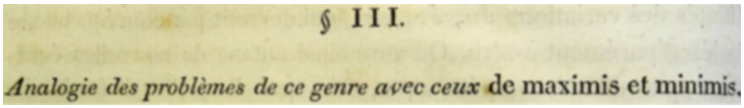
$$\boxed{P dp + Q dq + R dr + \ldots + \lambda dL + \mu dM + \nu dN + \ldots = 0} \tag{1.7}$$

(see Fig. 10). This discovery was called "Méthode très-simple" in Sect. IV of the first edition from 1788. Twenty-three years later, in [34], Lagrange stressed the importance of this idea by giving it the particular name "Méthode des Multiplicateurs" (see Fig. 11).

2 Problems of Maximum and Minimum

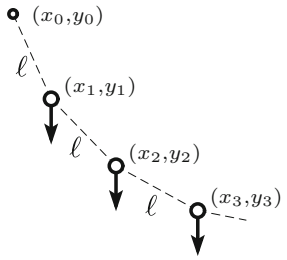
The above problems of *virtual velocities* are closely related to problems of maximizing or minimizing a function. This connection is mentioned briefly in Lagrange [33], but it was only in the second edition from 1811 that Lagrange stresses this important fact by an entire paragraph (see Fig. 12). If $U(x, y, z)$ is a "potential" function[6] satisfying $\frac{\partial U}{\partial x} = X$, $\frac{\partial U}{\partial y} = Y$ and $\frac{\partial U}{\partial z} = Z$, where X, Y and Z are as in (1.2), then the conditions (1.3) mean nothing else than

[6] Up to now, we have preserved all letters exactly as they appear in Lagrange, but we have changed this potential, denoted Π by Lagrange, to U, as it is usual now.


§ III.

Analogie des problèmes de ce genre avec ceux de maximis et minimis.

Fig. 12 Heading of §3 in Sect. IV of Lagrange [34]



$$\begin{aligned}&(x_0-x_1)^2+(y_0-y_1)^2-\ell^2=0\\&(x_1-x_2)^2+(y_1-y_2)^2-\ell^2=0\\&(x_2-x_3)^2+(y_2-y_3)^2-\ell^2=0\\&\cdots\end{aligned}$$

Fig. 13 The Catenary as a constrained mechanical system

$$U(x,y,z)\longrightarrow \text{min or max}. \tag{2.1}$$

Similarly, in the case where we have to minimize or maximize a function $U(x,y,z)$ under a constraint $L(x,y,z)=0$, the corresponding equations (1.5) and (1.6) would mean that we have to minimize or maximize

$$U(x,y,z)+\lambda L(x,y,z)\longrightarrow \text{min or max} \tag{2.2}$$

without constraints. This is the *Lagrange multiplier method for constrained optimization*. The geometric meaning of the term $\lambda L(x,y,z)$ is the following: it twists the function $U(x,y,z)$, without changing its values on the surface $L=0$, such that $U+\lambda L$ becomes flat in all directions at the minimal position.

For additional constraints, we add additional multipliers, and for higher dimensions, we add additional variables.

Example: The Catenary. One of the examples Lagrange discusses in detail (Part I, Sect. V) is a chain of particles attached by cords of constant length in an arbitrary force field. If we assume the forces to be constant downwards, we have the situation as in Fig. 13, for which (1.7) becomes

$$dy_1+dy_2+\ldots+\lambda_0\cdot d((x_0-x_1)^2+(y_0-y_1)^2-\ell^2)+\lambda_1\cdot d(\ldots)+\ldots=0. \tag{2.3}$$

Differentiating the constraints and collecting the coefficients of, say, dx_2, dy_2, we obtain

$$\begin{aligned}\lambda_2(x_2-x_3)&=\lambda_1(x_1-x_2)\\ \lambda_2(y_2-y_3)&=\lambda_1(y_1-y_2)-1\end{aligned}\quad\Rightarrow\quad\frac{y_2-y_3}{x_2-x_3}=\frac{y_1-y_2}{x_1-x_2}+\text{const.},$$

which means that the slope is a linear function of the arc length. This fact is in accordance with "... les formules connues de la chainette".

The Catenary as Optimization Problem. If we ask for the chain with $y_1+y_2+y_3+\ldots\longrightarrow$ min under the same constraints as in Fig. 13, i.e. if we seek the chain with the lowest center of gravity, (2.2) becomes

$$y_1+y_2+\ldots+\lambda_0\cdot((x_0-x_1)^2+(y_0-y_1)^2-\ell^2)+\lambda_1\cdot(\ldots)+\ldots\longrightarrow\min. \tag{2.4}$$

This equation, when differentiated, gives precisely the formula (2.3). We thus see that *the catenary is the curve with the lowest center of gravity for a given arc length*, a result Euler [20, Chap. V] found in a much more complicated way, as we will see below.

2.1 Variational Problems

Variational problems are optimization problems where not only some values, but an entire function $y(x)$, is unknown, for example

$$J=\int_a^b Z(x,y,p)\,dx\longrightarrow\text{min or max, where }\ p=\frac{dy}{dx} \tag{2.5}$$

and $Z(x,y,p)$ is a given function. We refer to Gander-Wanner ([28] SIREV 2013, formula (1.3), (1.4) and Sect. 9.1) to see how Euler [20, Chap. 2] turned this problem into a differential equation

$$\boxed{N-\frac{d}{dx}P=0}\qquad\textit{where}\quad N=\frac{\partial Z}{\partial y},\ \ P=\frac{\partial Z}{\partial p}, \tag{2.6}$$

and, in the case where $Z(y,p)$ is independent of x, how this equation can be simplified to

$$\boxed{Z-p\cdot\frac{\partial Z}{\partial p}=\text{Const.}} \tag{2.7}$$

2.2 Variational Problems with Constraints

The oldest problem of this type, the so-called "isoperimetric problem", was a challenge from Jakob Bernoulli to his brother Johann in 1697: *Given two points B and C (see Fig. 14), find a curve BaC of a given length L such that the area BMETNB is maximal; here, for any distance aN = y, the distance MN = g(y) is a given function of y*. In formulas, this means

$$\int_B^T g(y(x))\,dx \longrightarrow \max \quad \text{subject to} \quad \int_B^T \sqrt{1+p^2}\,dx = L\,. \tag{2.8}$$

Solution. Johann, who had accumulated success after success in the years before, thought that he could solve this seemingly simple problem in "three minutes". The 3 min turned into decades until Johann Bernoulli published an extensive paper in 1718 (Mémoires de l'Acad. Roy. des Sciences de Paris, p. 100). The collection of all the solutions of Jakob and himself fills more than 50 pages in Johann's *Opera Omnia* [4, vol. 2, p. 214–269]. Finally, Euler ([20], in Chap. 5 of E65) developed his general theory for such constrained problems. While in Chap. 2, Euler arrived at (2.6) by "virtual" displacements of the function values of the unknown function one-by-one (see Fig. 15, left), he was unable to displace the function values independently for constrained problems of the type (1.4). Instead, he varied the values *two by two*

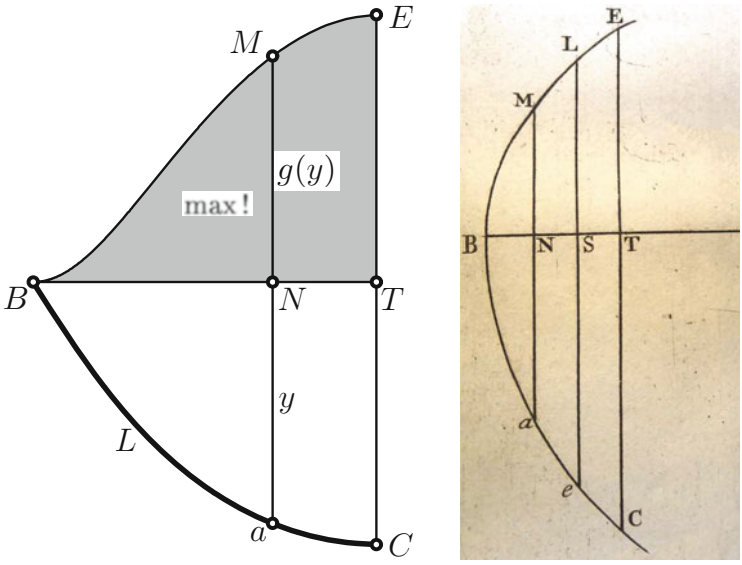


Fig. 14 The isoperimetric problem of Jakob (*left*, the drawing is for $g(y) = y^2$); the same picture in Johann's *Opera Omnia* from 1742, vol. 2, p. 270 (*right*)

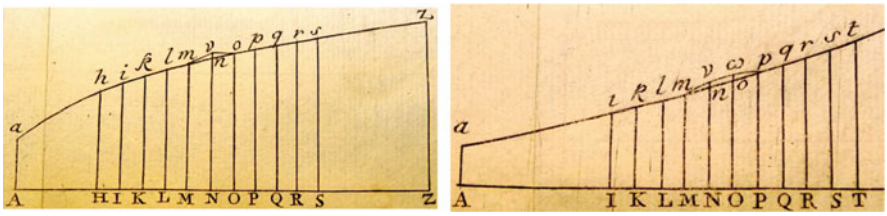


Fig. 15 Euler's solution of variational problems; unconstrained (left), constrained (right)

$n \mapsto \nu, o \mapsto \omega$ (see Fig. 15, right) and had to build an entirely new theory (16 pages; §1 through §39 of Chap. 5).

As Lagrange demonstrates proudly in many examples (in Sect. V), the idea of using multipliers to deal with constraints extends straightforwardly to these new problems. For the historical example (2.8), this turns into (for $B = 0, T = 1$)

$$J = \int_0^1 \Big(g(y) + \lambda\big(\sqrt{1+p^2} - L\big)\Big)\, dx \longrightarrow \max. \tag{2.9}$$

For this problem, condition (2.7) becomes, after simplification,

$$g(y) + \frac{\lambda}{\sqrt{1+p^2}} = C + \lambda L\,.$$

We set $C + \lambda L = -K$, solve for $p = \frac{dy}{dx}$ and separate the variables. This gives the solution (compared to the one from Johann's *Opera Omnia*, vol. 2, p. 244)

$$\int \frac{g(y) + K}{\sqrt{\lambda^2 - (g(y)+K)^2}}\, dy = x + c\,. \qquad y = \int \frac{(X \pm c)\, dx}{\sqrt{(aa - (X \pm c)^2)}} \tag{2.10}$$

This integral only has an elementary solution for $g(y) = y$, i.e. the problem of finding the maximal area surrounded by a curve of prescribed length. As Euler shows in §41 of [20] E65, Caput V, the integral then leads, not surprisingly, to a circular solution (quae est aequatio generalis pro Circulo). The drawing for $g(y) = y^2$ in Fig. 14 (left) has been produced by numerical integrations.

An Example with Two Constraints. For problems with *two* constraints ("Pluribus Proprietatibus"), Euler developed again an entirely new theory (E65, Chap. VI). With Lagrange, we just have to add a second multiplier. We demonstrate this on Euler's very last example (§24 in Chap. 6): We seek a curve $y(x)$ (the curve *DMAMD* in Fig. 16, right) of a given length L, as well as a constant a (the distance

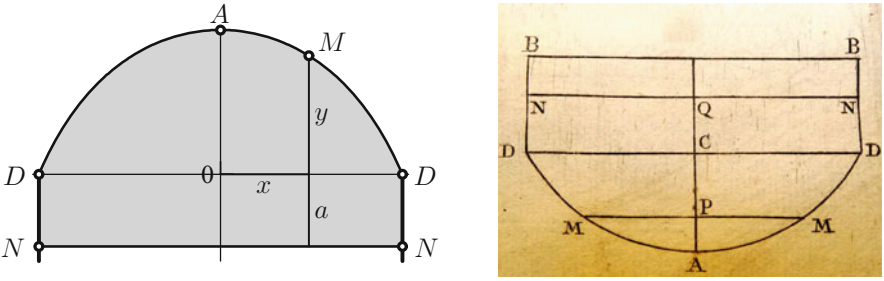


Fig. 16 Euler's problem from E65 with two constraints

CQ), such that the area of *NDMAMDNQN* has a given value M, and the center of gravity of this figure should be as low as possible. Expressed in formulas we have (we choose C as origin and take the curve upside down)

$$\int_{-1}^{1} \sqrt{1+p^2}\,dx = L\,, \quad \int_{-1}^{1} (y+a)\,dx = M\,, \quad \int_{-1}^{1} (y+a)\cdot\frac{y-a}{2}\,dx \longrightarrow \max.$$

Here, we introduce two multipliers λ and μ and get

$$J = \int_{-1}^{1} \Big((y^2-a^2) + \lambda\big(\sqrt{1+p^2}-L\big) + \mu\big((y+a)-M\big)\Big)\,dx \longrightarrow \min \text{ or } \max.$$

Since we have two unknowns y and a here, we cannot work with the simplified equation (2.7). Instead, we have to use (2.6) for each of them:

$$\text{for } y\colon\ 2y + \mu - \frac{d}{dx}\Big(\lambda\frac{p}{\sqrt{1+p^2}}\Big) = 0,$$
$$\text{for } a\colon\ -2a + \mu = 0 \quad \Rightarrow \quad \mu = 2a\,.$$

This, inserted into the first equation, gives

$$\frac{d}{dx}\Big(\frac{p}{\sqrt{1+p^2}}\Big) = k(y+a)\,.$$

If we think of a water basin, this result expresses the fact that *the curvature of the basin is proportional to the water pressure*.

2.3 Solving Optimal Control Problems with Lagrange Multipliers

Before explaining the invention of the maximum principle for control problems in the next section, we first show that the idea of Lagrange multipliers provides an elegant entry point to the treatment of certain classes of such problems. Let us look at a problem of the type

$$\int_a^b k(x, y, u)\, dx \longrightarrow \text{min or max}, \tag{2.11}$$

subject to

$$\frac{dy}{dx} = f(x, y, u), \qquad y(a) = A, \quad y(b) = B.$$

Here we have *two* types of functions to find: the values of $y_i(x)$, which are defined via a system of differential equations, and the so-called *controls* $u_j(x)$, which control the movement of the y's and with the help of which the *cost function* $k(x, y, u)$, when integrated over the interval $[a, b]$, is to be optimized.

Idea: since the differential equations in (2.11) represent an infinite number of constraints as x varies, we introduce Lagrange multipliers $\lambda_i(x)$ as *functions* multiplying the constraints $y_i' - f_i(x, y, u) = 0$. Inserting this into the integral, we thus obtain

$$\int_a^b \left\{k(x, y, u) + [p^T - f^T(x, y, u)] \cdot \lambda(x)\right\} dx \longrightarrow \text{min or max}. \tag{2.12}$$

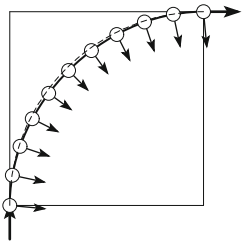
This is now an *unconstrained* variational problem with a "cost function" $Z(x, \lambda, y, p, u)$. Here we have three sets of unknowns, the Lagrange multipliers $\lambda_i(x)$, the differential equation solutions $y_i(x)$ together with their derivatives $p_i(x)$, and the control functions $u_j(x)$. For each of these, we apply Euler's equation (2.6):

$$\begin{aligned}
\boxed{\tfrac{\partial Z}{\partial \lambda} = 0} \quad &: \quad y'(x) = f(x, y, u) \\
\boxed{\tfrac{\partial Z}{\partial y} - \tfrac{d}{dx}\tfrac{\partial Z}{\partial p} = 0} \quad &: \quad \lambda'(x) = \tfrac{\partial k}{\partial y}(x, y, u) - \tfrac{\partial f}{\partial y}^T(x, y, u) \cdot \lambda(x) \\
\boxed{\tfrac{\partial Z}{\partial u} = 0} \quad &: \quad 0 \quad = \tfrac{\partial k}{\partial u}(x, y, u) - \tfrac{\partial f}{\partial u}^T(x, y, u) \cdot \lambda(x)
\end{aligned} \tag{2.13}$$

This is a system of differential algebraic equations (DAEs). The first set of equations are the desired constraints, the second set of equations is the so-called *adjoint system*, whose geometric meaning will be discussed below, and the third set consists of algebraic equations that determine the controls for every value of x.

Example. A body gliding in R^2 without friction should receive a new direction with the help of forces $(u_1(t), u_2(t)), 0 \le t \le T$ in such a way that this control uses as little energy as possible: $\int_0^T \frac{1}{2}(u_1^2 + u_2^2)\, dt \longrightarrow \min$.

Solution. With y_1, y_2 as the positions of the body and y_3, y_4 as velocities, the equations of motion together with the equations in (2.13) become



$$\begin{aligned} \dot{y}_1 &= y_3 & \dot{\lambda}_1 &= 0 & \\ \dot{y}_2 &= y_4 & \dot{\lambda}_2 &= 0 & u_1 - \lambda_3 &= 0 \\ \dot{y}_3 &= u_1 & \dot{\lambda}_3 &= -\lambda_1 & u_2 - \lambda_4 &= 0 \\ \dot{y}_4 &= u_2 & \dot{\lambda}_4 &= -\lambda_2 & \end{aligned}$$

We see that λ_1, λ_2 are constants, $\lambda_3 = u_1, \lambda_4 = u_2$ are linear, y_3, y_4 quadratic, and thus y_1, y_2 cubic; the solution curves are thus, not surprisingly, cubic splines. The time length T can be freely chosen. In the picture above, T is chosen to be that of a uniform circular movement, but the optimal solution is slightly different.

3 Optimal Control and the Maximum Principle

> An important case in applications is the one in which Ω [containing the controls] is a closed region [...]. In the case that Ω is an open set [...], the variational problem formulated here turns out to be a special case of the problem of Lagrange [47].

In the field of optimal control, there were historically two approaches: in the western world, researchers tried to tackle these problems using variational calculus and Lagrange multipliers, as we have already seen for a first example in Sect. 2.3. In Russia, a group of researchers led by Pontryagin tried to solve these problems using direct analysis and geometric arguments, with a particular emphasis on handling the important case of closed and bounded control sets. Their approach led to the invention of the maximum principle in 1956; they only later noticed the relation to Lagrange multipliers, see the quote above. To explain these two approaches historically, we first present the invention of Lagrange from Sect. 1.3 again, but now using matrix notation in preparation for its use in optimal control problems.

3.1 Invention of Lagrange Multipliers in Matrix Notation

> Lagrange, in his book from 1797: "Théorie des fonctions analytiques, contenant les principes du calcul différentiel, dégagés de toute considération d'infiniment petits, d'évanouissans, de limites ou de fluxions, et réduits à l'analyse algébrique des quantités finies"

Lagrange, who in his youth made his greatest triumphs by free and masterful manipulations of differentials, later in his life condemned them vigorously by replacing "differentials" by "derivatives" and "integrals" by "primitives", see the quote above. Under the influence of Cayley's matrix notation, the above theory subsequently took a different shape, the one we are used to seeing today: we first consider a finite dimensional optimization problem with constraints, and show how the Lagrange multipliers are none other than multipliers like in Gaussian elimination, but without using the notation of differentials that were essential in their invention, as we have seen earlier. This will also reveal a further advantage over the direct solution of the complete optimality system in the presence of constraints, since the system obtained with Lagrange multipliers is much smaller. Suppose we wish to solve the constrained optimization problem

$$f(\boldsymbol{x}) \longrightarrow \min, \quad \boldsymbol{g}(\boldsymbol{x}) = 0, \tag{3.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is the objective function and $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$ are the constraints, $m < n$. To eliminate the constraints, we partition the vector $\boldsymbol{x}$ into $\boldsymbol{x} = (\boldsymbol{y}, \boldsymbol{u})$, $\boldsymbol{y} \in \mathbb{R}^m$, $\boldsymbol{u} \in \mathbb{R}^{n-m}$, and invoke the implicit function theorem to obtain $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{u})$ from the constraint $\boldsymbol{g}(\boldsymbol{x}) = 0$. Substituting this into the objective function, we obtain the unconstrained optimization problem

$$f(\boldsymbol{y}(\boldsymbol{u}), \boldsymbol{u}) \longrightarrow \min . \tag{3.2}$$

A necessary condition for a local minimum is therefore

$$\frac{df}{d\boldsymbol{u}} = \frac{\partial f}{\partial \boldsymbol{y}} \cdot \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{u}} + \frac{\partial f}{\partial \boldsymbol{u}} := (Y_u^T \nabla_y f + \nabla_u f)^T = 0, \tag{3.3}$$

where $Y_u : \mathbb{R}^{n-m} \to \mathbb{R}^{m \times (n-m)}$ is the Jacobian of the implicit function $\boldsymbol{y}(\boldsymbol{u})$, and $\nabla_y f = f_y^T$ and $\nabla_u f = f_u^T$ are the gradients (column vectors) of the objective function with respect to the variables $\boldsymbol{y}$ and $\boldsymbol{u}$. The necessary optimality condition (3.3) is a small system involving the $n-m$ unknowns in the vector $\boldsymbol{u}$ only. However, only in very simple situations it is actually possible to explicitly form the function $\boldsymbol{y}(\boldsymbol{u})$ and differentiate it to obtain Y_u. In general, the Jacobian matrix Y_u is also unknown and depends implicitly on the solution $\boldsymbol{y}$, which must also be calculated. To obtain equations for $\boldsymbol{y}$, one can directly use the constraint $\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{u}) = 0$, and for the Jacobian, one can write the total derivative with respect to $\boldsymbol{u}$ of $\boldsymbol{g}(\boldsymbol{y}(\boldsymbol{u}), \boldsymbol{u}) = 0$. This leads to the complete optimality system

$$Y_u^T \nabla_y f + \nabla_u f = 0, \tag{3.4}$$

$$Y_u^T G_y^T + G_u^T = 0, \tag{3.5}$$

$$\boldsymbol{g} = 0, \tag{3.6}$$

where $G_y : \mathbb{R}^n \to \mathbb{R}^{m\times m}$ is the Jacobian matrix of $\boldsymbol{g}$ with respect to $\boldsymbol{y}$, and $G_u : \mathbb{R}^n \to \mathbb{R}^{m\times(n-m)}$ is the Jacobian matrix of $\boldsymbol{g}$ with respect to $\boldsymbol{u}$. Equation (3.4) contains $n-m$ equations, (3.5) is a matrix equation for the Jacobian matrix Y_u and contains a total of $m(n-m)$ equations, and (3.6) contains m equations from the constraints. This gives a total of $n + m(n-m)$ equations for the n unknowns in $\boldsymbol{y}$ and $\boldsymbol{u}$ combined, and the $m(n-m)$ unknowns in the Jacobian Y_u, a very big system. The key idea of Lagrange in this setting is that one can eliminate many of these equations using Gaussian elimination to arrive at a smaller, but equivalent system. If the Jacobian G_y is invertible, then multiplying the matrix-valued equation (3.5) by the vector-valued multiplier $\boldsymbol{\lambda} := -G_y^{-T}\nabla_y f$ from the right yields

$$Y_u^T G_y^T \boldsymbol{\lambda} + G_u^T \boldsymbol{\lambda} = -Y_u^T G_y^T G_y^{-T} \nabla_y f + G_u^T \boldsymbol{\lambda} = -Y_u^T \nabla_y f + G_u^T \boldsymbol{\lambda} = 0. \tag{3.7}$$

Adding this equation to (3.4), the cumbersome term with the large Jacobian matrix cancels and we obtain the smaller but equivalent optimality system

$$\nabla_u f + G_u^T \boldsymbol{\lambda} = 0, \tag{3.8}$$

$$\nabla_y f + G_y^T \boldsymbol{\lambda} = 0, \tag{3.9}$$

$$\boldsymbol{g} = 0, \tag{3.10}$$

which now contains $(n-m)+m+m = n+m$ equations for the n unknowns $\boldsymbol{y}$ and $\boldsymbol{u}$ combined, plus the m Lagrange multipliers $\boldsymbol{\lambda}$. The system (3.8–3.10) is equivalent to (3.4–3.6), and therefore represents the same necessary condition for a minimum of the original constraint problem (3.1), but it has the advantage of having many fewer unknowns to solve for. The key observation of Lagrange now was that this simpler necessary condition for optimality can be easily obtained from the function

$$\mathcal{L}(\boldsymbol{u}, \boldsymbol{y}, \boldsymbol{\lambda}) := f(\boldsymbol{y}, \boldsymbol{u}) + g(\boldsymbol{y}, \boldsymbol{u})^T \boldsymbol{\lambda}, \tag{3.11}$$

by simply taking derivatives with respect to its arguments. The function in (3.11), now known as the Lagrange function or the Lagrangian in honor of its inventor, is obtained by simply adding to the objective function the sum of the constraints, each multiplied by a Lagrange multiplier.

The new formulation, however, introduces an important difficulty when the remaining $\boldsymbol{u}$ variables are not allowed to vary freely, but are constrained to be in a closed set U. This is often the case in optimal control problems, since the controls may not be arbitrarily large. Then the necessary condition (3.3) for a minimum solution of (3.2) is only relevant if the minimum is in the interior of

U; when the minimum occurs on the boundary, which often happens in practice, the condition (3.3) need not be satisfied, i.e., the variation of the Lagrangian with respect to $\boldsymbol{u}$ in (3.8) need not vanish. One possibility in that case is to revert to the minimization condition of the Lagrangian with respect to $\boldsymbol{u}$, which leads to the necessary conditions for optimality

$$\mathcal{L}(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{\lambda}) \longrightarrow \min \quad \text{with respect to } \boldsymbol{u} \tag{3.12}$$

$$\nabla_y f + G_y^T \boldsymbol{\lambda} = 0, \tag{3.13}$$

$$\boldsymbol{g} = 0. \tag{3.14}$$

Since the constraint $\boldsymbol{g} = 0$ must be satisfied at the optimum, we have $\mathcal{L}(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{\lambda}) = f(\boldsymbol{y}, \boldsymbol{u})$ there, so (3.12) is equivalent to saying that

$$f(\boldsymbol{y}, \boldsymbol{u}) \longrightarrow \min \quad \text{with respect to } \boldsymbol{u}. \tag{3.15}$$

In this case, however, the equation (3.13) for the Lagrange multipliers is no longer needed, since they are not used anywhere in the system; if we remove it, we just get back the original problem formulation (3.1), except that one now sees explicitly that the minimization is only possible with respect the remaining "control" variables $\boldsymbol{u}$, since the other variables $\boldsymbol{y}$ are determined by the constraints. Nevertheless, the observation to replace the derivative condition again by the minimization condition points in the direction of results obtained by Pontryagin and his group and leads to the maximum principle for optimal control problems. We will see later that they chose a different function, a Hamiltonian, which has the same stationary points in $\boldsymbol{u}$ as the Lagrangian.[7]

A different way of characterizing minima on a closed set of controls U is to ensure that whenever the minimum occurs on the boundary, any variation in $\boldsymbol{u}$ that moves the point away from the boundary into the interior of the closed set must lead to an increase in the objective function, i.e.

$$(\nabla_u f + G_u^T \boldsymbol{\lambda})^T \delta \boldsymbol{u} \geq 0, \tag{3.16}$$

$$\nabla_y f + G_y^T \boldsymbol{\lambda} = 0, \tag{3.17}$$

$$\boldsymbol{g} = 0, \tag{3.18}$$

for all admissible variations $\delta \boldsymbol{u}$ such that $\boldsymbol{u} + \delta \boldsymbol{u}$ remains in the closed set of the admissible controls U. This approach became known under the name Karush–Kuhn–Tucker (KKT) conditions, which we will see again in Sect. 4.2.

[7] See also Carathéodory [16] for a general study of equivalent formulations.

3.2 Lagrange Multipliers for Optimal Control Problems

> Using what I had learned at Columbia about flights of airplanes, I set out to formulate this problem as a variational problem. I found that the usual variational formulation did not fit very well. It was too clumsy. And so I reformulated the Problem of Bolza so that it could be applied easily to the time-optimal problem at hand. It turns out that I had formulated what is now known as the general optimal control problem. I wrote it up as a RAND report [31] and it was widely circulated among engineers. (Hestenes, in a letter to Saunders Mac Lane, see [39])

Optimal control problems were becoming important with the invention of moving high-tech mechanical devices, especially in the context of war. A typical example is to guide an airplane along an optimal trajectory to reach a target, and this was precisely the problem considered by Hestenes in his famous RAND report [31], see also the quote above. Hestenes, who had obtained his Ph.D. on the calculus of variations under the direction of Bliss, was a young professor in Chicago during the Second World War and moved to UCLA afterward. He was also doing research for RAND, a nonprofit institution with the goal of improving policy and decision-making through research and analysis, which still exists today (www.rand.org). In his report, he formulated the problem of guiding an airplane in an optimal way from an initial position to a final position as an optimization problem with a constraint given by a differential equation. In modern notation, the problem reads

$$\int_0^T f(\boldsymbol{y},\boldsymbol{u})dt \longrightarrow \min, \tag{3.19}$$

$$\dot{\boldsymbol{y}} = \boldsymbol{g}(\boldsymbol{y},\boldsymbol{u}), \tag{3.20}$$

$$\boldsymbol{y}(0) = \boldsymbol{y}^0, \tag{3.21}$$

$$\boldsymbol{y}(T) = \boldsymbol{y}_T, \tag{3.22}$$

where the vector $\boldsymbol{y}(t)$ contains the position and velocity vectors of the airplane, and the vector $\boldsymbol{u}(t)$ contains the angles of the control vanes of the airplane and the thrust of the engines. Comparing this optimal control problem with the general constrained minimization problem (3.1), Hestenes noticed the striking similarity, so he applied the Lagrange multiplier technique we saw in Sect. 2.3 to obtain a necessary condition for optimality: he introduced the Lagrangian as in (3.11),

$$\mathcal{L}(\boldsymbol{y},\boldsymbol{u},\boldsymbol{\lambda}) := \int_0^T f(\boldsymbol{y},\boldsymbol{u})dt + \int_0^T (\dot{\boldsymbol{y}} - \boldsymbol{g}(\boldsymbol{y},\boldsymbol{u}))^T \boldsymbol{\lambda} dt, \tag{3.23}$$

where all the variables now depend on time, $\boldsymbol{y} = \boldsymbol{y}(t)$, $\boldsymbol{u} = \boldsymbol{u}(t)$, $\boldsymbol{\lambda} = \boldsymbol{\lambda}(t)$ [this is precisely equation (2.12) in the new notation]. In order to obtain necessary conditions for optimality, he computed the derivatives with respect to the variables $\boldsymbol{y}$, $\boldsymbol{u}$, and $\boldsymbol{\lambda}$ using variational calculus (as Euler did in E420, see [28]): if $\boldsymbol{y}$ is an optimum, then for an arbitrary variation $\boldsymbol{y} + \varepsilon \boldsymbol{z}$, the derivative of $\mathcal{L}(\boldsymbol{y} + \varepsilon \boldsymbol{z}, \boldsymbol{u}, \boldsymbol{\lambda})$

with respect to ε must vanish at $\varepsilon = 0$, regardless of what the variation $\boldsymbol{z}$ is. Thus, we obtain as the first necessary condition

$$\begin{aligned}\frac{d}{d\varepsilon}\mathcal{L}(\boldsymbol{y}+\varepsilon\boldsymbol{z},\boldsymbol{u},\boldsymbol{\lambda})|_{\varepsilon=0} &= \int_0^T \nabla_y f^T(\boldsymbol{y},\boldsymbol{u})\boldsymbol{z}dt + \int_0^T \left(\dot{\boldsymbol{z}} - G_y(\boldsymbol{y},\boldsymbol{u})\boldsymbol{z}\right)^T \boldsymbol{\lambda}\, dt \\ &= \int_0^T (\nabla_y f(\boldsymbol{y},\boldsymbol{u}) - \dot{\boldsymbol{\lambda}} - G_y^T(\boldsymbol{y},\boldsymbol{u})\boldsymbol{\lambda})^T \boldsymbol{z}dt + \boldsymbol{\lambda}^T\boldsymbol{z}|_0^T = 0,\end{aligned}$$

where we used integration by parts to factor out the arbitrary variation $\boldsymbol{z}$, and the fact that

$$(G_y\boldsymbol{z})^T\boldsymbol{\lambda} = \boldsymbol{z}^T G_y^T\boldsymbol{\lambda} = (\boldsymbol{z}^T G_y^T\boldsymbol{\lambda})^T = \boldsymbol{\lambda}^T G_y\boldsymbol{z} = (G_y^T\boldsymbol{\lambda})^T\boldsymbol{z}.$$

Now the variation $\boldsymbol{z}(t)$ must be zero for $t = 0$ and $t = T$, since the values of $\boldsymbol{y}$ are fixed there, see (3.21) and (3.22); thus, we have $\boldsymbol{z}(0) = \boldsymbol{z}(T) = 0$, so the boundary terms $\boldsymbol{\lambda}^T\boldsymbol{z}|_0^T$ in (3.24) must vanish as well. However, apart from the initial and final conditions, the variation $\boldsymbol{z}(t)$ is otherwise arbitrary, and hence from (3.24), the term multiplying $\boldsymbol{z}(t)$ under the integral must be zero. This leads to a differential equation for $\boldsymbol{\lambda}$, namely

$$\dot{\boldsymbol{\lambda}} = -G_y^T(\boldsymbol{y},\boldsymbol{u})\boldsymbol{\lambda} + \nabla_y f(\boldsymbol{y},\boldsymbol{u}), \tag{3.24}$$

without initial or final condition, since $\boldsymbol{y}$ was fixed at both ends. Similarly, since $\boldsymbol{u}$ is optimal, we can add an arbitrary variation $\boldsymbol{u} + \varepsilon\boldsymbol{v}$ and require the derivative of $\mathcal{L}(\boldsymbol{y},\boldsymbol{u}+\varepsilon\boldsymbol{v},\boldsymbol{\lambda})$ with respect to ε to vanish at $\varepsilon = 0$ for all variations $\boldsymbol{v}$. This yields the next necessary condition

$$\begin{aligned}\frac{d}{d\varepsilon}\mathcal{L}(\boldsymbol{y},\boldsymbol{u}+\varepsilon\boldsymbol{v},\boldsymbol{\lambda})|_{\varepsilon=0} &= \int_0^T \nabla_u f^T(\boldsymbol{y},\boldsymbol{u})\boldsymbol{v}dt + \int_0^T (-G_u(\boldsymbol{y},\boldsymbol{u})\boldsymbol{v})^T\boldsymbol{\lambda}\, dt \\ &= \int_0^T (\nabla_u f(\boldsymbol{y},\boldsymbol{u}) - G_u^T(\boldsymbol{y},\boldsymbol{u})\boldsymbol{\lambda})^T\boldsymbol{v}dt = 0.\end{aligned}$$

Since the variation $\boldsymbol{u}(t)$ is arbitrary, from (3.25), the term multiplying $\boldsymbol{v}(t)$ under the integral must be zero, which leads to an equation for $\boldsymbol{u}$, namely

$$G_u^T(\boldsymbol{y},\boldsymbol{u})\boldsymbol{\lambda} = \nabla_u f(\boldsymbol{y},\boldsymbol{u}). \tag{3.25}$$

Finally, adding an arbitrary variation $\boldsymbol{\lambda}+\varepsilon\boldsymbol{\mu}$, the derivative of $\mathcal{L}(\boldsymbol{y},\boldsymbol{u},\boldsymbol{\lambda}+\varepsilon\boldsymbol{\mu})$ with respect to ε must vanish at $\varepsilon = 0$ for all variations $\boldsymbol{\mu}$, and we obtain as the last necessary condition

$$\frac{d}{d\varepsilon}\mathcal{L}(\boldsymbol{y},\boldsymbol{u},\boldsymbol{\lambda}+\varepsilon\boldsymbol{\mu})|_{\varepsilon=0} = \int_0^T (\dot{\boldsymbol{y}} - \boldsymbol{g}(\boldsymbol{y},\boldsymbol{u}))^T\boldsymbol{\mu}dt = 0, \tag{3.26}$$

and we simply get back the equations of motion. Hence, for an optimal control problem, we get from the Lagrange multiplier rule a system of necessary conditions for optimality that is very similar to the classical conditions (3.8–3.10), and identical to (2.13):

$$\nabla_u f(\boldsymbol{y},\boldsymbol{u}) - G_u^T(\boldsymbol{y},\boldsymbol{u})\boldsymbol{\lambda} = 0, \tag{3.27}$$

$$\nabla_y f(\boldsymbol{y},\boldsymbol{u}) - G_y^T(\boldsymbol{y},\boldsymbol{u})\boldsymbol{\lambda} = \dot{\boldsymbol{\lambda}}, \tag{3.28}$$

$$\boldsymbol{g}(\boldsymbol{y},\boldsymbol{u}) = \dot{\boldsymbol{y}}, \tag{3.29}$$

the only difference is that the sign is flipped on the G terms, because this is how we introduced the constraints, and that a term with a time derivative appears on the right, because the constraint is an ordinary differential equation. This system contains precisely enough equations for the number of unknowns: there are as many algebraic equations in (3.27) as unknowns in $\boldsymbol{u}(t)$ for $t \in [0,T]$, and (3.28)–(3.29) is a coupled first-order system of ordinary differential equations in $\boldsymbol{y}(t)$ (optimal trajectory) and $\boldsymbol{\lambda}(t)$ (multipliers) with precisely two boundary conditions at $t = 0$ and $t = T$ (both on the unknown $\boldsymbol{y}$ in our case). Hestenes was therefore able to solve this coupled system numerically to obtain candidates for the optimal trajectory.

The optimality system (3.27–3.29) reveals a very interesting mathematical structure.[8] Defining the Hamiltonian function

$$H(\boldsymbol{y},\boldsymbol{u},\boldsymbol{\lambda}) := -f(\boldsymbol{y},\boldsymbol{u}) + \boldsymbol{g}(\boldsymbol{y},\boldsymbol{u})^T\boldsymbol{\lambda}, \tag{3.30}$$

we see that the boundary value problem (3.28), (3.29) is in fact given by

$$\begin{aligned} \dot{\boldsymbol{y}} &= \nabla_\lambda H(\boldsymbol{y},\boldsymbol{u},\boldsymbol{\lambda}) = \boldsymbol{g}(\boldsymbol{y},\boldsymbol{u}), \\ \dot{\boldsymbol{\lambda}} &= -\nabla_y H(\boldsymbol{y},\boldsymbol{u},\boldsymbol{\lambda}) = -G_y^T(\boldsymbol{y},\boldsymbol{u})\boldsymbol{\lambda} + \nabla_y f(\boldsymbol{y},\boldsymbol{u}), \end{aligned} \tag{3.31}$$

where $\nabla_y H = H_y^T$ and $\nabla_\lambda H = H_\lambda^T$. Therefore, we have a Hamiltonian system, which has the property that

$$\frac{d}{dt}H(\boldsymbol{y},\boldsymbol{u},\boldsymbol{\lambda}) = H_y\dot{\boldsymbol{y}} + H_u\dot{\boldsymbol{u}} + H_\lambda\dot{\boldsymbol{\lambda}} = H_y\nabla_\lambda H + H_u\dot{\boldsymbol{u}} + H_\lambda(-\nabla_y H) = 0 \tag{3.32}$$

along optimal trajectories, since $H_u^T = \nabla_u H = -\nabla_u f(\boldsymbol{y},\boldsymbol{u}) + G_u^T(\boldsymbol{y},\boldsymbol{u})\boldsymbol{\lambda} = 0$ whenever the optimality condition (3.27) holds. Thus, the Hamiltonian is conserved in this case. The fact that the derivative of the Hamiltonian (3.30) with respect to the controls $\boldsymbol{u}$ coincides with the corresponding derivatives of the Lagrangian in (3.23),

$$\nabla_u H = -\nabla_u f + G_u^T\boldsymbol{\lambda} = -\nabla_u \mathcal{L}, \tag{3.33}$$

[8] This was already discovered by Carathéodory [16], see also Sect. 3.7.

implies that an identical necessary condition for an interior minimum in the controls $\boldsymbol{u}$ can be obtained from both the Lagrangian and the Hamiltonian. Instead of minimizing the Lagrangian (3.23) with respect to the controls $\boldsymbol{u}$, which means minimizing the objective function on an optimal trajectory satisfying $\boldsymbol{g}(\boldsymbol{y},\boldsymbol{u}) = 0$

$$\int_0^T f(\boldsymbol{y},\boldsymbol{u})dt \longrightarrow \min \quad \text{with respect to } \boldsymbol{u}(t), \tag{3.34}$$

one could also maximize the Hamiltonian (3.30)

$$H(\boldsymbol{y},\boldsymbol{u},\boldsymbol{\lambda}) \longrightarrow \max \quad \text{with respect to } \boldsymbol{u}(t), \tag{3.35}$$

pointwise for each $t \in [0,T]$. Minimizing the Lagrangian (3.34) just leads back to the original problem formulation (3.19–3.22), since $\boldsymbol{\lambda}$ disappears from the optimality system (3.27–3.29) when (3.27) is replaced by (3.34). However, maximizing the Hamiltonian (3.35) leads to a new problem formulation

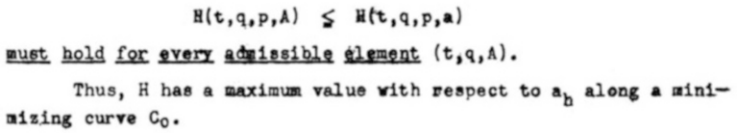
$$H(\boldsymbol{y},\boldsymbol{u},\boldsymbol{\lambda}) \longrightarrow \max \quad \text{with respect to } \boldsymbol{u}(t), \tag{3.36}$$

$$\dot{\boldsymbol{y}} = \nabla_{\lambda} H(\boldsymbol{y},\boldsymbol{u},\boldsymbol{\lambda}), \tag{3.37}$$

$$\dot{\boldsymbol{\lambda}} = -\nabla_{y} H(\boldsymbol{y},\boldsymbol{u},\boldsymbol{\lambda}), \tag{3.38}$$

since $\boldsymbol{\lambda}$ does not disappear from this new optimality system (3.36–3.38). This was already noticed by Hestenes in his famous RAND report from 1950, see Fig. 17. At the time, due to the lack of computing power, Hestenes was unable to solve the optimality system numerically. However, it was only a matter of time before digital computers became available, and Hestenes already anticipated this development in his manual to engineers, see Plail [46].

There is however a very important issue we did not address so far in the above attempt for optimizing the controls: the controls $\boldsymbol{u}$ of the airplane may not take on arbitrary values, but are instead confined to a closed and bounded set, since the thrust of the engine cannot be arbitrarily large, and the control vanes of the airplane cannot turn arbitrarily far. The optimality system (3.27–3.29) is therefore only a necessary condition if the solution lies in the interior of the domain of controls; the formulation in its present form cannot identify potential optima on the boundary of


H(t,q,p,A) ≤ H(t,q,p,a)

must hold for every admissible element (t,q,A).

Thus, H has a maximum value with respect to a_b along a minimizing curve C_0.

Fig. 17 Hestenes' discovery that the Hamiltonian must be maximized along a minimizing solution in the RAND report from 1950

the range of the controls because (3.27), which comes from requiring the derivative with respect to the controls $\boldsymbol{u}$ to be zero, need not hold on the boundary. We see however that the new optimality system (3.36–3.38), written with the Hamiltonian, does not have this problem and deals with the optimal trajectories correctly, even when the control $\boldsymbol{u}$ lies on the boundary, since the minimization is not characterized by a derivative. Next, we will see how this insight was found historically, and led to the famous maximum principle of Pontryagin.

3.3 Early Non-classical Optimal Control Problems

An interesting problem, very much related to the fact that the controls in many real applications must be bounded, was studied by Feldbaum in Russia in [22]: he considered the problem of guiding an object from one position to another with a control that can only take two states, a so-called "bang–bang system" of second order. This was modeled by the equation of motion

$$\ddot{y} = \pm M, \tag{3.39}$$

and the goal was to determine, for a given control strength constant M, when to choose the positive and when to choose the negative sign in order to go as quickly as possible from an initial position $y(0)$ at initial speed $\dot{y}(0)$ back to the origin at rest, i.e. $y(T) = \dot{y}(T) = 0$. Here, the controls are a discrete set, and depending on the sign chosen, we get the general solution branches by integration,

$$\begin{aligned}
\dot{y}^{\pm} &= \pm Mt + C_1^{\pm}, \\
y^{\pm} &= \pm\frac{1}{2M}(\pm Mt + C_1^{\pm})^2 + C_2^{\pm} \;=\; \pm\frac{1}{2M}(\dot{y}^{\pm})^2 + C_2^{\pm}.
\end{aligned}$$

Because $y^{\pm}$ is a quadratic function of $\dot{y}^{\pm}$, these solution branches are best drawn in phase space, where $y^{\pm}$ is a parabola as a function of $\dot{y}^{\pm}$ centered at $\dot{y}^{\pm} = 0$, as illustrated in Fig. 18 on the left.

On the red dashed curves, the control $-M$ is active, and we are moving from the right to the left. On the blue dashed-dotted curves, the control M is active, and we are moving from left to right. There are only two curves, shown as solid lines, that pass through the target $y(T) = \dot{y}(T) = 0$, namely $y^{\pm} = \pm\frac{1}{2M}(\dot{y}^{\pm})$, and from any point along these curves, the fastest is just to stay on these curves with the corresponding control. Now from any point in the phase space to the right of this solid curve, one can use the control $-M$ to arrive as quickly as possible on the blue solid curve, where the control has to be switched to M to arrive at the origin. An example of such a trajectory is shown in Fig. 18 in black. Similarly, from any point in the phase space to the left of the solid curve, one can use the control M to arrive as quickly as possible on the red solid curve, where the control has to be

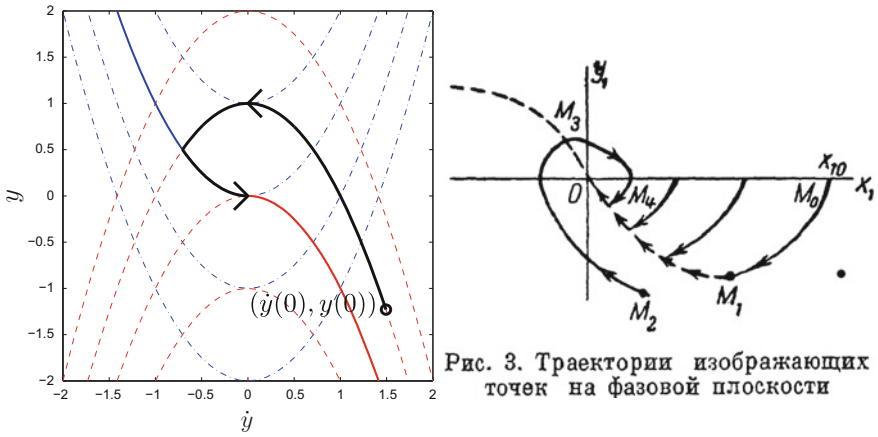


Fig. 18 Solutions of the bang–bang system of Feldbaum from 1949 on the *left*, and an original drawing of Feldbaum from 1949 leading to his understanding of the bang–bang solution

switched to $-M$ to arrive at the origin. In a follow-up paper [23] published 4 years later, Feldbaum made the key step of allowing not only the discrete set of controls $\{-M, M\}$, but the entire continuum of all controls in the closed interval $[-M, M]$, and the problem (3.39) became

$$\ddot{y} = \pm u, \quad |u| \leq M. \tag{3.40}$$

It was at this moment that the notational convention of using u for the control was born. Feldbaum gave a precise mathematical formulation of the minimum time problem for (3.40), and proved that for every initial point in the phase space, there exists a unique time-optimal control $u(t)$ which is still the bang–bang solution found for the control problem with only two discrete controls (3.39): on the optimal trajectory, the control is never used from within the interior of the interval $[-M, M]$! This was the first solution of what Boltyanski calls in his review [8] a non-classical variational problem. Bushaw made a similar discovery in his Ph.D. thesis [13], see also [14]. Feldbaum then generalized this result in two follow-up papers [24, 25] to higher order problems of the form

$$y^{(n)} = -\sum_{j=0}^{n-1} a_j y^{(j)} + u, \quad |u| \leq M,$$

and proved what he called the n-interval theorem, namely that the optimal control is still piecewise constant with values $\pm M$, and that there are no more than n distinct intervals where the control u is constant. Feldbaum was therefore undoubtedly one of the pioneers in the field of optimal control where the domain of the controls is a closed set.

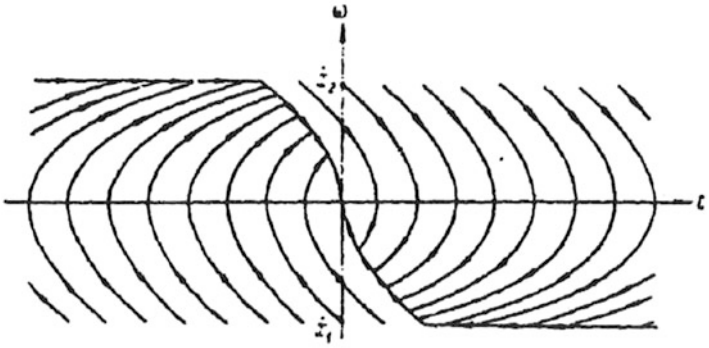


Fig. 19 Lerner's solution to problem (3.39) with an additional inequality constraint on the trajectory

Around the same time, Lerner, also in Russia, considered putting a constraint on the phase coordinates, restricting them to be in a closed set [35, 36]. He considered the same problem as Feldbaum (3.40), but now also with the additional constraint $a_1 \leq y \leq a_2$. Figure 19 shows the solution in that case from his publication [36]. Note that the trajectory constraint is sometimes active, and sometimes not, whereas the control is always on the boundary, i.e., its constraint is always active.

3.4 Invention of the Maximum Principle

> This fact appears in many cases as a general principle, which we call the *maximum principle* (translated from Boltyanski et al. [9], see Fig. 22 for the original)

It was in this context that Pontryagin started to work with his students Boltyanski and Gamkrelidze on optimal control.[9] Pontryagin was known worldwide at the time for his work on homotopic topology, even though he had become blind after an accident involving an explosion at the age of twelve. However, around the 1950s, his results in homotopic topology started to be surpassed by the achievements of the French school around Leray, Serre and Cartan [8], and Pontryagin decided to leave this area of research and focus on the very different area of optimal control. This was in part due to his friendship with A. Andronov, with whom Pontryagin had worked on rough systems, but also because the university administration and the communist party organization encouraged more applied research. Together with his students, Pontryagin started an active research seminar to which engineers were

[9] For more details on the historical context for this development, see Plail [46] and also [45].

also invited, and where the talks always had to have an applied side. Feldbaum also spoke several times at this seminar about his research on optimal control problems. In 1955, Pontryagin's group met Colonel Dobrohotov from the military academy of the Russian air force, and this contact led them to the important problem of guiding a flying object in minimal time in air combat. Even though the problems were not formulated as such, Pontryagin and his group realized immediately that the framework of optimal control was mathematically the correct one.

In their first publication in 1956, see [9], Pontryagin, Boltyanski and Gamkrelidze present the ideas which led them to formulate the maximum principle. There is only one reference in this paper, to Feldbaum's paper from 1955 [24], and the authors refer to the references given there. The problem they consider is to control in a time optimal way the system governed by the equations

$$\frac{d\boldsymbol{y}}{dt} = \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{u}), \quad \boldsymbol{y}(0) = \boldsymbol{y}^0, \quad \boldsymbol{y}(T) = \boldsymbol{y}_T, \tag{3.41}$$

which describe the trajectory $\boldsymbol{y} : \mathbb{R} \to \mathbb{R}^m$ of the object for a given set of control functions $\boldsymbol{u} : \mathbb{R} \to \mathbb{R}^{n-m}$. The precise problem formulation is to find among all admissible controls $\boldsymbol{u}(t)$ the one that leads to the shortest travel time, i.e. $T = T(\boldsymbol{u})$ should be minimized. The authors say right at the beginning that the controls often have to satisfy further constraints, for example $|u_j| \leq 1$. They therefore introduce an open set Ω where the controls live, and also its closure $\bar{\Omega}$, and carefully distinguish these two cases for the control. They start with the control in the open set Ω, where one could easily derive optimality conditions using Lagrange multipliers. However, since the group of Pontryagin had their roots in a different field from variational calculus, they derive the optimality conditions with their bare hands: they assume existence of an optimal control $\boldsymbol{u}$, and derive a necessary optimality condition by considering a variation of the control $\boldsymbol{u}(t) + \delta\boldsymbol{u}(t)$ and the associated variation in the trajectory $\boldsymbol{y}(t) + \delta\boldsymbol{y}(t)$. Inserting these variations into the equations of motion (3.41), we obtain

$$\frac{d\boldsymbol{y}}{dt} + \frac{d\,\delta\boldsymbol{y}}{dt} = \boldsymbol{g}(\boldsymbol{y} + \delta\boldsymbol{y}, \boldsymbol{u} + \delta\boldsymbol{u}) = \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{u}) + G_y\delta\boldsymbol{y} + G_u\delta\boldsymbol{u},$$

and therefore the variation in the trajectory satisfies the linear inhomogeneous system of ordinary differential equations

$$\frac{d\,\delta\boldsymbol{y}}{dt} = G_y\delta\boldsymbol{y} + G_u\delta\boldsymbol{u}, \tag{3.42}$$

where $G_u\delta\boldsymbol{u}$ plays the role of the forcing term. Now the initial condition for the motion is fixed, and therefore the initial variation $\delta\boldsymbol{y}(0)$ must vanish. Using the technique of variation of constants, we can solve the system (3.42) as follows: if we denote by the matrix $Y(t)$ the solution of the linear homogeneous system

$$\dot{Y} = G_y Y, \quad Y(0) = I \quad (I \text{ the identity}),$$

В силу линейности системы (2) точки $x(t_1) + \delta_1 x(t_1)$, соответствующие всевозможным, достаточно малым по модулю, возмущениям $\delta_1 u(t)$, заполняют область некоторого линейного многообразия P', проходящего через точку $x(t_1)$. Из оптимальности траектории $x(t)$ легко вытекает, что размерность многообразия P' не превосходит $n-1$ и P', вообще говоря, не касается траектории $x(t)$. Пусть $P(t_1)$ — некоторая $(n-1)$-мерная плоскость, содержащая P' и не касающаяся траектории $x(t)$. Ковариантные координаты $(n-1)$-мерной плоскости $P(t_1)$ обозначим через $a_1, \ldots, a_n$; тогда $a_\alpha \delta_1 x^\alpha(t_1) = 0$.

Fig. 20 Geometric idea of Pontryagin, leading to the adjoint equation without knowing about Lagrange multipliers (see text for a translation)

the general solution of the homogeneous part of (3.42) is given by $Y\boldsymbol{c}$ for an arbitrary constant vector $\boldsymbol{c}$. Now varying the constant by setting $z := Y\boldsymbol{c}(t)$, we get

$$\dot{z} = \dot{Y}\boldsymbol{c} + Y\dot{\boldsymbol{c}} = G_y z + Y\dot{\boldsymbol{c}}.$$

By letting $z = \delta\boldsymbol{y}$ and comparing with (3.42), we get $Y\dot{c} = G_u\delta\boldsymbol{u}$, and hence $\boldsymbol{c} = \boldsymbol{c}_0 + \int_0^t Y^{-1}(\tau)G_u\delta\boldsymbol{u}(\tau)d\tau$. The solution of (3.42) is thus given by $\delta\boldsymbol{y} = Y\boldsymbol{c}$, and with the zero initial condition, we obtain

$$\delta\boldsymbol{y}(t) = Y(t)\int_0^t Y^{-1}(\tau)G_u\delta\boldsymbol{u}(\tau)d\tau. \tag{3.43}$$

Now the end point is fixed as well, $\boldsymbol{y}(T) = \boldsymbol{y}_T$, but the time at which the solution trajectory passes through this endpoint is not. Pontryagin argues as shown in Fig. 20, which translated to English says (we use in the translation the symbols and equation numbers used in our presentation, instead of the original ones):

> Because of the linearity of system (3.42), the points $\boldsymbol{y}(T)+\delta\boldsymbol{y}(T)$ which correspond to any sufficiently small perturbation $\delta\boldsymbol{u}$ fill the whole range of some linear mapping P', which passes through $\boldsymbol{y}(T)$. From the optimality of the trajectory $\boldsymbol{y}(t)$, it is easy to see that the dimension of the range of P' does not exceed $m-1$, and P', in general, does not touch the trajectory $\boldsymbol{y}(t)$. Let $P(T)$ be some $m-1$ dimensional surface which contains P' and does not touch the trajectory $\boldsymbol{y}(t)$. Let the covariant coordinates of this $m-1$ dimensional surface $P(T)$ be $a_1, a_2, \ldots, a_m$. Then $\boldsymbol{a}^T\delta\boldsymbol{y}(T) = 0$.

It seems that this insight was obtained by Pontryagin very rapidly over two or three sleepless nights, see [27, 46].[10] To understand his argument, Fig. 21 is useful:

[10]Personal communication of Plail with Boltyanski, and explanation by Gamkrelidze in his paper about the discovery of the maximum principle:

> The first and the most important step toward the final solution was made by L.S. right after the formulation of the problem, during three days, or better to say, during three consecutive sleepless nights.

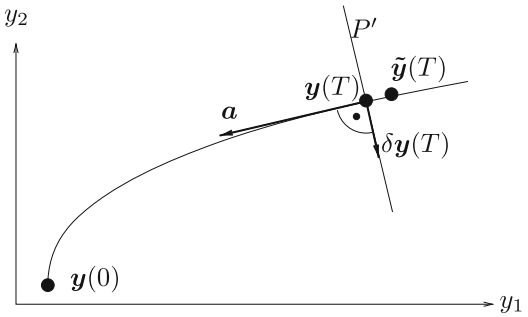


Fig. 21 Explanation of Pontryagin's geometric idea

If the trajectory $\boldsymbol{y}(t)$ is optimal, no variation $\delta\boldsymbol{u}(t)$ is allowed to produce a trajectory $\tilde{\boldsymbol{y}}(t)$ with $\tilde{\boldsymbol{y}}(T)$ beyond $\boldsymbol{y}(T)$, since otherwise this trajectory could have arrived at $\boldsymbol{y}(T)$ at a time $t < T$. Therefore, variations are only allowed to be orthogonal to the optimal trajectory,[11] in a manifold P' of dimension at most $m-1$, where $m = 2$ in the two dimensional example in Fig. 21. There must therefore exist a vector $\boldsymbol{a}$ orthogonal to this manifold, $\boldsymbol{a}^T\delta\boldsymbol{y}(T) = 0$. Since we know the solutions for the variations from (3.43), we can compute

$$\boldsymbol{a}^T\delta\boldsymbol{y}(T) = \boldsymbol{a}^T Y(T)\int_0^T Y^{-1}(\tau)G_u\delta u(\tau)d\tau = \int_0^T \boldsymbol{\psi}^T(\tau)G_u\delta u(\tau)d\tau = 0, \tag{3.44}$$

where we defined the vector $\boldsymbol{\psi}(t) := Y^{-T}(t)Y^T(T)\boldsymbol{a}$. This vector is solution to a differential equation: taking a time derivative of the identity $Y^{-1}Y = I$, we get

$$(\dot{Y^{-1}})Y + Y^{-1}\dot{Y} = 0 \implies (\dot{Y^{-1}}) = -Y^{-1}G_y \implies (\dot{Y^{-T}}) = -G_y^T Y^{-T},$$

and hence $\boldsymbol{\psi}$ is the solution of the differential equation

$$\dot{\boldsymbol{\psi}} = -G_y^T(\boldsymbol{y},\boldsymbol{u})\boldsymbol{\psi}, \tag{3.45}$$

with final condition $\boldsymbol{\psi}(T) = \boldsymbol{a}$. Since the variation $\delta\boldsymbol{u}$ is arbitrary in (3.44), the term under the integral sign must vanish, and Pontryagin and his students obtained the classical necessary conditions for an interior maximum

[11] In fact, since the endpoint is fixed as well, no variations are allowed at the endpoint either, but then Pontryagin could not have obtained the solution (3.43) of the then overdetermined system of ordinary differential equations (3.42), and thus he decided to first only fix the starting point [27, p. 442]. This flaw was only later fixed by Boltyanski, see the end of this subsection.

$$\boldsymbol{\psi}^T G_u(\boldsymbol{y},\boldsymbol{u}) = 0, \tag{3.46}$$

$$\dot{\boldsymbol{\psi}} = -G_y^T(\boldsymbol{y},\boldsymbol{u})\boldsymbol{\psi}, \tag{3.47}$$

$$\dot{\boldsymbol{y}} = \boldsymbol{g}(\boldsymbol{y},\boldsymbol{u}), \quad \boldsymbol{y}(0) = \boldsymbol{y}^0, \quad \boldsymbol{y}(T) = \boldsymbol{y}_T, \tag{3.48}$$

which is just a special case of (3.27–3.29),[12] with $\boldsymbol{\psi}$ playing the role of the Lagrange multiplier $\boldsymbol{\lambda}$, and with an objective function f that depends neither on $\boldsymbol{y}$ nor on $\boldsymbol{u}$. Pontryagin, however, did not know of the relation between this and the Lagrangian at the time of publication; according to Boltyanski [8], they only learned about this several months later when reading the Russian translation of Bliss' monograph [5] from 1946.

Next, the authors note that the functions $\boldsymbol{\psi}$ can be multiplied by a convenient constant in order to obtain $\boldsymbol{\psi}^T\boldsymbol{g}(\boldsymbol{y},\boldsymbol{u})|_{t=0} > 0$ without causing any changes to the necessary conditions for optimality (3.46–3.48), since this quantity is conserved along optimal trajectories, see (3.32). This then implies $\boldsymbol{\psi}^T\boldsymbol{g}(\boldsymbol{y},\boldsymbol{u}) > 0$ for all t. Now if the control $\boldsymbol{u}$ is only allowed to vary in the closed set $\bar{\Omega}$, the authors explain that the first condition (3.46) needs to be replaced by

$$\boldsymbol{\psi}^T G_u(\boldsymbol{y},\boldsymbol{u})\boldsymbol{\delta u} \leq 0 \tag{3.49}$$

for all admissible variations $\boldsymbol{u} + \boldsymbol{\delta u}$ that remain in $\bar{\Omega}$. With this modification, the optimal control may now also be on the boundary. This remark could have led them directly to the KKT system (3.16).

The second result in [9] is a sufficient condition for optimality, obtained according to [8] by Gamkrelidze, and again only for points in the interior of the control domain. The result is based on second variations of the function $\boldsymbol{\psi}^T\boldsymbol{g}(\boldsymbol{y},\boldsymbol{u})$, whose first derivative with respect to $\boldsymbol{u}$ was in the necessary condition for optimality in (3.46). With the change in sign such that $\boldsymbol{\psi}^T\boldsymbol{g}(\boldsymbol{y},\boldsymbol{u}) > 0$, Gamkrelidze showed that if, in addition to (3.46–3.48), the Hessian of $\boldsymbol{\psi}^T\boldsymbol{g}(\boldsymbol{y},\boldsymbol{u})$ with respect to $\boldsymbol{u}$ is negative definite at $t = 0$, then the control $\boldsymbol{u}(t)$ and associated trajectory $\boldsymbol{y}(t)$ are optimal in a neighborhood of $t = 0$. This sufficient condition was not a new result either, as it is a particular case of the sufficient condition of the Legendre type [5, Chap. IX], which the authors did not know at that time. They then however note that if the Hessian is indefinite, then there is no optimal control in the interior of Ω, so any optimal control inside the closed set $\bar{\Omega}$ of admissible controls must occur on the boundary.

[12] To solve the time optimal control problem correctly using Lagrange multipliers, we need to introduce the time variable as a state variable, $y_0(t) := t$, which implies $\dot{y}_0 = 1$, $y_0(0) = 0$. The correct Lagrangian then becomes $\mathcal{L}(\boldsymbol{y},\boldsymbol{\lambda},\boldsymbol{u}) = y_0(T) + \int_0^T \boldsymbol{\lambda}^T(\dot{\boldsymbol{y}} - \boldsymbol{g}(\boldsymbol{y},\boldsymbol{u}))dt$, where all vectors are now one element longer. Computing the variational derivative with respect to $\boldsymbol{y}$, we obtain now in addition to the earlier equations $\dot{\lambda}_0 = 0$ and $z_0(T) + \lambda_0(T)z_0(T) = 0$ for arbitrary variation z_0, which implies $\lambda_0(T) = -1$ and hence $\lambda_0(t) = -1$ to complete the time optimality system with $y_0(t) := t$.

The authors then conclude, based on the necessary conditions (3.46–3.48) and the fact that the Hessian of $\boldsymbol{\psi}^T \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{u})$ with respect to $\boldsymbol{u}$ must be negative definite for optimality, that the Hamiltonian $H(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{\psi}) := \boldsymbol{\psi}^T \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{u})$ must attain a local maximum in $\boldsymbol{u}(t)$ for fixed $\boldsymbol{y}(t)$ and $\boldsymbol{\psi}(t)$ satisfying (3.46–3.48): under the condition that the variations $\boldsymbol{\delta u}$ are admissible and small enough, the inequality

$$\boldsymbol{\psi}^T \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{u}) \geq \boldsymbol{\psi}^T \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{u} + \boldsymbol{\delta u}) \tag{3.50}$$

must hold for all time whenever (3.46–3.48) are satisfied and the Hessian is negative definite.

This was the historical moment of the invention of the maximum principle. The Hamiltonian could also be used to define the important differential equations involved, see Fig. 22 for the original paragraph in Russian, which translates as (we use again the notation from our text in the translation):

> This fact appears in many cases as a general principle, which we call the *maximum principle* (we have only proved this principle so far for several special cases): Let $H(\boldsymbol{y}, \boldsymbol{u}) = \boldsymbol{\psi}^T \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{u})$ have, for arbitrary but fixed $\boldsymbol{y}$, $\boldsymbol{\psi}$ a maximum as $\boldsymbol{u}$ varies within the closed set $\bar{\Omega}$. We denote this maximum by $M(\boldsymbol{y}, \boldsymbol{\psi})$. If the $2m$-dimensional vector $(\boldsymbol{y}, \boldsymbol{\psi})$ is a solution of the Hamiltonian system
>
> $$\begin{aligned} \dot{\boldsymbol{y}} &= \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{u}) = \nabla_{\boldsymbol{\psi}} H, \\ \dot{\boldsymbol{\psi}} &= -G_{\boldsymbol{y}}^T \boldsymbol{\psi} = -\nabla_{\boldsymbol{y}} H, \end{aligned}$$
>
> and a piecewise continuous vector $\boldsymbol{u}(t)$ satisfies for each point in time
>
> $$H(\boldsymbol{y}(t), \boldsymbol{\psi}(t), \boldsymbol{u}(t)) = M(\boldsymbol{y}(t), \boldsymbol{\psi}(t)) > 0,$$
>
> then $\boldsymbol{u}(t)$ is the optimal control and $\boldsymbol{y}(t)$ the corresponding (locally) optimal trajectory of system (3.41).

Этот факт является частным случаем следующего общего принципа, который мы называем принципом максимума (принцип этот доказан нами пока лишь в ряде частных случаев):

Пусть функция $H(x, \psi, u) = \psi_\alpha f^\alpha(x, u)$ *при любых фиксированных* x, ψ *имеет максимум по* u, *когда вектор* u *меняется в замкнутой области* $\bar{\Omega}$; *обозначим этот максимум через* $M(x, \psi)$. *Если* $2n$*-мерный вектор* (x, ψ) *является решением гамильтоновой системы*

$$\left.\begin{aligned} \dot{x}^i &= f^i(x, u) = \frac{\partial H}{\partial \psi_i}, \\ \dot{\psi}_i &= -\frac{\partial f^\alpha}{\partial x^i}\psi_\alpha = -\frac{\partial H}{\partial x^i}, \end{aligned}\right\} \quad i = 1, \ldots, n, \tag{8}$$

где кусочно-непрерывный вектор $u(t)$ *в каждый момент времени удовлетворяет условию* $H(x(t), \psi(t), u(t)) = M(x(t), \psi(t)) > 0$, *то* $u(t)$ *является оптимальным управлением, а* $x(t)$ *— соответствующей оптимальной (в малом) траекторией системы* (1).

Fig. 22 The historical moment when the maximum principle was invented

This first publication only gave a criterion for the solution of the time optimal control problem, and it was formulated as a sufficient condition. Pontryagin also hoped that the criterion would give the global optimal control, and put the word "locally" in parentheses [8], see also Fig. 22. The maximum principle allowed the authors to immediately solve the Bushaw–Feldbaum problem we have seen earlier,

$$\ddot{y} = u, \quad |u| \leq 1,$$

as follows: we first transform the system to first order

$$\dot{y}_1 = y_2, \quad \dot{y}_2 = u,$$

and the Hamiltonian becomes

$$H = \psi_1 y_2 + \psi_2 u.$$

For the auxiliary functions, we obtain the differential equations

$$\dot{\psi}_1 = 0, \quad \dot{\psi}_2 = -\psi_1.$$

These equations can be easily integrated to give $\psi_1(t) = C_1$ and $\psi_2(t) = C_2 - C_1 t$, where C_1 and C_2 are constants. To maximize H under the condition that $|u| \leq 1$, the control must satisfy

$$u(t) = \text{sign}(\psi_2(t)) = \text{sign}(C_2 - C_1 t),$$

and is therefore piecewise constant and can change at most once, since $\psi_2(t)$ is a linear function of t. We thus obtain precisely the bang–bang solution found by Feldbaum for this problem, but in a very simple way with the maximum principle. The maximum principle also worked very well for many similar problems that could not be solved earlier, which explains the high hopes Pontryagin had for it.

After this first publication, the work was divided by Pontryagin as follows: Gamkrelidze was asked to generalize the results obtained during the calculation of examples, and he quickly found the work by Bellman, Glicksberg and Gross [2], who had established a necessary and sufficient condition for the linear case

$$\dot{\boldsymbol{y}} = A\boldsymbol{y} + B\boldsymbol{u}, \quad |u_j| \leq 1,$$

and the time optimal control to get to $\boldsymbol{y} = 0$. For constant matrices A and B, where the eigenvalues of A have negative real parts, the optimal control is $\boldsymbol{u}^T(t) = \text{sign}(\boldsymbol{b}^T Y(t))$, where $Y(t) = X^{-1}(t)B$ and X solves the matrix equation $\dot{X} = AX$. Here $\boldsymbol{b}$ is an appropriately chosen vector, and the result holds under a general position condition, see [2]. Gamkrelidze managed to show that this necessary and sufficient condition coincides with the maximum principle, and hence for linear

problems, the maximum principle is indeed a necessary and sufficient condition for optimality.

Boltyanski was supposed to work out in detail the results in the first paper [9], and Pontryagin was supposed to find a general justification of the maximum principle. Boltyanski started working on the first result in [9] and tried to formulate it differently from the classical analysis textbook style in which the argument was given, and searched for a geometrical proof. After a more careful study of the second, sufficient condition in [9], Boltyanski finally arrived, "in a brilliant half hour" [8], at the conclusion that the maximum principle was only a necessary condition. He immediately called Pontryagin in his apartment and told him that the maximum principle was only a necessary condition, but a global one. Pontryagin was angry when he received the call because it had woken him up from his afternoon nap, but he called back 5 min later to say that if Boltyanski had really found a proof, this would be of great interest, so it had to be checked carefully. Gamkrelidze did the careful checking, and the argument was correct, so Boltyanski asked Pontryagin if he could publish the results [8]:

> It was proposed to publish it, as a joint paper of four authors. I refused point-blank. Then it was proposed (i) to name that theorem *Pontryagin's maximum principle*, and (ii) to add at the end of my paper a paragraph dictated by Pontryagin that pointed out his role in creation of the principle. Pontryagin was the head of the laboratory in the Steklov Mathematical Institute, and at that time could insist on his interests. I had to agree. After that, my paper was presented to Doklady AN SSSR [7].

Boltyanski indeed named the maximum principle after Pontryagin in the single authored paper [7]:

Высказанный Л. С. Понтрягиным в качестве гипотезы принцип максимума

The maximum principle suggested by Pontryagin as a hypothesis...

and we also show in Fig. 23 the final paragraph dictated by Pontryagin to Boltyanski from the end of the same paper. The literal translation of this paragraph is:

> I got the results which are published in this paper working in the Pontryagin seminar on the theory of oscillations and automatic regulation. Pontryagin pointed out to me one simplification in the proof of the maximum principle, and because of that my proof became applicable to arbitrary topological spaces U (the first variant of the proof contained an unnecessary, actually nowhere used, construction that forced the restriction on the case, when U is a closed domain in a vector space with piecewise-smooth boundary and convex inner corners in breaking points).

As we have seen already in footnote 11, the initial argument of Pontryagin, which allowed the end point to vary in a lower dimensional manifold, was not quite correct. To remove this flaw, Boltyanski resorted in [7] to the tool of needle variations, which already appeared in McShane in 1939 [40]; however, Boltyanski insists that he was unaware of McShane's work at the time and came up with the technique independently [11]. We show in Fig. 24 the hand drawing of Boltyanski from [8]. One can clearly see that a cone appears, instead of the variations orthogonal to the trajectory, and the role of the manifold is now played by Γ at the tip of the cone.

Публикуемые здесь результаты получены мною при работе в руководимом Л. С. Понтрягиным семинаре по теории колебаний и автоматического регулирования. Л. С. Понтрягин указал мне на одно упрощение в доказательстве принципа максимума, благодаря чему мое доказательство стало пригодным для произвольного топологического пространства U (первоначальный вариант доказательства содержал лишнюю, нигде фактически не использовавшуюся конструкцию, которая заставляла ограничиваться случаем, когда U есть замкнутая область векторного пространства с кусочно-гладкой границей и выпуклыми внутренними углами в точках перелома).

Fig. 23 The last paragraph Boltyanski had to add in his single authored paper, dictated by Pontryagin (see translation in the text)

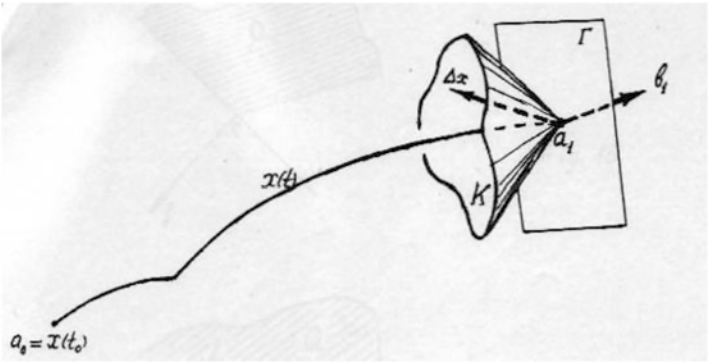


Fig. 24 Original drawing by Boltyanski removing the initial flaw of variations at the endpoint in the proof of the maximum principle

The complete original proof also relies on techniques from topology, the field of origin of the group. It is quite long and technical; details can be found in the historical book by the four authors from 1962 [48], which was quickly translated into many languages and made Pontryagin and the Russian school of optimal control famous with their maximum principle. However, from Boltyanski's point of view, it was he who formulated and proved the maximum principle correctly. Pontryagin's insistence on publishing the result as a joint paper led to a period of deep bitterness for Boltyanski, during which he could not even do mathematics any more, as he tells in [8].

3.5 General Formulation of the Maximum Principle

The times t_0 and t_1, in this statement of the problem, are not fixed. We only require that the object should be in state x_0 at the initial time, and at state x_1 at the final time, and that the functional should achieve a minimum [48].

Pontryagin and his students then generalized the problem of minimizing travel time to one of minimizing an arbitrary function [10]. The model for the technical object is again the system of ordinary differential equations

$$\frac{d\,\boldsymbol{y}}{dt} = \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{u}), \quad \boldsymbol{y}(t_0) = \boldsymbol{y}^0 \tag{3.51}$$

for the trajectory $\boldsymbol{y} : \mathbb{R} \to \mathbb{R}^m$ of the object, depending on the control functions $\boldsymbol{u} : \mathbb{R} \to \mathbb{R}^{n-m}$. These controls are supposed to be chosen such that when the object arrives at time t_1 at a given location $\boldsymbol{y}(t_1) = \boldsymbol{y}^1$, the general functional

$$J := \int_{t_0}^{t_1} g_0(\boldsymbol{y}(t), \boldsymbol{u}(t))dt \tag{3.52}$$

is minimized. Here the scalar function $g_0 : \mathbb{R}^m \times \mathbb{R}^{n-m} \to \mathbb{R}$ was on purpose denoted by the index zero, since a first step was then to define an additional ordinary differential equation

$$\frac{dy_0}{dt} = g_0(\boldsymbol{y}, \boldsymbol{u}), \quad y_0(t_0) = 0.$$

Appending this equation to the system of ordinary differential equations for the technical object as the zeroth coordinate, $\tilde{\boldsymbol{y}} := (y_0, y_1, \ldots, y_m)$, and similarly $\tilde{\boldsymbol{g}} := (g_0, g_1, \ldots, g_m)$, the new system of ordinary differential equations

$$\frac{d\,\tilde{\boldsymbol{y}}}{dt} = \tilde{\boldsymbol{g}}(\tilde{\boldsymbol{y}}, \boldsymbol{u}), \quad \tilde{\boldsymbol{y}}(t_0) = (0, \boldsymbol{y}^0) \tag{3.53}$$

encodes, in addition to the trajectory, also the current value of the objective function in its zeroth component:

$$y_0(t) = \int_{t_0}^{t} g_0(\boldsymbol{y}(t), \boldsymbol{u}(t))dt.$$

The authors now give a geometric interpretation of the optimal control problem in this higher dimensional space: given an initial point $\boldsymbol{y}^0$ and a target $\boldsymbol{y}^1$ in $\mathbb{R}^m$, as shown in Fig. 25, among all the trajectories solution of (3.53) and ending at $\boldsymbol{y}^1$ (dashed line examples in Fig. 25), find the one that crosses the vertical line in the y_0 direction above with the lowest coordinate value $y_0(t_1)$ possible (see solid line in Fig. 25). Next, they explain several properties of this optimal control problem: first, the problem is time invariant, since the right hand side of the state equation and the objective function do not depend on time. One can therefore do translations in time without changing the problem, see Fig. 26 from their book [48]. Because of this, one can also consider several points in phase space, and search for controls separately to move from one to the next sequentially, and then concatenate the controls in order to get a single control to go from the first to the last point in phase space. Doing this,

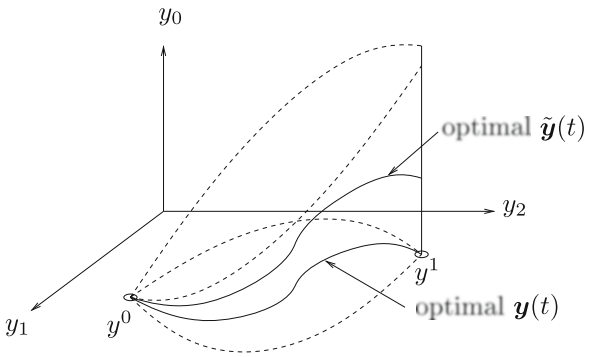


Fig. 25 Interpretation of the optimal control problem in the higher dimensional space including the objective function coordinate y_0

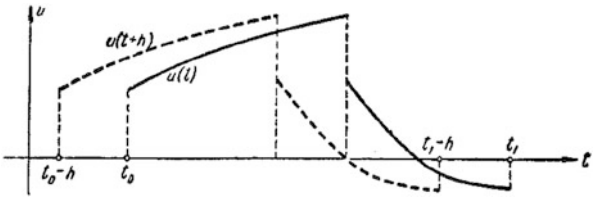


Fig. 26 Graph to illustrate time translation invariance from [48]

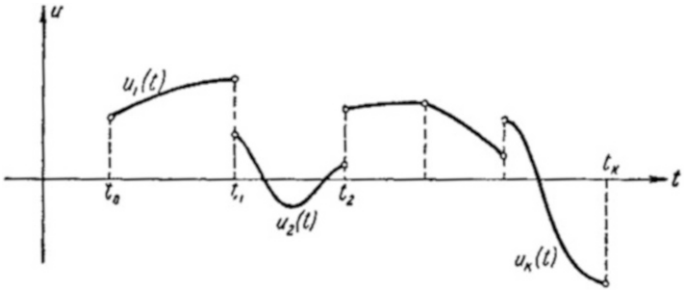


Fig. 27 Graph to illustrate that the optimal controls are piecewise continuous, from [48]

one just has to sum the local objective function values to obtain the global value of the objective function. Concatenating the controls this way, however, is not possible in the space of continuous controls in general, and therefore one must expect the optimal control to be piecewise continuous only, as illustrated in Fig. 27 from [48]. Finally, in preparation of their proof, they argue that the optimal trajectory must also be locally optimal: if it were not optimal on a sub-interval, then one could simply replace the control there by a better one, and since the objective functions are just summed, the global objective function would decrease, see Fig. 28 from [48] for an illustration of this.

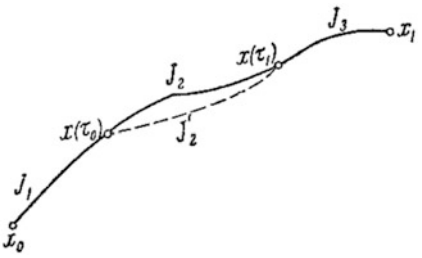


Fig. 28 Graph to illustrate that the solution must be locally optimal, from [48]

For the formal statement of the maximum principle, the authors introduce as before the adjoint system (but now without explanation)

$$\frac{d\tilde{\psi}_i}{dt} = -\sum_{j=0}^{m} \frac{\partial g_j(\boldsymbol{y},\boldsymbol{u})}{\partial y_i}\tilde{\psi}_j, \quad i = 0,1,\ldots,m \tag{3.54}$$

and the Hamiltonian

$$H(\tilde{\boldsymbol{\psi}},\tilde{\boldsymbol{y}},\boldsymbol{u}) := \tilde{\boldsymbol{\psi}}^T \tilde{\boldsymbol{g}}(\boldsymbol{y},\boldsymbol{u}), \tag{3.55}$$

but now the maximum principle is no longer stated as a sufficient condition: a necessary condition for the control $\boldsymbol{u}$ and associated trajectory $\boldsymbol{y}$ to be optimal is that there exist $\boldsymbol{\psi}$ such that the Hamiltonian system

$$\frac{dy_i}{dt} = \frac{\partial H}{\partial \tilde{\psi}_i}, \quad i = 0,1,\ldots,m \tag{3.56}$$

$$\frac{d\tilde{\psi}_i}{dt} = -\frac{\partial H}{\partial y_i}, \quad i = 0,1,\ldots,m \tag{3.57}$$

holds and that for each admissible control $\boldsymbol{v}$ the inequality

$$H(\tilde{\boldsymbol{\psi}},\tilde{\boldsymbol{y}},\boldsymbol{v}) \leq H(\tilde{\boldsymbol{\psi}},\tilde{\boldsymbol{y}},\boldsymbol{u}) \tag{3.58}$$

be satisfied, i.e. the optimal control $\boldsymbol{u}$ is the value of $\boldsymbol{v}$ maximizing the Hamiltonian.

Suppose now that the optimum is in the interior of the domain. Then the inequality (3.58) implies that we are at a stationary point, i.e. the derivative with respect to $\boldsymbol{u}$ must vanish,

$$\tilde{\boldsymbol{\psi}}^T \tilde{G}_u(\boldsymbol{y},\boldsymbol{u}) = 0 \quad \Longleftrightarrow \quad \psi_0 \nabla_u g_0(\boldsymbol{y},\boldsymbol{u}) + G_u^T(\boldsymbol{y},\boldsymbol{u})\boldsymbol{\psi} = 0.$$

Since the Hamiltonian does not depend on y_0, ψ_0 is just a constant, $\psi_0 = -1$ and we find naturally the condition (3.27) from the Lagrange multiplier approach.[13] So the maximum principle stating that the Hamiltonian has to be maximized is equivalent to stating explicitly that the Lagrangian has to be minimized, and not just at a stationary point, and the reason why it is a maximum for the Hamiltonian and a minimum for the Lagrangian comes just from the sign change in the definition of the Hamiltonian (3.30).

3.6 Example of an ODE Control Problem

We illustrate the use of Pontryagin's maximum principle on the following example. Suppose we have a system with a state variable $y = y(t) \in \mathbb{R}$ and a control variable $u = u(t) \in \mathbb{R}$ governed by

$$\dot{y} = u, \qquad y(0) = 0,$$

subject to the box constraints $|u(t)| \leq 1$ for all t. We would like to find the control $u(t)$ such that $y(1) = \frac{1}{2}$ and which minimizes the cost

$$J(y, u) = \frac{1}{2} \int_0^1 y^2 \, dt.$$

Without the constraint on the control, the optimality system (3.27–3.29) leads to $\dot{y} = u$, $\dot{\psi} = y$, $0 = 1 \cdot \psi$ and thus $\psi = 0$, $y = 0$ and $u = 0$. Since we must however have $y(1) = \frac{1}{2}$, one can force the solution in the last moment with a very large control to this value, and make the integral $\int y^2 dt$ arbitrarily small. With the constraint on the control, the best one can do is use $u = 1$, and we need to use this control over the second half of the interval to get $\dot{y} = 1$, in order to reach $y(1) = \frac{1}{2}$, which is the optimal solution, see Fig. 29.

Lets now see how Pontryagin's maximum principle guides us to this solution: it says that if $u(t)$ is the optimal control, then for every $t \in (0, 1)$, we have

$$H(y(t), u(t), \psi(t)) = \max_{|\xi| \leq 1} H(y(t), \xi, \psi(t)),$$

where $y(t)$ and $\psi(t)$ are the state and adjoint state of the optimal trajectory at time t, and H is the Hamiltonian

$$H(y, u, \psi) = \psi u - \frac{1}{2} y^2.$$

[13] See also Footnote 12.

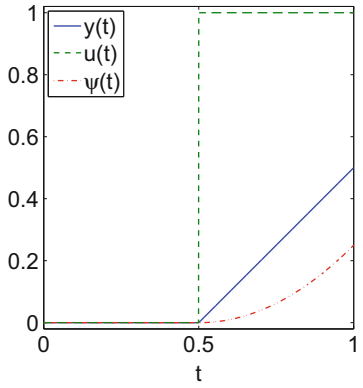


Fig. 29 Solution of the simple optimal control problem

Thus, by inspection, we have

$$u(t) = \begin{cases} 1, & \text{if } \psi(t) > 0, \\ -1 & \text{if } \psi(t) < 0. \end{cases}$$

If $\psi(t) = 0$, then we get no information from the maximum principle. We now deduce the optimal control and trajectory based on these properties.

1. We know that $y(1) = \frac{1}{2}$, so by the adjoint equation $\dot{\psi} = y$, we see that ψ has a positive slope in a neighborhood of $t = 1$, so it cannot vanish identically there. So if we assume that $\psi(1) \leq 0$, then $\psi(t) < 0$ in some interval $t \in (t_1, 1)$ with $t_1 = 1 - \delta, \delta > 0$, so $u(t) = -1$ there. This yields

$$y(t) = y(1) - \int_t^1 \dot{y}(\tau) d\tau = y(1) + 1 - t = \frac{3}{2} - t. \tag{3.59}$$

   Thus, $y(t) \geq \frac{1}{2}$ for all $t \in (t_1, 1)$, so $\psi(t)$ is a strictly increasing function with $\psi(1) \leq 0$, implying that $\psi(t) < 0$ for all $t \in (t_1, 1)$. In particular, $\psi(t_1) < 0$, so continuing this argument now over the interval $(t_1 - \delta, t_1)$, etc. shows that (3.59) in fact holds for the whole interval $(0, 1)$. This implies $y(0) = \frac{3}{2}$, which contradicts the initial condition $y(0) = 0$. Hence $\psi(1)$ cannot be negative (or zero).
2. Suppose now that $\psi(1) = \psi_1 > 0$. Then there exists a neighborhood around $t = 1$ in which $\psi(t) > 0$. Let $t^* \in [0, 1)$ be the smallest t such that $\psi(t) > 0$ whenever $t > t^*$. Then by the continuity of ψ, we have $\psi(t^*) = 0$. Moreover, $u = 1$ on $(t^*, 1)$, which implies

$$y(t) = y(1) - \int_t^1 u(\tau) d\tau = y(1) - 1 + t = t - \frac{1}{2} \tag{3.60}$$

   whenever $t \in (t^*, 1]$.

3. We show that $y(t^*) = 0$ by excluding both $y(t^*) > 0$ and $y(t^*) < 0$. If $y(t^*) > 0$, then $\psi(t^* - \delta) < 0$ for $\delta > 0$ small enough, so $u = -1$ on the interval $(t^* - \delta, t^*)$. This means $y(t^* - \delta) > y(t^*) > 0$; continuing this argument backwards in time, we obtain $y(0) > y(t^*) > 0$, a contradiction. On the other hand, if we assume that $y(t^*) < 0$, then $\dot{\psi}(t^*) < 0$ and $\psi(t^*) = 0$ together implies that $\psi(t^* + \delta) < 0$ for $\delta > 0$ small enough, which contradicts the definition of t^*. Thus, $y(t^*) = 0$. Since (3.60) is satisfied for all $t \in (t^*, 1]$, we deduce that $t^* = \frac{1}{2}$.
4. The optimal trajectory and control are now determined for the interval $[\frac{1}{2}, 1]$. Since $\int_{1/2}^{1} y^2\, dt$ is now fixed, we are left with the minimization problem

$$\int_0^{1/2} y^2\, dt \to \min \qquad \text{s.t. } y(0) = y(\tfrac{1}{2}) = 0,$$

where $\dot{y} = u$ and $|u(t)| \le 1$. The optimal solution is obviously

$$y(t) \equiv 0, \qquad u(t) \equiv 0 \qquad \forall t \in (0, \tfrac{1}{2}).$$

Note that the adjoint state must also vanish, since u would not be allowed to take on values different from ± 1 otherwise.

We thus obtain the same solution from Fig. 29. Note that unlike problems with a pure bang–bang solution, our optimal control contains both an interior part ($u = 0$ on $t \in (0, \frac{1}{2})$) and a boundary part ($u = 1$ on $t \in (\frac{1}{2}, 1)$). We also see that in this case, the maximum principle is useful in the sense that it guides us towards the optimal solution bit by bit, but it does not provide an algorithm for computing the optimal control directly.

3.7 *Caratheodory*

> Auf den folgenden Seiten soll auf das allgemeine Problem der Variationsrechnung in einem $(n + 1)$-dimensionalen Raum mit p gewöhnlichen Differentialgleichungen als Nebenbedingungen die Methode der geodätischen Äquidistanten angewandt werden[14] [16]

Constantin Carthéodory had already worked in his Ph.D. thesis on discontinuous solutions in the calculus of variations [15], and became one of the eminent researchers in this field. In a paper published in 1926, see also the quote above, he set out to solve precisely the same type of problem we have seen before, but 30 years earlier. He studied the minimization problem

$$I := \int_{t_1}^{t_2} L(t, \boldsymbol{x}, \dot{\boldsymbol{x}})dt \quad \longrightarrow \quad \min$$

[14] On the following pages we will solve the general problem of variational calculus in an $(n+1)$ dimensional space with p ordinary differential equations as constraints, using the method of geodesic equal distances.

under the constraints given by implicit differential equations

$$\boldsymbol{G}(t, \boldsymbol{x}, \dot{\boldsymbol{x}}) = 0, \tag{3.61}$$

where $L : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, and $\boldsymbol{G} : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^p$. Using geodesic arguments, he was led to define the scalar quantity

$$M(t, \boldsymbol{x}, \dot{\boldsymbol{x}}, \boldsymbol{\mu}) := L(t, \boldsymbol{x}, \dot{\boldsymbol{x}}) + \boldsymbol{\mu}^T \boldsymbol{G}(t, \boldsymbol{x}, \dot{\boldsymbol{x}}),$$

for some parameter functions $\boldsymbol{\mu}$. He then applied the Legendre transform to M, which led him to the Hamiltonian

$$H(t, \boldsymbol{x}, \boldsymbol{y}) := -M(t, \boldsymbol{x}, \boldsymbol{\varphi}, \boldsymbol{\chi}) + \boldsymbol{y}^T \boldsymbol{\varphi}.$$

Here, $\boldsymbol{\varphi}$ represents the right hand side when the implicit differential equation (3.61) is solved to obtain an explicit form $\dot{x}_i = \varphi_i(t, \boldsymbol{x})$, and $\boldsymbol{\chi} = \boldsymbol{\mu}$, which gives

$$H(t, \boldsymbol{x}, \boldsymbol{y}) = -L(t, \boldsymbol{x}, \boldsymbol{\varphi}) - \boldsymbol{\chi}^T \boldsymbol{G}(t, \boldsymbol{x}, \boldsymbol{\varphi}) + \boldsymbol{y}^T \boldsymbol{\varphi}.$$

Now along a solution satisfying the constraint, we have $\boldsymbol{G}(t, \boldsymbol{x}, \boldsymbol{\varphi}) = 0$, and Carathéodory obtains as the main result,[15] as we have seen earlier, that the solution candidates must satisfy the differential equations

$$\dot{\boldsymbol{x}} = \nabla_y H, \quad \dot{\boldsymbol{y}} = -\nabla_x H, \tag{3.62}$$

which he says play such a prominent role in mechanics, see also the original formulas in Fig. 30. In contrast to Pontryagin later, he does however only consider local optima in open sets. For more explanations on the derivation of the Hamiltonian formulation of Carathéodory, see [44], and also the very interesting description of the history of the maximum principle and optimal control in [46], see also [43, 45].

$$H(t, x_i, y_i) = -M(t, x_j, \varphi_j, \chi_{k'}) + \sum_j y_j \varphi_j,$$
$$\dot{x}_i = H_{y_i}, \quad \dot{y}_i = -H_{x_i}$$

Fig. 30 Formulation of necessary conditions using the Hamiltonian for optimal control problems already found in the work by Carathéodory from 1926

[15] Das Hauptresultat besteht darin, dass unsere Gefällkurven mit den Cauchyschen Charakteristiken zusammenfallen und Lösungen der kanonischen Differentialgleichungen (3.62) sind, die in der Mechanik eine so bedeutende Rolle spielen.

4 PDE Constrained Optimization

We have seen in the previous section how the desire to optimize the trajectory of a system governed by ODEs gave birth to the field of optimal control. In many applications, however, the system is not governed by ODEs, but by partial differential equations (PDEs), and the desire to optimize certain outputs leads to PDE constrained optimization problems. This field is nowadays an active research area, as attested by the many conferences and papers in recent years. Here we mention only three sample applications; other applications abound and new ones arise every day, so it is impossible to mention them all.

- Oil reservoir management: the flow of fluids in an oil field satisfies a system nonlinear PDEs that models the conservation of chemical species transported by different fluid phases. Here, the only interaction with the subsurface oil field is through wells, either by injecting fluid (water or gas) into the ground or by controlling how much fluid (typically a mixture of oil, water and gas) can come out of it. Thus, the goal could be, for instance, to optimize the oil output over the lifetime of the reservoir by optimizing over the control variables, such as the injection rate of water or gas at an injection well, or the fluid pressure or production rate at the production wells. Here the control variables can be functions of time, just like in the ODE case.
- Shape and topology optimization: consider the design of an airfoil. Depending on the purpose of the airfoil, one can maximize the lift, minimize the drag, or minimize the vortices created by the airfoil when air flows around it. Thus, the objective function depends on the solution of the PDE governing the flow of air around the airfoil, e.g., a Laplace-type potential flow equation, or the full Navier–Stokes equation. Here, the control variable is the "shape" of the airfoil, i.e., the function that defines the boundary of the domain, and the PDE constraint is the Laplace or Navier–Stokes equation.
- Inverse problems: consider an underground rock formation, of which we would like to understand its internal composition (types of rock, existence of layers and faults, etc.) One way of obtaining information without drilling is to send seismic or electromagnetic waves into the ground and install detectors on the surface to measure the reflected waves. If the rock parameters were known ahead of time, then the reflected waves can by calculated by solving a PDE (elasticity or wave equation). However, since our goal is precisely to estimate these parameters, we must solve an *optimization problem* by choosing the parameters that *minimize the discrepancy* between the predicted and measured waves, subject to the constraint that the waves satisfy a PDE.

4.1 Early Work

The discovery of Pontryagin's maximum principle and its ability to explain bang–bang type solutions generated great interest in the optimal control community. In particular, starting from the 1960s, there was a push to generalize both results to systems described by PDE rather than ODE constraints. The earliest reference appears to be a series of papers by Egorov [18, 19] starting in 1962, which contains a detailed study of the minimal time problem for the parabolic control problem of the type

$$\begin{aligned} \frac{\partial y}{\partial t} + Ay + b(u)y &= f + u \quad \text{on } \Omega \times (0,T), \\ y &= 0 \qquad \text{on } \partial\Omega \times (0,T), \end{aligned} \tag{4.1}$$

with initial condition $y(t_0;u) = y_0$ and target $y(t_1;u) = y_T$, but the arguments therein are rather opaque.[16]

Stateside, a proof of the bang–bang property when $b = 0$ and u is restricted to the set

$$U_{ad} = \{u : |u(t)| \leq 1 \text{ a.e.}\}$$

was given in 1964 by Fattorini [21], who wrote his Ph.D. thesis on the topic under the supervision of P. D. Lax. The proof proceeds in two steps. First, Fattorini writes $y(\tau;u)$ in terms of the Green's function

$$y(\tau;u) = G(\tau)y_0 + \int_0^\tau G(\tau-\sigma)u(\sigma)\,d\sigma.$$

Using this representation, he shows that if $|u(t)| \leq 1-\epsilon$ for some $\epsilon > 0$ *almost everywhere* in the interval $(0,\tau)$, then one can produce another control $v(t)$ such that $|v(t)| \leq 1$ and $y(s;v) = y_T$ with $s < \tau$, so that τ is not the optimal time. He then shows that even in the case where $|u(t)| \leq 1-\epsilon$ only on a *subset* $e \subset (0,\tau)$ of positive measure, u cannot optimal. To show this, let e be the subset in which $|u(t)| \leq 1-\epsilon$. Then using semi-group theory, Fattorini shows that there exists a control $\bar{g}(t)$ with bounded values and *support in* e such that $y(\tau;\bar{g}) = y(\tau;u) = y_T$. By taking a weighted average of u and $\bar{g}$, one obtains a new control $v = (1-\theta)u + \theta\bar{g}$ that satisfies $|v(t)| \leq 1-\hat{\epsilon}$ everywhere for some $\hat{\epsilon} > 0$, but without changing the target y_T, since $y(\tau,v) = (1-\theta)y(\tau;u) + \theta y(\tau;\bar{g}) = y_T$. Thus, by the previous argument, τ is not the shortest time necessary to arrive at y_T,

[16] According to J.-L. Lions: "Le travail de Yu. V. Egorov contient une étude détaillée de ce problème, mais nous n'avons pas pu comprendre tous les points des démonstrations de cet auteur, les résultats étant très probablement tous corrects."

so u is not time-optimal. This proof does not use any variant of the Pontryagin's maximum principle, so none was formulated in the paper.

Proofs of the bang–bang property for other systems, notably boundary control problems, appeared subsequently, see for instance Friedman [26]. However, it was a research monograph of Jacques-Louis Lions that launched the systematic study of optimal control under PDE constraints and shaped the field as we know it today.

4.2 Lions

> A new adventure began for Lions in the early 1960s, when he met (in spirit) another of his intellectual mentors, John von Neumann. By then, using computers built from his early designs, von Neumann was developing numerical methods for the solution of PDEs from fluid mechanics and meteorology. At a time when the French mathematical school was almost exclusively engaged in the development of the Bourbaki program, Lions — virtually alone in France — dreamed of an important future for mathematics in these new directions; he threw himself into this new work, while still continuing to produce high-level theoretical work on PDEs. (R. M. Temam, Obituary of Jacques-Louis Lions (SIAM News, July 10, 2001)

Jacques-Louis Lions (1928–2001) was one of the most influential figures of his time in applied mathematics in France and throughout the world. Under the influence of his Ph.D. supervisor, the Fields medalist L. Schwartz, Lions' early work was of a theoretical nature, emphasizing the use of distributions and appropriate function spaces in the study and solution of PDEs. During his time as scientific director at IRIA,[17] he discovered "systems theory", which subsequently became a new component of his research in the form of control theory. Given his expertise in PDEs and variational formulations, it is no surprise that his theory of PDE constrained optimization is heavily based on function (especially Sobolev) spaces and variational arguments.

Lions' first contribution in PDE constrained optimization was a research monograph entitled "Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles" [37]. It was published in 1968 and became the standard reference of the subject. In this volume, Lions developed his theory systematically by first considering the control of elliptic problems, and then moving on to time-dependent problems of the parabolic and hyperbolic types. The stated goals of the volume, which appear in the introduction, are as follows:

1. to obtain necessary (and maybe also sufficient) conditions for local extrema of the PDE constrained optimization problems;
2. to study the structure and properties of equations expressing such conditions;
3. to obtain constructive algorithms that can be used to calculate the optimal controls numerically.

[17] Institut de Recherche en Informatique et Automatique, the precursor of the modern INRIA.

This last point was particularly groundbreaking at a time when PDE research was mostly theoretical, see the quote above. It is especially fitting that variational formulations and Hilbert spaces play a fundamental role in the monograph, giving its results a natural algorithmic realization in the form of finite element methods, cf. [28].

To illustrate his approach, let us consider the problem of minimizing the cost functional

$$J(u) = \|Cy(u) - z_d\|_H^2 + (Nu, u)_U.$$

Here, the desired state z_d belongs to a Hilbert space H, where as the state variable $y = y(u)$ belongs to a possibly different Hilbert space V. The state variable $y(u)$ depends on the control variable u via the PDE

$$Ay = f + Bu, \tag{4.2}$$

where $A : V \to V'$ is generally taken to be a differential operator. The minimization is done over all controls u lying in the admissible set U_{ad}, a closed convex subset of a Hilbert space U. The quadratic form $(Nu, u)_U$, with N self-adjoint and semi-positive definite, penalizes large control variables u. From the definition of $J(u)$, we see that for all $v \in U_{ad}$, we have

$$\begin{aligned} J(v) &= (Cy(v) - z_d, Cy(v) - z_d)_H + (Nv, v)_U \\ &= \|Cy(u) - z_d\|_H^2 + 2(Cy(u) - z_d, C(y(v) - y(u)))_H + \|C(y(v) - y(u))\|_H^2 \\ &\quad + (Nu, u)_U + 2(Nu, v - u)_U + (N(v - u), v - u)_U \\ &= J(u) + 2(Cy(u) - z_d, C(y(v) - y(u)))_H + 2(Nu, v - u)_U \\ &\quad + \|C(y(v) - y(u))\|_H + (N(v - u), v - u)_U. \end{aligned}$$

Now since u is the minimizer, we must have $J(v) - J(u) \geq 0$, so that

$$\begin{aligned} 2(Cy(u) - z_d, C(y(v) - y(u)))_H + 2(Nu, v - u)_U \\ + \|C(y(v) - y(u))\|_H^2 + (N(v - u), v - u)_U \geq 0, \end{aligned}$$

which must hold for all $v \in U_{ad}$. So if $\|v - u\| = O(\epsilon)$ and we let ϵ tend to zero, the two quadratic terms become negligible, so we obtain after division by 2 the optimality condition

$$(Cy(u) - z_d, C(y(v) - y(u)))_H + (Nu, v - u)_U \geq 0 \qquad \forall v \in U_{ad}, \tag{4.3}$$

which is analogous to (3.16) in the KKT conditions. The inequality (4.3) can be rewritten as

$$(C^*\Lambda(Cy(u) - z_d), y(v) - y(u))_V + (Nu, v - u)_U \geq 0 \qquad \forall v \in U_{ad}, \tag{4.4}$$

where $\Lambda : H \to H'$ is the canonical isomorphism from H to its dual space H'. Lions then defines the *adjoint state* $p(v) \in V$ implicitly via

$$A^* p(v) = C^*\Lambda(Cy(v) - z_d), \tag{4.5}$$

where $A^* : V \to V'$ is the adjoint of A. Then substituting (4.5) into (4.4) yields

$$\begin{aligned}
&(C^*\Lambda(Cy(u) - z_d), y(v) - y(u))_V + (Nu, v - u)_U \\
&\qquad = (A^* p(u), y(v) - y(u))_V + (Nu, v - u)_U \\
&\qquad = (p(u), A(y(v) - y(u))_V + (Nu, v - u)_U \\
&\qquad = (p(u), B(v - u))_V + (Nu, v - u)_U \\
&\qquad = (\Lambda_U^{-1} B^* p(u) + Nu, v - u)_U \geq 0 \qquad \forall v \in U_{ad},
\end{aligned} \tag{4.6}$$

where $B^* : V \to U'$ is the adjoint of B, $\Lambda_U : U \to U'$ is the canonical isomorphism from U to U', and we have used the fact that

$$A(y(v) - y(u)) = f + Bv - (f + Bu) = B(v - u).$$

In other words, the definition of $p(v)$ in (4.5) can be seen as an intelligent guess that allows one to eliminate the state $y(u)$ from the optimality condition (4.4), similar to the way we chose the Lagrange multiplier $\boldsymbol{\lambda}$ in Sect. 3.1 to eliminate the state $\boldsymbol{y}$ in the finite-dimensional case. Inequality (4.6) can be reformulated as

$$(\Lambda_U^{-1} B^* p(u) + Nu, u)_U = \inf_{v \in U_{ad}} (\Lambda_U^{-1} B^* p(u) + Nu, v)_U,$$

which then looks like an elliptic analogue of Pontryagin's maximum principle.[18]

The advantage of the abstract Hilbert space approach is that the results are immediately applicable to many different types of control problems. For instance, consider a problem in which the control function is Neumann data on part of the boundary $\Gamma_0 \subset \Gamma = \partial\Omega$, and we want the Dirichlet trace on another part of the boundary $\Gamma_1 \subset \partial\Omega$, $\Gamma_0 \cap \Gamma_1 = \emptyset$ to be as close as possible to some desired trace z_d. Then the analogue of (4.6) in the boundary control case states that the optimal control $u \in U_{ad} \subset L^2(\Gamma)$ must satisfy

$$\int_\Gamma p(u)(v - u)\, d\Gamma \geq 0 \qquad \forall v \in U_{ad}. \tag{4.7}$$

[18] "La formulation (1.31) peut être considérée comme un analogue du «principe du maximum de Pontryagin», pour lequel nous référons [...] à PONTRYAGIN-BOLTYANSKI-GAMKRELIDZE-MISCHENKO" [37].

If the set of admissible controls is defined by pointwise box constraints, e.g., if

$$U_{ad} = \{v : \mathrm{Supp}(v) \subset \Gamma_0 \text{ and } |v(x)| \leq 1 \text{ a.e. on } \Gamma_0\},$$

then a standard argument allows one to convert the variational inequality (4.7) into a pointwise one of the form

$$p(x;u)(\xi - u(x)) \geq 0 \qquad \forall \xi \in [-1,1]. \tag{4.8}$$

Under some smoothness assumptions on the domain boundary Γ and the coefficients of the elliptic PDE, Lions shows that the optimal control $u \in U_{ad}$ satisfies either $p(x;u) \equiv 0$, in which case $y(u)|_{\Gamma_1} = z_d$, or $p(x;u) \neq 0$ almost everywhere. Then (4.8) implies

$$\begin{aligned} p(x;u) > 0 &\implies u(x) = -1, \\ p(x;u) < 0 &\implies u(x) = 1. \end{aligned}$$

Thus, we have a bang–bang property in the elliptic case, a result which, to Lions' knowledge, had not been published at the time.

4.3 Derivation by Lagrange Multipliers

It was never explicitly mentioned what motivated Lions to define the adjoint state p via (4.5). One possibility is that he was influenced by the work of Pontryagin; another reason could simply be that he wanted to eliminate the state variables $y(u)$ and $y(v)$ algebraically, just as we did in Sect. 3.1. Here, we show that the same variable p can be obtained using a formal Lagrange multiplier argument. Let the Lagrangian be defined by

$$\mathcal{L}(y,u,p) = \frac{1}{2}\|Cy - z_d\|_H^2 + \frac{1}{2}(Nu,u)_U - (Ay - f - Bu, p)_V,$$

where $p \in V$ now acts as the Lagrange multiplier. Next, we take the variational derivative with respect to y, i.e., we calculate

$$\frac{d}{d\epsilon}\,\mathcal{L}(y + \epsilon z, u, p)|_{\epsilon=0} = (Cz, Cy - z_d)_H - (Az, p)_V \stackrel{!}{=} 0$$

for all $z \in V$. We thus have

$$(Cz, Cy - z_d)_H - (Az, p)_V = (z, C^*\Lambda(Cy - z_d))_V - (z, A^*p)_V = 0,$$

which implies $A^*p = C^*\Lambda(Cy - z_d)$. So the adjoint state is nothing but the Lagrange multiplier for the constrained problem! We check that this formulation gives the same optimality condition for u: we want u to be a minimizer of $\mathcal{L}(y, u, p)$, i.e., for all $v \in U_{ad}$, we have

$$0 \leq \mathcal{L}(y, v, p) - \mathcal{L}(y, u, p) = (B(v-u), p)_V + (Nu, v-u)_U + \frac{1}{2}(N(v-u), v-u)_U$$
$$= (v-u, \Lambda_U^{-1} B^* p + Nu)_U + \frac{1}{2}(N(v-u), v-u)_U.$$

In particular, for $v = u + \epsilon w \in U_{ad}$, we have

$$\epsilon(w, Nu + \Lambda_U^{-1} B^* p)_U + \frac{\epsilon^2}{2}(Nw, w)_U \geq 0,$$

so by letting $\epsilon \to 0$, we obtain the same condition as (4.6). One can only speculate whether Lions had this derivation in mind.[19]

4.4 Later Developments

Lions' monograph only signaled the beginning of the rapid development of PDE constrained optimization as a modern field of research. Fueled by practical needs in industry and advances in other branches of applied mathematics, the field saw major progress in terms of both theory and algorithms—this is in addition to the number of application areas to which PDE constrained optimization is applied. The following list is by no means exhaustive; the goal is to show a sample of achievements in the intervening decades.

Theory. Much of the theory in Lions' monograph, including the existence and regularity of optimal controls and the maximum principle, has been extended to more general problems. For instance, Pontryagin's maximum principle for linear parabolic problems has been generalized to semilinear parabolic problems by von Wolfersdorf [52, 53]. It is also possible to include state constraints, i.e., constraints on the state variables y rather than on the control u. For a comprehensive modern introduction to the subject, see the recent book by Tröltzsch [50].

Another major theoretical development, related to the existence of optimal controls, is the theory of controllability, where the goal is to determine whether it is possible to find a control function that steers an object from any initial state y_0 to a given target state y_T. An important result, which appeared in [38] in 1988, was

[19] According to J. Blum, it was R. Glowinski, one of the former students of Lions, who once showed Lions on the board that the adjoint state can simply be interpreted as a Lagrange multiplier. This was confirmed by R. Glowinski (personal communication).

proved by Lions himself: he introduced what is known as the Hilbert Uniqueness Method. The method takes a linear time-reversible PDE (such as the wave equation), an initial state y_0 and a target state y_T, and constructs a control u (belonging to some specially chosen Hilbert space H) that steers y_0 to y_T, provided that the system is observable and the time horizon is long enough. For a more recent survey, see the articles by Zuazua [54, 55].

Algorithms. There has also been significant development on the algorithmic front: here, the goal is to discretize the infinite-dimensional PDE constrained problem, e.g. using finite element methods, in order to obtain a finite dimensional approximation, which can then be solved numerically. In principle, one can discretize the KKT formulation (3.16)–(3.18) and then use standard optimization routines, such as line search, trust region and interior point methods to solve the finite dimensional problem; however, one must be careful to discretize the forward and adjoint problems consistently to retain optimality in the discrete setting, see [12]. Using such routines allows one to take advantage of advances in sparse matrix factorizations and preconditioners that have been developed for general saddle-point problems, see for instance [3].

Shooting methods, or more precisely multiple shooting methods, were originally developed for solving two-point boundary value problems [32, 41, 42]. While the finite element method has become the method of choice for most boundary value problems (especially of the elliptic type), multiple shooting remained a viable approach for optimal control problems, since they are able to integrate systems that are highly unstable and very sensitive to changes in initial/final conditions, see the Ph.D. thesis by Bock [6]. More recently, multiple shooting has been applied successfully to problems with PDE constraints, see for example [29, 30, 49], and the recent work by Rannacher et al. [17].

With the rapid increase in computing power in the form of multi-core processors and parallel clusters, there is increasing interest in parallel algorithms for solving PDE constrained optimization and optimal control problems. Methods such as domain decomposition and multigrid, which have been developed and analyzed extensively for discretized PDE problems, are particularly suited for this purpose. For the use of domain decomposition in parabolic optimal control problems, see Heinkenschloss [29] and references therein.

Acknowledgements The authors are grateful to M. Mattmüller for providing us with a copy of Bernoulli's letter (Univ. Bibl. Basel, Handschriften-Signatur L I a 669, Nr. 50). We further thank Ph. Henry, C. Lubich and E. Hairer for helpful discussions which greatly improved the manuscript. We are also grateful to Armen Sergeev from the Steklov Institute in Moscow for his invaluable help to get the original sources of A.A. Feldbaum, and Peter Kloeden for obtaining the RAND report of Hestenes for us. We thank the Bibliothèque de Genève for granting permission to reproduce photographs from the original sources under catalogue numbers Kc62 (Varignon), Kc110 [33], Kc111 [34], Ka495 [4], Ka368 (Euler's *Methodus* E65), Ka459 (Archimedes) and also for Figs. 26, 27, and 28 from [48]. We also thank Tatiana Smirnova-Nagnibeda, Rinat Kashaev, Zdeněk Strakoš and Ivana Gander for their valuable help in translating several texts that originally appeared in Russian. The authors acknowledge support by the European ScienceFoundation, the Swiss National Science Foundation and the Centro Stefano Franscini.

References

1. Archimedes, *On the Equilibrium of Planes*. publ. Basel (Latin-Greek, 1544), Paris (Latin-Greek, 1615, p.145), Heath (Engl., 1897, p.189), Ver Eecke (French, 1921, I, p.237–299), 250 B.C., in *Opera of Archimedes*
2. R. Bellman, I. Glicksberg, O. Gross, On the 'bang-bang' control problem. Technical Report, DTIC Document, 1955
3. M. Benzi, G.H. Golub, J. Liesen, Numerical solution of saddle point problems. Acta Numerica, **14**(1), pp. 1–137 (2005)
4. J. Bernoulli, *Opera Omnia*, 4 vols. (Bousquet & Socios., Lausannae/Genevae, 1742)
5. G.A. Bliss, *Lectures on the Calculus of Variations*, vol. 850 (University of Chicago Press, Chicago, 1946)
6. H.G. Bock, Randwertproblemmethoden zur Parameteridentifizierung in Systemen nichtlinearer Differentialgleichungen. Ph.D. thesis, Rheinische Friedrich-Wilhelms-Universität, Bonn, 1987 (No. 183)
7. V.G. Boltyanski, The maximum principle in the theory of optimal processes (Russian). Doklady AN SSSR **119**(6), 1070–1073 (1958)
8. V.G. Boltyanski, *The Maximum Principle—How it Came to Be?* (Inst. für Mathematik, Technische Univ. München, München, 1994)
9. V.G. Boltyanski, R.V. Gamkrelidze, L.S. Pontryagin, On the theory of optimal processes (Russian). Doklady AN SSSR **110**, 7–10 (1956)
10. V.G. Boltyanski, R.V. Gamkrelidze, L.S. Pontryagin, The theory of optimal processes. I. The maximum principle. Izv. Akad. Nauk SSSR. Ser. Mat. **24**, 3–42 (1960)
11. V. Boltyanski, H. Martini, V. Soltan, *Geometric Methods and Optimization Problems*, vol. 4 (Springer, New York, 1999)
12. A. Borzì, V. Schulz, *Computational Optimization of Systems Governed by Partial Differential Equations*. Computational Science & Engineering (SIAM, Philadelphia, 2012)
13. D.W. Bushaw, Differential equations with a discontinuous forcing term. Ph.D. thesis, Department of Mathematics, Princeton University, 1952
14. D.W. Bushaw, Experimental towing tank. Technical Report, Stevens Institute of Technology, Reprint 169, Hoboken, 1953
15. C. Carathéodory, Über die diskontinuierlichen Lösungen in der Variationsrechnung. Ph.D. thesis, Universität Göttingen, 1904 (Gesammelte Mathematische Schriften, Band I), pp. 1–71
16. C. Carathéodory, Die Methode der geodätischen Äquidistanten und das Problem von Lagrange. Acta Mathematica **47**(3), 199–236 (1926)
17. T. Carraro, M. Geiger, R. Rannacher, Indirect multiple shooting for nonlinear parabolic optimal control problems with control constraints. SIAM J. Sci. Comput. **36**(2), A452–A481 (2014)
18. Y.V. Egorov, Some problems in the theory of optimal control. Dokl. Akad. Nauk. SSSR **145**, 720–723 (1962)
19. Y.V. Egorov, Sufficient conditions for optimal control in Banach spaces. Mat. Sbornik **64**, 79–101 (1964)
20. L. Euler, *Methodus inveniendi lineas curvas maximi minimive proprietate gaudentes sive solutio problematis isoperimetrici latissimo sensu accepti* (Bousquet & Socios., Lausannae/Genevae, 1744) [Enestr. 65, Opera Omnia, Ser.I, vol. 24]
21. H.O. Fattorini, Time-optimal control of solutions of operational differential equations. J. SIAM Control Ser. A **2**(1), 54–59 (1964)
22. A.A. Feldbaum, The simplest relay system of automatic control (Russian). Avtomatika i Telemehanika **10**(4), 249–266 (1949)
23. A.A. Feldbaum, Optimal processes in systems of automatic control (Russian). Avtomatika i Telemehanika **14**(6), 712–728 (1953)
24. A.A. Feldbaum, On synthesis of optimal systems of automatic control (Russian), in *Transactions of the 2nd National Conference on the Theory of Automatic Control, Izdat. AN SSSR*, vol. 2 (1955), pp. 325–360

25. A.A. Feldbaum, On synthesis of optimal systems with the aid of phase space (Russian). Avtomatika i Telemehanika **16**(2), 129–149 (1955)
26. A. Friedman, Optimal control for parabolic equations. J. Math. Anal. Appl. **18**, 479–491 (1967)
27. R.V. Gamkrelidze, Discovery of the maximum principle. J. Dyn. Control Syst. **5**(4), 437–451 (1999)
28. M.J. Gander, G. Wanner, From Euler, Ritz and Galerkin to modern computing. SIAM Rev. **54**, 627–666 (2013)
29. M. Heinkenschloss, A time-domain decomposition iterative method for the solution of distributed linear quadratic optimal control problems. J. Comput. Appl. Math. **173**, 169–198 (2005)
30. H.K. Hesse, G. Kanschat, Mesh adaptive multiple shooting for partial differential equations. Part I: linear quadratic optimal control problems. J. Numer. Math. **17**(3), 195–217 (2009)
31. M.R. Hestenes, A general problem in the calculus of variations with applications to paths of least time. Technical Report, RAND Memorandum RM-100, 1950. ASTIA Document Number AD 112382
32. H.B. Keller, *Numerical Methods for Two-Point Boundary Value Problems* (Waltham, Blaisdell, 1968)
33. J.L. Lagrange, *Méchanique analitique* (Chez la Veuve Desaint, A Paris, 1788)
34. J.L. Lagrange, *Mécanique analytique*. (Mme Ve Courcier, Paris, 1811/1815) [Second enlarged edition in two volumes; third edition 1853 publ. by J. Bertrand; fourth edition in Oeuvres de Lagrange, vol. 11,12, 1888]
35. A. Lerner, Improving of dynamic properties of automatic compensators with the aid of nonlinear connections I (Russian). Avtomatika i Telemehanika **13**(2), 134–144 (1952)
36. A. Lerner, Constructing of time-optimal systems of automatic control with constrained values of coordinates of controled object (Russian), in *Transactions of the 2nd National Conference on the Theory of Automatic Control, Izdat. AN SSSR*, vol. 2 (1955), pp. 305–324
37. J.L. Lions, *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles* (Dunod, Paris, 1968)
38. J.L. Lions, Exact controllability, stabilization and perturbations for distributed systems. SIAM Rev. **30**, 1–68 (1988)
39. S. Mac Lane, P.L. Duren, R.A. Askey, U.C. Merzbach, *Mathematics at the University of Chicago: A Brief History* (American Mathematical Society, Providence, 1989)
40. E.J. McShane, On multipliers for Lagrange problems. Am. J. Math. **61**(4), 809–819 (1939)
41. D.D. Morrison, J.D. Riley, J.F. Zancanaro, Multiple shooting method for two-point boundary value problems. Commun. ACM **5**(12), 613–614 (1962)
42. M.R. Osborne, On shooting methods for boundary value problems. J. Math. Anal. Appl. **27**(2), 417–433 (1969)
43. H.J. Pesch, Carathéodory's royal road of the calculus of variations: Missed exits to the maximum principle of optimal control theory. Numer. Algebra Control Optim. **3**(1), 161–173 (2013)
44. H.J. Pesch, R. Bulirsch, The maximum principle, Bellman's equation, and Carathéodory's work. J. Optim. Theory Appl. **80**(2), 199–225 (1994)
45. H.J. Pesch, M. Plail, The maximum principle of optimal control: a history of ingenious ideas and missed opportunities. Control Cybern. **38**(4A), 973–995 (2009)
46. M. Plail, *Die Entwicklung der optimalen Steuerungen: von den Anfängen bis zur eigenständigen Disziplin in der Mathematik* (Vandenhoeck und Ruprecht, Göttingen, 1998)
47. L.S. Pontryagin, Optimal regulation processes. Uspekhi Matematicheskikh Nauk **14**(1), 3–20 (1959)
48. L.S. Pontryagin, V.G. Boltyanski, R.V. Gamkrelidze, E.F. Mishchenko, *The Mathematical Theory of Optimal Processes* (Interscience Publishers/Wiley, New York, 1962)
49. R. Serban, S. Li, L.R. Petzold, Adaptive algorithms for optimal control of time-dependent partial differential-algebraic equation systems. Int. J. Numer. Methods Eng. **57**(10), 1457–1469 (2003)

50. F. Tröltzsch, *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*. Graduate studies in mathematics, vol. 112 (American Mathematical Society, Providence, 2010)
51. P. Varignon, *Nouvelle mechanique ou statique*, 2 vols. (Chez Claude Jombert, A Paris, 1725)
52. L. von Wolfersdorf, Optimal control for processes governed by mildly nonlinear differential equations of parabolic type I. ZAMM **56**, 531–538 (1976)
53. L. von Wolfersdorf, Optimal control for processes governed by mildly nonlinear differential equations of parabolic type II. ZAMM **57**, 11–17 (1977)
54. E. Zuazua, Some problems and results on the controllability of partial differential equations, in *Proceedings of the Second European Conference of Mathematics, Budapest, July 1996*. Progress in mathematics (Birkhäuser Verlag, Basel, 1998), pp. 276–311
55. E. Zuazua, Controllability of partial differential equations and its semi-discrete approximations. Discrete Continuous Dyn. Syst. **8**, 469–513 (2002)

Topology Design of Elastic Structures for a Contact Model

S.M. Giusti, Jan Sokołowski, and Jan Stebel

Abstract Contact problems are very important in the engineering design and the correct interpretation of the physical phenomena, and its influence in this process, is of paramount importance for the engineers. In this paper we employ the topological derivative concept for optimum design problems in contact solid mechanics. A nonlinear contact model governed by a variational inequality is considered. Beside the theoretical developments, some computational examples are included. The influence of the parameters of the contact model in the optimal results for the structures is studied. The numerical results show that the proposed method of optimum design can be applied to a broad class of engineering problems.

Keywords Asymptotic analysis • Frictionless contact problem • Optimum design problems • Topological derivative • Topological optimization

Mathematics Subject Classification (2010). Primary 49J40; Secondary 35J86, 74P15.

S.M. Giusti
Departamento de Ingeniería Civil, Facultad Regional Córdoba, Universidad Tecnológica Nacional (UTN/FRC - CONICET), Maestro M. López esq. Cruz Roja Argentina, X5016ZAA - Córdoba, Argentina
e-mail: sgiusti@civil.frc.utn.edu.ar

J. Sokołowski (✉)
Institut Élie Cartan, UMR7502 (Université Lorraine, CNRS, INRIA), Laboratoire de Mathématiques, Université de Lorraine, B.P.239, 54506 Vandoeuvre-lès-Nancy Cedex, France
e-mail: Jan.Sokolowski@univ-lorraine.fr

J. Stebel
Institute of Mathematics of the Academy of Sciences of the Czech Republic, Žitná 25, 115 67 Praha 1, Czech Republic
e-mail: stebel@math.cas.cz

R. Hoppe (ed.), *Optimization with PDE Constraints*, Lecture Notes in Computational Science and Engineering 101,
DOI 10.1007/978-3-319-08025-3_6

1 Introduction

An asymptotic expansion of a given shape functional, when a geometrical domain is singularly perturbed by the insertion of holes, can be obtained by performing a topological asymptotic analysis. This analysis is applied in the mathematical model that represents the physical phenomena under consideration. Asymptotic analysis of linear and nonlinear models in solid mechanics is considered in details in the recent monograph [23]. The related results can be also found in [5, 6, 10, 12, 13, 17, 22, 26, 27]. The main result of this analysis is the so-called topological derivative. This derivative measures the sensitivity of the shape functional when a singularity is introduced in an arbitrary point of the domain.

Classical shape optimization for contact problems is considered in [28] for the variational inequalities of the first and the second kind. The shape and material derivatives are determined in the framework of the conical differentiability of solutions to variational inequalities. Another branch of applied models with unilateral constraints are the crack models with nonlinear non-penetration conditions on the crack faces (lips) [19–21]. For such models the elastic energy is differentiated with respect to the crack length [9]. The stability of solutions to the evolution variational inequalities is analyzed in [18]. A new class of variational inequalities arises when a finite interpenetration is allowed in the potential contact region of the body with a rigid foundation, as proposed in [7].

In this work we present a closed form for the topological derivative when a small circular disc, with a material different than the surrounding medium, is introduced in an arbitrary point of the elastic body. We consider the energy shape functional associated to the frictionless contact problem allowing a finite interpenetration between an elastic body and a rigid foundation [7].

In order to apply the theoretical results, we present a computational procedure for topological optimization based on the topological derivative concept. The optimization procedure consists in minimizing the structural compliance for a given amount of material. This constraint in the volume of the optimized structure is introduced in the formulation of the optimization problem by means of an exact quadratic scheme. In this procedure, the topological derivative is used as a feasible descent direction. The robustness of the topological optimization technique presented in this work is demonstrated by a set of numerical examples, related to the topology design of elastic structures under this particular nonlinear contact condition. On the other hand, the formulation of the problem of topology optimization of structures in unilateral contact, with computational approaches such as SIMP (Solid Isotropic Microstructure with Penalization) and ESO (Evolutionary Structural Optimization), can be found in [8, 15, 24, 29].

This paper is organized as follows. Section 2 describes the frictionless contact model for finite interpenetration in two-dimensional elasticity. The topological derivative associated to this problem is presented in Sect. 3, where a simple and analytical formula is given. The compliance topology optimization procedure for

elastic structures subjected to a volume constraint is outlined in Sect. 4. A set of numerical experiments is presented in Sect. 5. The paper ends in Sect. 6 with some concluding remarks.

2 Static Contact Model for Finite Interpenetration

We consider the problem of an elastic body having contact with a rigid foundation. The domain of the body is denoted by $\Omega \subset \mathbb{R}^2$. The boundary $\partial\Omega$ of the body consists of three mutually disjoint parts with positive measures Γ_D, Γ_N and Γ_C, where different boundary conditions are prescribed. On the boundary Γ_D we prescribe Dirichlet boundary conditions (displacement), on Γ_N Neumann boundary conditions (traction) and, finally, on Γ_C the contact condition with the rigid foundation that admits an interpenetration, see Fig. 1a. For the contact model, we consider only a normal compliance law of the type

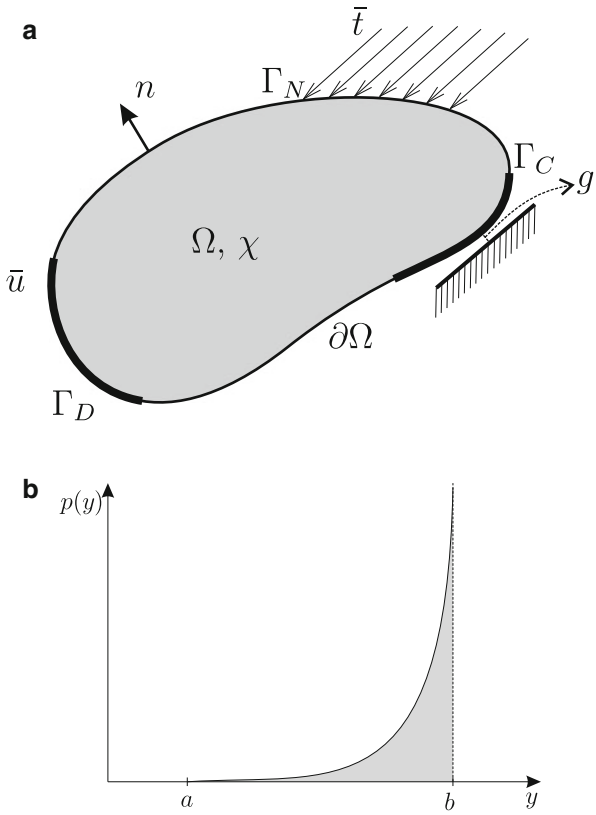


Fig. 1 Contact problem formulation. (**a**) Contact problem. (**b**) Example of function $p(y)$

$$\sigma_n(u) = -p(u_n - g)\,, \tag{2.1}$$

where $u_n := u \cdot n$ denotes the normal component of the displacement field u, n is the unit outward normal vector to the boundary $\partial\Omega$ and g the gap on the potential contact zone. Moreover, in (2.1), $\sigma_n(u)$ represents the normal component to the boundary of the stress tensor $\sigma(u)$, i.e. $\sigma_n(u) = \sigma(u)n \cdot n$. The Cauchy stress tensor $\sigma(u)$ is defined as:

$$\sigma(u) := \mathbb{C}\varepsilon(u)\,, \tag{2.2}$$

where $\varepsilon(u)$ is the symmetric part of the gradient of the displacement field u, i.e.

$$\varepsilon(u) := \frac{1}{2}(\nabla u + (\nabla u)^\top)\,, \tag{2.3}$$

and $\mathbb{C}$ denotes the fourth-order elastic tensor. For an isotropic elastic body, this tensor is given by:

$$\mathbb{C} = 2\mu\mathbb{I} + \lambda(\mathrm{I} \otimes \mathrm{I}), \tag{2.4}$$

with μ and λ denoting the Lamé coefficients. In the above expression, we use $\mathbb{I}$ and I to denote, respectively, the identities of fourth and second order. In terms of the engineering constant E (Young's modulus) and υ (Poisson's ratio) the above constitutive response can be written as:

$$\mathbb{C} = \frac{E}{1-\upsilon^2}[(1-\upsilon)\mathbb{I} + \upsilon(\mathrm{I} \otimes \mathrm{I})]\,. \tag{2.5}$$

The function $p : \mathbb{R} \to \overline{\mathbb{R}_+} = [0, +\infty]$ in (2.1) is used to model the interpenetration condition between the body and the foundation. This function p is monotone with the following properties:

$$\begin{cases} p(y) = 0 & \text{for } y \le a, \text{ with } a \text{ constant} \\ \lim\limits_{y \to b-} p(y) = +\infty & \text{for } y > a, \text{ with } b \text{ constant and } b > a \\ p(y) = +\infty & \text{for } y \ge b \end{cases}. \tag{2.6}$$

The parameter a indicates the initial contact and the value of b describes a limit such that no further interpenetration is possible, see Fig. 1b.

The strong form of the equilibrium equation under this contact condition is given by: find the displacement field $u : \Omega \to \mathbb{R}^2$ such that

$$\begin{cases} -\operatorname{div}\sigma(u) = 0 & \text{in } \Omega \\ u = \overline{u} & \text{on } \Gamma_D \\ \sigma(u)n = \overline{t} & \text{on } \Gamma_N \\ \sigma_n(u) = -p(u_n - g) & \text{on } \Gamma_C \\ \sigma_\tau(u) = 0 & \text{on } \Gamma_C \end{cases}. \tag{2.7}$$

The last condition in (2.7) indicates that the contact is without friction, where $\sigma_\tau(u) = \sigma(u)n - \sigma_n(u)n$ denotes the tangential component of the stress tensor $\sigma(u)$.

The weak formulation of the problem stated in (2.7) is given by the following variational equation: find $u \in \mathcal{U}$ with $(u_n - g) \in \mathrm{dom}(p)$, such that:

$$\int_\Omega \sigma(u) \cdot (\varepsilon(v) - \varepsilon(u)) + \int_{\Gamma_C} p(u_n - g)(v_n - u_n) = \int_{\Gamma_N} \bar{t} \cdot (v - u) \quad \forall v \in \mathcal{U}\,, \tag{2.8}$$

where the set of admissible functions $\mathcal{U}$ is given by:

$$\mathcal{U} := \{\varphi \in H^1(\Omega; \mathbb{R}^2) \,:\, \varphi = \bar{u} \quad \text{on} \quad \Gamma_D\}, \tag{2.9}$$

and the domain of definition of the function p, namely $\mathrm{dom}(p)$, is:

$$\mathrm{dom}(p) := \tag{2.10}$$
$$\left\{\varphi \in L^1(\Gamma_C) \,:\, p(\varphi) \in L^1(\Gamma_C), \exists C > 0 \,:\, \int_{\Gamma_C} p(\varphi)v \le C\,\|v\|_{H^{1/2}(\Gamma_C)}\right\}.$$

For a detailed description of this model, we refer the reader to [7].

3 Topological Derivative

In this section we obtain an asymptotic expansion for the energy shape functional when a small disc of radius ρ, with different constitutive property, is introduced in an arbitrary point $\hat{x}$ of the domain Ω, far enough from the potential contact region Γ_C, and denoted by $\mathcal{B}_\rho := \{x \in \mathbb{R}^2 \,:\, |x - \hat{x}| < \rho\}$, see Fig. 2. Thus, introducing a characteristic function $\chi = \mathbb{1}_\Omega$, associated to the unperturbed domain, it is possible to define the characteristic function associated to the topological perturbed domain χ_ρ. Particularly, when the topological perturbation is an inclusion, we have $\chi_\rho(\hat{x}) = \mathbb{1}_\Omega - (1-\gamma)\mathbb{1}_{\overline{\mathcal{B}_\rho(\hat{x})}}$, where $\gamma \in \mathbb{R}^+$ is the contrast parameter in the material property of the medium. Then we assume that a given shape functional $\psi(\chi_\rho(\hat{x}))$, associated to the topological perturbed domain Ω_ρ, admits the following topological asymptotic expansion

$$\psi(\chi_\rho(\hat{x})) = \psi(\chi) + f(\rho)\mathcal{T}\psi(\hat{x}) + o(f(\rho))\,, \tag{3.1}$$

where $\psi(\chi)$ it is the shape functional associated to the unperturbed domain, $f(\rho)$ it is a function such that $f(\rho) \to 0$, with $\rho \to 0^+$. The function $\hat{x} \mapsto \mathcal{T}\psi(\hat{x})$ is the so-called topological derivative of ψ in the point $\hat{x}$. Thus, the topological derivative

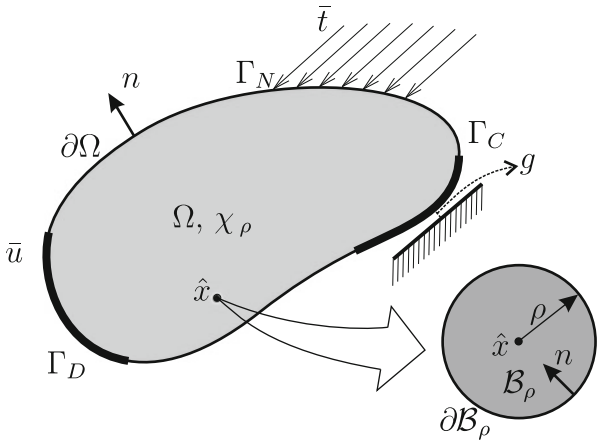


Fig. 2 Perturbed contact problem

can be seen as a first order correction factor over $\psi(\chi)$ to approximate $\psi(\chi_\rho(\hat{x}))$. In fact, after rearranging (3.1), we have

$$\frac{\psi(\chi_\rho(\hat{x})) - \psi(\chi)}{f(\rho)} = \mathcal{T}\psi(\hat{x}) + \frac{o(f(\rho))}{f(\rho)} \,. \tag{3.2}$$

Taking the limit $\rho \to 0^+$ in the above expression, we have the classical definition of the topological derivative [25] given by

$$\mathcal{T}\psi(\hat{x}) = \lim_{\rho \to 0^+} \frac{\psi(\chi_\rho(\hat{x})) - \psi(\chi)}{f(\rho)} \,. \tag{3.3}$$

Note that, the shape functionals $\psi(\chi_\rho(\hat{x}))$ and $\psi(\chi)$ are associated to domains with different topologies. Then, to calculate the limit $\rho \to 0^+$ in (3.3) it is necessary to perform a asymptotic expansion of the functional $\psi(\chi_\rho(\hat{x}))$ with respect to the parameter ρ.

In this work we are interested in the asymptotic expansion for the energy shape functional associated to the contact problem (2.8), given by [7]:

$$\mathcal{J}_\chi(u) := \frac{1}{2}\int_\Omega \sigma(u)\cdot\varepsilon(u) - \int_{\Gamma_N} \bar{t}\cdot u + \int_{\Gamma_C} P(u_n - g), \tag{3.4}$$

where the function $P(y)$ is given by:

$$P(y) := \int_{-\infty}^{y} p(z). \tag{3.5}$$

Considering the singular perturbation described above and denoted by $\mathcal{B}_\rho$, the energy shape functional associated to the perturbed domain is given by:

$$\mathcal{J}_{\chi_\rho}(u_\rho) := \frac{1}{2}\int_\Omega \sigma_\rho(u_\rho)\cdot\varepsilon(u_\rho) - \int_{\Gamma_N} \bar{t}\cdot u_\rho + \int_{\Gamma_C} P(u_{\rho n}-g), \tag{3.6}$$

where u_ρ is the solution of the problem in the singularly perturbed domain given by: find the displacement field $u_\rho : \Omega \to \mathbb{R}^2$ such that

$$\left\{\begin{array}{rcll} -\operatorname{div}\sigma_\rho(u_\rho) &=& 0 & \text{in } \Omega \\ u_\rho &=& \bar{u} & \text{on } \Gamma_D \\ \sigma(u_\rho)n &=& \bar{t} & \text{on } \Gamma_N \\ \sigma_n(u_\rho) &=& -p(u_{\rho n}-g) & \text{on } \Gamma_C \\ \sigma_\tau(u_\rho) &=& 0 & \text{on } \Gamma_C \\ [\![u_\rho]\!] &=& 0 & \text{on } \partial\mathcal{B}_\rho \\ [\![\sigma_\rho(u_\rho)]\!]n &=& 0 & \text{on } \partial\mathcal{B}_\rho \end{array}\right. , \tag{3.7}$$

since $u_{\rho n} := u_\rho \cdot n$ is used to denote the normal component of the displacement field u_ρ on the boundary Γ_C. The symbol $[\![(\cdot)]\!]$ in (3.7) denotes the jump of function $(\cdot)$ across the boundary $\partial\mathcal{B}_\rho$ and the stress operator $\sigma_\rho(\cdot)$ is defined as:

$$\sigma_\rho(\phi) := \gamma_\rho \mathbb{C}\varepsilon(\phi), \tag{3.8}$$

where the parameter γ_ρ is defined as:

$$\gamma_\rho := \left\{\begin{array}{lll} 1 & \text{in} & \Omega \setminus \overline{\mathcal{B}_\rho} \\ \gamma & \text{in} & \mathcal{B}_\rho \end{array}\right. . \tag{3.9}$$

Note that the domain Ω is topologically perturbed by the introduction of an inclusion $\mathcal{B}_\rho(\hat{x})$ of the same nature as the bulk material, but with contrast γ. Finally, the variational problem associated to (3.7) can be written as: find $u_\rho \in \mathcal{U}_\rho$ with $(u_{\rho n} - g) \in \operatorname{dom}(p)$, such that:

$$\int_\Omega \sigma_\rho(u_\rho)\cdot(\varepsilon(v)-\varepsilon(u_\rho)) + \int_{\Gamma_C} p(u_{\rho n}-g)(v_n - u_{\rho n}) = \int_{\Gamma_N} \bar{t}\cdot(v-u_\rho) \quad \forall v \in \mathcal{U}_\rho\,, \tag{3.10}$$

where the set of admissible functions $\mathcal{U}_\rho$ is given by:

$$\mathcal{U}_\rho := \{\varphi \in \mathcal{U} \,:\, [\![\varphi]\!] = 0 \quad \text{on} \quad \partial\mathcal{B}_\rho\}. \tag{3.11}$$

For an explicit and analytical formula for the topological derivative $\mathcal{T}_\mathcal{J}(\hat{x})$ of the functional (3.4) associated to the problem (2.7), we introduce the following result:

Theorem 1. *The energy shape functional of an elastic solid with a disc of radius ρ, centered at point $\hat{x} \in \Omega$ and with constitutive property characterized by the parameter γ, admits for $\rho \to 0^+$ the following asymptotic expansion:*

$$\mathcal{J}_{\chi_\rho}(u_\rho) = \mathcal{J}_\chi(u) + \rho^2 \pi \mathbb{H}_\gamma \sigma(u(\hat{x})) \cdot \varepsilon(u(\hat{x})) + o(\rho^2) \quad \forall \hat{x} \in \Omega, \tag{3.12}$$

where $u(\hat{x})$ is the solution of the problem (2.7) evaluated at $\hat{x}$ and $\mathbb{H}_\gamma$ is the fourth-order tensor defined as:

$$\mathbb{H}_\gamma := \frac{1}{4} \frac{(1-\gamma)^2}{1+\beta\gamma} \left(2 \frac{1+\beta}{1-\gamma} \mathbb{I} + \frac{\alpha-\beta}{1+\alpha\gamma} \mathrm{I} \otimes \mathrm{I} \right), \tag{3.13}$$

where I and $\mathbb{I}$ are the identities tensors of second- and fourth-order, respectively, and the parameters α and β depend exclusively on the Poisson's ratio of the elastic medium, given by

$$\alpha := \frac{1+\upsilon}{1-\upsilon} \qquad \text{and} \qquad \beta := \frac{3-\upsilon}{1+\upsilon}. \tag{3.14}$$

Proof. The reader interested in the proof of this result may refer to [14, 16, 23].

Corollary 2. *From the asymptotic expansion presented in Theorem 1, we can recognize the topological derivative of the functional $\mathcal{J}_\chi(u)$ given by:*

$$\mathcal{T}_\mathcal{J}(\hat{x}) := \mathbb{H}_\gamma \sigma(u(\hat{x})) \cdot \varepsilon(u(\hat{x})). \tag{3.15}$$

4 Topological Optimization Procedure

In order to illustrate the applicability of the topological asymptotic expansion (3.15), here we present an optimization procedure for elastic structures under the contact condition described in Sect. 2. The optimization procedure is based on the domain representation in a bi-material fashion, whose constituents properties are characterized by the Young modulus E and the phase contrast γ^*. Thus, as in (3.8) and (3.9), we have

$$E(x) = \begin{cases} E & \forall x \in \Omega^h, \\ \gamma^* E & \forall x \in \Omega^w, \end{cases} \tag{4.1}$$

where Ω^h and Ω^w denote the domains occupied by the two materials, the *hard* and *weak* materials, respectively.

The optimization problem consists in minimizing the structural compliance for a given amount of material. It can be written as

$$\begin{cases} \text{Minimize} & \psi(\chi) & = & -\mathcal{J}_\chi(u), \\ \text{Subjected to} & |\Omega^h| & \leq & V, \end{cases} \tag{4.2}$$

where $|\Omega^h|$ is the Lebesgue measure of the domain Ω^h and V is the required volume at the end of the optimization process. In order to solve the above problem, we use an exact quadratic penalization scheme. Thus, problem (4.2) is re-written as following

$$\underset{\Omega \subset \mathbb{R}^2}{\text{Minimize}}\ \mathcal{F}_\Omega(u) = -\mathcal{J}_\chi(u) + \lambda s_\Omega^2\,, \tag{4.3}$$

where λ is a positive parameter and the function s_Ω is defined as

$$s_\Omega := 1 - \frac{|\Omega^h|}{V}. \tag{4.4}$$

By considering the linearity property of the topological derivative operator, the topological derivative of the functional $\mathcal{F}_\Omega$ can be written as

$$\mathcal{T}_\mathcal{F}(\hat{x}) = -\mathcal{T}_\mathcal{J}(\hat{x}) - \frac{2\lambda}{V} s_\Omega\,. \tag{4.5}$$

From the definition of the Young modulus (4.1), we remark that (4.5) always measures the sensitivity of $\mathcal{T}_\mathcal{F}$ when the two materials are interchanged within the domain. Then, the computation of (4.5) is carried out using the expressions (3.15) with $\gamma = \gamma^*$ if $x \in \Omega^h$; and $\gamma = 1/\gamma^*$ if $x \in \Omega^w$. Having made the previous consideration and in order to solve the optimization problem (4.3), we use the topology optimization algorithm proposed in [1]. This algorithm is based on the concept of level-set domain representation and uses the topological derivative (4.5) as a feasible descent direction to minimize the cost function. This class of algorithms has been successfully applied in research areas related to topological optimization such as: microstructure of materials [2], load bearing structures [1], thermal conductors [11] and load bearing structures subjected to pointwise stress constraint [3,4]. For a detailed development of the algorithm we refer to the previous references.

5 Numerical Examples

Here we present five numerical examples associated to the topological optimization procedure outlined in the previous section. In all examples we set the Young modulus $E = 2.1\ GPa$, Poisson's ratio $\nu = 0.3$, the contrast parameter $\gamma = 1 \times 10^{-3}$ and the force $F = 1 \times 1^9\ N$. In the figures, the topology is identified by the strong material distribution (in black) and the inclusions of weak material (in white) are used to mimic the holes. Furthermore, the thick lines that appear

on the figures are used to denote clamped boundary conditions ($u|_{\Gamma_D} = 0$). The volume constraint is imposed with an exact quadratic penalization scheme. The function $p(y)$ used in the examples has the same behavior as presented in Fig. 1b. The variational equation (2.8) was solved using standard finite element technique. In particular, the three-node triangular elements are used to discretize the domain in a structured fashion.

5.1 Example 1

In this first example we consider a unit square panel submitted to a force F applied on its right upper corner, as shown in Fig. 3a. The volume constraint is of 50 % of the initial volume. In Fig. 3b we show the optimal topology without the contact condition. Then, a contact condition is applied in the bottom side with a gap of $g = 0.10$, see Fig. 3a where $c = 0.20$ and $d = 0.20$, and the parameter b is such

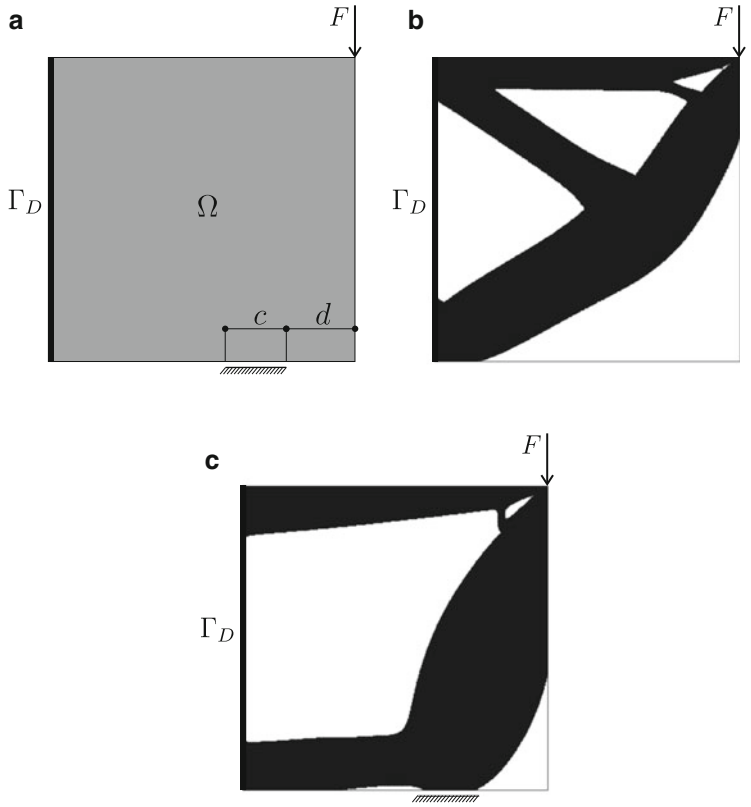


Fig. 3 Example 1. Results. (**a**) Contact problem. (**b**) Without contact. (**c**) With contact

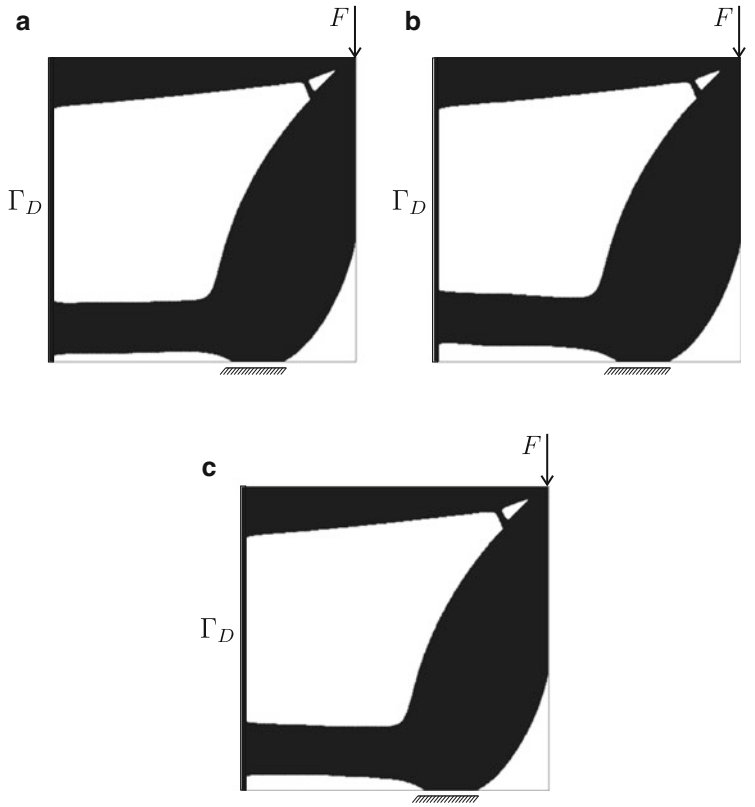


Fig. 4 Example 1. Results for different values of the gap. (**a**) $g = 0.15$. (**b**) $g = 0.20$. (**c**) $g = 0.25$

that the function p reaches the value of $p(y) = 1 \times 10^{15}$. In Fig. 3c is presented the obtained topology, where the effect of the contact condition is evident.

In Fig. 4, we present the obtained results for three different values for the gap, i.e. $g = \{0.15, 0.20, 0.25\}$ (the result for the gap $g = 0.10$ is shown in Fig. 3c).

In order to evaluate the effect of the function $p(y)$ in the optimal topology, in Fig. 5 the obtained results for different values of function $p(y)$ are presented. For this example, we set the gap in $g = 0.10$ and the parameter b, in each case, is such that the function p reaches the values $p = \{8 \times 10^{12}, 1 \times 10^{13}, 1 \times 10^{20}\}$ (the result for $p = 1 \times 10^{15}$ is shown in Fig. 3c).

For the volume fraction and set of parameters studied, the optimal topology (without the contact condition) is characterized by the classical four bars structure connecting the load F with the clamped boundary condition. When the contact condition is applied, the optimal topology shows that only three bars are needed. The influence of the value of the gap g and the function $p(y)$ is clear in all cases and tends to modify the shape of the lower bar of the structure.

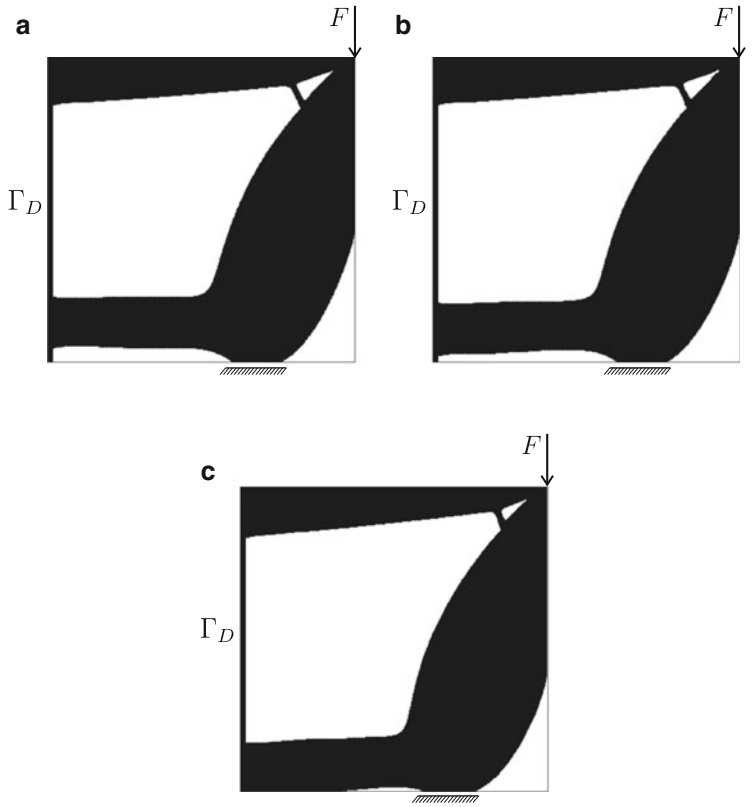


Fig. 5 Example 1. Results for different values of function $p(y)$. (**a**) $p(y) = 8^{12}$. (**b**) $p(y) = 1^{13}$. (**c**) $p(y) = 1^{20}$

5.2 Example 2

In this example we present the optimal topology design of a cantilever beam with a load F applied in the middle right side of its rectangular domain. The domain of the beam is a rectangular plane with dimensions of 2.00×1.00. The contact region is located in the bottom of the plane with length c, as shown in Fig. 6a. The volume constraint is of 40 % of the initial volume, the gap is $g = 0.1$ and the parameter b is such that the function p reaches the value of $p = 1 \times 10^{15}$. In this example, we study the influence of the length of the contact region in the optimal topology. In Fig. 6b, we present the result without considering the contact condition. In Fig. 6c–e is shown the results for three different values of parameter $c = \{0.20, 0.50, 0.70\}$.

In this example, the optimal topology (without consider the contact condition) is symmetric with respect to horizontal axis where the load is applied. However, when the contact condition is considered, the obtained structure loses its symmetry and became more complex. The influence of the length of the potential contact region is obvious in all studied cases.

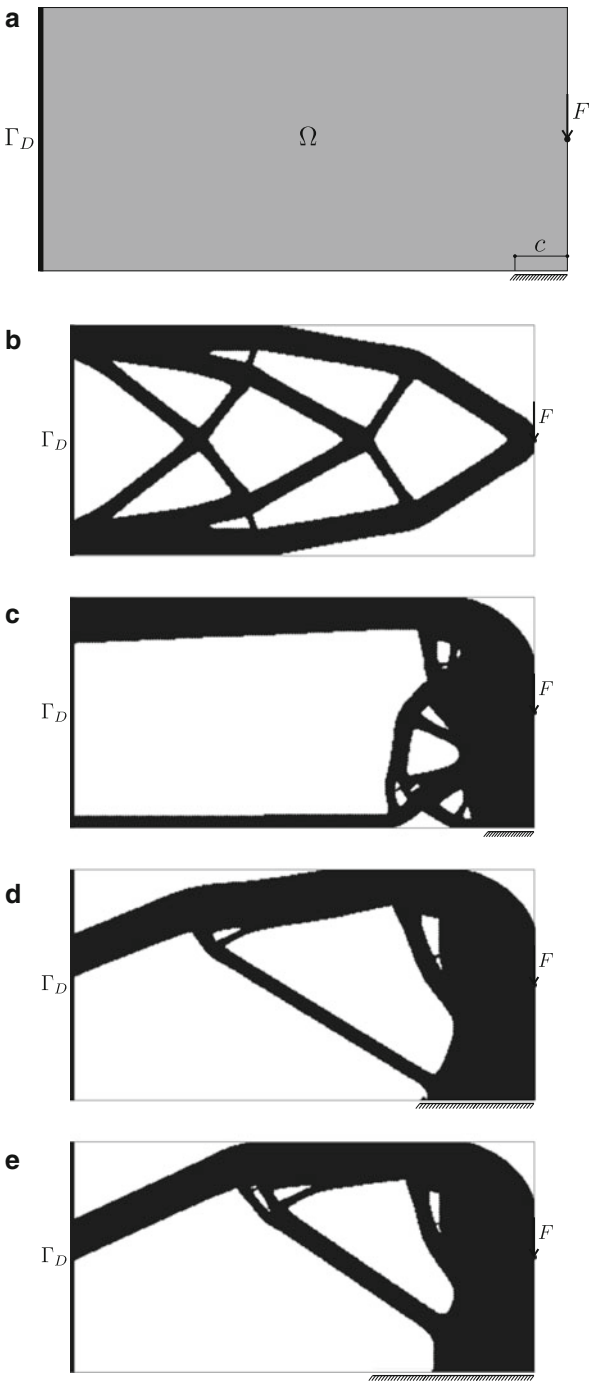


Fig. 6 Example 2. Results for different lengths of contact region. (**a**) Contact problem. (**b**) Without contact. (**c**) $c = 0.20$. (**d**) $c = 0.50$. (**e**) $c = 0.70$

5.3 Example 3

Now we consider the same domain and boundary conditions as in the previous example. Here we create a square hole of size 0.25×0.25 centered at the rectangular panel and the contact region is located on the top side of the hole, see Fig. 7a. The volume constraint is of 40 % of the initial volume and the gap is $g = 1 \times 10^{-5}$. The result for the case without the contact condition is presented in Fig. 7b. In Fig. 7c, we show the obtained topology considering the contact problem. Note the similarity in the results, without the boundary condition in the contact region, between this example and the previous (Fig. 6b). The conclusions and comments presented in the previous example remain valid for this one.

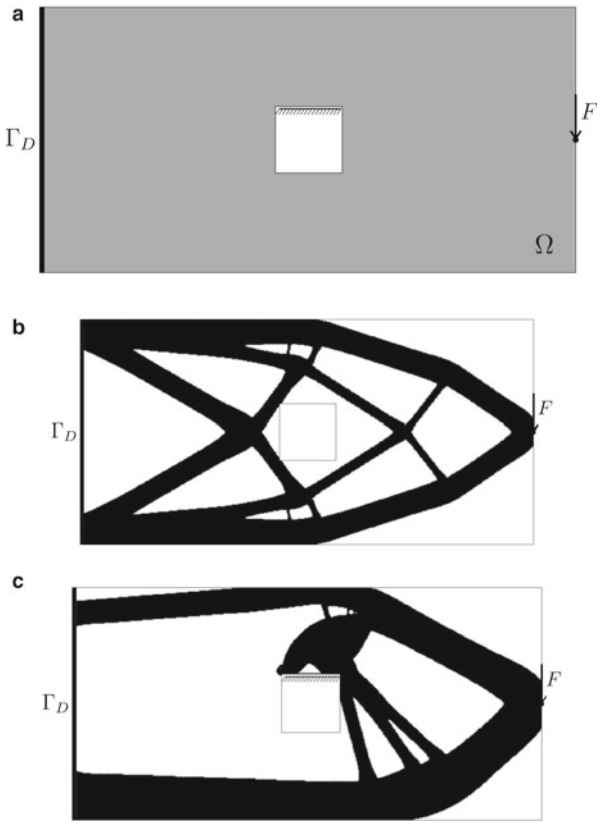


Fig. 7 Example 3. Results. (**a**) Contact problem. (**b**) Without contact. (**c**) With contact

5.4 Example 4

In this example, the design of a unit square panel subjected to two forces F applied at the corners of the top side with a volume constraint of 30 % of the initial volume is presented. The contact region is also in the top side of the panel, located a distance $d = 0.25$ from the right side and length $c = 0.50$. The gap considered is $g = 1 \times 10^{-3}$ and the parameter b is such that the function p reaches the value of $p = 1 \times 10^{15}$. The aim of this example is show the influence of the contact condition in the complexity of the final topology. The results with and without considering the contact condition are presented in Fig. 8c and b, respectively. As can be seen, topology changes from a very simple (two bars in the direction of the applied forces)

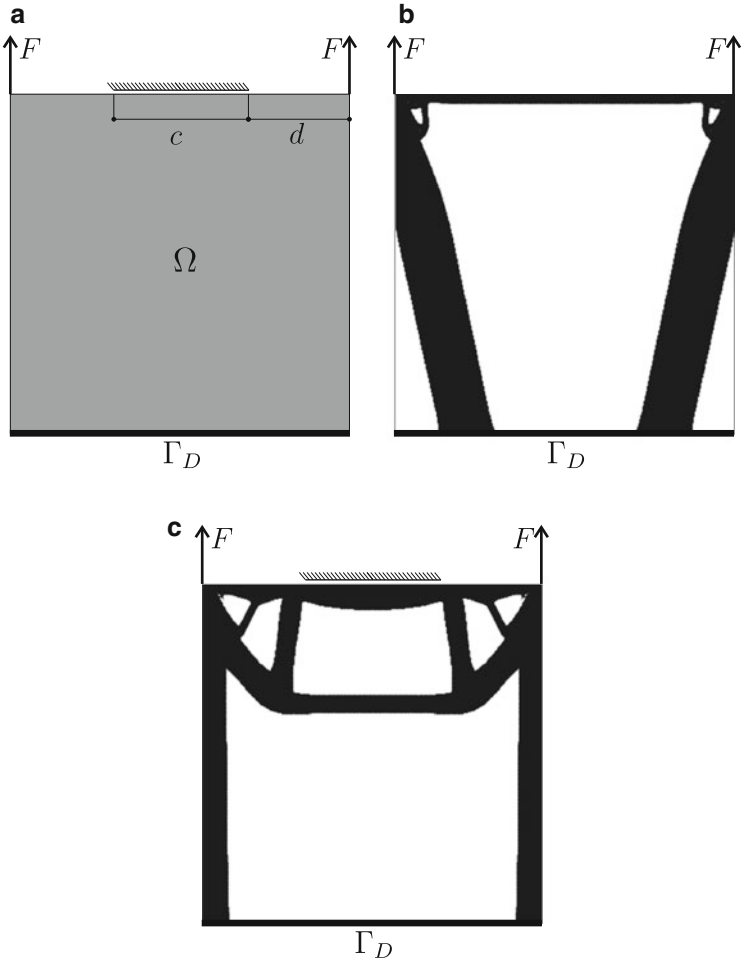


Fig. 8 Example 4. Results. (**a**) Contact problem. (**b**) Without contact. (**c**) With contact

to a more complex. The complexity is characterized by a structure of bars (similar to a small bridge) connecting the two bars in the direction of the applied forces.

5.5 Example 5

In this last example, we consider the topology design of a rectangular panel with height = 1.2 and width = 1.0, with a square hole in the right side of the domain. The design domain, boundary condition and the system of applied forces are presented in Fig. 9a, where $c = 0.20$ and $d = 0.40$. This example can be seen as the classical case of topology design of a gripping mechanism [1]. On the potential contact region the gap is $g = 1 \times 1^{-5}$ and the function p reaches the value of 1×1^{15}. The volume constraint imposed is of 50 % of the initial volume. The results are presented in Fig. 9c and b.

Again, in this example the effect of the contact model is manifested in the complexity of the optimal topology.

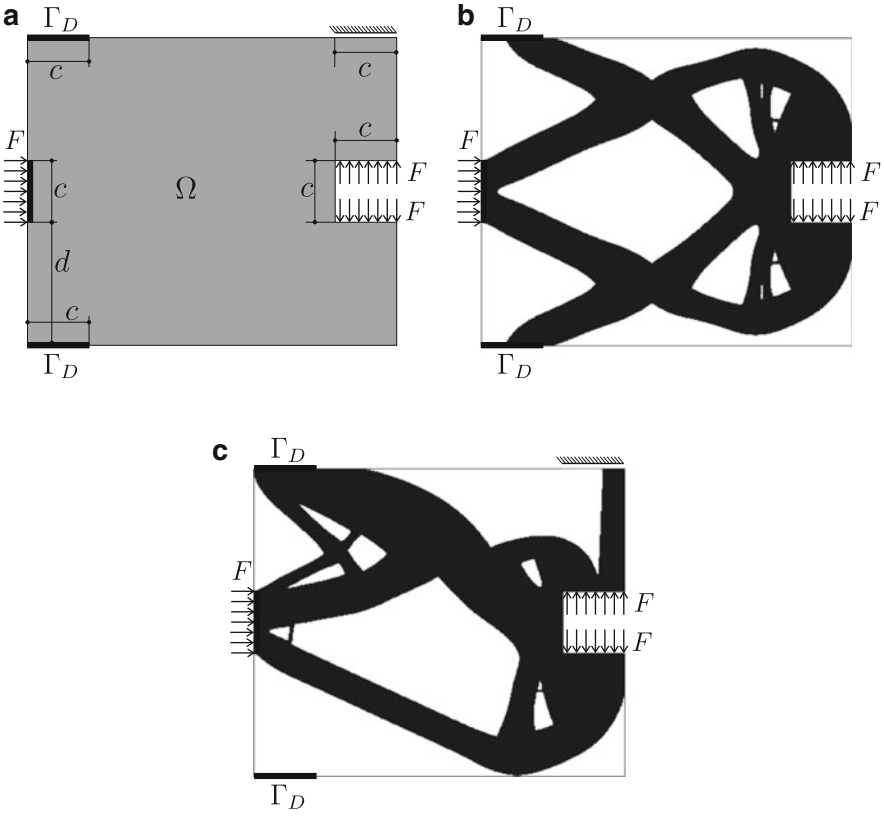


Fig. 9 Example 5. Results. (**a**) Contact problem. (**b**) Without contact. (**c**) With contact

6 Final Remarks

An analytical expression for the topological derivative of the energy shape functional associated to a frictionless contact model that allows a finite interpenetration between a two-dimensional elastic body and a rigid foundation has been presented. As topological perturbation, a disc with a different material has been considered in the analysis. The final formula is a general simple analytical expression in terms of the solution of the state equation and the constitutive parameters evaluated in each point of the unperturbed domain. The associated topological sensitivity has been used in a structural design algorithm based on the topological derivative and a level-set domain representation method. The robustness of the optimization procedure has been analyzed through some numerical experiments of compliance topology optimization of elastic structures subjected to volume constraint. Finally, we remark that the optimization procedure is conditioned by the contact model to produce more complex topologies that obtained by considering a unilateral contact condition and approaches such as SIMP-model.

Acknowledgements This research is partially supported by LabEx CARMIN–CIMPA SMV programme (France), CONICET (National Council for Scientific and Technical Research, Argentina) and PID-UTN (Research and Development Program of the National Technological University, Argentina) under grant PID/UTN 1420. The work of J. Stebel was supported by the ESF grant Optimization with PDE Constraints, by the Czech Science Foundation (GAČR) grant no. 201/09/0917 and RVO 67985840. The supports of these agencies are gratefully acknowledged.

Jan Sokolowski is supported by the Brazilian Research Council (CNPq), through the Special Visitor Researcher Framework of the Science Without Borders.

References

[1] S. Amstutz, H. Andrä, A new algorithm for topology optimization using a level-set method. J. Comput. Phys. **216**, 573–588 (2006)

[2] S. Amstutz, S.M. Giusti, A.A. Novotny, E.A. de Souza Neto, Topological derivative for multi-scale linear elasticity models applied to the synthesis of microstructures. Int. J. Numer. Methods Eng. **84**, 733–756 (2010)

[3] S. Amstutz, A.A. Novotny, Topological optimization of structures subject to von Mises stress constraints. Struct. Multidisciplinary Optim. **41**, 407–420 (2010)

[4] S. Amstutz, A.A. Novotny, E.A. de Souza Neto, Topological derivative-based topology optimization of structures subject to Drucker-Prager stress constraints. Comput. Methods Appl. Mech. Eng. **233–236**, 123–136 (2012)

[5] I.I. Argatov, J. Sokołowski, On asymptotic behavior of the energy functional for the Signorini problem under small singular perturbation of the domain. J. Comput. Math. Math. Phys. **43**, 742–756 (2003)

[6] G. Cardone, S.A. Nazarov, J. Sokołowski, Asymptotic analysis, polarization matrices, and topological derivatives for piezoelectric materials with small voids. SIAM J. Control Optim. **48**, 3925–3961 (2010)

[7] C. Eck, J. Jarušek, J. Starà, Normal compliance contact models with finite interpenetration. Technical Report Stuttgart Research Centre for Simulation Technology (SRC SimTech) (Universitĺat Stuttgart, Stuttgart 2012)

[8] E.A. Fancello, Topology optimization of minimum mass design considering local failure constraints and contact boundary conditions. Struct. Multidisciplinary Optim. **32**, 229–240 (2006)
[9] G. Frémiot, W. Horn, A. Laurain, M. Rao, J. Sokołowski, *On the Analysis of Boundary Value Problems in Nonsmooth Domainst'*. Dissertationes Mathematicae (Rozprawy Matematyczne), vol. 462 (Warsaw, Poland, 2009)
[10] P. Fulmanski, A. Lauraine, J.F. Scheid, J. Sokołowski, A level set method in shape and topology optimization for variational inequalities. Int. J. Appl. Math. Comput. Sci. **17**, 413–430 (2007)
[11] S.M. Giusti, A.A. Novotny, Topological derivative for an anisotropic and heterogeneous heat diffusion problem. Mech. Res. Commun. **46**, 26–33 (2012)
[12] S.M. Giusti, A.A. Novotny, E.A. de Souza Neto, R.A. Feijóo, Sensitivity of the macroscopic elasticity tensor to topological microstructural changes. J. Mech. Phys. Solids **57**, 555–570 (2009)
[13] S.M. Giusti, A.A. Novotny, J. Sokołowski, Topological derivative for steady-state orthotropic heat diffusion problem. Struct. Multidisciplinary Optim. **40**, 53–64 (2010)
[14] S.M. Giusti, J. Stebel, J. Sokołowski, On topological derivative for contact problem in elasticity. Technical Report HAL-00734652 (Institut Elie Cartan de Mathematiqt'es, Universitt'e de Lorraine, Nancy, 2012)
[15] D. Hilding, A. Klarbring, J. Petersson, Optimization of structures in unilateral contact. Appl. Mech. Rev. **52**, 139–160 (1999)
[16] I. Hlaváček, A.A. Novotny, J. Sokołowski, A. Żochowski, On topological derivatives for elastic solids with uncertain input data. J. Optim. Theory Appl. **141**, 569–595 (2009)
[17] M. Iguernane, S.A. Nazarov, J.-R. Roche, J. Sokołowski, K. Szulc, Topological derivatives for semilinear elliptic equations. Int. J. Appl. Math. Comput. Sci. **19**, 191–205 (2009)
[18] J. Jarušsek, M. Krbec, M. Rao, J. Sokołowski, Conical differentiability for evolution variational inequalities. J. Differ. Equat. **193**, 131–146 (2003)
[19] A.M. Khludnev, J. Sokołowski, *Modelling and Control in Solid Mechanics* (Birkhäuser, Basel/Boston/Berlin, 1997)
[20] A.M. Khludnev, J. Sokołowski, Griffith formulae for elasticity systems with unilateral conditions in domains with cracks. Eur. J. Mech. A/Solids **19**, 105–119 (2000)
[21] A.M. Khludnev, J. Sokołowski, On differentation of energy functionals in the crack theory with possible contact between crack faces. J. Appl. Math. Mech. **64**, 464–475 (2000)
[22] S.A. Nazarov, J. Sokołowski, Asymptotic analysis of shape functionals. J. de Mathématiques Pures et Appliquées **82**, 125–196 (2003)
[23] A.A. Novotny, J. Sokołowski, *Topological Derivatives in Shape Optimization*. Interaction of mechanics and mathematics (Springer, Berlin/Heidelberg/New York, 2013)
[24] J. Petersson, M. Patriksson, Topology optimization of sheets in contact by a subgradient method. Int. J. Numer. Methods Eng. **40**, 1295–1321 (1997)
[25] J. Sokołowski, A. Żochowski, On the topological derivative in shape optimization. SIAM J. Control Optim. **37**, 1251–1272 (1999)
[26] J. Sokołowski, A. Żochowski, Optimality conditions for simultaneous topology and shape optimization. SIAM J. Control Optim. **42**, 1198–1221 (2003)
[27] J. Sokołowski, A. Żochowski, Modelling of topological derivatives for contact problems. Numer. Math. **102**, 145–179 (2005)
[28] J. Sokołowski, J.P. Zolésio, *Introduction to Shape Optimization - Shape Sensitivity Analysis* (Springer, Berlin/Heidelberg/New York, 1992)
[29] N. Strömberg, A. Klarbring, Topology optimization of structures in unilateral contact. Struct. Multidisciplinary Optim. **41**, 57–64 (2010)

Convex Programming with Separable Ellipsoidal Constraints: Application in Contact Problems with Orthotropic Friction

Jaroslav Haslinger, Radek Kučera, and Tomáš Kozubek

Abstract This contribution presents an algorithm for constrained minimization of strictly convex quadratic functions subject to simple bounds and separable ellipsoidal constraints. The algorithm is used for numerical solution of discretized 3D contact problems with orthotropic friction. These problems have been solved by a polygonal approximation of the friction cone. Our algorithm enables us to use the original friction cone without any approximation. Results of model examples are shown.

Keywords Contact problems with orthotropic friction • Convex programming • Separable ellipsoidal constraints

Mathematics Subject Classification (2010). Primary 90C25; Secondary 35J86, 49M25, 74P10.

1 Introduction

Methods for numerical minimization of quadratic functions subject to convex constraints have been intensively developed in last decades [1–3,17] and nowadays they are an inherent part of many packages. These methods, however, are integrated into the packages in a fairly general setting. Therefore, they usually cannot be directly used in large scale problems arising, e.g., from finite element approximations. For this reason, the development of methods which take into account specifics of problems to be solved is important. Potential features which may be beneficial are

J. Haslinger (✉)
KNM MFF UK, Sokolovská 83, 18675 Praha 8, Czech Republic
e-mail: hasling@karlin.mff.cuni.cz

R. Kučera • T. Kozubek
IT4Innovations, VŠB-TUO, 17. listopadu 15, 70833 Ostrava, Czech Republic
e-mail: radek.kucera@vsb.cz; tomas.kozubek@vsb.cz

R. Hoppe (ed.), *Optimization with PDE Constraints*, Lecture Notes in Computational Science and Engineering 101,
DOI 10.1007/978-3-319-08025-3_7

the following: *a)* the number of variables subject to constraints is much lower than the total number of all variables; *b)* each variable appears in one constraint at most, i.e., the constraints are separable. In [13, 14], the author introduced and analyzed a new method for minimization of strictly convex quadratic functions with separable convex constraints. The separable character of constraints simplified the analysis that was based on the Karusch-Kuhn-Tucker (KKT) conditions. Their geometrical interpretation enabled to generalize an idea of the reduced gradient introduced originally for simple bound problems [4]. The resulting algorithm is closely related to the Rosen method [16]. Clearly, the efficient implementation of the algorithm strongly depends on the specific form of the constraint functions.

This study was motivated by necessity to solve numerically 3D contact problems with friction [10]. So far such problems have been solved by a polygonal approximation of the Coulomb friction cone [18]. The presented algorithm enables us to use the original Coulomb friction cone without any approximation. Since the number of unilateral constraints describing contact conditions is much smaller in comparison with the total number of all variables (hence *a)*), one of the efficient approaches for solving such problems is based on an appropriate discretization of the dual variational formulation, i.e., the formulation in terms of the Lagrange multipliers which are defined on the contact boundary. There are two vectors of the Lagrange multipliers in the discrete setting of frictional contact problems: one, denoted as $\bar{\boldsymbol{\lambda}}_\nu$, releases the unilateral constraints and is subject only to a sign condition; the second one, denoted as $\bar{\boldsymbol{\lambda}}_t := (\bar{\boldsymbol{\lambda}}_{t_1}, \bar{\boldsymbol{\lambda}}_{t_2})$, regularizes the non-smooth frictional term and is subject to convex constraints imposed on disjoint pairs of its components (hence *b)*). For an isotropic friction law, when frictional effects are the same in all directions, the constraints reduce to simple circular (spherical) ones, i.e., the zero level sets of the constraint functions are circles in $\mathbb{R}^2$.

The aim of the contribution is to extend this method to the case of separable ellipsoidal constraints. A simple change of variables permits to transform the ellipsoidal constraints to the circular ones. In computations, however, it turns out that the original setting (i.e., with the ellipsoidal constraints) is usually better for the performance of the algorithm, especially, in the case of strongly eccentric ellipses. Again, the minimization of functions with this type of constraints was motivated by practical needs. Indeed, the dual variational formulation of 3D contact problems with orthotropic friction (i.e., friction effects are now different in two à-priori given perpendicular directions) leads to separable ellipsoidal constraints for pairs made of the components of $\bar{\boldsymbol{\lambda}}_t$.

The paper is organized as follows. In Sect. 2 we shortly recall results from [13, 14]. The main attention will be paid to the numerical computation of the projection onto the ellipse which is an important ingredient of our algorithm. Unlike the projection onto the circle, this one is far from to be so simple. Finally in Sect. 3, we first derive the algebraic form of the dual formulation of 3D contact problems with orthotropic Coulomb friction and then, in Sect. 4, we apply our algorithms to several model examples.

2 Minimization Subject to Separable Ellipsoidal Constraints

In this section, we consider the following problem:

$$\text{find} \quad \bar{\mathbf{x}}^* = \arg\min\{q(\bar{\mathbf{x}}) : \bar{\mathbf{x}} \in \Lambda\}, \tag{2.1}$$

where $q(\bar{\mathbf{x}}) = \frac{1}{2}\bar{\mathbf{x}}^\top \mathbf{A}\bar{\mathbf{x}} - \bar{\mathbf{x}}^\top \bar{\mathbf{b}}$ with symmetric, positive definite $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\bar{\mathbf{b}}, \bar{\mathbf{x}} \in \mathbb{R}^n$, $\bar{\mathbf{x}} = (x_1, \ldots, x_n)^\top$, $n = 3m$, and $\Lambda = \Lambda_1 \times \cdots \times \Lambda_{2m}$ defined by

$$\Lambda_i = \{x_i \in \mathbb{R} : x_i \geq l_i\},$$

$$\Lambda_{i+m} = \{(x_{i+m}, x_{i+2m})^\top \in \mathbb{R}^2 : \left(\frac{x_{i+m} - c_i}{a_i}\right)^2 + \left(\frac{x_{i+2m} - c_{i+m}}{a_{i+m}}\right)^2 \leq g_i^2\}$$

with given $l_i, c_i, c_{i+m} \in \mathbb{R}$, $g_i, a_i, a_{i+m} \in \mathbb{R}_+$ for $i = 1, \ldots, m$. As q is strictly convex on the closed covex set Λ, there is a unique solution $\bar{\mathbf{x}}^* \in \Lambda$ to (2.1). Before we give ideas of the active-set KPRGP algorithm (KKT Proportioning with Reduced Gradient Projections) analyzed in [7, 14], we introduce notation.

Let $\mathcal{N} = \{1, \ldots, n\}$ be the set of all indices and let $\mathcal{A}(\bar{\mathbf{x}}) \subseteq \mathcal{N}$ be the subset of indices of active constraints at $\bar{\mathbf{x}} \in \Lambda$:

$$\mathcal{A}(\bar{\mathbf{x}}) = \{i : x_i = l_i,\ 1 \leq i \leq m\}$$

$$\bigcup\{j : j = i + m, \left(\frac{x_{i+m} - c_i}{a_i}\right)^2 + \left(\frac{x_{i+2m} - c_{i+m}}{a_{i+m}}\right)^2 = g_i^2,\ 1 \leq i \leq m\}$$

$$\bigcup\{j : j = i + 2m, \left(\frac{x_{i+m} - c_i}{a_i}\right)^2 + \left(\frac{x_{i+2m} - c_{i+m}}{a_{i+m}}\right)^2 = g_i^2, 1 \leq i \leq m\}.$$

Let $\bar{\mathbf{r}}(\bar{\mathbf{x}}) = \mathbf{A}\bar{\mathbf{x}} - \bar{\mathbf{b}}$ denote the gradient of q at $\bar{\mathbf{x}} \in \mathbb{R}^n$. The orthogonal projection $\mathbf{P}_\Lambda$ onto Λ at $\bar{\mathbf{x}} \in \mathbb{R}^n$ is defined by

$$\mathbf{P}_\Lambda(\bar{\mathbf{x}}) = \arg\min_{\bar{\mathbf{y}}\in\Lambda} \|\bar{\mathbf{y}} - \bar{\mathbf{x}}\|. \tag{2.2}$$

As Λ is separable, $\mathbf{P}_\Lambda$ may be split into single projections P_{Λ_i} onto Λ_i. Let us introduce the *reduced gradient* of q at $\bar{\mathbf{x}} \in \Lambda$ for a fixed $\alpha > 0$ by:

$$\bar{\mathbf{r}}^{red}(\bar{\mathbf{x}}) = \frac{1}{\alpha}(\bar{\mathbf{x}} - \mathbf{P}_\Lambda(\bar{\mathbf{x}} - \alpha\bar{\mathbf{r}}(\bar{\mathbf{x}}))).$$

Note that the reduced gradient characterizes the optimality criterion to (2.1). Indeed, $\bar{\mathbf{x}}^*$ is the solution to (2.1) iff $\bar{\mathbf{r}}^{red}(\bar{\mathbf{x}}^*) = \mathbf{0}$. Moreover, if $\bar{\mathbf{x}} \neq \bar{\mathbf{x}}^*$ and $\alpha > 0$

is sufficiently small, then the negative reduced gradient $-\bar{\mathbf{r}}^{red}(\bar{\mathbf{x}})$ is a decrease direction at $\bar{\mathbf{x}} \in \Lambda$. To change appropriately the active set, we decompose $\bar{\mathbf{r}}^{red} := \bar{\mathbf{r}}^{red}(\bar{\mathbf{x}})$ into the *free reduced gradient* $\bar{\boldsymbol{\varphi}} := \bar{\boldsymbol{\varphi}}(\bar{\mathbf{x}})$ and the *chopped reduced gradient* $\bar{\boldsymbol{\psi}} := \bar{\boldsymbol{\psi}}(\bar{\mathbf{x}})$ as follows:

$$\bar{\boldsymbol{\varphi}}_{\mathcal{A}} = \mathbf{0}, \quad \bar{\boldsymbol{\varphi}}_{\mathcal{N}\setminus\mathcal{A}} = \bar{\mathbf{r}}^{red}_{\mathcal{N}\setminus\mathcal{A}},$$
$$\bar{\boldsymbol{\psi}}_{\mathcal{A}} = \bar{\mathbf{r}}^{red}_{\mathcal{A}}, \ \bar{\boldsymbol{\psi}}_{\mathcal{M}\setminus\mathcal{A}} = \mathbf{0},$$

where $\bar{\boldsymbol{\varphi}}_{\mathcal{A}}$ and $\bar{\boldsymbol{\varphi}}_{\mathcal{N}\setminus\mathcal{A}}$ denote the sub-vectors of $\bar{\boldsymbol{\varphi}}$ with components determined by the indices of $\mathcal{A} := \mathcal{A}(\bar{\mathbf{x}})$ and $\mathcal{N} \setminus \mathcal{A}$, respectively (similarly for $\bar{\mathbf{r}}^{red}$ and $\bar{\boldsymbol{\psi}}$).

We combine the following three steps to generate a sequence $\{\bar{\mathbf{x}}^{(l)}\}$ that approximates the solution $\bar{\mathbf{x}}^*$:

- the *expansion step*: $\bar{\mathbf{x}}^{(l+1)} = \bar{\mathbf{x}}^{(l)} - \alpha\bar{\boldsymbol{\varphi}}(\bar{\mathbf{x}}^{(l)})$,
- the *proportioning step*: $\bar{\mathbf{x}}^{(l+1)} = \bar{\mathbf{x}}^{(l)} - \alpha\bar{\boldsymbol{\psi}}(\bar{\mathbf{x}}^{(l)})$,
- the *conjugate gradient step*: $\bar{\mathbf{x}}^{(l+1)} = \bar{\mathbf{x}}^{(l)} - \alpha^{(l)}_{cg}\bar{\mathbf{p}}^{(l)}$, where the step-length $\alpha^{(l)}_{cg}$ and the conjugate gradient directions $\bar{\mathbf{p}}^{(l)}$ are computed recurrently [8]; the recurrence starts from $\bar{\mathbf{x}}^{(s)}$ generated by the last expansion or the proportioning step and satisfies $\mathcal{A}(\bar{\mathbf{x}}^{(l+1)}) = \mathcal{A}(\bar{\mathbf{x}}^{(s)})$.

The expansion step may add while the proportioning step may remove indices to/from the current active set. The conjugate gradient steps are used to carry out efficiently the minimization of q in the interior of the face $W(\bar{\mathbf{x}}^{(s)}) = \{\bar{\mathbf{x}} \in \Lambda \,|\, \bar{\mathbf{x}}_{\mathcal{A}} = \bar{\mathbf{x}}^{(s)}_{\mathcal{A}}, \mathcal{A} := \mathcal{A}(\bar{\mathbf{x}}^{(s)})\}$. Moreover, the algorithm exploits a given constant $\Gamma > 0$ in the *proportioning criterion*

$$\bar{\boldsymbol{\psi}}(\bar{\mathbf{x}}^{(l)})^\top \bar{\mathbf{r}}(\bar{\mathbf{x}}^{(l)}) \le \Gamma\, \bar{\boldsymbol{\varphi}}(\bar{\mathbf{x}}^{(l)})^\top \bar{\mathbf{r}}(\bar{\mathbf{x}}^{(l)}) \tag{2.3}$$

to decide which of the steps will be performed.

Algorithm KPRGP

Let $\bar{\mathbf{x}}^{(0)} \in \Lambda$, $\Gamma > 0$, $\alpha \in (0, 2\|\mathbf{A}\|^{-1})$, and $\varepsilon > 0$ be given. For $\bar{\mathbf{x}}^{(l)}$, $\bar{\mathbf{x}}^{(s)}$ known, $0 \le s \le l$, where $\bar{\mathbf{x}}^{(s)}$ is computed by the last expansion or proportioning step, choose $\bar{\mathbf{x}}^{(l+1)}$ by the following rules:

(i). If $\|\bar{\mathbf{r}}^{red}(\bar{\mathbf{x}}^{(l)})\| \le \varepsilon$, return $\bar{\mathbf{x}} = \bar{\mathbf{x}}^{(l)}$.

(ii). If $\bar{\mathbf{x}}^{(l)}$ fulfils (2.3), try to generate $\bar{\mathbf{x}}^{(l+1)}$ by the conjugate gradient step. If $\bar{\mathbf{x}}^{(l+1)} \in \operatorname{Int} W(\bar{\mathbf{x}}^{(s)})$, accept it, otherwise generate $\bar{\mathbf{x}}^{(l+1)}$ by the expansion step.

(iii). If $\bar{\mathbf{x}}^{(l)}$ does not fulfil (2.3), generate $\bar{\mathbf{x}}^{(l+1)}$ by the proportioning step.

The convergence rate of this algorithm derived in [14] does not depend on the type of convex constraints. However, the implementation requires to compute the projection $\mathbf{P}_\Lambda$ via the single projections P_{Λ_i} and $P_{\Lambda_{i+m}}$, $1 \le i \le m$. In the rest of this section we show how to compute these projections.

The set Λ_i, $i = 1, \ldots, m$ represents the simple bound for which the projection is trivial:

$$P_{\Lambda_i}(x_i) = \begin{cases} x_i & \text{if } x_i \geq l_i, \\ l_i & \text{otherwise.} \end{cases}$$

The projection onto Λ_{i+m}, $i = 1, \ldots, m$ is more involved. To simplify our presentation we denote $\mathbf{x}_i = (x_{i+m}, x_{i+2m})^\top \in \mathbb{R}^2$ and $\mathbf{c}_i = (c_i, c_{i+m})^\top \in \mathbb{R}^2$. The corresponding projection is given by

$$P_{\Lambda_{i+m}}(\mathbf{x}_i) = \begin{cases} \mathbf{x}_i & \text{if } \left(\dfrac{x_{i+m} - c_i}{a_i}\right)^2 + \left(\dfrac{x_{i+2m} - c_{i+m}}{a_{i+m}}\right)^2 \leq g_i^2, \\ \mathbf{y}_i & \text{otherwise,} \end{cases}$$

where we will specify how to get $\mathbf{y}_i \in \mathbb{R}^2$. We distinguish two situations. If $a_i = a_{i+1}$, then Λ_{i+m} describes the circular constraint for which $\mathbf{y}_i$ is given by the explicit formula:

$$\mathbf{y}_i = \mathbf{c}_i + \frac{a_i g_i}{\|\mathbf{x}_i - \mathbf{c}_i\|}(\mathbf{x}_i - \mathbf{c}_i). \tag{2.4}$$

If $a_i \neq a_{i+1}$, then $\mathbf{y}_i$ is the closest point to $\mathbf{x}_i$ lying on the ellipse $\mathbf{e}_i := \mathbf{e}_i(t)$ (in the parametric representation):

$$\mathbf{e}_i(t) = \mathbf{c}_i + g_i \begin{pmatrix} a_i \cos t \\ a_{i+m} \sin t \end{pmatrix}, \quad t \in [0, 2\pi).$$

Let t^* be the value of t such that $\mathbf{y}_i = \mathbf{e}_i(t^*)$. Such t^* satisfies the following orthogonality condition:

$$(\mathbf{x}_i - \mathbf{e}_i(t))^\top \mathbf{e}_i'(t) = 0. \tag{2.5}$$

Although (2.5) is the equation in $\mathbb{R}^1$, its solution is not unique. The reason is that (2.5) is equivalent to the fourth degree polynomial equation with either two or four roots. Fortunately, one can recognize correct t^* characterized by the fact that $\mathbf{y}_i$ belongs to the same quadrant as $\mathbf{x}_i$, provided that the local coordinate system (in $\mathbb{R}^2$) coincides with the half-axes of the ellipse. To perform efficiently computations of t^* via (2.5), we combine the Newton and bisection methods (in $\mathbb{R}^1$). The resulting algorithm may benefit from fast convergence of the Newton iterations while the bisection steps ensure convergence to t^*. A long sequence of bisection steps are generated in situations when the root t^* is close to an inflection point of the function in (2.5) (that is not excluded in general).

Remark 2.1. Another way how to solve (2.1) by ALGORITHM KPRGP consists in transforming the ellipsoidal constraints to the circular ones using the substitution:

$$\bar{\mathbf{y}} = \mathbf{D}^{-1}\bar{\mathbf{x}},$$

where $\mathbf{D} = diag(1, \dots, 1, a_1, \dots, a_{2m}) \in \mathbb{R}^{n\times n}$. This leads to the problem in terms of the new variable $\bar{\mathbf{y}}$:

$$\text{find} \quad \bar{\mathbf{y}}^* = \arg\min\{q(\bar{\mathbf{y}}) : \ \bar{\mathbf{y}} \in \Lambda\}, \tag{2.6}$$

where $q(\bar{\mathbf{y}}) = \frac{1}{2}\,\bar{\mathbf{y}}^\top \mathbf{DAD}\bar{\mathbf{y}} - \bar{\mathbf{y}}^\top \mathbf{D}\bar{\mathbf{b}}$ and $\Lambda = \Lambda_1 \times \cdots \times \Lambda_{2m}$ is defined by

$$\Lambda_i = \{y_i \in \mathbb{R} : \ y_i \geq l_i\},$$

$$\Lambda_{i+m} = \{(y_{i+m}, y_{i+2m})^\top \in \mathbb{R}^2 : \ (y_{i+m} - d_i)^2 + (y_{i+2m} - d_{i+m})^2 \leq g_i^2\},$$

with $d_i = c_i/a_i$, $d_{i+m} = c_{i+m}/a_{i+m}$ for $i = 1, \dots, m$. Problem (2.6) is the special case of (2.1) for which the projections can be computed by (2.4). On the other hand, the condition number of **DAD** is usually greater than the one of **A**, especially, when the ellipses in the original problem are strongly eccentric. In this case, the convergence factor of ALGORITHM KPRGP derived in [14] is smaller for (2.1) that may result in a better performance of computations.

3 Numerical Solution of 3D Contact Problems with Orthotropic Coulomb Friction

The minimization algorithm from the previous section will be now used for the numerical solution of 3D contact problems with orthotropic Coulomb friction. Recall that contact mechanics is a branch of mechanics of solids which studies the behavior of loaded systems of deformable bodies being in mutual contact. Mathematical models of such problems are given by equations involving non-smooth multivalued mappings due to non-penetration and friction conditions on common parts of the boundary. In contrast to isotropic friction, effects of orthotropic friction are different in directions of two orthogonal orthotropy axis. We first present the weak formulation of such problems, then we give their finite element discretization and the transformation of the resulting algebraic problem into a new one having a structure required by the algorithm KPRGP.

Our system consists of two elastic bodies represented by *polyhedral* domains $\Omega^k \subset \mathbb{R}^3$ whose boundaries are split into three disjoint parts Γ_u^k, Γ_p^k, and Γ_c^k, $k = 1, 2$. Denote $\Omega = \Omega^1 \cup \Omega^2$, $\Gamma_u = \Gamma_u^1 \cup \Gamma_u^2$, $\Gamma_p = \Gamma_p^1 \cup \Gamma_p^2$, and $\Gamma_c = \Gamma_c^1 \cup \Gamma_c^2$. The zero displacements will be prescribed on Γ_u, while surface tractions of density $\mathbf{p} \in (L^2(\Gamma_p))^3$ act on Γ_p. Both bodies are in contact along Γ_c^1 and Γ_c^2 in

the undeformed state. In what follows we shall suppose that $\Gamma_u^k \neq \emptyset$, $k = 1, 2$ and $\Gamma_c^1 = \Gamma_c^2$, i.e. there is no gap between Ω^1 and Ω^2. On Γ_c unilateral and friction conditions will be prescribed. Finally, Ω is subject to body forces of density $\mathbf{f} \in (L^2(\Omega))^3$. Our aim is to find an equilibrium state of this system.

Before we give the weak formulation of this problem, we introduce several notation and function sets which will be needed. Let $\mathbf{u} : \Omega \mapsto \mathbb{R}^3$ be a deformation field in Ω and $\mathbf{u}^k := \mathbf{u}_{|\Omega^k}$ its restriction to Ω^k, $k = 1, 2$. By $\boldsymbol{\epsilon}(\mathbf{u}) = \frac{1}{2}(\nabla\mathbf{u} + (\nabla\mathbf{u})^\top)$ we denote the linearized strain tensor, while $\boldsymbol{\sigma}(\mathbf{u})$ is the stress tensor linked to $\boldsymbol{\epsilon}(\mathbf{u})$ by means of a linear Hooke's law whose coefficients satisfy the usual symmetry and ellipticity conditions [15]. The outward unit normal vector to $\partial\Omega^1$ at a point $\mathbf{x} \in \Gamma_c$ is denoted as $\boldsymbol{\nu}(\mathbf{x})$. The orthotropy axis of friction at $\mathbf{x} \in \Gamma_c$ are given by a pair of orthogonal vectors $\mathbf{t}_1(\mathbf{x})$ and $\mathbf{t}_2(\mathbf{x})$ lying in the tangent plane to Γ_c at $\mathbf{x}$. The relative normal contact displacement at $\mathbf{x} \in \Gamma_c$ is defined by $u_\nu(\mathbf{x}) := (\mathbf{u}^1(\mathbf{x}) - \mathbf{u}^2(\mathbf{x}))^\top \boldsymbol{\nu}(\mathbf{x})$ and $\sigma_\nu(\mathbf{u}(\mathbf{x})) := \boldsymbol{\nu}^\top(\mathbf{x})\boldsymbol{\sigma}(\mathbf{u}^1(\mathbf{x}))\boldsymbol{\nu}(\mathbf{x})$ is the normal contact stress. Similarly, $\mathbf{u}_t(\mathbf{x}) = (u_{t_1}(\mathbf{x}), u_{t_2}(\mathbf{x}))^\top$, $\boldsymbol{\sigma}_t(\mathbf{u}(\mathbf{x})) = (\sigma_{t_1}(\mathbf{u}(\mathbf{x})), \sigma_{t_2}(\mathbf{u}(\mathbf{x})))^\top$ are the relative tangential contact displacement and the tangential contact stress at $\mathbf{x} \in \Gamma_c$, respectively, whose components are $u_{t_i}(\mathbf{x}) := (\mathbf{u}^1(\mathbf{x}) - \mathbf{u}^2(\mathbf{x}))^\top \mathbf{t}_i(\mathbf{x})$ and $\sigma_{t_i}(\mathbf{u}(\mathbf{x})) := \mathbf{t}_i{}^\top(\mathbf{x})\boldsymbol{\sigma}(\mathbf{u}^1(\mathbf{x}))\boldsymbol{\nu}(\mathbf{x})$, $i = 1, 2$. In addition to orthotropy axis, friction will be characterized by two positive, bounded and continuous functions $\mathcal{F}_1$ and $\mathcal{F}_2$ whose values at $\mathbf{x} \in \Gamma_c$ define coefficients of friction in directions $\mathbf{t}_1(\mathbf{x})$ and $\mathbf{t}_2(\mathbf{x})$, respectively. The diagonal (2×2) matrix $diag\{\mathcal{F}_1, \mathcal{F}_2\}$ will be denoted by $\boldsymbol{\mathcal{F}}$. Finally, $\|\cdot\|$ stands for the Euclidean norm of vectors from $\mathbb{R}^2$.

Now we introduce the following function sets:

$$\mathbb{V} = \{\mathbf{v} = (\mathbf{v}^1, \mathbf{v}^2) \in (H^1(\Omega^1))^3 \times (H^1(\Omega^2))^3 |\ \mathbf{v} = \mathbf{0} \text{ on } \Gamma_u\},$$

$$\mathbb{K} = \{\mathbf{v} \in \mathbb{V} |\ v_\nu \leq 0 \text{ on } \Gamma_c\},$$

$$X_\nu = \{\varphi \in L^2(\Gamma_c) |\ \exists \mathbf{v} \in \mathbb{V} :\ \varphi = v_\nu \text{ on } \Gamma_c\},$$

$$X_\nu' = \text{ dual of } X_\nu,$$

$$X_\nu^+ = \{\varphi \in X_\nu |\ \varphi \geq 0 \text{ on } \Gamma_c\}.$$

The cone of all non-negative elements of X_ν' will be denoted by $X_{\nu+}'$ and $\langle\cdot,\cdot\rangle$ is a duality pairing on $X_\nu' \times X_\nu$. Next we shall suppose that $\|\boldsymbol{\mathcal{F}}\mathbf{v}_t\|$ belongs to $X_{\nu+}$ for any $\mathbf{v} \in \mathbb{V}$.

We start with the following auxiliary problem: given $g \in X_{\nu+}'$, find $\mathbf{u} := \mathbf{u}(g) \in \mathbb{K}$ satisfying

$$a(\mathbf{u}, \mathbf{v} - \mathbf{u}) + \langle g, \|\boldsymbol{\mathcal{F}}\mathbf{v}_t\| - \|\boldsymbol{\mathcal{F}}\mathbf{u}_t\|\rangle \geq L(\mathbf{v} - \mathbf{u}) \quad \forall \mathbf{v} \in \mathbb{K}, \tag{3.1}$$

where

$$a(\mathbf{u},\mathbf{v}) = \int_\Omega \boldsymbol{\sigma}(\mathbf{u}) : \boldsymbol{\epsilon}(\mathbf{v})\, d\mathbf{x} := \int_\Omega \sigma_{ij}(\mathbf{u})\epsilon_{ij}(\mathbf{v})\, d\mathbf{x},$$

$$L(\mathbf{v}) = \int_\Omega \mathbf{f}^\top \mathbf{v}\, d\mathbf{x} + \int_{\Gamma_p} \mathbf{p}^\top \mathbf{v}\, ds, \quad \mathbf{u},\mathbf{v} \in \mathbb{V}.$$

It is easy to show that (3.1) has a unique solution for any $g \in X'_{\nu+}$. In addition, (3.1) is equivalent to the following minimization problem:

$$\left.\begin{array}{l} \text{Find } \mathbf{u} \in \mathbb{K} \text{ such that} \\ J_g(\mathbf{u}) \le J_g(\mathbf{v}) \quad \forall \mathbf{v} \in \mathbb{K}, \end{array}\right\} \qquad (\mathcal{P}(g))$$

where $J_g(\mathbf{v}) = \frac{1}{2}a(\mathbf{v},\mathbf{v}) - L(\mathbf{v}) + \langle g, \|\mathcal{F}\mathbf{v}_t\|\rangle$. Problem $(\mathcal{P}(g))$ is the variational formulation of contact problems with orthotropic friction and a *given* slip bound g. Let us suppose that $-\sigma_\nu(\mathbf{u}(g)) \in X'_{\nu+}$ for every $g \in X'_{\nu+}$. Then one can define the mapping $\Psi : X'_{\nu+} \mapsto X'_{\nu+}$ by

$$\Psi : g \mapsto -\sigma_\nu(\mathbf{u}(g)) \quad \forall g \in X'_{\nu+}.$$

Definition 3.1. By a weak solution of 3D contact problems with orthotropic Coulomb friction we mean any $\mathbf{u} \in \mathbb{K}$ such that $\Psi(-\sigma_\nu(\mathbf{u})) = -\sigma_\nu(\mathbf{u})$, i.e., $-\sigma_\nu(\mathbf{u})$ is a fixed point of Ψ in $X'_{\nu+}$.

Remark 3.2. In the weak formulation of this problem, the following unilateral and friction conditions are hidden:

(unilateral conditions)

$$u_\nu \le 0,\ \sigma_\nu(\mathbf{u}) \le 0,\ u_\nu\sigma_\nu(\mathbf{u}) = 0 \quad \text{on } \Gamma_c,$$

(friction conditions)

$$\mathbf{u}_t(\mathbf{x}) = \mathbf{0} \Longrightarrow \|\mathcal{F}^{-1}\boldsymbol{\sigma}_t(\mathbf{u}(\mathbf{x}))\| \le -\sigma_\nu(\mathbf{u}(\mathbf{x})),$$

$$\mathbf{u}_t(\mathbf{x}) \ne \mathbf{0} \Longrightarrow \mathcal{F}^{-1}\boldsymbol{\sigma}_t(\mathbf{u}(\mathbf{x})) = \sigma_\nu(\mathbf{u}(\mathbf{x}))\frac{\mathcal{F}\mathbf{u}_t(\mathbf{x})}{\|\mathcal{F}\mathbf{u}_t(\mathbf{x})\|}, \quad \mathbf{x} \in \Gamma_c.$$

We use the method of *successive approximation* for finding fixed points of Ψ in $X'_{\nu+}$:

$$\left.\begin{array}{l} \text{given } g^{(0)} \in X'_{\nu+}; \\ \text{set } g^{(k+1)} = \Psi(g^{(k)}),\ k = 0, 1, \ldots. \end{array}\right\} \qquad (3.2)$$

To get the new iteration $g^{(k+1)}$ one has to solve problem $(\mathcal{P}(g^{(k)}))$.

Remark 3.3. Let us note that convergence of (3.2) in continuous setting of our problem is not guaranteed. The situation is somewhat different in the discrete case (for details see [9, 11]).

Since $(\mathcal{P}(g))$, $g \in X'_{v+}$ is the heart of (3.2), we focus in the subsequent part on its efficient numerical solution. For a discretization of $(\mathcal{P}(g))$ we use a finite element method. First we choose a finite dimensional space $\mathbb{V}_h \subset \mathbb{V}$, $\dim \mathbb{V}_h = n(h)$ of piecewise polynomial functions of the Lagrange type over partitions $\mathcal{T}_h^k$ of $\overline{\Omega}^k$, which are compatible with the decomposition of $\partial\Omega^k$ into Γ_u^k, Γ_p^k, and Γ_c, $k = 1, 2$. These partitions will be constructed in such a way that $\mathcal{T}^1_{h|\Gamma_c} = \mathcal{T}^2_{h|\Gamma_c}$. In particular this means that if $\mathbf{v}_h = (\mathbf{v}_h^1, \mathbf{v}_h^2) \in \mathbb{V}_h$, where $\mathbf{v}_h^k := \mathbf{v}_{h|\Omega^k}$, then the degrees of freedom (function values in our case) of $\mathbf{v}_h^1$ and $\mathbf{v}_h^2$ on $\overline{\Gamma}_c$ are prescribed at the same nodes of $\mathcal{T}_h^k$ on $\overline{\Gamma}_c$. Typically, $\mathbb{V}_h$ is made of P_1 tetrahedral elements. Finally set $\mathcal{T}_h = \mathcal{T}_h^1 \cup \mathcal{T}_h^2$ which is the partition of the whole $\overline{\Omega}$. By $\mathcal{C}$ we denote the set of all nodes $\mathbf{a}_1, \ldots, \mathbf{a}_m$ of $\mathcal{T}_h$ which are located on $\overline{\Gamma}_c$. To simplify our presentation we shall suppose that $\overline{\Gamma}_u^k \cap \overline{\Gamma}_c = \emptyset$, $k = 1, 2$ and Γ_c is a flat part of $\partial\Omega$. Then the discretization of $\mathbb{K}$ is defined by

$$\mathbb{K}_h = \{\mathbf{v}_h \in \mathbb{V}_h |\ v_{h\nu}(\mathbf{a}_i) \le 0 \quad \forall i = 1, \ldots, m\},$$

where $v_{h\nu}(\mathbf{a}_i) := (\mathbf{v}_h^1(\mathbf{a}_i) - \mathbf{v}_h^2(\mathbf{a}_i))^\top \boldsymbol{\nu}$, $i = 1, \ldots, m$, i.e., the non-penetration conditions in $\mathbb{K}_h$ are prescribed at the nodes of $\mathcal{C}$ only using the fact that $\boldsymbol{\nu}$ is constant along Γ_c. The approximation of $(\mathcal{P}(g))$ reads as follows:

$$\left.\begin{aligned} &\text{Find } \mathbf{u}_h := \mathbf{u}_h(g) \in \mathbb{K}_h \text{ such that} \\ &J_g(\mathbf{u}_h) \le J_g(\mathbf{v}_h) \quad \forall \mathbf{v}_h \in \mathbb{K}_h. \end{aligned}\right\} \qquad (\mathcal{P}(g))_h$$

Next we rewrite $(\mathcal{P}(g))_h$ into the algebraic form, i.e. the problem expressed by means of the nodal displacement vectors $\bar{\mathbf{v}} \in \mathbb{R}^n$ of $\mathbf{v}_h \in \mathbb{V}_h$, where $n := n(h) = \dim \mathbb{V}_h$. Since first two terms of J_g define the quadratic, coercive functional, its algebraic form leads to a quadratic function with a positive definite, symmetric, block diagonal matrix $\mathbf{K}$. The frictional term will be evaluated using an appropriate cubature formula. Suppose that the slip bound g is represented by a continuous function. Then

$$\langle g, \|\boldsymbol{\mathcal{F}} \mathbf{v}_{ht}\| \rangle = \int_{\Gamma_c} g \|\boldsymbol{\mathcal{F}} \mathbf{v}_{ht}\|\, ds \approx \sum_{r=1}^{m} \omega_r\, g(\mathbf{a}_r) \|\boldsymbol{\mathcal{F}}(\mathbf{a}_r) \mathbf{v}_{ht}(\mathbf{a}_r)\|, \qquad (3.3)$$

where $\omega_r \in \mathbb{R}^1$, $r = 1, \ldots, m$ are weights of the used cubature formula. To express (3.3) and the whole problem $(\mathcal{P}(g))_h$ in the algebraic form, the following notation will be used: by $\mathbf{N}$ we denote an $(m \times n)$ matrix representing the linear mapping $\mathbf{v}_h \mapsto (v_{h\nu}(\mathbf{a}_1), \ldots, v_{h\nu}(\mathbf{a}_m)) \in \mathbb{R}^m$, $\mathbf{v}_h \in \mathbb{V}_h$. Similarly, $\mathbf{T}_j$, $j = 1, 2$ are $(m \times n)$ matrices of the linear mappings $\mathbf{v}_h \mapsto (v_{ht_j}(\mathbf{a}_1), \ldots, v_{ht_j}(\mathbf{a}_m)) \in \mathbb{R}^m$,

$\mathbf{v}_h \in \mathbb{V}_h$, where $v_{ht_j}(\mathbf{a}_r) := (\mathbf{v}_h^1(\mathbf{a}_r) - \mathbf{v}_h^2(\mathbf{a}_r))^\top \mathbf{t}_j$. Let $\mathbf{T}_{jr}$ be the r-th row of $\mathbf{T}_j$. Then $\bar{\mathbf{v}}_t^r := (\mathbf{T}_{1r}\bar{\mathbf{v}}, \mathbf{T}_{2r}\bar{\mathbf{v}})^\top \in \mathbb{R}^2$ is the vector of the tangential displacements at the node $\mathbf{a}_r$. Finally set $\boldsymbol{\mathcal{F}}_r := \boldsymbol{\mathcal{F}}(\mathbf{a}_r)$, $g_r := g(\mathbf{a}_r)$, and $\bar{\mathbf{g}} = (g_1, \ldots, g_m)^\top$. Using this notation, $(\mathcal{P}(g))_h$ can be written as follows:

$$\left.\begin{aligned} &\text{Find } \bar{\mathbf{u}} \in \mathcal{K} \text{ such that} \\ &\mathcal{J}_{\bar{\mathbf{g}}}(\bar{\mathbf{u}}) \le \mathcal{J}_{\bar{\mathbf{g}}}(\bar{\mathbf{v}}) \quad \forall \bar{\mathbf{v}} \in \mathcal{K}, \end{aligned}\right\} \tag{$\mathcal{P}(\bar{\mathbf{g}})$}$$

where

$$\mathcal{J}_{\bar{\mathbf{g}}}(\bar{\mathbf{v}}) = \frac{1}{2}\bar{\mathbf{v}}^\top \mathbf{K}\bar{\mathbf{v}} - \bar{\mathbf{v}}^\top \bar{\mathbf{f}} + \sum_{r=1}^{m} \omega_r g_r \|\boldsymbol{\mathcal{F}}_r \bar{\mathbf{v}}_t^r\|$$

and

$$\mathcal{K} = \{\bar{\mathbf{v}} \in \mathbb{R}^n \,|\, \mathbf{N}\bar{\mathbf{v}} \le \mathbf{0}\}.$$

To release the constraints in $\mathcal{K}$ and to regularize the non-differentiable frictional term we use the duality approach. Let

$$\mathbf{X}(\bar{\mathbf{g}}) = \mathbb{R}^m_+ \times \mathbf{X}_t(\bar{\mathbf{g}})$$

be the set of the Lagrange multipliers, where

$$\mathbf{X}_t(\bar{\mathbf{g}}) = \{(\bar{\boldsymbol{\mu}}_{t_1}, \bar{\boldsymbol{\mu}}_{t_2}) \in \mathbb{R}^m \times \mathbb{R}^m \,|\, \|\boldsymbol{\mathcal{F}}_r^{-1} \bar{\boldsymbol{\mu}}_t^r\| \le \omega_r g_r,\ r = 1, \ldots, m\}$$

and $\bar{\boldsymbol{\mu}}_t^r = (\mu_{t_1 r}, \mu_{t_2 r})^\top \in \mathbb{R}^2$ is the vector made of the r-th components of $\bar{\boldsymbol{\mu}}_{t_1}$ and $\bar{\boldsymbol{\mu}}_{t_2}$. It is easy to verify that

$$\sum_{r=1}^{m} \omega_r g_r \|\boldsymbol{\mathcal{F}}_r \bar{\mathbf{v}}_t^r)\| = \max_{(\bar{\boldsymbol{\mu}}_{t_1}, \bar{\boldsymbol{\mu}}_{t_2}) \in \mathbf{X}_t(\bar{\mathbf{g}})} \sum_{r=1}^{m} (\bar{\boldsymbol{\mu}}_t^r)^\top \bar{\mathbf{v}}_t^r.$$

Thus

$$\min_{\bar{\mathbf{v}} \in \mathcal{K}} \mathcal{J}_{\bar{\mathbf{g}}}(\bar{\mathbf{v}}) = \min_{\bar{\mathbf{v}} \in \mathbb{R}^n} \sup_{\bar{\boldsymbol{\mu}} \in \mathbf{X}(\bar{\mathbf{g}})} \mathcal{L}(\bar{\mathbf{v}}, \bar{\boldsymbol{\mu}}),$$

where $\mathcal{L}(\bar{\mathbf{v}}, \bar{\boldsymbol{\mu}}) = \frac{1}{2}\bar{\mathbf{v}}^\top \mathbf{K}\bar{\mathbf{v}} - \bar{\mathbf{v}}^\top \bar{\mathbf{f}} + \bar{\boldsymbol{\mu}}^\top \mathbf{B}\bar{\mathbf{v}}$ is the Lagrangian, $\bar{\boldsymbol{\mu}} = (\bar{\boldsymbol{\mu}}_\nu^\top, \bar{\boldsymbol{\mu}}_{t_1}^\top, \bar{\boldsymbol{\mu}}_{t_2}^\top)^\top \in \mathbf{X}(\bar{\mathbf{g}})$, and $\mathbf{B} = (\mathbf{N}^\top, \mathbf{T}_1^\top, \mathbf{T}_2^\top)^\top$ is the $(3m \times n)$ matrix. Instead of $(\mathcal{P}(\bar{\mathbf{g}}))$ we shall use its *saddle-point formulation*:

$$\left.\begin{aligned} &\text{Find } (\bar{\mathbf{u}}, \bar{\boldsymbol{\lambda}}) \in \mathbb{R}^n \times \mathbf{X}(\bar{\mathbf{g}}) \text{ such that} \\ &\mathcal{L}(\bar{\mathbf{u}}, \bar{\boldsymbol{\mu}}) \le \mathcal{L}(\bar{\mathbf{u}}, \bar{\boldsymbol{\lambda}}) \le \mathcal{L}(\bar{\mathbf{v}}, \bar{\boldsymbol{\lambda}}) \quad \forall (\bar{\mathbf{v}}, \bar{\boldsymbol{\mu}}) \in \mathbb{R}^n \times \mathbf{X}(\bar{\mathbf{g}}), \end{aligned}\right\}$$

or, equivalently,

$$\left.\begin{aligned} &\text{Find } (\bar{\mathbf{u}}, \bar{\boldsymbol{\lambda}}) \in \mathbb{R}^n \times \mathbf{X}(\bar{\mathbf{g}}) \text{ satisfying} \\ &\mathbf{K}\bar{\mathbf{u}} = \bar{\mathbf{f}} - \mathbf{B}^\top \bar{\boldsymbol{\lambda}}, \\ &(\bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\lambda}})^\top \mathbf{B}\bar{\mathbf{u}} \le 0 \quad \forall \bar{\boldsymbol{\mu}} \in \mathbf{X}(\bar{\mathbf{g}}). \end{aligned}\right\} \qquad (\mathcal{M}(\bar{\mathbf{g}}))$$

One can easily show that $(\mathcal{M}(\bar{\mathbf{g}}))$ has a unique solution. Moreover its first component $\bar{\mathbf{u}}$ solves $(\mathcal{P}(\bar{\mathbf{g}}))$. Now we eliminate $\bar{\mathbf{u}}$ from the first equation: $\bar{\mathbf{u}} = \mathbf{K}^{-1}(\bar{\mathbf{f}} - \mathbf{B}^\top \bar{\boldsymbol{\lambda}})$ and substitute it into the second inequality. The resulting problem in terms of the Lagrange multipliers is equivalent to the following minimization problem:

$$\left.\begin{aligned} &\text{Find } \bar{\boldsymbol{\lambda}} \in \mathbf{X}(\bar{\mathbf{g}}) \text{ such that} \\ &\mathcal{S}(\bar{\boldsymbol{\lambda}}) \le \mathcal{S}(\bar{\boldsymbol{\mu}}) \quad \forall \bar{\boldsymbol{\mu}} \in \mathbf{X}(\bar{\mathbf{g}}), \end{aligned}\right\} \qquad (\mathcal{D}(\bar{\mathbf{g}}))$$

where $\mathcal{S}$ is the quadratic function with the symmetric, positive definite matrix $\mathbf{B}\mathbf{K}^{-1}\mathbf{B}^\top$ and the linear term $\bar{\mathbf{h}} = \mathbf{B}\mathbf{K}^{-1}\bar{\mathbf{f}}$. Notice that $(\mathcal{D}(\bar{\mathbf{g}}))$ has already the structure required by the algorithm KPRGP: the separated lower bounds for the components of $\bar{\boldsymbol{\lambda}}_\nu$ and the ellipsoidal constraints for the components of $(\bar{\boldsymbol{\lambda}}_{t_1}, \bar{\boldsymbol{\lambda}}_{t_2})$ as it follows from the definition of $\mathbf{X}(\bar{\mathbf{g}})$. Having $\bar{\boldsymbol{\lambda}}$ at our disposal we easily obtain $\bar{\mathbf{u}}$.

Remark 3.4. Model examples are solved by MatSol library [12] which uses the TFETI domain decomposition approach: each Ω^k, $k = 1, 2$ is divided into a finite number of subdomains involving "floating" blocks. To ensure continuity across subdomain interfaces and to satisfy the Dirichlet boundary conditions at the nodes of $\mathcal{T}_h$ on $\overline{\Gamma}_u$, the additional Lagrange multipliers are introduced. Then the resulting dual problem is given by the minimization of the quadratic function as in $(\mathcal{D}(\bar{\mathbf{g}}))$ but the set $\mathbf{X}(\bar{\mathbf{g}})$ contains, in addition, linear equality constraints. This fact requires an extension of the KPRGP algorithm called the SMALSE-M algorithm (for details see [7]). For realization of the problem with orthotropic Coulomb friction, we use an inexact implementation of the method of successive approximations (3.2) which performs only one iteration of the SMALSE-M in each step. In other words, the SAMLSE-M iterations and the successive approximations are performed by one outer loop.

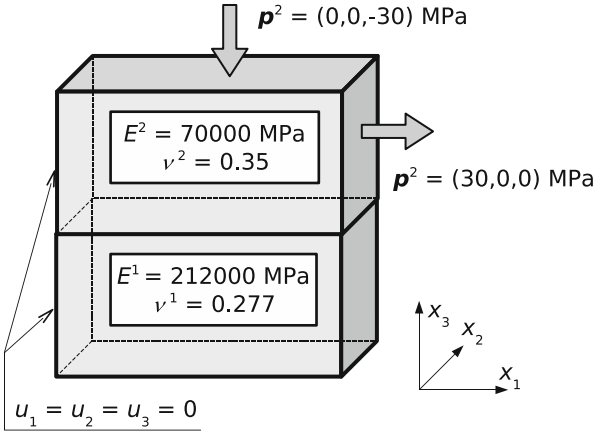


Fig. 1 Geometry of the problem

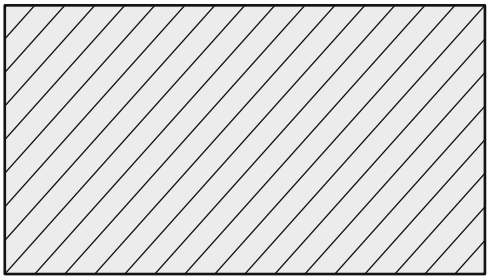


Fig. 2 Milled contact surface

4 Numerical Examples

Let us consider a 3D contact problem of two cantilever beams of sizes $2\times1\times1$ [m] in mutual contact with different coefficients of friction in two orthogonal directions to describe specially milled contact surface. The geometry, the prescribed boundary conditions, and material properties are specified in Fig. 1. The milled surface is depicted in Fig. 2. Finally, the volume forces are neglected and the coefficients of friction $\mathcal{F}_1$ and $\mathcal{F}_2$ on the contact interface are chosen in four different ways:

(a) Frictionless case: friction is neglected (Example 1);
(b) Isotropic case: $\mathcal{F}_1 = \mathcal{F}_2 = 0.3$ (Example 2);
(c) Orthotropic case: $\mathcal{F}_1 = 0.5$ in the direction $\mathbf{t}_1 = (1,0,0)^\top$ and $\mathcal{F}_2 = 0.1$ in $\mathbf{t}_2 = (0,1,0)^\top$ (Example 3);
(d) Orthotropic case: $\mathcal{F}_1 = 0.5$ in the direction $\mathbf{t}_1 = (\sqrt{2}/2, -\sqrt{2}/2, 0)^\top$ and $\mathcal{F}_2 = 0.1$ in $\mathbf{t}_2 = (\sqrt{2}/2, \sqrt{2}/2, 0)^\top$ (Example 4).

Case (d) corresponds to the real measurements, case (c) to incorrectly chosen tangential directions, case (b) to averaged coefficients which is a routinely used approach in engineering practise, and case (a) is for comparison purposes.

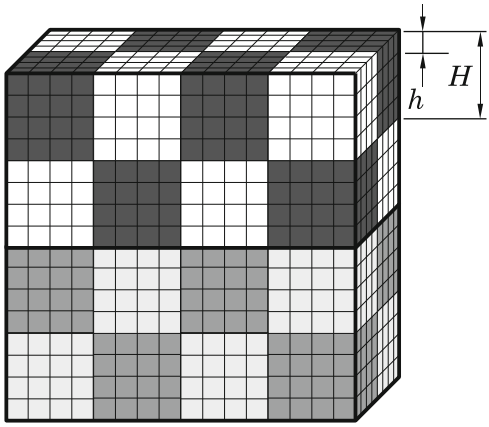


Fig. 3 Domain decomposition and the discretization

Each beam is divided into the same number of cubic subdomains with the decomposition step H and each subdomain is then decomposed into hexahedrons with the discretization step h; see Fig. 3. To demonstrate the behavior of our algorithms, we resolved the problem with varying discretizations and decompositions keeping $H/h = 10$.

The optimal choice of the parameters in the KPRGP is based on the analysis in [14] and on numerical experiments: we use $\Gamma = 1$, $\alpha \approx 2\|\mathbf{A}\|^{-1}$, adaptive values of ε depends on the precision achieved in the outer loop, and $\bar{\mathbf{x}}^{(0)}$ is determined by results from the previous outer iteration. The parameters of the SMALSE-M are chosen in agreement with [7]. The final relative stopping tolerance terminating the outer loop is 10^{-4} and the initial slip bound value in the discrete version of (3.2) is $\bar{\mathbf{g}}^{(0)} := \mathbf{0}$. The examples were computed by MatSol library [12] developed in Matlab environment and parallelized by Matlab Distributed Computing Server. For all computations we used 24CPUs of the HP Blade system, model BLc7000.

Example 1. We start with the frictionless case. The solution characteristics are summarized on the top of Table 1. We observe that the number of matrix–vector multiplications increases only moderately in agreement with the theory of [7]. The distribution of the normal contact stress Example 1 along the contact interface is depicted in Fig. 4.

Example 2. Let us consider the isotropic case *(b)*. This choice corresponds to the averaged friction coefficients of the real measurements for the surface from Fig. 2. The solution characteristics are summarized in the next part of Table 1. One can see that the number of outer iterations increases modestly with the size of the problem and the solution is more expensive compared with the previous example as follows from a higher number of the Hessian multiplications. The distribution of the normal contact stress along the contact interface is depicted in Fig. 5. In Figs. 6 and 7, we show the distributions of the Euclidean norm of the tangential contact stress and of displacements. The behavior of the contact stress inside the contact zone is seen in

Table 1 Solution characteristics for all examples

Number of subdomains	4	32	108	256
Primal variables	15,972	127,776	431,244	1,022,208
Dual variables	2,145	24,519	90,957	225,291
Equality constraints	24	192	648	1,536
Frictionless problem (Example 1)				
Bound constraints (active)	231 (11)	861 (15)	1,891 (20)	3,321 (35)
Outer iterations	11	11	9	9
Hessian multiplications	87	147	211	210
Isotropic case (Example 2)				
Bound constraints (active)	231 (11)	861 (64)	1,891 (135)	3,321 (246)
Circular constraints (active)	231 (220)	861 (830)	1,891 (1,847)	3,321 (3,270)
Outer iterations	11	15	19	22
Hessian multiplications	121	222	415	721
Orthotropic case, circular constraints (Example 3)				
Bound constraints (active)	231 (11)	861 (63)	1,891 (155)	3,321 (281)
Circular constraints (active)	231 (211)	861 (796)	1,891 (1,771)	3,321 (3,141)
Outer iterations	11	11	14	15
Hessian multiplications	149	262	487	665
Orthotropic case, ellipsoidal constraints (Example 3)				
Bound constraints (active)	231 (11)	861 (63)	1,891 (156)	3,321 (284)
Ellipsoidal constraints (active)	231 (213)	861 (798)	1,891 (1,784)	3,321 (3,151)
Outer iterations	11	10	16	17
Hessian multiplications	121	221	363	469
Orthotropic case, circular constraints (Example 4)				
Bound constraints (active)	231 (13)	861 (47)	1,891 (113)	3,321 (203)
Circular constraints (active)	231 (229)	861 (847)	1,891 (1,869)	3,321 (3,301)
Outer iterations	10	12	11	12
Hessian multiplications	126	315	332	344
Orthotropic case, ellipsoidal constraints (Example 4)				
Bound constraints (active)	231 (8)	861 (41)	1,891 (102)	3,321 (188)
Ellipsoidal constraints (active)	231 (220)	861 (846)	1,891 (1,860)	3,321 (3,301)
Outer iterations	15	24	28	29
Hessian multiplications	208	376	617	738

Fig. 8. The radiuses of small circles are given by the slip bound values $\mathcal{F}_1\lambda_{\nu,i}$, where $\mathcal{F}_1 = \mathcal{F}_2 = 0.3$ and $\lambda_{\nu i}$ is the component of $\bar{\boldsymbol{\lambda}}_\nu$ at the i-th contact node. The arrows in the circles represent the tangential contact stress.

Example 3. In this example we consider the orthotropic case *(c)* with the coefficients of friction $\mathcal{F}_1 = 0.5$ and $\mathcal{F}_2 = 0.1$ in the incorrectly chosen tangential directions $\mathbf{t}_1 = (1, 0, 0)^\top$ and $\mathbf{t}_2 = (0, 1, 0)^\top$, respectively. The results in Table 1 show that the computations with the circular constraints are more expensive than

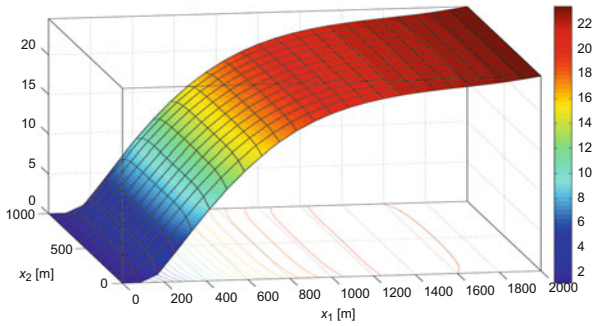


Fig. 4 Normal contact stress

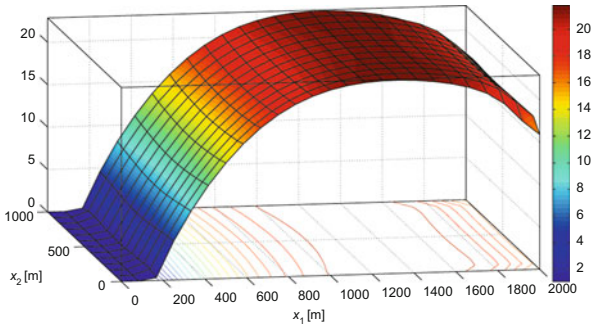


Fig. 5 Normal contact stress

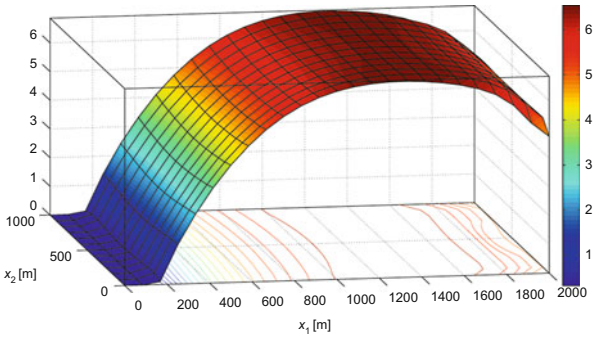


Fig. 6 The norm of the tangential contact stress

the ones with the original ellipsoidal constraints. This may be due to worse spectral properties of the Hessian matrix which increase the bound on the number of iterations; see Remark 2.1. In Figs. 9, 10, and 11, we depict the distributions of the normal contact stress and the standard and scaled Euclidean norms of the tangential

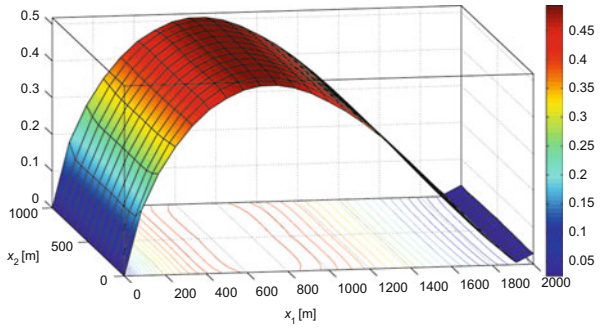


Fig. 7 The norm of the relative tangential contact displacements

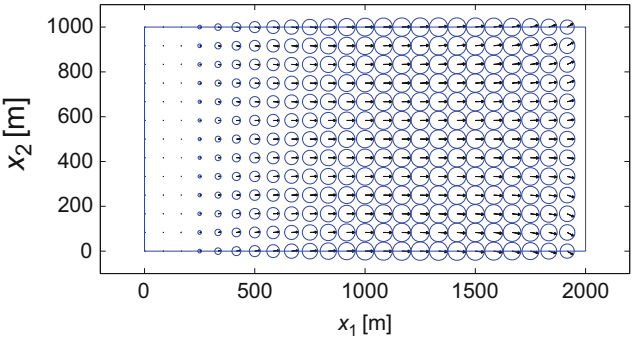


Fig. 8 Contact zone

contact stresses, respectively. The value of the scaled norm at the i-th contact node is defined as $\|\mathcal{F}_i^{-1}\bar{\boldsymbol{\lambda}}_t^i\|$, where $\bar{\boldsymbol{\lambda}}_t^i = (\lambda_{t_1 i}, \lambda_{t_2 i})^\top \in \mathbb{R}^2$ is the vector made of the i-th components of $\bar{\boldsymbol{\lambda}}_{t_1}$ and $\bar{\boldsymbol{\lambda}}_{t_2}$. The Euclidean norm of the relative tangential contact displacements is seen in Fig. 12. Finally, Figs. 13 and 14 show the behavior inside the contact zone. The semi-axes of ellipses are oriented by the directions $\mathbf{t}_1$ and $\mathbf{t}_2$ and their lengths are $\mathcal{F}_1\lambda_{vi}$ and $\mathcal{F}_2\lambda_{vi}$, respectively. Again, λ_{vi} are the components of $\bar{\boldsymbol{\lambda}}_v$ and the arrows in the ellipses represent the tangential contact stress.

Example 4. Finally, let us consider the orthotropic case *(d)* with the coefficients of friction $\mathcal{F}_1 = 0.5$ and $\mathcal{F}_2 = 0.1$ in the directions $\mathbf{t}_1 = (\sqrt{2}/2, -\sqrt{2}/2, 0)^\top$ and $\mathbf{t}_2 = (\sqrt{2}/2, \sqrt{2}/2, 0)^\top$, respectively. This setting corresponds to the milled surface depicted in Fig. 2. From Table 1 we can see that the circular constraints require less computations than the ellipsoidal ones. A heuristic argument explaining this fact is that the spectral properties of the new matrix after transformation of the ellipsoidal

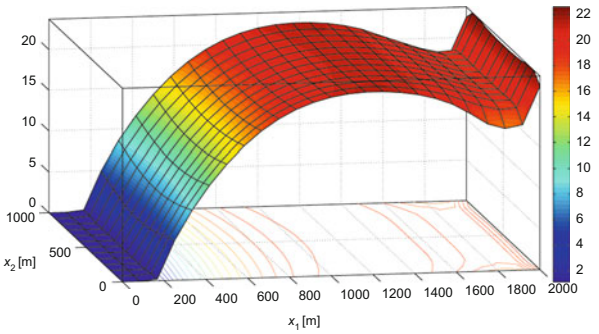


Fig. 9 Normal contact stress

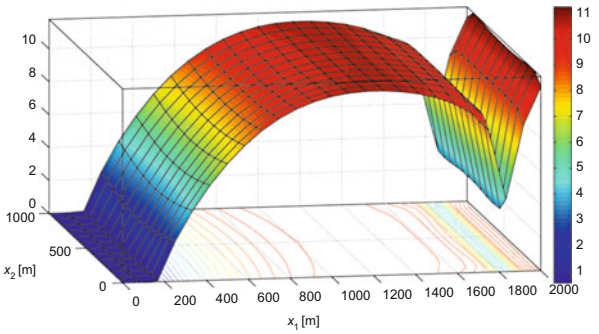


Fig. 10 The norm of the tangential contact stress

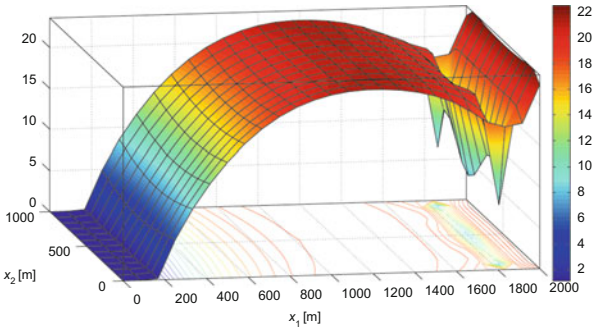


Fig. 11 The scaled norm of the tangential contact stress

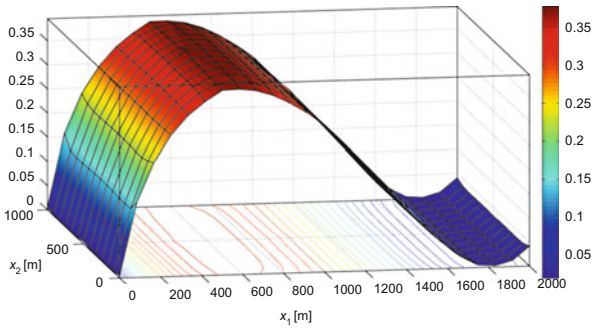


Fig. 12 The norm of the relative tangential contact displacements

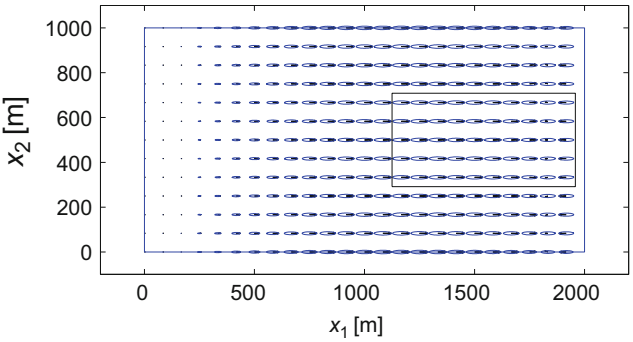


Fig. 13 Contact zone

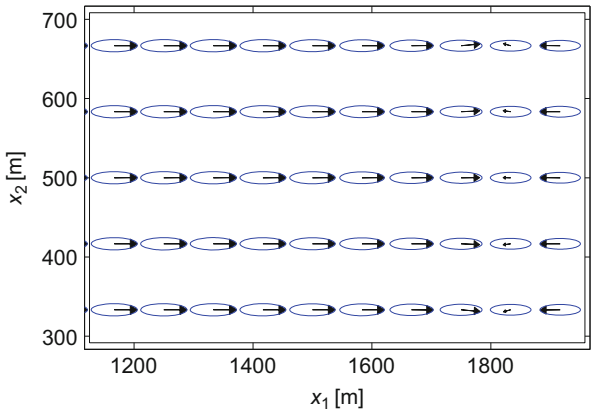


Fig. 14 Contact zone zoom

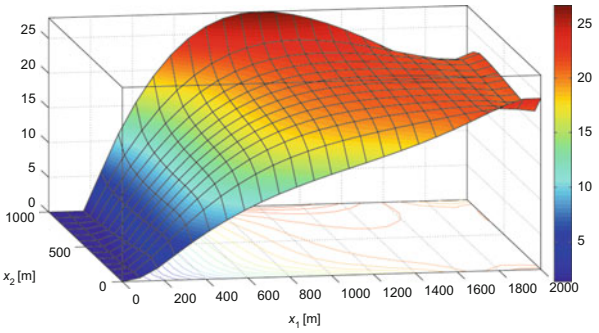


Fig. 15 Normal contact stress

constraints into the circular ones are sensitive to the orientation of the ellipses. In Figs. 15, 16, 17, 18, 19, and 20 we depict the same characteristics of the solution as in Example 3.

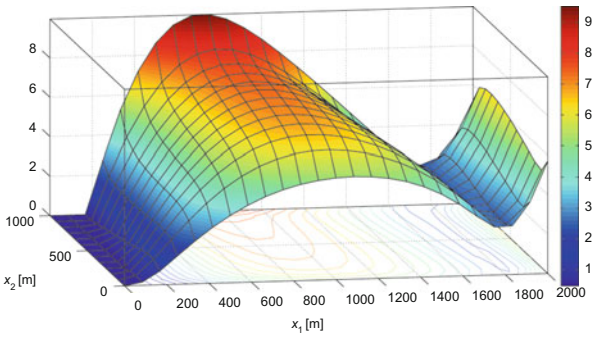


Fig. 16 The norm of the tangential contact stress

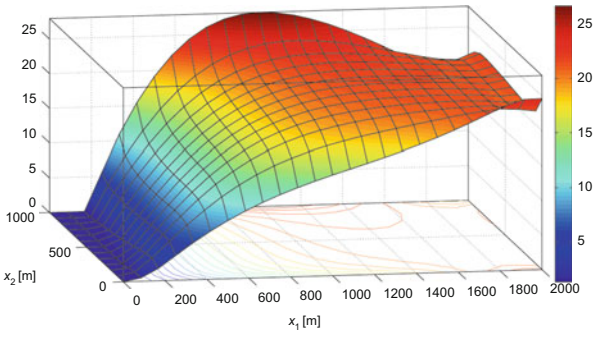


Fig. 17 The scaled norm of the tangential contact stress

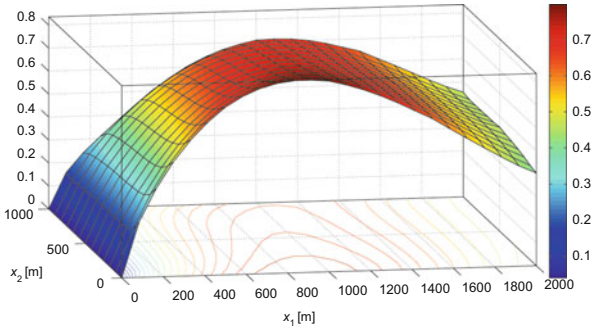


Fig. 18 The norm of the relative tangential contact displacements

Table 2 compares the computed displacements for different friction models. One can see the significant dependence of the results on the used friction model. Using the orthotropic friction law with correctly chosen tangential directions we get the results which are closer to the reality. An industrial application for the isotropic case may be found in [6].

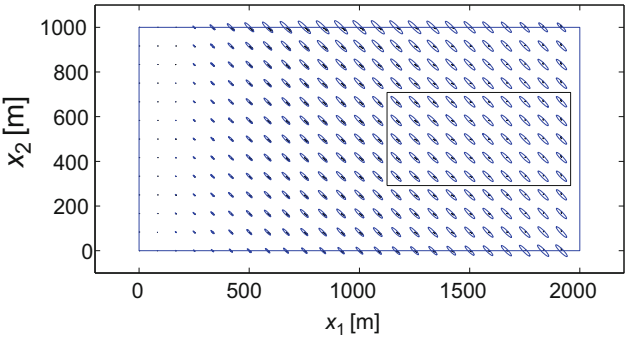


Fig. 19 Contact zone

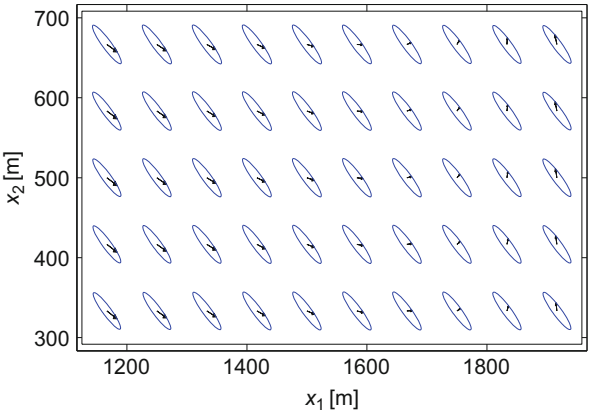


Fig. 20 Contact zone zoom

Table 2 Characteristic of the displacements for different friction models in the whole configuration

Friction model	$\max(\lvert u_1\rvert)$	$\max(\lvert u_2\rvert)$	$\max(\lvert u_3\rvert)$	$\max(\lVert \mathbf{u}\rVert)$
Case (*a*)	1.93765	0.13952	3.64374	4.12538
Case (*b*)	1.83178	0.12977	3.02874	3.53673
Case (*c*)	1.80919	0.12659	2.85510	3.37758
Case (*d*)	1.89417	0.30280	3.29204	3.80103

5 Conclusions

The paper deals with the KPRGP algorithm [14] for constrained minimization of strictly convex quadratic functions subject to simple bounds and separable ellipsoidal constraints. Since the algorithm uses the reduced gradient defined by the projection on the feasible set, the implementation requires to compute the

projections on ellipses. These projections are computed by a combination of the Newton and bisection methods.

Our study is motivated by the numerical solution of contact problems in linear elasticity with orthotropic Coulomb friction. The presented approach uses the method of successive approximations that requires to solve auxiliary contact problems with orthotropic Tresca friction in each iterative step. The algebraic dual formulation of the Tresca problem leads to the constrained minimization for which the KPRGP may be used. As an alternative to KPRGP one can use MPGP algorithm described in [5]. In order to increase the computational efficiency, we apply the finite element discretization based on the TFETI domain decomposition method. Since the TFETI introduces additional equality constraints in the algebraic problem, we apply the SMALSE-M algorithm [7] in which the KPRGP is included as the inner loop. The outer loop of the SMALSE-M is based on the augmented Lagrangian method. The important property of the SMALSE-M is the fact that the number of iterations needed to get a solution with a given accuracy is uniformly bounded (with respect to the size of the problem) provided that the spectrum of the Hessian is confined in a given interval (i.e., the algorithm is scalable). This assumption is satisfied, if the ratio between the maximal diameter of the subdomains H and the norm of the finite element partitions h is fixed and the Hessian is normalized in the spectral norm [6]. Let us recall that the scalability can be proven only for the frictionless case and Tresca friction but we observed it experimentally also for some examples with Coulomb fiction.

Acknowledgements The work was supported by the ESF OPTPDE Research Programme. The first author acknowledges also the support of the grant GAČR P201/12/0671. The second and the third author acknowledge the support of the European Regional Development Fund in the IT4 Innovations Centre of Excellence project (CZ.1.05/1.1. 00/02.0070).

References

[1] A. Ben-Tal, A. Nemirovski, *Lectures on Modern Convex Optimization* (SIAM, Philadelphia, 2001)
[2] S. Boyd, L. Vandenberghe, *Convex Optimization* (Cambridge University Press, Cambridge, 2004)
[3] A. Conn, N. Gould, P. Toint, *Trust Region Methods* (SIAM, Philadelphia, 2000)
[4] Z. Dostál, *Optimal Quadratic Programming Algorithms: With Applications to Variational Inequalities* (Springer, Berlin/Heidelberg/New York, 2009)
[5] Z. Dostál, T. Kozubek, An optimal algorithm and superrelaxation for minimization of a quadratic function subject to separable convex constraints with applications. Math. Programming **135**, 195–220 (2012)
[6] Z. Dostál, T. Kozubek, A. Markopoulos, T. Brzobohatý, V. Vondrák, P. Horyl, Theoretically supported scalable TFETI algorithm for the solution of multibody 3D contact problems with friction. Comput. Methods Appl. Mech. Eng. **205**, 110–120 (2012)
[7] Z. Dostál, R. Kučera, An optimal algorithm for minimization of quadratic functions with bounded spectrum subject to separable convex inequality and linear equaity constraints. SIAM J. Opt. **20**, 2913–2938, 2010.

[8] G.H. Golub, C. F. Van Loan, *Matrix Computation* (The Johns Hopkins University Press, Baltimore, 1996)
[9] J. Haslinger, Approximation of the Signorini problem with friction, obeying Coulomb law. Math. Meth. Appl. **5**, 422–437 (1983)
[10] J. Haslinger, I. Hlaváček, J. Nečas, Numerical methods for unilateral problems in solid mechanics, in *Handbook of Numerical Analysis*, vol. IV, ed. by P.G. Ciarlet, J.-L. Lions (North-Holland, Amsterdam, 1996), 313–485
[11] J. Haslinger, R. Kučera, T. Ligurský, Qualitative analysis of 3D elastostatic contact problems with orthotropic Coulomb friction and a solution dependent coefficients of friction. J. Comput. Appl. Math. **235**, 3464–3480 (2011)
[12] T. Kozubek, A. Markopoulos, T. Brzobohatý, R. Kučera, V. Vondrák, Z. Dostál, *MatSol - MATLAB Efficient Solvers for Problems in Engineering*. http://industry.it4i.cz/en/products/matsol
[13] R. Kučera, Minimizing quadratic functions with separable quadratic constraints. Optim. Meth. Soft. **22**, 453–467 (2007)
[14] R. Kučera, Convergence rate of an optimization algorithm for minimizing quadratic functions with separable convex constraints. SIAM J. Optim. **19**, 846–862 (2008)
[15] J. Nečas, I. Hlaváček, *Mathematical Theory of Elastic and Elasto-Plastic Bodies: An Introduction*. Studies in applied mechanics, vol. 3 (Elsevier Scientific Publishing Co., Amsterdam/New York, 1980)
[16] J.B. Rosen, The gradient projection method for nonlinear programming. Part II. Nonlinear constraints. J. Soc. Indust. Appl. Math. **9**, 515–532 (1961)
[17] S.J. Wright, *Primal-Dual Interior-Point Methods* (SIAM, Philadelphia, 1997)
[18] H.W. Zhang, A.H. Liao, Z.Q. Xie, B.S. Chen, H. Wang, Some advances in mathematical programming methods for numerical simulation of contact problems, in *Proceedings of the IUTAM Symposium on Computational Methods in Contact Mechanics*, ed. by P. Wriggers, U. Nackenhorst (Springer, Berlin/Heidelberg/New York, 2007), 33–56

Shape-Topological Differentiability of Energy Functionals for Unilateral Problems in Domains with Cracks and Applications

Günter Leugering, Jan Sokołowski, and Antoni Żochowski

Abstract A review of results on first order shape-topological differentiability of energy functionals for a class of variational inequalities of elliptic type is presented.

The *velocity method in shape sensitivity analysis* for solutions of elliptic unilateral problems is established in the monograph (Sokołowski and Zolésio, Introduction to Shape Optimization: Shape Sensitivity Analysis, Springer, Berlin/Heidelberg/New York, 1992). The *shape and material derivatives* of solutions to frictionless contact problems in solid mechanics are obtained. In this way the *shape gradients* of the associated integral functionals are derived within the framework of nonsmooth analysis. In the case of the energy type functionals classical differentiability results can be obtained, because the shape differentiability of solutions is not required to obtain the shape gradient of the shape functional (Sokołowski and Zolésio, Introduction to Shape Optimization: Shape Sensitivity Analysis, Springer, Berlin/Heidelberg/New York, 1992). Therefore, for cracks the strong continuity of solutions with respect to boundary variations is sufficient in order to obtain first order shape differentiability of the associated energy functional. This simple observation which is used in Sokołowski and Zolésio (Introduction to Shape Optimization: Shape Sensitivity Analysis, Springer, Berlin/Heidelberg/New York, 1992) for the shape differentiability of multiple

G. Leugering
Department Mathematik, Friedrich-Alexander-Universität Erlangen-Nürnberg, Cauerstr. 11 (03.322) 91058 Erlangen, Germany
e-mail: leugering@math.fau.de

J. Sokołowski (✉)
Laboratoire de Mathématiques, Institut Élie Cartan, UMR 7502
Nancy-Université-CNRS-INRIA, Université Henri Poincaré Nancy 1, B.P. 239, 54506 Vandoeuvre Lès Nancy Cedex, France

Systems Research Institute of the Polish Academy of Sciences, ul. Newelska 6, 01-447 Warszawa, Poland
e-mail: Jan.Sokolowski@iecn.u-nancy.fr

A. Zochowski
Systems Research Institute of the Polish Academy of Sciences, ul. Newelska 6, 01-447 Warszawa, Poland
e-mail: Antoni.Zochowski@ibspan.waw.pl

R. Hoppe (ed.), *Optimization with PDE Constraints*, Lecture Notes in Computational Science and Engineering 101,
DOI 10.1007/978-3-319-08025-3_8

eigenvalues is further applied in Khludnev and Sokołowski (Eur. J. Appl. Math. 10:379–394, 1999; Eur. J. Mech. A Solids 19:105–120, 2000) to derive the first order shape gradient of the energy functional with respect to perturbations of the crack tip. A domain decomposition technique in shape-topology sensitivity analysis for problems with unilateral constraints on the crack faces (lips) is presented for the shape functionals.

We introduce the *Griffith shape functional* as the distributed shape derivative of the elastic energy evaluated in a domain with a crack, with respect to the crack length. We are interested in the dependence of this functional on domain perturbations far from the crack. As a result, the directional shape and topological derivatives of the nonsmooth Griffith shape functional are obtained with respect to boundary variations of an inclusion.

Keywords Conical differential of metric projection • Dirichlet Sobolev space • Griffith criterium for crack propagation • Hadamard shape differentiability • Nonsmooth analysis • Shape gradient • Shape Hessian • Signorini variational inequality

Mathematics Subject Classification (2010). Primary 35J86; Secondary 35R35, 49J40, 74R99.

1 Introduction

First order shape sensitivity analysis of the energy functional for an elliptic boundary value problem with unilateral constraints defined in domains with cracks is of broad interest and, therefore, it is named *Griffith shape functional*. In order to introduce the Griffith shape functional we make use of

- the crack model within an elastic body, represented by an elliptic variational inequality with the unilateral constraints representing the first order linear approximation of the non-penetration condition;
- the energy shape functional defined for the solutions of the variational inequality depending on the *shape of the crack*;
- an abstract result on the directional differentiability of the optimal value for constrained optimization problems over convex sets with respect to a parameter $t \to 0$,

$$t \to j(t, v^*(t)) := \inf_{v \in K} j(t, v)$$

  which requires only the strong convergence of the minimizers $v^*(t) \to v^*(0)$ with respect to the parameter as well as the existence of the partial derivative of the mapping $\mathbb{R} \ni t \to j(t, v) \in \mathbb{R}$;
- a technical result on linear transformations of the displacement field in the elasticity model obtained in [25] which provides the convex cone K, invariant

under the change of variables of the velocity method; it means that in order to apply the abstract sensitivity result for optimal values, we have in hand the linear transformation of the unknown solution to the variational inequality such that we could analyze the variational inequality transformed to the fixed geometrical domain with the parameter independent convex cone K.

Therefore, the Griffith shape functional is the first order shape derivative of the elastic energy with respect to the perturbation of the crack tip for a given direction of the velocity vector field. In addition, the second order shape derivative of the energy functional, whenever it does exist, becomes the first order shape derivative of the Griffith shape functional. But it is not our primary concern, since we are more interested in the influence of elastic inclusions far from the crack on the behaviour of the Griffith shape functional. We believe that such an influence is possible and can be used for the control of crack propagation in elastic media. Indeed, the dependence of the Griffith functional with respect to shape changes of an elastic or rigid inclusion has been considered in [8,17]. This research has been triggered by numerical studies on optimization an control of crack growth also for the case of cohesive crack theories in [18,21,22]. See also [7,19].

We recall also that the second order shape differentiability of the energy functional with respect to the perturbations of the crack tip is known for the Signorini type variational inequalities which governs frictionless contact problems [6]. This result can be extended to the crack problems with non-penetration contact conditions on the crack faces (lips), but this is a subject of the forthcoming paper.

1.1 Interface Problems in Lipschitz Domains

In this paper a class of models with defects in solids is introduced. The defect takes the form of a cut in the geometrical domain. The cut is a part of a curve in two spatial dimensions, and the unilateral boundary conditions for displacements and the tractions are prescribed for the jumps from both sides of the cut. The variational formulation of the model include the unilateral conditions for the displacements imposed in the convex cone constraints for admissible displacements. The variational inequality for displacements is obtained for the minimization problem of the energy functional over a convex cone. In the specific case of our setting, the solution operator is Lipschitz continuous with respect to the right-hand side of the variational inequality. This property leads usually to the Lipschitz continuity of the solution with respect to the regular boundary variations in the framework of the velocity method of shape sensitivity analysis. On the other hand, the asymptotic analysis of solutions to singular perturbations of the geometrical domain can be performed for linear problems or a restricted class of nonlinear problems. Since the technique of compound asymptotic expansions cannot be directly applied to the variational inequalities under considerations, a domain decomposition technique is used in order to obtain the first order asymptotic expansion of the energy functional and to obtain the topological derivatives of the energy functionals for the variational inequalities.

In this section the framework is introduced for the crack problem in the bounded domain Ω in two spatial dimensions. It is assumed [8–17] that a crack in Ω is a part Σ_l of the Lipschitz interface Σ. By an interface we mean a Lipschitz, closed curve without intersections $\Sigma \Subset \Omega$ such that the jumps $[u]$ of values for traces of Sobolev functions u from both sides of the interface are allowed.

In addition, in our model the interface, thus, also the crack are supposed to be sufficiently smooth, say Σ is a $C^{1,1}$ closed curves without intersections. This regularity assumption is added in order to use the standard properties of traces of Sobolev functions on the interface.

However, the shape sensitivity analysis is performed in our framework by the bi-Lipschitz changes of variables, we refer to [25] for all details necessary for such a construction.

Let us consider the Lipschitz domain Ω with the boundary $\Gamma = \partial\Omega$ decomposed into two Lipschitz subdomains Ω', Ω'' and the interface $\Sigma \subset \Omega$, i.e., $\Omega := \Omega' \cup \Sigma \cup \Omega''$. For the decomposition of functions in $v \in H^1_0(\Omega)$, we use the notation for restrictions to subdomains $v' \in H^1_0(\Omega')$ and $v'' \in H^1_0(\Omega'')$. Thus, the traces on Σ are well defined

$$v|_\Sigma := v'|_\Sigma = v''|_\Sigma \in H^{1/2}(\Sigma)\,.$$

Now, we define a broader space $H^1_0(\Omega) \subset H^1_\Gamma(\Omega_\Sigma) \subset L^2(\Omega)$ of functions which admit the jump

$$[v] := v'|_\Sigma - v''|_\Sigma \in H^{1/2}(\Sigma)$$

over the interface Σ. This leads also to the boundary value problems in Ω with the prescribed jump over the interface, which is not our primary interest. We are interested in the cracks $\Sigma_l \subset \Sigma$ modeled by closed subsets of the interface, with $\Omega_l := \Omega \setminus \overline{\Sigma}_l$, thus, in solutions of the boundary value problems in the convex set

$$K(\Omega_l) := \{v \in H^1_\Gamma(\Omega_\Sigma) \,:\, [v] \geqslant 0 \quad \text{on } \Sigma_l\,, \quad [v] = 0 \quad \text{on } \Sigma \setminus \overline{\Sigma}_l\}\,.$$

The primary interest of such a function space setting for the crack problems with unilateral non-penetration conditions on the crack faces (lips) is the so-called polyhedricity of the set $K(\Omega_l)$. In other words, polyhedral convex sets admit the Hadamard differential of the metric projection [6, 25]. This property is inherited from the polyhedricity of the positive cone in the fractional Sobolev space $H^{1/2}(\Sigma)$, since the space $H^{1/2}(\Sigma)$ is the so-called Dirichlet space with respect to the natural order. Let us recall the known facts [6].

Proposition 1.1. *The scalar product* $(\cdot,\cdot)_\Sigma$ *in the Dirichlet space* $H^{1/2}(\Sigma)$ *satisfies the condition*

$$(v^+, v^-)_\Sigma \leqslant 0 \quad \forall v \in H^{1/2}(\Sigma),$$

therefore, the metric projection in $H^{1/2}(\Sigma)$ *onto the positive cone of* $H^{1/2}(\Sigma)$ *is conically differentiable.*

This implies

Corollary 1.2. *The metric projection in* $H^1_\Gamma(\Omega_\Sigma)$ *onto the closed, convex cone* $K(\Omega_l)$ *is conically differentiable.*

The above results lead to the first order shape derivatives of the Griffith shape functional for the cracks with the nonlinear non-penetration conditions prescribed on the crack lips (or faces in three spatial dimensions).

Remark 1.3. The Griffith shape functional of the crack $\Sigma_l := \{(x_1, 0) \in \mathbb{R}^2, \quad 0 < x_1 < l\}$ at the tip $P_l := (l, 0)$ is defined by the shape derivative which is denoted by

$$J(\Omega_l) := \frac{d\,\Pi(\Omega_l; u_l)}{d\,l}$$

of the energy functional

$$l \to \Pi(\Omega_l; u_l) = \inf_{v \in K(\Omega_l)} \int_{\Omega_l} \left(\frac{1}{2} |\nabla v|^2 - f\, v \right)$$

where $u_l \in K(\Omega_l)$ is the minimizer for a given length $l > 0$ of the crack, and $f \in L^2(\Omega)$ is a given element.

We are going to extend such results to elastic bodies Ω_l with cracks Σ_l and unilateral conditions on the crack lips (faces) $\Sigma_l^{\pm}$. Then, we consider the differentiability properties of the Griffith functional

- evaluation of the first order shape derivative with respect to the perturbations of the crack;
- asymptotic analysis of the Griffith functional with respect to singular perturbations of the geometrical domain far from the crack;

2 Modeling of Cracks in Elastic Bodies

2.1 Non-Penetration Conditions on the Crack Faces

It is well known that classical crack theory in elasticity is characterized by linear boundary conditions which leads to linear boundary value problems. This approach has a clear shortcoming from a mechanical standpoint, since opposite crack faces can penetrate each other. We consider nonlinear boundary conditions on crack faces, the so-called non-penetration conditions, written in terms of inequalities. From the standpoint of applications, these boundary conditions are preferable since they provide a mutual non-penetration between crack faces. As a result, a free boundary

problem is obtained which means that a concrete boundary condition at a given point can be found provided that we have a solution of the problem.

The main attention in this paper is paid to dependence of solutions of the problem on domain perturbations, and in particular, on the crack shape.

Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with smooth boundary Γ, and $\Gamma_c \subset \Omega$ be a smooth curve without self-intersections, $\Omega_c = \Omega \setminus \overline{\Gamma}_c$.

It is assumed that Γ_c can be extended in such a way that this extension crosses Γ at two points, and Ω_c is divided into two subdomains D_1 and D_2 with Lipschitz boundaries ∂D_1, ∂D_2, $meas(\Gamma \cap \partial D_i) > 0$, $i = 1, 2$. Denote by $\nu = (\nu_1, \nu_2)$ a unit normal vector to Γ_c. We assume that Γ_c does not contain its tip points, i.e. $\Gamma_c = \overline{\Gamma}_c \setminus \partial\Gamma_c$.

The equilibrium problem for a linear elastic body occupying Ω_c is as follows. In the domain Ω_c we have to find a displacement field $u = (u_1, u_2)$ and stress tensor components $\sigma = \{\sigma_{ij}\}$, $i, j = 1, 2$, such that

$$-\mathrm{div}\sigma = f \quad \text{in} \quad \Omega_c, \tag{1}$$

$$\sigma = A\varepsilon(u) \quad \text{in} \quad \Omega_c, \tag{2}$$

$$u = 0 \quad \text{on} \quad \Gamma, \tag{3}$$

$$[u]\nu \geqslant 0, \quad [\sigma_\nu] = 0, \quad \sigma_\nu \cdot [u]\nu = 0 \quad \text{on} \quad \Gamma_c, \tag{4}$$

$$\sigma_\nu \leqslant 0, \quad \sigma_\tau = 0 \quad \text{on} \quad \Gamma_c^{\pm}. \tag{5}$$

Here $[v] = v^+ - v^-$ is a jump of v on Γ_c, and signs $\pm$ correspond to positive and negative crack faces with respect to ν, $f = (f_1, f_2) \in L^2(\Omega_c)$ is a given function,

$$\sigma_\nu = \sigma_{ij}\nu_j\nu_i, \quad \sigma_\tau = \sigma\nu - \sigma_\nu \cdot \nu, \quad \sigma_\tau = (\sigma_\tau^1, \sigma_\tau^2),$$
$$\sigma\nu = (\sigma_{1j}\nu_j, \sigma_{2j}\nu_j),$$

the strain tensor components are denoted by $\varepsilon_{ij}(u)$,

$$\varepsilon_{ij}(u) = \frac{1}{2}(u_{i,j} + u_{j,i}), \quad \varepsilon(u) = \{\varepsilon_{ij}(u)\}, \quad i, j = 1, 2.$$

Elasticity tensor $A = \{a_{ijkl}\}$, $i, j, k, l = 1, 2$, is given and satisfies the usual properties of symmetry and positive definiteness

$$a_{ijkl}\xi_{kl}\xi_{ij} \geqslant c_0|\xi|^2, \quad \forall\ \xi_{ij},\ \xi_{ij} = \xi_{ji}, \quad c_0 = const,$$

$a_{ijkl} = a_{klij} = a_{jikl}$, $a_{ijkl} \in L^\infty(\Omega)$.

Relations (1) are equilibrium equations, and (2) is Hooke's law, $u_{i,j} = \frac{\partial u_i}{\partial u_j}$, $(x_1, x_2) \in \Omega_c$. All functions with two below indices are symmetric in those indices, i.e. $\sigma_{ij} = \sigma_{ji}$ etc. Summation convention is assumed over repeated indices throughout the paper.

The first condition in (4) is called the non-penetration condition. It provides a mutual non-penetration between the crack faces $\Gamma_c^{\pm}$. The second condition of (5) provides zero friction on Γ_c. For simplicity we assume a clamping condition (3) at the external boundary Γ.

Note that a priori we do not know points on Γ_c where strict inequalities in (4), (5) are fulfilled. Due to this, the problem (1)–(5) is a free boundary value problem. If we have $\sigma_\nu = 0$ then, together with $\sigma_\tau = 0$, the classical boundary condition $\sigma\nu = 0$ follows which is used in linear crack theory. On the other hand, due to (4), the condition $\sigma_\nu < 0$ implies $[u]\nu = 0$, i.e. we have a contact between the crack faces at a given point. The strict inequality $[u]\nu > 0$ at a given point means that we have no contact between the crack faces.

Hence, the first difficulty in studying the problem (1)–(5) is concerned with boundary conditions (4)–(5). The second one is related to the general crack problem difficulty—a presence of nonsmooth boundaries. We refer the reader to [6] for related results on boundary value problems defined in domains with cracks.

2.2 *Existence of Solutions*

First of all we note that problem (1)–(5) admits several equivalent formulations. In particular, it corresponds to the minimization of the energy functional. To check this, introduce the Sobolev space

$$H^1_\Gamma(\Omega_c) = \{v = (v_1, v_2) \mid v_i \in H^1(\Omega_c),\ v_i = 0 \text{ on } \Gamma,\ i = 1, 2\}$$

and the closed convex set of admissible displacements

$$K = \{v \in H^1_\Gamma(\Omega_c) \mid [v]\nu \geqslant 0 \quad \text{a.e. on} \quad \Gamma_c\}. \tag{6}$$

In this case, due to the Weierstrass theorem, the problem

$$\min_{v \in K} \left\{ \frac{1}{2} \int_{\Omega_c} \sigma_{ij}(v)\varepsilon_{ij}(v) - \int_{\Omega_c} f_i v_i \right\}$$

has (a unique) solution u satisfying the variational inequality

$$u \in K, \tag{7}$$

$$\int_{\Omega_c} \sigma_{ij}(u)\varepsilon_{ij}(v - u) \geqslant \int_{\Omega_c} f_i(v_i - u_i), \quad \forall v \in K, \tag{8}$$

where $\sigma_{ij}(u) = \sigma_{ij}$ are defined from (2).

Problem formulations (1)–(5) and (7)–(8) are equivalent. We shall use in Sect. 47 the abstract form (144) of the variational inequality (7)–(8).

Remark 2.1. It follows from the coercivity on the energy space $H^1_\Gamma(\Omega_c)$ of the symmetric bilinear form

$$H^1_\Gamma(\Omega_c) \times H^1_\Gamma(\Omega_c) \ni (u, v) \to a(u, v) := \int\limits_{\Omega_c} \sigma_{ij}(u)\varepsilon_{ij}(v) \in \mathbb{R}$$

that the solution u to (7)–(8) is Lipschitz continuous in the energy space with respect to the right-hand side f in the dual space $\left(H^1_\Gamma(\Omega_c)\right)^*$.

Any smooth solution of (1)–(5) satisfies (7)–(8) and, conversely, from (7)–(8) it follows (1)–(5).

Below we provide two more equivalent formulations for the problem (1)–(5), the so-called mixed and smooth domain formulations. To this end, we first discuss in what sense boundary conditions (4)–(5) are fulfilled. Denote by Σ a closed curve without self-intersections of the class $C^{1,1}$, which is an extension of Γ_c such that $\Sigma \subset \Omega$, and the domain Ω is divided into two subdomains Ω_1 and Ω_2. In this case Σ is the boundary of the domain Ω_1, and the boundary of Ω_2 is $\Sigma \cup \Gamma$.

Introduce the space $H^{\frac{1}{2}}(\Sigma)$ with the norm

$$\|v\|^2_{H^{\frac{1}{2}}(\Sigma)} = \|v\|^2_{L^2(\Sigma)} + \int\limits_\Sigma \int\limits_\Sigma \frac{|v(x) - v(y)|^2}{|x - y|^2} dxdy \tag{9}$$

and denote by $H^{-\frac{1}{2}}(\Sigma)$ a space dual of $H^{\frac{1}{2}}(\Sigma)$. Also, consider the space

$$H^{1/2}_{00}(\Gamma_c) = \left\{ v \in H^{\frac{1}{2}}(\Gamma_c) \mid \frac{v}{\sqrt{\rho}} \in L^2(\Gamma_c) \right\}$$

with the norm

$$\|v\|^2_{1/2,00} = \|v\|^2_{1/2} + \int\limits_{\Gamma_c} \rho^{-1} v^2,$$

where $\rho(x) = dist(x; \partial\Gamma_c)$ and $\|v\|_{1/2}$ is the norm in the space $H^{1/2}(\Gamma_c)$. It is known that functions from $H^{1/2}_{00}(\Gamma_c)$ can be extended to Σ by zero values, and moreover this extension belongs to $H^{1/2}(\Sigma)$. More precisely, let v be defined at Γ_c, and $\bar{v}$ be the extension of v by zero, i.e.

$$\bar{v}(x) = \begin{cases} v(x), & x \in \Gamma_c \\ 0, & x \in \Sigma \setminus \Gamma_c. \end{cases}$$

Then

$$v \in H_{00}^{1/2}(\Gamma_c) \quad \text{if and only if} \quad \bar{v} \in H^{1/2}(\Sigma).$$

With the above notations, it is possible to describe in what sense boundary conditions (4)–(5) are fulfilled. Namely, the condition $\sigma_\nu \leqslant 0$ in (5) means that

$$\langle \sigma_\nu, \phi \rangle_{1/2,00} \leqslant 0, \quad \forall\, \phi \in H_{00}^{1/2}(\Gamma_c), \quad \phi \geqslant 0 \text{ a.e. on } \Gamma_c,$$

where $\langle \cdot, \cdot \rangle_{1/2,00}$ is a duality pairing between $H_{00}^{-1/2}(\Gamma_c)$ and $H_{00}^{1/2}(\Gamma_c)$. The condition $\sigma_\tau = 0$ in (5) means that

$$\langle \sigma_\nu, \phi \rangle_{1/2,00} = 0, \quad \forall\, \phi = (\phi_1, \phi_2) \in H_{00}^{1/2}(\Gamma_c).$$

The last condition of (4) holds in the following sense

$$\langle \sigma_\nu, [u]\nu \rangle_{1/2,00} = 0.$$

2.3 Mixed Formulation of the Problem

Now we are interested to give a mixed formulation of the problem (1)–(5). Introduce the space for stresses

$$H(\mathrm{div}) = \left\{ \sigma = \{\sigma_{ij}\} \mid \sigma \in L^2(\Omega_c), \mathrm{div}\sigma \in L^2(\Omega_c) \right\}$$

with the norm

$$\|\sigma\|^2_{H(\mathrm{div})} = \|\sigma\|^2_{L^2(\Omega_c)} + \|\mathrm{div}\sigma\|^2_{L^2(\Omega_c)}$$

and the set of admissible stresses

$$H(\mathrm{div}; \Gamma_c) = \left\{ \sigma \in H(\mathrm{div}) \mid [\sigma\nu] = 0 \text{ on } \Gamma_c; \quad \sigma_\nu \leqslant 0, \quad \sigma_\tau = 0 \text{ on } \Gamma_c^{\pm} \right\}.$$

We should note at this step that for $\sigma \in H(\mathrm{div})$ the traces $(\sigma\nu)^{\pm}$ are correctly defined on $\Sigma^{\pm}$ as elements of $H^{-1/2}(\Sigma)$. The first condition in the definition of $H(\mathrm{div}; \Gamma_c)$ is fulfilled in the following sense

$$(\sigma\nu)^+ = (\sigma\nu)^- \text{ on } \Sigma$$

for any curve Σ with the prescribed properties. Relations $\sigma \leqslant 0, \sigma_\tau = 0$ on $\Gamma_c^{\pm}$ also make sense. The values σ_ν, σ_τ are defined as elements of the space $H_{00}^{-1/2}(\Gamma_c)$.

The mixed formulation of the problem (1)–(5) is as follows. We have to find a displacement field $u = (u_1, u_2)$ and stress tensor components $\sigma = \{\sigma_{ij}\}, i, j = 1, 2$, such that

$$u \in L^2(\Omega_c), \quad \sigma \in H(\mathrm{div}; \Gamma_c), \tag{10}$$

$$-\mathrm{div}\sigma = f \quad \text{in} \quad \Omega_c, \tag{11}$$

$$\int\limits_{\Omega_c} C\sigma(\overline{\sigma} - \sigma) + \int\limits_{\Omega_c} u(\mathrm{div}\overline{\sigma} - \mathrm{div}\sigma) \geqslant 0 \quad \forall \overline{\sigma} \in H(\mathrm{div}; \Gamma_c). \tag{12}$$

The tensor C is obtained by inverting the Hooke's law (2), i.e.

$$C\sigma = \varepsilon(u).$$

It is possible to establish the existence of a solution to the problem (10)–(12) and check that (10)–(12) is formally equivalent to (1)–(5) (see [16]). Existence of solutions to (10)–(12) can be proved independently of (1)–(5). On the other hand, the solution exists due to the equivalence, and we already have the solution to the problem (1)–(5).

2.4 Smooth Domain Formulation

Along with the mixed formulation (10)–(12), the so-called smooth domain formulation of the problem (1)–(5) can be provided. In this case the solution of the problem is defined in the smooth domain Ω. To do this, we should notice that the solution of the problem (1)–(5) satisfies (7)–(8), thus, the condition

$$[\sigma\nu] = 0 \quad \text{on} \quad \Gamma_c$$

holds, and, therefore, it can be proved that in the distributional sense

$$-\mathrm{div}\sigma = f \quad \text{in} \quad \Omega.$$

Hence, the equilibrium equations (1) hold in the smooth domain Ω.

Introduce the space for stresses defined in Ω,

$$\mathcal{H}(\mathrm{div}) = \{\sigma = \{\sigma_{ij}\} \mid \sigma, \ \mathrm{div}\sigma \in L^2(\Omega)\}$$

and the set of admissible stresses

$$\mathcal{H}(\mathrm{div}; \Gamma_c) = \{\sigma \in \mathcal{H}(\mathrm{div}) \mid \sigma_\tau = 0, \quad \sigma_\nu \leqslant 0 \text{ on } \Gamma_c\}.$$

The norm in the space $\mathcal{H}(\mathrm{div})$ is defined as follows

$$\|\sigma\|^2_{\mathcal{H}(\mathrm{div})} = \|\sigma\|^2_{L^2(\Omega)} + \|\mathrm{div}\sigma\|^2_{L^2(\Omega)}.$$

We see that for $\sigma \in \mathcal{H}(\mathrm{div})$, the boundary condition $\sigma_\tau = 0$, $\sigma_\nu \leqslant 0$ on Γ_c are correctly defined in the sense $H_{00}^{-1/2}(\Gamma_c)$. Thus, we can provide the smooth domain formulation for the problem (1)–(5). It is necessary to find a displacement field $u = (u_1, u_2)$ and stress tensor components $\sigma = \{\sigma_{ij}\}$, $i, j = 1, 2$, such that

$$u \in L^2(\Omega), \quad \sigma \in \mathcal{H}(\mathrm{div}; \Gamma_c), \tag{13}$$

$$-\mathrm{div}\sigma = f \quad \text{in} \quad \Omega, \tag{14}$$

$$\int_\Omega C\sigma(\overline{\sigma} - \sigma) + \int_\Omega u(\mathrm{div}\overline{\sigma} - \mathrm{div}\sigma) \geqslant 0 \quad \forall \overline{\sigma} \in \mathcal{H}(\mathrm{div}; \Gamma_c). \tag{15}$$

It is possible to prove existence of a solution to the problem (13)–(15) (see [14]). Moreover, any smooth solution of (1)–(5) satisfies (13)–(15) and, conversely, from (13)–(15) it follows (1)–(5). Advantage of the formulation (13)–(15) is that it is given in the smooth domain. This formulation reminds contact problems with thin obstacle when restrictions are imposed on sets of small dimensions.

Numerical aspects for the problems like (1)–(5) can be found, for example, in [2, 3].

2.5 *Fictitious Domain Method*

In this section we provide a connection between the problem (1)–(5) and the Signorini contact problem. It turns out that the Signorini problem is a limit problem for a family of problems like (1)–(5). First we give a formulation of the Signorini problem. Let $\Omega_1 \subset \mathbb{R}^2$ be a bounded domain with smooth boundary Γ_1, $\Gamma_1 = \Gamma_c \cup \Gamma_0$, $\Gamma_c \cap \Gamma_0 = \emptyset$, $meas\Gamma_0 > 0$.

For simplicity, we assume that Γ_c is a smooth curve (without its tip points). Denote by $\nu = (\nu_1, \nu_2)$ a unit normal inward vector to Γ_c. We have to find a displacement field $u = (u_1, u_2)$ and stress tensor components $\sigma = \{\sigma_{ij}\}$, $i, j = 1, 2$, such that

$$-\mathrm{div}\sigma = f \quad \text{in} \quad \Omega_1, \tag{16}$$

$$\sigma = A\varepsilon(u) \quad \text{in} \quad \Omega_1, \tag{17}$$

$$u = 0 \quad \text{on} \quad \Gamma_0, \tag{18}$$

$$u\nu \geqslant 0,\ \sigma_\nu \leqslant 0,\ \sigma_\tau = 0,\ u\nu \cdot \sigma_\nu = 0 \quad \text{on} \quad \Gamma_c. \tag{19}$$

Here $f = (f_1, f_2) \in L^2_{loc}(\mathbb{R}^2)$ is a given function, $A = \{a_{ijkl}\}$, $i, j, k, l = 1, 2$ is a given elasticity tensor, $a_{ijkl} \in L^\infty_{loc}(\mathbb{R}^2)$, with the usual properties of symmetry and positive definiteness.

It is well known (see [4, 5]) that the problem (16)–(19) has a variational formulation providing a solution existence. Namely, denote

$$H^1_{\Gamma_0}(\Omega_1) = \{v = (v_1, v_2) \in H^1(\Omega_1) \mid v_i = 0 \text{ on } \Gamma_0, \quad i = 1, 2\}$$

and introduce the set of admissible displacements

$$K_c = \{v = (v_1, v_2) \in H^1_{\Gamma_0}(\Omega_1) \mid \nu v \geqslant 0 \text{ a.e. on } \Gamma_c\}.$$

In this case the problem (16)–(19) is equivalent to minimization of the functional

$$\frac{1}{2} \int\limits_{\Omega_1} \sigma_{ij}(v)\varepsilon_{ij}(v) - \int\limits_{\Omega_1} f_i v_i$$

over the set K_c and can be written in the form of the variational inequality

$$u \in K_c, \tag{20}$$

$$\int\limits_{\Omega_1} \sigma_{ij}(u)\varepsilon_{ij}(v - u) \geqslant \int\limits_{\Omega_1} f_i (v_i - u_i) \quad \forall v \in K_c. \tag{21}$$

Here $\sigma_{ij}(u) = \sigma_{ij}$ are defined from the Hooke's law (17). Variational inequality (20)–(21) is equivalent to (16)–(19) and, conversely, i.e., any smooth solution of (16)–(19) satisfies (20)–(21) and from (20)–(21) it follows (16)–(19). Along with variational formulation (20)–(21), the problem (16)–(19) admits a mixed formulation which is omitted here.

The aim of this section is to prove that the problem (16)–(19) is a limit problem for a family of problems like (1)–(5). In what follows we provide explanation of this statement.

First of all we extend the domain Ω_1 by adding a domain Ω_2 with smooth boundary Γ_2. An extended domain is denoted by Ω_c, and it has a crack (cut) Γ_c. Boundary of Ω_c is $\Gamma \cup \Gamma_c^{\pm}$. Denote $\Sigma_0 = \Gamma_1 \cap \Gamma_2$, $\Sigma = \Sigma_0 \setminus \Gamma$, thus Σ does not contain its tip points.

We introduce a family of elasticity tensors with a positive parameter λ,

$$a^\lambda_{ijkl} = \begin{cases} a_{ijkl} & \text{in } \Omega_1 \\ \lambda^{-1} a_{ijkl} & \text{in } \Omega_2. \end{cases}$$

Denote $A^\lambda = \{a^\lambda_{ijkl}\}$, and in the extended domain Ω_c, consider a family of the crack problems. Find a displacement field $u^\lambda = (u^\lambda_1, u^\lambda_2)$, and stress tensor components

$\sigma^\lambda = \{\sigma_{ij}^\lambda\}$, $i, j = 1, 2$, such that

$$-\mathrm{div}\sigma^\lambda = f \quad \text{in} \quad \Omega_c, \tag{22}$$

$$\sigma^\lambda = A^\lambda \varepsilon(u^\lambda) \quad \text{in} \quad \Omega_c, \tag{23}$$

$$u^\lambda = 0 \quad \text{on} \quad \Gamma, \tag{24}$$

$$[u^\lambda]\nu \geqslant 0,\ [\sigma_\nu^\lambda] = 0,\ \sigma_\nu^\lambda \cdot [u]\nu = 0 \quad \text{on} \quad \Gamma_c, \tag{25}$$

$$\sigma_\nu^\lambda \leqslant 0,\ \sigma_\tau^\lambda = 0 \quad \text{on} \quad \Gamma_c^\pm. \tag{26}$$

As before, $[v] = v^+ - v^-$ is the jump of v through Γ_c, where $\pm$ fit positive and negative crack faces $\Gamma_c^\pm$. All the remaining notations correspond to those of Sect. 1. We see that for any fixed $\lambda > 0$ the problem (22)–(26) describes an equilibrium state of linear elastic body with the crack Γ_c where non-penetration conditions are prescribed. Hence, the problem (22)–(26) is exactly the problem like (1)–(5), and we are interested in passage to the limit as $\lambda \to 0$. In particular, the problem (22)–(26) admits a variational formulation. Boundary conditions (25)–(26) are fulfilled in the form as it is explained in Sect. 1. It can be shown that the following convergence takes place as $\lambda \to 0$

$$u^\lambda \to u^0 \quad \text{strongly in} \quad H^1_\Gamma(\Omega_c), \tag{27}$$

$$\frac{u^\lambda}{\sqrt{\lambda}} \to 0 \quad \text{strongly in} \quad H^1(\Omega_2), \tag{28}$$

where $u^0 = u$ on Ω_1, i.e. a restriction of the limit function from (27) to Ω_1 coincides with the unique solution of the Signorini problem (16)–(19). From (27)–(28) it is seen that the limit function u^0 is zero in Ω_2. On the other hand, there is no limit passage for σ^λ in Ω_2 as $\lambda \to 0$. Thus, the domain Ω_2 can be understood as undeformable body, and the stresses are not defined in Ω_2. This means that the Signorini problem is, in fact, a crack problem with non-penetration condition between crack faces, where the crack Γ_c is located between the elastic body Ω_1 and non-deformable (rigid) body Ω_2. It is worth noting that, in fact, we can write the problem (22)–(26) in the equivalent form in the smooth domain $\Omega_c \cup \overline{\Gamma}_c$ by using the smooth domain formulation of Sect. 2.4.

3 Griffith Functionals Evaluation by the Shape Sensitivity Analysis of Energy Functionals

The velocity method [6, 25] is used in the shape sensitivity analysis of the energy functionals with respect to perturbations of a crack tip in two spatial dimensions. In Frémiot et al. [6] the Hadamard structure [25] theorem for the first and the second

order shape derivatives of differentiable shape functionals in domains with cracks is given with full proof. We use the distributed form of the shape gradient for the energy functional with respect to the crack tip perturbations in order to define the Griffith shape functional which is further considered in Sect. 47. In applications, the Griffith functional can be used, it seems, to control the crack propagation in elastic body with elastic and/or rigid inclusions.

In the crack theory, the Griffith criterion can be used for the prediction of crack propagation. This criterion says that a crack propagates provided that the derivative of the energy functional with respect to the crack length reaches a critical value. In this section we discuss the Griffith criterion and the associated Griffith functional for the model (1)–(5).

The general point of view is that we should consider a perturbed problem with respect to (1)–(5). In particular, a crack length may be perturbed. Perturbation will be characterized by a small parameter t, and $t = 0$ corresponds to the unperturbed problem, i.e. to the problem (1)–(5). To describe properly a perturbation of the problem, we should define a perturbation of the domain Ω_c. This can be done in the framework of the sensitivity analysis by the so-called velocity method (see [25]). We briefly recall this method in a way useful for our purposes.

Let us consider a given velocity field V defined in $\mathbb{R}^2$ and describe a perturbation of Ω_c by solving a Cauchy problem for a system of ODE. Namely, let $V \in W^{1,\infty}(\mathbb{R}^2)^2$ be a given field, $V = (V_1, V_2)$. Consider a Cauchy problem for finding a function $\Phi = (\Phi_1, \Phi_2)$, with x the spatial variable,

$$\frac{d\Phi}{dt}(t,x) = V(\Phi(t,x)) \quad \text{for} \quad t \neq 0, \quad \Phi(0,x) = x. \tag{29}$$

There exists a unique solution Φ to (29) such that

$$\Phi = (\Phi_1, \Phi_2)(t,x) \in C^1([0,t_0]; W^{1,\infty}_{loc}(\mathbb{R}^2)^2), \quad |t_0| > 0. \tag{30}$$

Simultaneously, we can find a solution $\Psi = (\Psi_1, \Psi_2)$ to the following Cauchy problem

$$\frac{d\Psi}{dt}(t,y) = -V(\Psi(t,y)) \quad \text{for} \quad t \neq 0, \quad \Psi(0,y) = y \tag{31}$$

with the some regularity

$$\Psi = (\Psi_1, \Psi_2)(t,y) \in C^1([0,t_0]; W^{1,\infty}_{loc}(\mathbb{R}^2)^2), \quad |t_0| > 0. \tag{32}$$

It can be proved that for any fixed t, the inverse function of $\Phi(t,\cdot)$ is the function $\Psi(t,\cdot)$, thus

$$y = \Phi(t, \Psi(t,y)), \quad x = \Psi(t, \Phi(t,x)), \quad x, y \in \mathbb{R}^2.$$

Due to this, we have a one-to-one mapping between the domain Ω_c and a perturbed domain Ω_c^t, namely

$$y = \Phi(t, x) : \Omega_c \to \Omega_c^t,$$
$$x = \Psi(t, y) : \Omega_c^t \to \Omega_c.$$

Moreover, by (30), (32), we have the following asymptotic expansions (I denotes the identity operator)

$$\Phi(t, x) = x + tV(x) + r_1(t), \tag{33}$$
$$\Psi(t, y) = y - tV(y) + r_2(t), \tag{34}$$
$$\frac{\partial \Phi(t)}{\partial x} = I + t\frac{\partial V}{\partial x} + r_3(t), \tag{35}$$
$$\frac{\partial \Psi(t)}{\partial y} = I - t\frac{\partial V}{\partial y} + r_4(t), \tag{36}$$
$$\|r_i(t)\|_{W^{1,\infty}_{loc}(\mathbb{R}^2)^2} = o(t), \quad i = 1, 2,$$
$$\|r_i(t)\|_{L^{\infty}_{loc}(\mathbb{R}^2)^{2\times 2}} = o(t), \quad i = 3, 4.$$

Hence, in the domain Ω_c^t it is possible to consider the following boundary value problem (perturbed with respect to (1)–(5)). Find a displacement field $u^t = (u_1^t, u_2^t)$, and stress tensor components $\sigma^t = \{\sigma_{ij}^t\}$, $i, j = 1, 2$, such that

$$-\mathrm{div}\sigma^t = f \quad \text{in} \quad \Omega_c^t, \tag{37}$$
$$\sigma^t = A\varepsilon(u^t) \quad \text{in} \quad \Omega_c^t, \tag{38}$$
$$u^t = 0 \quad \text{on} \quad \Gamma^t, \tag{39}$$
$$[u^t]\nu^t \geqslant 0,\ [\sigma_{\nu^t}^t] = 0,\ \sigma_{\nu^t}^t \cdot [u^t]\nu^t = 0 \quad \text{on} \quad \Gamma_c^t, \tag{40}$$
$$\sigma_{\nu^t}^t \leqslant 0,\ \sigma_{\tau^t}^t = 0 \quad \text{on} \quad \Gamma_c^{t\pm}. \tag{41}$$

Here,

$$y = \Phi(t, x) : \Gamma \to \Gamma^t, \quad \Gamma_c \to \Gamma_c^t,$$

and we assume in this section that $f = (f_1, f_2) \in C^1(\mathbb{R}^2)$ and that $a_{ijkl} = const$, $i, j, k, l = 1, 2$. All the rest notations in (37)–(41) remind those of (1)–(5), in particular, $\nu^t = (\nu_1^t, \nu_2^t)$ is a unit normal vector to Γ_c^t.

We can provide a variational formulation of the problem (37)–(41). Indeed, introduce the Sobolev space

$$H^1_{\Gamma^t}(\Omega_c^t) = \{v = (v_1, v_2) \mid v_i \in H^1(\Omega_c^t),\ v_i = 0 \text{ on } \Gamma^t, \quad i = 1, 2\}$$

and the set of admissible displacements

$$K^t = \{v \in H^1_{\Gamma^t}(\Omega^t_c) \mid [v]v^t \geqslant 0 \quad \text{a.e. on} \quad \Gamma^t_c\}.$$

Consider the functional

$$\Pi(\Omega^t_c; v) = \frac{1}{2}\int\limits_{\Omega^t_c} \sigma^t_{ij}(v)\varepsilon_{ij}(v) - \int\limits_{\Omega^t_c} f_i v_i$$

and the minimization problem

$$\min_{v \in K^t} \Pi(\Omega^t_c; v). \tag{42}$$

Here, $\sigma^t_{ij}(v)$ are defined from Hooke's law similar to (38). Solution of the problem (42) exists and it satisfies the variational inequality

$$u^t \in K^t, \tag{43}$$

$$\int\limits_{\Omega^t_c} \sigma^t_{ij}(u^t)\varepsilon_{ij}(v - u^t) \geqslant \int\limits_{\Omega^t_c} f_i (v_i - u^t_i) \quad \forall v \in K^t. \tag{44}$$

Having found a solution of the problem (43)–(44) we can define the energy functional

$$\Pi(\Omega^t_c; u^t) = \frac{1}{2}\int\limits_{\Omega^t_c} \sigma^t_{ij}(u^t)\varepsilon_{ij}(u^t) - \int\limits_{\Omega^t_c} f_i u^t_i.$$

Note that for $t = 0$, we have $\Omega^0_c = \Omega_c$ and $u^0 = u$, where u is the solution of the unperturbed problem (7), (8). The question arises whether the functional $t \to \Pi(\Omega^t_c; u^t)$ is differentiable at $t = 0$. Thus, we consider the existence of The question whether

$$\frac{d}{dt}\Pi(\Omega^t_c; u^t)|_{t=0} = \lim_{t \to 0} \frac{\Pi(\Omega^t_c; u^t) - \Pi(\Omega_c; u)}{t}.$$

The answer is positive in many practical situations. We consider two cases, where the derivative

$$I = \frac{d}{dt}\Pi(\Omega^t_c; u^t)|_{t=0} \tag{45}$$

can be evaluated.

3.1 Griffith Functionals for Rectilinear Cracks

Assume for simplicity that the normal vector ν to Γ_c keeps its value under the mapping $x \to \Phi(t, x)$, i.e. $\nu^t = \nu$. In this case,

$$I = \frac{1}{2}\int\limits_{\Omega_c} \left\{ \mathrm{div} V \cdot \varepsilon_{ij}(u) - 2E_{ij}(V;u) \right\} \sigma_{ij}(u) - \int\limits_{\Omega_c} \mathrm{div}(Vf_i)u_i\,, \tag{46}$$

where

$$E_{ij}(U;v) = \frac{1}{2}(v_{i,k}U_{k,j} + v_{j,k}U_{k,i}), \quad U = \{U_{ij}\}, \quad i,j = 1,2.$$

Note that the assumption concerning the normal vector ν holds for rectilinear cracks Γ_c and vector fields V tangential to Γ_c. In this situation, (46) provides a formula for the derivative of the energy functional with respect to the crack length what is practically needed for using the Griffith criterion.

- It will be the case when $V = 1$ in a vicinity of the right crack tip and the support denoted by $\mathrm{supp} V$ belongs to a small neighborhood of this tip.
- Formula (46) for the shape derivative of the energy functional with respect to the crack length is called the *distributed shape gradient*. More precisely, by the shape gradient we understand the mapping

$$V \to \frac{1}{2}\int\limits_{\Omega_c} \left\{ \mathrm{div} V \cdot \varepsilon_{ij}(u) - 2E_{ij}(V;u) \right\} \sigma_{ij}(u) - \int\limits_{\Omega_c} \mathrm{div}(Vf_i)u_i\,. \tag{47}$$

- In Sect. 7 the expression of the distributed gradient (47) is shown to be differentiable with respect to the perturbations of the linear boundary conditions for the displacement field. In this way the shape derivative of the Griffith functional with respect to the boundary variations of an inclusion far from the crack is determined.

3.2 Griffith Functionals for Curvilinear Cracks

The formula for the derivative (45) can be derived for curvilinear cracks if the simplified assumption on the normal vector ν is not fulfilled by using an appropriate transformation of unknown functions i.e., of the displacement field [25]. We provide here the formula (45) for the crack Γ_c which is defined by a graph of a smooth function.

Let $\psi \in H^3(0, l_1)$ be a given function, $l_1 > 0$, and

$$\Sigma = \{(x_1, x_2) \mid x_2 = \psi(x_1), \quad 0 < x_1 < l_1\}.$$

Consider a crack Γ_l, $\Gamma_l \subset \Sigma$, as a graph of the function ψ,

$$\Gamma_l = \{(x_1, x_2) \mid x_2 = \psi(x_1), \quad 0 < x_1 < l\}, \quad 0 < l < l_1.$$

Here, l is a parameter that characterizes the length of the projection of the crack Γ_l onto x_1 axis. Consider a smooth cut-off function θ with a support in a vicinity of the crack tip $(l, \psi(l))$, moreover, we assume that $\theta = 1$ in a small neighborhood of $(l, \psi(l))$. We can consider a perturbation of the crack Γ_l along Σ via a small parameter t. Denote $\Omega_l = \Omega \setminus \overline{\Gamma}_l$. Perturbed crack Γ_l^t has a tip $(l + t, \psi(l + t))$, and we consider a perturbed domain $\Omega_l^t = \Omega \setminus \overline{\Gamma}_l^t$. It is possible to establish a one-to-one correspondence between Ω_l and Ω_l^t by formulas

$$\begin{aligned} y_1 &= x_1 + t\theta(x), \\ y_2 &= x_2 + \psi(x_1 + t\theta(x)) - \psi(x_1), \end{aligned} \quad (x_1, x_2) \in \Omega_l, \ (y_1, y_2) \in \Omega_l^t. \tag{48}$$

Transformation (48) is equivalent to the following (cf. (33))

$$y = x + tV(x) + r(t, x)$$

with the velocity field

$$V(x) = (\theta(x), \psi'(x_1)\theta(x)). \tag{49}$$

In the domain Ω_l^t, we can consider a perturbed problem formulation. Namely, it is necessary to find a displacement field $u^t = (u_1^t, u_2^t)$ and the stress tensor components $\sigma^t = \{\sigma_{ij}^t\}$, $i, j = 1, 2$, such that

$$-\mathrm{div}\sigma^t = f \quad \text{in} \quad \Omega_l^t, \tag{50}$$

$$\sigma^t = A\varepsilon(u^t) \quad \text{in} \quad \Omega_l^t, \tag{51}$$

$$u^t = 0 \quad \text{on} \quad \Gamma, \tag{52}$$

$$[u^t]\nu^t \geqslant 0, \ [\sigma_{\nu^t}^t] = 0, \ \sigma_{\nu^t}^t \cdot [u^t]\nu^t = 0 \quad \text{on} \quad \Gamma_l^t, \tag{53}$$

$$\sigma_{\nu^t}^t \leqslant 0, \ \sigma_{\tau^t}^t = 0 \quad \text{on} \quad \Gamma_l^{t\pm}. \tag{54}$$

Here, $\nu^t = (\nu_1^t, \nu_2^t)$ is a unit normal vector to Γ_l^t. For a solution u^t of (50)–(54) it is possible to define the energy functional

$$\Pi(\Omega_l^t; u^t) = \frac{1}{2} \int\limits_{\Omega_l^t} \sigma_{ij}^t(u^t)\varepsilon_{ij}(u^t) - \int\limits_{\Omega_l^t} f_i u_i^t$$

and to find the derivative

$$\Pi'(l) = \frac{d\,\Pi(\Omega_l^t; u^t)}{dt}|_{t=0}$$

with the formula

$$\begin{aligned}\Pi'(l) = \frac{1}{2}\int\limits_{\Omega_l} \{\mathrm{div} V \cdot \varepsilon_{ij}(u) - 2E_{ij}(V;u)\}\sigma_{ij}(u) \\ - \int\limits_{\Omega_l} \mathrm{div}(Vf_i)u_i + \int\limits_{\Omega_l} \sigma_{ij}(u)\varepsilon_{ij}(w) - \int\limits_{\Omega_l} f_i w_i\,,\end{aligned} \tag{55}$$

where the vector field V is defined in (49) and $w = (0, \theta\psi'' u_1)$ is a given function. Note that the formula (55) contains the function θ, but in fact there is no dependence of the right-hand side of (55) on θ. In particular, if $\psi'' = 0$, the formula (55) reduces to (46) with $\Omega_c = \Omega_l$. In this case we have a rectilinear crack and $\nu^t = \nu$. Formula (55) defines a derivative of the energy functional with respect to the length of the projection of the crack Γ_l onto the x_1 axis. Hence, the derivative of the energy functional with respect to the length of the curvilinear crack is as follows

$$\Pi'(s) = \Pi'(l)(\psi'(l)^2 + 1)^{-1/2},$$

where

$$s = \int\limits_0^l \sqrt{\psi'(t)^2 + 1}$$

is the length of the crack Γ_l.

To conclude this section we briefly discuss the existence of so-called invariant integrals in crack theory. It is turned out that the formula (46) for the derivative of the energy functional can be rewritten as an integral over closed curve surrounding the crack tip.

Consider the most simple case of a rectilinear crack $\Gamma_c = (0,1) \times \{0\}$ assuming that $\overline{\Gamma}_c \subset \Omega$. Let θ be a smooth cut-off function equal to 1 near the point $(1,0)$, and $\mathrm{supp}\theta$ belong to a small neighborhood of the point $(1,0)$. Then we can take the vector field

$$V = (\theta, 0)$$

in (29), (31) which, according to (33), corresponds to the following change of independent variables

$$\begin{aligned} y_1 &= x_1 + t\theta(x) + r_{11}(t), \\ y_2 &= x_2. \end{aligned}$$

In this case the formula (46) (or the formula (55) in a particular case $\psi = 0$) provides a derivative of the energy functional with respect to the crack length. This formula can be rewritten [13] as an integral over curve L surrounding the crack tip $(1, 0)$,

$$I = \int_L \left\{ \frac{1}{2} \nu_1 \sigma_{ij}(u) \varepsilon_{ij}(u) - \sigma_{ij}(u) u_{i,1} \nu_j \right\} \tag{56}$$

provided that f is equal to zero in a neighborhood of the point $(1, 0)$. We should underline two important points. First, the formula (56) is independent of L, and second, the right-hand side of (56) is equal to the derivative of the energy functional with respect to the crack length.

In fact, invariant integrals like (56) can be obtained in more complex situations. For example, we can assume that the crack Γ_c is situated on the interface between two media which means that the elasticity tensor $A = \{a_{ijkl}\}$ is as follows

$$a_{ijkl} = \begin{cases} a^1_{ijkl} \text{ for } x_2 > 0 \\ a^2_{ijkl} \text{ for } x_2 < 0. \end{cases}$$

Here, $a^1_{ijkl} = const$, $a^2_{ijkl} = const$, $i, j, k, l = 1, 2$, and $\{a^1_{ijkl}\}$, $\{a^2_{ijkl}\}$ satisfy the usual properties of symmetry and positive definiteness. In this case, formula (46) for the derivative of the energy functional holds true provided that V is tangential to Γ_c. This formula provides an existence of invariant integral of the form (56). We should remark at this point that while the integral (56) is calculated, the values $\sigma_{ij}(u) u_{i,1} \nu_j$ can be taken at Γ_c^+ or at Γ_c^-. It gives the same value of the integral (56) due to the equality

$$[\sigma_{ij}(u) u_{i,1} \nu_j] = 0 \text{ on } \Gamma_c.$$

On the other hand, we can analyze the case when a rigidity of the elastic body part $\Omega_c \cap \{x_2 < 0\}$ goes to infinity. Indeed, consider the following elasticity tensor for a positive parameter $\lambda > 0$,

$$a^\lambda_{ijkl} = \begin{cases} a^1_{ijkl} & \text{for } x_2 > 0 \\ \lambda^{-1} a^2_{ijkl} & \text{for } x_2 < 0. \end{cases}$$

Then for any fixed $\lambda > 0$, the solution of the equilibrium problem like (1)–(5) exists, and a passage to the limit as $\lambda \to 0$ can be fulfilled. As we already noted in Sect. 3, in the limit the following contact Signorini problem is obtained. Find a displacement field $u = (u_1, u_2)$ and stress tensor components $\sigma = \{\sigma_{ij}\}$, $i, j = 1, 2$, such that

$$-\text{div}\sigma = f \quad \text{in} \quad \Omega_c \cap \{x_2 > 0\}, \tag{57}$$

$$\sigma = A\varepsilon(u) \quad \text{in} \quad \Omega_c \cap \{x_2 > 0\}, \tag{58}$$

$$u = 0 \quad \text{on} \quad \partial(\Omega_c \cap \{x_2 > 0\}) \setminus \Gamma_c, \tag{59}$$

$$u\nu \geqslant 0,\ \sigma_\nu \leqslant 0,\ \sigma_\tau = 0,\ \sigma_\nu \cdot u\nu = 0 \quad \text{on} \quad \Gamma_c. \tag{60}$$

For the problem (57)–(60) it is possible to differentiate the energy functional in the direction of the vector field $V = (\theta, 0)$, where the properties of θ are described above. The formula for the derivative has the following form (cf. (46))

$$I = \frac{1}{2}\int\limits_{\Omega_1} \{\mathrm{div} V \cdot \sigma_{ij}(u) - 2E_{ij}(V,u)\}\sigma_{ij}(u) - \int\limits_{\Omega_1} \mathrm{div}(Vf_i)u_i. \tag{61}$$

Assume that $f = 0$ in a neighborhood of the point $(1,0)$. In this case, formula (61) can be rewritten in the form of invariant integral

$$I = \int\limits_{L_1} \left\{ \frac{1}{2}\nu_1 \sigma_{ij}(u)\varepsilon_{ij}(u) - \sigma_{ij}(u)u_{i,1}\nu_j \right\}, \tag{62}$$

where L_1 is a smooth curve "covering" the point $(1,0)$. Like for invariant integrals in the crack problems, formula (62) is independent of a choice of L_1.

4 Domain Decomposition Technique for Singularly Perturbed Elliptic Boundary Value Problems

Our primary concern is the domain decomposition technique [20, 23, 24] in application to the shape sensitivity analysis of the Griffith shape functional. However, the precise results on the shape sensitivity analysis of the Griffith shape functional are given in a forthcoming paper. In this paper we collect all the results recently obtained for shape-topological sensitivity analysis of the broad class of variational inequalities for elastic bodies with cracks. The asymptotic analysis in singularly perturbed geometrical domain is performed by domain decomposition technique. The boundary variations are used far from the defect, and the influence of the domain perturbations is imposed on the variational inequality by means of the Steklov–Poincaré operator defined within the domain decomposition technique. In this way the conical differentiability of solutions to the variational inequality with respect to the regular perturbations of the boundary conditions can be employed for shape-topological sensitivity analysis of the specific functional defined in the subdomain which contains the crack. This is the case of the Griffith shape functional evaluated for a crack with nonlinear boundary conditions prescribed on the crack lips.

The reference domain $\Omega \setminus \overline{\Gamma}_c$ of the elastic body under considerations is divided into two subdomains Ω_c with a crack Γ_c inside and Ω_i with an elastic inclusion ω inside. The domains are coupled within the nonlinear elasticity boundary value problem with the nonlocal boundary conditions defined on the interface $\Gamma_{sp} := \overline{\Omega}_i \cap \overline{\Omega}_c$ by an appropriate Steklov–Poincaré operator. In this section, however, we introduce the domain decomposition technique for the evaluation of the topological derivatives [20, 23, 24].

Let us consider the linear elliptic boundary value problems, and describe the domain decomposition technique for asymptotic analysis of the energy functional in singularly perturbed geometrical domains. The method is presented for simplicity for circular holes and for the Laplacian with Neumann conditions on the hole, and the Dirichlet condition on the outer boundary. In such a case the function $f(\varepsilon) = \varepsilon^2$ is used in asymptotic analysis. The shape functional is defined by the associated energy functional to the boundary value problem.

The domain decomposition technique and the Steklov–Poincaré nonlocal boundary operators are used in the topological sensitivity analysis of nonlinear variational problems. We start with a scalar linear boundary value problem in order to present the outline of the method. Therefore, given domains Ω and $\Omega_\varepsilon(\hat{x}) = \Omega \setminus \overline{B_\varepsilon(\hat{x})} \subset \mathbb{R}^2$, where $B_\varepsilon(\hat{x})$ is a ball of radius $\varepsilon \to 0$ and center at a point $\hat{x} \in \Omega$ far from the boundary $\Gamma = \partial\Omega$, with $\overline{B_\varepsilon} \Subset \Omega$. By u_ε we denote a unique classical solution of the Poisson equation in singularly perturbed domain:

$$\begin{cases} \text{Find } u_\varepsilon \text{ such that} \\ -\Delta u_\varepsilon = b \text{ in } \Omega_\varepsilon \,, \\ u_\varepsilon = 0 \text{ on } \partial\Omega \,, \\ \partial_n u_\varepsilon = 0 \text{ on } \partial B_\varepsilon \,, \end{cases} \tag{63}$$

where $b \in C^{0,\alpha}(\overline{\Omega})$, with $\alpha \in (0,1)$, is a given element which vanishes in the vicinity of the point $\hat{x} \in \Omega$. The solution u_ε of the boundary value problem (63) is variational, since $u_\varepsilon \in \mathcal{V}_\varepsilon \subset H^1(\Omega_\varepsilon)$ minimizes the quadratic functional

$$\mathcal{I}_\varepsilon(\varphi) = \frac{1}{2}\int_{\Omega_\varepsilon} \|\nabla\varphi\|^2 - \int_{\Omega_\varepsilon} b\varphi \tag{64}$$

over the linear subspace $\mathcal{V}_\varepsilon \subset H^1(\Omega_\varepsilon)$, where $\mathcal{V}_\varepsilon$ is defined as

$$\mathcal{V}_\varepsilon := \{\varphi \in H^1(\Omega_\varepsilon) : \varphi_{|\Gamma} = 0\} \,. \tag{65}$$

The shape functional

$$\mathcal{J}(\Omega_\varepsilon) := \mathcal{J}(\Omega_\varepsilon; u_\varepsilon)\frac{1}{2}\int_{\Omega_\varepsilon} \|\nabla u_\varepsilon\|^2 - \int_{\Omega_\varepsilon} b u_\varepsilon = -\frac{1}{2}\int_{\Omega_\varepsilon} b u_\varepsilon \tag{66}$$

defined by the equality

$$\mathcal{J}(\Omega_\varepsilon; u_\varepsilon) := \mathcal{I}_\varepsilon(u_\varepsilon) \tag{67}$$

is the energy functional evaluated for the solution of the boundary value problem (63) posed in the singularly perturbed domain Ω_ε.

Proposition 4.1. *The energy admits the expansion with respect to the small parameter $\varepsilon \to 0$ of the following form:*

$$\mathcal{J}(\Omega_\varepsilon) = \mathcal{J}_\Omega(u) - \pi\varepsilon^2 \|\nabla u(\hat{x})\|^2 + o(\varepsilon^2)\ , \tag{68}$$

where $\|\nabla u(\hat{x})\|^2$ is the bulk energy density at the point $\hat{x} \in \Omega$ and u is a solution to (63) for $\varepsilon = 0$.

Remark 4.2. The bulk energy density functional $H^1(\Omega) \ni \varphi \mapsto \|\nabla\varphi(\hat{x})\|^2 \in \mathbb{R}$, in general, is not continuous at a point $\hat{x} \in \Omega$. Therefore, the bulk energy density is replaced by a continuous bilinear form $H^1(\Omega) \ni \varphi \mapsto \langle \mathcal{B}(\varphi), \varphi\rangle_{\Gamma_R} \in \mathbb{R}$. For the Laplacian in two spatial dimensions and the solution of unperturbed problem u which is harmonic in a neighborhood of $\hat{x}$, the appropriate continuous bilinear form with respect to $H^1(\Omega)$ norm, such that there is equality for u,

$$\|\nabla u(\hat{x})\|^2 = \langle \mathcal{B}(u), u\rangle_{\Gamma_R}$$

is given by (72) or (74). This replacement of $\|\nabla\varphi(\hat{x})\|^2$ by $\langle \mathcal{B}(\varphi), \varphi\rangle_{\Gamma_R}$ in the energy functional for problem (63) has been introduced in [23, 24] for the purposes of topological derivatives evaluation in the framework of domain decomposition method.

Note 4.1. If we combine (64) with (68), we arrive at the conclusion that the modified energy functional

$$H^1(\Omega) \ni \varphi \to \frac{1}{2}\int_\Omega \|\nabla\varphi\|^2 - \int_\Omega b\varphi - \pi\varepsilon^2\langle \mathcal{B}(\varphi), \varphi\rangle_{\Gamma_R} \in \mathbb{R}$$

is an approximation of (64) which furnishes the topological derivative (68) but with the minimization over unperturbed space $H^1(\Omega)$. This observation is in fact used in the domain decomposition method for unilateral problems.

4.1 Domain Decomposition Technique

Now, we are going to decompose the linear elliptic problem (63) into two parts, defined in two disjoint domains Ω_R and $C(R, \varepsilon) := B_R \setminus \overline{B_\varepsilon} \subset \Omega$, $R > \varepsilon > 0$. Two non-overlapping subdomains $\Omega_R, C(R, \varepsilon)$ of Ω_ε are selected $\Omega_\varepsilon = \Omega_R \cup \Gamma_R \cup$

$C(R,\varepsilon)$, where we assume that $R > \varepsilon_0$, $\varepsilon \in (0,\varepsilon_0]$ and Γ_R stands for the exterior boundary ∂B_R of $C(R,\varepsilon)$. Since the gradient of Sobolev functions is not continuous for test functions in $H^1(\Omega)$, but it is the case for harmonic functions, we replace the pointwise values of the gradient of test functions by a representation formula valid only for the pointwise values of the gradient of a harmonic function.

Proposition 4.3. *If the function u is harmonic in a ball $B_R \subset \mathbb{R}^2$, of radius $R > 0$ and center at $\hat{x} \in \Omega$, then the gradient of u evaluated at $\hat{x}$ is given by*

$$\nabla u(\hat{x}) = \frac{1}{\pi R^3} \int_{\Gamma_R} (x - \hat{x}) u(x) \, . \tag{69}$$

Proof. The proof of this result we leave as an exercise. □

In view of (69), since $b \equiv 0$ in B_R for sufficiently small $R > \varepsilon_0$, expansion (68) can be rewritten in the equivalent form

$$\mathcal{J}(\Omega_\varepsilon) = \mathcal{J}(\Omega) - \frac{\varepsilon^2}{\pi R^6} \left[\left(\int_{\Gamma_R} u\, x_1 \right)^2 + \left(\int_{\Gamma_R} u\, x_2 \right)^2 \right] + o(\varepsilon^2) \, , \tag{70}$$

where $x - \hat{x} = (x_1, x_2)$. As observed in [23, 24], it is interesting to note that (70) can be rewritten as follows

$$\mathcal{J}(\Omega_\varepsilon) = \mathcal{J}(\Omega) - \varepsilon^2 \langle \mathcal{B}(u), u \rangle_{\Gamma_R} + o(\varepsilon^2) \, . \tag{71}$$

with the nonlocal, positive and self-adjoint boundary operator $\mathcal{B}$ uniquely determined by its bilinear form

$$\langle \mathcal{B}(u), u \rangle_{\Gamma_R} = \frac{1}{\pi R^6} \left[\left(\int_{\Gamma_R} u\, x_1 \right)^2 + \left(\int_{\Gamma_R} u\, x_2 \right)^2 \right] \, . \tag{72}$$

From the above representation, since the line integrals on Γ_R are well defined for functions in $L^1(\Gamma_R)$, it follows that the operator $\mathcal{B}$ can be extended e.g., to a bounded operator on $L^2(\Gamma_R)$, namely

$$\mathcal{B} \in \mathcal{L}(L^2(\Gamma_R); L^2(\Gamma_R)) \, , \tag{73}$$

with the same symmetric bilinear form

$$\langle \mathcal{B}(\varphi), \phi \rangle_{\Gamma_R} = \frac{1}{\pi R^6} \left[\int_{\Gamma_R} \varphi\, x_1 \int_{\Gamma_R} \phi\, x_1 + \int_{\Gamma_R} \varphi\, x_2 \int_{\Gamma_R} \phi\, x_2 \right] \, , \tag{74}$$

which is continuous for all $\varphi, \phi \in L^2(\Gamma_R)$. We observe that the bilinear form

$$L^2(\Gamma_R) \times L^2(\Gamma_R) \ni (\varphi, \phi) \mapsto \langle \mathcal{B}(\varphi), \phi \rangle_{\Gamma_R} \in \mathbb{R} \tag{75}$$

is continuous with respect to the weak convergence since it has the simple structure

$$\langle \mathcal{B}(\varphi), \phi \rangle_{\Gamma_R} = L_1(\varphi) L_1(\phi) + L_2(\varphi) L_2(\phi) \quad \varphi, \phi \in L^1(\Gamma_R) \tag{76}$$

with two linear forms $\varphi \mapsto L_1(\varphi)$ and $\phi \mapsto L_2(\phi)$, given by the line integrals on Γ_R.

4.2 Steklov–Poincaré Pseudodifferential Boundary Operators

Note 4.2. We determine the family Steklov–Poincaré boundary operators on the *outer boundary* Γ_R of the domain $C(R, \varepsilon)$, if there is a hole B_ε inside of $C(R, \varepsilon)$.

We select $R > 0$ such that the circle (or the ball for $d = 3$) B_R contains the hole B_ε and introduce the truncated domain Ω_R. For the boundary value problem defined in Ω_ε, we introduce its approximation in Ω_R. The singular perturbation Ω_ε of the geometrical domain Ω is replaced by a regular perturbation of the Steklov–Poincaré boundary operator living on the interface, which coincides with the interior boundary Γ_R of Ω_R.

Definition 4.4. The Steklov–Poincaré boundary operator

$$\mathcal{A}_\varepsilon : H^{1/2}(\Gamma_R) \to H^{-1/2}(\Gamma_R) \tag{77}$$

is defined for the Poisson equation in the domain $C(R, \varepsilon)$. For a fixed parameter $\varepsilon > 0$ and a given element $v \in H^{1/2}(\Gamma_R)$, the corresponding element in the range of the operator $\mathcal{A}_\varepsilon$ is given by the Neumann trace of a unique solution to the boundary value problem

$$\begin{cases} \text{Find } w_\varepsilon \text{ such that} \\ -\Delta w_\varepsilon = 0 \text{ in } C(R, \varepsilon) \,, \\ \quad w_\varepsilon = v \text{ on } \Gamma_R \,, \\ \partial_n w_\varepsilon = 0 \text{ on } \partial B_\varepsilon \,. \end{cases} \tag{78}$$

Then we set

$$\mathcal{A}_\varepsilon(v) = \partial_n w_\varepsilon \quad \text{on} \quad \Gamma_R \,, \tag{79}$$

where n is the unit exterior normal vector on $\partial C(R, \varepsilon)$.

Remark 4.5. Let us note that, in absence of the source term b, the energy shape functional in $C(R, \varepsilon)$ evaluated for the harmonic function w_ε coincides with the

boundary energy of the Steklov–Poincaré operator on Γ_R evaluated for the Dirichlet trace of the solution w_ε, namely

$$\int_{C(R,\varepsilon)} \|\nabla w_\varepsilon\|^2 = \langle \mathcal{A}_\varepsilon(v), v\rangle_{\Gamma_R} \ . \tag{80}$$

Therefore, the asymptotics of the energy shape functional in $C(R,\varepsilon)$ for $\varepsilon \to 0$, gives rise to the regular expansion of the Steklov–Poincaré operator:

$$\mathcal{A}_\varepsilon = \mathcal{A} - 2\varepsilon^2\mathcal{B} + \mathcal{R}_\varepsilon \ , \tag{81}$$

where the remainder denoted by $\mathcal{R}_\varepsilon$ in the above expansion is of order $o(\varepsilon^2)$ in the operator norm $\mathcal{L}(H^{1/2}(\Gamma_R); H^{-1/2}(\Gamma_R))$.

By Remark 4.5 we obtain the strong convergence of solutions in the truncated domain. In fact, let us state the following important result:

Proposition 4.6. *The sequence of solutions u_ε converges as $\varepsilon \to 0$ in the following sense. For any $R > 0$,*

$$u_\varepsilon^R \to u^R \quad \textit{strongly in} \quad H^1(\Omega_R) \ , \tag{82}$$

where $\Omega_R := \Omega \setminus \overline{B_R}$, $\varepsilon \in (0, \varepsilon_0]$, and $R > \varepsilon_0 > 0$, where B_R is a ball of radius R and center at $\hat{x} \in \Omega$.

Proof. Let u_ε^R be the restriction to Ω_R of the solution u_ε to (63), namely

$$u_\varepsilon^R \in H^1_\Gamma(\Omega_R) : \int_{\Omega_R} \nabla u_\varepsilon^R \cdot \nabla\eta + \int_{\Gamma_R} \mathcal{A}_\varepsilon(u_\varepsilon^R)\eta = \int_{\Omega_R} b\eta \quad \forall \eta \in H^1_\Gamma(\Omega_R) \ . \tag{83}$$

In the same way, for $\varepsilon = 0$ we have

$$u^R \in H^1_\Gamma(\Omega_R) : \int_{\Omega_R} \nabla u^R \cdot \nabla\eta + \int_{\Gamma_R} \mathcal{A}(u^R)\eta = \int_{\Omega_R} b\eta \quad \forall \eta \in H^1_\Gamma(\Omega_R) \ , \tag{84}$$

where u^R is the restriction to Ω_R of the solution to (63) for $\varepsilon = 0$. In addition, $H^1_\Gamma(\Omega_R)$ is a subset of $H^1(\Omega_R)$, which is defined as

$$H^1_\Gamma(\Omega_R) := \{\varphi \in H^1(\Omega_R) : \varphi_{|\Gamma} = 0\} \ . \tag{85}$$

By taking $\eta = u_\varepsilon^R - u^R$ and after subtracting the second equation from the first one we get

$$\int_{\Omega_R} \|\nabla(u_\varepsilon^R - u^R)\|^2 + \int_{\Gamma_R} (\mathcal{A}_\varepsilon(u_\varepsilon^R) - \mathcal{A}(u^R))(u_\varepsilon^R - u^R) = 0 \ . \tag{86}$$

By taking into account the expansion (81) we observe that

$$\int_{\Omega_R} \|\nabla(u_\varepsilon^R - u^R)\|^2 = \int_{\Gamma_R} (2\varepsilon^2 \mathcal{B}(u^R) - \mathcal{R}_\varepsilon(u^R))(u_\varepsilon^R - u^R) \, . \tag{87}$$

From the *Cauchy–Schwarz inequality* we obtain

$$\begin{aligned} \int_{\Omega_R} \|\nabla(u_\varepsilon^R - u^R)\|^2 \leqslant\ & 2\varepsilon^2 \|\mathcal{B}(u^R)\|_{H^{-1/2}(\Gamma_R)} \|u_\varepsilon^R - u^R\|_{H^{1/2}(\Gamma_R)} \\ & + \|\mathcal{R}_\varepsilon(u^R)\|_{H^{-1/2}(\Gamma_R)} \|u_\varepsilon^R - u^R\|_{H^{1/2}(\Gamma_R)} \, . \end{aligned} \tag{88}$$

Taking into account the *trace theorem* and the compactness of the remainder $\mathcal{R}_\varepsilon$, we have

$$\int_{\Omega_R} \|\nabla(u_\varepsilon^R - u^R)\|^2 \leqslant \varepsilon^2 C_1 \|u_\varepsilon^R - u^R\|_{H^1(\Omega_R)} \, . \tag{89}$$

Finally, from the *coercivity* of the bilinear form on the left hand side of the above inequality, namely,

$$c\|u_\varepsilon^R - u^R\|^2_{H^1(\Omega_R)} \leqslant \int_{\Omega_R} \|\nabla(u_\varepsilon^R - u^R)\|^2 \, , \tag{90}$$

we obtain

$$\|u_\varepsilon^R - u^R\|_{H^1(\Omega_R)} \leqslant C\varepsilon^2 \, , \tag{91}$$

which leads to the result, with $C = C_1/c$. □

Now, we make use of the Steklov–Poincaré operator defined above for the annulus $C(R, \varepsilon)$ in order to rewrite the energy shape functional in Ω_ε as a sum of integrals over Ω_R and of the boundary bilinear form on Γ_R,

$$\mathcal{J}(\Omega_\varepsilon) = \frac{1}{2} \int_{\Omega_R} \|\nabla u_\varepsilon\|^2 - \int_{\Omega_R} b u_\varepsilon + \frac{1}{2} \langle \mathcal{A}_\varepsilon(u_\varepsilon), u_\varepsilon \rangle_{\Gamma_R} \, , \tag{92}$$

which is possible since the source term b vanishes in the small ball B_R around the point $\hat{x} \in \Omega$.

In conclusion, another method of evaluation of the topological derivative for the energy shape functional is now available. We have the energy shape functional in the form

$$\mathcal{J}(\Omega_\varepsilon) = \inf_{\varphi \in H^1_\Gamma(\Omega_R)} \left\{ \frac{1}{2} \int_{\Omega_R} \|\nabla \varphi\|^2 - \int_{\Omega_R} b\varphi + \frac{1}{2} \langle \mathcal{A}_\varepsilon(\varphi), \varphi \rangle_{\Gamma_R} \right\} \, , \tag{93}$$

where $H^1_\Gamma(\Omega_R)$ is defined trough (85). Taking into account expansion (81), from (93) it follows by an elementary argument that

$$\mathcal{J}(\Omega_\varepsilon) = \inf_{\varphi \in H^1_\Gamma(\Omega_R)} \left\{ \frac{1}{2} \int_{\Omega_R} \|\nabla \varphi\|^2 - \int_{\Omega_R} b\varphi + \frac{1}{2} \langle \mathcal{A}(\varphi), \varphi \rangle_{\Gamma_R} \right\} - \varepsilon^2 \langle \mathcal{B}(u), u \rangle_{\Gamma_R} + o(\varepsilon^2) \,, \tag{94}$$

where (94) coincides with (71). The range of applications of the presented method is not limited to linear problems only. In fact, this is the only available method without any strict complementarity type assumptions on the unknown solution of the variational inequality, for evaluation of topological derivatives of the energy shape functional for unilateral problems.

5 Domain Decomposition Technique for Topological Derivatives Evaluation

The method of compound asymptotic expansions is usually used for the purposes of asymptotic analysis of elliptic boundary value problems in singularly perturbed geometrical domains. The application of this method requires the linearization of the boundary value problem under considerations which becomes quite involved in the case of variational inequalities [1]. Therefore, the domain decomposition technique was proposed and used in [23, 24], as well as used in [20] for the purposes of *topological derivation* for variational inequalities which describe the static frictionless contact between an elastic body and a rigid foundation as well as for cracks with the unilateral non-penetration condition.

We recall that the Sobolev space $H^1(\Omega)$ is the Dirichlet space for the natural order, we refer the reader e.g. to Frémiot et al. [6] for further details in the case of contact problems in linear elasticity. By the Dirichlet-Sobolev space we mean the ordered Sobolev spaces e.g., $H^1(\Omega)$ or $H^{1/2}(\partial\Omega)$ with the following property for the natural order. If the function $x \mapsto u(x)$ is in the Sobolev space, then the function $x \mapsto u^+(x) := \max\{u(x), 0\}$ belongs to the Sobolev space.

5.1 Problem Formulation

Let us consider the new boundary value problem, with nonlinear boundary conditions on $\Gamma_c \subset \Omega$. For the domain with a hole $B_\varepsilon(\hat{x})$, where $\hat{x} \in \Omega$, the *boundary value problem* takes the following form:

$$\left\{\begin{array}{l}\text{Find } u_\varepsilon \text{ such that}\\ \quad -\Delta u_\varepsilon = b \quad \text{in } \Omega_\varepsilon\ ,\\ \quad\quad u_\varepsilon = 0 \quad \text{on } \Gamma\ ,\\ \quad \partial_n u_\varepsilon = 0 \quad \text{on } \partial B_\varepsilon\ ,\\ \left.\begin{array}{l} u_\varepsilon \geqslant 0\\ \partial_n u_\varepsilon \leqslant 0\\ u_\varepsilon\, \partial_n u_\varepsilon = 0\end{array}\right\} \text{on } \Gamma_c\ ,\end{array}\right. \tag{95}$$

where the source term $b \in C^{0,\alpha}(\overline{\Omega})$ vanishes in the neighborhood of the point $\hat{x} \in \Omega$. A weak solution u_ε of problem (95) minimizes the energy functional (64) over a cone in the Sobolev space, and the shape energy functional takes the form

$$\mathcal{J}(\Omega_\varepsilon) = \inf_{\varphi \in \{\mathcal{V}_\varepsilon : \varphi_{|\Gamma_c} \geqslant 0\}} \left\{ \frac{1}{2} \int_{\Omega_\varepsilon} \|\nabla \varphi\|^2 - \int_{\Omega_\varepsilon} b\varphi \right\}\ , \tag{96}$$

where the liner space $\mathcal{V}_\varepsilon$ is defined by (65).

Now, let us consider the domain decomposition method for (95), assuming that $\Gamma_c \subset \Omega_R$. In particular, this means that the linear space $H^1_\Gamma(\Omega_R)$ defined through (85) is replaced in (93) by the *convex and closed subset*

$$\mathcal{K} := \{\varphi \in H^1_\Gamma(\Omega_R) : \varphi_{|\Gamma_c} \geqslant 0\}\ , \tag{97}$$

and the functional including the Steklov–Poincaré operator is as follows

$$\mathcal{I}^R_\varepsilon(u^R_\varepsilon) = \inf_{\varphi \in \mathcal{K}} \left\{ \frac{1}{2} \int_{\Omega_R} \|\nabla \varphi\|^2 - \int_{\Omega_R} b\varphi + \frac{1}{2} \langle \mathcal{A}_\varepsilon(\varphi), \varphi \rangle_{\Gamma_R} \right\}\ . \tag{98}$$

In order to establish the equality

$$\mathcal{I}^R_\varepsilon(u^R_\varepsilon) \equiv \mathcal{J}(\Omega_\varepsilon)\ , \tag{99}$$

it is sufficient to show that the minimizer u^R_ε in (98) coincides with the restriction to Ω_R of the minimizer u_ε of the corresponding quadratic functional defined in the whole singularly perturbed domain Ω_ε, which is left as an exercise. In this way we obtain

$$\begin{aligned}\mathcal{J}(\Omega_\varepsilon) &= \frac{1}{2} \int_{\Omega_\varepsilon} \|\nabla u_\varepsilon\|^2 - \int_{\Omega_\varepsilon} b u_\varepsilon\\ &= \frac{1}{2} \int_{\Omega_R} \|\nabla u_\varepsilon\|^2 - \int_{\Omega_R} b u_\varepsilon + \frac{1}{2} \langle \mathcal{A}_\varepsilon(u_\varepsilon), u_\varepsilon \rangle_{\Gamma_R}\\ &= \mathcal{I}^R_\varepsilon(u^R_\varepsilon)\\ &= \inf_{\varphi \in \mathcal{K}} \left\{ \frac{1}{2} \int_{\Omega_R} \|\nabla \varphi\|^2 - \int_{\Omega_R} b\varphi + \frac{1}{2} \langle \mathcal{A}_\varepsilon(\varphi), \varphi \rangle_{\Gamma_R} \right\}\ ,\end{aligned} \tag{100}$$

thus, the topological derivative of $\mathcal{J}(\Omega)$ can be evaluated by using the expansion of $\mathcal{I}_\varepsilon^R(u_\varepsilon^R)$. The assumption required for the derivation of $\mathcal{I}_\varepsilon^R(u_\varepsilon^R)$ with respect to the parameter ε at $\varepsilon = 0^+$ is only the strong convergence as $\varepsilon \to 0$ for fixed $R > 0$, namely $u_\varepsilon^R \to u^R$ strongly in $H^1(\Omega_R)$, i.e., there is no need for differentiability properties of the minimizer $u_\varepsilon^R \in H^1(\Omega_R)$ with respect to ε (see the proof of Proposition 4.6).

5.2 *Hadamard Differentiability of Minimizer for Parametric Programming in Function Spaces*

The existence of the conical differential for the mapping

$$[0, \varepsilon_0) \ni \varepsilon \mapsto u_\varepsilon^R \in H^1(\Omega_R) \tag{101}$$

is established.

We introduce:

- The quadratic functional

$$\mathcal{G}^R(\varphi) := \frac{1}{2} a^R(\varphi, \varphi) - l^R(\varphi) + \frac{1}{2} \langle \mathcal{A}(\varphi), \varphi \rangle_{\Gamma_R} - \varepsilon^2 \langle \mathcal{B}(\varphi), \varphi \rangle_{\Gamma_R} , \tag{102}$$

  where

$$a^R(\varphi, \varphi) = \int_{\Omega_R} \|\nabla \varphi\|^2 \quad \text{and} \quad l^R(\varphi) = \int_{\Omega_R} b\varphi \, . \tag{103}$$

- The coincidence set

$$\Xi := \{x \in \Gamma_c : u^R = 0\} \, . \tag{104}$$

- The linear form (non-negative measure)

$$\langle \mu_c, \varphi \rangle := a^R(u^R, \varphi) - l^R(\varphi) + \langle \mathcal{A}(u^R), \varphi \rangle_{\Gamma_R} \, . \tag{105}$$

- The convex cone

$$\mathcal{S}_K(u^R) = \{\varphi \in H^1_\Gamma(\Omega_R) : \varphi \geqslant 0 \quad \text{q.e. on} \quad \Xi, \quad \langle \mu_c, \varphi \rangle = 0\} \, . \tag{106}$$

We recall that the symbol q.e. reads “quasi everywhere” and it means, everywhere, with possible exception on a set of *null capacity*.

Theorem 5.1. *For fixed $R > 0$ we have*

$$\|u_\varepsilon^R - u^R\|_{H^1_\Gamma(\Omega_R)} \leqslant C_R \varepsilon^2 \,. \tag{107}$$

Furthermore, there is an expansion with respect to $\varepsilon \to 0^+$,

$$u_\varepsilon^R = u^R + \varepsilon^2 v^R + o^R(\varepsilon^2) \quad in \quad H^1(\Omega_R) \,. \tag{108}$$

The element $v^R \in H^1(\Omega_R)$ is uniquely determined by a solution to the following quadratic minimization problem

$$\mathcal{G}^R(v^R) = \inf_{\varphi \in \mathcal{S}_K(u^R)} \mathcal{G}^R(\varphi) \,. \tag{109}$$

Remark 5.2. The result established in Theorem 5.1 can be obtained as well for a class of contact problems by an application of general results given in [6, 25].

5.3 Topological Derivatives

In this section the outline of the domain decomposition method for variational inequalities is given. The topological derivative can be evaluated for the energy shape functional. The scalar elliptic equation as well as the linear elasticity system in two spatial dimensions with the unilateral conditions far from the hole are considered. The case of three spatial dimensions can be described in the same manner. The unilateral conditions are imposed for the weak solutions of elliptic boundary value problems by a cone constraint for the minimization of the quadratic energy functional. We recall that the cone of admissible displacements in contact problems of linear elasticity is defined by the non-penetration condition. The unilateral condition is only an approximation of the real condition and it is prescribed for normal displacements in the contact zone. Thus the normal displacements in the contact zone belong to a positive cone in the space of traces.

In this part we restrict ourselves to the circular holes. Let us recall the notation for the domain decomposition technique. Given a domain $\Omega_\varepsilon = \Omega \setminus \overline{B_\varepsilon} \subset \mathbb{R}^2$, with a small hole $B_\varepsilon \subset B_R$ of radius $\varepsilon \to 0$ and center at $\hat{x} \in \Omega$, we denote by $\Omega_R = \Omega \setminus \overline{B_R}$ the domain without the hole B_ε, and by $C(R, \varepsilon) = B_R \setminus \overline{B_\varepsilon}$ the ring with the small hole B_ε inside. It means that the domain Ω_ε is decomposed into two subdomains, the truncated one Ω_R and the ring $C(R, \varepsilon)$. The main idea which is employed here is to perform the asymptotic analysis for a linear problem and then apply the result to the nonlinear problem in a smaller domain called truncated domain. This is possible for unilateral conditions prescribed on $\Gamma_c \subset \Omega_R$, where the set Γ_c is far from the hole B_ε, and therefore far from the ball B_R.

Under this geometrical assumption it is possible to restrict the asymptotic analysis to the ring $C(R, \varepsilon)$. Then the obtained result on the asymptotic behavior of

the associated solution to the boundary value problem defined in the ring is applied to the variational inequality considered in the truncated domain Ω_R. In this way the singular domain perturbation in the ring influences, by a regular perturbation, the boundary conditions on the interface for variational inequality. The regular perturbation is governed by a nonlocal, pseudodifferential, self-adjoint boundary operator of Steklov–Poincaré type. The nonlocal Steklov–Poincaré operator is introduced on the interface between two subdomains, it is the exterior boundary Γ_R of the ring, which is exactly the interior boundary of the truncated domain Ω_R. The subproblem to be solved in the truncated domain is a variational inequality associated to the constrained minimization problem over a closed convex cone $\mathcal{K} \subset H^1(\Omega_R)$:

Find a unique minimizer $u_\varepsilon \in \mathcal{K}$ of the quadratic energy functional

$$\mathcal{I}_\varepsilon^R(\varphi) = \frac{1}{2} a^R(\varphi, \varphi) - l^R(\varphi) + \frac{1}{2} \langle \mathcal{A}_\varepsilon(\varphi), \varphi \rangle_{\Gamma_R} , \tag{110}$$

where $\mathcal{A}_\varepsilon$ stands for the Steklov–Poincaré operator for the ring $C(R, \varepsilon)$ and $\langle \cdot, \cdot \rangle_{\Gamma_R}$ is the duality pairing defined for the fractional Sobolev spaces $H^{-1/2}(\Gamma_R) \times H^{1/2}(\Gamma_R)$ on the interface Γ_R, associated with the corresponding Steklov–Poincaré operator $\mathcal{A}_\varepsilon : H^{1/2}(\Gamma_R) \mapsto H^{-1/2}(\Gamma_R)$. We need an assumption on its asymptotic behavior, which is:

Condition 5.3 *The Steklov–Poincaré operator for the ring $C(R, \varepsilon)$ admits the expansion for $\varepsilon > 0$, ε small enough,*

$$\mathcal{A}_\varepsilon = \mathcal{A} - 2 f(\varepsilon) \mathcal{B} + \mathcal{R}_\varepsilon , \tag{111}$$

with an appropriate function $f(\varepsilon) \to 0$, when $\varepsilon \to 0$, depending on the boundary conditions on the hole, where the remainder $\mathcal{R}_\varepsilon$ is of order $o(f(\varepsilon))$ in the operator norm $\mathcal{L}(H^{1/2}(\Gamma_R); H^{-1/2}(\Gamma_R))$.

Remark 5.4. In the scalar case the operator $\mathcal{B}$ is defined by the bilinear form (74). From (81) it follows that $f(\varepsilon) = \varepsilon^2$ for the Neumann boundary conditions on the hole B_ε. For our specific applications, expansion (111) results from the asymptotics of the shape energy functional in the ring $C(R, \varepsilon)$, as it is for the scalar problem. If the form of operator $\mathcal{B}$ in (111) is known, in order to apply the general scheme the only assumption to check is the compactness condition for the remainder in the operator norm $\mathcal{L}(H^{1/2}(\Gamma_R); H^{-1/2}(\Gamma_R))$.

Therefore, the original variational inequality defined in the domain Ω_ε is replaced by the variational inequality defined in the truncated domain Ω_R. In this way, for the purposes of asymptotic analysis the original quadratic functional defined in the domain of integration Ω_ε, namely $\mathcal{J}(\Omega_\varepsilon; \varphi)$, is replaced by the functional $\mathcal{I}_\varepsilon^R(\varphi)$ defined in the truncated domain without any hole. Two problems are equivalent under the following assumption on the minimizers u_ε and u_ε^R of $\mathcal{J}(\Omega_\varepsilon; \varphi)$ and $\mathcal{I}_\varepsilon^R(\varphi)$, respectively.

Condition 5.5 *For $\varepsilon > 0$, with ε small enough, the minimizer u_ε^R in the truncated domain coincides with the restriction to the truncated domain Ω_R of the minimizer u_ε in the singularly perturbed domain Ω_ε.*

If Conditions 5.3 and 5.5 are fulfilled, then the topological asymptotic expansion of the energy functional

$$\mathcal{J}(\Omega_\varepsilon; u_\varepsilon) = \frac{1}{2}\int_{\Omega_\varepsilon} \|\nabla u_\varepsilon\|^2 - \int_{\Omega_\varepsilon} b u_\varepsilon \tag{112}$$

can be determined from the expansion of the energy functional in the truncated domain, namely

$$\mathcal{I}_\varepsilon^R(u_\varepsilon^R) = \frac{1}{2}a^R(u_\varepsilon^R, u_\varepsilon^R) - l^R(u_\varepsilon^R) + \frac{1}{2}\langle \mathcal{A}_\varepsilon(u_\varepsilon^R), u_\varepsilon^R\rangle_{\Gamma_R}\ , \tag{113}$$

where u_ε^R is the restriction to the truncated domain Ω_R of the solution u_ε to the variational inequality in the perturbed domain Ω_ε. Under our assumptions, the solution u_ε coincides with the solution obtained by the domain decomposition method.

The evaluation of the topological asymptotic expansion for the energy functional (112) is based on the equality (99), so we have $\mathcal{J}(\Omega_\varepsilon; u_\varepsilon) = \mathcal{I}_\varepsilon^R(u_\varepsilon^R)$, combined with the following characterization of the energy functional

$$\mathcal{I}_\varepsilon^R(u_\varepsilon^R) = \inf_{\varphi\in\mathcal{K}}\left\{\frac{1}{2}a^R(\varphi,\varphi) - l^R(\varphi) + \frac{1}{2}\langle \mathcal{A}_\varepsilon(\varphi),\varphi\rangle_{\Gamma_R}\right\}\ . \tag{114}$$

The quadratic term $\varphi \mapsto \frac{1}{2}\langle \mathcal{A}_\varepsilon(\varphi),\varphi\rangle_{\Gamma_R}$ of the functional $\mathcal{I}_\varepsilon^R(\varphi)$ is, in view of assumption (111) or of Condition 5.3, the regular perturbation of the bilinear form in the quadratic functional $\mathcal{I}_\varepsilon^R(\varphi)$. Therefore, we obtain the result on the differentiability of the optimal value in (113) with respect to the parameter ε.

Proposition 5.6. *Assume that:*

- *The Condition 5.3 given by (111) holds in the operator norm.*
- *The strong convergence takes place $u_\varepsilon^R \to u^R$ in the norm of the space $H^1(\Omega_R)$, which also defines the energy norm for the functional (114).*

Then, the energy in the truncated domain Ω^R has the following topological asymptotic expansion

$$\mathcal{I}_\varepsilon^R(u_\varepsilon^R) = \mathcal{I}^R(u^R) - f(\varepsilon)\langle \mathcal{B}(u^R), u^R\rangle_{\Gamma_R} + o(f(\varepsilon))\ , \tag{115}$$

where u^R is the restriction to the truncated domain Ω_R of the solution u to the original variational inequality in the unperturbed domain Ω. Therefore, the

topological derivative of the energy shape functional is obtained from the asymptotic expansion

$$\mathcal{J}(\Omega_\varepsilon; u_\varepsilon) = \mathcal{J}(\Omega; u) - f(\varepsilon)\langle \mathcal{B}(u), u\rangle_{\Gamma_R} + o(f(\varepsilon)) \,. \tag{116}$$

Proof. There are inequalities

$$\frac{\mathcal{I}^R_\varepsilon(u^R_\varepsilon) - \mathcal{I}^R(u^R_\varepsilon)}{f(\varepsilon)} \leqslant \frac{\mathcal{I}^R_\varepsilon(u^R_\varepsilon) - \mathcal{I}^R(u^R)}{f(\varepsilon)} \leqslant \frac{\mathcal{I}^R_\varepsilon(u^R) - \mathcal{I}^R(u^R)}{f(\varepsilon)} \,, \tag{117}$$

which imply the existence of the limit

$$\begin{aligned} &\limsup_{f(\varepsilon)\to 0} \frac{\mathcal{I}^R_\varepsilon(u^R_\varepsilon) - \mathcal{I}^R(u^R_\varepsilon)}{f(\varepsilon)} = \\ &\lim_{f(\varepsilon)\to 0} \frac{\mathcal{I}^R_\varepsilon(u^R_\varepsilon) - \mathcal{I}^R(u^R)}{f(\varepsilon)} = \\ &\liminf_{f(\varepsilon)\to 0} \frac{\mathcal{I}^R_\varepsilon(u^R) - \mathcal{I}^R(u^R)}{f(\varepsilon)} = \langle \mathcal{B}(u^R), u^R\rangle_{\Gamma_R} \,. \end{aligned} \tag{118}$$

From (115), in view of (99), it follows (116). □

We can conclude the analysis for the Signorini problem, and confirm that the topological derivative of the energy shape functional is given by the same formula as it is in the linear case.

Theorem 5.7. *The energy functional for the Signorini problem admits the expansion*

$$\mathcal{J}(\Omega_\varepsilon; u_\varepsilon) = \mathcal{J}(\Omega; u) - \pi\varepsilon^2 \|\nabla u\|^2 + o(\varepsilon^2) \,, \tag{119}$$

where the topological derivative $\mathcal{T}(\hat{x}) = -\|\nabla u(\hat{x})\|^2$ *is the negative bulk energy density at the point* $\hat{x} \in \Omega$. *Since the solution of the Signorini problem is harmonic in a vicinity of* $\hat{x}$, *the expansion is well defined. Therefore, the topological derivative of the energy shape functional is given by the same expression as it is in the case of linear problem.*

6 Conical Differentiability of Metric Projections in Dirichlet Spaces onto Positive Cones and Applications to the Shape Sensitivity Analysis of Variational Inequalities

The conical differentiability of the metric projection onto the positive cone in the Dirichlet space is considered in [6, 25] with applications to the sensitivity analysis of variational inequalities. There are numerous applications of such results for

the shape sensitivity analysis of the Signorini problem and frictionless contact problems in elasticity [25], crack models with unilateral non-penetration condition [6]. We recall that the shape differentiability of the energy functional for cracks with unilateral non-penetration condition which is established in [12], does require only the appropriate strong shape continuity of solutions to variational inequalities and can be obtained under mild regularity assumptions on the governing variational inequality [6]. In Sect. 6.3 the topological derivative of the energy functional is given for the elastic body with a rigid inclusion, weakened by a crack on the boundary of the inclusion. It is assumed that on the crack the unilateral non-penetration condition is prescribed which makes the analysis more involved [20] compared to the linear case.

For the convenience of the reader we recall here the abstract result [25] which is a generalization of the implicit function theorem for variational inequalities. We use the result on the Hadamard differentiability of the metric projection on polyhedral convex sets in Hilbert spaces due to Mignot and Haraux, we refer the reader to [6] for a simple proof of such a result.

6.1 *Generalization of Implicit Function Theorem for Variational Inequalities. Hadamard Differentiability of Solutions to Variational Inequalities.*

Let $\mathcal{K} \subset \mathcal{V}$ be a convex and closed subset of a Hilbert space $\mathcal{V}$, and let $\langle \cdot, \cdot \rangle$ denote the duality pairing between $\mathcal{V}'$ and $\mathcal{V}$, where $\mathcal{V}'$ denotes the dual of $\mathcal{V}$. We shall consider the following family of variational inequalities depending on a parameter $t \in [0, t_0)$, $t_0 > 0$,

$$u_t \in \mathcal{K} \ : \ a_t(u_t, \varphi - u_t) \geq \langle b_t, \varphi - u_t \rangle \qquad \forall \varphi \in \mathcal{K} \, . \tag{120}$$

Moreover, let $u_t = \mathcal{P}_t(b_t)$ be a solution to (120). For $t = 0$ we denote

$$u \in \mathcal{K} \ : \ a(u, \varphi - u) \geq \langle b, \varphi - u \rangle \qquad \forall \varphi \in \mathcal{K} \, , \tag{121}$$

with $u = \mathcal{P}(b)$ solution to (121).

Theorem 6.1. *Let us assume that:*

- *The bilinear form $a_t(\cdot, \cdot) : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is coercive and continuous uniformly with respect to $t \in [0, t_0)$. Let $\mathcal{Q}_t \in \mathcal{L}(\mathcal{V}; \mathcal{V}')$ be the linear operator defined as follows $a_t(\phi, \varphi) = \langle \mathcal{Q}_t(\phi), \varphi \rangle \ \forall \phi, \varphi \in \mathcal{V}$; it is supposed that there exists $\mathcal{Q}' \in \mathcal{L}(\mathcal{V}; \mathcal{V}')$ such that*

$$\mathcal{Q}_t = \mathcal{Q} + t\mathcal{Q}' + o(t) \quad \text{in} \quad \mathcal{L}(\mathcal{V}; \mathcal{V}') \, . \tag{122}$$

- *For $t > 0$, t small enough, the following equality holds*

$$b_t = b + tb' + o(t) \quad in \quad \mathcal{V}' , \tag{123}$$

where $b_t, b, b' \in \mathcal{V}'$.
- *The set $\mathcal{K} \subset \mathcal{V}$ is convex and closed, and for the solutions to the variational inequality*

$$\Pi b = \mathcal{P}(b) \in \mathcal{K} : \quad a(\Pi b, \varphi - \Pi b) \geq \langle b, \varphi - \Pi b \rangle \quad \forall \varphi \in \mathcal{K} \tag{124}$$

the following differential stability result holds

$$\forall h \in \mathcal{V}' : \quad \Pi(b + sh) = \Pi b + s\Pi' h + o(s) \quad in \quad \mathcal{V} \tag{125}$$

for $s > 0$, s small enough, where the mapping $\Pi' : \mathcal{V}' \to \mathcal{V}$ is continuous and positively homogeneous and $o(s)$ is uniform, with respect to $h \in \mathcal{V}'$, on compact subsets of $\mathcal{V}'$.

Then, the solutions to the variational inequality (120) *are right-differentiable with respect to t at $t = 0$, i.e. for $t > 0$, t small enough,*

$$u_t = u + tu' + o(t) \quad in \quad \mathcal{V} , \tag{126}$$

where

$$u' = \Pi'(b' - \mathcal{Q}'u) . \tag{127}$$

Let us note, that for $b_t = 0$ and $u_t = \mathcal{P}_t(0)$ we obtain $u' = \Pi'(-\mathcal{Q}'u)$.

6.2 Applications to Unilateral Contact Problems

We recall a result on the topological derivatives of the energy functional for elastic bodies with rigid inclusions with cracks on the interfaces. We refer to [20] for the proof.

Let us introduce the description of the convex cone $\mathcal{S}_K(u)$,

$$\mathcal{S}_K(u) = \left\{ \varphi \in H^{1,\omega}_{\Gamma}(\Omega_\Upsilon) : [\![\varphi]\!] \cdot n \geqslant 0 \text{ on } \Upsilon_0; \int_{\Omega \setminus \overline{\omega}} \sigma(u) \cdot \nabla \varphi^s = \int_{\Omega_\Upsilon} b \cdot \varphi \right\} \tag{128}$$

where $\Upsilon_0 = \{x \in \Upsilon : (u - \rho_0) \cdot n = 0\}$, where $\rho_0 := u_{|_\omega}$. We have the following result:

Theorem 6.2. *Let there be given the right hand side $b_t = b + th$ of the variational inequality which governs the unilateral contact problem under investigations, then*

the unique solution $u_t \in \mathcal{K}_\omega$ is Lipschitz continuous

$$\|u_t - u\|_{H^1(\Omega_\Upsilon;\mathbb{R}^2)} \leqslant C\,t \tag{129}$$

and conically differentiable in $H^1(\Omega_\Upsilon;\mathbb{R}^2)$, that is, for $t > 0$, t small enough,

$$u_t = u + tv + o(t)\ , \tag{130}$$

where the conical differential solves the variational inequality

$$v \in \mathcal{S}_K(u) : \int_{\Omega\setminus\overline{\omega}} \sigma(v)\cdot\nabla(\eta - v)^s \geqslant \int_{\Omega_\Upsilon} h\cdot(\eta - v) \quad \forall\,\eta \in \mathcal{S}_K(u)\ . \tag{131}$$

The remainder converges to zero

$$\frac{1}{t}\|o(t)\|_{H^1(\Omega_\Upsilon;\mathbb{R}^2)} \underset{t\to 0}{\to} 0 \tag{132}$$

uniformly with respect to the direction h on the compact sets of the dual space $(H^{1,\omega}_\Gamma(\Omega_\Upsilon))'$. Thus, v is the Hadamard directional derivative of the solution to the variational inequality with respect to the right hand side.

6.3 Example: Topological Derivative of Energy Functional for the Crack on Boundaries of Rigid Inclusions

We present an example of shape-topological sensitivity analysis for a crack located on the boundary of a rigid inclusion. The rigid inclusion can be considered as the limit case of elastic inclusions. In this particular case the general theory applies and we are able to present the topological derivative of the energy functional following [20].

Let us now consider a singularly perturbed domain $\Omega_\varepsilon(\hat{x}) = \Omega \setminus \overline{B_\varepsilon(\hat{x})}$, where $B_\varepsilon(\hat{x})$ is a ball of radius $\varepsilon > 0$, $\varepsilon \to 0$, and center at $\hat{x} \in \Omega \setminus \overline{\omega}$. We assume that the hole B_ε do not touch the rigid inclusion ω, namely $\overline{B_\varepsilon} \Subset \Omega \setminus \overline{\omega}$.

We are interested in the topological asymptotic expansion of the energy shape functional of the form

$$\mathcal{J}(\Omega_\varepsilon;\varphi) = \frac{1}{2}\int_{\Omega_\varepsilon\setminus\overline{\omega}} \sigma(\varphi)\cdot\nabla\varphi^s - \int_{\Omega_\Upsilon} b\cdot\varphi\ , \tag{133}$$

with $\varphi = u_\varepsilon$ solution to the following *nonlinear system*:

$$
\begin{cases}
\text{Find } u_\varepsilon \text{ such that} \\
\qquad -\mathrm{div}\sigma(u_\varepsilon) = b \qquad \text{in } \Omega_\varepsilon \setminus \overline{\omega}\ , \\
\qquad \sigma(u_\varepsilon) = \mathbb{C}\nabla u_\varepsilon^s\ , \\
\qquad u_\varepsilon = 0 \qquad \text{on } \Gamma\ , \\
\qquad \sigma(u_\varepsilon)n = 0 \qquad \text{on } \partial B_\varepsilon\ , \\
\left.\begin{array}{r} (u_\varepsilon - \rho_0)\cdot n \geqslant 0 \\ \sigma^\tau(u_\varepsilon) = 0 \\ \sigma^{nn}(u_\varepsilon) \leqslant 0 \\ \sigma^{nn}(u_\varepsilon)(u_\varepsilon - \rho_0)\cdot n = 0 \end{array}\right\} \quad \text{on } \Upsilon^+\ , \\
\qquad -\displaystyle\int_{\partial\omega} \sigma(u_\varepsilon)n\cdot\rho = \int_\omega b\cdot\rho \ \ \forall\ \ \rho\in\mathcal{R}(\omega)\ .
\end{cases}
\tag{134}
$$

Since the problem is nonlinear, let us introduce two disjoint domains Ω_R and $C(R,\varepsilon)$, with $\Omega_R = \Omega\setminus\overline{B_R(\hat{x})}$ and $C(R,\varepsilon) = B_R\setminus\overline{B_\varepsilon} \Subset \Omega\setminus\overline{\omega}$, where $B_R(\hat{x})$ is a ball of radius $R>\varepsilon$ and center at $\hat{x}\in\Omega\setminus\overline{\omega}$. For the sake of simplicity, we assume that $b=0$ in $B_R(\hat{x})$, that is, the source term b vanishes in the neighborhood of the point $\hat{x}\in\Omega\setminus\overline{\omega}$. Thus, we have the following linear elasticity system defined in the ring $C(R,\varepsilon)$:

$$
\begin{cases}
\text{Find } w_\varepsilon \text{ such that} \\
-\mathrm{div}\sigma(w_\varepsilon) = 0 \qquad \text{in } C(R,\varepsilon)\ , \\
\qquad \sigma(w_\varepsilon) = \mathbb{C}\nabla w_\varepsilon^s\ , \\
\qquad w_\varepsilon = v \qquad \text{on } \Gamma_R\ , \\
\qquad \sigma(u_\varepsilon)n = 0 \qquad \text{on } \partial B_\varepsilon\ ,
\end{cases}
\tag{135}
$$

where Γ_R is used to denote the exterior boundary ∂B_R of the ring $C(R,\varepsilon)$. We are interested in the Steklov–Poincaré operator on Γ_R, that is

$$
\mathcal{A}_\varepsilon : v\in H^{1/2}(\Gamma_R;\mathbb{R}^2) \to \sigma(w_\varepsilon)n \in H^{-1/2}(\Gamma_R;\mathbb{R}^2)\ . \tag{136}
$$

Then we have $\sigma(u_\varepsilon^R)n = \mathcal{A}_\varepsilon(u_\varepsilon^R)$ on Γ_R, where u_ε^R is solution of the variational inequality in Ω_R, that is

$$
\begin{aligned}
u_\varepsilon^R\in\mathcal{K}_\omega : \int_{\Omega_R}\sigma(u_\varepsilon^R)\cdot\nabla(\eta-u_\varepsilon^R) + \int_{\Gamma_R}\mathcal{A}_\varepsilon(u_\varepsilon^R)\cdot(\eta-u_\varepsilon^R) \\
\geqslant \int_{\Omega_\Upsilon\setminus\overline{B_R}} b\cdot(\eta-u_\varepsilon^R) \quad \forall\eta\in\mathcal{K}_\omega\ .
\end{aligned}
\tag{137}
$$

Finally, in the ring $C(R,\varepsilon)$ we have

$$
\int_{C(R,\varepsilon)}\sigma(w_\varepsilon)\cdot\nabla w_\varepsilon^s = \int_{\Gamma_R}\mathcal{A}_\varepsilon(w_\varepsilon)\cdot w_\varepsilon\ , \tag{138}
$$

where w_ε is the solution of the elasticity system in the ring (135). Therefore the solutions u_ε^R and w_ε are defined as restriction of u_ε to the truncated domain Ω_R and to the ring $C(R,\varepsilon)$, respectively.

In particular, in the neighborhood of $\hat{x} \in \Omega \setminus \overline{\omega}$, the energy in the ring $C(R,\varepsilon)$ admits the following topological asymptotic expansion

$$\int_{C(R,\varepsilon)} \sigma(w_\varepsilon)\cdot\nabla w_\varepsilon^s = \int_{B_R} \sigma(w)\cdot\nabla w^s - 2\pi\varepsilon^2\mathbb{P}\sigma(w(\hat{x}))\cdot\nabla w^s(\hat{x}) + o(\varepsilon^2)\ . \tag{139}$$

where w is solution to (135) for $\varepsilon = 0$ and $\mathbb{P}$ is the polarization tensor. It means that w is the restriction to the disk B_R of the solution u to the nonlinear system defined in the unperturbed domain Ω_Υ. Therefore, we have that the Steklov–Poincaré operator defined by (136) admits the expansion for $\varepsilon > 0$, with ε small enough,

$$\mathcal{A}_\varepsilon = \mathcal{A} - 2\varepsilon^2\mathcal{B} + o(\varepsilon^2)\ , \tag{140}$$

where the operator $\mathcal{B}$ is determined by its bilinear form

$$\langle\mathcal{B}(w), w\rangle_{\Gamma_R} = \pi\mathbb{P}\sigma(w(\hat{x}))\cdot\nabla w^s(\hat{x})\ . \tag{141}$$

From the above results, we have that the energy shape functional associated to the cracks on boundaries of rigid inclusions embedded in elastic bodies has the following topological asymptotic expansion

$$\mathcal{J}(\Omega_\varepsilon) = \mathcal{J}(\Omega) - \pi\varepsilon^2\mathbb{P}\sigma(u(\hat{x}))\cdot\nabla u^s(\hat{x}) + o(\varepsilon^2)\ , \tag{142}$$

with the *topological derivative* $\mathcal{T}(\hat{x})$ given by

$$\mathcal{T}(\hat{x}) = -\mathbb{P}\sigma(u(\hat{x}))\cdot\nabla u^s(\hat{x})\ , \tag{143}$$

where u is solution of the variational inequality in the unperturbed domain Ω_Υ and $\mathbb{P}$ is the Pólya–Szegö polarization tensor.

Remark 6.3. From equality (138) we observe that the bilinear form (141) represents the topological derivative of the Steklov–Poincaré operator (136). In addition, since solution $u \in \mathcal{K}_\omega$ of the variational inequality is a $H^1(\Omega_\Upsilon;\mathbb{R}^2)$ function, then it is convenient to compute the topological derivative from quantities evaluated on the boundary Γ_R in similar way as for the scalar case.

7 Shape Sensitivity Analysis of the Griffith Functional

In a forthcoming paper the first order shape-topological sensitivity analysis of energy functionals is used to establish the shape differentiability of the so-called Griffith shape functional. We are going to describe briefly a result of this sort.

Example 7.1. Let $\Omega := \Omega_c \cup \Gamma \cup \Omega_i$ be an elastic body with the rectilinear crack $\Gamma_c \subset \Sigma \subset \Omega_c$, thus $\partial\Omega := \Gamma_c \cup \partial\Omega$. We consider the shape functional defined by (46) which is called the Griffith functional

$$J(\Omega) := \frac{1}{2} \int\limits_{\Omega_c} \left\{ \mathrm{div} V \cdot \varepsilon_{ij}(u) - 2E_{ij}(V;u) \right\} \sigma_{ij}(u) - \int\limits_{\Omega_c} \mathrm{div}(Vf_i)u_i \,,$$

where the displacement field u is given by the unique solution of the variational inequality

$$u \in K \ : \ a(u, v-u) \geqslant (f, v-u) \quad \forall u \in K \,, \tag{144}$$

and the velocity vector field V is compactly supported in Ω_c. We need the decomposition of Ω into Ω_c and Ω_i for the purposes of the domain decomposition technique to our problem. Let $\omega \subset \Omega_i$ be an elastic inclusion.

Proposition 7.2. *Assume that the energy shape functional $\mathcal{E}(\Omega_i)$ is shape differentiable in the direction of the velocity field W compactly supported in a neighborhood of the inclusion $\omega \subset \Omega_i$, then the Griffith functional is directionally differentiable in the direction of the velocity field W.*

The result is proved by the domain decomposition technique with a linear problem in Ω_i which is used to determine the expansion of the energy functional with respect to the boundary variations of an inclusion and the nonlinear problem in cracked subdomain Ω_c which is used to obtain the conical differentiability of the solution with respect to the variations of the Steklov–Poincaré operator:

- the differentiability of the energy functional in the subdomain Ω_i implies the differentiability of the associated Steklov–Poincaré operator defined on the Lipschitz curve given by the interface $\overline{\Omega}_i \cap \overline{\Omega}_c$ with respect to the scalar parameter $t \to 0$ which governs the boundary variations of the inclusion ω;
- the expansion of the Steklov–Poincaré nonlocal boundary pseudodifferential operator obtained in the subdomain Ω_i is used in the boundary conditions for the variational inequality defined in the cracked subdomain Ω_c and leads to the conical differential of the solution to the unilateral problem in the subdomain;
- the one term expansion of the solution to the unilateral problem is used in the Griffith functional in order to obtain the directional derivative with respect to the boundary variations of the inclusion.

Acknowledgements The authors acknowledge support by the European Science Foundation within the Programme 'Optimization with PDE Constraints (OPTPDE)'.

Jan Sokolowski is supported by the Brazilian Research Council (CNPq), through the Special Visitor Researcher Framework of the Science Without Borders Programme, under grant 400273/2012-8.

This work has been supported by the DFG-EC315 "Engineering of Advanced Materials" and by DFG-SFB 814 "Additive Manufacturing".

References

[1] I.I. Argatov, J. Sokołowski, Asymptotics of the energy functional in the Signorini problem under small singular perturbation of the domain. (Russian) Zh. Vychisl. Mat. Mat. Fiz. **43**, 744–758 (2003); translation in Comput. Math. Math. Phys. **43**(5), 710–724 (2003)

[2] Z. Belhachmi, J.-M. Sac-Epée, J. Sokołowski, Approximation par la méthode des élement finit de la formulation en domaine régulière de problems de fissures. C. R. Acad. Sci. Paris Ser. I **338**, 499–504 (2004)

[3] Z. Belhachmi, J.M. Sac-Epée, J. Sokołowski, Mixed finite element methods for smooth domain formulation of crack problems. SIAM J. Numer. Anal. **43**, 1295–1320 (2005)

[4] G. Fichera, Existence theorems in elasticity, in *Festkörpermechanik/Mechanics of Solids, Handbuch der Physik (Encyclopedia of Physics)*, ed. by S. Flügge, C.A. Truesdell, VIa/2. (Springer, Berlin/Heidelberg/New York, 1984), pp. 347–389

[5] G. Fichera, Boundary value problems of elasticity with unilateral constraints, in *Festkörpermechanik/Mechanics of Solids, Handbuch der Physik (Encyclopedia of Physics)*, ed. by S. Flügge, C.A. Truesdell, VIa/2.(Springer, Berlin/Heidelberg/New York, 1984), pp. 391–424.

[6] G. Frémiot, W. Horn, A. Laurain, M. Rao, J. Sokołowski, On the analysis of boundary value problems in nonsmooth domains. Dissertationes Math. **462**, 149 (2009)

[7] P. Hild, A. Münch, Y. Ousset, On the control of crack growth in elastic media. C. R. Mecanique **336**, 422–427 (2008)

[8] A.M. Khludnev, G. Leugering, Optimal control of cracks in elastic bodies with thin rigid inclusions. Z. Angew. Math. Mech. **91**, 125–137 (2011)

[9] A.M. Khludnev, J. Sokołowski, *Modelling and Control in Solid Mechanics*. (Birkhäuser, Basel/Boston/Berlin, 1997)

[10] A.M. Khludnev, J. Sokołowski, On solvability of boundary value problems in elastoplasicity. Control Cybern. **27**, 311–330 (1997)

[11] A.M. Khludnev, J. Sokołowski, Griffith formula and Rice integral for elliptic equations with unilateral conditions in nonsmooth domains. Eur. J. Appl. Math. **10**, 379–394 (1999)

[12] A.M. Khludnev, J. Sokołowski, Griffith formulae for elasticity systems with unilateral conditions in domains with cracks. Eur. J. Mech. A Solids **19**, 105–120 (2000)

[13] A.M. Khludnev, J. Sokołowski, On differentation of energy functionals in the crack theory with possible contact between crack faces. J. Appl. Math. Mechs. **64**, 464–475 (2000)

[14] A.M. Khludnev, J. Sokołowski, Smooth domain method for crack problem. Q. Appl. Math. **62**, 401–422 (2004)

[15] A.M. Khludnev, K. Ohtsuka, J. Sokołowski, On derivative of energy functional for elastic bodies with a crack and unilateral conditions. Q. Appl. Math. **60**, 99–109 (2002)

[16] A.M. Khludnev, J. Sokołowski, K. Szulc, Shape and topological sensitivity analysis in domains with cracks. Appl. Math. **55**, 433–469 (2010)

[17] A.M. Khludnev, G. Leugering, M. Specovius-Neugebauer, Optimal control of inclusion and crack shapes in elastic bodies. J. Optim. Theory Appl. **155**, 54–78 (2012)

[18] G. Leugering, M. Prechtel, P. Steinmann, M. Stingl, A cohesive crack propagation model: mathematical theory and numerical solution. Commun. Pure Appl. Anal. **12**, 1705–1729 (2013)

[19] A. Münch, P. Pedregal, Relaxation of an optimal design problem in fracture mechanic: the anti-plane case. ESAIM Control Optim. Calc. Var. **16**, 719–743 (2010)

[20] A.A. Novotny, J. Sokołowski, in *Topological Derivatives in Shape Optimization*. Series: Interaction of Mechanics and Mathematics. (Springer, Berlin/Heidelberg/New York, 2013)

[21] M. Prechtel, P. Leiva Ronda, R. Janisch, A. Hartmaier, G. Leugering, P. Steinmann, M. Stingl, Simulation of fracture in heterogeneous elastic materials with cohesive zone models. Int. J. Fract. **168**, 15–29 (2010)

[22] M. Prechtel, G. Leugering, P. Steinmann, M. Stingl, Towards optimization of crack resistance of composite materials by adjusting of fiber shapes. Eng. Fract. Mech. **78**, 944–960 (2011)

[23] J. Sokołowski, A. Żochowski, Modeling of topological derivatives for contact problerms. Numer. Math. **102**, 145–179 (2005)
[24] J. Sokołowski, A. Żochowski, Topological derivatives for optimization of plane elasticity contact problems. Eng. Anal. Bound. Elem. **32**, 900–908 (2008)
[25] J. Sokołowski, J.-P. Zolésio, *Introduction to Shape Optimization. Shape Sensitivity Analysis*. (Springer, Berlin/Heidelberg/New York, 1992)

Boundary Stabilization of Numerical Approximations of the 1-D Variable Coefficients Wave Equation: A Numerical Viscosity Approach

Aurora Marica and Enrique Zuazua

Abstract In this paper, we consider the boundary stabilization problem associated to the $1-d$ *wave equation* with both *variable density and diffusion coefficients* and to its *finite difference* semi-discretizations. It is well-known that, for the finite difference semi-discretization of the constant coefficients wave equation on uniform meshes (Tébou and Zuazua, Adv. Comput. Math. 26:337–365, 2007) or on some non-uniform meshes (Marica and Zuazua, BCAM, 2013, preprint), the discrete decay rate fails to be uniform with respect to the mesh-size parameter. We prove that, under suitable regularity assumptions on the coefficients and after adding an appropriate *artificial viscosity* to the numerical scheme, the decay rate is uniform as the mesh-size tends to zero. This extends previous results in Tébou and Zuazua (Adv. Comput. Math. 26:337–365, 2007) on the constant coefficient wave equation. The methodology of proof consists in applying the classical *multiplier technique* at the discrete level, with a multiplier adapted to the variable coefficients.

Keywords 1-d wave equation • Artificial viscosity • Boundary stabilization • Variable density and diffusion coefficients

Mathematics Subject Classification (2010). Primary 49K40; Secondary 35L05

A. Marica (✉)
BCAM - Basque Center for Applied Mathematics, Alameda Mazarredo 14, 48009, Bilbao, Basque Country, Spain
e-mail: auroramarica@yahoo.com

E. Zuazua
BCAM - Basque Center for Applied Mathematics, Alameda Mazarredo 14, 48009, Bilbao, Basque Country, Spain

Ikerbasque - Basque Foundation for Science, Alameda Urquijo 36-5, Plaza Bizkaia, 48011, Bilbao, Basque Country, Spain
e-mail: zuazua@bcamath.org

R. Hoppe (ed.), *Optimization with PDE Constraints*, Lecture Notes in Computational Science and Engineering 101,
DOI 10.1007/978-3-319-08025-3_9

1 Preliminaries

Let us consider the following initial boundary value problem associated to the $1-d$ wave equation with variable coefficients and with a *damping mechanism* acting on the right endpoint of the space interval:

$$\begin{cases} \rho(x)v_{tt} - (\sigma(x)v_x)_x = 0, & x \in (0,1), t \in (0,T] \\ v(0,t) = \sigma(1)v_x(1,t) + v_t(1,t) = 0, & t \in [0,T] \\ v(x,0) = v^0(x), v_t(x,0) = v^1(x), & x \in (0,1), \end{cases} \tag{1.1}$$

where $(v^0, v^1) \in H_l^1 \times L^2(0,1)$ and $H_l^1(0,1) = \{f \in H^1(0,1), f(0) = 0\}$.

Here we have taken the dissipative boundary condition

$$\sigma(1)v_x(1,t) + v_t(1,t) = 0,$$

but similar results can be proved for more general feedback terms as, for instance,

$$\sigma(1)v_x(1,t) + kv_t(1,t) = 0,$$

with $k > 0$ and $k \neq 1$. In fact, problem (1.1) with the second dissipative condition can be reduced to (1.1) by scaling the time variable.

The energy corresponding to the solution of (1.1),

$$\mathcal{E}_{\rho,\sigma}(v(\cdot,t), v_t(\cdot,t)) := \frac{1}{2}(||\sqrt{\rho}v_t(\cdot,t)||_{L^2}^2 + ||\sqrt{\sigma}v_x(\cdot,t)||_{L^2}^2),$$

obeys the following dissipation law:

$$\frac{d}{dt}\mathcal{E}_{\rho,\sigma}(v(\cdot,t), v_t(\cdot,t)) = -|v_t(1,t)|^2 \quad \text{or} \tag{1.2}$$

$$\mathcal{E}_{\rho,\sigma}(v^0, v^1) - \mathcal{E}_{\rho,\sigma}(v(\cdot,t), v_t(\cdot,t)) = \int_0^t |v_t(1,t')|^2 \, dt'.$$

When the variable coefficients ρ and σ belong to the $BV(0,1)$ class of functions with bounded variation, the following stabilization property holds, ensuring the *exponential decay* of the energy $\mathcal{E}_{\rho,\sigma}(v(\cdot,t), v_t(\cdot,t))$ of the solutions of (1.1), i.e., the existence of two constants $M,\ \omega > 0$ such that the following estimate holds for any solution v of (1.1) corresponding to the initial data $(v^0, v^1) \in H_l^1 \times L^2(0,1)$ and any $t > 0$:

$$\mathcal{E}_{\rho,\sigma}(v(\cdot,t), v_t(\cdot,t)) \leq M \exp(-t\omega)\mathcal{E}_{\rho,\sigma}(v^0, v^1). \tag{1.3}$$

There is an extensive literature on the exponential decay of solutions of damped wave equations. In the $1-d$ case under consideration, a careful spectral analysis allows showing that, most often, the eigenfunctions of the associated generator of the semigroup constitute a Riesz basis. This allows characterizing the decay rate in terms of the spectral abscissa. In the case $\rho = \sigma = 1$, the solutions of (1.1) vanish in finite time for all $t > 2$. But this only occurs in this very exceptional case as pointed out in [5].

For multi-dimensional problems, the analysis of the exponential decay rate cannot be performed by spectral analysis methods and it requires of tools such as *microlocal analysis* (see the Appendix 2 of [20] by Bardos–Lebeau–Rauch and [1] where the exponential decay is proved by microlocal tools under the so-called *Geometric Control Condition*), *multiplier techniques* [18, 29, 30], and *Carleman inequalities* [25].

For one-dimensional problems as the one we are considering here, the exponential decay can also be proved by means of the so-called *sidewise energy estimates* or *Liouville transformations* [13, 27, 28].

In this paper, we are mainly interested in the exponential decay of the finite difference approximations and the multiplier technique seems to be the one that is better adapted to this goal. In [4], in the continuous setting, the variable coefficients case was analyzed and *multipliers adapted to the two variable coefficients* ρ and σ in (1.1) were introduced. Note that the multiplier technique requires some minimal regularity of the coefficients, say, one derivative. This is not just a technical requirement, but rather a necessary condition. Indeed, BV is the minimal regularity assumption on ρ and σ to ensure that stabilization holds since, as proved in [3] and [12], the observability property may fail at infinite order when $\sigma = 1$ for some pathological $\rho \in C^{0,\alpha}(0,1)$ for any $\alpha \in (0,1)$ or at finite order for some ρ in the log-Lipschitz or log-Zygmund class.

Consequently, one can say that the exponential decay of solutions of the dissipative wave equation with variable coefficients can be handled in a satisfactory manner using multiplier techniques. But this is far from being the case for numerical approximation schemes and this is the subject we address here. Indeed, the propagation, controllability, and observability properties of the numerical schemes for the wave equation are usually much more delicate to be analyzed due to the existence of *fictitious numerical solutions* concentrated on the high-frequency modes. These solutions do not affect the convergence of the numerical solutions in the classical sense of the numerical analysis since they are highly oscillatory and weakly converge to zero, but they become important from a stabilization point of view since the standard feedback mechanisms are not capable of dissipating their energy efficiently. Roughly, one can say that boundary feedbacks are inefficient with waves that do not reach the boundary. For the numerical schemes of the constant coefficients wave equations on uniform meshes, the theory is almost complete. Indeed, by now the full panorama of the numerical pathologies and the corresponding remedies to get uniform observability estimates as the mesh size parameter tends to zero are well-understood (see the survey paper [31] or the more recent one [10]).

The paper [26] deals with the finite difference discrete version of the exponential decay estimate (1.3) under the assumptions that the coefficients ρ, σ in the continuous model (1.1) are constant and the grid is uniform. In particular, an uniform exponential decay rate is proved to be recovered uniformly as the mesh size parameter goes to zero by adding an artificial damping in the form of a *numerical viscosity*. This artificial viscosity method is similar to the *Tychonoff regularization* one introduced by Glowinski–Li–Lions in [15] and proved to be efficient for computing convergent numerical controls for the wave equation. The efficiency of the vanishing viscosity method was proved in some other contexts as well. For example, to recover the uniform controllability properties of a semi-discrete finite difference scheme for the $1-d$ wave equation by means of *moment theory* tools (fine *biorthogonal estimates*) in [23]; to obtain uniform stabilization properties of the finite difference semi-discretizations for the *perfectly matched layer* (PML) approximation of the wave equation in [7] or of some time discretizations of a general class of exponentially stable systems by using the so-called *resolvent estimates method* in [8] and [9].

The aim of the present paper is to extend the results in [26] to the three-point finite difference semi-discretizations of the wave equation (1.1) with variable and smooth coefficients ρ, σ on uniform meshes. We will also see that a numerical scheme on a non-uniform mesh obtained as a diffeomorphic transformation of a uniform one can be written as a numerical scheme for a different variable coefficients wave equation on a uniform mesh.

Let us also highlight some recent literature on the behaviour of the numerical waves on *quasi-uniform meshes* or on non-uniform ones obtained by *diffeomorphic transformations*. In [6], uniform observability properties are obtained for mixed finite element approximations of the constant coefficients wave equation on quasi-uniform meshes. This analysis uses the particular structure of the numerical scheme allowing a full description of the spectrum, even in the non-uniform mesh case. In [2], the spectral distribution of the eigenvalues of sequences of locally Toeplitz matrices arising in numerical approximations of elliptic operators with variable coefficients on non-uniform meshes obtained by diffeomorphic transformations is analyzed. In [22], using pseudo-differential calculus tools, we rigorously define the Fourier symbol and analyze fine geometric properties of the bi-characteristic rays for the finite difference approximations of the variable coefficient wave equation on non-uniform meshes obtained by regular transformations. The spectral distribution in [2] turns out to be the integral on the phase-space domain of the Fourier symbol in [22].

This paper is organized as follows. In Sect. 2, we recall the stabilization properties of the continuous model. In Sect. 3.1, we state our main result concerning the uniform stabilization of the viscous finite difference approximations of (1.1) and we discuss the analogy between the numerical schemes for the wave equation on non-uniform meshes and those for variable coefficients wave equations on uniform meshes. In Sect. 3.2, we prove the main result in Sect. 3, Theorem 1, and in Sect. 3.3, we test numerically our theoretical result in Sect. 3.1. Finally, in Sect. 4 we provide some final comments on our results and related open problems. Additionally, for

the sake of completeness, we have included Appendix A. containing the proof of the main result in Sect. 2 concerning the continuous model.

2 Stabilization of the Continuous Model

The exact statement of the exponential decay property (1.3) is the following one:

Theorem 2.1 (Appendix in [5]). *For any strictly positive coefficients* $\rho, \sigma \in BV(0,1)$ *and any initial data* $(v^0, v^1) \in H_l^1 \times L^2(0,1)$*, the solution* v *of the damped wave equation (1.1) satisfies the estimate (1.3).*

By density arguments, it is enough to prove the decay property (1.3) for dense subsets of coefficients and initial data, with constants M and ω depending on the total variation of the coefficients ρ and σ. We will consider here the dense subclasses given by strictly positive coefficients $\rho, \sigma \in C^1(0,1)$ and initial data $(v^0, v^1) \in \mathcal{V}$, where

$$\mathcal{V} := \{(f^0, f^1) \in H_l^1 \times L^2(0,1), (\partial_{\rho,\sigma}^2 f^0, \partial_{\rho,\sigma}^2 f^1) \in H_l^1 \times L^2(0,1)\}.$$

We denote by $\partial_{\rho,\sigma}^2 f := (\sigma f_x)_x/\rho$ the weighted Laplace operator involved in the wave equation (1.1). Indeed, set v_ϵ to be the solution of (1.1) corresponding to the strictly positive coefficients $\rho_\epsilon, \sigma_\epsilon \in C^1(0,1)$ and to the initial data $(v_\epsilon^0, v_\epsilon^1) \in \mathcal{V}$ and assume that $\rho_\epsilon \longrightarrow \rho$, $\sigma_\epsilon \longrightarrow \sigma$ strongly in $BV(0,1)$ and $(v_\epsilon^0, v_\epsilon^1) \longrightarrow (v^0, v^1)$ strongly in $H_l^1 \times L^2(0,1)$ as $\epsilon \to 0$. Taking into account the dissipation law of the energy (1.2) and the uniform positivity of the coefficients $(\rho_\epsilon, \sigma_\epsilon)$, we see that v_ϵ is uniformly bounded in $L^\infty(0,\infty; H_l^1(0,1)) \cap W^{1,\infty}(0,\infty; L^2(0,1))$ as $\epsilon \to 0$ so that

$$v_\epsilon \rightharpoonup v \text{ weakly star in } L^\infty(0,\infty; H_l^1(0,1)) \cap W^{1,\infty}(0,\infty; L^2(0,1)).$$

For any $\theta \in L^1(0,T; H_l^1(0,1)) \cap W^{1,1}(0,T; L^2(0,1))$, with a finite $T > 0$, v_ϵ satisfies the weak formulation

$$\int_0^1 \rho_\epsilon v_{\epsilon,t} \theta \, dx \Big|_0^T - \int_0^T \int_0^1 \rho_\epsilon v_{\epsilon,t} \theta_{\epsilon,t} \, dx \, dt + \int_0^T v_{\epsilon,t}(1,t)\theta(1,t) \, dt + \int_0^T \int_0^1 \sigma_\epsilon v_{\epsilon,x} \theta_{\epsilon,t} \, dx \, dt = 0. \quad (2.1)$$

Since $BV(0,1) \subset L^\infty(0,1)$ continuously, from the strong convergences in BV, we get strong convergence $\rho_\epsilon \longrightarrow \rho$ and $\sigma_\epsilon \longrightarrow \sigma$ in $L^\infty(0,1)$. On the other hand, from the dissipation law (1.2) for v_ϵ, we see that $v_{\epsilon,t}(1,t)$ is bounded in $L^2(0,T)$ and

then, after extracting subsequences, there exists a function $f \in L^2(0,T)$ such that $v_{\epsilon,t}(1,\cdot) \rightharpoonup f$ weakly in $L^2(0,T)$. To identify f as $v_t(1,\cdot)$, observe that $v_\epsilon(1,t) = \int_0^1 v_{\epsilon,x}(x,t)\,dx$, so that $v_\epsilon(1,\cdot) \rightharpoonup v(1,\cdot)$ weakly in $L^2(0,T)$, $v_{\epsilon,t}(1,\cdot) \rightharpoonup v_t(1,\cdot)$ in $H^{-1}(0,T)$ and then $f = v_t(1,\cdot) \in L^2(0,T)$. We can now rigorously pass to the limit into the weak formulation (2.1) and we obtain that the limit v satisfies the wave equation (1.1) with coefficients ρ, σ and initial data (v^0, v^1). Taking into account the weak star convergence of v_ϵ to v, we obtain $\mathcal{E}_{\rho,\sigma}(v(\cdot,t), v_t(\cdot,t)) \leq \liminf_\epsilon \mathcal{E}_{\rho_\epsilon,\sigma_\epsilon}(v_\epsilon(\cdot,t), v_{\epsilon,t}(\cdot,t))$, which, combined with the passing to the limit in the exponential decay property (1.3) for the regular coefficients $\rho_\epsilon, \sigma_\epsilon$ and data $(v_\epsilon^0, v_\epsilon^1)$, concludes the exponential decay property (1.3) for all $\rho, \sigma \in BV(0,1)$ and $(v^0, v^1) \in H_l^1 \times L^2(0,1)$.

Taking into account the above observation, in what follows, we will focus only on the decay property (1.3) for strictly positive coefficients $\rho, \sigma \in C^1(0,1)$ and $(v^0, v^1) \in \mathcal{V}$. As we will see, the exponential decay property (1.3) is equivalent to the existence of a time $T > 0$ and a positive constant $C > 0$ such that the following observability inequality holds for any $(v^0, v^1) \in \mathcal{V}$ and v the corresponding solution of (1.1):

$$\mathcal{E}_{\rho,\sigma}(v^0, v^1) \leq C \int_0^T |v_t(1,t)|^2\,dt. \tag{2.2}$$

Also (2.2) is equivalent to a similar observability property for the corresponding conservative system with initial data $(u^0, u^1) \in \mathcal{V}$:

$$\begin{cases} \rho(x)u_{tt} - (\sigma(x)u_x)_x = 0, & x \in (0,1), t \in (0,T] \\ u(0,t) = u_x(1,t) = 0, & t \in [0,T] \\ u(x,0) = u^0(x), u_t(x,0) = u^1(x), & x \in (0,1), \end{cases} \tag{2.3}$$

namely,

$$\mathcal{E}_{\rho,\sigma}(u^0, u^1) \leq C' \int_0^T |u_t(1,t)|^2\,dt. \tag{2.4}$$

We obtain the following result for which we will give two proofs in Appendix A.:

Theorem 2.2. *a) For all initial data $(u^0, u^1) \in \mathcal{V}$ and all strictly positive coefficients $\rho, \sigma \in C^1(0,1)$, the observability inequality (2.4) for the solution of (2.3) holds for any time T satisfying*

$$T > \tilde{T}_\star := 2\ell, \text{ with } \ell := \int_0^1 \sqrt{\frac{\rho(x)}{\sigma(x)}}\,dx. \tag{2.5}$$

b) For all initial data $(u^0, u^1) \in \mathcal{V}$ *in (2.3), all strictly positive coefficients* $\rho, \sigma \in C^1(0,1)$ *and all* $T > 0$ *the following direct inequality for the solution of (2.3) holds:*

$$\int_0^T |u_t(1,t)|^2 \, dt \leq C'' \mathcal{E}_{\rho,\sigma}(u^0, u^1). \tag{2.6}$$

c) Equivalently, the observability inequality (2.2) and the exponential decay estimate (1.3) also hold.

Remark 1. The optimal time $\tilde{T}_\star := 2\ell$ in (2.5) can be obtained by using the Liouville transformation $\tilde{x} = L(x) := \int_0^x \sqrt{\rho/\sigma}(x')\, dx'$ as in [19], [21] or [27]. In this way, (2.3) is transformed into the wave equation $\tilde{u}_{tt} - \tilde{u}_{\tilde{x}\tilde{x}} - f_1(L^{-1}(\tilde{x}))\tilde{u}_{\tilde{x}} - f_0(L^{-1}(\tilde{x}))\tilde{u} = 0$ on the space interval $\tilde{x} \in (0, \ell)$. Here, the unknown is $\tilde{u}(\tilde{x}, t) := u(x,t)/f(x)$ and f can be any strictly positive function such that $f_1 := ((f\sqrt{\rho\sigma})' + \sqrt{\rho\sigma} f')/(f\rho)$ and $f_0 := (\sigma f')'/(f\rho)$ belong to $L^\infty(0,1)$. The principal part of this new equation being the d'Alembert operator, the optimal observability time is indeed $\tilde{T}_\star := 2\ell$.

In this paper, the optimal time $\tilde{T}_\star$ in (2.5) is rigorously obtained by the method of *sidewise energy estimates* in Appendix A.. However, since this method does not seem to be adaptable at the discrete level, we will present a second proof using a multiplier technique, that can be adapted to the discrete case. Note however that the multiplier technique provides the observability property (2.4) in a non-optimal time $T > T_\star$,

$$T_\star := 2||\varphi \sqrt{\rho/\sigma}||_{L^\infty}. \tag{2.7}$$

In order to define the function φ in (2.7), we first introduce some notations and results concerning the $BV(0,1)$ class. For strictly positive coefficients $\rho, \sigma \in L^\infty(0,1)$, we denote by $\rho_\star, \rho^\star$ and $\sigma_\star, \sigma^\star$ the four constants with the following properties:

$$0 < a_\star := \inf_{\tilde{x} \in [0,1]} a(\tilde{x}) \leq a(x) \leq a^\star := \sup_{\tilde{x} \in [0,1]} a(\tilde{x}) < \infty \tag{2.8}$$

for all $x \in [0,1]$ and $a \in \{\rho, \sigma\}$. Since $\rho, \sigma \in BV(0,1)$, they admit the Jordan decomposition $a(x) := a(0) + a^+(x) - a^-(x)$ for all $a \in \{\rho, \sigma\}$ (cf. [14]), where

$$a^+(x) := TV(a, 0, x) \text{ and } a^-(x) := -a(x) + a(0) + TV(a, 0, x). \tag{2.9}$$

By $TV(a, \alpha, \beta)$, we denote the total variation of the function a on the interval $(\alpha, \beta) \subset (0,1)$. Moreover, $\rho^+, \sigma^+ \geq 0$ and $\rho^-, \sigma^- \geq 0$ are increasing functions. If $a \in W^{1,1}(0,1) \subset BV(0,1)$, then $TV(a, \alpha, \beta) = \int_\alpha^\beta |a'(\tilde{x})|\, d\tilde{x}$. Here, $'$ denotes the derivative of a function depending on one variable.

As indicated above, one of the techniques used here to prove the observability inequality (2.4) with observability time given by (2.7) is the multiplier one (cf. [20]). For $\rho_\star$, $\sigma_\star$ and $\rho^\pm$, $\sigma^\pm$ as in (2.8) and (2.9) and inspired by [4], let us define the adapted multiplier $\varphi(x)u_x$, where φ is given as follows:

$$\varphi(x) := \exp(\psi(x)), \text{ with } \psi(x) := x + \frac{\sigma^+(x)}{\sigma_\star} + \frac{\rho^-(x)}{\rho_\star}. \tag{2.10}$$

Remark 2. Note that the classical multiplier xu_x used to prove observability for the constant coefficient wave equation in [20] is not appropriate for the variable coefficients case. Indeed, by multiplying (2.3) by xu_x and integrating in $(x,t) \in (0,1)\times(0,T)$, we obtain the following multiplier identity:

$$\begin{aligned} T\mathcal{E}_{\rho,\sigma}(u^0,u^1) = {} & \frac{\rho(1)}{2}\int_0^T |u_t(1,t)|^2\,dt - \mathcal{X}_{x\rho}(t)\Big|_0^T \\ & + \frac{1}{2}\int_0^T\int_0^1 (x\sigma'(x)|u_x(x,t)|^2 - x\rho'(x)|u_t(x,t)|^2)\,dx\,dt, \end{aligned} \tag{2.11}$$

where

$$\mathcal{X}_\alpha(t) := \int_0^1 \alpha(x)u_t(x,t)u_x(x,t)\,dx. \tag{2.12}$$

Using the Cauchy–Schwarz inequality, we get

$$|\mathcal{X}_{x\rho}(t)| \le ||x\sqrt{\rho/\sigma}||_{L^\infty}\mathcal{E}_{\rho,\sigma}(u^0,u^1),$$

so that

$$|\mathcal{X}_{x\rho}(t)|_0^T| \le 2||x\sqrt{\rho/\sigma}||_{L^\infty}\mathcal{E}_{\rho,\sigma}(u^0,u^1). \tag{2.13}$$

We also easily obtain

$$\frac{1}{2}\int_0^T\int_0^1 x(\sigma'(x)|u_x(x,t)|^2 - \rho'(x)|u_t(x,t)|^2)\,dx\,dt \le m_{\rho,\sigma}T\mathcal{E}_{\rho,\sigma}(u^0,u^1), \tag{2.14}$$

where

- $m_{\rho,\sigma} := 0$, if $\sigma' \le 0$ and $\rho' \ge 0$,
- $m_{\rho,\sigma} := \max\{||x\rho'/\rho||_{L^\infty}, ||x\sigma'/\sigma||_{L^\infty}\}$, if $\sigma' \ge 0$ and $\rho' \le 0$,

- $m_{\rho,\sigma} := ||x\rho'/\rho||_{L^\infty}$, if $\sigma', \rho' \leq 0$,
- $m_{\rho,\sigma} := ||x\sigma'/\sigma||_{L^\infty}$, if $\sigma', \rho' \geq 0$.

Combining identity (2.11) with inequalities (2.13) and (2.14), we get

$$\left(T(1-m_{\rho,\sigma}) - 2||x\sqrt{\rho/\sigma}||_{L^\infty}\right)\mathcal{E}_{\rho,\sigma}(u^0,u^1) \leq \frac{\rho(1)}{2}\int_0^T |u_t(1,t)|^2\,dt, \tag{2.15}$$

so that observability holds under the requirement $m_{\rho,\sigma} < 1$ on the coefficients. But this artificially reduces the class of admissible coefficients in Theorem 2.2.

3 Stabilization of the Numerical Approximation Scheme: A Vanishing Viscosity Method

3.1 Main Result

Consider $N \in \mathbb{N}$, $h := 1/(N+1) > 0$ to be the mesh size parameter and $\mathcal{G}^h := \{x_j = jh, 0 \leq j \leq N+1\}$ an uniform grid of size h of the interval $[0,1]$. For σ as in (1.1), set $\sigma_{j+1/2} := \sigma(x_{j+1/2})$, with $0 \leq j \leq N$, and $\rho_j := \rho(x_j)$, with $0 \leq j \leq N+1$. We denote by $\mathbf{f}^h$ the column vector of components f_j, $0 \leq j \leq N+1$. Also define the forward, backward and centered first-order finite differences ∂_+^h, ∂_-^h and ∂^h as $\partial^{h,\pm} f_j := \pm(f_{j\pm1} - f_j)/h$ and $\partial^h f_j := (f_{j+1} - f_{j-1})/(2h)$. Let $\partial_\sigma^{h,2}$ be the three-points finite difference approximation of the weighted second-order derivative $\partial_\sigma^2 := \partial_x(\sigma\partial_x)$ defined as follows:

$$\partial_\sigma^{h,2} f_j := \frac{\sigma_{j+1/2}\partial_+^h f_j - \sigma_{j-1/2}\partial_-^h f_j}{h}. \tag{3.1}$$

We also set $\partial^{h,2} := \partial_1^{h,2}$ to be the classical centered three-point finite difference operator approximating the Laplacian ∂_{xx}.

Let us now consider the following viscous finite difference semi-discretization of the damped wave equation (1.1) on the uniform grid $\mathcal{G}^h$:

$$\begin{cases} \rho_j v_j''(t) - \partial_\sigma^{h,2} v_j(t) = h^2\partial^{h,2} v_j'(t), & 1 \leq j \leq N, t \in (0,T] \\ v_0(t) = \sigma_{N+1/2}\partial_+^h v_N(t) + v_{N+1}'(t) = 0, & t \in [0,T] \\ \mathbf{v}^h(0) = \mathbf{v}^{h,0}, (\mathbf{v}^h)'(0) = \mathbf{v}^{h,1}. \end{cases} \tag{3.2}$$

In this section, $'$ denotes the time derivative. The energy associated to the damped wave equation (3.2),

$$\tilde{\mathcal{E}}^h_{\rho,\sigma}(\mathbf{v}^h(t),\mathbf{v}^h_t(t)):=\frac{h}{2}\sum_{j=1}^{N}\rho_j|v_j'(t)|^2+\frac{h}{2}\sum_{j=0}^{N}\sigma_{j+1/2}|\partial_+^h v_j(t)|^2+\frac{h^2}{2\sigma_{N+1/2}}|v_{N+1}'(t)|^2,$$

is decreasing in time and satisfies the following dissipation law:

$$\frac{d}{dt}\tilde{\mathcal{E}}^h_{\rho,\sigma}(\mathbf{v}^h(t),\mathbf{v}^h_t(t)) = -|v_{N+1}'(t)|^2 - h^3\sum_{j=0}^{N}|\partial_+^h v_j'(t)|^2 \tag{3.3}$$

or

$$\tilde{\mathcal{E}}^h_{\rho,\sigma}(\mathbf{v}^h(0),\mathbf{v}^h_t(0))-\tilde{\mathcal{E}}^h_{\rho,\sigma}(\mathbf{v}^h(t),\mathbf{v}^h_t(t))=\int_0^t|v_{N+1}'(t')|^2\,dt'+h^3\sum_{j=0}^{N}\int_0^t|\partial_+^h v_j'(t')|^2dt'.$$

As in the continuous case, we are interested in the exponential decay property of the discrete wave equation (3.2), i.e., in the existence of two positive constants $\tilde{M},\tilde{\omega}>0$ independent of h such that the following energy estimate holds for all $t>0$ and for all initial data $(\mathbf{v}^{h,0},\mathbf{v}^{h,1})$ in (3.2):

$$\tilde{\mathcal{E}}^h_{\rho,\sigma}(\mathbf{v}^h(t),\mathbf{v}^h_t(t)) \le \tilde{M}\exp(-t\tilde{\omega})\tilde{\mathcal{E}}^h_{\rho,\sigma}(\mathbf{v}^{h,0},\mathbf{v}^{h,1}). \tag{3.4}$$

Also, as for the continuous model, the decay property (3.4) is equivalent to the existence of a finite time $T>0$ and of a constant $C>0$ such that the following observability estimate holds for any solution $\mathbf{v}^h(t)$ of the damped system (3.2), uniformly as $h\to 0$:

$$\tilde{\mathcal{E}}^h_{\rho,\sigma}(\mathbf{v}^h(0),\mathbf{v}^h_t(0)) \le C\left(\int_0^T|v_{N+1}'(t)|^2\,dt+h^3\sum_{j=0}^{N}\int_0^T|\partial_+^h v_j'(t)|^2\,dt\right). \tag{3.5}$$

Consider the following finite difference approximation of (2.3):

$$\begin{cases}\rho_j u_j''(t)-\partial_\sigma^{h,2}u_j(t)=0, & 1\le j\le N, t\in(0,T]\\ u_0(t)=\partial_+^h u_N(t)=0, & t\in[0,T]\\ \mathbf{u}^h(0)=\mathbf{u}^{h,0},(\mathbf{u}^h)'(0)=\mathbf{u}^{h,1},\end{cases} \tag{3.6}$$

for which the total energy of the solutions given below is time conservative

$$\mathcal{E}^h_{\rho,\sigma}(\mathbf{u}^h(t),\mathbf{u}^h_t(t)) := \frac{h}{2}\sum_{j=1}^{N}\rho_j|u_j'(t)|^2+\frac{h}{2}\sum_{j=0}^{N}\sigma_{j+1/2}|\partial_+^h u_j(t)|^2.$$

At the same time, the decay property (3.4) is equivalent to the fact that the following observability inequality holds for any solution $\mathbf{u}^h(t)$ of the conservative

system (3.6):

$$\mathcal{E}^h_{\rho,\sigma}(\mathbf{u}^h(0),\mathbf{u}^h_t(0)) \le C'\left(\int_0^T |u'_{N+1}(t)|^2\,dt + h^3\sum_{j=0}^{N-1}\int_0^T |\partial^h_+ u'_j(t)|^2\,dt\right). \tag{3.7}$$

Let us now state the main result of this paper:

Theorem 1. *a) For all strictly positive coefficients $\rho,\sigma \in C^2(0,1)$ and all initial data $(\mathbf{u}^{h,0},\mathbf{u}^{h,1}) \in \mathbb{R}^N\times\mathbb{R}^{N+1}$ in (3.6) such that their total energy $\mathcal{E}^h_{\rho,\sigma}(\mathbf{u}^{h,0},\mathbf{u}^{h,1})$ is finite, the discrete observability inequality (3.7) holds uniformly in a finite time $T^h_\star$ that tends to $T_\star$ in (2.7) as $h\to 0$.*

b) Equivalently, (3.5) and (3.4) also hold true, uniformly with respect to $h>0$.

Remark 3. Consider $g:[0,1]\to[0,1]$ to be an increasing $C^3(0,1)$ function such that $g'(x) > g_\star > 0$ for all $x\in[0,1]$, a non-uniform grid $\mathcal{G}^h_g$ of the interval $[0,1]$ constituted by the nodes $g_j := g(x_j)$ and the following viscous numerical approximation of the constant coefficients wave equation $v_{tt}-v_{yy}=0$ on the non-uniform mesh $\mathcal{G}^h_g$:

$$\begin{cases} v''_j(t) - \dfrac{\frac{\partial^h_+ v_j(t)}{\partial^h_+ g_j} - \frac{\partial^h_- v_j(t)}{\partial^h_- g_j}}{\partial^h g_j} = \dfrac{h^2\partial^h_2 v'_j(t)}{\partial^h g_j},\ 1\le j\le N, t\in(0,T] \\ v_0(t) = \dfrac{\partial^h_+ v_N(t)}{\partial^h_+ g_N} + v'_{N+1}(t) = 0, \quad t\in[0,T] \\ \mathbf{v}^h(0)=\mathbf{v}^{h,0}, (\mathbf{v}^h)'(0)=\mathbf{v}^{h,1}. \end{cases} \tag{3.8}$$

Observe that (3.8) can be also seen as a semi-discretization of the variable coefficients wave equation (1.1) with $\partial^h g \sim g' =: \rho$ and $1/\partial^h_+ g \sim 1/g' =: \sigma$ on the uniform mesh $\mathcal{G}^h$, so that our result in Theorem 1 also applies to system (3.8) and to its conservative version in which the two damping mechanisms represented by the right hand side in the first equation and the second term $v'_{N+1}(t)$ in the second equation on the boundary are eliminated.

3.2 *Proof of the Main Result, Theorem 1*

a) Set $\varphi_j := \varphi(x_j)$, with φ as in (2.10). Let us multiply (3.6) by $\varphi_j\,\partial^h u_j(t)$, add in $1\le j\le N$ and integrate in $t\in(0,T)$. Then

$$h\sum_{j=1}^N\int_0^T \rho_j u''_j(t)\varphi_j\,\partial^h u_j(t)\,dt - h\sum_{j=1}^N\int_0^T \partial^{h,2}_\sigma u_j(t)\varphi_j\,\partial^h u_j(t)\,dt = 0. \tag{3.9}$$

Step 1. Processing of the first term in the left hand side of (3.9). After integration by parts in time, taking real parts and using the identity $2ab = |a|^2 + |b|^2 - |a-b|^2$ for all $a, b \in \mathbb{R}$, we obtain

$$h\sum_{j=1}^{N}\int_0^T \rho_j u_j''(t)\varphi_j \partial^h u_j(t)\,dt = \mathcal{X}^h_{\varphi\rho}(t)\Big|_0^T - \frac{1}{4}\sum_{j=1}^{N}\int_0^T (\varphi\rho)_j (w_{j+1}(t) - w_j(t))\,dt, \tag{3.10}$$

where $w_j(t) := |u_j'(t)|^2 + |u_{j-1}'(t)|^2 - |u_j'(t) - u_{j-1}'(t)|^2$ and

$$\mathcal{X}^h_\alpha(t) := h\sum_{j=1}^{N}\alpha_j u_j'(t)\partial^h u_j(t).$$

In what follows, we will often apply the Abel summation by parts formula:

$$\sum_{j=\alpha}^{\beta}(a_{j+1}-a_j)b_j = a_{\beta+1}b_{\beta+1} - a_\alpha b_{\alpha-1} - \sum_{j=\alpha-1}^{\beta} a_{j+1}(b_{j+1}-b_j) \tag{3.11}$$

for all $\alpha \le \beta \in \mathbb{Z}$. We first apply formula (3.11) for the last term in the right hand side of (3.10) in the particular case of $\alpha = 1$, $\beta = N$, $a_j = w_j(t)$ and $b_j = (\varphi\rho)_j$. Taking into account the boundary conditions in (3.6), we get $w_{N+1} = 2|u_{N+1}'|^2$ (since $u_{N+1}' = u_N'$) and $w_1 = 0$ (since $u_0' = 0$). In this way,

$$\sum_{j=1}^{N}(\varphi\rho)_j(w_{j+1}-w_j) = 2(\varphi\rho)_{N+1}|u_{N+1}'|^2 - \sum_{j=0}^{N}((\varphi\rho)_{j+1}-(\varphi\rho)_j)(|u_{j+1}'|^2+|u_j'|^2)$$
$$+\sum_{j=0}^{N-1}((\varphi\rho)_{j+1}-(\varphi\rho)_j)|u_{j+1}'-u_j'|^2. \tag{3.12}$$

Due to the boundary condition $u_0' = 0$, the second term in the right hand side of (3.12) verifies the identity

$$\sum_{j=0}^{N}((\varphi\rho)_{j+1}-(\varphi\rho)_j)(|u_{j+1}'|^2+|u_j'|^2) = ((\varphi\rho)_{N+1}-(\varphi\rho)_N)|u_{N+1}'|^2$$
$$+\sum_{j=1}^{N}((\varphi\rho)_{j+1}-(\varphi\rho)_{j-1})|u_j'|^2,$$

so that we get the following expression for the first term in the left hand side of (3.9):

$$h\sum_{j=1}^{N}\int_0^T \rho_j u_j''(t)\varphi_j\partial^h u_j(t)\,dt = \mathcal{X}^h_{\varphi\rho}(t)\Big|_0^T - \frac{(\varphi\rho)_{N+1}+(\varphi\rho)_N}{4}\int_0^T |u'_{N+1}(t)|^2\,dt$$

$$+\frac{h}{2}\sum_{j=1}^{N}\int_0^T \partial^h(\varphi\rho)_j|u'_j(t)|^2\,dt - R_1,\ R_1 = \frac{h^3}{4}\sum_{j=0}^{N-1}\int_0^T \partial^h_+(\varphi\rho)_j|\partial^h_+u'_j(t)|^2\,dt. \tag{3.13}$$

Step 2. Processing of the second term in the left hand side of (3.9). First, by simply taking into account that $2\partial^h u_j = \partial^h_+ u_j + \partial^h_+ u_{j-1}$, we obtain

$$-h\sum_{j=1}^{N}\int_0^T \partial^{h,2}_\sigma u_j(t)\varphi_j\partial^h u_j(t)\,dt$$

$$= -\frac{1}{2}\sum_{j=1}^{N}\int_0^T \varphi_j(\sigma_{j+1/2}|\partial^h_+u_j(t)|^2 - \sigma_{j-1/2}|\partial^h_+u_{j-1}(t)|^2)\,dt$$

$$-\frac{1}{2}\sum_{j=1}^{N}\int_0^T \varphi_j(\sigma_{j+1/2}-\sigma_{j-1/2})\partial^h_+u_j(t)\partial^h_-u_j(t)\,dt. \tag{3.14}$$

For the first term in the right hand side of (3.14), we use the Abel formula (3.11) with $\alpha = 1$, $\beta = N$, $a_j = \sigma_{j-1/2}|\partial^h_+u_{j-1}|^2$ and $b_j = \varphi_j$. Taking into account that $a_{N+1} = 0$ (since $\partial^h_+u_N = 0$), we get

$$\sum_{j=1}^{N}\varphi_j(\sigma_{j+1/2}|\partial^h_+u_j|^2 - \sigma_{j-1/2}|\partial^h_+u_{j-1}|^2) =$$

$$-\varphi_0\sigma_{1/2}|\partial^h_+u_0|^2 - \sum_{j=0}^{N-1}\sigma_{j+1/2}(\varphi_{j+1}-\varphi_j)|\partial^h_+u_j|^2. \tag{3.15}$$

On the other hand, using the identity $2ab = |a|^2 + |b|^2 - |a-b|^2$, holding for all $a, b \in \mathbb{R}$, in the particular case $a = \sigma_{j+1/2}\partial^h_+u_j$ and $b = \sigma_{j-1/2}\partial^h_-u_j$, we can transform the second term in the right hand side of (3.14) as follows:

$$\sum_{j=1}^{N}\varphi_j(\sigma_{j+1/2}-\sigma_{j-1/2})\partial^h_+u_j\partial^h_-u_j = \frac{1}{2}\sum_{j=1}^{N}\gamma_j(\sigma^2_{j+1/2}|\partial^h_+u_j|^2 + \sigma^2_{j-1/2}|\partial^h_-u_j|^2)$$

$$-\frac{h^2}{2}\sum_{j=1}^{N}\gamma_j|\partial^{h,2}_\sigma u_j|^2,\ \text{ with }\ \gamma_j := \varphi_j\frac{\sigma_{j+1/2}-\sigma_{j-1/2}}{\sigma_{j+1/2}\sigma_{j-1/2}}\quad \forall 1\le j\le N. \tag{3.16}$$

Since $\partial_-^h u_j = \partial_+^h u_{j-1}$ and $\partial_+^h u_N = 0$, we get

$$\sum_{j=1}^{N} \gamma_j (\sigma_{j+1/2}^2 |\partial_+^h u_j|^2 + \sigma_{j-1/2}^2 |\partial_-^h u_j|^2) =$$

$$\gamma_1 \sigma_{1/2}^2 |\partial_+^h u_0|^2 + \sum_{j=1}^{N-1} (\gamma_{j+1} + \gamma_j) \sigma_{j+1/2}^2 |\partial_+^h u_j|^2. \tag{3.17}$$

Finally, inserting (3.15)–(3.17) into (3.14), we obtain

$$-h \sum_{j=1}^{N} \int_0^T \partial_\sigma^{h,2} u_j(t) \varphi_j \partial^h u_j(t)\, dt = \frac{\varphi_1 \sigma_{1/2}^2}{4} \left(\frac{1}{\sigma_{1/2}} + \frac{1}{\sigma_{3/2}} \right) \int_0^T |\partial_+^h u_0(t)|^2\, dt$$

$$+ \frac{h}{2} \sum_{j=1}^{N-1} \int_0^T \sigma_{j+1/2}^2 \partial_+^h \left[\frac{\varphi_j}{2} \left(\frac{1}{\sigma_{j+1/2}} + \frac{1}{\sigma_{j-1/2}} \right) \right] |\partial_+^h u_j(t)|^2\, dt$$

$$+ \frac{h^2}{4} \sum_{j=1}^{N} \int_0^T \gamma_j |\partial_\sigma^{h,2} u_j(t)|^2\, dt. \tag{3.18}$$

Step 3. Equipartition of energy. Note that the last terms in the right hand sides of (3.13) and (3.18) are reminder terms with respect to formulas (A.4) and (A.5) (in Appendix A.) corresponding to the continuous case. The last term in (3.13) is a discrete version of $-h^2 \int_0^T \int_0^1 (\varphi\rho)'(x) |u_{tx}(x,t)|^2\, dx\, dt/4$, while the last one in (3.18) is of the form

$$-h^2 \int_0^T \int_0^1 \varphi(x) (1/\sigma)'(x) |(\sigma u_x)_x(x,t)|^2\, dx\, dt/4.$$

However, in the right hand side of the discrete observability inequality (3.7) only the discrete version of the term $h^2 \int_0^T \int_0^1 |u_{tx}(x,t)|^2\, dx\, dt$ appears, so that firstly we have to express the last terms in (3.13) and (3.18) in terms of the last one in (3.7), modulo eventually some additive reminders. To obtain an equivalent expression of the last term in (3.18), we multiply (3.6) by $\gamma_j \partial_\sigma^{h,2} u_j(t)$, add in $1 \le j \le N$ and integrate in $t \in (0,T)$. After integration by parts in time, we get

$$h^2 \sum_{j=1}^{N} \int_0^T \gamma_j |\partial_\sigma^{h,2} u_j(t)|^2\, dt = h^2 \sum_{j=1}^{N} \int_0^T \gamma_j \rho_j u_j''(t) \partial_\sigma^{h,2} u_j(t)\, dt$$

$$= h^2 \sum_{j=1}^{N} \gamma_j \rho_j u_j'(t) \partial_\sigma^{h,2} u_j(t) \Big|_0^T - h^2 \sum_{j=1}^{N} \int_0^T \gamma_j \rho_j u_j'(t) \partial_\sigma^{h,2} u_j'(t)\, dt. \tag{3.19}$$

If a_j is defined only for values of j between 1 and N, then, by applying the Abel formula (3.11) for $\alpha = 2$, $\beta = N-1$, $a_j = \sigma_{j-1/2}\partial_+^h g_{j-1}$ and $b_j = f_j$, we get:

$$\begin{aligned}
h^2 \sum_{j=1}^{N} f_j \partial_\sigma^{h,2} g_j &= h \sum_{j=1}^{N} f_j (\sigma_{j+1/2}\partial_+^h g_j - \sigma_{j-1/2}\partial_+^h g_{j-1}) \\
&= h f_N (\sigma_{N+1/2}\partial_+^h g_N - \sigma_{N-1/2}\partial_+^h g_{N-1}) + h f_1 (\sigma_{3/2}\partial_+^h g_1 - \sigma_{1/2}\partial_+^h g_0) \\
&+ h \sum_{j=2}^{N-1} f_j (\sigma_{j+1/2}\partial_+^h g_j - \sigma_{j-1/2}\partial_+^h g_{j-1}) \\
&= h f_N (\sigma_{N+1/2}\partial_+^h g_N - \sigma_{N-1/2}\partial_+^h g_{N-1}) + h f_1 (\sigma_{3/2}\partial_+^h g_1 - \sigma_{1/2}\partial_+^h g_0) \\
&+ h f_N \sigma_{N-1/2}\partial_+^h g_{N-1} - h f_1 \sigma_{3/2}\partial_+^h g_1 - h \sum_{j=1}^{N-1} (f_{j+1} - f_j)\sigma_{j+1/2}\partial_+^h g_j
\end{aligned}$$

and, finally,

$$\begin{aligned}
&h^2 \sum_{j=1}^{N} f_j \partial_\sigma^{h,2} g_j = \\
&h f_N \sigma_{N+1/2}\partial_+^h g_N - h f_1 \sigma_{1/2}\partial_+^h g_0 - h \sum_{j=1}^{N-1} (f_{j+1} - f_j)\sigma_{j+1/2}\partial_+^h g_j.
\end{aligned} \tag{3.20}$$

Apply (3.20) for $f_j = \gamma_j \rho_j u_j'$ and $g_j = u_j$. Taking into account that $\partial_+^h u_N = 0$, we transform the first term in the right hand side of (3.19) as follows:

$$h^2 \sum_{j=1}^{N} \gamma_j \rho_j u_j' \partial_\sigma^{h,2} u_j = -\mathcal{Y}_{\rho,\sigma}^h, \tag{3.21}$$

with

$$\begin{aligned}
\mathcal{Y}_{\rho,\sigma}^h(t) :&= h\gamma_1\rho_1 u_1'(t)\sigma_{1/2}\partial_+^h u_0(t) + h \sum_{j=1}^{N-1} (\gamma_{j+1}\rho_{j+1}u_{j+1}'(t) \\
&- \gamma_j \rho_j u_j'(t))\sigma_{j+1/2}\partial_+^h u_j(t).
\end{aligned}$$

Also apply (3.20) for $f_j = \gamma_j \rho_j u'_j$ and $g_j = u'_j$. Taking into account that $\partial_+^h u'_N = 0$, from the second term in the right hand side of (3.19) we get:

$$h^2 \sum_{j=1}^{N} \gamma_j \rho_j u'_j \partial_\sigma^{h,2} u'_j = -h^2 \gamma_1 \rho_1 \sigma_{1/2} |\partial_+^h u'_0|^2$$

$$- h \sum_{j=1}^{N-1} (\gamma_{j+1}\rho_{j+1} u'_{j+1} - \gamma_j \rho_j u'_j) \sigma_{j+1/2} \partial_+^h u'_j$$

$$= -h^2 \gamma_1 \rho_1 \sigma_{1/2} |\partial_+^h u'_0|^2 - \frac{h^2}{2} \sum_{j=1}^{N-1} (\gamma_{j+1}\rho_{j+1} + \gamma_j \rho_j) \sigma_{j+1/2} |\partial_+^h u'_j|^2$$

$$- \frac{1}{2} \sum_{j=1}^{N-1} (\gamma_{j+1}\rho_{j+1} - \gamma_j \rho_j) \sigma_{j+1/2} (|u'_{j+1}|^2 - |u'_j|^2). \tag{3.22}$$

In the last term in the right hand side of (3.22), for the terms of indices from $j = 2$ to $j = N-2$, we apply the Abel formula (3.11) for $\alpha = 2, \beta = N-2, a_j = |u'_j|^2$, $b_j = (\gamma_{j+1}\rho_{j+1} - \gamma_j \rho_j)\sigma_{j+1/2}$ and, taking into account that $u'_N = u'_{N+1}$, we obtain

$$\sum_{j=1}^{N-1} ((\gamma\rho)_{j+1} - (\gamma\rho)_j) \sigma_{j+1/2} (|u'_{j+1}|^2 - |u'_j|^2)$$

$$= ((\gamma\rho)_N - (\gamma\rho)_{N-1}) \sigma_{N-1/2} |u'_{N+1}|^2 - ((\gamma\rho)_2 - (\gamma\rho)_1) \sigma_{3/2} |u'_1|^2$$

$$- \sum_{j=2}^{N-1} [\big((\gamma\rho)_{j+1} - (\gamma\rho)_j\big)\sigma_{j+1/2} - \big((\gamma\rho)_j - (\gamma\rho)_{j-1}\big)\sigma_{j-1/2}] |u'_j|^2. \tag{3.23}$$

Finally, by inserting (3.20)–(3.23) into (3.19), we get the following equipartition of energy identity:

$$h^2 \sum_{j=1}^{N} \int_0^T \gamma_j |\partial_\sigma^{h,2} u_j(t)|^2 \, dt = -\mathcal{Y}_{\rho,\sigma}^h(t)|_0^T + R_2 + R_3 + R_4, \tag{3.24}$$

where

$$R_2 := h^2 \gamma_1 \rho_1 \sigma_{1/2} \int_0^T |\partial_+^h u'_0(t)|^2 dt + \frac{h^2}{2} \sum_{j=1}^{N-1} \int_0^T (\gamma_{j+1}\rho_{j+1} + \gamma_j \rho_j) \sigma_{j+1/2} |\partial_+^h u'_j(t)|^2 dt,$$

$$R_3 := \frac{h^2}{2} \frac{\partial_+^h (\gamma\rho)_{N-1} \sigma_{N-1/2}}{h} \int_0^T |u'_{N+1}(t)|^2 \, dt$$

and

$$R_4 := -\frac{h^2}{2}\frac{\partial_+^h(\gamma\rho)_1\sigma_{3/2}}{h}\int_0^T |u_1'(t)|^2\,dt - \frac{h^2}{2}\sum_{j=2}^{N-1}\int_0^T \partial_\sigma^{h,2}(\gamma\rho)_j\,|u_j'(t)|^2\,dt.$$

Step 4. The discrete multiplier identity. By putting together (3.9), (3.13), (3.18) and (3.24), we obtain the following multiplier identity:

$$E_c + E_p + E_r = \frac{(\varphi\rho)_{N+1} + (\varphi\rho)_N}{4}\int_0^T |u_{N+1}'(t)|^2\,dt - \mathcal{X}_{\varphi\rho}^h(t)|_0^T + \frac{1}{4}\mathcal{Y}_{\rho,\sigma}^h(t)|_0^T + R_1 - \frac{R_2}{4} - \frac{R_3}{4} - \frac{R_4}{4}, \tag{3.25}$$

where $\mathcal{X}_{\varphi\rho}^h$ and $\mathcal{Y}_{\rho,\sigma}^h$ are as in (3.10) and (3.21), the reminder terms R_1 to R_4 are defined in (3.13) and (3.24) and the energy terms E_c, E_p, E_r are as follows:

$$E_c := \frac{h}{2}\sum_{j=1}^{N}\int_0^T \partial^h(\varphi\rho)_j\,|u_j'(t)|^2\,dt,\ E_p := \frac{h}{2}\sum_{j=1}^{N-1}\int_0^T \sigma_{j+1/2}^2\partial_+^h\Big[\frac{\varphi_j}{2}\Big(\frac{1}{\sigma_{j+1/2}} + \frac{1}{\sigma_{j-1/2}}\Big)\Big]|\partial_+^h u_j(t)|^2\,dt + \frac{h}{2}\int_0^T \sigma_{1/2}|\partial_+^h u_0(t)|^2\,dt$$

and

$$E_r := \frac{\varphi_1\sigma_{1/2}^2}{4}\Big(\frac{1}{\sigma_{1/2}} + \frac{1}{\sigma_{3/2}}\Big)\int_0^T |\partial_+^h u_0(t)|^2\,dt - \frac{h}{2}\int_0^T \sigma_{1/2}|\partial_+^h u_0(t)|^2\,dt.$$

Step 5. Estimates on the energy terms E_c, E_p and E_r. Due to the structure of φ in (2.10), even if the coefficients ρ, σ in the continuous model are very smooth, φ cannot have more than two derivatives in $L^\infty(0,1)$ since $\varphi' = \psi'\varphi$ and ψ' involves the absolute values of ρ' and σ' (excepting the situation when both ρ and σ are monotonic). By Taylor expansions, for some $\hat{x}_{j+1/2} \in (x_j, x_{j+1})$, $\hat{x}_{j+3/4} \in (x_{j+1/2}, x_{j+1})$, $\hat{x}_{j+1/4} \in (x_j, x_{j+1/2})$, $\hat{x}_j \in (x_{j-1/2}, x_{j+1/2})$, we obtain

$$\partial^h(\varphi\rho)_j = (\varphi\rho)'(x_j) + \frac{h}{4}\big((\varphi\rho)''(\hat{x}_{j+1/2}) - (\varphi\rho)''(\hat{x}_{j-1/2})\big),$$

if $(\varphi\rho)'' \in L^\infty(0,1)$, and

$$\partial_+^h\Big[\frac{\varphi_j}{2}\Big(\frac{1}{\sigma_{j+1/2}} + \frac{1}{\sigma_{j-1/2}}\Big)\Big] = \Big(\frac{\varphi}{\sigma}\Big)'(x_{j+1/2}) + r_j^1,$$

if $\varphi'', \sigma'' \in L^\infty(0,1)$, where

$$r_j^1 := \frac{h}{4}\varphi_{j+1/2}\Big[\Big(\frac{1}{\sigma}\Big)''(\hat{x}_{j+1}) - \Big(\frac{1}{\sigma}\Big)''(\hat{x}_j)\Big] + \frac{h}{8}\Big(\frac{1}{\sigma}\Big)_{j+1/2}\big(\varphi''(\hat{x}_{j+3/4}) - \varphi''(\hat{x}_{j+1/4})\big)$$
$$+ \frac{h^2}{8}(\varphi')_{j+1/2}\Big[\Big(\frac{1}{\sigma}\Big)''(\hat{x}_{j+1}) + \Big(\frac{1}{\sigma}\Big)''(\hat{x}_j)\Big] + \frac{h^2}{16}\Big(\frac{1}{\sigma}\Big)'_{j+1/2}\big(\varphi''(\hat{x}_{j+3/4}) + \varphi''(\hat{x}_{j+1/4})\big)$$
$$+ \frac{h^3}{32}\Big[\varphi''(\hat{x}_{j+3/4})\Big(\frac{1}{\sigma}\Big)''(\hat{x}_{j+1}) - \varphi''(\hat{x}_{j+1/4})\Big(\frac{1}{\sigma}\Big)''(\hat{x}_j)\Big].$$

At the points $x \in (0,1)$ where $\sigma'(x) = 0$ or $\rho'(x) = 0$, the second-order derivative $\varphi''(x)$ has jumps since it depends on $(|\rho'|)'(x)$ and $(|\sigma'|)'(x)$. We define then rigorously $\varphi''(x)$ as the generalized derivative (or subderivative, cf. [24], pp. 213) of φ' taking at the jump points x the set value given by the closed convex hull of $\{\varphi''(x-), \varphi''(x+)\}$, where $f(x\pm)$ is the value of f to the left/right of the point x.

Thus, by taking into account that φ satisfies the second and third inequalities in (A.7), we get the following lower bound on E_c and E_p:

$$E_c \geq \frac{C_c^h h}{2}\sum_{j=1}^{N}\int_0^T \rho_j |u_j'(t)|^2\, dt, \text{ with } C_c^h := 1 - \frac{h}{2}\frac{||(\varphi\rho)''||_{L^\infty}}{\rho_\star}, \tag{3.26}$$

and

$$E_p \geq \frac{C_p^h h}{2}\sum_{j=0}^{N-1}\int_0^T \sigma_{j+1/2}|\partial_+^h u_j(t)|^2\, dt, \text{ with } C_p^h := 1 - p^1(h) \tag{3.27}$$

and

$$p^1(h) := \frac{h}{2}\frac{||\varphi||_{L^\infty}||(1/\sigma)''||_{L^\infty}}{\sigma_\star} + \frac{h}{4}\frac{||\varphi''||_{L^\infty}}{\sigma_\star^2} + O(h^2).$$

Set $C_\star^h := \min\{C_c^h, C_p^h\} = 1 - O(h)$. Taking into account (3.26) and (3.27) and the time conservation of the energy of the solutions of (3.6), we get

$$E_c + E_p \geq C_\star^h T \mathcal{E}_{\rho,\sigma}^h(\mathbf{u}^{h,0}, \mathbf{u}^{h,1}). \tag{3.28}$$

Moreover, using the first inequality in (A.7), it is easy to check that, for any $h \leq (1 + \sigma_\star/\sigma^\star)/2$, we get that E_r mimics the positivity of the last term in the left hand side of (A.6), i.e.,

$$E_r \geq 0. \tag{3.29}$$

Step 6. Estimates on the remainder terms R_1 and R_2. Observe that R_1 and R_2 are of the same nature. Therefore, we can write their sum as follows:

$$R_1 - \frac{R_2}{4} = \frac{h^3}{4}\sum_{j=0}^{N-1}\int_0^T c_j|\partial_+^h u_j'(t)|^2\,dt, \tag{3.30}$$

with

$$\begin{aligned} c_0 &:= \partial_+^h(\varphi\rho)_0 - \frac{\gamma_1\rho_0\sigma_{1/2}}{h},\\ c_j &:= \partial_+^h(\varphi\rho)_j - \frac{(\gamma_{j+1}\rho_{j+1}+\gamma_j\rho_j)\sigma_{j+1/2}}{2h}, \quad 1\le j\le N-1, \end{aligned}$$

and φ as in (2.10). By taking into account the expression of γ_j, $1\le j\le N$, in (3.16), we obtain the following equivalent expressions of the coefficients c_j:

$$\begin{aligned} c_0 &= \frac{\varphi_1\rho_1\frac{\sigma_{1/2}}{\sigma_{3/2}} - \varphi_0\rho_0}{h} \quad \text{and}\\ c_j &= \frac{\varphi_{j+1}\rho_{j+1}\left(1+\frac{\sigma_{j+1/2}}{\sigma_{j+3/2}}\right) - \varphi_j\rho_j\left(1+\frac{\sigma_{j+1/2}}{\sigma_{j-1/2}}\right)}{2h}, \quad 1\le j\le N-1. \end{aligned}$$

By Taylor expansions around $x_{j+1/2}$, we obtain the following expressions of the coefficients c_j, $0\le j\le N-1$, in which $\hat{x}_{j+1/4}\in(x_j,x_{j+1/2})$, $\hat{x}_{j+3/4}\in(x_{j+1/2},x_{j+1})$ and $\hat{x}_j\in(x_{j-1/2},x_{j+1/2})$:

$$\begin{aligned} c_0 &= \Big[\sigma\Big(\frac{\varphi\rho}{\sigma}\Big)'\Big]_{1/2} + \frac{h}{2}(\sigma\varphi\rho)_{1/2}\Big(\frac{1}{\sigma}\Big)'(\hat{x}_1) + \frac{h}{8}\big((\varphi\rho)''(\hat{x}_{3/4})\\ &\quad - (\varphi\rho)''(\hat{x}_{1/4})\big) + \frac{h}{2}\Big[\sigma(\varphi\rho)'\Big(\frac{1}{\sigma}\Big)'\Big]_{1/2} + \frac{h^2}{4}\Big[\sigma(\varphi\rho)'\Big]_{1/2}\Big(\frac{1}{\sigma}\Big)''(\hat{x}_1)\\ &\quad + \frac{h^2}{8}(\varphi\rho)''(\hat{x}_{3/4})\Big[\sigma\Big(\frac{1}{\sigma}\Big)'\Big]_{1/2} + \frac{h^3}{16}\sigma_{1/2}(\varphi\rho)''(\hat{x}_{3/4})\Big(\frac{1}{\sigma}\Big)''(\hat{x}_1) \end{aligned} \tag{3.31}$$

and, for $1\le j\le N-1$,

$$\begin{aligned} c_j &= \Big[\sigma\Big(\frac{\varphi\rho}{\sigma}\Big)'\Big]_{j+1/2} + \frac{h}{4}(\sigma\varphi\rho)_{j+1/2}\Big[\Big(\frac{1}{\sigma}\Big)''(\hat{x}_{j+1}) - \Big(\frac{1}{\sigma}\Big)''(\hat{x}_j)\Big]\\ &\quad + \frac{h}{8}\big((\varphi\rho)''(\hat{x}_{j+3/4}) - (\varphi\rho)''(\hat{x}_{j+1/4})\big)\\ &\quad + \frac{h^2}{8}\big[\sigma(\varphi\rho)'\big]_{j+1/2}\Big[\Big(\frac{1}{\sigma}\Big)''(\hat{x}_{j+1}) + \Big(\frac{1}{\sigma}\Big)''(\hat{x}_j)\Big] \end{aligned}$$

$$+\frac{h^2}{16}\big((\varphi\rho)''(\hat{x}_{j+3/4})+(\varphi\rho)''(\hat{x}_{j+1/4})\big)\Big[\sigma\Big(\frac{1}{\sigma}\Big)'\Big]_{j+1/2}$$
$$+\frac{h^3}{32}\sigma_{j+1/2}\Big[(\varphi\rho)''(\hat{x}_{j+3/4})\Big(\frac{1}{\sigma}\Big)''(\hat{x}_{j+1})-(\varphi\rho)''(\hat{x}_{j+1/4})\Big(\frac{1}{\sigma}\Big)''(\hat{x}_j)\Big].$$

The following inequality shows that the main term in each coefficient c_j is strictly positive for any ρ,σ and φ as in (2.10). Thus, for h small enough, all c_j are strictly positive. We use here the fact that $\rho\geq\rho_\star>0$, $\sigma\geq\sigma_\star>0$ and $|\sigma'|,\ |\rho'|-\rho'\geq 0$:

$$\Big(\frac{\varphi\rho}{\sigma}\Big)'=\frac{\varphi}{\sigma^2}(\psi'\rho\sigma+\rho'\sigma-\rho\sigma')=$$
$$\frac{\varphi}{\sigma^2}\Big[\rho\sigma+\rho\Big(\frac{|\sigma'|\sigma}{\sigma_\star}-\sigma'\Big)+\sigma\Big(\frac{(|\rho'|-\rho')\rho}{\rho_\star}+\rho'\Big)\Big]\geq\frac{\varphi\rho}{\sigma}>0. \tag{3.32}$$

Consequently, we get the following estimate on $R_1-R_2/4$:

$$R_1-\frac{R_2}{4}\leq\frac{C_1^h h^3}{4}\sum_{j=0}^{N-1}\int_0^T|\partial_+^h u_j'(t)|^2\,dt, \tag{3.33}$$

with

$$C_1^h:=\Big|\Big|\sigma\Big(\frac{\varphi\rho}{\sigma}\Big)'\Big|\Big|_{L^\infty}+O(h).$$

Step 7. Estimates on $\mathcal{X}^h_{\varphi\rho}(t)$ and $\mathcal{Y}^h_{\rho,\sigma}(t)$. Observe first that, since $\partial^h u_j=(\partial_+^h u_j+\partial_+^h u_{j-1})/2$, $\mathcal{X}^h_{\varphi\rho}$ in (3.10) can be rewritten as

$$\mathcal{X}^h_{\varphi\rho}(t)=\frac{h}{2}(\varphi\rho)_1 u_1'(t)\partial_+^h u_0(t)+\frac{h}{2}\sum_{j=1}^{N-1}\big((\varphi\rho)_{j+1}u_{j+1}'(t)+(\varphi\rho)_j u_j'(t)\big)\partial_+^h u_j(t).$$

By applying twice the Cauchy–Schwarz inequality, we obtain

$$|\mathcal{X}^h_{\varphi\rho}|\leq\frac{h}{2}\sqrt{\frac{\varphi_1^2\rho_1}{\sigma_{1/2}}}|\sqrt{\rho_1}u_1'||\sqrt{\sigma_{1/2}}\partial_+^h u_0|+$$
$$h\sum_{j=1}^{N-1}\alpha_{j+1/2}\sqrt{\frac{\rho_{j+1}|u_{j+1}'|^2+\rho_j|u_j'|^2}{2}}|\sqrt{\sigma_{j+1/2}}\partial_+^h u_j|$$
$$\leq\frac{h}{2}\sum_{j=0}^{N-1}\beta_{j+1/2}\sigma_{j+1/2}|\partial_+^h u_j|^2+\frac{h}{2}\sum_{j=1}^{N}\beta_j\rho_j|u_j'|^2, \tag{3.34}$$

where

$$\alpha_{j+1/2} := \sqrt{\frac{\varphi_{j+1}^2\rho_{j+1} + \varphi_j^2\rho_j}{2\sigma_{j+1/2}}}, \quad \beta_{j+1/2} := \alpha_{j+1/2}, \quad \beta_{1/2} := \frac{1}{2}\sqrt{\frac{\varphi_1^2\rho_1}{\sigma_{1/2}}}$$

for all $1 \le j \le N-1$, and

$$\beta_1 := \frac{1}{2}\Big(\alpha_{3/2} + \sqrt{\frac{\varphi_1^2\rho_1}{\sigma_{1/2}}}\Big), \quad \beta_j := \frac{\alpha_{j+1/2} + \alpha_{j-1/2}}{2}, \quad \beta_N := \frac{1}{2}\alpha_{N-1/2}$$

for all $2 \le j \le N-1$. We use the following Taylor expansions of $\alpha_{j+1/2}$, the first one around $x_{j+1/2}$ to estimate $\beta_{j+1/2}$ and the second one around x_j to estimate β_j:

$$\alpha_{j+1/2}^2 = \Big(\frac{\varphi^2\rho}{\sigma}\Big)_{j+1/2} + \frac{h^2}{16\sigma_{j+1/2}}\big[(\varphi^2\rho)''(\hat{x}_{j+3/4}) + (\varphi^2\rho)''(\hat{x}_{j+1/4})\big]$$

and

$$\alpha_{j\pm1/2}^2 = \Big(\frac{\varphi^2\rho}{\sigma}\Big)_j \pm \frac{h}{2}(\varphi^2\rho)_j\Big(\frac{1}{\sigma}\Big)'(\hat{x}_{j\pm1/4})\pm$$
$$\frac{h}{2\sigma_j}(\varphi^2\rho)'(\hat{x}_{j\pm1/2}) + \frac{h^2}{4}(\varphi^2\rho)'(\hat{x}_{j\pm1/2})\Big(\frac{1}{\sigma}\Big)'(\hat{x}_{j\pm1/4}).$$

Then taking into account that $\varphi^2\rho/\sigma > 0$ on $(0,1)$, we obtain the following upper bound for all β_j, $1 \le j \le N$, and $\beta_{j+1/2}$, $0 \le j \le N-1$:

$$\beta_{j/2} \le C_2^h := \Big|\Big|\varphi\sqrt{\frac{\rho}{\sigma}}\Big|\Big|_{L^\infty} + O(h) \quad \forall 1 \le j \le 2N. \tag{3.35}$$

Consequently,

$$|\mathcal{X}_{\varphi\rho}^h(t)|_0^T| \le 2C_2^h\mathcal{E}_{\rho,\sigma}^h(\mathbf{u}^{h,0}, \mathbf{u}^{h,1}). \tag{3.36}$$

Remark that $2C_2^h \to T_\star$ as $h \to 0$, with $T_\star$ as in (2.7).

Note that $\mathcal{Y}_{\rho,\sigma}^h(t)$ in (3.21) is of the same nature as $\mathcal{X}_{\varphi\rho}^h(t)$ we just estimated, so that the same kind of techniques will be applied. Indeed, by applying the Cauchy–Schwarz inequality as before, we obtain

$$|\mathcal{Y}_{\rho,\sigma}^h| \le h^2\sqrt{\frac{\gamma_1^2\rho_1\sigma_{1/2}}{h^2}}|\sqrt{\rho_1}u_1'||\sqrt{\sigma_{1/2}}\partial_+^h u_0|$$

$$+ 2h^2 \sum_{j=1}^{N-1} \tilde{\alpha}_{j+1/2} \sqrt{\frac{\rho_{j+1}|u'_{j+1}|^2 + \rho_j |u'_j|^2}{2}} |\sqrt{\sigma_{j+1/2}} \partial_+^h u_j|$$

$$\leq h^2 \sum_{j=0}^{N-1} \tilde{\beta}_{j+1/2} \sigma_{j+1/2} |\partial_+^h u_j|^2 + h^2 \sum_{j=1}^{N} \tilde{\beta}_j \rho_j |u'_j|^2, \tag{3.37}$$

where

$$\tilde{\alpha}_{j+1/2} := \sqrt{\frac{(\gamma_{j+1}^2 \rho_{j+1} + \gamma_j^2 \rho_j)\sigma_{j+1/2}}{2h^2}}, \ \tilde{\beta}_{j+1/2} := \tilde{\alpha}_{j+1/2}, \ \tilde{\beta}_{1/2} := \frac{1}{2} \sqrt{\frac{\gamma_1^2 \rho_1 \sigma_{1/2}}{h^2}}$$

for all $1 \leq j \leq N-1$, and

$$\tilde{\beta}_1 := \frac{1}{2} \Big(\tilde{\alpha}_{3/2} + \sqrt{\frac{\gamma_1^2 \rho_1 \sigma_{1/2}}{h^2}} \Big), \quad \tilde{\beta}_j := \frac{\tilde{\alpha}_{j+1/2} + \tilde{\alpha}_{j-1/2}}{2}, \quad \tilde{\beta}_N := \frac{1}{2} \tilde{\alpha}_{N-1/2}$$

for all $2 \leq j \leq N-1$. All $\tilde{\beta}_{j/2}$, $1 \leq j \leq 2N$, are finite as $h \to 0$ since γ_j in (3.16) approximates $\gamma_j \sim h\varphi_j (1/\sigma)'_j = O(h)$. More precisely, by the same kind of Taylor expansions used above for $\alpha_{j+1/2}$, we can prove that

$$\tilde{\beta}_{j/2} \leq \tilde{C}_2^h := \Big|\Big| \varphi \Big(\frac{1}{\sigma} \Big)' \sqrt{\rho\sigma} \Big|\Big|_{L^\infty} + \begin{cases} O(h), & \text{if } |\sigma'| > 0 \ \forall x \in [0,1] \\ O(\sqrt{h}), & \text{if } |\sigma'| = 0 \text{ at some } x \end{cases} \tag{3.38}$$

for all $1 \leq j \leq 2N$. Finally, we get

$$\frac{1}{4} |\mathcal{Y}_{\rho,\sigma}^h(t)|_0^T| \leq h \tilde{C}_2^h \mathcal{E}_{\rho,\sigma}^h(\mathbf{u}^{h,0}, \mathbf{u}^{h,1}). \tag{3.39}$$

Step 8. Estimate on the remainder term R_4 and on the energy on the boundary. The following identity is obvious:

$$\partial_\sigma^{h,2} (\gamma\rho)_j = \partial^{h,2} (\gamma\rho)_j \frac{\sigma_{j+1/2} + \sigma_{j-1/2}}{2} + \partial^h (\gamma\rho)_j \frac{\sigma_{j+1/2} - \sigma_{j-1/2}}{h},$$

so that

$$|\partial_\sigma^{h,2} (\gamma\rho)_j| \leq \sigma^\star |\partial^{h,2} (\gamma\rho)_j| + ||\sigma'||_{L^\infty} |\partial^h (\gamma\rho)_j| \quad \forall 2 \leq j \leq N-1. \tag{3.40}$$

To estimate the terms $|\partial^{h,2} (\gamma\rho)_j|$ and $|\partial^h (\gamma\rho)_j|$, we first note that, for γ_j as in (3.16), we obtain the following identity:

$$(\gamma\rho)_j = -(\varphi\rho)_j\delta_j, \text{ with } \delta_j := \frac{1}{\sigma_{j+1/2}} - \frac{1}{\sigma_{j-1/2}}$$

and then, using the formula for the discrete Laplacian $\partial^{h,2}$ of the product of the sequences $\varphi\rho$ and δ, we obtain

$$\begin{aligned}\partial^{h,2}(\gamma\rho)_j = &-\partial^{h,2}(\varphi\rho)_j\frac{\delta_{j+1}+2\delta_j+\delta_{j-1}}{4}\\ &-2\partial^h(\varphi\rho)_j\partial^h\delta_j - \partial^{h,2}\delta_j\frac{(\varphi\rho)_{j+1}+2(\varphi\rho)_j+(\varphi\rho)_{j-1}}{4}.\end{aligned}$$

Since δ_j is also a finite difference and, by the hypotheses of Theorem 1, $\sigma \in C^2(0,1)$, we obtain

$$\begin{aligned}\partial^h\delta_j =&\frac{h}{2}\Big[\partial_2^h\Big(\frac{1}{\sigma}\Big)_{j+1/2} + \partial_2^h\Big(\frac{1}{\sigma}\Big)_{j-1/2}\Big],\\ \partial^{h,2}\delta_j =&\partial^{h,2}\Big(\frac{1}{\sigma}\Big)_{j+1/2} - \partial^{h,2}\Big(\frac{1}{\sigma}\Big)_{j-1/2}.\end{aligned}$$

Thus

$$\begin{aligned}|\partial^{h,2}(\gamma\rho)_j| \le\ & 2||\varphi\rho||_{L^\infty}\Big|\Big|\Big(\frac{1}{\sigma}\Big)''\Big|\Big|_{L^\infty}\\ &+h\Big[2||(\varphi\rho)'||_{L^\infty}\Big|\Big|\Big(\frac{1}{\sigma}\Big)''\Big|\Big|_{L^\infty} + ||(\varphi\rho)''||_{L^\infty}\Big|\Big|\Big(\frac{1}{\sigma}\Big)'\Big|\Big|_{L^\infty}\Big].\end{aligned} \tag{3.41}$$

Similarly, to compute the centered derivative for the product of $\varphi\rho$ and δ, we use the identity

$$\partial^h(\gamma\rho)_j = -\partial^h((\varphi\rho)\delta)_j = -\partial^h(\varphi\rho)_j\frac{\delta_{j+1}+\delta_{j-1}}{2} - \partial^h\delta_j\frac{(\varphi\rho)_{j+1}+(\varphi\rho)_{j-1}}{2},$$

so that

$$|\partial^h(\gamma\rho)_j| \le h\Big[||(\varphi\rho)'||_{L^\infty}\Big|\Big|\Big(\frac{1}{\sigma}\Big)'\Big|\Big|_{L^\infty} + ||\varphi\rho||_{L^\infty}\Big|\Big|\Big(\frac{1}{\sigma}\Big)''\Big|\Big|_{L^\infty}\Big]. \tag{3.42}$$

After combining (3.40)–(3.42), we get

$$\frac{|\partial_\sigma^{h,2}(\gamma\rho)_j|}{\rho_j} \le \frac{2\sigma^\star||\varphi\rho||_{L^\infty}}{\rho_\star}\Big|\Big|\Big(\frac{1}{\sigma}\Big)''\Big|\Big|_{L^\infty} + O(h) \quad \forall 2 \le j \le N-1. \tag{3.43}$$

In a similar way, we obtain

$$\frac{\sigma_{3/2}|\partial_+^h(\gamma\rho)_1|}{h\rho_1} \leq \frac{\sigma^\star}{\rho_\star}\Big[||\varphi\rho||_{L^\infty}\Big|\Big|\Big(\frac{1}{\sigma}\Big)''\Big|\Big|_{L^\infty} + ||(\varphi\rho)'||_{L^\infty}\Big|\Big|\Big(\frac{1}{\sigma}\Big)'\Big|\Big|_{L^\infty}\Big] + O(h). \tag{3.44}$$

Finally, we obtain the following estimate of R_4:

$$\frac{|R_4|}{4} \leq hC_3^h T\mathcal{E}_{\rho,\sigma}^h(\mathbf{u}^{h,0},\mathbf{u}^{h,1}), \tag{3.45}$$

$$\text{with } C_3^h := \frac{\sigma^\star}{4\rho_\star}\Big[2||\varphi\rho||_{L^\infty}\Big|\Big|\Big(\frac{1}{\sigma}\Big)''\Big|\Big|_{L^\infty} + ||(\varphi\rho)'||_{L^\infty}\Big|\Big|\Big(\frac{1}{\sigma}\Big)'\Big|\Big|_{L^\infty}\Big] + O(h).$$

Set $C_N := ((\varphi\rho)_{N+1} + (\varphi\rho)_N)/4$ and $C_4^h := ||(\varphi\rho)'||_{L^\infty}/4 + h\rho_\star C_3^h/2$. Similarly to (3.44), we get

$$C_N\int_0^T |u'_{N+1}(t)|^2\,dt + \frac{|R_3|}{4} \leq \Big(\frac{(\varphi\rho)(1)}{2} + hC_4^h\Big)\int_0^T |u'_{N+1}(t)|^2\,dt. \tag{3.46}$$

Replacing the inequalities (3.28), (3.29), (3.33), (3.36), (3.39), (3.45) and (3.46) into the multiplier identity (3.25), we conclude the part a) of Theorem 1, by taking the observability time $T > T_\star^h$ and the observability constant C' in (3.7) as follows:

$$T_\star^h := \frac{2C_2^h + h\tilde{C}_2^h}{C_\star^h - hC_3^h} \text{ and } C' := \frac{\max\Big\{\frac{C_1^h}{4}, \frac{(\varphi\rho)(1)}{2} + hC_4^h\Big\}}{(C_\star^h - hC_3^h)T - (2C_2^h + h\tilde{C}_2^h)}.$$

3.3 Numerical Experiments

As we pointed out in Remark 3, our result in Theorem 1 is also valid in the context of numerical discretizations of the constant coefficients wave equation on non-uniform meshes obtained as diffeomorphic transformations of uniform ones through smooth mappings g. As we know from [8] or [9], our main result in Theorem 1 applied to system (3.8) is also true for the implicit midpoint fully discrete scheme

$$\begin{cases} \frac{V_j^{n+1}-2V_j^n+V_j^{n-1}}{\delta t^2} - \frac{\frac{\partial_+^h V_j^{n+1}}{\partial_+^h g_j} - \frac{\partial_-^h V_j^{n+1}}{\partial_-^h g_j}}{2\partial^h g_j} - \frac{\frac{\partial_+^h V_j^{n-1}}{\partial_+^h g_j} - \frac{\partial_-^h V_j^{n-1}}{\partial_-^h g_j}}{2\partial^h g_j} = \\ \frac{h^2(\partial_2^h V_j^{n+1} - \partial_2^h V_j^{n-1})}{2\delta t\,\partial^h g_j},\ 1 \leq j \leq N \\ v_0^n = \frac{\partial_+^h V_N^{n+1} + \partial_+^h V_N^{n-1}}{2\partial_+^h g_N} + \frac{V_{N+1}^{n+1} - V_{N+1}^{n-1}}{2\delta t} = 0, \\ V_j^0 = v_j^0,\ V_j^1 = V_j^0 + \delta t v_j^1,\ 0 \leq j \leq N+1. \end{cases} \tag{3.47}$$

Here, δt is the time step, V_j^n is an approximation of $v_j(n\delta t)$, $n \geq 0$, and $\mathbf{v}^{h,i} := (v_j^i)_{0\leq j\leq N+1}$ are the initial data in (3.8) ($i = 0, 1$, $v_0^0 = 0$). The total energy of the solutions of (3.47),

$$\tilde{\mathcal{E}}_g^h(\mathbf{V}^{h,n}, \mathbf{V}^{h,n-1}) := \frac{h}{2}\sum_{j=1}^{N} \partial^h g_j \left|\frac{V_j^n - V_j^{n-1}}{\delta t}\right|^2$$

$$+ \frac{h}{4}\sum_{j=0}^{N} \partial_+^h g_j \left(\left|\frac{\partial_+^h V_j^n}{\partial_+^h g_j}\right|^2 + \left|\frac{\partial_+^h V_j^{n-1}}{\partial_+^h g_j}\right|^2\right) + \frac{h^2\partial_+^h g_N}{2}\left(\left|\frac{\partial_+^h V_N^n}{\partial_+^h g_N}\right|^2 + \left|\frac{\partial_+^h V_N^{n-1}}{\partial_+^h g_N}\right|^2\right),$$

satisfies the following dissipation law:

$$\frac{\tilde{\mathcal{E}}_g^h(\mathbf{V}^{h,n+1}, \mathbf{V}^{h,n}) - \tilde{\mathcal{E}}_g^h(\mathbf{V}^{h,n}, \mathbf{V}^{h,n-1})}{\delta t} =$$

$$-\left|\frac{V_{N+1}^{n+1} - V_{N+1}^{n-1}}{2\delta t}\right|^2 - h\sum_{j=0}^{N}\left|\frac{V_{j+1}^{n+1} - V_{j+1}^{n-1}}{2\delta t} - \frac{V_j^{n+1} - V_j^{n-1}}{2\delta t}\right|^2.$$

In the numerical simulations, we will consider

$$v_j^0 = \exp(-\gamma(x_j - g^{-1}(y_0)))\cos(x_j\eta_0/h),$$

$\gamma = h^{-0.9}$, $h = 1/400$, $v_j^1 = 0$, the final time $T = 20$, the time step $\delta t = 0.005$ and $\eta_0 \in \{\pi/5, \pi/3, \pi/2, 2\pi/3\}$. We will take two non-uniform meshes, $g(x) = g_1(x) := \tan(\pi x/4)$ and $g(x) = g_2(x) := 2\sin(\pi x/6)$. The grid given by g_1 is finer close to $x = 0$ and coarser to the endpoint $x = 1$, where the natural damping is acting, while for the grid g_2 the situation is opposite (see Fig. 1).

In Figs. 2 and 4, on the left column, we consider the numerical scheme without the artificial damping given by the right hand side in (3.47) (in that case, we also have to eliminate the last term in the energy $\tilde{\mathcal{E}}_g^h$), while on the right column we plot the quotient of the energies at time $t = n\delta t$ and $t = 0$ for the numerical scheme with the artificial damping. Note that, when the artificial damping term is not added, for both grids, the quotient of energies has a stepped structure which is due to the fact that the energy of solutions is essentially conserved along the rays of Geometric Optics, except for those time instances at which the support of the solution interacts with the endpoint $x = 1$ of the space interval, where the dissipative boundary condition is imposed.

Also remark that the size of the flat areas increases with the frequency, so that the energy of solutions associated to the highest frequency ($\eta_0 = \pi/2, 2\pi/3$) wave packets remains essentially constant until the final time $T = 20$. This is due to the fact that, as predicted in [22], the corresponding solutions of (3.47) remain concentrated along rays that propagate with a negligible group velocity and that, accordingly, do not interact with the dissipative boundary. In Fig. 3e, we observe that the initial wave packet splits into two parts, one going to the left and one to

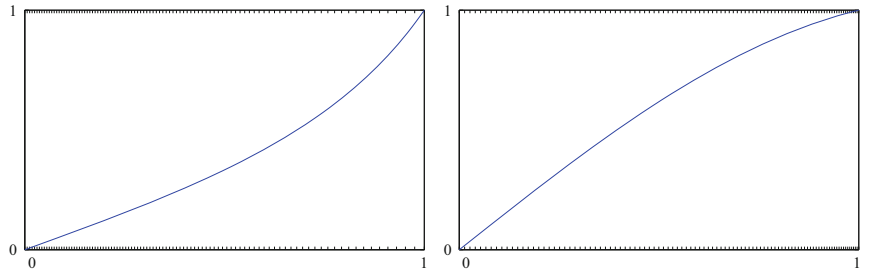


Fig. 1 The two grids obtained through the applications g_1 (*left*) and g_2 (*right*)

the right, and both of them touch the endpoint $x = 1$ once, but in a diffractive manner, so that their behaviour is not influenced by the dissipative boundary. In Fig. 3g, no one of the two wave packets reaches the dissipative boundary, in Fig. 5e, g, the right wave packet touches the endpoint $x = 1$ and is dissipated, while the left wave packet reaches the dissipative boundary approximately at the final time in Fig. 5e and does not reaches the dissipative boundary until the final time in Fig. 5g. On the right columns of Figs. 2, 3, 4, 5 we observe that the damping mechanism is indeed efficient for both grids. Moreover, in the case of added numerical viscosity, the lowest frequencies $\eta_0 = \pi/5, \pi/3$ are dissipated more than exponentially. This is due to the fact that we considered an approximation of the wave equation (1.1) with $\rho = \sigma \equiv 1$, for which the solutions vanish in finite time.

4 Comments and Open Problems

In this paper, we extend to the numerical approximations of the wave equation with regular (C^2) variable coefficients on smooth non-uniform diffeomorphic meshes the results in [26] concerning the efficiency of the vanishing viscosity method to stabilize the numerical schemes for the constant coefficients wave equation on uniform meshes. The method of proof uses multipliers adapted to the variable coefficients as in [4].

We list some open problems related to the content of this article:

- *The stabilization problem for the numerical approximations of the variable coefficients multi-dimensional wave equation.* Extending the results of this paper to the multi-dimensional case is a challenging open problem. Some of the difficulties encountered when doing that are related to the fact that sidewise energy estimates and multipliers do not yield satisfactory results in the continuous context and that they are hard to adapt to multi-dimensional numerical grids.
- *The efficiency of the bi-grid techniques in the stabilization and controllability of the numerical schemes of the wave equation on non-uniform meshes.* Bi-grid algorithm was shown to be useful for observability and control problems (cf. [10]

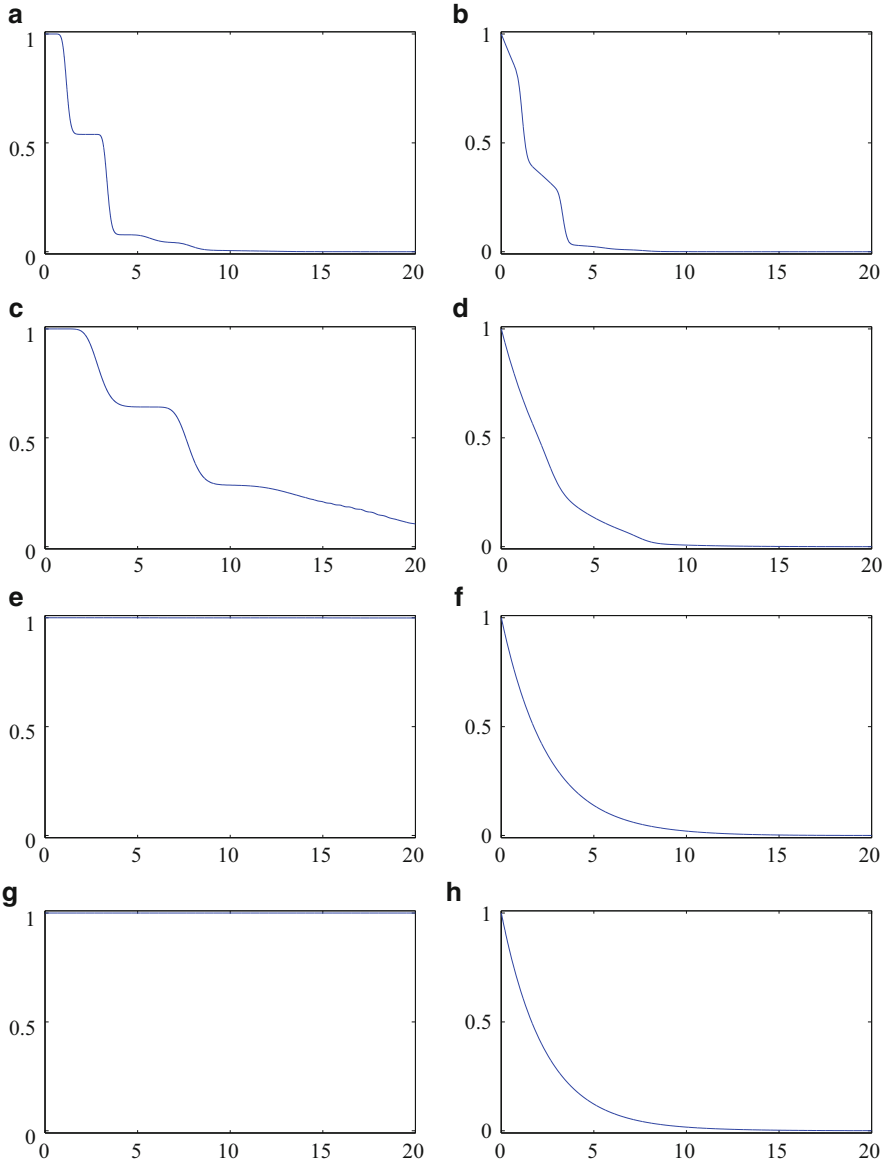


Fig. 2 The quotient between the energy at time $n\delta t$, $\tilde{\mathcal{E}}_g^h(\mathbf{V}^{h,n}, \mathbf{V}^{h,n-1})$, and the energy at the initial time, $\tilde{\mathcal{E}}_g^h(\mathbf{V}^{h,1}, \mathbf{V}^{h,0})$, for the tangential mesh $g_1(x) = \tan(\pi x/4)$. Each row corresponds to a high frequency oscillation $\eta_0 \in \{\pi/5, \pi/3, \pi/2, 2\pi/3\}$ and the *left/right* column to the numerical approximation without/with numerical viscosity. (**a**) $\eta_0 = \pi/5$, no artificial viscosity, (**b**) $\eta_0 = \pi/5$, with artificial viscosity, (**c**) $\eta_0 = \pi/3$, no artificial viscosity, (**d**) $\eta_0 = \pi/3$, with artificial viscosity, (**e**) $\eta_0 = \pi/2$, no artificial viscosity, (**f**) $\eta_0 = \pi/2$, with artificial viscosity, (**g**) $\eta_0 = 2\pi/3$, no artificial viscosity, (**h**) $\eta_0 = 2\pi/3$, with artificial viscosity

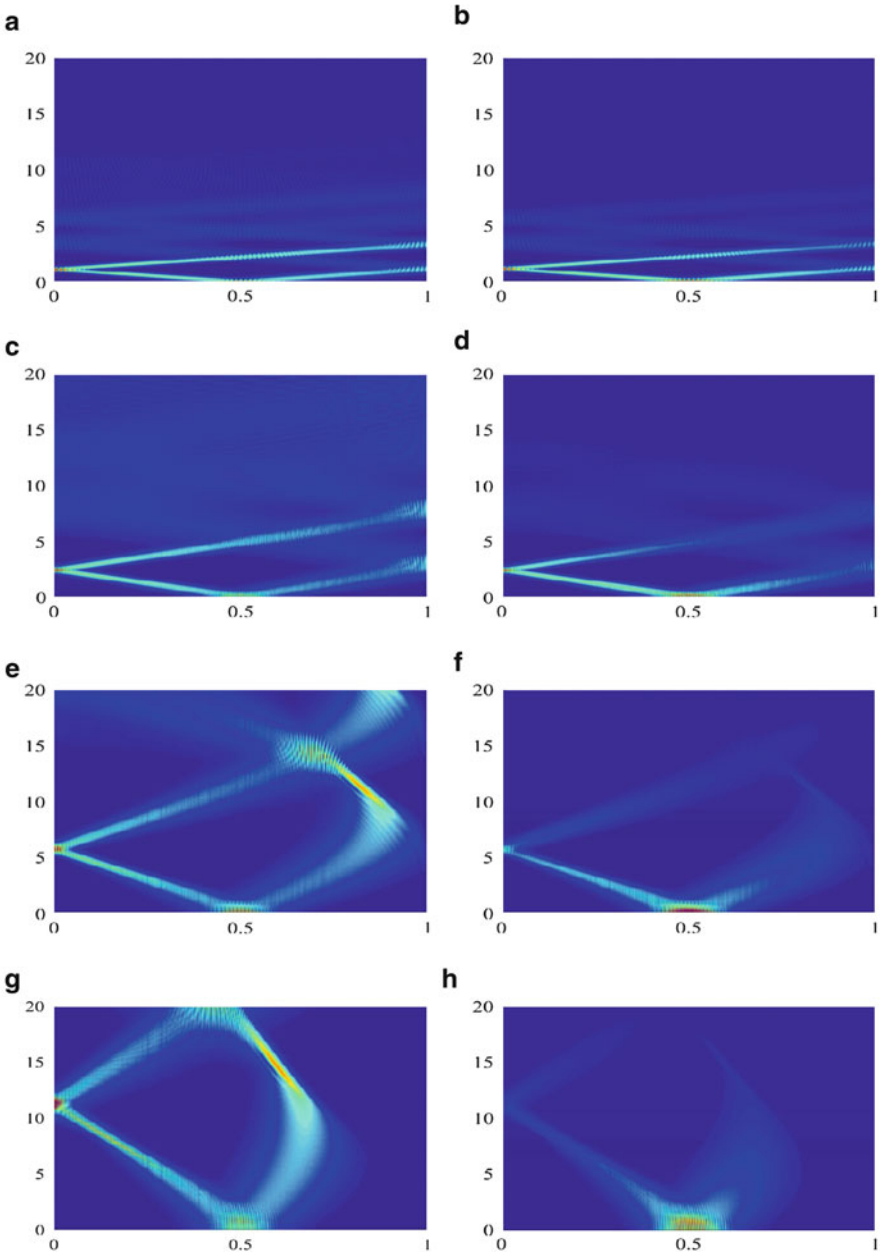


Fig. 3 The solutions of (3.47) on the tangential mesh $g_1(x) = \tan(\pi x/4)$. Each row corresponds to a high frequency oscillation $\eta_0 \in \{\pi/5, \pi/3, \pi/2, 2\pi/3\}$ and the *left/right* column to the numerical approximation without/with numerical viscosity. (**a**) $\eta_0 = \pi/5$, no artificial viscosity, (**b**) $\eta_0 = \pi/5$, with artificial viscosity, (**c**) $\eta_0 = \pi/3$, no artificial viscosity, (**d**) $\eta_0 = \pi/3$, with artificial viscosity, (**e**) $\eta_0 = \pi/2$, no artificial viscosity, (**f**) $\eta_0 = \pi/2$, with artificial viscosity, (**g**) $\eta_0 = 2\pi/3$, no artificial viscosity, (**h**) $\eta_0 = 2\pi/3$, with artificial viscosity

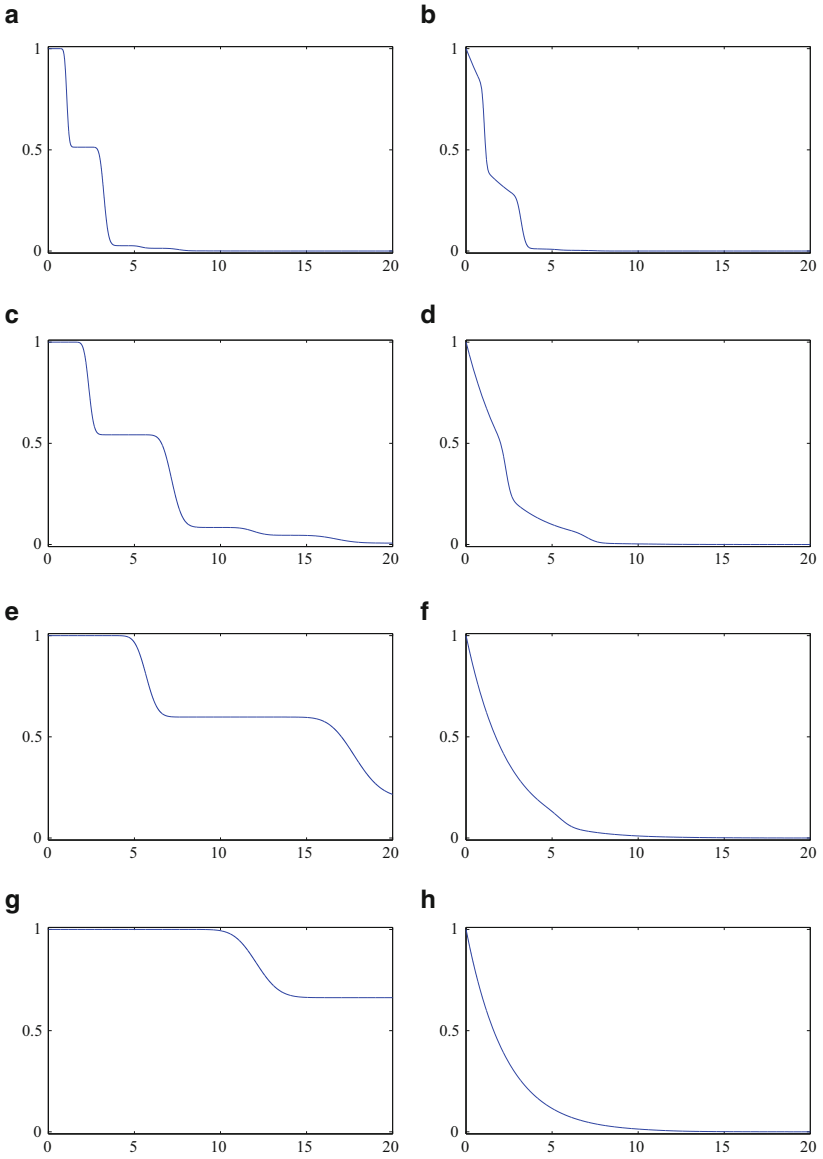


Fig. 4 The quotient between the energy at time $n\delta t$, $\widetilde{\mathcal{E}}_g^h(\mathbf{V}^{h,n}, \mathbf{V}^{h,n-1})$, and the energy at the initial time, $\widetilde{\mathcal{E}}_g^h(\mathbf{V}^{h,1}, \mathbf{V}^{h,0})$, for the sinusoidal mesh $g_2(x) = 2\sin(\pi x/6)$. Each row corresponds to a high frequency oscillation $\eta_0 \in \{\pi/5, \pi/3, \pi/2, 2\pi/3\}$ and the *left/right* column to the numerical approximation without/with numerical viscosity. (**a**) $\eta_0 = \pi/5$, no artificial viscosity, (**b**) $\eta_0 = \pi/5$, with artificial viscosity, (**c**) $\eta_0 = \pi/3$, no artificial viscosity, (**d**) $\eta_0 = \pi/3$, with artificial viscosity, (**e**) $\eta_0 = \pi/2$, no artificial viscosity, (**f**) $\eta_0 = \pi/2$, with artificial viscosity, (**g**) $\eta_0 = 2\pi/3$, no artificial viscosity, (**h**) $\eta_0 = 2\pi/3$, with artificial viscosity

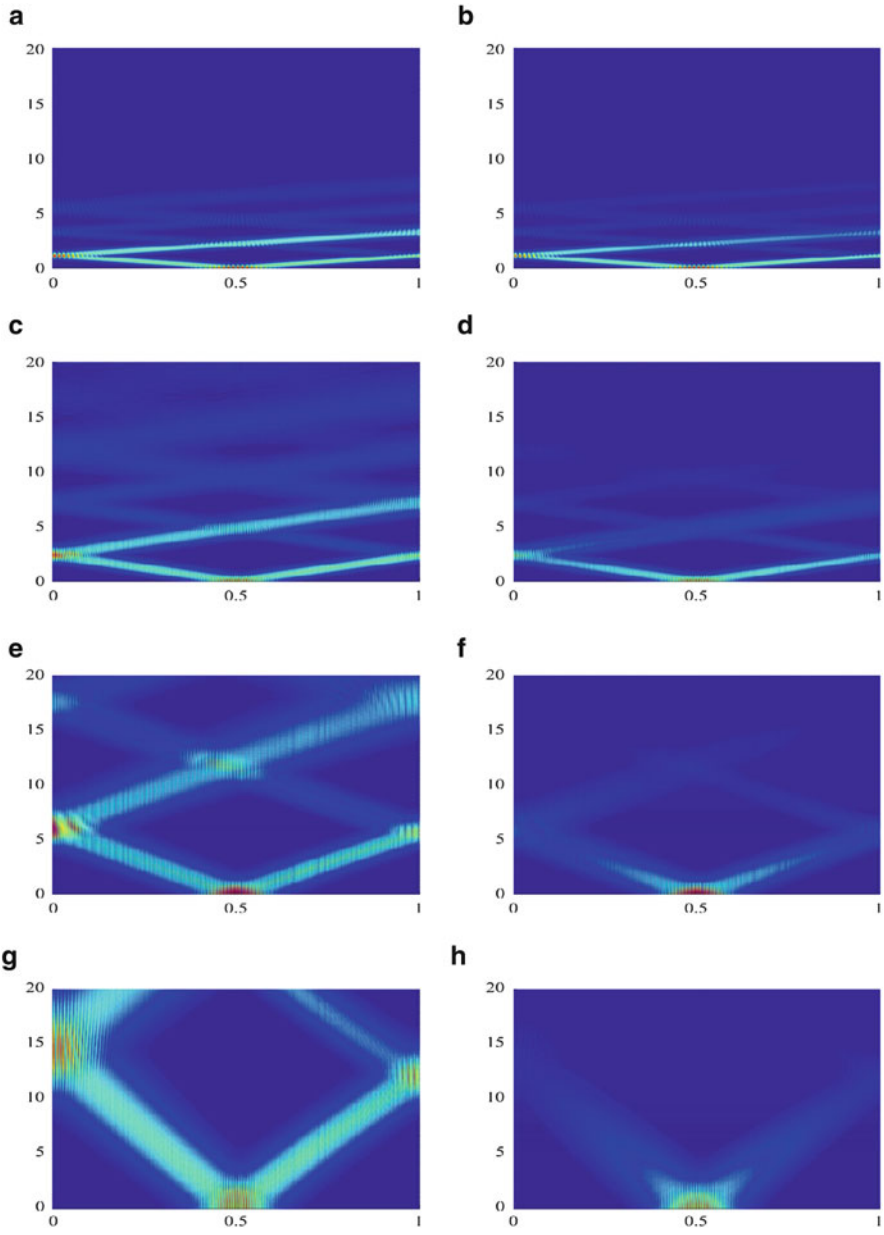


Fig. 5 The solutions of (3.47) on the sinusoidal mesh $g_2(x) = 2\sin(\pi x/6)$. Each row corresponds to a high frequency oscillation $\eta_0 \in \{\pi/5, \pi/3, \pi/2, 2\pi/3\}$ and the *left/right* column to the numerical approximation without/with numerical viscosity. (**a**) $\eta_0 = \pi/5$, no artificial viscosity, (**b**) $\eta_0 = \pi/5$, with artificial viscosity, (**c**) $\eta_0 = \pi/3$, no artificial viscosity, (**d**) $\eta_0 = \pi/3$, with artificial viscosity, (**e**) $\eta_0 = \pi/2$, no artificial viscosity, (**f**) $\eta_0 = \pi/2$, with artificial viscosity, (**g**) $\eta_0 = 2\pi/3$, no artificial viscosity, (**h**) $\eta_0 = 2\pi/3$, with artificial viscosity

and the references therein). But its analysis has been confuted mainly for uniform grids. On the other hand, its use in stabilization problems is to be developed.

- *Construct classes of meshes adapted to the damping or control mechanism*, for which there is no need to filter the high frequency spurious modes to obtain uniform controllability/stabilization properties of the numerical schemes. For instance, in [11], we prove that, when the mesh transformation is strictly concave, the discrete version of the observability inequality for the finite difference semi-discretization of the $1-d$ wave equation with observation on the right endpoint of the space interval holds uniformly with respect to the mesh size parameter. Consequently, on this particular class of concave meshes, a similar uniform result should hold for the boundary stabilization problem from the right endpoint without necessity of adding a filtering mechanism of the high frequency components.
- *Numerics for rough coefficients.* As explained above, the $1-d$ wave equation with BV coefficients can be observed from the boundary, while here we obtained uniform observability properties for the numerical approximation of the wave equation with smoother coefficients ($C^2(0,1)$). It remains to analyze *the uniform stabilization/control properties of the numerical approximations for the wave equation with less regular coefficients*, between $BV(0,1)$ and $C^2(0,1)$.

Appendix A. Some Technical Proofs

Proof of Theorem 2.2. a) **By the method of sidewise energy estimates** as in [13, 16, 27] or [28]. Recall that we work under the assumption that the coefficients and initial data in (2.3) are smooth, i.e., $\rho, \sigma \in C^1(0,1)$ and $(u^0, u^1) \in \mathcal{V}$. For $\varepsilon[u](x,t) := (\rho(x)|u_t(x,t)|^2 + \sigma(x)|u_x(x,t)|^2)/2$ being the energy density of the solution u of (2.3), we define the sideways energy

$$F(x) := \int_{e_-(x)}^{e_+(x)} \varepsilon[u](x,t)\, dt,$$

where $e_+ \geq e_-$ are two functions to be determined. By applying Leibniz formula for differentiation under the integral sign [17], for which $\varepsilon[u]$ must belong to $C((0,1)\times(0,T)) \cap C^1(0,1)$ and $e^{\pm} \in C^1(0,1)$, we obtain

$$F'(x) = A(x) + A_+(x) + A_-(x),$$

where

$$A(x) := \frac{1}{2} \int_{e_-(x)}^{e_+(x)} (\rho'(x)|u_t(x,t)|^2 - \sigma'(x)|u_x(x,t)|^2)\, dt$$

$$A_\pm(x) = \pm (e_\pm)'(x)\varepsilon[u](x, e_\pm(x)) \pm \rho(x)u_t(x, e_\pm(x))u_x(x, e_\pm(x)).$$

By choosing $(e_\pm)' \equiv \pm\sqrt{\rho/\sigma}$, we get

$$A_\pm = 0.5\sqrt{\rho/\sigma}|\sqrt{\rho}u_t(\cdot, e_\pm) \pm \sqrt{\sigma}u_x(\cdot, e_\pm)|^2 \geq 0.$$

Since $A \geq -mF$, with $m := \max\{|\rho'|/\rho, |\sigma'|/\sigma\}$, we get $F'(x) \geq -m(x)F(x)$. Thus, $F(x)\exp(\int_0^x m(\tilde{x})\,d\tilde{x})$ is an increasing function, i.e.,

$$F(x) \leq F(1)\exp\left(\int_x^1 m(\tilde{x})\,d\tilde{x}\right)$$

and, by integration in $x \in (0, 1)$, we get

$$\int_0^1 F(x)\,dx \leq F(1)\int_0^1 \exp\left(\int_x^1 m(\tilde{x})\,d\tilde{x}\right)dx \leq F(1)\exp\left(\int_0^1 m(x)\,dx\right).$$

We choose $e_+(0) = T - \ell$ and $e_-(0) = \ell$, so that $e_+(1) = T$, $e_-(1) = 0$ and $F(1) = \rho(1)\int_0^T |u_t(1, t)|^2\,dt/2$ (since $u_x(1, t) = 0$). In order to ensure the positivity of F, we ask $e_+(0) \geq e_-(0)$, i.e. $T \geq 2\ell$ (and then, since e^+ and e^- are increasing/decreasing, $e^+ > e^-$ for all $x \in (0, 1]$). However, in view of the time conservation of the energy of u, in order to prove the observability inequality (2.4), it is sufficient to ask the existence of a rectangle of the form $(0, 1)\times(t_-, t_+)$, with $t_- < t_+$, included in the curved trapezoidal region $Trap := \{(x, t), x \in (0, 1), e_-(x) < t < e_+(x)\}$ (see the bold curved trapezoidal region in Fig. 6, left). Then $\int_0^1 F(x)\,dx \geq (t_+ - t_-)\mathcal{E}_{\rho,\sigma}(u^0, u^1)$. Of course, the optimal choice of this rectangle is given by $t_\pm = e_\pm(0)$, so that $0 < t_+ - t_- = T - 2\ell$. Since $\exp(\int_0^1 m(x)\,dx) \leq \exp(TV(\rho, 0, 1)/\rho_\star + TV(\sigma, 0, 1)/\sigma_\star)$, (2.4) holds with

$$C' := \frac{\rho(1)\exp(TV(\rho, 0, 1)/\rho_\star + TV(\sigma, 0, 1)/\sigma_\star)}{2(T - 2\ell)}. \tag{A.1}$$

The requirement $\varepsilon[u] \in C^1(0, 1)$ is ensured by the extra regularity assumption imposed on the initial data (u^0, u^1) and on the coefficients. Indeed, when the initial data $(u^0, u^1) \in \mathcal{V}$, the second-order energy $\mathcal{E}^2_{\rho,\sigma}(u(\cdot, t), u_t(\cdot, t))$ below is also time conservative and finite

$$\mathcal{E}^2_{\rho,\sigma}(u(\cdot, t), u_t(\cdot, t)) := \mathcal{E}_{\rho,\sigma}(w(\cdot, t), w_t(\cdot, t)) =$$

$$\frac{1}{2}\int_0^1 \left[\rho(x)\left|\partial^2_{\rho,\sigma}u_t(x, t)\right|^2 + \sigma(x)\left|(\partial^2_{\rho,\sigma}u)_x(x, t)\right|^2\right]dx. \tag{A.2}$$

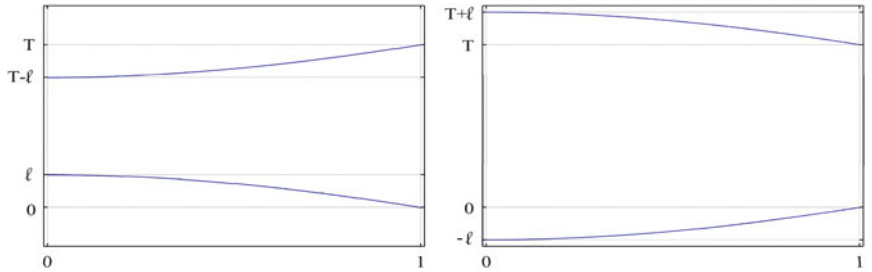


Fig. 6 The two trapezoidal integration domains where the sidewise energy estimates method is applied to prove the observability inequality (*left*) or the direct inequality (*right*)

In particular, there exist two functions $f^0,\ f^1 \in L^\infty(0,T;L^2(0,1))$ such that

$$(\sigma u_{tx})_x = f^1 \text{ and } \left(\frac{1}{\rho}(\sigma u_x)_x\right)_x = f^0.$$

Thus,

$$(\sigma u_x)_x(x,t) = \rho(x)\int_0^x f^0(x',t)\,dx' \in H^1(0,1),$$

$$u_{t,x}(x,t) = \frac{1}{\sigma(x)}\int_0^x f^1(x',t)\,dx' \in H^1(0,1)$$

and, finally, $\sigma u_x, u_t \in H^2(0,1) \subset C^1(0,1)$. Since $\sigma,\rho \in C^1(0,1)$, we get $|\sigma u_x|^2/\sigma,\ \rho|u_t|^2,\ \varepsilon[u] \in C^1(0,1)$.

The direct inequality (2.6) can be obtained by applying the same method of sideways energy estimates within the same class of regular coefficients $\sigma,\rho \in C^1(0,1)$ and initial data $(u^0,u^1) \in \mathcal{V}$ in (2.3). We only choose $e^{\pm}$ such that $(e_\pm)' \equiv \mp\sqrt{\rho/\sigma}$ and the initial data in (2.3) at time $t_0 = -\ell$. Thus, $A_\pm \le 0$ and $F' \le mF$, so that, at the end, (2.6) holds with

$$C'' := \frac{2}{\rho(1)}\exp(TV(\rho,0,1)/\rho_\star + TV(\sigma,0,1)/\sigma_\star)(T+2\ell).$$

By the method of adapted multipliers. Let us consider (2.3) with $(u^0,u^1) \in \mathcal{V}$ and strictly positive coefficients $\rho,\sigma \in C^1(0,1)$. We multiply (2.3) by $\varphi(x)u_x$, with φ as in (2.10), and integrate in $(0,1)\times(0,T)$:

$$0 = \int_0^T \int_0^1 \rho(x) u_{tt}(x,t)\varphi(x) u_x(x,t)\, dx\, dt - \int_0^T \int_0^1 (\sigma u_x)_x(x,t)\varphi(x) u_x(x,t)\, dx\, dt. \tag{A.3}$$

With $\mathcal{X}_{\rho\varphi}(t)$ as in (2.12) and using integrations by parts in both time and space variables in the first term on the right hand side of the above identity, we get:

$$\int_0^T \int_0^1 \rho(x) u_{tt}(x,t)\varphi(x) u_x(x,t)\, dx\, dt = \mathcal{X}_{\rho\varphi}(t)\Big|_0^T - \frac{1}{2}\int_0^T \int_0^1 \rho(x)\varphi(|u_t|^2)_x(x,t)\, dx\, dt$$

$$= \mathcal{X}_{\rho\varphi}(t)\Big|_0^T - \frac{(\varphi\rho)(1)}{2}\int_0^T |u_t(1,t)|^2\, dt + \frac{1}{2}\int_0^T \int_0^1 (\rho\varphi)'(x)|u_t(x,t)|^2\, dx\, dt. \tag{A.4}$$

From the second term in (A.3), we get

$$-\int_0^T \int_0^1 (\sigma u_x)_x(x,t)\varphi(x) u_x(x,t)\, dx\, dt = -\frac{1}{2}\int_0^T \int_0^1 \left(\frac{\varphi}{\sigma}\right)(x)(|\sigma u_x|^2)_x(x,t)\, dx\, dt$$

$$= \frac{(\varphi\sigma)(0)}{2}\int_0^T |u_x(0,t)|^2\, dt + \frac{1}{2}\int_0^T \int_0^1 \left(\frac{\varphi}{\sigma}\right)'(x)\sigma^2(x)|u_x(x,t)|^2\, dx\, dt. \tag{A.5}$$

Putting together the two identities (A.4) and (A.5), we obtain

$$\frac{1}{2}\int_0^T \int_0^1 \left[(\rho\varphi)'(x)|u_t(x,t)|^2 + \left(\frac{\varphi}{\sigma}\right)'(x)\sigma^2(x)|u_x(x,t)|^2\right] dx\, dt +$$

$$+ \frac{(\varphi\sigma)(0)}{2}\int_0^T |u_x(0,t)|^2\, dt = \frac{(\varphi\rho)(1)}{2}\int_0^T |u_t(1,t)|^2\, dt - \mathcal{X}_{\rho\varphi}(t)\Big|_0^T. \tag{A.6}$$

Let us verify that, for φ as in (2.10), the following three inequalities are verified in the sense of $\mathcal{M}(0,1)$, the set of Radon measures on $(0,1)$, which is the dual space of positive $C_c^1(0,1)$ functions:

$$\varphi \geq 1, \quad (\varphi\rho)' \geq \rho \text{ and } \left(\frac{\varphi}{\sigma}\right)'\sigma \geq 1 \quad \forall x \in (0,1). \tag{A.7}$$

Since the derivative of the total variation function $TV(a,0,x)$ is $|a'(x)|$ in a measure sense for any $a \in W^{1,1}(0,1)$, we note that ψ in (2.10) is increasing since

$\psi' = 1 + |\sigma'|/\sigma_\star + (|\rho'| - \rho')/\rho_\star \geq 1$ and $\psi(0) = 0$, so that $\varphi(x) \geq \varphi(0) = 1$. On the other hand, since $\rho/\rho_\star \geq 1$, then

$$(\varphi\rho)' = \varphi(\psi'\rho + \rho') \geq \varphi[\rho + (\rho/\rho_\star)(|\rho'| - \rho') + \rho'] \geq \varphi\rho \geq \rho.$$

Similarly, we get the third inequality in (A.7). Using (A.7) and the time conservation of the total energy for the solution of (2.3), we get the positivity of the second term in the left hand side of (A.6) and the following lower bound on the first term:

$$\frac{1}{2}\int_0^T\int_0^1 \left[(\rho\varphi)'(x)|u_t(x,t)|^2 + \left(\frac{\varphi}{\sigma}\right)'(x)\sigma^2(x)|u_x(x,t)|^2\right] dx\,dt \geq T\mathcal{E}_{\rho,\sigma}(u^0,u^1). \tag{A.8}$$

Using the Cauchy–Schwarz inequality, we get

$$|\mathcal{X}_{\varphi\rho}(t)| \leq ||\varphi\sqrt{\rho/\sigma}||_{L^\infty}\mathcal{E}_{\rho,\sigma}(u^0,u^1), \tag{A.9}$$

so that

$$|\mathcal{X}_{\varphi\rho}(t)|_0^T| \leq 2||\varphi\sqrt{\rho/\sigma}||_{L^\infty}\mathcal{E}_{\rho,\sigma}(u^0,u^1). \tag{A.10}$$

Combining identity (A.6) with the inequalities (A.8) and (A.10), we get

$$(T - 2||\varphi\sqrt{\rho/\sigma}||_{L^\infty})\mathcal{E}_{\rho,\sigma}(u^0,u^1) \leq \frac{(\varphi\rho)(1)}{2}\int_0^T |u_t(1,t)|^2\,dt, \tag{A.11}$$

which concludes the part a) of our result.

c) Let us prove now that (2.4) implies (2.2). We argue by means of a decomposition argument, i.e. we consider (2.3) with the same initial data (v^0, v^1) as in (1.1) and the following problem satisfied by the difference $z = v - u$:

$$\begin{cases} \rho(x)z_{tt} - (\sigma(x)z_x)_x = 0, & x \in (0,1), t \in (0,T] \\ z(0,t) = \sigma(1)z_x(1,t) + v_t(1,t) = 0, & t \in [0,T] \\ z(x,0) = z_t(x,0) = 0, & x \in (0,1). \end{cases} \tag{A.12}$$

From (2.4) and the fact that $u = v - z$, we get

$$\mathcal{E}_{\rho,\sigma}(v^0,v^1) \leq 2C'\int_0^T |v_t(1,t)|^2\,dt + 2C'\int_0^T |z_t(1,t)|^2\,dt. \tag{A.13}$$

It is enough to prove that

$$\int_0^T |z_t(1,t)|^2\,dt \le C'' \int_0^T |v_t(1,t)|^2\,dt. \tag{A.14}$$

To obtain (A.14), we first multiply (A.12) by z_t, integrate in $x \in (0,1)$ and from 0 to t, with $t \in (0,T)$, and, taking into account that the initial data in (A.12) is the trivial one, we get the identity

$$\mathcal{E}_{\rho,\sigma}(z(\cdot,t), z_t(\cdot,t)) = -\int_0^t z_t(1,t')v_t(1,t')\,dt'. \tag{A.15}$$

The second step is to use sideways energy estimates. More precisely, set $G(x) := \int_0^T \varepsilon[z](x,t)\,dt$, where $\varepsilon[z] := (\rho|z_t|^2 + \sigma|z_x|^2)/2$ is the energy density of the solution of (A.12). Then, since the initial data in (A.12) are the trivial ones, we get

$$G'(x) = \frac{1}{2}\int_0^T (\rho'(x)|z_t(x,t)|^2 - \sigma'(x)|z_x(x,t)|^2)\,dt + \rho(x)z_t(x,T)z_x(x,T).$$

For $m(x) := \max\{|\rho'(x)|/\rho(x), |\sigma'(x)|/\sigma(x)\}$, from the previous identity we obtain

$$G'(x) \le m(x)G(x) + \rho(x)|z_t(x,T)||z_x(x,T)| \tag{A.16}$$

and

$$G(1) \le \exp\Big(\int_0^1 m(x)\,dx\Big)\Big(G(x) + ||\sqrt{\rho/\sigma}||_{L^\infty}\mathcal{E}_{\rho,\sigma}(z(\cdot,T), z_t(\cdot,T))\Big). \tag{A.17}$$

After integrating inequality (A.17) in $x \in (0,1)$, we get

$$\begin{aligned} G(1) \le\ & \exp\Big(\int_0^1 m(x)\,dx\Big)\Big(\int_0^T \mathcal{E}_{\rho,\sigma}(z(\cdot,t), z_t(\cdot,t))\,dt \\ & + ||\sqrt{\rho/\sigma}||_{L^\infty}\mathcal{E}_{\rho,\sigma}(z(\cdot,T), z_t(\cdot,T))\Big). \end{aligned} \tag{A.18}$$

Note that, by using the boundary condition at $x = 1$ in (A.12), we obtain

$$G(1) = \frac{\rho(1)}{2}\int_0^T |z_t(1,t)|^2\,dt + \frac{1}{2\sigma(1)}\int_0^T |v_t(1,t)|^2\,dt. \tag{A.19}$$

Using the fact that

$$\exp\left(\int_0^1 m(x)\,dx\right) \le \alpha := \exp\left(TV(\rho,0,1)/\rho_\star + TV(\sigma,0,1)/\sigma_\star\right)$$

and replacing (A.15) in (A.18), we get

$$G(1) \le -\alpha\Big[\int_0^T\int_0^t z_t(1,t')v_t(1,t')\,dt'\,dt + ||\sqrt{\rho/\sigma}||_{L^\infty}\int_0^T z_t(1,t)v_t(1,t)\,dt\Big]. \tag{A.20}$$

By applying Cauchy–Schwarz inequality in (A.20), we get

$$G(1) \le \frac{c\epsilon}{2}\int_0^T |z_t(1,t)|^2\,dt + \frac{c}{2\epsilon}\int_0^T |v_t(1,t)|^2\,dt \quad \forall \epsilon > 0, \tag{A.21}$$

with $c := \alpha(T + ||\sqrt{\rho/\sigma}||_{L^\infty})$.

Choosing $\epsilon = \rho(1)/(2c)$ in (A.21) and taking into account (A.19), we obtain (A.14) with

$$C'' := \frac{4}{\rho^2(1)}\Big(c^2 - \frac{1}{2}\frac{\rho(1)}{\sigma(1)}\Big) > \frac{2||\sqrt{\rho/\sigma}||^2_{L^\infty}}{\rho^2(1)} > 0.$$

To prove that (2.2) implies (2.4), we argue similarly. That is, we consider the same initial data (u^0, u^1) in both problems (1.1) and (2.3) and use the same decomposition $v = u + z$ as in the direct implication. It is enough to prove the estimate

$$\int_0^T |z_t(1,t)|^2\,dt \le C''\int_0^T |u_t(1,t)|^2\,dt. \tag{A.22}$$

Since $v = u + z$, (A.20) becomes

$$\Big(\frac{\rho(1)}{2} + \frac{1}{2\sigma(1)} + \alpha||\sqrt{\rho/\sigma}||_{L^\infty}\Big)\int_0^T |z_t(1,t)|^2\,dt + \frac{1}{2\sigma(1)}\int_0^T |u_t(1,t)|^2\,dt$$

$$+\alpha\int_0^T\int_0^t |z_t(1,t')|^2\,dt'\,dt \le -\frac{1}{\sigma(1)}\int_0^T z_t(1,t)u_t(1,t)\,dt-$$

$$-\alpha\Big(\int_0^T\int_0^t z_t(1,t')u_t(1,t')\,dt'\,dt + ||\sqrt{\rho/\sigma}||_{L^\infty}\int_0^T z_t(1,t)u_t(1,t)\,dt\Big).$$

Then, for all $\epsilon > 0$ and $c = 1/\sigma(1) + \alpha(T + ||\sqrt{\rho/\sigma}||_{L^\infty})$, we get

$$\Big(\frac{\rho(1)}{2} + \frac{1}{2\sigma(1)} + \alpha||\sqrt{\rho/\sigma}||_{L^\infty}\Big)\int_0^T |z_t(1,t)|^2\,dt + \frac{1}{2\sigma(1)}\int_0^T |u_t(1,t)|^2\,dt$$

$$\le \frac{c\epsilon}{2}\int_0^T |z_t(1,t)|^2\,dt + \frac{c}{2\epsilon}\int_0^T |v_t(1,t)|^2\,dt.$$

By taking $\epsilon = (\rho(1)/2 + 1/(2\sigma(1)) + \alpha||\sqrt{\rho/\sigma}||_{L^\infty})/c$ in the above inequality, we obtain (A.22) with

$$C'' := \frac{4\Big(c^2 - \frac{\rho(1)+1/\sigma(1)+2\alpha||\sqrt{\rho/\sigma}||_{L^\infty}}{2\sigma(1)}\Big)}{(\rho(1) + 1/\sigma(1) + 2\alpha||\sqrt{\rho/\sigma}||_{L^\infty})^2} > 0.$$

To prove that (2.2) implies (1.3), first observe that, due to the dissipation law (1.2) and to (2.2), we get

$$\mathcal{E}_{\rho,\sigma}(v(\cdot,T),v_t(\cdot,T)) \le C\int_0^T |v_t(1,t)|^2\,dt = C(\mathcal{E}_{\rho,\sigma}(v^0,v^1) - \mathcal{E}_{\rho,\sigma}(v(\cdot,T),v_t(\cdot,T))),$$

so that

$$\mathcal{E}_{\rho,\sigma}(v(\cdot,T),v_t(\cdot,T)) \le \gamma\mathcal{E}_{\rho,\sigma}(v(\cdot,0),v_t(\cdot,0)), \quad \gamma := \frac{C}{C+1} \in (0,1) \qquad \text{(A.23)}$$

and we obtain (1.3) with $M := 1/\gamma$ and $\omega := \ln(1/\gamma)/T$.

To prove that (1.3) implies (2.2), we combine (1.3) with the dissipation law (1.2) and chose T such that $M\exp(-\omega T) < 1$. □

Acknowledgements This work is supported by the Advanced Grant NUMERIWAVES/FP7-246775 of the European Research Council Executive Agency, FA9550-14-1-0214 of the EOARD-AFOSR, MTM2011-29306-C02-00 and SEV-2013-0323 Grants of the MINECO Spain, the PI2010-04 Grant and the BERC 2014-2017 Program of the Basque Government. The details of this work were elaborated during the postdoc stay of the first author at Institute of Mathematics and Scientific Computing and University of Graz, funded by the MOBIS - Mathematical

Optimization and Applications in Biomedical Sciences grant of the FWF - Austrian Science Fund. Additionally, the first author's work was supported by two grants of the Romanian Ministry of National Education (CNCS-UEFISCDI), i.e., projects PN-II-ID-PCE-2012-4-0021 *Variable Exponent Analysis: Partial Differential Equations and Calculus of Variations* and PN-II-ID-PCE-2011-3-0075 *Analysis, control and numerical approximations of PDEs*. Both authors thank the CIMI - Toulouse for the hospitality and support during the preparation of this work in the context of the Excellence Chair in *PDE, Control and Numerics*.

References

[1] C. Bardos, G. Lebeau, J. Rauch, Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary. SIAM J. Control Optim. **30**, 1024–1065 (1992)
[2] B. Beckermann, S. Serra-Capizzano, On the asymptotic spectrum of finite element matrix sequences. SIAM J. Numer. Anal. **45**, 746–769 (2007)
[3] C. Castro, E. Zuazua, Concentration and lack of observability of waves in highly heterogeneous media. Arch. Ration. Mech. Anal. **164**, 39–72 (2002)
[4] F. Conrad, J. Leblond, J.-P. Marmorat, Boundary control and stabilization of the one-dimensional wave equation, in *Boundary Control and Boundary Variation, Proceedings of IFIP WG 7.2 Conference*, Sophia-Antipolis, France, 15–17 October 1990, ed. by J.P. Zolésio (1990), pp. 142–162
[5] S. Cox, E. Zuazua, The rate at which the energy decays in a string damped at one end. Indiana Univ. Math. J. **44**, 545–573 (1995)
[6] S. Ervedoza, On the mixed finite element method for the $1-d$ wave equation on non-uniform meshes. ESAIM COCV **2**, 298–326 (2010)
[7] S. Ervedoza, E. Zuazua, On the perfectly matched layers: energy decay for $1-d$ continuous and semi-discrete waves. Numer. Math. **109**, 597–634 (2008)
[8] S. Ervedoza, E. Zuazua, Uniformly exponentially stable approximations for a class of damped systems. J. Math. Pures Appl. **91**, 20–48 (2009)
[9] S. Ervedoza, E. Zuazua, Uniform exponential decay for viscous damped systems, in *Advances in Phase Space Analysis of Partial Differential Equations, Progress in Nonlinear Differential Equations and their Applications*, vol. 78. (Birkhäuser Boston, 2009), pp. 95–112
[10] S. Ervedoza, E. Zuazua, The wave equation: control and numerics, in *Control and Stabilization of PDEs*, eds. by P.M. Cannarsa, J.M. Coron. Lecture Notes in Mathematics, vol. 2048. (Springer, Berlin/Heidelberg/New York, 2012), pp. 245–339
[11] S. Ervedoza, A. Marica, E. Zuazua, Uniform observability property for discrete waves on strictly concave non-uniform meshes: a multiplier approach (in preparation)
[12] F. Fanelli, E. Zuazua, Weak observability estimates for 1-D wave equations with rough coefficients, to appear in Annales de l'Institut Henri Poincare (C) Non Linear Analysis
[13] E. Fernández-Cara, E. Zuazua, On the null controllability of the one-dimensional heat equation with BV coefficients. Special issue in memory of Jacques-Louis Lions. Comput. Appl. Math. **21**, 167–190 (2002)
[14] S.V. Fomin, A.N. Kolmogorov, *Introductory Real Analysis*. (Dover Publications, New York, 1970)
[15] R. Glowinski, C.H. Li, J.-L. Lions, A numerical approach to the exact boundary controllability of the wave equation (I). Dirichlet controls: Description of the numerical methods. Jpn. J. Appl. Math. **7**, 1–76 (1990)
[16] A. Haraux, A generalized internal control for the wave equation in a rectangle. J. Math. Anal. Appl. **153**, 190–216 (1990)
[17] F. Harley, Differentiation under the integral sign. Am. Math. Mon. **80**, 615–627 (1973)
[18] V. Komornik, E. Zuazua, A direct method for the boundary stabilization of the wave equation. J. Math. Pures Appl. **69**, 33–54 (1990)

[19] J. Lagnese, Control of wave processes with distributed controls supported on a subregion. SIAM J. Cont. Optim. **21**, 68–85 (1993)
[20] J.-L. Lions, *Contrôlabilité Exacte, Perturbations et Stabilisation des Systèmes Distribués*, vol. 1. (Masson, Paris, 1988)
[21] A. López, E. Zuazua, Uniform null controllability for the one dimensional heat equation with rapidly oscillating periodic density. Ann. l'Institut Henri Poincare Anal Non Linéaire **19**, 543–580 (2002)
[22] A. Marica, E. Zuazua, Propagation of $1-d$ waves in regular discrete heterogeneous media: A Wigner measure approach. Accepted in FoCM - Foundations of Computational Mathematics
[23] S. Micu, Uniform boundary controllability of a semi-discrete $1-d$ wave equation with vanishing viscosity. SIAM J. Cont. Optim. **47**, 2857–2885 (2008)
[24] R.T. Rockafellar, *Convex Analysis*. (Princeton University Press, Princeton, 1970)
[25] L.R.T. Tébou, A Carleman estimates based approach for the stabilization of some locally damped semilinear hyperbolic equations. ESAIM COCV **14**, 561–574 (2008)
[26] L.R.T. Tébou, E. Zuazua, Uniform boundary stabilization of the finite difference space discretization of the $1-d$ wave equation. Adv. Comput. Math. **26**, 337–365 (2007)
[27] J. Vancostenoble, Stabilisation non monotone de systèmes vibrants et contrôlabilité, Ph.D. thesis, Université de Rennes I, 1998
[28] E. Zuazua, *An introduction to the exact controllability for distributed systems*. Textos e Notas, vol. 44. C.M.A.F. (Universidades de Lisboa, Lisboa, 1990)
[29] E. Zuazua, Exponential decay for the semilinear wave equation with locally distributed damping. Commun. Partial Differ. Equ. **15**, 205–235 (1990)
[30] E. Zuazua, Exponential decay for the semilinear wave equation with localized damping in unbounded domains. J. Math. Pures Appl. **70**, 513–529 (1991)
[31] E. Zuazua, Propagation, observation, control and numerical approximation of waves. SIAM Rev. **47**, 197–243 (2005)

Two-Sided Guaranteed Estimates of the Cost Functional for Optimal Control Problems with Elliptic State Equations

Pekka Neittaanmäki and Sergey Repin

Abstract In the paper, we discuss error estimation methods for optimal control problems with distributed control functions entering the right-hand side of the corresponding elliptic state equations. Our analysis is based on a posteriori error estimates of the functional type, which were derived in the last decade for many boundary value problems. They provide guaranteed two-sided bounds of approximation errors for any conforming approximation. If they are applied to approximate solutions of state equations, then we obtain new variational formulations of optimal control problems and guaranteed bounds of the cost functional. Moreover, for problems with linear state equations this procedure leads to guaranteed and computable error estimates for the state and control functions.

Keywords A posteriori error estimates • Elliptic boundary value problems • Guaranteed error bounds • Optimal control problems

Mathematics Subject Classification (2010). Primary 65K15; Secondary 49M99, 65K15.

1 Introduction

Optimal control problems with distributed control arises in many scientific and industrial problems. The corresponding mathematical theory is well developed (see, e.g., [9] and for numerical methods [1,16]). In the majority of cases, these problems can be stated in the following abstract form. Consider a functional

$$J(\eta, \mathsf{v}) : \Xi \times U \to \mathbb{R},$$

P. Neittaanmäki • S. Repin (✉)
Department Mathematical Information Technology, University of Jyväskylä, P.O. Box 35 (Agora), FI-40014, Jyväskylä, Finland
e-mail: pekka.neittaanmaki@jyu.fi; sergey.repin@mit.jyu.fi

R. Hoppe (ed.), *Optimization with PDE Constraints*, Lecture Notes in Computational Science and Engineering 101,
DOI 10.1007/978-3-319-08025-3_10

where Ξ and U are reflexive Banach spaces (associated with state and control functions, respectively). The goal is to find $\mathsf{u} \in K_V \subset U$ such that

$$J(\eta_{\mathsf{u}}, \mathsf{u}) = \inf J := \inf_{\mathsf{v} \in K_U} J(\eta_{\mathsf{v}}, \mathsf{v}), \tag{1.1}$$

where K_U is a closed set of admissible control functions and $\eta_{\mathsf{v}} \in \Xi$ solves the problem

$$\mathcal{A}(\eta_{\mathsf{v}}, \mathsf{v}) = 0. \tag{1.2}$$

Here, $\mathcal{A}$ is a certain (linear or nonlinear) differential operator. It is assumed that the problem (1.2) is well posed and the cost functional J is bounded from below and continuous with respect to both variables.

We consider a subclass of optimal control problems, in which the control function v enters the source term of the equation, i.e., problems of the type

$$\mathcal{A}(\eta_{\mathsf{v}}) = \mathsf{v} + f, \tag{1.3}$$

where f is a given function in the image space of the operator $\mathcal{A}$. In shape optimization problems and problems of topological (structural) optimization the function v may also enter the differential operator. Then, existence of a solution is not guaranteed and a special closure of the respective operator set (so-called G-closure) may be necessary to obtain a mathematically correct statement. Here, we do not consider this class of problems and refer to, e.g., [10,15–18] where the reader can find a consequent exposition and numerous references.

In the simplest case, (1.3) is generated by the linear boundary value problem: Find

$$\eta_{\mathsf{v}} \in V_0(\Omega) := \{\eta \in H^1(\Omega) \mid \eta = 0 \text{ on } \Gamma\}$$

such that

$$\int_\Omega A \nabla \eta_{\mathsf{v}} \cdot \nabla w \, dx = \int_\Omega (\mathsf{v} + f) w \, dx \quad \forall w \in V_0(\Omega). \tag{1.4}$$

Here Ω is a bounded connected domain in $\mathbb{R}^d$ with Lipschitz boundary Γ, A is a symmetric positive definite matrix such that $\nu_1 |\xi|^2 \le A\xi \cdot \xi \le \nu_2 |\xi|^2$, and the functions v and f belong to $L^2(\Omega)$. Let $\sigma(\eta) := A \nabla \eta$ be the flux associated with η. For this problem, we consider integral type cost functionals

$$J_1(\eta, \mathsf{v}) := \frac{1}{2} \left\| \sigma(\eta) - \sigma^d \right\|^2_{A^{-1}} + \frac{a}{2} \left\| \mathsf{v} - \mathsf{u}^d \right\|^2, \quad a > 0. \tag{1.5}$$

and

$$J_2(\eta, \mathsf{v}) := \frac{1}{2} \left\| \eta - \eta^d \right\|^2 + \frac{a}{2} \left\| \mathsf{v} - \mathsf{u}^d \right\|^2. \tag{1.6}$$

Here $\| \cdot \|$ denotes the norm of L^2 (since no confusion may arise we use the same notation for scalar and vector valued functions), u^d, η^d, and $\sigma^d \in L^2(\Omega, \mathbb{R}^d)$ are given functions representing the desired flux and the control function, respectively. In this case, $\Xi = V_0(\Omega)$ and $U = L^2(\Omega)$.

In more complicated cases, $\mathcal{A}$ can be represented by a nonlinear boundary value problem (e.g., by a variational inequality).

We consider a class of optimal control problems, in which the set of admissible control functions includes constraints, namely,

$$K_U := \left\{ \mathsf{v} \in L^2(\Omega) \,|\, \mathsf{v} \le \mathsf{v}^+ \text{ a.e. in} \Omega \right\}, \quad \mathsf{v}^+ \in L^\infty(\Omega) \tag{1.7}$$

It is well known that under the above assumptions Problems 1 and 2 have unique solutions (e.g., see [9]).

Approximation methods, a priori and a posteriori estimates, and adaptive numerical methods were intensively studied in the last decade. In this short note, we cannot present a consequent overview of these results and refer to [2–5, 7, 8, 12] and other publications cited in these papers.

Our goal and the corresponding mathematical approach are different. They are motivated by specific features and difficulties related numerical analysis of optimal control problems of the considered type. One of them comes from the fact that the set of admissible pairs η_v is the exact solution of a boundary value problem. In general this function is unknown and, therefore, the value of $J(\mathsf{v}, \eta_v v)$ is difficult to compute. This fact makes optimal control problems more complicated than classical variational problems, in which convex functionals are explicitly defined and minimized on convex sets of admissible functions.

Also, it is worth outlining a specific feature of optimal control problems, which makes them rather different with respect to variational problems generated by elliptic type equations. The latter problems are focused on finding the minimizer, which coincides with the solution of a boundary value problem. For this reason, we need to find an approximation close in the corresponding energy space. In optimal control problems we are mainly interested in the function u while η_v (solution of the differential problem) plays a subsidiary role. Moreover, from the practical point of view it is often enough to find an ϵ-solution $\mathsf{v}_\epsilon \in K_U$ such that

$$J(\mathsf{v}_\epsilon, \eta_{\mathsf{v}_\epsilon}) \le \inf J + \epsilon, \tag{1.8}$$

where ϵ is a small positive number. Indeed, if we can guarantee that a control function v_ϵ generates a value of the cost functional, which is very close to the best possible, then v_ϵ can be efficiently used instead of u (even if v_ϵ is not close to u in U). In other words, even if the best function u is unique, it may happen that a wide variety of "almost optimal" control functions provide practically the same value of

the cost functional. If v_ϵ is much simpler than u (e.g., if u is a complicated function and v_ϵ is a piecewise constant function), then using v_ϵ may be preferable from the practical point of view.

All said above leads us to the following important question: *How to guarantee that an approximate control function found by a certain numerical procedure provides the value of the cost functional, which is indeed close to* $\inf J$*?*.

In this paper, we discuss methods able to give an adequate answer to this question. With the paradigm of the problem (1.5)–(1.7), we show a way to deduce such estimates. Moreover, we show that they generate guaranteed bounds of errors associated with the state and control functions.

The key mathematical tools used to establish two-sided and guaranteed bounds of cost functionals comes from the theory of *functional type a posteriori error estimates*, which provides a guaranteed bound of the difference between the exact solution of a boundary value problem and any conforming approximation from the corresponding energy class (see [14, 20–23]). In terms of (1.1)–(1.2) [and a special class of problems presented by (1.5)–(1.7)], these estimates reads as follows:

$$\mathsf{M}^{-}(\eta, \mathsf{v}, \mathcal{D}) \le \|\eta_{\mathsf{v}} - \eta\|_{\Xi} \le \mathsf{M}^{+}(\eta, \mathsf{v}, \mathcal{D}). \tag{1.9}$$

Here $\mathsf{M}^{-}(\eta, \mathsf{v}, \mathcal{D})$ and M^{+} are explicitly computable functionals. They depend on the control function v, the corresponding approximate solution η, and other explicitly known data $\mathcal{D}$. In the last decade, estimates (1.9) has been derived for many problems generated by elliptic and parabolic differential equations (a consequent exposition of the mathematical theory is presented in [23] and the book [11] is focused on the corresponding numerical methods and algorithms).

For example, for the problem ($g \in L^2(\Omega)$)

$$\operatorname{div} A\nabla\eta_g + g = 0 \quad \text{in } \Omega, \quad \eta_g = \mu \text{ on } \Gamma \tag{1.10}$$

it was established that the difference between exact solution η_f and any conforming approximation $\eta \in H^1(\Omega)$ is controlled by the following estimates (see [20,21,23]).

Let $\eta \in \overset{\circ}{H}{}^1(\Omega) + \mu$ be an approximation of η_g, which satisfies the Dirichlet boundary condition. Then,

$$\big\|\nabla(\eta - \eta_g)\big\|_A \le \|A\nabla\eta - \tau\|_{A^{-1}} + C_{F\Omega}\|\mathrm{r}(\tau)\|, \tag{1.11}$$

where τ is an arbitrary vector-valued function in $H(\Omega, \operatorname{div})$,

$$\mathrm{r}(\tau) := \operatorname{div}\tau + f,$$

$$\|\tau\|_A^2 := \int_\Omega A\tau \cdot \tau \, dx, \quad \|\tau\|_{A^{-1}}^2 := \int_\Omega A^{-1}\tau \cdot \tau \, dx \qquad \forall \tau \in L^2(\Omega, \mathbb{R}^d),$$

and $C_{F\Omega}$ is the constant in the Friedrichs inequality (for functions vanishing at the boundary).

If Ω is divided into a collection of N nonoverlapping subdomains Ω_i and τ additionally satisfies the relations

$$\int_{\Omega_i} (\mathrm{div}\tau + f)dx = 0, \qquad i = 1, 2, \ldots, N \tag{1.12}$$

then we have a modified form of (1.11), in which the global constant $C_{F\Omega}$ is replaced by a set of local constants

$$\left\|\nabla(\eta - \eta_g)\right\|_A \le \|A\nabla\eta - \tau\|_{A^{-1}} + \sqrt{\sum_{i=1}^{N} C^2_{P\Omega_i} \|\mathrm{r}(\tau)\|^2_{\Omega_i}}, \tag{1.13}$$

where $C_{P\Omega_i}$ are constant in the Poincaré inequalities for Ω_i.

For a convex Ω_i, we know that $C_{P\Omega_i} \le \frac{\mathrm{diam}\Omega_i}{\pi}$ (see [19]), Then, (1.13) implies the estimate

$$\left\|\nabla(\eta - \eta_g)\right\|_A \le \|A\nabla\eta - \tau\|_{A^{-1}} + C_P \|\mathrm{r}(\tau)\|_\Omega, \tag{1.14}$$

where

$$C_P := \frac{1}{\pi} \max_i \{\mathrm{diam}\Omega_i\}.$$

We know (see [23]) that the estimate (1.14) is also valid for problems with mixed boundary conditions.

Estimates (1.11)–(1.14) can be applied to the state equation. For example, if we apply (1.11) to (1.4), then we obtain

$$\left\|\nabla(\eta - \eta_g)\right\|_A \le \|A\nabla\eta - \tau\|_{A^{-1}} + C_{F\Omega} \|\mathrm{div}\tau + \mathsf{v} + f\|, \tag{1.15}$$

By means of (1.14), we deduce another estimate

$$\left\|\nabla(\eta - \eta_g)\right\|_A \le \|A\nabla\eta - \tau\|_{A^{-1}} + C_P \|\mathrm{div}\tau + \mathsf{v} + f\|, \tag{1.16}$$

provided that

$$\int_{\Omega_i} (\mathrm{div}\tau + \mathsf{v} + f)dx = 0, \qquad i = 1, 2, \ldots, N \tag{1.17}$$

Majorants and minorants of the functional type derived for many linear and also nonlinear problems possess the following important properties, namely, they are continuous with respect to both variables and for any $\mathsf{v} \in U$ and $\eta \in V_0$

$$\mathsf{M}^-(\eta, \mathsf{v}, \mathcal{D}) \text{ and } \mathsf{M}^+(\eta, \mathsf{v}, \mathcal{D}) \text{ are nonnegative functionals;} \tag{1.18}$$

$$\mathsf{M}^{-}(\eta_{\mathsf{v}}, \mathsf{v}, \mathcal{D}) = \mathsf{M}^{+}(\eta_{\mathsf{v}}, \mathsf{v}, \mathcal{D}) = 0. \tag{1.19}$$

2 Two–Sided Estimates of the Cost Functional: General Case

First, we present a general result associated with the setting (1.1) and (1.2). Assume that the cost functional satisfies the following condition: for any $\mathsf{v} \in K_U$

$$J(\eta, \mathsf{v}) - \Psi(\|\vartheta\|_{\Xi}) \leq J(\eta + \vartheta, \mathsf{v}) \leq J(\eta, \mathsf{v}) + \Phi(\|\vartheta\|_{\Xi}), \tag{2.1}$$

where Φ and Ψ are some known (continuous) functions vanishing at zero. We note that (2.1) can be viewed as a continuity condition with respect to the state function. In the majority of cases, this condition holds (see Remark 2.2).

Theorem 2.1. *Let (2.1) and (1.9) hold. Then*

$$\inf J = \inf_{\eta \in \Xi} \inf_{\mathsf{v} \in K_U} J^{+}(\eta, \mathsf{v}) \tag{2.2}$$

and

$$\inf J \geq \sup_{\eta \in \Xi} \inf_{\mathsf{v} \in K_U} J^{-}(\eta, \mathsf{v}), \tag{2.3}$$

where

$$J^{+}(\eta, \mathsf{v}) := J(\eta, \mathsf{v}) + \Phi(\mathsf{M}^{+}(\eta, \mathsf{v}, \mathcal{D})),$$
$$J^{-}(\eta, \mathsf{v}) := J(\eta, \mathsf{v}) - \Psi(\mathsf{M}^{-}(\eta, \mathsf{v}, \mathcal{D})).$$

Proof. For any $\eta \in \Xi$, we have

$$\inf J = \inf_{\mathsf{v} \in K_U} J(\eta_{\mathsf{v}}, \mathsf{v}) \leq \inf_{\mathsf{v} \in K_U} \{J(\eta, \mathsf{v}) + \Phi(\|\eta_{\mathsf{v}} - \eta\|_{\Xi})\} \leq$$
$$\inf_{\mathsf{v} \in K_U} \left\{J(\eta, \mathsf{v}) + \Phi(\mathsf{M}^{+}(\eta, \mathsf{v}, \mathcal{D})\right\}.$$

Thus,

$$\inf J \leq \inf_{\substack{\mathsf{v} \in K_U \\ \eta \in \Xi}}, J^{+}(\eta, \mathsf{v}),$$

It is easy to see that the above relation holds as equality. Indeed, if $\mathsf{v} = \mathsf{u}$ and $\eta = \eta_{\mathsf{u}}$, then $\mathsf{M}^{+}(\eta, \mathsf{v}, \mathcal{D}) = 0$ and the second term vanishes while the first one equals $\inf J$.

Analogously,

$$\inf J = \inf_{\mathsf{v}\in K_U} J(\eta_{\mathsf{v}}, \mathsf{v}) \geq \inf_{\mathsf{v}\in K_U} \{J(\eta, \mathsf{v}) - \Psi(\|\eta_{\mathsf{v}} - \eta\|_{\Xi})\}$$
$$\geq \inf_{\mathsf{v}\in K_U} \{J(\eta, \mathsf{v}) - \Psi(\mathsf{M}^{-}(\eta, \mathsf{v}, \mathcal{D})\}$$

and we conclude that

$$\inf J \geq \sup_{\eta\in\Xi} \inf_{\mathsf{v}\in K_U} \{J(\eta, \mathsf{v}) - \Psi(\mathsf{M}^{-}(\eta, \mathsf{v}, \mathcal{D})\}.$$

□

Remark 2.2. In many cases, the condition (2.1) is not difficult to verify. For example, if J is Lipschitz continuous with respect to the state function, i.e., there exists a constant L such that

$$|J(\eta + \vartheta, \mathsf{v}) - J(\eta, \mathsf{v})| \leq L \|\vartheta\|_{\Xi},$$

then (2.1) is obviously satisfied.

Another example is related to the quadratic functional J_1 [cf. (1.5)], we have

$$\left\|A\nabla(\eta + \vartheta) - \sigma^d\right\|^2_{A^{-1}} - \left\|A\nabla\eta - \sigma^d\right\|^2_{A^{-1}}$$
$$= \|A\nabla\vartheta\|^2_{A^{-1}} + 2\int_{\Omega} \nabla\vartheta \cdot (A\nabla\eta - \sigma^d)dx \leq 2\kappa \|\nabla\vartheta\|_A + \|\nabla\vartheta\|^2_A,$$

where $\kappa = \left\|A\nabla\eta - \sigma^d\right\|_{A^{-1}}$. Analogously,

$$\left\|\nabla(\eta + \vartheta) - \sigma^d\right\|^2_{A^{-1}} - \left\|\nabla\eta - \sigma^d\right\|^2_{A^{-1}} \geq \|\nabla\vartheta\|^2_A - 2\kappa \|\nabla\vartheta\|_A.$$

Therefore, if the space Ξ is defined as H^1 endowed with the norm $\|\ \|_A$, then

$$\Psi(t) = \frac{1}{2}t^2 - \kappa\, t \quad \text{and} \quad \Phi(t) = \frac{1}{2}t^2 + \kappa\, t.$$

Similar estimates can be derived for cost functionals satisfying the Hölder continuity condition.

Remark 2.3. In view of (2.2),

$$\inf J \leq J^{+}(\eta, \mathsf{v}) \quad \forall\, \mathsf{v} \in K_U,\ \eta \in \Xi,$$

and the exact lower bound is achieved if we minimize $J^{+}(\eta, \mathsf{v})$ over $K_U \times \Xi$. This means that the problem (1.1) and (1.2) can be represented in a form, where the state and control functions are formally independent.

Analogously

$$\inf J \geq \inf_{\mathsf{v}\in K_U} J^+(\eta, \mathsf{v}) \quad \forall\, \eta \in \Xi.$$

If the minimizer can be found analytically, then the corresponding lower bound is also directly computable. In Sect. 4, we deduce such an computable estimate for the problem (1.3)–(1.5). For relatively simple problems [e.g., for (1.3)–(1.5)], it is also possible to derive guaranteed upper bounds for the norms of $\mathsf{u}-\mathsf{v}$ and $\eta_{\mathsf{u}}-\eta$, i.e., to find computable a posteriori estimates for the state and control functions. In Sect. 5, we discuss these results.

3 Majorant of the Cost Functional: Problem (1.3)–(1.5)

Now, we consider the problem (1.3), (1.5), and (1.7). Let $\mathsf{v} \in K_U$ be an approximation of u. By η_{v} we denote the corresponding exact solution of the state equation. In general, η_{v} is unknown and we use a certain approximation $\eta \in V_0(\Omega)$ instead. It is easy to see that

$$J_1(\eta_{\mathsf{v}}, \mathsf{v}) \leq \frac{1}{2}\left(\|\sigma(\eta) - \sigma^d\|_{A^{-1}} + \|\sigma(\eta_{\mathsf{v}}) - \sigma(\eta)\|_{A^{-1}}\right)^2 + \frac{a}{2}\|\mathsf{v} - \mathsf{u}^d\|^2.$$

We apply (1.15) or (1.16) and find that

$$\begin{aligned}\|\sigma(\eta_{\mathsf{v}}) - \sigma(\eta)\|_{A^{-1}} &= \|\sigma(\eta_{\mathsf{v}} - \eta)\|_{A^{-1}} = \|\nabla(\eta_{\mathsf{v}} - \eta)\|_A \\ &\leq \|\tau - \nabla\eta\|_{A^{-1}} + \mathrm{C}\,\|\mathrm{div}\tau + \mathsf{v} + f\|,\end{aligned} \tag{3.1}$$

where

$$\mathrm{C} = \begin{cases} C_{F\Omega} & \text{if } \tau \in H(\Omega, \mathrm{div}) \\ C_P; & \text{if } \tau \in \tilde{H}^N(\Omega, \mathrm{div}) \end{cases}$$

and $\tau \in H(\Omega, \mathrm{div})$. Here

$$\tilde{H}^N(\Omega, \mathrm{div}) := \left\{ \tau \in H(\Omega, \mathrm{div}) \;|\; \int_{\Omega_i} (\mathrm{div}\tau + \mathsf{v} + f)dx = 0,\ i = 1, \ldots, N \right\}.$$

In view of (3.1), we find that

$$\begin{aligned} &J_1(\eta_{\mathsf{v}}, \mathsf{v}) \\ &\quad\leq \frac{1}{2}\left(\|\sigma(\eta) - \sigma^d\|_{A^{-1}} + \|\tau - A\nabla\eta\|_{A^{-1}} + \mathrm{C}\,\|\mathrm{div}\tau + \mathsf{v} + f\|\right)^2 \\ &\quad+\frac{a}{2}\|\mathsf{v} - \mathsf{u}^d\|^2, \qquad \forall \eta \in V_0, \mathsf{v} \in K_U. \end{aligned}$$

For technical reasons, it is convenient to represent the first term in the right-hand side as the sum of squared norms. Therefore, we introduce positive parameters α and $\beta > 0$, apply Young's inequalities and obtain

$$J_1(\eta_{\mathsf{u}}, \mathsf{u}) \le J_1(\eta_{\mathsf{v}}, \mathsf{v}) \le J^+_{1,\alpha,\beta}(\tau, \mathsf{v}), \qquad \forall \mathsf{v} \in K, \tag{3.2}$$

where

$$\begin{aligned} J^+_{1,\alpha,\beta}(\eta, \tau, \mathsf{v}) := {} & \frac{1+\alpha}{2} \left\| \sigma(\eta) - \sigma^d \right\|^2_{A^{-1}} + \\ & \frac{(1+\alpha)(1+\beta)}{2\alpha} \| \tau - A\nabla\eta \|^2_{A^{-1}} + \\ & + \frac{(1+\alpha)(1+\beta)}{2\alpha\beta} \mathrm{C}^2 \| \mathrm{div}\tau + \mathsf{v} + f \|^2 + \frac{a}{2} \| \mathsf{v} - \mathsf{u}^d \|^2. \end{aligned}$$

Set

$$\mathsf{v} = \mathsf{u}, \quad \tau = A\nabla\eta_{\mathsf{u}}, \quad \text{and} \quad \eta = \eta_{\mathsf{u}}.$$

Then

$$J^+_{1,\alpha,\beta}(\eta_{\mathsf{u}}, A\nabla\eta_{\mathsf{u}}, \mathsf{u}) := \frac{1+\alpha}{2} \left\| \sigma(\eta_{\mathsf{u}}) - \sigma^d \right\|^2_{A^{-1}} + \frac{a}{2} \| \mathsf{u} - \mathsf{u}^d \|^2 \tag{3.3}$$

and we arrive at the important conclusion:

$$\inf J_1 = J_1(\eta_{\mathsf{u}}, \mathsf{u}) = \inf_{\substack{\eta \in V_0,\ \mathsf{v} \in K, \\ \tau \in H(\Omega, \mathrm{div}), \\ \alpha, \beta \in \mathrm{R}_+}} J^+_{1,\alpha,\beta}(\eta, \tau, \mathsf{v}). \tag{3.4}$$

In other words, we have reformulated our problem as an unconstrained minimization problem for a quadratic functional $J^+_{1,\alpha,\beta}$. This functional is explicitly computable and its lower bound *coincides with the exact minimal value of the cost functional*. Therefore, the functional $J^+_{1,\alpha,\beta}(\eta, \tau, \mathsf{v})$ can be used for finding *guaranteed upper bounds* for the cost functional when the minimization problem is solved by known direct minimization methods. Indeed, since the functions η and v are arbitrary, we can take them as approximate solutions computed by certain optimization procedure and minimize $J^+_{1,\alpha,\beta}(\eta, \tau, \mathsf{v})$ with respect to the function τ and parameters α and β. The respective value of β shows an upper bound of the cost functional obtained with these data. In order to obtain a sharper bound, the functions and parameters in the majorant $J^+_{1,\alpha,\beta}(\eta, \tau, \mathsf{v})$ should be changes, e.g., by minimization on finite-dimensional subspaces selected for the state and control functions. The latter subspaces are independent and, in general, may use different meshes and approximations.

Guaranteed upper bounds of the cost functional for problems with state relations defined by the Poisson equation were derived and tested in [6]. The theory applicable for a more general class of problems is presented in [23].

Remark 3.1. For J_2, the majorant can be easily derived by applying the same techniques.

Finding the sharpest upper bound for the cost functional requires the minimization of J_1^+ over η, τ, v, α, and β, where the variables are taken in the above stated sets and are formally independent. Below, we show that the amount of independent variables in J_1^+ can be reduced. For this purpose, we represent the functional in the form

$$J^+_{1,\alpha,\beta}(\eta,\tau,\mathsf{v}) = \mathrm{j}_{11}(\alpha;\eta) + \mathrm{j}_{12}(\alpha,\beta;\eta,\tau) + \mathrm{j}_{13}(\alpha,\beta;\tau,\mathsf{v}), \tag{3.5}$$

where

$$\begin{aligned}
\mathrm{j}_{11}(\alpha;\eta) &:= \frac{1+\alpha}{2}\left\|\sigma(\eta)-\sigma^d\right\|^2_{A^{-1}},\\
\mathrm{j}_{12}(\alpha,\beta;\eta,\tau) &:= \frac{(1+\alpha)(1+\beta)}{2\alpha}\left\|\tau - A\nabla\eta\right\|^2_{A^{-1}},\\
\mathrm{j}_{13}(\alpha,\beta;\tau,\mathsf{v}) &:= \frac{C_{\alpha\beta}}{2}\|\mathrm{div}\tau + \mathsf{v} + f\|^2 + \frac{a}{2}\left\|\mathsf{v}-\mathsf{u}^d\right\|^2,
\end{aligned}$$

and $C_{\alpha\beta} = \mathrm{C}^2\frac{(1+\alpha)(1+\beta)}{\alpha\beta}$.

It is easy to observe that the minimization of J_1^+ with respect to v is equivalent to the problem

$$\inf_{\mathsf{v}\in K}\mathrm{j}_{13}(\alpha,\beta;\tau,\mathsf{v}) := \mathrm{j}_{13}(\alpha,\beta;\tau),$$

which can be solved by minimizing the integrand of j_{13} at almost all $x \in \Omega$. If no constraints are imposed on the control function (i.e., $K_U = L^2(\Omega)$), then the respective minimizer v_{opt} is easy to find. It satisfies the relation

$$\mathsf{v}^0_{opt}(x) = \frac{1}{C_{\alpha\beta}+a}\left[a\mathsf{u}^d(x) - C_{\alpha\beta}(\mathrm{div}\tau(x) + f(x))\right]$$

and results in

$$\hat{\mathrm{j}}_{13}(\alpha,\beta;\tau) = \frac{a}{2}\frac{C_{\alpha\beta}}{C_{\alpha\beta}+a}\|\mathrm{div}\tau + \mathsf{u}^d + f\|^2. \tag{3.6}$$

If K_U is defined by (1.7), then

$$\mathsf{v}_{opt}(x) = \begin{cases} \mathsf{v}^0_{opt}(x) & \text{if } x \in \Omega_1, \\ \\ \mathsf{v}^+(x) & \text{if } x \in \Omega_2, \end{cases}$$

where

$$\Omega_2 := \{x \in \Omega \,|\, \mathsf{v}^0_{opt}(x) > \mathsf{v}^+(x)\} \qquad \text{and } \Omega_1 := \Omega \setminus \Omega_2.$$

In this case,

$$\hat{\mathrm{j}}_{13}(\alpha, \beta; \tau) = \frac{a}{2} \frac{C_{\alpha\beta}}{C_{\alpha\beta} + a} \|\mathrm{div}\tau + \mathsf{u}^d + f\|^2_{\Omega_1} + \tag{3.7}$$

$$+ \frac{C_{\alpha\beta}}{2} \|\mathrm{div}\tau + \mathsf{v}^+ + f\|^2_{\Omega_2} + \frac{a}{2} \|\mathsf{v}^+ - \mathsf{u}^d\|^2_{\Omega_2}.$$

Thus, we arrive at the following result:

Theorem 3.2. *For any $\eta \in V_0$, $\tau \in H(\Omega, \mathrm{div})$ (or $\tau \in \tilde{H}^N(\Omega, \mathrm{div})$), and positive α and β*

$$\inf J_1 \le \mathrm{j}_{11}(\alpha; \eta) + \mathrm{j}_{12}(\alpha, \beta; \eta, \tau) + \hat{\mathrm{j}}_{13}(\alpha, \beta; \tau, \mathsf{v}). \tag{3.8}$$

Moreover,

$$\inf J_1 = \inf_{\substack{\eta \in V_0, \alpha, \beta > 0 \\ \tau \in H(\Omega, \mathrm{div})}} \Big\{ \mathrm{j}_{11}(\alpha; \eta) + \mathrm{j}_{12}(\alpha, \beta; \eta, \tau) + \hat{\mathrm{j}}_{13}(\alpha, \beta; \tau, \mathsf{v}) \Big\}. \tag{3.9}$$

Remark 3.3. Let v be an approximation of u computed by some numerical procedure and η be an approximation of the respective state function η_{v}. Then (3.5)–(3.7) show the value of the cost functional, which is definitely achievable. To find it we should take τ as a post-processed flux $\nabla\eta_{\mathsf{v}}$ and perform a simple minimization with respect to α and β. It is worth noting, that $J_1(\mathsf{v}, \eta)$ does not show a guaranteed upper bound because η is not the exact solution of (1.3).

4 Minorant of the Cost Functional: Problem (1.3)–(1.5)

Now, our goal is to deduce a directly computable minorant of the cost functional. Assume that $\sigma^d = \nabla\eta^d$, where $\eta^d \in V_0$. This assumption does not lead to a loss of generality (see Remark 4.1). Then J_1 has the form

$$J_1(\eta, \mathsf{v}) := \frac{1}{2} \left\| \nabla(\eta - \eta^d) \right\|^2_A + \frac{a}{2} \|\mathsf{v} - \mathsf{u}^d\|^2, \tag{4.1}$$

For any $\eta \in V_0$, we have

$$
\begin{aligned}
J_1(\eta_{\mathsf{v}}, \mathsf{v}) &= \frac{1}{2}\|\nabla(\eta_{\mathsf{v}} - \eta)\|_A^2 + \frac{1}{2}\left\|\nabla(\eta - \eta^d)\right\|_A^2 \\
&+ \int_\Omega A(\nabla(\eta_{\mathsf{v}} - \eta)) \cdot (\nabla(\eta - \eta^d))dx + \frac{a}{2}\|\mathsf{v} - \mathsf{u}^d\|^2 \\
&= \frac{1}{2}\|\nabla(\eta_{\mathsf{v}} - \eta)\|_A^2 + \frac{1}{2}\left\|\nabla(\eta - \eta^d)\right\|_A^2 \\
&+ \int_\Omega (f + \mathsf{v})(\eta - \eta^d)\, dx - \int_\Omega A\nabla\eta \cdot \nabla(\eta - \eta^d)dx \\
&+ \frac{a}{2}\|\mathsf{v} - \mathsf{u}^d\|^2.
\end{aligned} \tag{4.2}
$$

Hence,

$$
\begin{aligned}
J_1(\eta_{\mathsf{u}}, \mathsf{u}) = \inf_{\mathsf{v}\in K_U} J_1(\eta_{\mathsf{v}}, \mathsf{v}) &= \frac{1}{2}\|\nabla(\eta_{\mathsf{v}} - \eta)\|_A^2 + \\
\frac{1}{2}\left\|\nabla(\eta - \eta^d)\right\|_A^2 &+ \int_\Omega (f(\eta - \eta^d) - A\nabla\eta \cdot \nabla(\eta - \eta^d))dx + \\
+ \inf_{\mathsf{v}\in K_U} &\left\{ \int_\Omega \mathsf{v}(\eta - \eta^d)\, dx + \frac{a}{2}\|\mathsf{v} - \mathsf{u}^d\|^2 \right\}.
\end{aligned} \tag{4.3}
$$

It remains to estimate the first term in the right-hand side of (4.3) from below. For this purpose, we use the error minorant $\mathsf{M}^-(\eta)$, which for the considered class of problems reads as follows (see, e.g., [23]): for any $w \in V_0$,

$$
\frac{1}{2}\|\nabla(\eta_{\mathsf{v}} - \eta)\|_A^2 \geq \mathsf{M}^-(\eta, w),
$$

where

$$
\mathsf{M}^-(\eta_{\mathsf{v}}, w) := G(\eta, w) + \int_\Omega \mathsf{v}w\, dx,
$$

$$
G(\eta, w) = \int_\Omega \left(-\frac{1}{2}A\nabla w \cdot \nabla w - A\nabla w \cdot \nabla\eta + fw\right) dx.
$$

Moreover,

$$
\frac{1}{2}\|\nabla(\eta_{\mathsf{v}} - \eta)\|_A^2 = \sup_{w\in V_0} \mathsf{M}^-(\eta, w). \tag{4.4}
$$

We conclude that

$$J_1(\eta_{\mathsf{u}}, \mathsf{u}) = \frac{1}{2}\left\|\nabla(\eta - \eta^d)\right\|_A^2 + \int_\Omega (f(\eta - \eta^d) - A\nabla\eta\cdot\nabla(\eta - \eta^d))dx + \sup_{w\in V_0}\inf_{\mathsf{v}\in K_U}\left\{G(\eta, w) + \int_\Omega \mathsf{v}(w + \eta - \eta^d)\,dx + \frac{a}{2}\|\mathsf{v} - \mathsf{u}^d\|^2\right\} \tag{4.5}$$

and for any $w \in V_0$,

$$J_1(\eta_{\mathsf{u}}, \mathsf{u}) \geq G(\eta, w) + \frac{1}{2}\left\|\nabla(\eta - \eta^d)\right\|_A^2 + \int_\Omega (f(\eta - \eta^d) - A\nabla\eta\cdot\nabla(\eta - \eta^d))dx + \inf_{\mathsf{v}\in K_U}\left\{\int_\Omega \mathsf{v}(w + \eta - \eta^d)\,dx + \frac{a}{2}\|\mathsf{v} - \mathsf{u}^d\|^2\right\}. \tag{4.6}$$

The right-hand side contains an auxiliary variational problem, which has a simple solution. Indeed,

$$\inf_{\mathsf{v}\in K_U}\int_\Omega \left(g\mathsf{v} + \frac{a}{2}|\mathsf{v} - \mathsf{u}^d|^2\right)dx = \int_\Omega \mathcal{H}(a, \mathsf{u}^d, \mathsf{v}^+, g)dx, \tag{4.7}$$

where

$$\mathcal{H}(a, \mathsf{u}^d, \mathsf{v}^+, g)dx := \begin{cases} \mathsf{u}^d g - \frac{1}{2a}g^2 & \text{if } \bar{\mathsf{v}} := \mathsf{u}^d - \frac{g}{a} \leq \mathsf{v}^+, \\ \mathsf{v}^+ g + \frac{a}{2}(\mathsf{v}^+ - \mathsf{u}^d)^2 & \text{if } \bar{\mathsf{v}} > \mathsf{v}^+. \end{cases}$$

Thus, (4.3)–(4.7) imply

$$J_1(\eta_{\mathsf{u}}, \mathsf{u}) \geq J_1^-(\eta, w) \qquad \forall \eta \in V_0, \tag{4.8}$$

where

$$J_1^-(\eta, w) := G(\eta, w) + \frac{1}{2}\|\nabla(\eta - \eta^d)\|_A^2 + \int_\Omega \Big(f(\eta - \eta^d) - A\nabla\eta\cdot\nabla(\eta - \eta^d) + \mathcal{H}(a, \mathsf{u}^d, \mathsf{v}^+, w + \eta - \eta^d)\Big)dx.$$

In other words, for any η and w, the functional $J_1^-(\eta, w)$ is a lower bound of the cost functional. Since all the functions entering J_1^- are known, this minorant is directly computable.

Now we confine ourselves to the case $K_U = L^2(\Omega)$. Then,

$$J_1^-(\eta, w) = G(\eta, w) + \frac{1}{2}\|\nabla(\eta - \eta^d)\|_A^2 \tag{4.9}$$
$$+ \int_\Omega \Big(f(\eta - \eta^d) - A\nabla\eta \cdot \nabla(\eta - \eta^d)$$
$$+ \mathsf{u}^d(w + \eta - \eta^d) - \frac{1}{2a}(w + \eta - \eta^d)^2 \Big) dx$$

Let $\delta\mathsf{u}$ be a small variation of u and the corresponding variation of the state equation be defined by the integral relation

$$\int_\Omega A\nabla(\eta_\mathsf{u} + \delta\eta_\mathsf{u}) \cdot \nabla w \, dx = \int_\Omega (f + \mathsf{u} + \delta\mathsf{u}) w \, dx \quad \forall w \in V_0.$$

Then

$$\int_\Omega A\nabla\delta\eta_\mathsf{u} \cdot \nabla w \, dx = \int_\Omega \delta\mathsf{u} w \, dx. \tag{4.10}$$

Since

$$J_1(\mathsf{u}, \eta_\mathsf{u}) \le J_1(\mathsf{u} + \delta\mathsf{u}, \eta_\mathsf{u} + \delta\eta_\mathsf{u}),$$

we apply usual variational arguments, neglect the quadratic terms of $\delta\mathsf{u}$ and η and find that

$$\int_\Omega (A\nabla(\eta_\mathsf{u} - \eta^d) \cdot \nabla\delta\eta_\mathsf{u} + a(\mathsf{u} - \mathsf{u}^d)\delta\mathsf{u}) \, dx = 0.$$

Using (4.10), we see that for all $\delta\mathsf{u}$

$$\int_\Omega ((\eta_\mathsf{u} - \eta^d) + a(\mathsf{u} - \mathsf{u}^d))\delta\mathsf{u} \, dx = 0,$$

which implies $(\eta_\mathsf{u} - \eta^d) = a(\mathsf{u}^d - \mathsf{u})$.

Let us set $w = 0$ and $\eta = \eta_\mathsf{u}$ in (4.9). We have

$$J_1^-(\eta_\mathsf{u}, 0) = \frac{1}{2}\|\nabla(\eta_\mathsf{u} - \eta^d)\|_A^2 \tag{4.11}$$
$$+ \int_\Omega \Big(f(\eta_\mathsf{u} - \eta^d) - A\nabla\eta_\mathsf{u} \cdot \nabla(\eta_\mathsf{u} - \eta^d) + \mathsf{u}^d(\eta_\mathsf{u} - \eta^d) - \frac{1}{2a}(\eta_\mathsf{u} - \eta^d)^2 \Big) dx$$

$$= \frac{1}{2}\|\nabla(\eta_{\mathsf{u}} - \eta^d)\|_A^2 + \int_\Omega \Big((\mathsf{u}^d - \mathsf{u})(\eta_{\mathsf{u}} - \eta^d) - \frac{1}{2a}(\eta_{\mathsf{u}} - \eta^d)^2\Big)dx$$

$$= \frac{1}{2}\left(\|\nabla(\eta_{\mathsf{u}} - \eta^d)\|_A^2 + a\|\mathsf{u}^d - \mathsf{u}\|^2\right) = J(\mathsf{u}, \eta_{\mathsf{u}}).$$

We have proved that the minorant is sharp, i.e., using majorants and minorants we can find as accurate two sided bounds of the cost functional as it is required.

Remark 4.1. If σ^d does not have the form $A\nabla\eta^d$, then the optimization problem can be reduced to the above considered case. Indeed, let $\hat{\eta}^d \in V_0(\Omega)$ solve the problem

$$\int_\Omega (A\nabla\hat{\eta}^{\,d} - \sigma^d)\cdot\nabla w\,dx = 0 \quad \forall w \in V_0. \tag{4.12}$$

Then

$$\int_\Omega (A\nabla\hat{\eta}^{\,d} - \sigma^d)\cdot\nabla(\eta - \hat{\eta}^{\,d})\,dx = 0$$

and we find that

$$\left\|A\nabla\eta - \sigma^d\right\|_{A^{-1}}^2 = \left\|A\nabla\eta - A\nabla\hat{\eta}^{\,d}\right\|_{A^{-1}}^2 + \left\|A\nabla\hat{\eta}^{\,d} - \sigma^d\right\|_{A^{-1}}^2.$$

In view of this fact,

$$J(\eta, \mathsf{u}) = \frac{1}{2}\|A\nabla\eta - \nabla\hat{\eta}^{\,d}\|^2 + \frac{a}{2}\|\mathsf{u} - \mathsf{u}^d\|^2 + c, \tag{4.13}$$

where $c = \|A\nabla\hat{\eta}^{\,d} - \sigma^d\|_{A^{-1}}^2$ is a certain measure of the distance from η^d to the set V_0. Since c does not depend on the state and control function, we see that the cost functional is reduced to the form (4.1).

5 Estimates for the State and Control Functions

In the final section, we derive guaranteed upper estimate indexguaranteed error boundsfor the error of the approximate solution measured in terms of a *combined norm*

$$|[\mathsf{u} - \mathsf{v}]|^2 := \frac{1}{2}\|\nabla(\eta_{\mathsf{u}} - \eta_{\mathsf{v}})\|_A^2 + \frac{a}{2}\|\mathsf{u} - \mathsf{v}\|^2.$$

The derivation is based on the following result.

Theorem 5.1. *Let $K = L^2(\Omega)$. For any control function $\mathsf{v} \in K$*

$$|[\mathsf{u} - \mathsf{v}]|^2 = J_1(\eta_\mathsf{v}, \mathsf{v}) - J_1(\eta_\mathsf{u}, \mathsf{u}). \tag{5.1}$$

Proof. Since

$$(\eta_\mathsf{u} - \eta^d) + a(\mathsf{u} - \mathsf{u}^d) = 0, \tag{5.2}$$

we have

$$\int_\Omega (\eta_\mathsf{u} - \eta^d)\mathsf{w}dx + a\int_\Omega (\mathsf{u} - \mathsf{u}^d)\mathsf{w}dx = 0, \quad \mathsf{w} \in K. \tag{5.3}$$

Let $\bar{\eta}_\mathsf{w}$ be such that

$$\int_\Omega A\nabla\bar{\eta}_\mathsf{w} \cdot \nabla\xi dx = \int_\Omega \mathsf{w}\,\xi\,dx \quad \forall\xi \in V_0. \tag{5.4}$$

From (5.3) and (5.4) with $\xi = \eta_\mathsf{u} - \eta^d$ it follows that for any $\mathsf{w} \in K$

$$\int_\Omega A\nabla(\eta_\mathsf{u} - \eta^d) \cdot \nabla\bar{\eta}_\mathsf{w}dx + a\int_\Omega \mathsf{w}(\mathsf{u} - \mathsf{u}^d)\,dx = 0.$$

For arbitrary $\mathsf{v} \in K$ we have

$$J(\eta_\mathsf{v}, \mathsf{v}) - J(\eta_\mathsf{u}, \mathsf{u}) = \frac{1}{2}\|\nabla(\eta_\mathsf{v} - \eta_\mathsf{u})\|_A^2 + \frac{a}{2}\|\mathsf{v} - \mathsf{u}\|^2 +$$
$$+ \int_\Omega A\nabla(\eta_\mathsf{u} - \eta^d) \cdot \nabla(\eta_\mathsf{v} - \eta_\mathsf{u})dx + a\int_\Omega (\mathsf{u} - \mathsf{u}^d)(\mathsf{v} - \mathsf{u})dx.$$

Set $\mathsf{w} = \mathsf{v} - \mathsf{u}$. Since

$$\int_\Omega A(\nabla\eta_\mathsf{v} - \nabla\eta_\mathsf{u}) \cdot \nabla\xi\,dx = \int_\Omega (\mathsf{v} - \mathsf{u})\xi dx,$$

we observe that $\bar{\eta}_\mathsf{w} = \eta_\mathsf{v} - \eta_\mathsf{u}$. Therefore, the last two terms vanish and we arrive at (5.1). □

Remark 5.2. The estimate (5.1) can be viewed as a generalization of the Mikhlin's estimate, which was derived for variational problems generated by quadratic functionals $\frac{1}{2}a(v, v) + (f, v)$ in [13]. In [23], it was shown that analogous estimates hold for some classes of optimal control problems. Theorem 5.1 is a generalized version of this result proved by the same method.

Theorem 5.1 and estimates (3.2) and (4.8) yield the following majorant of the combined state–control norm:

Theorem 5.3. *For any* $\mathsf{v} \in L^2(\Omega)$,

$$|[\mathsf{v}-\mathsf{u}]|^2 \leq \mathsf{M}^+(\alpha,\beta,\eta,\tau,w,\mathsf{v}), \tag{5.5}$$

where $\tau \in H(\Omega,\mathrm{div})$, $w \in V_0(\Omega)$, *and*

$$\mathsf{M}^+(\alpha,\beta,\eta,\tau,w,\mathsf{v}) := J^+_{1,\alpha,\beta}(\eta,\tau,\mathsf{v}) - J_1^-(\eta,w) \geq 0.$$

It is not difficult to show that if $\mathsf{v}=\mathsf{u}$, then there exist parameters, which make the majorant zero. Indeed, let $\eta=\eta_{\mathsf{u}}$, $\tau=\sigma(\eta_{\mathsf{u}})$, and $w=0$. In view of (3.3)

$$J^+_{1,\alpha,\beta}(\eta_{\mathsf{u}}, A\nabla\eta_{\mathsf{u}}, \mathsf{u}) = \frac{1+\alpha}{2}\left\|\nabla(\eta_{\mathsf{u}}-\eta^d)\right\|^2_A + \frac{a}{2}\left\|\mathsf{u}-\mathsf{u}^d\right\|^2.$$

On the other hand, in view of (4.11)

$$J_1^-(\eta_{\mathsf{u}},\mathsf{u}) = \frac{1}{2}\|\nabla(\eta_{\mathsf{u}}-\eta^d)\|^2_A + \frac{a}{2}\|\mathsf{u}^d-\mathsf{u}\|^2.$$

Thus,

$$\mathsf{M}^+(\alpha,\beta,\eta_{\mathsf{u}},A\nabla\eta_{\mathsf{u}},0,\mathsf{u}) = \frac{\alpha}{2}\|\nabla(\eta_{\mathsf{u}}-\eta^d)\|^2.$$

We can set α arbitrary small. Therefore the majorant is smaller than any positive number, i.e., it is equal to zero.

References

[1] V. Arnautu, P. Neittaanmäki, *Optimal Control from Theory to Computer Programs* (Kluwer, Dordrecht, 2003)

[2] R. Becker, R. Rannacher, An optimal control approach to error estimation and mesh adaptation, in *Acta Numerica 2000*, ed. by A. Iserles (Cambridge University Press, Cambridge, 2001), pp. 1–102.

[3] R. Becker, B. Vexler, A posteriori error estimation for finite element discretization of parameter identification problems. Numer. Math. **96**, 435–459 (2004)

[4] R. Becker, H. Kapp, R. Rannacher, Adaptive finite element methods for optimal control of partial differential equations: Basic concept. SIAM J. Control Optim. **39**, 113–132 (2000)

[5] A. Gaevskaya, R.H.W. Hoppe, Y. Iliash, M. Kieweg, Convergence analysis of an adaptive finite element method for distributed control problems with control constraints, in Control of Coupled Partial Differential Equations, International series of numerical mathematics, vol. 155 (Birkhäuser, Basel, 2007), pp. 47–68

[6] A. Gaevskaya, W.H. Hoppe, S. Repin, A posteriori error estimation for elliptic optimal control problems with distributed control. J. Math. Sci. **144**, 1–14 (2007)

[7] M. Hintermüller, A primal-dual active set algorithm for bilaterally control constrained optimal control problems. Q. Appl. Math. **LXI**, 131–161 (2003)
[8] R.H.W. Hoppe, Y. Iliash, C. Iyyunni, N.H. Sweilam, A posteriori error estimates for adaptive finite element discretizations of boundary control problems. J. Numer. Math. **14**, 57–82 (2006)
[9] J.L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations* (Springer, Berlin/Heidelberg/New York, 1971)
[10] W.B. Liu, P. Neittaanmäki, D. Tiba, Existence for shape optimization problems in arbitrary dimension. SIAM J. Control Optim. **41**, 1440–1454 (2002)
[11] O. Mali, P. Neittaanäki, S. Repin, *Accuracy Verification Methods. Theory and Algorithms* (Springer, Berlin/Heidelberg/New York, 2013)
[12] D. Meidner, B. Vexler, Adaptive space-time finite element methods for parabolic optimization problems. SIAM J. Control Optim. **46**, 116–142 (2007)
[13] S.G. Mikhlin, *Variational Methods in Mathematical Physics* (Pergamon Press, Oxford, 1964)
[14] P. Neittaanmäki, S. Repin, *Reliable Methods for Computer Simulation, Error Control and A Posteriori Estimates* (Elsevier, New York, 2004)
[15] P. Neittaanmäki, D. Tiba, *Optimal Control of Nonlinear Parabolic Systems: Theory, Algorithms and Applications* (Marcel Dekker, New York, 1994)
[16] P. Neittaanmäki, D. Tiba, Fixed domain approaches in shape optimization problems. Inverse Probl. **28** (2012). doi:10.1088/0266-5611/28/9/093001
[17] P. Neittaanmäki, J. Sprekls, D. Tiba, *Optimization of Elliptic Systems. Theory and Applications* (Springer, Berlin/Heidelberg/New York, 2006)
[18] P. Neittaanmäki, A. Pennanen, D. Tiba, Fixed domain approaches in shape optimization problems with Dirichlet boundary conditions. Inverse Probl. **25**, 055003 (2009)
[19] L.E. Payne, H.F. Weinberger, An optimal Poincaré inequality for convex domains. Arch. Rat. Mech. Anal. **5**, 286–292 (1960)
[20] S. Repin, A posteriori error estimation for nonlinear variational problems by duality theory. Zapiski Nauchn, Semin, V.A. Steklov Mathematical Institute in St.-Petersburg (POMI) **243**, 201–214 (1997)
[21] S. Repin, A posteriori error estimates for variational problems with uniformly convex functionals. Math. Comput. **69**, 481–500 (2000)
[22] S. Repin, Two-sided estimates of deviation from exact solutions of uniformly elliptic equations. Proc. St. Petersburg Math. Society **IX**, 143–171 (2001) [Translation in Am. Math. Soc. Transl. Ser. 2 **209**. American Mathematical Society, Providence, 2003]
[23] S. Repin, *A Posteriori Estimates for Partial Differential Equations* (De Gruyter, Berlin, 2008)

Shape Sensitivity Analysis of the Work Functional for the Compressible Navier–Stokes Equations

Pavel I. Plotnikov and Jan Sokołowski

Abstract The compressible Navier–Stokes equations with nonhomogeneous Dirichlet conditions in a bounded domain with an obstacle are considered (P.I. Plotnikov, J. Sokolowski, *Compressible Navier-Stokes Equations. Theory and Shape Optimization*, Birkhäuser, Basel, 2012). The dependence of local solutions on the shape of an obstacle is analyzed (P.I. Plotnikov, E.V. Ruban, J. Sokołowski, SIAM J. Math. Anal. 40:1152–1200, 2008; P.I. Plotnikov, E.V. Ruban, J. Sokołowski, J. Math. Pures Appl. 92:113–162, 2009; P.I. Plotnikov, J. Sokołowski, Dokl. Akad. Nauk 397:166–169, 2004; P.I. Plotnikov, J. Sokołowski, J. Math. Fluid Mech. 7:529–573, 2005; P.I. Plotnikov, J. Sokołowski, Comm. Math. Phys. 258:567–608, 2005; P.I. Plotnikov, J. Sokołowski, SIAM J. Control Optim. 45:1165–1197, 2006; P.I. Plotnikov, J. Sokołowski, Uspekhi Mat. Nauk 62:117–148, 2007; P.I. Plotnikov, J. Sokołowski, Stationary boundary value problems for compressible Navier-Stokes equations, in *Handbook of Differential Equations: Stationary Partial Differential Equations*, vol. VI, Elsevier/North-Holland, Amsterdam, 2008, pp. 313–410; P.I. Plotnikov, J. Sokołowski, SIAM J. Control Optim. 48:4680–4706, 2010; P.I. Plotnikov, J. Sokołowski, J. Math. Sci. 170:34–130, 2010). The shape derivatives (J. Sokołowski, J.-P. Zolésio, *Introduction to Shape Optimization. Shape Sensitivity Analysis*, Springer, Berlin/Heidelberg/New York, 1992) of solutions to the compressible Navier–Stokes equations are derived. The shape gradient (J. Sokołowski, J.-P. Zolésio, *Introduction to Shape Optimization. Shape Sensitivity Analysis*, Springer, Berlin/Heidelberg/New York, 1992) of the work functional is obtained. In this way the framework for numerical methods of shape optimization (P. Plotnikov, J. Sokołowski, A. Żochowski, Numerical experiments in drag minimization for compressible Navier-Stokes flows in bounded domains, in *Proceedings of the 14th International IEEE/IFAC*

P.I. Plotnikov
Lavrentyev Institute of Hydrodynamics, Lavrentyev pr. 15, 630090 Novosibirsk, Russia
e-mail: plotnikov@hydro.nsc.ru

J. Sokołowski (✉)
Laboratoire de Mathématiques, Institut Élie Cartan Nancy, UMR7502 (Université Lorraine, CNRS, INRIA), Université de Lorraine, B.P.239, 54506 Vandoeuvre-lès-Nancy Cedex, France
e-mail: Jan.Sokolowski@univ-lorraine.fr

R. Hoppe (ed.), *Optimization with PDE Constraints*, Lecture Notes in Computational Science and Engineering 101,
DOI 10.1007/978-3-319-08025-3_11

Conference on Methods and Models in Automation and Robotics, MMAR'09, 2009, 4 pp; A. Kaźmierczak, P.I. Plotnikov, J. Sokołowski, A. Żochowski, Numerical method for drag minimization in compressible flows, in 15th International Conference on Methods and Models in Automation and Robotics (MMAR), pp. 97–101, 2009. doi:10.1109/MMAR.2010.5587258) is established.

Keywords Compressible Navier–Stokes equations • Drag functional • Nonhomogeneous Dirichlet boundary value problem for nonstationary • Rigid obstacle • Shape derivative • Shape functional • Shape gradient • Work functional

Mathematics Subject Classification (2010). Primary 49K20; Secondary 35K55, 76N25.

1 Introduction

The shape sensitivity analysis of the work functional for the compressible Navier–Stokes equations is performed in this paper. The shape derivatives of the solutions to the equations and the shape gradient of the work functional are obtained in the framework of boundary variational techniques [11, 15].

The recent monograph [11] is devoted to the study of boundary value problems for equations of viscous gas dynamics, named compressible Navier–Stokes equations. The principal significance of the mathematical theory of the Navier–Stokes equations lies in the central role they now play in fluid dynamics. In [11] we focus on existence results for the inhomogeneous in/out flow problem, in particular the problem of the flow around a body placed in a finite domain, on the stability of solutions with respect to domain perturbations, on the domain dependence of solutions to compressible Navier–Stokes equations, and on the drag optimization problem. We refer the reader to [2–8, 10, 12–14] for the related results on modeling and shape optimization for compressible Navier–Stokes equations.

We recall briefly the main topics considered in the mathematical monograph [11] on compressible Navier–Stokes equations which is our main reference in this paper.

Existence Theory. The problem of the flow of a viscous gas around a moving rigid body $S \in \mathbb{R}^d$, $d = 2, 3$, can be formulated as follows. Choose an arbitrary hold-all $B \subset \mathbb{R}^3$, for instance, a sufficiently large ball, such that $S \subset B$. Next, we transfer the boundary conditions from infinity to ∂B and arrive at the following boundary value problem for the velocity $\mathbf{v}$ and the density ρ. Find functions $(\mathbf{v}, \rho)$ satisfying

$$\partial_t(\rho\,\mathbf{v}) + \operatorname{div}(\rho\,\mathbf{v}\otimes\mathbf{v}) - \frac{1}{\mathrm{Re}}\operatorname{div}\mathbb{S}(\mathbf{v}) + \frac{1}{\mathrm{Ma}^2}\nabla p(\rho) + \mathbb{C}\,\mathbf{v} = \rho\,\mathbf{f} \quad \text{in } \Omega\times(0,T),$$

$$\partial_t\rho + \operatorname{div}(\rho\mathbf{v}) = 0 \quad \text{in } \Omega\times(0,T),$$

$$\mathbf{v} = 0 \quad \text{on } \partial S \times (0,T), \quad \mathbf{v} = \mathbf{V} \quad \text{on } \partial B \times (0,T),$$
$$\rho = \varrho_\infty \quad \text{on } \Sigma_{\text{in}},$$
$$\mathbf{v}(x,0) = \mathbf{V}(x,0) \quad \text{in } \Omega, \quad \rho(y,0) = \varrho_\infty(y) \quad \text{in } \Omega,$$

where $\mathbf{V}, \mathbf{f} : \mathbb{R}^d \times [0,T]$ are given smooth vector fields, $\varrho_\infty : \mathbb{R}^d \to \mathbf{R}^+$ is a given nonnegative bounded function, $\mathbb{C}$ is a skew-symmetric matrix,

$$\Omega = B \setminus S, \quad \Sigma_{\text{in}} = \{(x,t) \in \partial B \times (0,T) : \mathbf{V}(x,t) \cdot \mathbf{n}(y) > 0\},$$
$$\mathbb{S}(\mathbf{v}) = \nabla \mathbf{v} + (\nabla \mathbf{v})^\top + (\lambda - 1)\text{div}\, \mathbf{v}\, \mathbb{I}.$$

The peculiarity of this problem is that we deal with the boundary value problem for the mass balance equations. We prove that for the adiabatic exponent $\gamma > d/2$, the problem has a renormalized solution. We follow the multilevel regularization scheme proposed by P.L. Lions and E. Feireisl, but with a different regularization technique. We show that the solution admits the energy estimate and the pressure $p(\rho)$ is locally integrable with some exponent greater than 1.

Stability of Solutions with Respect to Nonsmooth Data and Domain Perturbations. Propagation of Rapid Oscillations in Compressible Fluids. In compressible viscous flows, any irregularities in the initial and boundary data are transferred inside the flow domain along fluid particle trajectories. We develop a new method for the study of the propagation of rapid oscillations of the density, which can be regarded as acoustic waves. The main idea is that any rapidly oscillating sequence is associated with a parametrized family μ_{xt} of probability measures on the real line named the Young measure. We establish that the distribution function $f(x,t,s) = \mu_{x,t}(-\infty, s]$ satisfies a differential relation named a kinetic equation. A remarkable property of compressible Navier–Stokes equations is that in this particular case the kinetic equation can be written in closed form as

$$\partial_t f + \text{div}\,(f\mathbf{v}) - \partial_s \left(s f \,\text{div}\, \mathbf{v} + \frac{s}{\lambda + 1} \int_{(-\infty, s]} (p(\tau) - \overline{p})\, d_\tau f(x,t,\tau) \right) = 0.$$

The kinetic equation being combined with the momentum balance equations gives a closed system of integro-differential equations which describes the propagation of rapid oscillations in a compressible viscous flow. Notice that oscillations can be induced not only by oscillations of initial and boundary data, but also by irregularities of the boundary of the flow domain. We also prove that if the data are deterministic and the function f satisfies some integrability condition, then any solution to the kinetic equation satisfying some integrability conditions is deterministic.

Domain Dependence of Solutions to Compressible Navier–Stokes Equations. We apply the kinetic equation method to the analysis of the domain dependence of

solutions to compressible Navier–Stokes equations. We restrict our considerations to the problem of the flow around an obstacle placed in a fixed domain. Recall that in this problem, the flow domain $\Omega = B \setminus S$ is a condenser type domain, B is a fixed hold all domain and S is a compact obstacle. We introduce the notion of the Kuratowski-Mosco. To this end we denote by $C_S^\infty(B)$ the set of all smooth functions defined in B and vanishing on $S \subset B$. Let $W_S^{1,2}(B)$ be the closure of $C_S^\infty(B)$ in the $W^{1,2}(B)$-norm. A sequence of compact sets $S_n \subset B$ is said to converge to S in the Kuratowski-Mosco sense if

- there is a compact set $B' \subset B$ such that $S_n, S \subset B'$;
- for any sequence $u_n \rightharpoonup u$ weakly convergent in $W^{1,2}(B)$ with $u_n \in W_{S_n}^{1,2}(B)$, the limit element u belongs to $W_S^{1,2}(B)$;
- whenever $u \in W_S^{1,2}(B)$, there is a sequence $u_n \in W_{S_n}^{1,2}(B)$ with $u_n \to u$ strongly in $W^{1,2}(B)$.

We show that if a sequence S_n of compact obstacles converges to a compact obstacle S in the Hausdorff and the Kuratowski-Mosco sense, then the sequence of corresponding solutions to the in/out flow problem contains a subsequence which converges to a solution to the in/out flow problem in the limiting domain. Moreover, we prove that the typical cost functionals, such as the work of hydrodynamical forces, are continuous with respect to $\mathcal{S}$-convergence. As a conclusion we establish the solvability of the problem of minimization of the work of hydrodynamical forces in the class of obstacles with a given fixed volume.

2 Boundary Variations Technique for Shape Sensitivity Analysis of Work Functional

Beside the existence of an optimal obstacle for the work and drag shape optimization problems [11], it is important for applications to provide necessary optimality conditions and to devise a numerical method for the solution of the shape optimization problems under consideration. The numerical methods of gradient or steepest descent types require the local information on the behavior of the shape functional to be minimized. The precise information on the shape gradient of the cost functional can be obtained as a result from the appropriate shape sensitivity analysis of the functional. The shape sensitivity analysis requires some regularity of solutions to the governing equations like the Lipschitz continuity with respect to boundary perturbations of the obstacle. The shape sensitivity analysis is performed in [11] for local solutions defined by small perturbations of a class of approximate solutions to the stationary problem. The shape optimization problem with the drag functional for stationary problems as well as the work functional are considered for nonstationary problems in [11].

2.1 Boundary and Distributed Shape Functionals

We recall here, that the following notation is used for the shape functionals under consideration for the shape sensitivity analysis.

We consider the integral shape functionals denoted by $J(S)$ or by $J(\Omega)$, with $S \Subset B$ a compact obstacle in a hold all domain B, and $\Omega := B \setminus S$. There are two different cases of governing equations under considerations, the stationary compressible Navier–Stokes equations and the nonstationary compressible Navier–Stokes equations. In the stationary case $J(S) \equiv J(\Omega)$, stands for the drag functional. In the nonstationary case the same symbols $J(S) \equiv J(\Omega)$ are used for the work functional. There is however a difference between $J(S)$ and $J(\Omega)$, e.g., in the stationary case the drag functional $J(S)$ is given by an integral over the obstacle boundary ∂S, and the same drag functional $J(\Omega)$ takes a form of a volume integral. Similarly, in the nonstationary case the work functional $J(S)$ contains an integral over the lateral boundary $\partial S \times (0, T)$, and the same work functional $J(\Omega)$ contains an integral over the cylinder $\Omega \times (0, T)$. Usually, the functional $J(\Omega)$ is obtained from the functional $J(S)$ by an integration by parts formulae.

It is clear that the distributed shape functionals $J(\Omega)$ require less regularity from the solutions to the governing equations compared to the boundary shape functionals $J(S)$. On the other hand the distributed shape functionals formally depend on a choice of a function denoted by η which is required in the integration by parts formulae, however in view of the identity $J(S) \equiv J(\Omega)$ the values of the shape functional $J(\Omega)$ are independent of the choice of η.

2.2 Shape Sensitivity Analysis Within Boundary Variations Technique

Our goal now is to develop the shape sensitivity analysis which results in the shape derivatives of solutions to the governing equations and in the shape gradients of $J(\Omega)$ obtained for stationary and nonstationary governing equations by introduction of appropriate adjoint state equations.

Two different types of velocity fields can be employed. The physical field is the state variable $\mathbf{u} := \mathbf{u}(\Omega)$, $\Omega = B \setminus S$ determined from the governing equations for a given obstacle S. This field is in general non-unique, thus the local classical solutions of governing equations are considered for the purposes of the shape sensitivity analysis. The artificial velocity field $\mathfrak{V} := \mathfrak{V}(\varepsilon, x)$, $x \in B$, is introduced for the purposes of the shape sensitivity analysis with respect to the small perturbations of the obstacle boundary in the normal direction. This field depends on the small shape parameter $\varepsilon \to 0$ and it is associated with the domain transformation mapping $\mathfrak{T}_\varepsilon : S \mapsto S_\varepsilon$,

$$\mathfrak{V}(\varepsilon, x) = \left(\frac{\partial}{\partial \varepsilon} \mathfrak{T}_\varepsilon \right) \circ \mathfrak{T}_\varepsilon^{-1}(x). \tag{2.1}$$

Now, the form of the mapping $\mathfrak{T}_\varepsilon$ is specified,

$$\mathfrak{T}_\varepsilon(x) := x + \varepsilon \mathbf{T}(x), \tag{2.2}$$

where the field $\mathbf{T}(x)$, $x \in B$, is compactly supported in a small neighborhood of the obstacle S and the support of $\mathbf{T}$ is disjoint with the boundary $\Sigma = \partial B$. This means that the boundary Σ is invariant under transformation (2.2).

In order to evaluate the shape gradient of the functional $J(\Omega)$ the method of boundary variations is applied and the Eulerian semi-derivative of the shape functional $dJ(\Omega; \mathfrak{V})$ is obtained in the direction of a vector field $\mathfrak{V}$ associated with the change of the variables $\mathfrak{T}_\varepsilon$.

This means that for the mapping (2.2) the family of perturbed obstacles is defined by $S_\varepsilon := \mathfrak{T}_\varepsilon(S)$, where $\varepsilon \to 0$ stands for the shape parameter. As a result, the differentiability of the real valued function $\varepsilon \mapsto J(\Omega_\varepsilon)$, with $\Omega_\varepsilon = B \setminus S_\varepsilon$, is considered at $\varepsilon = 0$, and the existence of the derivative is established.

2.3 Stationary Navier–Stokes State Equations

We assume that the viscous gas occupies the double-connected domain $\Omega = B \setminus S$, where $B \subset \mathbb{R}^d$, $d = 2, 3$, is a hold all domain with the smooth boundary $\Sigma = \partial B$, and $S \subset B$ is a compact obstacle. The boundary of the obstacle is denoted by ∂S.

Furthermore, we assume that the velocity of the gas coincides with a given constant vector field $\mathbf{U}$ on the surface Σ. The state variables include the velocity field $\mathbf{u}$ and the gas density ϱ, and satisfy the following equations along with the boundary conditions

$$\operatorname{div}(\varrho\, \mathbf{u} \otimes \mathbf{u}) - \operatorname{div} \mathbb{S}(\mathbf{u}) + \nabla p(\varrho) - \varrho\, \mathbf{f} = 0 \quad \text{in } \Omega\,, \tag{2.3a}$$

$$\operatorname{div}(\varrho \mathbf{u}) = 0 \quad \text{in } \Omega\,, \tag{2.3b}$$

$$\mathbf{u} = \mathbf{U} \text{ on } \Sigma, \qquad \mathbf{u} = 0 \text{ on } \partial S, \tag{2.3c}$$

$$\varrho = \varrho_0 \text{ on } \Sigma_{\text{in}}, \tag{2.3d}$$

with the viscous stress tensor of the form

$$\mathbb{S}(\mathbf{u}) = \nabla \mathbf{u} + \nabla \mathbf{u}^\top + (\lambda - 1) \operatorname{div} \mathbf{u} \mathbb{I}\,, \tag{2.3e}$$

where the pressure $p = p(\varrho)$ is a smooth, strictly monotone function of the density, the Mach, Strouhal, and Reynolds numbers are fixed, $\mathsf{Ma}^2 = \mathsf{Sr} = \mathsf{Re} = 1$, λ is the viscosity ratio, ϱ_0 is a positive constant, and the inlet Σ_{in} and the outlet Σ_{out} are defined by

$$\Sigma_{\text{in}} = \{x \in \Sigma : \mathbf{U} \cdot \mathbf{n} < 0\}, \quad \Sigma_{\text{out}} = \{x \in \Sigma : \mathbf{U} \cdot \mathbf{n} > 0\},$$

respectively. To avoid the technical difficulties at this stage of formal analysis we assume that the intersection of the inlet and of the outlet is an empty set. Here $\mathbf{n}$ stands for the outward normal to $\partial\Omega = \Sigma \cup S$.

2.4 Drag Functional

The boundary value problem (2.3) can be regarded as a mathematical model of viscous gas flow around an airfoil S tested in a wind tunnel. In our notation the stress tensor is equal to

$$\mathbb{T}(\mathbf{u}) := \nabla\mathbf{u} + \nabla\mathbf{u}^{\top} + (\lambda - 1)\,\mathrm{div}\,\mathbf{u}\,\mathbb{I} - p\mathbb{I} = \mathbb{S}(\mathbf{u}) - p\mathbb{I}\,,$$

and the hydrodynamic force acting on the element dS of the obstacle boundary ∂S is $-\mathbb{T}\mathbf{n}\,dS$. Hence the hydrodynamic force acting on the body S is given by a boundary integral,

$$\mathbf{J}(S) := -\int_{\partial S} \mathbb{T}\mathbf{n}\,ds = \tag{2.4}$$

$$-\int_{\partial S} \left(\nabla\mathbf{u} + (\nabla\mathbf{u})^{\top} + (\lambda - 1)\,\mathrm{div}\,\mathbf{u}\mathbb{I} - p\mathbb{I}\right)\mathbf{n}\,ds.$$

Note that (2.4) can be equivalently rewritten in the form of a volume integral. To this end we fix an arbitrary function $\eta \in C^{\infty}(\Omega)$ such that $\eta = 1$ in an open neighborhood of the obstacle S and $\eta = 0$ in a vicinity of Σ. Using the identities

$$\int_{\partial S} \mathbb{T}\mathbf{n}\,ds = \int_{\Omega} (\eta\,\mathrm{div}\,\mathbb{T} + \mathbb{T}\nabla\eta)\,dx, \quad \mathrm{div}\,\mathbb{T} = \varrho\mathbf{u}\nabla\mathbf{u},$$

we introduce the drag functional

$$J(\Omega) := \mathbf{U}_{\infty} \cdot \mathbf{J}(S) = \int_{\Omega} \mathfrak{F}(\mathbf{u}, \nabla\mathbf{u}, p, \eta, \nabla\eta)dx\,, \tag{2.5}$$

where $\Omega := B \setminus S$ and

$$\mathfrak{F}(\mathbf{u}, \nabla\mathbf{u}, p, \eta, \nabla\eta) = \tag{2.6}$$

$$\mathbf{U}_{\infty} \cdot \left[-\left(\nabla\mathbf{u} + (\nabla\mathbf{u})^{\top} + (\lambda - 1)\,\mathrm{div}\,\mathbf{u}\mathbb{I} - p\mathbb{I}\right)\nabla\eta - \eta\varrho\mathbf{u}\nabla\mathbf{u}\right],$$

here $\mathbf{U}_{\infty}$ stands for the airfoil speed. The value of $\mathbf{J}$ is independent of the choice of η. In nonstationary case the drag is a work in unit time developed by the component of $\mathbf{J}(S)$ parallel to the airfoil speed $\mathbf{U}_{\infty}$. In the stationary case it is just a component of the force. The associated shape functional $S \mapsto J(\Omega)$, $\Omega := B \setminus S$ is considered

as a shape functional defined for an admissible family of obstacles $S \in \mathcal{U}_{\text{ad}}$. We need a family $\mathcal{U}_{\text{ad}}$ to assure the existence of an optimal obstacle. It is probable that the drag functional in stationary case is defined for an obstacle S which is a closed set included in the open set B, provided that the existence of a solution for the governing equations in the stationary case with the nonhomogeneous Dirichlet boundary conditions is shown, which is still an open problem, it seems.

2.5 The Velocity Method of Shape Sensitivity Analysis

In [11] the local strong solutions of stationary compressible Navier–Stokes equations are considered for the purposes of the shape sensitivity analysis. Such solutions are uniquely determined and are stable with respect to shape perturbations within the boundary variations technique.

In this chapter the general framework of shape sensitivity analysis adapted to the specific case of Navier–Stokes equations is established. This means in particular, that the Piola transform of the velocity field is employed in order to determine the material derivatives in a reasonable way.

For the general purposes of shape sensitivity analysis the family of perturbations S_ε for an obstacle $S \Subset B$ is introduced, depending on the small parameter $\varepsilon \to 0$. To this end the perturbations of the domain Ω are defined by an appropriate change of variables (2.2),

$$\mathfrak{T}_\varepsilon \, : \, \mathbb{R}^d \ni x \mapsto y(\varepsilon, x) \in \mathbb{R}^d \, , \tag{2.7}$$

or equivalently, the boundary of the obstacle S_ε is given by the perturbation of the sufficiently smooth boundary ∂S in the normal direction depending on a function $f(\omega), \omega \in \partial S$,

$$\partial S_\varepsilon = \{x = \omega + \varepsilon f(\omega)\mathbf{n}(\omega), \quad \omega \in \partial S\} \, , \tag{2.8}$$

where $\mathbf{n}$ stands for the unit outward normal vector on ∂S, and $f(\omega)$, $\omega \in \partial S$ is a given function which defines the boundary variations of ∂S in normal direction.

Let us observe that f determines the mapping $\mathfrak{T}_\varepsilon$ only on the boundary ∂S. By the Hadamard representation theorem [15] the knowledge of f is also sufficient for determination of the shape gradient $\mathfrak{G}$ of the differentiable shape functional $J(\Omega)$. The explicit form of the shape gradient of the drag functional $J(\Omega)$ is required in particular for the numerical methods of shape optimization. For example, the level set method of shape optimization is based on the knowledge of the shape gradient given by a boundary integral in (2.9). It means that, with applications to numerical methods in mind, we investigate the existence of the following limit possibly given by a boundary integral

$$dJ(\Omega; \mathfrak{V}(0)) := \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}(J(\Omega_\varepsilon) - J(\Omega)) = \int\limits_{\partial S} f(\omega) G_S(\omega) ds(\omega) \, , \tag{2.9}$$

where $G_S(\omega), \omega \in \partial S$, stands for the so-called *boundary shape gradient of drag functional* in the direction of the *shape velocity vector field* depending on two variables $(\varepsilon, x) \mapsto \mathfrak{V}(\varepsilon, x)$, the first variable is the small shape parameter $\varepsilon \to 0$, and x is the spatial variable. We point out that in terms of the mapping $\mathfrak{T}_\varepsilon$ given in (2.7), the associated shape velocity field [15] takes the form (2.1).

In order to show the existence of the shape gradient given by a function $G_S(\omega), \omega \in \partial S$, and to identify its form, the general procedure described in [15] essentially for linear elliptic equations cannot be directly applied due to the nonlinear nature of the governing equations. Some modifications of the method are proposed here, such modifications are in fact necessary in view of the complexity of nonlinear model. As a result, a new method of shape sensitivity analysis well adapted to the analysis of compressible Navier–Stokes equations is proposed in the monograph.

First, the shape sensitivity analysis of the state equation is performed in order to determine the shape gradient G_S. To this end we evaluate the derivatives of solutions $(\mathbf{u}(\Omega_\varepsilon), \varrho(\Omega_\varepsilon))$ defined in the domain $B \setminus S_\varepsilon$, and extended by zero over the obstacles S_ε. The solutions are differentiated with respect to the parameter ε and the derivative of the drag functional $\varepsilon \mapsto J(\Omega_\varepsilon)$ at $\varepsilon = 0$ is determined.

The procedure proposed in the monograph for differentiation of the mapping

$$\varepsilon \mapsto (\mathbf{u}(\Omega_\varepsilon), \varrho(\Omega_\varepsilon)), \tag{2.10}$$

results in the material derivatives as well as in the shape derivatives of the state $(\mathbf{u}, \varrho)$ with respect to the small shape parameter ε.

In general, the algorithm for evaluation of material derivatives is simple [15], first make the change the variables, then differentiate the composed mapping

$$\varepsilon \mapsto (\mathbf{u}(\Omega_\varepsilon) \circ \mathfrak{T}_\varepsilon, \varrho(\Omega_\varepsilon) \circ \mathfrak{T}_\varepsilon) \tag{2.11}$$

in the fixed function space over the unperturbed domain Ω e.g., in the stationary case. Then the form of shape derivatives can be deduced from the material derivatives. Such a procedure is straightforward for linear partial differential equations [15], unfortunately it becomes difficult to apply for the nonlinear problems. The formal evaluation of the shape gradient for an integral shape functional in terms of the shape derivatives of solutions is easy to apply even in the case of nonlinear problems, but the obtained result is more difficult to justify when compared to linear problems. Such formal evaluation uses the direct differentiation of the mapping (2.10) by only a formal application of the Implicit Function Theorem. We point out at this point, that for the compressible Navier–Stokes equations there is no general result on the regularity of solutions, and such a regularity result is required for the justification of the formal procedure.

Let us introduce the material and the shape derivatives for the density, the same objects are defined for the velocity field later on. We denote by

$$\dot{\varrho} = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} (\varrho(\Omega_\varepsilon) \circ \mathfrak{T}_\varepsilon - \varrho(\Omega)) \tag{2.12}$$

the so-called material derivative of the density with respect to the shape parameter ε. However, the material derivative is different from the limit

$$\varrho' = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}(\varrho(\Omega_\varepsilon) - \varrho(\Omega)) \tag{2.13}$$

which is the so-called shape derivative of the density with respect to the shape parameter ε, the relation between material and shape derivatives takes the form

$$\dot{\varrho} = \varrho' + \nabla \varrho \cdot \mathfrak{V}\,. \tag{2.14}$$

Once, the material derivatives $(\dot{\mathbf{u}}, \dot{\varrho})$ of the state $(\mathbf{u}, \varrho)$ are determined, the first order necessary optimality conditions for the shape optimization of compressible Navier–Stokes equations can be obtained. This means that for given $(\dot{\mathbf{u}}, \dot{\varrho})$ the shape gradient of the drag functional can be identified.

We can conclude that the shape differentiability of solutions to the governing equations can be achieved by the material derivative technique developed in [15], provided some additional regularity of the weak solutions is known. Concerning the shape differentiability of the velocity field we need an additional transformation, which is called the Piola transform. To this end the new velocity field $\mathbf{u}_\varepsilon(x), x \in \Omega = B \setminus S$ (here we assume that $S := \overline{S}$ is a compact subset in order to simplify the notation) in the fixed domain is introduced for all sufficiently small $\varepsilon > 0$. The field $\mathbf{u}_\varepsilon$ is defined in (5.11) by using the Piola transform for the fluid velocity fields. The Piola transform assures the invariance of the divergence operator for the change of variables (2.7) from the fixed domain Ω to the variable domain Ω_ε. Therefore, by the change of variables combined with the Piola transformation the unknown velocity field $\mathbf{u}(\Omega_\varepsilon)(x), x \in B \setminus S_\varepsilon$, is replaced by the new unknown function $\mathbf{u}_\varepsilon$ defined in the fixed reference domain $B \setminus S$. The function $\mathbf{u}_\varepsilon$ can be extended to the hold all domain B. However, the derivative of $\mathbf{u}_\varepsilon$ with respect to ε is different from the material derivative $\dot{\mathbf{u}}$, the relation between such a derivative and the material derivative depends on the specific choice of matrix $\mathbf{N}$ in the transformation of the velocity fields defined in (2.2) or in (5.11).

It turns out, that in general the material derivatives, $\dot{\varrho}$ for the density ϱ, and $\dot{\mathbf{u}}$ for the velocity $\mathbf{u}$, are given by an auxiliary boundary value problem depending on the shape velocity field $\mathfrak{V}(0, x)$ at $\varepsilon = 0$ and for $x \in \Omega$. On the other hand the shape derivatives ϱ' and $\mathbf{u}'$, under some regularity assumptions, depend only on the normal component of the field $\mathfrak{V}(0, \omega)$ at $\varepsilon = 0$ and for $\omega \in \partial S$, i.e. on the function $f(\omega), \omega \in \partial S$ in (2.8). It is clear that the function f defines the normal component of the shape velocity field on the boundary of the obstacle $f(\omega) = \mathfrak{V}(0, \omega) \cdot \mathbf{n}(\omega)$, $\omega \in \partial S$.

Therefore, the general strategy of shape derivation consists in two steps. First, the material derivatives are obtained and used in order to show that the drag functional is shape differentiable. Then, the shape derivatives are employed, possibly with an appropriate adjoint state, in order to identify the form of the shape gradient of the drag functional. The shape gradient $\mathfrak{G}$ is a distribution, obtained by the Hadamard

representation formula (see e.g., [15] for a proof of the representation formula), which lives on the boundary ∂S of the obstacle. In our case $\mathfrak{G}$ is given by a function denoted G_S. Such a strategy is in fact applied in [11], where the complete proofs are given for the stationary governing equations.

2.6 Shape Derivatives of Solutions to Governing Equations

We briefly describe the specific shape optimization problem which is analyzed in this chapter. The problem is considered in three spatial dimensions, the particular case of two spatial dimensions is also covered by our framework. The question is to find an optimal shape of the obstacle S included in a large computational hold all domain B. The optimal shape if any, minimizes the drag functional $J(\Omega)$ within a family of admissible obstacles, or admissible shapes $\mathcal{U}_{\text{ad}}$. The hold all domain B can be selected e.g., as a ball of the radius $R \gg 1$, to fix the ideas. For such an optimization problem we have already the existence of an optimal shape. Since the velocity field $\mathbf{u}$ is in the Sobolev space $H^1(B \setminus S)$ and vanishes on the boundary ∂S of the obstacle, the extension of $\mathbf{u}$ by zero over the obstacle S, still denoted by the same symbol $\mathbf{u}$, is in the Sobolev space $H^1(B)$. Therefore, the analysis of the differential properties of the mapping

$$\mathcal{U}_{\text{ad}} \ni S \mapsto \mathbf{u}(\Omega) \in H^1(B)\,,$$

with respect to the shape of the obstacle can be performed in the fixed function space $H^1(B)$ for the extended function $\mathbf{u}$ which is defined all over the fixed hold all domain B and vanishes on S.

Now, the outline of the shape sensitivity analysis of the state equation is presented. If the shape of the obstacle S is perturbed by the boundary variations technique, and the perturbed obstacle is denoted by S_ε, where $\varepsilon \to 0$ is a shape parameter, the extended velocity field determined from the governing equations depends on the small parameter

$$\mathcal{U}_{\text{ad}} \ni S_\varepsilon \mapsto \mathbf{u}(\Omega_\varepsilon) \in H^1(B)\,,$$

where $S_\varepsilon = \mathfrak{T}_\varepsilon(S)$ is the image of the unperturbed obstacle S under transformation (2.2). In general, the mapping of (2.2) $\mathfrak{T}_\varepsilon : \mathbb{R}^3 \to \mathbb{R}^3$ is associated with the shape velocity field $\mathfrak{V}(\varepsilon, x)$ given by (2.1).

We give more details concerning the specific form of the mapping $\mathfrak{T}_\varepsilon$. Let $\mathfrak{x} = \mathfrak{x}(t, X)$ denote the solution to the system of ordinary differential equations parametrized by the initial condition,

$$\frac{d}{dt}\mathfrak{x}(t, X) = \mathfrak{V}(t, \mathfrak{x}(t, X))\,,$$
$$\mathfrak{x}(0, X) = X\,,$$

then the mapping associated with the shape velocity field $\mathfrak{V}$ takes the form

$$\mathfrak{T}_\varepsilon(X) := \mathfrak{x}(\varepsilon, X)\,.$$

Furthermore, by our construction, the hold all domain is invariant for the mapping

$$B = \mathfrak{T}_\varepsilon(B)\,,$$

since $\mathfrak{T}_\varepsilon$ is reduced to the identity mapping in a neighborhood of the exterior boundary ∂B of the hold all domain B, under the assumption, e.g., that $\mathfrak{V} \cdot \mathbf{n} = 0$ on $\Sigma = \partial B$.

By construction the extended velocity field of Navier–Stokes equations satisfies $\mathbf{u}(\Omega_\varepsilon)(x) = 0$ on S_ε since on the boundary of the obstacle S_ε the non-slip boundary condition is prescribed for the velocity field $\mathbf{u}(\Omega_\varepsilon)$. If the shape derivative $\mathbf{u}'$ of the velocity field $\mathbf{u}(\Omega)$ in the direction of the field $\mathfrak{V}$ does exist, it is given by the following limit

$$\mathbf{u}'(x) := \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \left(\mathbf{u}(\Omega_\varepsilon)(x) - \mathbf{u}(\Omega)(x)\right) \qquad \text{in } B \setminus S\,,$$

with the limit passage $\varepsilon \to 0$ taken with respect to the weak or strong convergence in the associated Sobolev space, in our case the limit is taken with respect to the weak convergence only, we refer to [11] for all details.

Therefore, the limit is called *the weak (or the strong) shape derivative of* $\mathbf{u}(\Omega)(x)$, $x \in \Omega = B \setminus S$, *in the direction of the velocity field* $\mathfrak{V}(\varepsilon, x)$ *associated with the family of mappings* $\mathfrak{T}_\varepsilon$, the family of mappings being parameterized by the small parameter $\varepsilon \to 0$. In the similar way, the shape derivative of the density in the direction of the velocity field $\mathfrak{V}$ is defined by

$$\varrho'(x) := \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \left(\varrho(\Omega_\varepsilon)(x) - \varrho(\Omega)(x)\right) \qquad \text{in } B \setminus S\,.$$

2.7 Shape Gradients of Functionals

The knowledge of the shape derivatives $\mathbf{u}'(x)$ and $\varrho'(x)$ is sufficient to determine the shape gradient of the drag functional given by the formula

$$dJ(\Omega; \mathfrak{V}(0)) := \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \left(J(\Omega_\varepsilon) - J(\Omega)\right)\,,$$

and to obtain the first order necessary optimality conditions for the drag minimization problem. The structure of the shape gradient $\mathfrak{G}$ for the differentiable shape functional $J(\Omega)$ is characterized by the so-called Hadamard formula, we refer

to [15] for a simple proof of the result for C^2 domains, or to [1] for a proof of such a formula in the case of a class of nonsmooth domains.

Theorem 2.1. *If the boundary ∂S of the obstacle is C^2 and the mapping*

$$C(-\delta, \delta; C_0^1(B; \mathbb{R}^3)) \ni \mathfrak{V} \mapsto dJ(\Omega; \mathfrak{V}(0)) \in \mathbb{R}\,, \quad \delta > 0\,,$$

is continuous, then there is a distribution $\mathfrak{G} \in \mathcal{D}_1'(\partial S)$ supported on the boundary of the obstacle, such that

$$dJ(\Omega; \mathfrak{V}) = \langle \mathfrak{G}, \mathfrak{V}(0) \cdot \mathbf{n} \rangle_{\partial S}\,.$$

If the distribution $\mathfrak{G}$ is given by a function called the boundary gradient $G_S(\omega)$, then there is the boundary integral representation of the shape gradient

$$dJ(\Omega; \mathfrak{V}) = \int_{\partial S} G_S(\omega)\, (\mathfrak{V}(0, \omega) \cdot \mathbf{n}(\omega)) ds(\omega)\,, \tag{2.15}$$

which implies (2.9)*.*

It is shown in [9] that for the drag functional, the distribution $\mathfrak{G}$ in (2.9) actually is given by a function, the function is explicitly determined in terms of an appropriate adjoint state. The regularity of the shape gradient is an important issue e.g., from the point of view of numerical methods of shape optimization. Namely, if the shape gradient is given by a function, then the level set type numerical methods can be used for computations of an optimal shape, the shape gradient being a coefficient of the Hamilton–Jacobi equation for the level set function.

2.8 Material Derivatives of Solutions in Reference Domain

The existence of the shape derivatives of the velocity and the density in the Navier–Stokes equations combines the material derivatives and the formulae of the type (2.14) for the relation between the shape and the material derivatives. Therefore, it is convenient to introduce the material derivatives $\dot{\mathbf{u}}(\Omega; \mathfrak{V})$, $\dot{\varrho}(\Omega; \mathfrak{V})$ of $\mathbf{u}(\Omega_\varepsilon)(x)$, and $\varrho(\Omega_\varepsilon)(x)$, $x \in B$, which are intermediary objects determined in the fixed or reference domain $B \setminus S$. The word fixed means that the domain is shape parameter independent, the word reference means that small boundary variations of the fixed domain are considered. The knowledge of the material derivatives $(\dot{\mathbf{u}}, \dot{\varrho})$ is also sufficient in order to determine the shape gradient of the drag functional. In order to define the material derivatives we use the change of variables (5.7) parameterized by $\varepsilon \to 0$, denoted by $x \mapsto y(\varepsilon, x) := \mathfrak{T}_\varepsilon(x)$, with the property that $x \equiv y(0, x)$, and that the image of the obstacle S under this transformation is exactly the perturbed obstacle:

$$S_\varepsilon = y(\varepsilon, S)\,, \quad \text{and} \quad \Omega_\varepsilon = y(\varepsilon, \Omega)\,.$$

In this way, the composed extended velocity vector field

$$B \ni x \mapsto \mathbf{u}(\Omega_\varepsilon)(y(\varepsilon, x)) \in H^1(B)$$

vanishes on the fixed obstacle

$$\mathbf{u}(\Omega_\varepsilon)(y(\varepsilon, x)) = 0 \text{ for } x \in S\,.$$

The material derivative $\dot{\mathbf{u}}(x)$ of $\mathbf{u}(\Omega_\varepsilon)(y(\varepsilon, x))$ is given by the following limit if the limit exists

$$\dot{\mathbf{u}} = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \left(\mathbf{u}(\Omega_\varepsilon)(y(\varepsilon, x)) - \mathbf{u}(\Omega)(x)\right) \quad \text{in } B \setminus S\,. \tag{2.16}$$

In the same way, the material derivative of the density is given by the formula

$$\dot{\varrho} := \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \left(\varrho(\Omega_\varepsilon)(y(\varepsilon, x)) - \varrho(\Omega)(x)\right) \quad \text{in } B \setminus S\,. \tag{2.17}$$

By the same change of variables (5.7), denoted for the sake of simplicity of our notation by

$$x \mapsto y(x)\,,$$

so the dependence on ε is not explicitly indicated, the governing equations are transformed to the equations defined in the fixed domain $B \setminus S$. The material derivatives are then determined for the transformed equations by the stability theorem of solutions to governing equations in the reference domain with respect to the operator coefficients.

3 Decomposition of Shape Gradient

Shape derivatives of weak solutions leads to the decomposition of the shape gradient of the cost functional into its geometrical and dynamical components. This decomposition is interesting on its own, since the evaluation of the dynamical part of the shape gradient requires the solution of the appropriate adjoint state equations and becomes complicated for practical applications.

Let us consider now the drag functional $J(\Omega_\varepsilon)$ in the domain $\Omega_\varepsilon := B \setminus S_\varepsilon = \mathfrak{T}_\varepsilon(\Omega)$, where $\mathfrak{T}_\varepsilon : \Omega \to \Omega$ is a smooth mapping, $\varepsilon \to 0$ is a parameter, and the shape functional

$$J(\Omega_\varepsilon) := \mathbf{U}_\infty \cdot \mathbf{J}(S_\varepsilon) = \int_{\Omega_\varepsilon} \mathfrak{F}(\overline{\mathbf{u}}_\varepsilon, \nabla \overline{\mathbf{u}}_\varepsilon, \overline{p}_\varepsilon, \eta, \nabla \eta) dx\,. \tag{3.1}$$

The obtained formula for the shape gradient is justified in [11].

3.1 Geometrical and Dynamical Parts of Shape Gradient

Using the Reynolds transport theorem and the shape derivatives $\mathbf{u}', \varrho'$ of solutions to the state equations $\mathbf{u}, \varrho$ we obtain the shape gradient of the drag functional in the direction of the vector field

$$\mathfrak{V}(\varepsilon, x) = \left(\frac{\partial}{\partial \varepsilon}\mathfrak{T}_\varepsilon\right) \circ \mathfrak{T}_\varepsilon^{-1}(x)\,.$$

given by the expressionc

$$dJ(\Omega;\mathfrak{V}) := \int_{\partial S} \mathbf{U}_\infty \cdot [-\varrho \mathbf{u}\nabla \mathbf{u}]\,(\mathfrak{V}\cdot \mathbf{n}) ds(x) \tag{3.2}$$

$$+ \int_\Omega \mathbf{U}_\infty \cdot \left[-\left(\nabla \mathbf{u}' + (\nabla \mathbf{u}')^\top + (\lambda - 1)\,\mathrm{div}\,\mathbf{u}'\mathbb{I} - p'(\varrho)\varrho'\mathbb{I}\right)\nabla \eta\right] dx$$

$$+ \int_\Omega \mathbf{U}_\infty \cdot \left[-\eta(\varrho' \mathbf{u}\nabla \mathbf{u} + \varrho \mathbf{u}'\nabla \mathbf{u} + \varrho \mathbf{u}\nabla \mathbf{u}')\right] dx\,,$$

It is convenient to integrate by parts the terms depending on the first order derivatives of the shape derivative $\mathbf{u}'$,

$$-\int_\Omega \mathbf{U}_\infty \cdot \left[\left(\nabla \mathbf{u}' + (\nabla \mathbf{u}')^\top + (\lambda - 1)\,\mathrm{div}\,\mathbf{u}'\mathbb{I}\right)\nabla \eta\right] dx - \int_\Omega \mathbf{U}_\infty \cdot \left[\eta\varrho \mathbf{u}\nabla \mathbf{u}'\right] dx = \tag{3.3}$$

$$\int_\Omega \mathbf{u}' \cdot \mathrm{div}(\nabla \eta \otimes \mathbf{U}_\infty + \mathbf{U}_\infty \otimes \nabla \eta) + \mathbf{u}' \cdot \nabla(\mathrm{tr}(\nabla \eta \otimes \mathbf{U}_\infty))dx +$$

$$\int_\Omega \mathrm{div}(\eta\varrho \mathbf{u})\mathbf{u}' \cdot \mathbf{U}_\infty dx - \int_{\partial S} \varrho(\mathbf{u}\cdot \mathbf{n})(\mathbf{u}' \cdot \mathbf{U}_\infty) ds(x)\,,$$

and we denote

$$L_{\mathrm{dyn},2}(\mathbf{v}) := -\int_{\partial S} \varrho(\mathbf{u}\cdot \mathbf{n})(\mathbf{v}\cdot \mathbf{U}_\infty) ds(x)\,. \tag{3.4}$$

The shape derivative becomes

$$dJ(\Omega;\mathfrak{V}) := -\int\limits_{\partial S} \mathbf{U}_\infty \cdot [\eta\varrho\mathbf{u}\nabla\mathbf{u}]\,(\mathfrak{V}\cdot\mathbf{n})ds(x) \tag{3.5}$$

$$+\int\limits_{\Omega} p'(\varrho)\varrho'\nabla\eta\cdot\mathbf{U}_\infty dx - \int\limits_{\Omega}\eta\mathbf{U}_\infty\cdot\left[\varrho'\mathbf{u}\nabla\mathbf{u}+\varrho\mathbf{u}'\nabla\mathbf{u}\right]dx+$$

$$\int\limits_{\Omega}\mathbf{u}'\cdot\mathrm{div}(\nabla\eta\otimes\mathbf{U}_\infty+\ \mathbf{U}_\infty\otimes\nabla\eta)+\mathbf{u}'\cdot\nabla(\mathrm{tr}(\nabla\eta\otimes\ \mathbf{U}_\infty))dx+$$

$$\int\limits_{\Omega}\mathrm{div}(\eta\varrho\mathbf{u})\mathbf{u}'\cdot\ \mathbf{U}_\infty dx - \int\limits_{\partial S}\varrho(\mathbf{u}\cdot\mathbf{n})(\mathbf{u}'\cdot\mathbf{U}_\infty)ds(x)\,.$$

The following expression is called the geometrical part of the shape derivative

$$dJ_{\mathrm{geom}}(\Omega;\mathfrak{V}) := -\int\limits_{\partial S}\mathbf{U}_\infty\cdot[\varrho\mathbf{u}\nabla\mathbf{u}]\,(\mathfrak{V}\cdot\mathbf{n})ds(x)\,. \tag{3.6}$$

We introduce the right hand sides for the adjoint state equations, so we denote

$$L_{\mathrm{dens}}(\pi) := \int\limits_{\Omega} p'(\varrho)\pi\nabla\eta\cdot\mathbf{U}_\infty dx - \int\limits_{\Omega}\eta\mathbf{U}_\infty\cdot[\pi\mathbf{u}\nabla\mathbf{u}]\,dx\,, \tag{3.7}$$

and

$$L_{\mathrm{vel}}(\mathbf{v}) := -\int\limits_{\Omega}\eta\mathbf{U}_\infty\cdot[\varrho\mathbf{v}\nabla\mathbf{u}]\,dx+ \tag{3.8}$$

$$\int\limits_{\Omega}\mathbf{v}\cdot\mathrm{div}(\nabla\eta\otimes\mathbf{U}_\infty+\ \mathbf{U}_\infty\otimes\nabla\eta)+\mathbf{v}\cdot\nabla(\mathrm{tr}(\nabla\eta\otimes\ \mathbf{U}_\infty))dx+$$

$$\int\limits_{\Omega}\mathrm{div}(\eta\varrho\mathbf{u})\mathbf{v}\cdot\mathbf{U}_\infty dx\,.$$

The shape derivatives $\varrho' := \pi$ and $\mathbf{u}' := \mathbf{v}$ are given by the following system of linearized equations,

$$\mathrm{div}\,(\pi\,\mathbf{u}\otimes\mathbf{u}+\varrho\,\mathbf{v}\otimes\mathbf{u}+\varrho\,\mathbf{u}\otimes\mathbf{v})-\mathrm{div}\,\mathbb{S}(\mathbf{v})+\nabla p'(\varrho)\pi-\pi\,\mathbf{f}=0\quad\text{in }\ \Omega\,, \tag{3.9a}$$

$$\mathrm{div}\,(\pi\mathbf{u}+\varrho\mathbf{v})=0\quad\text{in }\ \Omega\,, \tag{3.9b}$$

$$\mathbf{v} = 0 \ \text{ on } \ \Sigma, \qquad \mathbf{v} = -\frac{\partial \mathbf{u}}{\partial n}(\mathfrak{V} \cdot \mathbf{n}) \ \text{ on } \ \partial S, \tag{3.9c}$$

$$\varrho = 0 \ \text{ on } \ \Sigma_{\text{in}}\,, \tag{3.9d}$$

where $\mathbb{S}(\mathbf{v}) = (\nabla \mathbf{v} + \nabla \mathbf{v}^{\top} + (\lambda - 1)\,\text{div}\,\mathbf{v}\mathbb{I})$.

3.2 Linearized and Adjoint State Equations

We multiply (2.3a) and (2.3b) by the smooth test functions $\boldsymbol{\phi}$, φ, respectively, $\boldsymbol{\phi}$ with compact support in Ω and φ which vanishes on Σ_{out}, and integrate by parts. It follows that

$$\int_{\Omega} [\text{div}\,(\varrho\,\mathbf{u} \otimes \mathbf{u}) - \text{div}\ \mathbb{S}(\mathbf{u}) + \nabla p(\varrho) - \varrho\,\mathbf{f}] \cdot \boldsymbol{\phi} dx =$$

$$\int_{\Omega} [-(\varrho\,\mathbf{u} \otimes \mathbf{u}) : D\boldsymbol{\phi} + \mathbb{S}(\mathbf{u}) : D\boldsymbol{\phi} - p(\varrho)\,\text{div}\ \boldsymbol{\phi} - \varrho\,\mathbf{f} \cdot \boldsymbol{\phi}]\,dx$$

and for the mass balance equation

$$\int_{\Omega} \text{div}\,(\varrho \mathbf{u})\varphi dx = -\int_{\Omega} \varrho \mathbf{v} \cdot \nabla \varphi dx\,.$$

Now, denote by $(\pi, \mathbf{v})$ a solution to the linearized system at the sufficiently smooth solution $(\varrho, \mathbf{u})$ of the nonlinear system. Hence the linearized balance of momentum system takes the form of the integral identities satisfied for all test functions $\boldsymbol{\phi}$,

$$\int_{\Omega} [-(\pi\,\mathbf{u} \otimes \mathbf{u}) : D\boldsymbol{\phi} - (\varrho\,\mathbf{v} \otimes \mathbf{u}) : D\boldsymbol{\phi} - (\varrho\,\mathbf{u} \otimes \mathbf{v}) : D\boldsymbol{\phi}]\,dx$$

$$+ \int_{\Omega} \left[\mathbb{S}(\mathbf{v}) : D\boldsymbol{\phi} - p'(\varrho)\pi\,\text{div}\ \boldsymbol{\phi} - \pi\,\mathbf{f} \cdot \boldsymbol{\phi}\right] dx\,.$$

The linearized balance of momentum system is satisfied by the shape derivatives $(\varrho', \mathbf{u}')$.

It is also useful to perform the integration by parts for the functions $\mathbf{v}$ which vanish on ∂B and are non-null on the obstacle boundary ∂S, this leads to

$$\int\limits_\Omega \mathbb{S}(\mathbf{v}) : D\boldsymbol{\phi}\, dx = \int\limits_\Omega \left[\nabla \mathbf{v} + (\nabla \mathbf{v})^\top\right] : D\boldsymbol{\phi}\, dx + (\lambda - 1) \int\limits_\Omega \operatorname{div} \mathbf{v} \operatorname{tr}(D\boldsymbol{\phi})\, dx =$$

$$-\int\limits_\Omega \mathbf{v} \cdot \operatorname{div}\left[D\boldsymbol{\phi} + D\boldsymbol{\phi}^\top\right] dx + \int\limits_{\partial S} [\mathbf{v} \otimes n + n \otimes \mathbf{v}] : D\boldsymbol{\phi} ds(x)$$

$$-(\lambda - 1) \int\limits_\Omega \mathbf{v} \cdot \nabla [\operatorname{tr}(D\boldsymbol{\phi})]\, dx + \int\limits_{\partial S} (\mathbf{v} \cdot \mathbf{n}) \operatorname{tr}(D\boldsymbol{\phi}) ds(x)\,.$$

We denote by

$$L_{\mathrm{dyn},1}(\mathbf{v}) := \int\limits_{\partial S} [\mathbf{v} \otimes \mathbf{n} + \mathbf{n} \otimes \mathbf{v}] : D\boldsymbol{\phi} ds(x) \tag{3.10}$$

$$+ \int\limits_{\partial S} (\mathbf{v} \cdot \mathbf{n}) \operatorname{tr}(D\boldsymbol{\phi}) ds(x)\,.$$

the boundary integrals on ∂S which furnish one part of the dynamical shape gradient $L_{\mathrm{dyn},1}(\mathbf{u}')$. It is clear that $L_{\mathrm{dyn},1}(\mathbf{v})$ depends only on the trace of $\mathbf{v}$ on the obstacle boundary ∂S, and the expression is nontrivial when used with the shape derivative $\mathbf{v} := \mathbf{u}'$. There are the nonhomogeneous Dirichlet conditions on ∂S for the shape derivative of the velocity field $\mathbf{u}' = -\dfrac{\partial \mathbf{u}}{\partial n}(\mathfrak{V} \cdot \mathbf{n})$, such conditions result from the homogeneous Dirichlet condition for the velocity field $\mathbf{u} = 0$ prescribed on ∂S.

In the same way, for the mass balance equation and all test functions φ,

$$\int\limits_\Omega [-\pi \mathbf{u} \cdot \nabla \varphi - \varrho \mathbf{v} \cdot \nabla \varphi]\, dx = 0\,,$$

where φ is a smooth test function, and $\varphi = 0$ on Σ_{out}.

We introduce the following notation for the bilinear forms defined for linearized operators, evaluated for the smooth functions such that, $\mathbf{v} = 0$ on $\partial\Omega$, and $\pi = 0$ on Σ_{in}, and defined for all smooth test functions $\varphi, \boldsymbol{\phi}$ which satisfy the boundary conditions $\varphi = 0$ on Σ_{out}, and $\boldsymbol{\phi} = 0$ on $\partial\Omega$. The first bilinear form is associated with the linearized momentum balance equations,

$$\langle \mathcal{L}_1(\pi, \mathbf{v}), (\varphi, \boldsymbol{\phi}) \rangle := \int\limits_\Omega \left[-p'(\varrho)\pi \operatorname{div} \boldsymbol{\phi} - \pi\, \mathbf{f} \cdot \boldsymbol{\phi}\right] dx$$

$$+ \int\limits_\Omega [-(\pi\, \mathbf{u} \otimes \mathbf{u}) : D\boldsymbol{\phi} - (\varrho\, \mathbf{v} \otimes \mathbf{u}) : D\boldsymbol{\phi} - (\varrho\, \mathbf{u} \otimes \mathbf{v}) : D\boldsymbol{\phi}]\, dx$$

$$-\int\limits_\Omega \mathbf{v} \cdot \operatorname{div}\left[D\boldsymbol{\phi} + D\boldsymbol{\phi}^\top\right] dx - (\lambda - 1) \int\limits_\Omega \mathbf{v} \cdot \nabla [\operatorname{tr}(D\boldsymbol{\phi})]\, dx\,.$$

We denote also by

$$\langle \mathcal{L}_2(\pi, \mathbf{v}), (\varphi, \boldsymbol{\phi}) \rangle = \int_\Omega [-\pi \mathbf{u} \cdot \nabla \varphi - \varrho \mathbf{v} \cdot \nabla \varphi] \, dx = 0$$

the bilinear form associated with the linearized mass balance equation.

Now we take the sum of bilinear forms and define its decomposition in order to identify the adjoint operators $\mathcal{L}_\pi$ and $\mathcal{L}_\mathbf{v}$,

$$\langle \mathcal{L}_\pi(\varphi, \boldsymbol{\phi}), \pi \rangle + \langle \mathcal{L}_\mathbf{v}(\varphi, \boldsymbol{\phi}), \mathbf{v} \rangle := \langle \mathcal{L}_1(\pi, \mathbf{v}), (\varphi, \boldsymbol{\phi}) \rangle + \langle \mathcal{L}_2(\pi, \mathbf{v}), (\varphi, \boldsymbol{\phi}) \rangle \, . \tag{3.11}$$

In view of decomposition (3.11) we can define the following adjoint operators, the first is obtained for $\pi := 0$ in (3.11),

$$\langle \mathcal{L}_\mathbf{v}(\varphi, \boldsymbol{\phi}), \mathbf{v} \rangle := \tag{3.12}$$

$$- \int_\Omega [\varrho \mathbf{v} \cdot \nabla \varphi + (\varrho \, \mathbf{v} \otimes \mathbf{u}) : D\boldsymbol{\phi} + (\varrho \, \mathbf{u} \otimes \mathbf{v}) : D\boldsymbol{\phi}] \, dx$$

$$- \int_\Omega \mathbf{v} \cdot \operatorname{div} \left[D\boldsymbol{\phi} + D\boldsymbol{\phi}^\top \right] dx - (\lambda - 1) \int_\Omega \mathbf{v} \cdot \nabla \left[\operatorname{tr} (D\boldsymbol{\phi}) \right] dx \, ,$$

then the second is obtained for $\mathbf{v} := 0$ in (3.11),

$$\langle \mathcal{L}_\pi(\varphi, \boldsymbol{\phi}), \pi \rangle = \tag{3.13}$$

$$- \int_\Omega \left[\pi \mathbf{u} \cdot \nabla \varphi + (\pi \, \mathbf{u} \otimes \mathbf{u}) : D\boldsymbol{\phi} + p'(\varrho) \pi \operatorname{div} \boldsymbol{\phi} + \pi \, \mathbf{f} \cdot \boldsymbol{\phi} \right] dx \, .$$

Finally, the adjoint state equations are introduced:

Find φ and $\boldsymbol{\phi}$ with $\varphi = 0$ on Σ_{out} and $\boldsymbol{\phi}$ on $\partial\Omega$, such that

$$\langle \mathcal{L}_\pi(\varphi, \boldsymbol{\phi}), \pi \rangle = L_{\text{dens}}(\pi) \ \text{ for all test functions } \ \pi \, , \tag{3.14}$$

$$\langle \mathcal{L}_\mathbf{v}(\varphi, \boldsymbol{\phi}), \mathbf{v} \rangle = L_{\text{vel}}(\mathbf{v}) \ \text{ for all test functions } \ \mathbf{v} \, , \tag{3.15}$$

where the bilinear forms are defined by (3.12) and (3.13). The smooth test functions satisfy the following boundary conditions, $\pi = 0$ on Σ_{in}, $\mathbf{v} = 0$ on $\partial\Omega$.

Now, let us note that by the adjoint state equations we have the identity

$$L_{\text{dens}}(\varrho') + L_{\text{vel}}(\mathbf{u}') = \langle \mathcal{L}_\pi(\varphi, \boldsymbol{\phi}), \varrho' \rangle + \langle \mathcal{L}_\mathbf{v}(\varphi, \boldsymbol{\phi}), \mathbf{u}' \rangle \, , \tag{3.16}$$

and by the linearized equations written for the shape derivatives $(\varrho', \mathbf{u}')$ it follows that we have the second identity

$$\langle \mathcal{L}_\pi(\varphi, \boldsymbol{\phi}), \varrho' \rangle + \langle \mathcal{L}_{\mathbf{v}}(\varphi, \boldsymbol{\phi}), \mathbf{u}' \rangle = \tag{3.17}$$
$$\langle \mathcal{L}_1(\varrho', \mathbf{u}'), (\varphi, \boldsymbol{\phi}) \rangle + \langle \mathcal{L}_2(\varrho', \mathbf{u}'), (\varphi, \boldsymbol{\phi}) \rangle + L_{\mathrm{dyn},1}(\mathbf{u}')) \, .$$

We can combine the above equalities and as a result the dynamical part of the shape gradient is obtained in the form

$$dJ_{\mathrm{dyn}}(\Omega; \mathbf{T}) = L_{\mathrm{dens}}(\varrho') + L_{\mathrm{vel}}(\mathbf{u}') + L_{\mathrm{dyn},2}(\mathbf{u}') = \tag{3.18}$$
$$\langle \mathcal{L}_\pi(\varphi, \boldsymbol{\phi}), \varrho' \rangle + \langle \mathcal{L}_{\mathbf{v}}(\varphi, \boldsymbol{\phi}), \mathbf{u}' \rangle + L_{\mathrm{dyn},2}(\mathbf{u}') =$$
$$L_{\mathrm{dyn},1}(\mathbf{u}') + L_{\mathrm{dyn},2}(\mathbf{u}') = -\int\limits_{\partial S} \left[\frac{\partial \mathbf{u}}{\partial n} \otimes \mathbf{n} + \mathbf{n} \otimes \frac{\partial \mathbf{u}}{\partial n} \right] : D\, \boldsymbol{\phi}\, (\mathfrak{V} \cdot \mathbf{n})\, ds(x)$$
$$- \int\limits_{\partial S} \left[\frac{\partial \mathbf{u}}{\partial n} \cdot \mathbf{n} \right] \, \mathrm{tr}\, (D\boldsymbol{\phi})(\mathfrak{V} \cdot \mathbf{n})\, ds(x) + \int\limits_{\partial S} \varrho(\mathbf{u} \cdot \mathbf{n}) \left[\frac{\partial \mathbf{u}}{\partial n} \cdot \mathbf{U}_\infty \right] (\mathfrak{V} \cdot \mathbf{n})\, ds(x) \, .$$

4 Shape Sensitivity Analysis of Navier–Stokes Equations

4.1 Preliminaries

In order to perform the shape sensitivity analysis of the work functional for the non-stationary equations, first, the framework is established. In the governing equations, most of physical constants are posed to be equal to one, therefore the only constant is $\lambda > 0$. It is also assumed at this stage of analysis that there is no intersection between the inlet and the outlet on the boundary of the hold all domain B. The boundary variations technique is applied in order to investigate the dependence of the shape functional $J(\Omega_\varepsilon)$ on the shape of the obstacle S_ε in the variable domain $\Omega_\varepsilon = B \setminus S_\varepsilon$ for $\varepsilon \to 0$.

The tools we are going to discuss in this chapter include the shape derivatives of the solutions to the non-stationary, compressible Navier–Stokes equations, the shape gradient of the work functional $J(\Omega)$ and its decompositions into the geometrical and dynamical parts, and the adjoint state equations associated with the dynamical part of the shape gradient. The proofs of the results are given in [11] in the case of local solutions to the stationary, compressible Navier–Stokes equations.

4.2 Navier–Stokes State Equations

In this chapter we are going to consider the general model.

The state equation defined in the reference domain $\Omega \times (0, T)$ takes the form

$$-\partial_t(\varrho\,\mathbf{u}) + \Delta\mathbf{u} + \lambda\nabla\,\mathrm{div}\,\mathbf{u} = \varrho\mathbf{u}\cdot\nabla\mathbf{u} + \nabla p(\varrho) - \mathbb{C}\mathbf{u} + \varrho\mathbf{f} \text{ in } \Omega\times(0,T)\,, \tag{4.1a}$$

$$\partial_t\varrho + \mathrm{div}(\varrho\mathbf{u}) = 0 \text{ in } \Omega\times(0,T)\,, \tag{4.1b}$$

$$\begin{aligned}
\mathbf{u} &= 0 \text{ on } \partial S\times(0,T),\\
\mathbf{u} &= \mathbf{U} \text{ on } \partial B\times(0,T),\\
\varrho &= \varrho_\infty \text{ on } \Sigma_{\mathrm{in}}\,,\\
\mathbf{u}(x,0) &= \mathbf{u}_0(x) \text{ in } \Omega,\\
\varrho(x,0) &= \varrho_0(x) \text{ in } \Omega\,.
\end{aligned} \tag{4.1c}$$

It is also convenient to introduce the *effective viscous pressure*

$$q = p(\varrho) - \lambda\,\mathrm{div}\,\mathbf{u},$$

and rewrite the state equation in the equivalent form useful for numerical methods,

$$-\partial_t(\varrho\,\mathbf{u}) + \Delta\mathbf{u} - \nabla q = \varrho\mathbf{u}\cdot\nabla\mathbf{u} - \mathbb{C}\mathbf{u} + \varrho\mathbf{f} \text{ in } \Omega\times(0,T)\,, \tag{4.2a}$$

$$\mathrm{div}\,\mathbf{u} = \frac{1}{\lambda}p(\varrho) - \frac{1}{\lambda}q\,, \text{ in } \Omega\times(0,T)\,, \tag{4.2b}$$

$$\partial_t\varrho + \mathrm{div}(\varrho\mathbf{u}) = 0 \text{ in } \Omega\times(0,T)\,, \tag{4.2c}$$

$$\begin{aligned}
\mathbf{u} &= 0 \text{ on } \partial S\times(0,T),\\
\mathbf{u} &= \mathbf{U} \text{ on } \partial B\times(0,T),\\
\varrho &= \varrho_\infty \text{ on } \Sigma_{\mathrm{in}}\,,\\
\mathbf{u}(x,0) &= \mathbf{u}_0(x) \text{ in } \Omega,\\
\varrho(x,0) &= \varrho_0(x) \text{ in } \Omega\,.
\end{aligned} \tag{4.2d}$$

4.3 Linearized and Adjoint State Equations: Material and Shape Derivatives of Solutions

Material and shape derivatives of solutions to the governing equations are given by solutions to the appropriate linearized equations. We are going to derive the equations for the shape derivatives of solutions to the Navier–Stokes equations.

For the sake of simplicity we assume that the intersection of the inlet $\overline{\Sigma}_{\mathrm{in}}$ with the outlet $\overline{\Sigma}_{\mathrm{out}}$ is empty. We assume also that the primal variables π and $\mathbf{v}$ for the density and the velocity in the linearized equations vanish for $t = 0$ and on Σ_{in} and

$\partial\Omega$, respectively. The only exception from the homogeneous initial and boundary conditions is the nonhomogeneous Dirichlet condition for the shape derivative $\mathbf{u}'$ of the velocity field on the obstacle boundary ∂S. The dual variables denoted by φ and $\boldsymbol{\phi}$ for the density and the velocity vanish on Σ_{out} and on $\partial\Omega$, respectively.

In order to differentiate the solutions of the state equations with respect to the shape the linearized and the adjoint equations are introduced. To this end it is convenient to rewrite equations (4.1a) and (4.1b) in the following form

$$\partial_t(\varrho\,\mathbf{u}) + \operatorname{div}(\varrho\,\mathbf{u}\otimes\mathbf{u}) - \operatorname{div}\mathbb{S}(\mathbf{u}) + \nabla p(\varrho) + \mathbb{C}\,\mathbf{u} - \varrho\,\mathbf{f} = 0 \quad \text{in } \Omega\times(0,T), \tag{4.3}$$

$$\partial_t\varrho + \operatorname{div}(\varrho\mathbf{u}) = 0 \quad \text{in } \Omega\times(0,T), \tag{4.4}$$

$$\text{where } \mathbb{S}(\mathbf{u}) = (\nabla\mathbf{u} + \nabla\mathbf{u}^{\top} + (\lambda-1)\operatorname{div}\mathbf{u}\mathbb{I})\,. \tag{4.5}$$

We multiply (4.3) and (4.4) by the smooth test functions $\boldsymbol{\phi}$, φ, respectively, $\boldsymbol{\phi}$ with compact support in Ω and φ which vanishes on Σ_{out}, and integrate by parts. It follows that

$$\int_0^T\int_\Omega [\partial_t(\varrho\,\mathbf{u}) + \operatorname{div}(\varrho\,\mathbf{u}\otimes\mathbf{u}) - \operatorname{div}\mathbb{S}(\mathbf{u}) + \nabla p(\varrho) + \mathbb{C}\,\mathbf{u} - \varrho\,\mathbf{f}]\cdot\boldsymbol{\phi}\,dxdt =$$

$$\int_0^T\int_\Omega \Big(-\varrho\,\mathbf{u}\cdot\partial_t\boldsymbol{\phi} - (\varrho\,\mathbf{u}\otimes\mathbf{u}) : D\boldsymbol{\phi} + \mathbb{S}(\mathbf{u}) : D\boldsymbol{\phi} - p(\varrho)\operatorname{div}\boldsymbol{\phi} +$$

$$(\mathbb{C}\,\mathbf{u} - \varrho\,\mathbf{f})\cdot\boldsymbol{\phi}\Big)dxdt + \int_\Omega [\varrho(T)\,\mathbf{u}(T)\cdot\boldsymbol{\phi}(T) - \varrho(0)\,\mathbf{u}(0)\cdot\boldsymbol{\phi}(0)]\,dx$$

and for the mass balance equation

$$\int_0^T\int_\Omega [\partial_t\varrho + \operatorname{div}(\varrho\mathbf{u})]\,\varphi dxdt =$$

$$\int_0^T\int_\Omega [-\varrho\partial_t\varphi - \varrho\mathbf{v}\cdot\nabla\varphi]\,dxdt + \int_\Omega [\varrho(T)\,\varphi(T) - \varrho(0)\,\varphi(0)]\,dx\,.$$

Now, denote by $(\pi,\mathbf{v})$ a solution to the linearized system at the sufficiently smooth solution $(\varrho,\mathbf{u})$ of the nonlinear system, the same linearized system is derived for the so-called shape derivatives $(\varrho',\mathbf{u}')$, hence we obtain the integral identities satisfied for all test functions $\boldsymbol{\phi}$,

$$\int_0^T \int_\Omega \Big(- \varrho\, \mathbf{v} \cdot \partial_t \boldsymbol{\phi} - \pi\, \mathbf{u} \cdot \partial_t \boldsymbol{\phi} - (\pi\, \mathbf{u} \otimes \mathbf{u}) : D\boldsymbol{\phi} - (\varrho\, \mathbf{v} \otimes \mathbf{u}) : D\boldsymbol{\phi} -$$

$$(\varrho\, \mathbf{u} \otimes \mathbf{v}) : D\boldsymbol{\phi} \Big) dxdt + \int_0^T \int_\Omega \Big(\mathbb{S}(\mathbf{v}) : D\boldsymbol{\phi} - p'(\varrho)\pi \operatorname{div} \boldsymbol{\phi} +$$

$$(\mathbb{C}\, \mathbf{v} - \pi\, \mathbf{f}) \cdot \boldsymbol{\phi} \Big) dxdt + \int_\Omega \Big(\pi(T)\, \mathbf{u}(T) \cdot \boldsymbol{\phi}(T) + \varrho(T)\, \mathbf{v}(T) \cdot \boldsymbol{\phi}(T) -$$

$$\pi(0)\, \mathbf{u}(0) \cdot \boldsymbol{\phi}(0) - \varrho(0)\, \mathbf{v}(0) \cdot \boldsymbol{\phi}(0) \Big) dx = 0\,.$$

It is also useful to perform the integration by parts for the test functions $\mathbf{v}$ which vanish on ∂B and are non-null on the obstacle boundary ∂S, this leads to

$$\int_0^T \int_\Omega \mathbb{S}(\mathbf{v}) : D\boldsymbol{\phi}\, dxdt = \int_0^T \int_\Omega \left[\nabla \mathbf{v} + (\nabla \mathbf{v})^\top \right] : D\boldsymbol{\phi}\, dxdt +$$

$$(\lambda - 1) \int_0^T \int_\Omega \operatorname{div} \mathbf{v}\, \operatorname{tr}(D\boldsymbol{\phi})\, dxdt = - \int_0^T \int_\Omega \mathbf{v} \cdot \operatorname{div} \left[D\boldsymbol{\phi} + D\boldsymbol{\phi}^\top \right] dxdt +$$

$$\int_0^T \int_{\partial S} [\mathbf{v} \otimes n + n \otimes \mathbf{v}] : D\boldsymbol{\phi} ds(x)dt - (\lambda - 1) \int_0^T \int_\Omega \mathbf{v} \cdot \nabla \left[\operatorname{tr}(D\boldsymbol{\phi}) \right] dxdt +$$

$$\int_0^T \int_{\partial S} (\mathbf{v} \cdot \mathbf{n})\, \operatorname{tr}(D\boldsymbol{\phi}) ds(x)dt\,.$$

We denote by

$$L_{\mathrm{dyn},1}(\mathbf{v}) := \int_0^T \int_{\partial S} [\mathbf{v} \otimes n + n \otimes \mathbf{v}] : D\boldsymbol{\phi} ds(x)dt \tag{4.6}$$

$$+ \int_0^T \int_{\partial S} (\mathbf{v} \cdot \mathbf{n})\, \operatorname{tr}(D\boldsymbol{\phi}) ds(x)dt\,.$$

the boundary integrals on ∂S which furnish one part of the dynamical shape gradient. It is clear that $L_{\mathrm{dyn},1}(\mathbf{v})$ depends only on the trace of $\mathbf{v}$ on the obstacle boundary ∂S, and the expression is nontrivial when used with the shape derivative

$\mathbf{v} := \mathbf{u}'$. There are the nonhomogeneous Dirichlet conditions on ∂S for the shape derivative of the velocity field $\mathbf{u}' = -\dfrac{\partial \mathbf{u}}{\partial n}(\mathbf{T}\cdot\mathbf{n})$, such conditions result from the homogeneous Dirichlet condition for the velocity field $\mathbf{u} = 0$ prescribed on ∂S.

In the same way the linearized equation is obtained for the mass balance equation,

$$\int\limits_0^T \int\limits_\Omega [-\pi\,\partial_t\varphi - \pi\mathbf{u}\cdot\nabla\varphi - \varrho\mathbf{v}\cdot\nabla\varphi]\,dxdt + \int\limits_\Omega [\pi(T)\,\varphi(T) - \pi(0)\,\varphi(0)]\,dx = 0\,.$$

for all test functions φ.

We introduce the following notation for the bilinear forms defined for linearized operators, acting on the smooth functions such that, $\mathbf{v} = 0$ on $\partial\Omega$, and $\pi = 0$ on Σ_{in},

$$\langle \mathcal{L}_1(\pi,\mathbf{v}),(\varphi,\boldsymbol{\phi})\rangle := \int\limits_0^T \int\limits_\Omega \Big(-\varrho\,\mathbf{v}\cdot\partial_t\boldsymbol{\phi} -$$

$$\pi\,\mathbf{u}\cdot\partial_t\boldsymbol{\phi} - (\pi\,\mathbf{u}\otimes\mathbf{u}) : D\boldsymbol{\phi} - (\varrho\,\mathbf{v}\otimes\mathbf{u}) : D\boldsymbol{\phi} - (\varrho\,\mathbf{u}\otimes\mathbf{v}) : D\boldsymbol{\phi}\Big)dxdt$$

$$-\int\limits_0^T \int\limits_\Omega \mathbf{v}\cdot\operatorname{div}\Big(D\boldsymbol{\phi} + D\boldsymbol{\phi}^\top\Big)dxdt - (\lambda-1)\int\limits_0^T \int\limits_\Omega \mathbf{v}\cdot\nabla[\operatorname{tr}(D\boldsymbol{\phi})]\,dxdt$$

$$+\int\limits_0^T \int\limits_\Omega \Big(-p'(\varrho)\pi\operatorname{div}\boldsymbol{\phi} + (\mathbb{C}\,\mathbf{v} - \pi\,\mathbf{f})\cdot\boldsymbol{\phi}\Big)dxdt +$$

$$\int\limits_\Omega \Big(\pi(T)\,\mathbf{u}(T)\cdot\boldsymbol{\phi}(T) + \varrho(T)\,\mathbf{v}(T)\cdot\boldsymbol{\phi}(T) -$$

$$\pi(0)\,\mathbf{u}(0)\cdot\boldsymbol{\phi}(0) - \varrho(0)\,\mathbf{v}(0)\cdot\boldsymbol{\phi}(0)\Big)dx\,,$$

the above expression can be slightly simplified assuming in addition that the initial values for $t = 0$ also vanish, $\pi(0) = 0$ and $\mathbf{v}(0) = 0$.

We denote also

$$\langle \mathcal{L}_2(\pi,\mathbf{v}),(\varphi,\boldsymbol{\phi})\rangle =$$

$$\int\limits_0^T \int\limits_\Omega [-\pi\,\partial_t\varphi - \pi\mathbf{u}\cdot\nabla\varphi - \varrho\mathbf{v}\cdot\nabla\varphi]\,dxdt + \int\limits_\Omega [\pi(T)\,\varphi(T) - \pi(0)\,\varphi(0)]\,dx = 0\,.$$

Now we take the sum of bilinear form and decompose in the following way in order to identify the adjoint operators

$$\langle \mathcal{L}_\pi(\varphi, \boldsymbol{\phi}), \pi \rangle + \langle \mathcal{L}_{\mathbf{v}}(\varphi, \boldsymbol{\phi}), \mathbf{v} \rangle := \langle \mathcal{L}_1(\pi, \mathbf{v}), (\varphi, \boldsymbol{\phi}) \rangle + \langle \mathcal{L}_2(\pi, \mathbf{v}), (\varphi, \boldsymbol{\phi}) \rangle \,. \tag{4.7}$$

Assuming that for $t = 0$, we have $\pi(0) = 0$, $\mathbf{v}(0) = 0$, and that the values of $\varphi(T)$ and $\boldsymbol{\phi}(T)$ are prescribed, in view of decomposition (4.7) we define the following adjoint operators, first for $\pi := 0$ in (4.7),

$$\langle \mathcal{L}_{\mathbf{v}}(\varphi, \boldsymbol{\phi}), \mathbf{v} \rangle := \int_\Omega \varrho(T)\, \mathbf{v}(T) \cdot \boldsymbol{\phi}(T) dx - \tag{4.8}$$

$$\int_0^T \int_\Omega \left[\varrho \mathbf{v} \cdot \nabla \varphi + \varrho\, \mathbf{v} \cdot \partial_t \boldsymbol{\phi} + (\varrho\, \mathbf{v} \otimes \mathbf{u}) : D\boldsymbol{\phi} + (\varrho\, \mathbf{u} \otimes \mathbf{v}) : D\boldsymbol{\phi} - \mathbb{C}\, \mathbf{v} \cdot \boldsymbol{\phi} \right] dxdt$$

$$- \int_0^T \int_\Omega \mathbf{v} \cdot \operatorname{div} \left[D\boldsymbol{\phi} + D\boldsymbol{\phi}^\top \right] dxdt - (\lambda - 1) \int_0^T \int_\Omega \mathbf{v} \cdot \nabla \left[\operatorname{tr} (D\boldsymbol{\phi}) \right] dxdt \,,$$

then for $\mathbf{v} := 0$ in (4.7),

$$\langle \mathcal{L}_\pi(\varphi, \boldsymbol{\phi}), \pi \rangle = \int_\Omega \pi(T)\, \varphi(T) dx - \tag{4.9}$$

$$\int_0^T \int_\Omega \Big(\pi\, \partial_t \varphi + \pi \mathbf{u} \cdot \nabla \varphi + \pi\, \mathbf{u} \cdot \partial_t \boldsymbol{\phi} +$$

$$(\pi\, \mathbf{u} \otimes \mathbf{u}) : D\boldsymbol{\phi} + p'(\varrho) \pi \operatorname{div} \boldsymbol{\phi} + \pi\, \mathbf{f} \cdot \boldsymbol{\phi} \Big) dxdt \,. \tag{4.10}$$

5 Decomposition of the Shape Gradient

The decomposition of the shape gradient into the geometrical and dynamical components seems to be useful for the numerical methods of shape optimization. The first component of this decomposition describes the direct influence of the geometry variations on the variations of the functional. The second dynamical component of this decomposition measures the influence of the variations of solutions to the governing equations resulting form the geometry variations on the variations of the shape functional. The dynamical component actually depends on the shape derivatives of solutions with respect to the boundary variations.

5.1 Work Shape Functional

Let us consider a shape functional depending on $\Omega = B \setminus S$ and defined to be the work $J(\Omega) := W_S$ of hydrodynamic forces acting on a moving obstacle S

$$J(\Omega) = -\int\limits_{\Omega} \eta \Big\{ (\varrho \mathbf{u} \cdot \mathbf{W})(x,T) - \varrho_0(x)\mathbf{U}(x) \cdot \mathbf{W}(x,0) \Big\} dx + \int\limits_0^T \int\limits_{\Omega} \Big\{ \varrho \eta \mathbf{u} \cdot \partial_t \mathbf{W} + \big(\varrho(\mathbf{u} \otimes \mathbf{u}) - \mathbb{T}\big) : \nabla(\eta \mathbf{W}) + \eta\big(\varrho \mathbf{f} - \mathbb{C}\mathbf{u}\big) \cdot \mathbf{W} \Big\} dxdt, \tag{5.1}$$

where η is a smooth function with a compact support which contains the obstacle boundary ∂S, and $\eta(x) \equiv 1$ in a vicinity of the obstacle boundary, furthermore

$$\mathbb{T} = \nabla \mathbf{u} + (\nabla \mathbf{u})^{\top} + (\lambda - 1) \operatorname{div} \mathbf{u} \mathbb{I} - p(\varrho)\, \mathbb{I}.$$

Recall that if an obstacle S moves in atmosphere like a solid body then its physical position S_t at time t is defined by

$$S_t = \mathbb{U}(t)S + \mathbf{a}(t). \tag{5.2}$$

Here a unitary matrix $\mathbb{U}$ and a vector field $\mathbf{a}(t)$ can be defined by an appropriate flight planning scenario. In this case the vector field $\mathbf{W}$ is given by formulae

$$\mathbf{W}(x,t) = \mathbb{U}^{\top}(t)\dot{\mathbb{U}}(t)x + \mathbb{U}^{\top}(t)\dot{\mathbf{a}}(t), \tag{5.3}$$

and we have

$$\mathbb{C}\mathbf{u} = \operatorname{rot} \mathbf{W} \times \mathbf{u}.$$

It is convenient for our purposes to introduce the following notation for the shape functional

$$J(\Omega) := \int\limits_{\Omega} \mathfrak{f}(\varrho_0, \mathbf{U}, \mathbf{W}(0), \eta) dx - \int\limits_{\Omega} \mathfrak{f}(\varrho(T), \mathbf{u}(T), \mathbf{W}(T), \eta) dx + \int\limits_0^T \int\limits_{\Omega} \mathfrak{F}(\varrho, \mathbf{u}, \mathbb{T}, \mathbf{W}, \partial_t \mathbf{W}, \nabla \mathbf{W}, \mathbf{f}, \eta, \nabla \eta) dxdt\,, \tag{5.4}$$

where

$$\mathfrak{f}(\varrho, \mathbf{u}, \mathbf{W}, \eta) := \eta(x)\varrho(x)\mathbf{u}(x) \cdot \mathbf{W}(x), \tag{5.5}$$

$$\mathfrak{F}(\varrho, \mathbf{u}, \mathbb{T}, \mathbf{W}, \partial_t \mathbf{W}, \nabla \mathbf{W}, \mathbf{f}, \eta, \nabla\eta) := \eta\varrho\mathbf{u}\cdot\partial_t\mathbf{W} + \tag{5.6}$$
$$\eta\big(\varrho(\mathbf{u}\otimes\mathbf{u}) - \mathbb{T}\big) : \nabla\mathbf{W} + \big(\varrho(\mathbf{u}\otimes\mathbf{u}) - \mathbb{T}\big) : (\nabla\eta\otimes\mathbf{W}) + \eta\big(\varrho\mathbf{f} - \mathbb{C}\mathbf{u}\big)\cdot\mathbf{W}.$$

5.2 State Equation and Shape Functional in a Variable Domain

We are going to apply the boundary variations technique [15], however in the framework adapted to our problem. The reference domain $\Omega = B\setminus S$ is transformed onto the perturbed domain $\Omega_\varepsilon = B\setminus S_\varepsilon$, $\varepsilon\to 0$, by means of the change of variables

$$y = x + \varepsilon\mathbf{T}(x), \quad x\in\Omega, \quad y\in\Omega_\varepsilon, \tag{5.7}$$

with the appropriate vector field $\mathbf{T}$ supported in a neighborhood of the obstacle S, since we are only interested in the boundary variations of the obstacle S. Therefore, the state equation in Ω_ε takes the same form (4.1) with Ω replaced by Ω_ε, with the new unknown functions denoted by $(\bar\varrho_\varepsilon, \bar{\mathbf{u}}_\varepsilon)$, and with $\bar q_\varepsilon := p(\bar\varrho_\varepsilon) - \lambda\,\mathrm{div}\,\bar{\mathbf{u}}_\varepsilon$, the new unknown functions depend on the small parameter $\varepsilon\to 0$ which is omitted in the equations,

$$-\partial_t(\varrho\,\mathbf{u}) + \Delta\mathbf{u} + \lambda\nabla\,\mathrm{div}\,\mathbf{u} = \varrho\mathbf{u}\cdot\nabla\mathbf{u} + \nabla p(\varrho) - \mathbb{C}\mathbf{u} + \varrho\mathbf{f} \ \text{ in } \ \Omega_\varepsilon\times(0,T)\,, \tag{5.8a}$$

$$\partial_t\varrho + \mathrm{div}(\varrho\mathbf{u}) = 0 \ \text{ in } \ \Omega_\varepsilon\times(0,T)\,, \tag{5.8b}$$

$$\begin{aligned}
\mathbf{u} &= 0 \ \text{ on } \ \partial S_\varepsilon\times(0,T),\\
\mathbf{u} &= \mathbf{U} \ \text{ on } \ \partial B\times(0,T),\\
\varrho &= \varrho_\infty \ \text{ on } \ \Sigma_{\mathrm{in}}\,,\\
\mathbf{u}(x,0) &= \mathbf{u}_0(x) \ \text{ in } \ \Omega_\varepsilon,\\
\varrho(x,0) &= \varrho_0(x) \ \text{ in } \ \Omega_\varepsilon\,.
\end{aligned} \tag{5.8c}$$

The expression for the shape functional in $\Omega_\varepsilon := B\setminus S_\varepsilon$ takes the form

$$J(\Omega_\varepsilon) := \int_{\Omega_\varepsilon} \mathfrak{f}(\varrho_0, \mathbf{U}, \mathbf{W}(0), \eta)dx - \int_{\Omega_\varepsilon} \mathfrak{f}(\overline{\varrho}_\varepsilon(T), \overline{\mathbf{u}}_\varepsilon(T), \mathbf{W}(T), \eta)dx + \tag{5.9}$$
$$\int_0^T\int_{\Omega_\varepsilon} \mathfrak{F}(\overline{\varrho}_\varepsilon, \overline{\mathbf{u}}_\varepsilon, \overline{\mathbb{T}}_\varepsilon, \mathbf{W}, \partial_t\mathbf{W}, \nabla\mathbf{W}, \mathbf{f}, \eta, \nabla\eta)dxdt\,,$$

where the functions $\varrho_0(x)$, $\mathbf{f}(x,t)$, $x \in \Omega_\varepsilon, t \in (0,T)$, are given by the restrictions to Ω_ε of the functions which are defined for $x \in \mathbb{R}^d$, the plan of the flight for the deformed obstacle S_ε takes the form of the matrix function given by (5.3), now $\mathbf{W}$ depends on $x \in \Omega_\varepsilon$. The functions $\overline{\varrho}_\varepsilon$, $\overline{\mathbf{u}}_\varepsilon$, $\overline{\mathbb{T}}_\varepsilon$ are given by the solutions to state equations, therefore the functions depend implicitly on $\varepsilon \to 0$.

5.3 *Shape Derivatives of Solutions*

We are interested in the form of the derivative for the mapping

$$\varepsilon \mapsto J(\Omega_\varepsilon).$$

From the general theory of shape optimization [15] it follows by the Hadamard structure theorem of the shape gradient 2.1 that the first order derivative of this mapping, under appropriate regularity assumptions on the domain and on the solutions is of the form

$$\int\limits_0^T \int\limits_{\partial S} \mathfrak{G}(\mathbf{T}\cdot\mathbf{n}) ds(x) = \lim_{\varepsilon\to 0} \frac{J(\Omega_\varepsilon) - J(\Omega)}{\varepsilon}, \tag{5.10}$$

where the shape gradient $\mathfrak{G}$ is in general given by a distribution which lives on the boundary. In [9, 11] it is shown that for the drag functional the shape gradient is given by a function G_S.

Therefore, $\mathfrak{G}$ is the so-called shape gradient of the shape functional $J(\Omega)$ in the direction of the vector field $\mathbf{T}$. There are two distinct parts of the shape gradient, the geometrical part, and the dynamical part, it is shown in [9] that the geometrical part of the shape gradient vanishes in the case of the drag functional.

Actually, the shape gradient $\mathfrak{G}$ can be decomposed into two parts, one which is easy to evaluate numerically which we call the geometrical part, and another which is very difficult to evaluate by numerical methods which is called the dynamical part. The reason is that in order to evaluate the dynamical part, it is required to solve not only the state equation, but also the so-called adjoint state equation which is a linearized variant of the state equations depending on the right hand sides on the derivative of the shape functional. This type of decomposition is easy for the linear problems which are well posed, and extremely difficult for the nonlinear problems we consider in the monograph. Now, we explain briefly such a decomposition as well as the concepts of the material and shape derivatives for the solutions of our state equation. In the description it is assumed that the solutions are sufficiently smooth which in a specific application should be justified.

The dynamical part of the shape gradient contains the so-called shape derivatives of the fields $\varrho, \mathbf{u}$. The remaining part of the shape gradient is called the geometrical part. Roughly speaking, the dynamical part of the shape gradient takes into account only the variations of the state with respect to the boundary variations of the

obstacle. In other words, the geometrical part of the shape gradient is obtained for the solutions of the state equation replaced by restrictions to the variable domain of given functions defined e.g., in all the hold all domain, which means that the shape derivatives of the density and of the velocity are set to be null. The decomposition into two parts of the shape gradient is obtained at the final stage of our procedure.

In order to deduce the shape gradient $\mathfrak{G}$ some methods are available. One possibility is the change of variables $\Omega_\varepsilon \ni y(x) = x + \varepsilon \mathbf{T}(x) \mapsto x \in \Omega$ in the state equation, and derivation in the reference domain Ω with respect to $\varepsilon \to 0$. Now, we give some details on the change of variables. It is convenient to change also the unknown velocity field using the Piola transformation in the following way

$$\mathbf{u}_\varepsilon(x,t) = \mathbb{N}\overline{\mathbf{u}}_\varepsilon(x + \varepsilon \mathbf{T}(x), t), \quad \varrho_\varepsilon(x,t) = \overline{\varrho}_\varepsilon(x + \varepsilon \mathbf{T}(x), t), \tag{5.11}$$

where we denote

$$\mathbb{N}(x) = \det(\mathbb{I} + \varepsilon \mathrm{D}\mathbf{T}(x))(\mathbb{I} + \varepsilon \mathrm{D}\mathbf{T}(x))^{-1}. \tag{5.12}$$

The new unknown functions $\Omega \ni x \to (\mathbf{u}_\varepsilon(x), \varrho_\varepsilon(x)) \in \mathbb{R}^{d+1}$ are defined in the fixed reference domain, therefore the mapping $\varepsilon \mapsto (\mathbf{u}_\varepsilon, \varrho_\varepsilon)$ can be differentiated in classical way in an appropriate function space. We use the following notation for the derivatives with respect to the shape parameter $\varepsilon \to 0$, the limits are taken with respect to the strong or the weak convergence in an appropriate function space, the functions are extended to the hold all domain B if necessary, for the evaluation of the shape derivatives,

- Derivatives of the solutions to the state equation

$$\frac{d\varrho_\varepsilon}{d\varepsilon}(x,t) := \lim_{\varepsilon \to 0} \frac{\varrho_\varepsilon(x,t) - \varrho(x,t)}{\varepsilon}, \tag{5.13}$$

$$\frac{d\mathbf{u}_\varepsilon}{d\varepsilon}(x,t) := \lim_{\varepsilon \to 0} \frac{\mathbf{u}_\varepsilon(x,t) - \mathbf{u}(x,t)}{\varepsilon}, \tag{5.14}$$

- Material derivatives of the solutions to the state equation

$$\dot{\varrho}(x,t) := \lim_{\varepsilon \to 0} \frac{\overline{\varrho}_\varepsilon(x + \varepsilon \mathbf{T}(x), t) - \varrho(x,t)}{\varepsilon}, \tag{5.15}$$

$$\dot{\mathbf{u}}(x,t) := \lim_{\varepsilon \to 0} \frac{\overline{\mathbf{u}}_\varepsilon(x + \varepsilon \mathbf{T}(x), t) - \mathbf{u}(x,t)}{\varepsilon}, \tag{5.16}$$

- Shape derivatives of the solutions to the state equation

$$\varrho'(x,t) := \lim_{\varepsilon \to 0} \frac{\overline{\varrho}_\varepsilon(x,t) - \varrho(x,t)}{\varepsilon}, \tag{5.17}$$

$$\mathbf{u}'(x,t) := \lim_{\varepsilon \to 0} \frac{\overline{\mathbf{u}}_\varepsilon(x,t) - \mathbf{u}(x,t)}{\varepsilon}. \tag{5.18}$$

There is natural decomposition of the material derivatives $\dot{\varrho}$, $\dot{\mathbf{u}}$ into the shape derivatives ϱ', $\mathbf{u}'$ and the remainders [15], which can be written formally for $(x,t) \in \Omega \times (0,T)$ as follows:

$$\dot{\varrho}(x,t) := \varrho'(x,t) + \nabla \varrho(x,t) \cdot \mathbf{T}(x)\,, \tag{5.19}$$

$$\dot{\mathbf{u}}(x,t) := \mathbf{u}'(x,t) + \nabla \mathbf{u}(x,t)\mathbf{T}(x)\,, \tag{5.20}$$

but this decomposition is unfortunately difficult to be used in our context in order to determine the shape derivatives.

Remark 5.1. Notice that the functions $\overline{\varrho}_\varepsilon(x,t)$ and $\overline{\mathbf{u}}_\varepsilon(x,t)$ are only defined for $x \in \Omega_\varepsilon$, where $\varepsilon \to 0$, therefore the limits (5.17) and (5.18) are well defined in the open set Ω. □

Remark 5.2. Formally, the equations for the shape derivatives $(\varrho', \mathbf{u}')$ are obtained by the linearization of the state equation at the reference domain $\Omega \times (0,T)$, and the formal system is of the following form

$$-\partial_t(\varrho'\,\mathbf{u}) - \partial_t(\varrho\,\mathbf{u}') + \Delta \mathbf{u}' - \nabla q' = \tag{5.21a}$$

$$\varrho'\mathbf{u}\cdot\nabla\mathbf{u} + \varrho\mathbf{u}'\cdot\nabla\mathbf{u} + \varrho\mathbf{u}\cdot\nabla\mathbf{u}' - \mathbb{C}'\mathbf{u} - \mathbb{C}\mathbf{u}' + \varrho'\mathbf{f} + \varrho\mathbf{f}' \ \text{ in } \ \Omega\times(0,T)\,,$$

$$\operatorname{div}\mathbf{u}' = \frac{1}{\lambda}p'(\varrho)\varrho' - \frac{1}{\lambda}q' \ \text{ in } \ \Omega\times(0,T)\,, \tag{5.21b}$$

$$\partial_t\varrho' + \operatorname{div}(\varrho'\mathbf{u}) + \operatorname{div}(\varrho\mathbf{u}') = 0 \ \text{ in } \ \Omega\times(0,T)\,, \tag{5.21c}$$

$$\mathbf{u}' = -\frac{\partial \mathbf{u}}{\partial n}(\mathbf{T}\cdot\mathbf{n}) \ \text{ on } \ \partial S\times(0,T), \tag{5.21d}$$

$$\mathbf{u}' = \mathbf{U}' \ \text{ on } \ \partial B\times(0,T), \tag{5.21e}$$

$$\varrho' = \varrho'_\infty \ \text{ on } \ \Sigma_{\text{in}}\,, \tag{5.21f}$$

$$\mathbf{u}'(x,0) = \mathbf{u}'_0(x) \ \text{ in } \ \Omega\,, \tag{5.21g}$$

$$\varrho'(x,0) = \varrho'_0(x) \ \text{ in } \ \Omega\,, \tag{5.21h}$$

where we assume that the data in the state equation in the reference domain $\mathbf{f}, \mathbf{U}, \varrho_\infty$ admit the shape derivatives $\mathbb{C}', \mathbf{f}', \mathbf{U}', \varrho'_\infty, \mathbf{u}'_0, \varrho'_0$. □

5.4 *Geometrical and Dynamical Components of Shape Gradient Decomposition*

Now, since the data of our state equations are fixed, then the shape derivatives of the data $\mathbb{C}', \mathbf{f}', \mathbf{U}', \varrho'_\infty, \mathbf{u}'_0, \varrho'_0$ are null, and formal differentiation of the shape functional (5.9), making use of the Reynolds transport theorem, leads to

$$dJ(\Omega;\mathbf{T}) := \lim_{\varepsilon\to 0}\frac{J(\Omega_\varepsilon)-J(\Omega)}{\varepsilon} = dJ_{\text{geom}}(\Omega;\mathbf{T}) + dJ_{\text{dyn}}(\Omega;\mathbf{T})\,, \tag{5.22}$$

with the dynamical part of the shape derivative

$$\begin{aligned} dJ_{\text{dyn}}(\Omega;\mathbf{T}) := &\int_\Omega \varrho'(T)\mathfrak{f}_\varrho(\varrho(T),\mathbf{u}(T),\mathbf{W}(T),\eta)dx- \\ &\int_\Omega \mathbf{u}'(T)\cdot\mathfrak{f}_{\mathbf{u}}(\varrho(T),\mathbf{u}(T),\mathbf{W}(T),\eta)dx+ \\ &\int_0^T\int_\Omega \varrho'\mathfrak{F}_\varrho(\varrho,\mathbf{u},\mathbb{T},\mathbf{W},\partial_t\mathbf{W},\nabla\mathbf{W},\mathbf{f},\eta,\nabla\eta)dxdt+ \\ &\int_0^T\int_\Omega \mathbf{u}'\cdot\mathfrak{F}_{\mathbf{u}}(\varrho,\mathbf{u},\mathbb{T},\mathbf{W},\partial_t\mathbf{W},\nabla\mathbf{W},\mathbf{f},\eta,\nabla\eta)dxdt+ \\ &\int_0^T\int_\Omega \mathbb{T}' : \mathfrak{F}_{\mathbb{T}}(\varrho,\mathbf{u},\mathbb{T},\mathbf{W},\partial_t\mathbf{W},\nabla\mathbf{W},\mathbf{f},\eta,\nabla\eta)dxdt\,, \end{aligned} \tag{5.23}$$

and, in view of (5.5), (5.6), we obtain

$$\begin{aligned} dJ_{\text{dyn}}(\Omega;\mathbf{T}) := &\int_\Omega \eta(x)\varrho'(x,T)\mathbf{u}(x,T)\cdot\mathbf{W}(x,T)dx- \\ &\int_\Omega \eta(x)\varrho(x,T)\mathbf{u}'(x,T)\cdot\mathbf{W}(x,T)dx+ \\ &\int_0^T\int_\Omega \Big(\eta\varrho'\mathbf{u}\cdot\partial_t\mathbf{W}+\eta\big(\varrho'\mathbf{u}\otimes\mathbf{u}\big):\nabla\mathbf{W}+\varrho'(\mathbf{u}\otimes\mathbf{u}):(\nabla\eta\otimes\mathbf{W}) \\ &+\varrho'\mathbf{f}\cdot\mathbf{W}\Big)dxdt+\int_0^T\int_\Omega \Big(\eta\varrho\mathbf{u}'\cdot\partial_t\mathbf{W}+\eta\varrho\big((\mathbf{u}'\otimes\mathbf{u})+(\mathbf{u}\otimes\mathbf{u}')\big):(\nabla\mathbf{W}+ \\ &\nabla\eta\otimes\mathbf{W})-\eta\mathbb{C}\mathbf{u}'\cdot\mathbf{W}\Big)dxdt\int_0^T\int_\Omega \mathbb{T}':\nabla(\eta\mathbf{W})dxdt\,. \end{aligned} \tag{5.24}$$

The geometrical part of the shape derivative takes the form

$$dJ_{\text{geom}}(\Omega;\mathbf{T}) := \int\limits_{\partial S} \mathfrak{f}(\varrho_0,\mathbf{U},\mathbf{W}(0),\eta)(\mathbf{T}\cdot\mathbf{n})ds(x) - \tag{5.25}$$

$$\int\limits_{\partial S} \mathfrak{f}(\varrho(T),\mathbf{u}(T),\mathbf{W}(T),\eta)(\mathbf{T}\cdot\mathbf{n})ds(x) -$$

$$\int\limits_0^T\int\limits_{\partial S} \mathfrak{F}(\varrho,\mathbf{u},\mathbb{T},\mathbf{W},\partial_t\mathbf{W},\nabla\mathbf{W},\mathbf{f},\eta,\nabla\eta)(\mathbf{T}\cdot\mathbf{n})ds(x)dt\,,$$

where we denote

$$\mathbb{T}' = \mathbb{S}' - p'(\varrho)\varrho'\,\mathbb{I}. \tag{5.26}$$

$$\mathbb{S}' = \nabla\mathbf{u}' + (\nabla\mathbf{u}')^{\top} + (\lambda-1)\,\text{div}\,\mathbf{u}'\mathbb{I}\,. \tag{5.27}$$

□

We develop further the last term in (5.24) in view of (5.26), (5.27), so we have

$$\int\limits_0^T\int\limits_{\Omega} \mathbb{T}' : \mathfrak{F}_{\mathbb{T}}(\varrho,\mathbf{u},\mathbb{T},\mathbf{W},\partial_t\mathbf{W},\nabla\mathbf{W},\mathbf{f},\eta,\nabla\eta)dxdt =$$

$$\int\limits_0^T\int\limits_{\Omega} \mathbb{T}' : \nabla(\eta\mathbf{W})dxdt = \int\limits_0^T\int\limits_{\Omega} \left[\nabla\mathbf{u}' + (\nabla\mathbf{u}')^{\top}\right] : \nabla(\eta\mathbf{W})dxdt + \tag{5.28}$$

$$(\lambda-1)\int\limits_0^T\int\limits_{\Omega} \text{div}\;\mathbf{u}'\;\text{tr}\,(\nabla(\eta\mathbf{W}))dxdt - \tag{5.29}$$

$$\int\limits_0^T\int\limits_{\Omega} p'(\varrho)\varrho'\;\text{tr}\,(\nabla(\eta\mathbf{W}))dxdt = \tag{5.30}$$

$$-\int\limits_0^T\int\limits_{\Omega} \mathbf{u}'\cdot\text{div}\left[\nabla(\eta\mathbf{W}) + \nabla(\eta\mathbf{W})^{\top}\right]dxdt +$$

$$\int\limits_0^T\int\limits_{\partial S} \left[\mathbf{u}'\otimes n + n\otimes\mathbf{u}'\right] : \nabla(\eta\mathbf{W})ds(x)dt$$

$$-(\lambda-1)\int_0^T\int_\Omega \mathbf{u}'\cdot\nabla\left[\operatorname{tr}\left(\nabla(\eta\mathbf{W})\right)\right]dxdt+$$

$$\int_0^T\int_{\partial S}\mathbf{u}'\cdot\mathbf{n}\operatorname{tr}\left(\nabla(\eta\mathbf{W})\right)ds(x)dt+$$

$$\int_0^T\int_\Omega p'(\varrho)\varrho'\operatorname{tr}\left(\nabla(\eta\mathbf{W})\right)dxdt\,.$$

From the above expression we can deduce the second part of the dynamical shape gradient,

$$L_{\mathrm{dyn},2}(\mathbf{v}):= \tag{5.31}$$

$$\int_0^T\int_{\partial S}[\mathbf{v}\otimes n+n\otimes\mathbf{v}]:\nabla(\eta\mathbf{W})ds(x)dt+\int_0^T\int_{\partial S}\mathbf{v}\cdot\mathbf{n}\operatorname{tr}\left(\nabla(\eta\mathbf{W})\right)ds(x)dt\,.$$

5.5 Adjoint State Equations

The dynamical part of the shape derivative is further simplified by introduction of an appropriate adjoint state equations. We introduce two linear forms in order to decompose the dynamical part of the shape gradient. The first linear form for the density shape derivative,

$$L_{\mathrm{dens}}(\pi):=\int_\Omega\eta(x)\pi(x,T)\mathbf{u}(x,T)\cdot\mathbf{W}(x,T)dx- \tag{5.32}$$

$$\int_0^T\int_\Omega\Big(\eta\pi\mathbf{u}\cdot\partial_t\mathbf{W}+\eta\big(\pi\mathbf{u}\otimes\mathbf{u}\big):\nabla\mathbf{W}+$$

$$\pi(\mathbf{u}\otimes\mathbf{u}):(\nabla\eta\otimes\mathbf{W})+\pi\mathbf{f}\cdot\mathbf{W}\Big)dxdt+\int_0^T\int_\Omega p'(\varrho)\pi\operatorname{tr}\left(\nabla(\eta\mathbf{W})\right)dxdt\,.$$

and the second linear form for the velocity shape derivative

$$L_{\mathrm{vel}}(\mathbf{v}):=\int_\Omega\eta(x)\varrho(x,T)\mathbf{v}(x,T)\cdot\mathbf{W}(x,T)dx+ \tag{5.33}$$

$$\int\limits_0^T \int\limits_\Omega \Big(\eta \varrho \mathbf{v} \cdot \partial_t \mathbf{W} + \eta \varrho \big((\mathbf{v} \otimes \mathbf{u}) + (\mathbf{u} \otimes \mathbf{v}) \big) : (\nabla \mathbf{W} + \nabla \eta \otimes \mathbf{W})$$

$$- \eta \mathbb{C} \mathbf{v} \cdot \mathbf{W} \Big) dxdt - \int\limits_0^T \int\limits_\Omega \mathbf{v} \cdot \operatorname{div} \big[\nabla(\eta \mathbf{W}) + \nabla(\eta \mathbf{W})^\top \big] dxdt$$

$$- (\lambda - 1) \int\limits_0^T \int\limits_\Omega \mathbf{v} \cdot \nabla \left[\operatorname{tr} (\nabla(\eta \mathbf{W})) \right] dxdt \, .$$

With the notation we have

$$dJ_{\text{dyn}}(\Omega; \mathbf{T}) := L_{\text{dens}}(\varrho') + L_{\text{vel}}(\mathbf{u}') + L_{\text{dyn},2}(\mathbf{u}') \tag{5.34}$$

Now, the adjoint state equations are defined as follows:

Find φ and $\boldsymbol{\phi}$ such that $\varphi = 0$ on $\Sigma_{\text{out}} \times (0, T)$ and $\boldsymbol{\phi}$ on $\partial\Omega \times (0, T)$,

$$\langle \mathcal{L}_\pi(\varphi, \boldsymbol{\phi}), \pi \rangle = L_{\text{dens}}(\pi) \ \text{ for all test functions } \ \pi \, , \tag{5.35}$$

$$\langle \mathcal{L}_{\mathbf{v}}(\varphi, \boldsymbol{\phi}), \mathbf{v} \rangle = L_{\text{vel}}(\mathbf{v}) \ \text{ for all test functions } \ \mathbf{v} \, , \tag{5.36}$$

$$\varphi(T) = (\eta - 1)\mathbf{u}(T) \cdot \mathbf{W}(T), \quad \boldsymbol{\phi}(T) = \eta \mathbf{W}(T) \, , \tag{5.37}$$

where the bilinear forms are defined by (4.8) and (4.9). The smooth test functions satisfy the following boundary conditions, $\pi = 0$ on $\Sigma_{\text{in}} \times (0, T)$, $\mathbf{v} = 0$ on $\partial\Omega \times (0, T)$, $\mathbf{v}(0) = 0$ and $\pi(0) = 0$ in Ω.

Now, let us note that by the adjoint state equations we have the identity

$$L_{\text{dens}}(\varrho') + L_{\text{vel}}(\mathbf{u}') = \langle \mathcal{L}_\pi(\varphi, \boldsymbol{\phi}), \varrho' \rangle + \langle \mathcal{L}_{\mathbf{v}}(\varphi, \boldsymbol{\phi}), \mathbf{u}' \rangle \, , \tag{5.38}$$

and by the linearized equations for the shape derivatives it follows that we have the second identity

$$\langle \mathcal{L}_\pi(\varphi, \boldsymbol{\phi}), \varrho' \rangle + \langle \mathcal{L}_{\mathbf{v}}(\varphi, \boldsymbol{\phi}), \mathbf{u}' \rangle = \tag{5.39}$$

$$\langle \mathcal{L}_1(\varrho', \mathbf{u}'), (\varphi, \boldsymbol{\phi}) \rangle + \langle \mathcal{L}_2(\varrho', \mathbf{u}'), (\varphi, \boldsymbol{\phi}) \rangle +$$

$$\int\limits_0^T \int\limits_{\partial S} \left[\frac{\partial \mathbf{u}}{\partial n} (\mathbf{T} \cdot \mathbf{n}) \otimes n + n \otimes \frac{\partial \mathbf{u}}{\partial n} \right] : D\boldsymbol{\phi}(\mathbf{T} \cdot \mathbf{n}) ds(x) dt$$

$$+ \int\limits_0^T \int\limits_{\partial S} \left(\frac{\partial \mathbf{u}}{\partial n} \cdot \mathbf{n} \right) \operatorname{tr} (D\boldsymbol{\phi})(\mathbf{T} \cdot \mathbf{n}) ds(x) dt \, ,$$

We can combine the above equalities and in this way we obtain the dynamical part of the shape gradient in the form

$$
\begin{aligned}
dJ_{\mathrm{dyn}}(\Omega;\mathbf{T}) = L_{\mathrm{dens}}(\varrho') + L_{\mathrm{vel}}(\mathbf{u}') + L_{\mathrm{dyn},2}(\mathbf{u}') = \qquad (5.40)\\
\langle \mathcal{L}_{\pi}(\varphi,\boldsymbol{\phi}),\varrho'\rangle + \langle \mathcal{L}_{\mathbf{v}}(\varphi,\boldsymbol{\phi}),\mathbf{u}'\rangle + L_{\mathrm{dyn},2}(\mathbf{u}') =\\
\int_0^T\int_{\partial S}\left[\frac{\partial \mathbf{u}}{\partial n}\otimes n + n\otimes\frac{\partial \mathbf{u}}{\partial n}\right] : D\boldsymbol{\phi}\,(\mathbf{T}\cdot\mathbf{n})ds(x)dt +\\
\int_0^T\int_{\partial S}\left(\frac{\partial \mathbf{u}}{\partial n}\cdot\mathbf{n}\right)\,\mathrm{tr}\,(D\boldsymbol{\phi})(\mathbf{T}\cdot\mathbf{n})ds(x)dt -\\
\int_0^T\int_{\partial S}\left[\frac{\partial \mathbf{u}}{\partial n}\otimes n + n\otimes\frac{\partial \mathbf{u}}{\partial n}\right] : \nabla(\mathbf{W})(\mathbf{T}\cdot\mathbf{n})ds(x)dt -\\
\int_0^T\int_{\partial S}\frac{\partial \mathbf{u}}{\partial n}\cdot\mathbf{n}\;\mathrm{tr}\,(\nabla(\mathbf{W}))(\mathbf{T}\cdot\mathbf{n})ds(x)dt\,,
\end{aligned}
$$

where the element $\boldsymbol{\phi}$ is given by the solution of the adjoint state equations (5.35)–(5.37).

Acknowledgements The authors acknowledge support by the European Science Foundation within the Programme 'Optimization with PDE Constraints'.

Jan Sokolowski is supported by the Brazilian Research Council (CNPq), through the Special Visitor Researcher Framework of the Science Without Borders.

References

[1] G. Frémiot, W. Horn, A. Laurain, M. Rao, J. Sokołowski, On the analysis of boundary value problems in nonsmooth domains. Dissertationes Math. **462**, 149 (2009)

[2] A. Kaźmierczak, P.I. Plotnikov, J. Sokołowski, A. Żochowski, Numerical method for drag minimization in compressible flows, in 15th International Conference on Methods and Models in Automation and Robotics (MMAR), (pp. 97–101) (2009). doi:10.1109/MMAR.2010.5587258

[3] P.I. Plotnikov, J. Sokołowski, Stationary boundary value problems for Navier-Stokes equations with adiabatic exponent $\gamma < 3/2$ (in Russian). Dokl. Akad. Nauk **397**, 166–169 (2004)

[4] P.I. Plotnikov, J. Sokołowski, On compactness, domain dependence and existence of steady state solutions to compressible isothermal Navier-Stokes equations. J. Math. Fluid Mech. **7**, 529–573 (2005)

[5] P.I. Plotnikov, J. Sokołowski, Concentrations of solutions to time-discretized compressible Navier -Stokes equations. Comm. Math. Phys. **258**, 567–608 (2005)

[6] P.I. Plotnikov, J. Sokołowski, Domain dependence of solutions to compressible Navier-Stokes equations. SIAM J. Control Optim. **45**, 1165–1197 (2006)

[7] P.I. Plotnikov, J. Sokołowski, Stationary solutions of Navier-Stokes equations for diatomic gases (in Russian). Uspekhi Mat. Nauk **62**, 117–148 (2007)
[8] P.I. Plotnikov, J. Sokołowski, Stationary boundary value problems for compressible Navier-Stokes equations, in *Handbook of Differential Equations: Stationary Partial Differential Equations*, vol. VI (Elsevier/North-Holland, Amsterdam, 2008), pp. 313–410
[9] P.I. Plotnikov, J. Sokołowski, Shape derivative of drag functional. SIAM J. Control Optim. **48**, 4680–4706 (2010)
[10] P.I. Plotnikov, J. Sokołowski, Inhomogeneous boundary value problem for nonstationary compressible Navier-Stokes equations. J. Math. Sci. **170**, 34–130 (2010)
[11] P.I. Plotnikov, J. Sokolowski, *Compressible Navier-Stokes Equations. Theory and Shape Optimization* (Birkhäuser, Basel, 2012)
[12] P.I. Plotnikov, E.V. Ruban, J. Sokołowski, Inhomogeneous boundary value problems for compressible Navier-Stokes equations, well-posedness and sensitivity analysis. SIAM J. Math. Anal. **40**, 1152–1200 (2008)
[13] P.I. Plotnikov, E.V. Ruban, J. Sokołowski, Inhomogeneous boundary value problems for compressible Navier-Stokes and transport equations. J. Math. Pures Appl. **92**, 113–162 (2009)
[14] P. Plotnikov, J. Sokołowski, A. Żochowski, Numerical experiments in drag minimization for compressible Navier-Stokes flows in bounded domains, in *Proceedings of the 14th International IEEE/IFAC Conference on Methods and Models in Automation and Robotics* (MMAR'09, 2009), 4 pp
[15] J. Sokołowski, J.-P. Zolésio, *Introduction to Shape Optimization. Shape Sensitivity Analysis* (Springer, Berlin/Heidelberg/New York, 1992)

Controllability of Navier–Stokes Equations

Jean-Pierre Puel

Abstract These notes correspond to a course taught at BCAM, Bilbao in March-April 2012 and at CMM, Santiago de Chile in November 2012. Most of them follow the lines of published articles on the subject, essentially the basic article by Fernandez-Cara, Guerrero, Imanuvilov, Puel (J. Math. Pures Appl. 83:1501–1542) and the article by Imanuvilov, Puel, Yamamoto (Chin. Ann. Math. 30:333–378, 2009), and also an article by Imanuvilov, Puel, Yamamoto (Carleman estimates for second order non homogeneous parabolic equations) which is not yet published. But, strictly speaking, the results presented here are new in the sense that they are not written anywhere in this generality. Moreover the method of proof is different from the ones given in the above articles.

Keywords Carleman estimate • Local exact controllability • Navier–Stokes equations

Mathematics Subject Classification (2010). Primary 65K15; Secondary 49M99, 65K15.

1 Introduction

The present monograph intends to give a precise description of the latest method used to study the local exact controllability of Navier–Stokes equations. It relies on a new global Carleman estimate for the Stokes equations for the linearized problem and it follows essentially the lines of [3] in a slightly different functional class for the nonlinear problem. The result obtained here seems quite optimal and is stronger than the one in [6] and a little stronger than the one obtained in [3].

A relevant objective for the control of a viscous fluid is not straightforward as the system is dissipative and not reversible. The notion of Exact Controllability to

J.-P. Puel (✉)
Laboratoire de Mathématiques de Versailles, Université de Versailles Saint-Quentin, Versailles, France
e-mail: jppuel@math.uvsq.fr

R. Hoppe (ed.), *Optimization with PDE Constraints*, Lecture Notes in Computational Science and Engineering 101,
DOI 10.1007/978-3-319-08025-3_12

Trajectories seems to be the most interesting one and it is described in the first chapter. In the second chapter we treat the linearized problem, which corresponds to a null controllability problem and requires the main mathematical tools. The third chapter deals with the full nonlinear problem, which is studied in a quite complex functional class even if it is simplified from the one considered in [3]. The argument uses regularity properties for the Stokes system and some fine interpolation results.

2 Exact Controllability to Trajectories

2.1 Introduction

The physical domain is a regular bounded connected open set Ω of $\mathbb{R}^N$ with $N = 2$ or $N = 3$, whose boundary is denoted by Γ. We denote by ν the unit exterior normal vector at a point of Γ. The time variable t will be taken in the interval $(0, T)$ with $T > 0$. We consider the following controlled Navier–Stokes system in vectorial form

$$\frac{\partial y}{\partial t} - \Delta y + (y.\nabla)y + \nabla p = f + v\mathbb{1}_\omega \text{ in } \Omega \times (0, T), \tag{2.1}$$

$$\text{div } y = 0 \text{ in } \Omega \times (0, T), \tag{2.2}$$

$$y = 0 \text{ on } \Gamma \times (0, T), \tag{2.3}$$

$$y(0) = y_0 \text{ in } \Omega. \tag{2.4}$$

Here y is the velocity of a viscous incompressible fluid satisfying the no-slip boundary condition on the boundary, p is the pressure, y_0 is the initial velocity of the fluid, f is a given external force exerted on the fluid and v is the control acting on a (small) subdomain ω of the physical domain Ω, ω being a non empty open subset of Ω. We consider here the case of a distributed control in order to avoid additional technicalities, but the case of a boundary control, which can be seen as more realistic, can be treated a posteriori from the present situation. For simplicity we have taken the viscosity to be equal to 1. Even if the present description is for the moment formal, we already have to say that, due to the fact that uniqueness is an open problem for Navier–Stokes equations in dimension 3, for a given control v we cannot speak of the solution $y(v)$ but of one solution $y(v)$.

The controllability question is then to try to describe the set of reachable states at time T, i.e. the set $\{y(v, T)\}$ when the control v varies in an appropriate functional class. This question cannot be answered exactly here. The dissipative character of the system and the "regularizing" effect imply that we cannot expect exact controllability for this system, which would mean for example that, starting from an initial data in some Hilbert space H, we could reach any target in H by choosing adequately the control v. This is hopeless here. Approximate controllability is here a relevant question. This question corresponds to the following: given any target

$y_1 \in H$, and any neighborhood of y_1 in H, can we choose a control v such that $y(v, T)$ reaches this neighborhood? This is an interesting problem which has the drawback that it does not say what to do after time T in order to stay close to the target. Anyway we will not consider this question here.

Another controllability objective can be introduced. The idea is that by controlling the motion of a viscous incompressible fluid we can only hope to obtain another motion of a viscous incompressible fluid following the same physical laws. This is the basis of what is now called **exact controllability to trajectories** (ECT) which is described below.

Let us consider an "ideal" (uncontrolled) trajectory of the same operator $(\bar{y}, \bar{p})$ starting from a different initial data $\bar{y}_0$ and satisfying (for simplicity we take the same external force f for the moment)

$$\frac{\partial \bar{y}}{\partial t} - \Delta \bar{y} + (\bar{y}.\nabla)\bar{y} + \nabla \bar{p} = f \text{ in } \Omega \times (0, T), \tag{2.5}$$

$$\operatorname{div} \bar{y} = 0 \text{ in } \Omega \times (0, T), \tag{2.6}$$

$$\bar{y} = 0 \text{ on } \Gamma \times (0, T), \tag{2.7}$$

$$\bar{y}(0) = \bar{y}_0 \text{ in } \Omega. \tag{2.8}$$

The global exact controllability to trajectory question is the following: can we find a control v such that we have at time T

$$y(v, T) = \bar{y}(T) \ ?$$

The local version of this question is the following: does there exist $\eta > 0$ such that if $||y_0 - \bar{y}_0||_X \leq \eta$, we can find a control v such that

$$y(v, T) = \bar{y}(T) \ ?$$

Of course if one of these properties occur, the two trajectories exactly meet at time T and after time T we can switch off the control and follow the/one ideal solution $\bar{y}$.

In order to point out the strength of this notion, let us consider the particular case of $\bar{y}$ being a stationary solution (with f independent of t of course). It is known that there might exist an unstable stationary solution, so let us assume that $\bar{y}$ is such an unstable stationary solution. Of course in that case we have $\bar{y}_0 = \bar{y}$. If now we take y_0 different from $\bar{y}$, even close to $\bar{y}$ and if we do not exert any control, from the instability of $\bar{y}$ we see that $y(t)$ would diverge from $\bar{y}$. If we can prove (local) exact controllability to trajectories, it says that by choosing a suitable control v, not only the/one solution $y(v, t)$ will stay close to $\bar{y}$ but at time T it will meet exactly $\bar{y}$. This says that we have been able to perform a strong stabilization of the unstable solution $\bar{y}$.

At the moment, global exact controllability for Navier–Stokes equations is essentially an open problem and a very important one. Only the case of a control

acting on the whole boundary Γ of the domain can be solved using a result by Coron on approximate controllability in [2] (when the control acts on the whole boundary the type of boundary condition which is considered does not change anything) and a local controllability result like the one which will be given later on here.

The following sections will be devoted to the proof of a result of local exact controllability to trajectories for Navier–Stokes equations.

2.2 Result and Strategy

We have to introduce the classical functional spaces entering the study of Navier–Stokes equations. Let us define

$$H = \{y \in (L^2(\Omega))^N,\ \text{div}\, y = 0,\ y.\nu = 0 \text{ on } \Gamma\} \tag{2.9}$$

and

$$V = \{y \in (H_0^1(\Omega))^N,\ \text{div}\, y = 0\}. \tag{2.10}$$

We will denote by Q and Q_ω the cylinders

$$Q = \Omega \times (0,T),\ Q_\omega = \omega \times (0,T),$$

and by Σ the cylindrical boundary

$$\Sigma = \Gamma \times (0,T).$$

We will prove the following result of local exact controllability to trajectories

Theorem 2.1. *Let us assume that ω is a non empty open subset of Ω and that $T > 0$. We suppose that $\bar{y}$ satisfies $\bar{y} \in L^2(0,T;V) \cap L^\infty(Q)^N$. Then there exists $\eta > 0$ such that if $||y_0 - \bar{y}_0||_{L^4(\Omega)^N} \leq \eta$, there exists a control $v \in L^2(Q_\omega)^N$ and a solution y of (2.1) such that $y(T) = \bar{y}(T)$.*

The proof of this result will require several steps. First of all if we write

$$z = y - \bar{y},\ q = p - \bar{p},\ \ z_0 = y_0 - \bar{y}_0,$$

we have

$$\frac{\partial z}{\partial t} - \Delta z + \nabla \cdot (z \otimes \bar{y} + \bar{y} \otimes z) \tag{2.11}$$
$$+\nabla \cdot (z \otimes z) + \nabla q = v \mathbb{1}_\omega \text{ in } \Omega \times (0,T),$$
$$\text{div}\, z = 0 \text{ in } \Omega \times (0,T), \tag{2.12}$$

$$z = 0 \text{ on } \Gamma \times (0, T), \tag{2.13}$$

$$z(0) = z_0 \text{ in } \Omega. \tag{2.14}$$

We now want to find v and a corresponding solution z such that

$$z(T) = 0. \tag{2.15}$$

We will first of all consider the following linearized controllability problem. For a given tensor g (in a functional class which will be made precise later on) and a given initial data z_0, we consider the linearized Navier–Stokes problem

$$\frac{\partial z}{\partial t} - \Delta z + \nabla \cdot (z \otimes \bar{y} + \bar{y} \otimes z) + \tag{2.16}$$

$$\nabla q = \nabla \cdot g + v \mathbb{1}_\omega \text{ in } \Omega \times (0, T),$$

$$\operatorname{div} z = 0 \text{ in } \Omega \times (0, T), \tag{2.17}$$

$$z = 0 \text{ on } \Gamma \times (0, T), \tag{2.18}$$

$$z(0) = z_0 \text{ in } \Omega. \tag{2.19}$$

We want to find the control v such that at time T we have

$$z(T) = 0.$$

This linear null controllability problem will be studied in the next chapter.

Next, we will go back to the nonlinear problem. We will then need refined estimates and regularity results together with a variant of the inverse mapping theorem.

3 The Linearized Controllability Problem

For $g = (g_{ij}) \in L^2(Q)^{N^2}$ and $z_0 \in H$ we consider for every $v \in L^2(Q_\omega)^N$ the solution z of the following linear problem

$$\frac{\partial z}{\partial t} - \Delta z + \nabla \cdot (z \otimes \bar{y} + \bar{y} \otimes z) + \tag{3.1}$$

$$\nabla q = \nabla \cdot g + v \mathbb{1}_\omega \text{ in } \Omega \times (0, T),$$

$$\operatorname{div} z = 0 \text{ in } \Omega \times (0, T), \tag{3.2}$$

$$z = 0 \text{ on } \Gamma \times (0, T), \tag{3.3}$$

$$z(0) = z_0 \text{ in } \Omega. \tag{3.4}$$

Lemma 3.1. *This problem has a unique solution* $z = z(v) \in C([0,T];H) \cap L^2(0,T;V)$.

We leave the proof of this lemma to the reader. Using the assumption $\bar{y} \in L^\infty(Q)^N$, it is a classical extension of the existence result for Stokes problem which can be found for example in [14] or [11].

We now want to find a control $v \in L^2(Q_\omega)^N$ such that

$$z(T) = 0. \tag{3.5}$$

Of course this will require some conditions on g which will appear along the lines of the method.

We will give two different methods to obtain admissible controls which rely on the same key estimate.

3.1 Penalty Method

Let ϵ be a strictly positive number and for ϵ fixed let us consider the following optimal control problem:

$$\min_{v \in (L^2(Q_\omega))^N} J_\epsilon(v) \tag{3.6}$$

where

$$J_\epsilon(v) = \frac{1}{2\epsilon}|z(T)|_H^2 + \frac{1}{2}\int_{Q_\omega} |v|^2 dxdt. \tag{3.7}$$

From classical arguments (see for example [10]), this optimal control problem has a unique solution v_ϵ to which corresponds a state $z_\epsilon = z(v_\epsilon)$ and the optimality condition can be written as

$$DJ_\epsilon(v_\epsilon)[w] = 0 \ \forall w \in L^2(Q_\omega)^N. \tag{3.8}$$

If we write the derivative of the (affine) map $v \to z(v)$ at the point v_ϵ in the direction w as z_w it satisfies the system

$$\frac{\partial z_w}{\partial t} - \Delta z_w + \nabla \cdot (z_w \otimes \bar{y} + \bar{y} \otimes z_w) + \tag{3.9}$$

$$\nabla q_w = w \mathbb{1}_\omega \text{ in } \Omega \times (0,T),$$

$$\operatorname{div} z_w = 0 \text{ in } \Omega \times (0,T), \tag{3.10}$$

$$z_w = 0 \text{ on } \Gamma \times (0,T), \tag{3.11}$$

$$z_w(0) = 0 \text{ in } \Omega. \tag{3.12}$$

Using this, the optimality condition can be written as

$$(\frac{1}{\epsilon}z_\epsilon(T), z_w(T))_H + \int_{Q_\omega} v_\epsilon w dxdt = 0 \ \forall w \in L^2(Q_\omega)^N.$$

Let us introduce the adjoint state φ which, together with a pressure π satisfy the system (in components form)

$$-\frac{\partial \varphi_i}{\partial t} - \Delta\varphi_i - \sum_{j=1}^{N} \bar{y}_j D_{i,j}(\varphi) + \frac{\partial \pi}{\partial x_i} = 0 \text{ in } \Omega \times (0,T), \tag{3.13}$$

$$\operatorname{div} \varphi = 0 \text{ in } \Omega \times (0,T), \tag{3.14}$$

$$\varphi = 0 \text{ on } \Gamma \times (0,T), \tag{3.15}$$

$$\varphi(T) = \frac{1}{\epsilon} z_\epsilon(T) \text{ in } \Omega, \tag{3.16}$$

where

$$D_{i,j}(\varphi) = (\frac{\partial \varphi_i}{\partial x_j} + \frac{\partial \varphi_j}{\partial x_i}).$$

With the help of this adjoint state, the optimality condition can be written as

$$\int_{Q_\omega} (v_\epsilon + \varphi) w dxdt = 0 \ \forall w \in L^2(Q_\omega)^N,$$

or

$$v_\epsilon + \varphi = 0 \text{ in } Q_\omega. \tag{3.17}$$

Let us now try to pass to the limit when $\epsilon \to 0$. Multiplying the state equation for z_ϵ by φ, we obtain

$$(z_\epsilon(T), \varphi(T))_H - (z_0, \varphi(0))_H = -\sum_{i,j=1}^{N} \int_Q g_{ij} \frac{\partial \varphi_i}{\partial x_j} dxdt + \int_{Q_\omega} v_\epsilon \varphi dxdt,$$

so that

$$\frac{1}{\epsilon}|z_\epsilon(T)|_H^2 + \int_{Q_\omega} |v_\epsilon|^2 dxdt = (z_0, \varphi(0))_H - \sum_{i,j=1}^{N} \int_Q g_{ij} \frac{\partial \varphi_i}{\partial x_j} dxdt.$$

Let us assume for the moment that we know an estimate (Observability Inequality) like

$$|\varphi(0)|_H^2 + \int_Q \rho^2 |\nabla \varphi]^2 dxdt \le C \int_{Q_\omega} |\varphi|^2 dxdt = C \int_{Q_\omega} |v_\epsilon|^2 dxdt, \tag{3.18}$$

for some suitable weight function ρ. Then, if we assume that g satisfies

$$\int_Q \frac{1}{\rho^2}|g|^2 dxdt < +\infty,$$

we have

$$\frac{1}{\epsilon}|z_\epsilon(T)|_H^2 + \frac{1}{2}\int_{Q_\omega} |v_\epsilon|^2 dxdt \leq C(|z_0|_H^2 + \int_Q \frac{1}{\rho^2}|g|^2 dxdt).$$

Then, v_ϵ is bounded in $L^2(Q_\omega)^N$ and $\frac{1}{\sqrt{\epsilon}} z_\epsilon(T)$ is bounded in $L^2(\Omega)^N$ independently of ϵ.

After extraction of a subsequence we have

$$v_\epsilon \rightharpoonup v \text{ in } L^2(Q_\omega)^N \text{ weak},$$
$$z_\epsilon = z(v_\epsilon) \rightharpoonup z = z(v) \text{ in } C([0,T];H) \cap L^2(0,T;V) \text{ weak},$$
$$z_\epsilon(T) \rightharpoonup z(T) \text{ in } H \text{ weak}.$$

Therefore we must have

$$z(T) = 0,$$

and this proves that v is a control solving the null controllability problem.

In fact it is easy to show that $v_\epsilon \to v$ in $L^2(Q_\omega)^N$ strongly where v a solution to the null controllability problem which minimizes the $L^2(Q_\omega)^N$-norm among admissible controls.

We still have to prove the Observability Inequality.

3.2 Observability Inequality

The only way we know to obtain such an inequality is the use of a global Carleman estimate for the Stokes system. Let us consider the Stokes system

$$\frac{\partial z}{\partial t} - \Delta z + \nabla q = h \text{ in } \Omega \times (0,T), \tag{3.19}$$
$$\operatorname{div} z = 0 \text{ in } \Omega \times (0,T), \tag{3.20}$$
$$z = 0 \text{ on } \Gamma \times (0,T), \tag{3.21}$$
$$z(0) = z_0 \text{ in } \Omega, \tag{3.22}$$

with $z_0 \in V$ and $h \in L^2(0,T;(L^2(\Omega)^N)$.

Then, from classical regularity results (see for example [14]), we have

$$z \in C([0,T];V) \cap L^2(0,T;(H^2(\Omega))^N),$$
$$\frac{\partial z}{\partial t} \in (L^2(Q))^N,\ q \in L^2(0,T;H^1(\Omega)).$$

Let us define

$$w = \operatorname{curl} z.$$

Then

$$\frac{\partial w}{\partial t} - \Delta w = \operatorname{curl} h \text{ in } \Omega \times (0,T), \tag{3.23}$$
$$\Delta z(t) = \operatorname{curl} w(t) \text{ in } \Omega,\ \text{a.e. in } t \in (0,T), \tag{3.24}$$
$$z(t) = 0 \text{ on } \Gamma,\ \text{a.e. in } t \in (0,T). \tag{3.25}$$

Notice that on Γ, because $z_{/\Gamma} = 0$ we have

$$\nabla z = (\nabla z.\nu)\nu.$$

Therefore, $w_{/\Gamma}$ can be expressed in terms of $(\nabla z.\nu)$. The vector function w satisfies a system of non homogeneous heat equations. We can use the global Carleman estimate given in [8] or better the improvement proved in [9]. In order to write down this estimate we first have to define some suitable weights.

We know from [4] Lemma 1.1 that there exists a function $\psi \in C^2(\overline{\Omega})$ such that

$$\psi = 0 \text{ on } \Gamma,\ \psi(x) > 0\ \forall x \in \Omega,\ |\nabla\psi| \geq c_0 > 0 \text{ in } \Omega \setminus \overline{\omega}.$$

Let us define $l \in C^\infty([0,T])$ such that

$$l(t) = t \text{ on } [0,\frac{T}{4}],\quad l(t) = T - t \text{ on } [\frac{3T}{4},T],\quad l(t) \geq \frac{T}{4} \text{ on } [\frac{T}{4},\frac{3T}{4}].$$

We now define

$$\xi(x,t) = \frac{e^{\lambda(\psi(x)+m_1)}}{l^4(t)}, \tag{3.26}$$
$$\alpha(x,t) = \frac{e^{\lambda(\psi(x)+m_1)} - e^{\lambda(|\psi|_{L^\infty(\Omega)}+m_2)}}{l^4(t)}, \tag{3.27}$$

where $\lambda \geq 1$ and $m_1 \leq m_2$ are two constants chosen for the moment such that (it is easy to show that this is possible)

$$|\frac{\partial\alpha}{\partial t}| \leq C\xi^{\frac{5}{4}},\quad \text{and}\quad |\frac{\partial^2\alpha}{\partial t^2}| \leq C\xi^{\frac{3}{2}}.$$

We use a classical notation to define the space

$$H^{\frac{1}{4},\frac{1}{2}}(\Sigma) = H^{\frac{1}{4}}(0,T;L^2(\Gamma)) \cap L^2(0,T;H^{\frac{1}{2}}(\Gamma)).$$

The following result of global Carleman estimate for non homogeneous parabolic equations has been proved in [9]

Theorem 3.2. *There exist $\lambda_0 \geq 1$, $s_0 \geq 0$ and $C > 0$ such that for every $\lambda \geq \lambda_0$ and $s \geq s_0$, we have*

$$\frac{1}{s}\int_Q \frac{e^{2s\alpha}}{\xi}|\nabla w|^2 dxdt + s\lambda^2 \int_Q \xi e^{2s\alpha}|w|^2 dxdt \leq \tag{3.28}$$
$$C\left(s^{-\frac{1}{2}}||\xi^{-\frac{1}{4}}e^{s\alpha}w||^2_{H^{\frac{1}{4},\frac{1}{2}}(\Sigma)} + \int_Q e^{2s\alpha}|h|^2 dxdt\right) +$$
$$Cs\lambda^2 \int_{Q_\omega} \xi e^{2s\alpha}|w|^2 dxdt.$$

We recall that on Γ the trace of w can be expressed in terms of the normal derivative of z.

Now z is solution of an elliptic equation with right hand side w. We can use the estimate for elliptic equations given in [7] with the weight

$$\beta(x) = e^{\lambda(\psi(x)+m_1)}.$$

Theorem 3.3. *There exist $\tau_0 \geq 0$, $\lambda_0 \geq 1$ and $C > 0$ such that for every $\tau \geq \tau_0$ and $\lambda \geq \lambda_0$ the function $z(t)$ satisfies for almost every $t \in (0,T)$*

$$\int_\Omega e^{2\tau\beta}\left(|\nabla z(t)|^2 + \tau^2\lambda^2\beta^2|z(t)|^2\right)dx \leq \tag{3.29}$$
$$C\left(\tau\int_\Omega \beta e^{2\tau\beta}|w(t)|^2 dx + \tau^2\lambda^2\int_\omega \beta^2 e^{2\tau\beta}|z(t)|^2 dx\right).$$

Now let us take $\tau = \frac{s}{l^4(t)}$ and after multiplication by a suitable function of time we integrate in time on $(0,T)$ to obtain

$$\int_Q e^{2s\alpha}\left(\lambda^2|\nabla z|^2 + s^2\lambda^2\xi^2|z|^2\right)dxdt \leq \tag{3.30}$$
$$C\left(s\lambda^2\int_Q \xi e^{2s\alpha}|w|^2 dxdt + s^2\lambda^4\int_{Q_\omega}\xi^2 e^{2s\alpha}|z|^2 dxdt\right).$$

Combining the above parabolic and elliptic estimates we get

$$\int_Q e^{2s\alpha}\Big(\frac{|\nabla w|^2}{s\xi}+s\lambda^2\xi|w|^2+\lambda^2|\nabla z|^2+s^2\lambda^4\xi^2|z|^2\Big)dxdt \leq \tag{3.31}$$
$$C\Big(s^{-\frac{1}{2}}||\xi^{-\frac{1}{4}}e^{s\alpha}\frac{\partial z}{\partial \nu}||^2_{H^{\frac{1}{4},\frac{1}{2}}(\Sigma)}+\int_Q e^{2s\alpha}|h|^2dxdt\Big)$$
$$+C\int_{Q_\omega} e^{2s\alpha}\Big(s\lambda^2\xi|w|^2+s^2\lambda^4\xi^2|z|^2\Big)dxdt.$$

We now want to get rid of the boundary terms in the right hand side.

Let us define

$$\check{\alpha}(t)=\min_{x\in\Omega}\alpha(x,t)=\alpha_{/\Gamma}(t)=\frac{e^{\lambda m_1}-e^{\lambda(|\psi|_{L^\infty(\Omega)}+m_2)}}{l^4(t)}, \tag{3.32}$$

$$\check{\xi}(t)=\min_{x\in\Omega}\xi(x,t)=\xi_{/\Gamma}(t)=\frac{e^{\lambda m_1}}{l^4(t)}; \tag{3.33}$$

and let us write

$$(u,r)(x,t)=\check{\xi}(t)^{-\frac{1}{4}}(z(x,t)e^{s\check{\alpha}(t)},q(x,t)e^{s\check{\alpha}(t)}).$$

We have

$$\forall(x,t)\in\Omega\times(0,T),\ \check{\alpha}(t)\leq\alpha(x,t),\ \check{\xi}(t)\leq\xi(x,t),\ \check{\xi}(t)\geq C_0>0,$$

and

$$\frac{\partial u}{\partial t}-\Delta u+\nabla r=\frac{e^{s\check{\alpha}}}{\check{\xi}^{\frac{1}{4}}}h+s\frac{\check{\alpha}'}{\check{\xi}^{\frac{1}{4}}}e^{s\check{\alpha}}z-\frac{1}{4}\frac{\check{\xi}'}{\check{\xi}^{\frac{5}{4}}}e^{s\check{\alpha}}z, \tag{3.34}$$
$$\operatorname{div}u=0, \tag{3.35}$$
$$u_{/\Sigma}=0, \tag{3.36}$$
$$u(0)=0. \tag{3.37}$$

We now use the regularity result for Stokes equation (see [14]) to obtain

$$||u||^2_{H^{1,2}(Q)}\leq C(|e^{s\check{\alpha}}h|^2_{L^2(Q)}+s^2|\check{\xi}e^{s\check{\alpha}}z|^2_{L^2(Q)}),$$

where

$$H^{1,2}(Q)=H^1(0,T;L^2(\Omega))\cap L^2(0,T;H^2(\Omega)).$$

From [12] we know that

$$||\frac{\partial u}{\partial \nu}||^2_{H^{\frac{1}{4},\frac{1}{2}}(\Sigma)} \leq C\,||u||^2_{H^{1,2}(Q)}.$$

As

$$\frac{\partial u}{\partial \nu} = \check{\xi}^{-\frac{1}{4}} e^{s\check{\alpha}} \frac{\partial z}{\partial \nu}$$

we obtain

$$s^{-\frac{1}{2}}||\check{\xi}^{-\frac{1}{4}} e^{s\check{\alpha}} \frac{\partial z}{\partial \nu}||^2_{H^{\frac{1}{4},\frac{1}{2}}(\Sigma)} \leq C(s^{-\frac{1}{2}}|e^{s\check{\alpha}} h|^2_{L^2(Q)} + s^{\frac{3}{2}}|\check{\xi} e^{s\check{\alpha}} z|^2_{L^2(Q)}).$$

We use this estimate in (3.31). Taking s_0 large enough (we have $s \geq s_0$), we can absorb the term $s^{\frac{3}{2}}|\check{\xi} e^{s\check{\alpha}} z|^2_{L^2(Q)}$ from the left hand side and we obtain the first Carleman estimate for the Stokes system (recall that $w = \operatorname{curl} z$)

$$\int_Q e^{2s\alpha}\Big(\frac{|\nabla w|^2}{s\xi} + s\lambda^2\xi|w|^2 + \lambda^2|\nabla z|^2 + s^2\lambda^4\xi^2|z|^2\Big)dxdt \leq \tag{3.38}$$

$$C\Big(\int_Q e^{2s\alpha}|h|^2 dxdt + \int_{Q_\omega} e^{2s\alpha}\Big(s\lambda^2\xi|w|^2 + s^2\lambda^4\xi^2|z|^2\Big)dxdt.$$

Now we would like to remove the local term $s\lambda^2\int_{Q_\omega} e^{2s\alpha}\xi|w|^2 dxdt$ from the right hand side.

First of all we obtain the previous inequality with ω replaced by ω_0 with $\omega_0 \neq \emptyset$, and $\overline{\omega_0} \subset \omega$. Then we take a function $\theta \in C_0^\infty(\omega)$, $0 \leq \theta \leq 1$, $\theta = 1$ on ω_0. We have

$$s\lambda^2\int_{Q_{\omega_0}} e^{2s\alpha}\xi|w|^2 dxdt \leq s\lambda^2\int_{Q_\omega} \theta e^{2s\alpha}\xi|w|^2 dxdt =$$

$$s\lambda^2\int_{Q_\omega} \theta e^{2s\alpha}\xi w.\operatorname{curl} z dxdt = s\lambda^2\int_{Q_\omega} \operatorname{curl}(\theta\xi e^{2s\alpha} w).z dxdt =$$

$$s\lambda^2\Big(\int_{Q_\omega} \theta\xi e^{2s\alpha}\operatorname{curl} w.z dxdt + \int_{Q_\omega} e^{2s\alpha} w.z[2s(D\alpha)\theta\xi +$$

$$(D\theta)\xi + \lambda(D\psi)\xi]dxdt\Big),$$

where D stands for various first order differential operators. We know that $D\alpha = \lambda\xi D\psi$, $D\psi$ and $D\theta$ are bounded and $\operatorname{curl} w \sim \nabla w$. Therefore

$$s\lambda^2\int_{Q_{\omega_0}} e^{2s\alpha}\xi|w|^2dxdt \leq$$

$$C\Big(\int_{Q_\omega}\frac{e^{2s\alpha}}{s\xi}|\nabla w|^2dxdt\Big)^{\frac{1}{2}}\Big(\int_{Q_\omega}s^3\lambda^4\xi^3e^{2s\alpha}|z|^2dxdt\Big)^{\frac{1}{2}}+$$

$$C\Big(\int_{Q_\omega}s\lambda^2\xi e^{2s\alpha}|w|^2dxdt\Big)^{\frac{1}{2}}\Big(\int_{Q_\omega}s^2\lambda^4\xi^3e^{2s\alpha}|z|^2dxdt\Big)^{\frac{1}{2}}+$$

$$C\Big(\int_{Q_\omega}s\lambda^2\xi e^{2s\alpha}|w|^2dxdt\Big)^{\frac{1}{2}}\Big(\int_{Q_\omega}s\lambda^2\xi e^{2s\alpha}|z|^2dxdt\Big)^{\frac{1}{2}}+$$

$$C\Big(\int_{Q_\omega}s\lambda^2\xi e^{2s\alpha}|w|^2dxdt\Big)^{\frac{1}{2}}\Big(\int_{Q_\omega}s\lambda^4\xi e^{2s\alpha}|z|^2dxdt\Big)^{\frac{1}{2}}\leq$$

$$\frac{1}{2}\int_{Q_\omega}s\lambda^2\xi e^{2s\alpha}|w|^2dxdt+\frac{1}{2}\int_{Q_\omega}\frac{e^{2s\alpha}}{s\xi}|\nabla w|^2dxdt+$$

$$C\int_{Q_\omega}s^3\lambda^4\xi^3e^{2s\alpha}|z|^2dxdt.$$

We can absorb the first two terms from the left hand side of (3.38) to obtain the following new Carleman estimate for the Stokes system.

Theorem 3.4. *The exist $s_0 \geq 0$, $\lambda_0 \geq 1$ and $C > 0$ such that for $s \geq s_0$ and $\lambda \geq \lambda_0$ and for every solution z of the Stokes system (with $w = \operatorname{curl} z$) we have*

$$\int_Q e^{2s\alpha}\Big(\frac{|\nabla w|^2}{s\xi}+s\lambda^2\xi|w|^2+\lambda^2|\nabla z|^2+s^2\lambda^4\xi^2|z|^2\Big)dxdt\leq \tag{3.39}$$

$$C\Big(\int_Q e^{2s\alpha}|h|^2dxdt+s^3\lambda^4\int_{Q_\omega}e^{2s\alpha}\xi^3|z|^2dxdt\Big).$$

Let us now notice that our adjoint system (3.13) can be viewed as the previous Stokes equation with t replaced by $T-t$ and $h = \bar{y}D(z)$ with by hypothesis $\bar{y} \in (L^\infty(Q))^N$. Therefore, by choosing λ_0 large enough, we can absorb the term in ∇z from the left hand side in order to obtain the following inequality for the adjoint state φ.

$$\int_Q e^{2s\alpha}\Big(\frac{|\nabla\operatorname{curl}\varphi|^2}{s\xi}+s\lambda^2\xi|\operatorname{curl}\varphi|^2+\lambda^2|\nabla\varphi|^2+s^2\lambda^4\xi^2|\varphi|^2\Big)dxdt$$

$$\leq Cs^3\lambda^4\int_{Q_\omega}e^{2s\alpha}\xi^3|\varphi|^2dxdt. \tag{3.40}$$

From now on we fix $s \geq s_0$ and $\lambda \geq \lambda_0$.

Let us define

$$\tilde{\alpha}(t) = \alpha(t) \text{ if } t \in [\frac{T}{2}, T], \quad \tilde{\alpha}(t) = \alpha(\frac{T}{2}) \text{ if } t \in [0, \frac{T}{2}], \tag{3.41}$$

$$\tilde{\xi}(t) = \xi(t) \text{ if } t \in [\frac{T}{2}, T], \quad \tilde{\xi}(t) = \xi(\frac{T}{2}) \text{ if } t \in [0, \frac{T}{2}]. \tag{3.42}$$

Notice that the new functions $\tilde{\alpha}$ and $\tilde{\xi}$ are no longer degenerate in the neighborhood of $t = 0$.

Using standard energy estimates for the (backward) Stokes system, we can obtain the same inequality replacing α and ξ by $\tilde{\alpha}$ and $\tilde{\xi}$ so that

$$\int_Q e^{2s\tilde{\alpha}} \Big(\frac{|\nabla \text{curl}\, \varphi|^2}{s\tilde{\xi}} + s\lambda^2 \tilde{\xi} |\text{curl}\, \varphi|^2 + \lambda^2 |\nabla \varphi|^2 + s^2 \lambda^4 \tilde{\xi}^2 |\varphi|^2 \Big) dxdt$$
$$\leq C s^3 \lambda^4 \int_{Q_\omega} e^{2s\tilde{\alpha}} \tilde{\xi}^3 |\varphi|^2 dxdt. \tag{3.43}$$

Again with the help of standard energy estimates we also obtain what is called the Observability inequality

$$|\varphi(0)|_H^2 + \int_Q \tilde{\xi}^2 e^{2s\tilde{\alpha}} |\varphi|^2 dxdt + \int_Q e^{2s\tilde{\alpha}} |\nabla \varphi|^2 dxdt$$
$$\leq C \int_{Q_\omega} e^{2s\tilde{\alpha}} \tilde{\xi}^3 |\varphi|^2 dxdt. \tag{3.44}$$

Going back to our null controllability problem we see that we can solve this problem provided the initial data z_0 and the right hand side g satisfy

$$z_0 \in H, \quad \int_Q e^{-2s\tilde{\alpha}} |g|^2 dxdt < +\infty.$$

We then obtain the following controllability result for the linearized Navier–Stokes system.

Theorem 3.5. *If $z_0 \in H$ and g satisfies $\int_Q e^{-2s\tilde{\alpha}} |g|^2 dxdt < +\infty$, then there exists $v \in (L^2(Q_\omega))^N$ such that the solution z of (3.1) satisfies*

$$z(T) = 0.$$

3.3 Exponentially Decreasing Controls and Solutions

We will show here that we can find a control v which is exponentially decreasing when $t \to T$ and such that not only $z(T) = 0$ but z is also exponentially decreasing when $t \to T$.

Let us define

$$X_0 = \{(z,q) \in C^\infty(\overline{Q})^{N+1},\ \mathrm{div}\, z = 0 \text{ in } Q,\ z = 0 \text{ on } \Sigma,$$
$$\int_\omega q(t)dx = 0\ a.e.\ \text{in } (0,T)\}.$$

If we set

$$L^*z = -\frac{\partial z}{\partial t} - \Delta z - \bar{y}.D(z),$$

we can define on X_0 the bilinear form

$$a((z,q),(\tilde{z},\tilde{q})) = \int_Q e^{2s\tilde{\alpha}}(L^*z + \nabla q)(L^*\tilde{z} + \nabla\tilde{q})dxdt + \int_{Q_\omega} e^{2s\tilde{\alpha}}\tilde{\xi}^3 z.\tilde{z}dxdt.$$

As $\bar{y} \in (L^\infty(Q)^N$, because of the Carleman estimate (3.39), we see that this is a scalar product on X_0. Let us define X to be the completion of X_0 with respect to this scalar product. Then of course X is a Hilbert space for this scalar product and we have

$$\forall (z,q) \in X,\ |z(0)|_H^2 + \int_Q e^{2s\tilde{\alpha}}(|\nabla z|^2 + \tilde{\xi}^2|z|^2)dxdt \le Ca((z,q),(z,q)).$$

Let us now consider the linear form $l : X \to \mathbb{R}$ defined by

$$\forall (\tilde{z},\tilde{q}) \in X,\ < l,(\tilde{z},\tilde{q} >= (z_0,\tilde{z}(0))_H - \sum_{i,j=1}^N \int_Q g_{ij}\frac{\partial \tilde{z}_i}{\partial x_j}dxdt.$$

Then, if $z_0 \in H$ and $\int_Q e^{-2s\tilde{\alpha}}|g|^2dxdt < +\infty$, we can see that l is a continuous linear form on X. From Lax–Milgram Theorem, there exists a unique solution $(z,q) \in X$ of the problem

$$a((z,q),(\tilde{z},\tilde{q})) =< l,(\tilde{z},\tilde{q}) >,\ \forall (\tilde{z},\tilde{q}) \in X.$$

Let us now define

$$y = e^{2s\tilde{\alpha}}(L^*z + \nabla q) \text{ in } Q,$$
$$v = -e^{2s\tilde{\alpha}}\tilde{\xi}^3 z_{/\omega} \text{ in } Q_\omega.$$

Then $y \in (L^2(Q))^N$, $v \in (L^2(Q_\omega))^N$ and we have

$$\int_Q y.(L^*\tilde{z} + \nabla\tilde{q})dxdt =$$

$$(z_0, \tilde{z}(0))_H - \sum_{i,j=1}^{N} \int_Q g_{ij} \frac{\partial \tilde{z}_i}{\partial x_j} dxdt + \int_{Q_\omega} v\tilde{z}dxdt, \ \forall (\tilde{z}, \tilde{q}) \in X.$$

Therefore, y, together with a pressure p is the (unique!) solution defined by transposition of the problem

$$\frac{\partial y}{\partial t} - \Delta y + \nabla \cdot (y \otimes \bar{y} + \bar{y} \otimes y) + \tag{3.45}$$

$$\nabla p = \nabla \cdot g + v\mathbb{1}_\omega \text{ in } \Omega \times (0, T),$$

$$\text{div } y = 0 \text{ in } \Omega \times (0, T), \tag{3.46}$$

$$y = 0 \text{ on } \Gamma \times (0, T), \tag{3.47}$$

$$y(0) = z_0 \text{ in } \Omega. \tag{3.48}$$

But as $g \in (L^2(Q))^{N^2}$ and $v \in (L^2(Q_\omega)^N$ this problem has a solution (which has to be the same one by uniqueness) $y \in C([0, T]; H) \cap L^2(0, T; V)$.
On the other hand we have

$$\int_Q e^{-2s\tilde{\alpha}} |y|^2 dxdt =$$

$$\int_Q e^{-2s\tilde{\alpha}} e^{4s\tilde{\alpha}} |L^*z + \nabla q|^2 dxdt = \int_Q e^{2s\tilde{\alpha}} |L^*z + \nabla q|^2 dxdt < +\infty$$

and

$$\int_{Q_\omega} \frac{e^{-2s\tilde{\alpha}}}{\tilde{\xi}^3} |v|^2 dxdt = \int_{Q_\omega} e^{2s\tilde{\alpha}} \tilde{\xi}^3 |z|^2 dxdt < +\infty.$$

Of course this says that $y(T) = 0$ and that y and v are exponentially decreasing when $t \to T$. We can summarize this result as follows.

Theorem 3.6. *If $z_0 \in H$ and g satisfies $\int_Q e^{-2s\tilde{\alpha}} |g|^2 dxdt < +\infty$, then there exists a control v and a solution y of (3.1) such that $y(T) = 0$ and*

$$\int_Q e^{-2s\tilde{\alpha}} |y|^2 dxdt < +\infty, \text{ and } \int_{Q_\omega} \frac{e^{-2s\tilde{\alpha}}}{\tilde{\xi}^3} |v|^2 dxdt < +\infty. \tag{3.49}$$

4 The Nonlinear Problem

4.1 Choice of Special Weights

From now on, in order to simplify, we will omit the notation˜for the weights. In the weights ξ and α we still have some choice for the constants m_1 and m_2. We will make a special choice for these constants. First of all we define for $t \in [\frac{T}{2}, T]$ (the case $t \in [0, \frac{T}{2}]$ is straightforward)

$$\check{\alpha}(t) = \min_{x\in\overline{\Omega}} \alpha(x,t) = \frac{e^{\lambda m_1} - e^{\lambda(|\psi|_{L^\infty(\Omega)}+m_2)}}{l^4(t)} < 0, \tag{4.1}$$

$$\check{\xi}(t) = \min_{x\in\overline{\Omega}} \xi(x,t) = \frac{e^{\lambda m_1}}{l^4(t)}, \tag{4.2}$$

$$\hat{\alpha}(t) = \max_{x\in\overline{\Omega}} \alpha(x,t) = \frac{e^{\lambda(|\psi|_{L^\infty(\Omega)}+m_1)} - e^{\lambda(|\psi|_{L^\infty(\Omega)}+m_2)}}{l^4(t)} < 0, \tag{4.3}$$

$$\hat{\xi}(t) = \max_{x\in\overline{\Omega}} \xi(x,t) = \frac{e^{\lambda(|\psi|_{L^\infty(\Omega)}+m_1)}}{l^4(t)}. \tag{4.4}$$

We then have.

Lemma 4.1. *We can choose the constants m_1 and m_2 with $m_1 \le m_2$ such that for λ_0 large enough, we have for $\lambda \ge \lambda_0$*

$$|\frac{\partial\alpha}{\partial t}| \le C\xi^{\frac{5}{4}},\ |\frac{\partial^2\alpha}{\partial t^2}| \le C\xi^{\frac{3}{2}}$$

and

$$\frac{3}{2}\hat{\alpha} \le \check{\alpha}\ or\ -\check{\alpha} \le -\frac{3}{2}\hat{\alpha}.$$

Proof. Let us take

$$m_1 = (m_0+4)|\psi|_{L^\infty(\Omega)}\ ,\ m_2 = m_3|\psi|_{L^\infty(\Omega)}.$$

It is easy to see that all conditions are fulfilled if

$$m_0 + 4 < m_3 < \frac{5}{4}m_0 + 4$$

for λ_0 large enough. Now such a choice of m_0 and m_3 is obviously possible.

4.2 Functional Class and Solution of the Nonlinear Problem

We recall the setting of the nonlinear problem.

$$\frac{\partial z}{\partial t} - \Delta z + \nabla \cdot (z \otimes \bar{y} + \bar{y} \otimes z) + \tag{4.5}$$
$$\nabla \cdot (z \otimes z) + \nabla q = v \mathbb{1}_\omega \text{ in } \Omega \times (0,T),$$
$$\operatorname{div} z = 0 \text{ in } \Omega \times (0,T), \tag{4.6}$$
$$z = 0 \text{ on } \Gamma \times (0,T), \tag{4.7}$$
$$z(0) = z_0 \text{ in } \Omega. \tag{4.8}$$

We now want to find v and a corresponding solution z such that

$$z(T) = 0. \tag{4.9}$$

We have to define a correct functional class in order to apply some inverse mapping argument. Let us define

$$Lz = \frac{\partial z}{\partial t} - \Delta z + \nabla \cdot (z \otimes \bar{y} + \bar{y} \otimes z)$$

and

$$E = \{(z,v),\ e^{-s\alpha} z \in (L^2(Q))^N,\ \frac{e^{-s\alpha}}{\xi^{\frac{3}{2}}} v \in (L^2(Q_\omega))^N, \tag{4.10}$$
$$e^{-\frac{3s}{4}\hat{\alpha}} z \in L^4(0,T;(L^{12}(\Omega))^N) \cap L^2(0,T;V) \cap L^\infty(0,T;H),$$
$$\exists q, \exists k,\ e^{-s\alpha} k \in L^2(0,T;L^6(\Omega)^{N^2}),\ Lz + \nabla q - v\mathbb{1}_\omega = \nabla . k,$$
$$z(0) \in H \cap L^4(\Omega)^N\},$$

equipped with the norm

$$||(z,v)||_E^2 = |e^{-s\alpha} z|^2_{L^2(Q)^N} + |\frac{e^{-s\alpha}}{\xi^{\frac{3}{2}}} v|^2_{L^2(Q_\omega)^N} +$$
$$|e^{-\frac{3s}{4}\hat{\alpha}} z|^2_{L^4(0,T;L^{12}(\Omega)^N)} + ||e^{-\frac{3s}{4}\hat{\alpha}} z||^2_{L^2(0,T;V)\cap L^\infty(0,T;H)} +$$
$$|e^{-s\alpha} k|^2_{L^2(0,T;L^6(\Omega)^{N^2})} + ||z(0)||^2_{L^4(\Omega)^N}.$$

This class E is non empty and is a Banach space.

On the other hand we define

$$G = \{(\nabla . k, z_0),\ e^{-s\alpha} k \in L^2(0,T;L^6(\Omega)^{N^2}),\ z_0 \in H \cap L^4(\Omega)^N\} \tag{4.11}$$

equipped with the norm

$$||(\nabla.k, z_0)||_G^2 = |e^{-s\alpha}k|^2_{L^2(0,T;L^6(\Omega)^{N^2}} + ||z_0||^2_{L^4(\Omega)^N}.$$

The space G is a Banach space. We now define the operator A as follows: for $(z, v) \in E$

$$A(z, v) = (Lz + \nabla.(z \otimes z) + \nabla q - v\mathbb{1}_\omega, z(0)). \tag{4.12}$$

From the definition of E it is clear that $z(0) \in H \cap L^4(\Omega)^N$.

On the other hand we have

$$e^{-\frac{3s}{2}\hat{\alpha}}(z \otimes z) \in L^2(0, T; L^6(\Omega)^{N^2})$$

and

$$-\alpha \leq -\check{\alpha} \leq -\frac{3}{2}\hat{\alpha}$$

so that

$$e^{-s\alpha}(z \otimes z) \in L^2(0, T; L^6(\Omega)^{N^2}).$$

Therefore, A maps E into G and it is obviously continuous. As the first component of $A(z, v)$ is linear plus quadratic and this quadratic part comes from a continuous bilinear map from $E \times E$ to G, it is clear that A is a C^1 map from E to G. Let us compute the derivative of A at the point $(0, 0) \in E$. For $(y, w) \in E$

$$A'(0, 0)[y, w] = (Ly + \nabla p - w\mathbb{1}_\omega, y(0)).$$

Given $(\nabla.h, y_0) \in G$, we know from Theorem 3.6 that there exists (y, w) such that

$$Ly + \nabla p = w\mathbb{1}_\omega + \nabla.h, \text{ in } \Omega \times (0, T) \tag{4.13}$$

$$\text{div } y = 0, \text{ in } \Omega \times (0, T) \tag{4.14}$$

$$y = 0 \text{ on } \Gamma \times (0, T), \tag{4.15}$$

$$y(0) = y_0 \text{ in } \Omega. \tag{4.16}$$

with $y(T) = 0$ and

$$y \in L^2(0, T; V) \cap L^\infty(0, T; H),$$

$$e^{-s\alpha}y \in L^2(Q)^N, \quad \frac{e^{-s\alpha}}{\xi^{\frac{3}{2}}}w \in L^2(Q_\omega)^N.$$

Let us show that $(y, w) \in E$. It remains to prove that

$$e^{-\frac{3s}{4}\hat{\alpha}} y \in L^2(0,T;V) \cap L^\infty(0,T;H) \cap L^4(0,T;L^{12}(\Omega)^N).$$

Let us define

$$\tilde{y} = e^{-\frac{3s}{4}\hat{\alpha}} y, \quad \tilde{p} = e^{-\frac{3s}{4}\hat{\alpha}} p, \quad \tilde{h} = e^{-\frac{3s}{4}\hat{\alpha}} h, \quad \tilde{w} = e^{-\frac{3s}{4}\hat{\alpha}} w.$$

It is easy to show that

$$\tilde{h} \in L^2(0,T;L^6(\Omega)^{N^2}) \text{ , and } \tilde{w} \in L^2(0,T;L^2(Q_\omega)^N).$$

On the other hand we have

$$\begin{aligned}
&\frac{\partial \tilde{y}}{\partial t} - \Delta \tilde{y} + \nabla.(\bar{y} \otimes \tilde{y} + \tilde{y} \otimes \bar{y}) + \nabla \tilde{p} = \\
&\nabla.\tilde{h} + \tilde{w}\mathbb{1}_\omega - \frac{3s}{4}\frac{\partial \hat{\alpha}}{\partial t}\tilde{y}, \text{ in } \Omega \times (0,T) \\
&\text{div } \tilde{y} = 0, \text{ in } \Omega \times (0,T) \\
&\tilde{y} = 0 \text{ on } \Gamma \times (0,T), \\
&\tilde{y}(0) = e^{-\frac{3s}{4}\hat{\alpha}(0)} y_0 \text{ in } \Omega.
\end{aligned}$$

Notice that

$$\frac{\partial \hat{\alpha}}{\partial t}\tilde{y} = \frac{\partial \hat{\alpha}}{\partial t} e^{-\frac{3s}{4}\hat{\alpha}} y = (\frac{\partial \hat{\alpha}}{\partial t} e^{\frac{s}{4}\hat{\alpha}})(e^{-s\hat{\alpha}} y) \in L^2(0,T;L^2(\Omega)^N).$$

Therefore, from the existence result and as $N \leq 3$, we have $\tilde{y} \in L^2(0,T;V) \subset L^2(0,T;L^6(\Omega)^N)$. As $\bar{y} \in L^\infty(Q)$ we have $\bar{y} \otimes \tilde{y} \in L^2(0,T;L^6(\Omega)^{N^2})$ and $\tilde{y} \otimes \bar{y} \in L^2(0,T;L^6(\Omega)^{N^2})$ and these terms can be incorporated in $\tilde{h}$ so that without loss of generality we can write (without taking a different notation)

$$\frac{\partial \tilde{y}}{\partial t} - \Delta \tilde{y} + \nabla \tilde{p} = \nabla.\tilde{h} + \tilde{k}, \text{ in } \Omega \times (0,T) \tag{4.17}$$

$$\text{div } \tilde{y} = 0, \text{ in } \Omega \times (0,T) \tag{4.18}$$

$$\tilde{y} = 0 \text{ on } \Gamma \times (0,T), \tag{4.19}$$

$$\tilde{y}(0) = e^{-\frac{3s}{4}\hat{\alpha}(0)} y_0 \text{ in } \Omega, \tag{4.20}$$

with

$$\tilde{k} = \tilde{w}\mathbb{1}_\omega - \frac{3s}{4}\frac{\partial \hat{\alpha}}{\partial t}\tilde{y} \in L^2(Q)^N$$

and

$$\tilde{h} \in L^2(0,T;L^6(\Omega)^{N^2}).$$

Lemma 4.2. *Let us assume that* $y_0 \in H \cap L^4(\Omega)^N$ *and that* $\tilde{h} \in L^2(0,T;L^6(\Omega)^{N^2})$, $\tilde{k} \in L^2(Q)^N$. *Then* $\tilde{y} \in L^4(0,T;L^{12}(\Omega)^N)$.

This lemma will be a consequence of the following one by the transposition method.

Lemma 4.3. *Let* $k \in L^{\frac{4}{3}}(0,T;L^{\frac{12}{11}}(\Omega)^N)$. *Then there exists a unique solution* (z,q) *to the Stokes system*

$$\begin{aligned}
&-\frac{\partial z}{\partial t} - \Delta z + \nabla q = k, \ \textit{in } \Omega \times (0,T)\\
&\operatorname{div} z = 0, \ \textit{in } \Omega \times (0,T)\\
&z = 0 \ \textit{on } \Gamma \times (0,T),\\
&z(T) = 0 \ \textit{in } \Omega,
\end{aligned}$$

with

$$z \in C([0,T];L^{\frac{4}{3}}(\Omega)^N) \cap L^2(0,T;W^{1,\frac{6}{5}}(\Omega)^N).$$

Assume Lemma 4.3 is proved. We then have

$$\int_Q \tilde{y}k dxdt = \int_\Omega e^{-\frac{3s}{4}\hat{\alpha}(0)} y_0 z(0) dx - \sum_{i,j}^N \int_Q \tilde{h}_{ij} \frac{\partial z_i}{\partial x_j} dxdt + \int_Q \tilde{k} z dxdt.$$

We know that

$$z(0) \in L^{\frac{4}{3}}(\Omega)^N, \ \frac{\partial z_i}{\partial x_j} \in L^2(0,T;L^{\frac{6}{5}}(\Omega)) \text{ and } W^{1,\frac{6}{5}}(\Omega) \subset L^2(\Omega).$$

As

$$y_0 \in L^4(\Omega)^N, \ \tilde{h}_{ij} \in L^2(0,T;L^6(\Omega)) \text{ and } \tilde{k} \in L^2(Q)^N,$$

the right hand side is a linear continuous form on k and therefore defines a unique element $\tilde{y}$ in $\left(L^{\frac{4}{3}}(0,T;L^{\frac{12}{11}}(\Omega)^N)\right)' = L^4(0,T;L^{12}(\Omega)^N)$. This gives the proof of Lemma 4.2.

The proof of Lemma 4.3 follows the proof given in [3].

Proof of Lemma 4.3. From a Giga–Sohr regularity result on the Stokes problem (see [5]) we have

$$z \in L^{\frac{4}{3}}(0,T;W^{2,\frac{12}{11}}(\Omega)^N), \text{ and } \frac{\partial z}{\partial t} \in L^{\frac{4}{3}}(0,T;L^{\frac{12}{11}}(\Omega)^N).$$

We have

$$W^{2,\frac{12}{11}}(\Omega) \subset W^{1,\frac{12}{7}}(\Omega) \subset L^4(\Omega).$$

Then

$$z \in L^{\frac{4}{3}}(0,T;L^4(\Omega)^N)\ ,\ \frac{\partial z}{\partial t} \in L^{\frac{4}{3}}(0,T;L^{\frac{12}{11}}(\Omega)^N).$$

By interpolation results (see [13]) we obtain

$$z \in C([0,T];\left(L^4(\Omega),L^{\frac{12}{11}}(\Omega)\right)^N_{\frac{3}{4},\frac{4}{5}}),$$

and $\left(L^4(\Omega),L^{\frac{12}{11}}(\Omega)\right)_{\frac{3}{4},\frac{4}{5}}$ is the Lorentz space $L^{\frac{4}{3},\frac{4}{3}}(\Omega) = L^{\frac{4}{3}}(\Omega)$. Then

$$z \in C([0,T];L^{\frac{4}{3}}(\Omega)^N).$$

On the other hand we have

$$z \in L^{\frac{4}{3}}(0,T;W^{2,\frac{12}{11}}(\Omega)^N \cap W^{1,\frac{12}{7}}(\Omega)^N)\ ,\ \text{and}\ ,\ z \in L^\infty(0,T;L^{\frac{12}{11}}(\Omega)^N).$$

By interpolation we have

$$z \in L^2(0,T;\left(W^{2,\frac{12}{11}}(\Omega),L^{\frac{12}{11}}(\Omega)\right)^N_{\frac{1}{3},\frac{12}{11}}).$$

But

$$\left(W^{2,\frac{12}{11}}(\Omega),L^{\frac{12}{11}}(\Omega)\right)_{\frac{1}{3},\frac{12}{11}} = W^{\frac{4}{3},\frac{12}{11}}(\Omega) \subset W^{1,\frac{6}{5}}(\Omega).$$

As $z = 0$ on the boundary $\Gamma \times (0,T)$, we finally obtain

$$z \in L^2(0,T;W_0^{1,\frac{6}{5}}(\Omega)^N)$$

and this finishes the proof of Lemma 4.3 and then of Lemma 4.2.

This also proves that the solution (y,w) to the controllability problem (4.13) is element of E and therefore that $A'(0,0)$ is a surjective linear map from E to G. We can now apply the following epimorphism theorem (see [1]).

Theorem 4.4. *Let E and G be two Banach spaces and let $\mathbf{A} : E \mapsto G$ satisfy $\mathbf{A} \in C^1(E; G)$. Assume that $e_0 \in E$, $\mathbf{A}(e_0) = h_0$ and $\mathbf{A}'(e_0) : E \mapsto G$ is surjective. Then, there exists $\delta > 0$ such that, for every $h \in G$ satisfying $\|h - h_0\|_G < \delta$, there exists a solution of the equation*

$$\mathbf{A}(e) = h, \quad e \in E.$$

Taking here $e_0 = (0, 0)$ and $h_0 = (0, 0)$ this theorem gives exactly the conclusions of Theorem 2.1 if we take $h \in G$ of the form $(0, z_0)$.

Remark. We could have taken in the equation for $\bar{y}$ an external force $\bar{f}$ different from f. In fact, following the argument it is easy to show that we can take

$$f = \bar{f} + \tilde{f}$$

with $\frac{e^{-s\tilde{\alpha}}}{\tilde{\xi}^{\frac{3}{2}}} \tilde{f}$ finite and small enough.

Acknowledgements This work has been supported by the ESF within the programme OPTPDE.

References

[1] V.M. Alekseev, V.M. Tikhomirov, S.V. Fomin, in *Optimal Control*, translated from the Russian by V.M. Volosov. Contemporary Soviet Mathematics (Consultants Bureau, New York, 1987)
[2] J.-M. Coron, On the controllability of the 2-D incompressible Navier–Stokes equations with the Navier slip boundary conditions. ESAIM Control Optim. Calc. Var. **1**, 35–75 (1996)
[3] E. Fernández-Cara, S. Guerrero, O.Yu. Imanuvilov, J.-P. Puel, Local exact controllability to the trajectories of the Navier-Stokes equations. J. Math. Pures Appl. **83**, 1501–1542
[4] A. Fursikov, O.Yu. Imanuvilov, *Controllability of Evolution Equations*. Lecture Notes #34. (Seoul National University, Seoul, 1996)
[5] Y. Giga, H. Sohr, Abstract L^p estimates for the Cauchy problem with applications to the Navier-Stokes equations in exterior domains. J. Funct. Anal. **102**, 72–94 (1991)
[6] O.Yu. Imanuvilov, Remarks on exact controllability for the Navier-Stokes equations. ESAIM Control Optim. Calc. Var. **6**, 39–72 (2001)
[7] O.Yu. Imanuvilov, J.-P. Puel, Global Carleman estimates for weak elliptic non homogeneous Dirichlet problem. Int. Math. Res. Not. **16**, 883–913 (2003)
[8] O. Imanuvilov, J.-P. Puel, M. Yamamoto, Carleman estimates for parabolic equations with nonhomogeneous boundary conditions. Chin. Ann. Math. **30**, 333–378 (2009)
[9] O. Imanuvilov, J.-P. Puel, M. Yamamoto, Carleman estimates for second order non homogeneous parabolic equations. (To appear)
[10] J.-L. Lions, Contrôle Optimal De Systèmes Gouvernés Par Des Équations Aux Dérivées Partielles. (Dunod, Paris, 1968)
[11] J.-L. Lions, Quelques Méthodes De Résolution Des Problèmes Aux Limites Non Linéaires. (Dunod, Paris, 1969)
[12] J.-L. Lions, E. Magenes, Problèmes Aux Limites Non Homogènes Et Applications, vol. 2. (Dunod, Paris, 1968)

[13] L. Tartar, *An Introduction to Sobolev Spaces and Interpolation Spaces* (Springer, Berlin/Heidelberg/New York, 2007)
[14] R. Temam, *Navier-Stokes Equations: Theory and Numerical Analysis* (North Holland, Amsterdam/New York/Oxford, 1977)

Editorial Policy

1. Volumes in the following three categories will be published in LNCSE:

i) Research monographs
ii) Tutorials
iii) Conference proceedings

Those considering a book which might be suitable for the series are strongly advised to contact the publisher or the series editors at an early stage.

2. Categories i) and ii). Tutorials are lecture notes typically arising via summer schools or similar events, which are used to teach graduate students. These categories will be emphasized by Lecture Notes in Computational Science and Engineering. **Submissions by interdisciplinary teams of authors are encouraged.** The goal is to report new developments – quickly, informally, and in a way that will make them accessible to non-specialists. In the evaluation of submissions timeliness of the work is an important criterion. Texts should be well-rounded, well-written and reasonably self-contained. In most cases the work will contain results of others as well as those of the author(s). In each case the author(s) should provide sufficient motivation, examples, and applications. In this respect, Ph.D. theses will usually be deemed unsuitable for the Lecture Notes series. Proposals for volumes in these categories should be submitted either to one of the series editors or to Springer-Verlag, Heidelberg, and will be refereed. A provisional judgement on the acceptability of a project can be based on partial information about the work: a detailed outline describing the contents of each chapter, the estimated length, a bibliography, and one or two sample chapters – or a first draft. A final decision whether to accept will rest on an evaluation of the completed work which should include

– at least 100 pages of text;
– a table of contents;
– an informative introduction perhaps with some historical remarks which should be accessible to readers unfamiliar with the topic treated;
– a subject index.

3. Category iii). Conference proceedings will be considered for publication provided that they are both of exceptional interest and devoted to a single topic. One (or more) expert participants will act as the scientific editor(s) of the volume. They select the papers which are suitable for inclusion and have them individually refereed as for a journal. Papers not closely related to the central topic are to be excluded. Organizers should contact the Editor for CSE at Springer at the planning stage, see *Addresses* below.

In exceptional cases some other multi-author-volumes may be considered in this category.

4. Only works in English will be considered. For evaluation purposes, manuscripts may be submitted in print or electronic form, in the latter case, preferably as pdf- or zipped ps-files. Authors are requested to use the LaTeX style files available from Springer at http://www.springer.com/gp/authors-editors/book-authors-editors/book-manuscript-guidelines/manuscript-preparation/5636 (Click on LaTeX → monographs or contributed books).

For categories ii) and iii) we strongly recommend that all contributions in a volume be written in the same LaTeX version, preferably LaTeX2e. Electronic material can be included if appropriate. Please contact the publisher.

Careful preparation of the manuscripts will help keep production time short besides ensuring satisfactory appearance of the finished book in print and online.

5. The following terms and conditions hold. Categories i), ii) and iii):

Authors receive 50 free copies of their book. No royalty is paid.
Volume editors receive a total of 50 free copies of their volume to be shared with authors, but no royalties.

Authors and volume editors are entitled to a discount of 33.3 % on the price of Springer books purchased for their personal use, if ordering directly from Springer.

6. Springer secures the copyright for each volume.

Addresses:

Timothy J. Barth
NASA Ames Research Center
NAS Division
Moffett Field, CA 94035, USA
barth@nas.nasa.gov

Michael Griebel
Institut für Numerische Simulation
der Universität Bonn
Wegelerstr. 6
53115 Bonn, Germany
griebel@ins.uni-bonn.de

David E. Keyes
Mathematical and Computer Sciences
and Engineering
King Abdullah University of Science
and Technology
P.O. Box 55455
Jeddah 21534, Saudi Arabia
david.keyes@kaust.edu.sa

and

Department of Applied Physics
and Applied Mathematics
Columbia University
500 W. 120 th Street
New York, NY 10027, USA
kd2112@columbia.edu

Risto M. Nieminen
Department of Applied Physics
Aalto University School of Science
and Technology
00076 Aalto, Finland
risto.nieminen@aalto.fi

Dirk Roose
Department of Computer Science
Katholieke Universiteit Leuven
Celestijnenlaan 200A
3001 Leuven-Heverlee, Belgium
dirk.roose@cs.kuleuven.be

Tamar Schlick
Department of Chemistry
and Courant Institute
of Mathematical Sciences
New York University
251 Mercer Street
New York, NY 10012, USA
schlick@nyu.edu

Editor for Computational Science
and Engineering at Springer:
Martin Peters
Springer-Verlag
Mathematics Editorial IV
Tiergartenstrasse 17
69121 Heidelberg, Germany
martin.peters@springer.com

Lecture Notes in Computational Science and Engineering

1. D. Funaro, *Spectral Elements for Transport-Dominated Equations.*
2. H.P. Langtangen, *Computational Partial Differential Equations.* Numerical Methods and Diffpack Programming.
3. W. Hackbusch, G. Wittum (eds.), *Multigrid Methods V.*
4. P. Deuflhard, J. Hermans, B. Leimkuhler, A.E. Mark, S. Reich, R.D. Skeel (eds.), *Computational Molecular Dynamics: Challenges, Methods, Ideas.*
5. D. Kröner, M. Ohlberger, C. Rohde (eds.), *An Introduction to Recent Developments in Theory and Numerics for Conservation Laws.*
6. S. Turek, *Efficient Solvers for Incompressible Flow Problems.* An Algorithmic and Computational Approach.
7. R. von Schwerin, ***M**ulti **B**ody **S**ystem **SIM**ulation.* Numerical Methods, Algorithms, and Software.
8. H.-J. Bungartz, F. Durst, C. Zenger (eds.), *High Performance Scientific and Engineering Computing.*
9. T.J. Barth, H. Deconinck (eds.), *High-Order Methods for Computational Physics.*
10. H.P. Langtangen, A.M. Bruaset, E. Quak (eds.), *Advances in Software Tools for Scientific Computing.*
11. B. Cockburn, G.E. Karniadakis, C.-W. Shu (eds.), *Discontinuous Galerkin Methods.* Theory, Computation and Applications.
12. U. van Rienen, *Numerical Methods in Computational Electrodynamics.* Linear Systems in Practical Applications.
13. B. Engquist, L. Johnsson, M. Hammill, F. Short (eds.), *Simulation and Visualization on the Grid.*
14. E. Dick, K. Riemslagh, J. Vierendeels (eds.), *Multigrid Methods VI.*
15. A. Frommer, T. Lippert, B. Medeke, K. Schilling (eds.), *Numerical Challenges in Lattice Quantum Chromodynamics.*
16. J. Lang, *Adaptive Multilevel Solution of Nonlinear Parabolic PDE Systems.* Theory, Algorithm, and Applications.
17. B.I. Wohlmuth, *Discretization Methods and Iterative Solvers Based on Domain Decomposition.*
18. U. van Rienen, M. Günther, D. Hecht (eds.), *Scientific Computing in Electrical Engineering.*
19. I. Babuška, P.G. Ciarlet, T. Miyoshi (eds.), *Mathematical Modeling and Numerical Simulation in Continuum Mechanics.*
20. T.J. Barth, T. Chan, R. Haimes (eds.), *Multiscale and Multiresolution Methods.* Theory and Applications.
21. M. Breuer, F. Durst, C. Zenger (eds.), *High Performance Scientific and Engineering Computing.*
22. K. Urban, *Wavelets in Numerical Simulation.* Problem Adapted Construction and Applications.
23. L.F. Pavarino, A. Toselli (eds.), *Recent Developments in Domain Decomposition Methods.*

24. T. Schlick, H.H. Gan (eds.), *Computational Methods for Macromolecules: Challenges and Applications.*

25. T.J. Barth, H. Deconinck (eds.), *Error Estimation and Adaptive Discretization Methods in Computational Fluid Dynamics*.

26. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations*.

27. S. Müller, *Adaptive Multiscale Schemes for Conservation Laws*.

28. C. Carstensen, S. Funken, W. Hackbusch, R.H.W. Hoppe, P. Monk (eds.), *Computational Electromagnetics*.

29. M.A. Schweitzer, *A Parallel Multilevel Partition of Unity Method for Elliptic Partial Differential Equations*.

30. T. Biegler, O. Ghattas, M. Heinkenschloss, B. van Bloemen Waanders (eds.), *Large-Scale PDE-Constrained Optimization*.

31. M. Ainsworth, P. Davies, D. Duncan, P. Martin, B. Rynne (eds.), *Topics in Computational Wave Propagation*. Direct and Inverse Problems.

32. H. Emmerich, B. Nestler, M. Schreckenberg (eds.), *Interface and Transport Dynamics.* Computational Modelling.

33. H.P. Langtangen, A. Tveito (eds.), *Advanced Topics in Computational Partial Differential Equations.* Numerical Methods and Diffpack Programming.

34. V. John, *Large Eddy Simulation of Turbulent Incompressible Flows.* Analytical and Numerical Results for a Class of LES Models.

35. E. Bänsch (ed.), *Challenges in Scientific Computing - CISC 2002.*

36. B.N. Khoromskij, G. Wittum, *Numerical Solution of Elliptic Differential Equations by Reduction to the Interface.*

37. A. Iske, *Multiresolution Methods in Scattered Data Modelling.*

38. S.-I. Niculescu, K. Gu (eds.), *Advances in Time-Delay Systems.*

39. S. Attinger, P. Koumoutsakos (eds.), *Multiscale Modelling and Simulation.*

40. R. Kornhuber, R. Hoppe, J. Périaux, O. Pironneau, O. Wildlund, J. Xu (eds.), *Domain Decomposition Methods in Science and Engineering.*

41. T. Plewa, T. Linde, V.G. Weirs (eds.), *Adaptive Mesh Refinement – Theory and Applications.*

42. A. Schmidt, K.G. Siebert, *Design of Adaptive Finite Element Software.* The Finite Element Toolbox ALBERTA.

43. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations II.*

44. B. Engquist, P. Lötstedt, O. Runborg (eds.), *Multiscale Methods in Science and Engineering.*

45. P. Benner, V. Mehrmann, D.C. Sorensen (eds.), *Dimension Reduction of Large-Scale Systems.*

46. D. Kressner, *Numerical Methods for General and Structured Eigenvalue Problems.*

47. A. Boriçi, A. Frommer, B. Joó, A. Kennedy, B. Pendleton (eds.), *QCD and Numerical Analysis III*.

48. F. Graziani (ed.), *Computational Methods in Transport.*

49. B. Leimkuhler, C. Chipot, R. Elber, A. Laaksonen, A. Mark, T. Schlick, C. Schütte, R. Skeel (eds.), *New Algorithms for Macromolecular Simulation.*

50. M. Bücker, G. Corliss, P. Hovland, U. Naumann, B. Norris (eds.), *Automatic Differentiation: Applications, Theory, and Implementations.*

51. A.M. Bruaset, A. Tveito (eds.), *Numerical Solution of Partial Differential Equations on Parallel Computers.*

52. K.H. Hoffmann, A. Meyer (eds.), *Parallel Algorithms and Cluster Computing.*

53. H.-J. Bungartz, M. Schäfer (eds.), *Fluid-Structure Interaction.*

54. J. Behrens, *Adaptive Atmospheric Modeling.*

55. O. Widlund, D. Keyes (eds.), *Domain Decomposition Methods in Science and Engineering XVI.*

56. S. Kassinos, C. Langer, G. Iaccarino, P. Moin (eds.), *Complex Effects in Large Eddy Simulations.*

57. M. Griebel, M.A Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations III.*

58. A.N. Gorban, B. Kégl, D.C. Wunsch, A. Zinovyev (eds.), *Principal Manifolds for Data Visualization and Dimension Reduction.*

59. H. Ammari (ed.), *Modeling and Computations in Electromagnetics: A Volume Dedicated to Jean-Claude Nédélec.*

60. U. Langer, M. Discacciati, D. Keyes, O. Widlund, W. Zulehner (eds.), *Domain Decomposition Methods in Science and Engineering XVII.*

61. T. Mathew, *Domain Decomposition Methods for the Numerical Solution of Partial Differential Equations.*

62. F. Graziani (ed.), *Computational Methods in Transport: Verification and Validation.*

63. M. Bebendorf, *Hierarchical Matrices.* A Means to Efficiently Solve Elliptic Boundary Value Problems.

64. C.H. Bischof, H.M. Bücker, P. Hovland, U. Naumann, J. Utke (eds.), *Advances in Automatic Differentiation.*

65. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations IV*.

66. B. Engquist, P. Lötstedt, O. Runborg (eds.), *Multiscale Modeling and Simulation in Science.*

67. I.H. Tuncer, Ü. Gülcat, D.R. Emerson, K. Matsuno (eds.), *Parallel Computational Fluid Dynamics 2007.*

68. S. Yip, T. Diaz de la Rubia (eds.), *Scientific Modeling and Simulations*.

69. A. Hegarty, N. Kopteva, E. O'Riordan, M. Stynes (eds.), *BAIL* 2008 – *Boundary and Interior Layers.*

70. M. Bercovier, M.J. Gander, R. Kornhuber, O. Widlund (eds.), *Domain Decomposition Methods in Science and Engineering XVIII.*

71. B. Koren, C. Vuik (eds.), *Advanced Computational Methods in Science and Engineering.*

72. M. Peters (ed.), *Computational Fluid Dynamics for Sport Simulation.*

73. H.-J. Bungartz, M. Mehl, M. Schäfer (eds.), *Fluid Structure Interaction II - Modelling, Simulation, Optimization.*

74. D. Tromeur-Dervout, G. Brenner, D.R. Emerson, J. Erhel (eds.), *Parallel Computational Fluid Dynamics 2008.*

75. A.N. Gorban, D. Roose (eds.), *Coping with Complexity: Model Reduction and Data Analysis.*

76. J.S. Hesthaven, E.M. Rønquist (eds.), *Spectral and High Order Methods for Partial Differential Equations.*

77. M. Holtz, *Sparse Grid Quadrature in High Dimensions with Applications in Finance and Insurance*.

78. Y. Huang, R. Kornhuber, O.Widlund, J. Xu (eds.), *Domain Decomposition Methods in Science and Engineering XIX*.

79. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations V*.

80. P.H. Lauritzen, C. Jablonowski, M.A. Taylor, R.D. Nair (eds.), *Numerical Techniques for Global Atmospheric Models*.

81. C. Clavero, J.L. Gracia, F.J. Lisbona (eds.), *BAIL 2010 – Boundary and Interior Layers, Computational and Asymptotic Methods*.

82. B. Engquist, O. Runborg, Y.R. Tsai (eds.), *Numerical Analysis and Multiscale Computations*.

83. I.G. Graham, T.Y. Hou, O. Lakkis, R. Scheichl (eds.), *Numerical Analysis of Multiscale Problems*.

84. A. Logg, K.-A. Mardal, G. Wells (eds.), *Automated Solution of Differential Equations by the Finite Element Method*.

85. J. Blowey, M. Jensen (eds.), *Frontiers in Numerical Analysis - Durham 2010*.

86. O. Kolditz, U.-J. Gorke, H. Shao, W. Wang (eds.), *Thermo-Hydro-Mechanical-Chemical Processes in Fractured Porous Media - Benchmarks and Examples*.

87. S. Forth, P. Hovland, E. Phipps, J. Utke, A. Walther (eds.), *Recent Advances in Algorithmic Differentiation*.

88. J. Garcke, M. Griebel (eds.), *Sparse Grids and Applications*.

89. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations VI*.

90. C. Pechstein, *Finite and Boundary Element Tearing and Interconnecting Solvers for Multiscale Problems*.

91. R. Bank, M. Holst, O. Widlund, J. Xu (eds.), *Domain Decomposition Methods in Science and Engineering XX*.

92. H. Bijl, D. Lucor, S. Mishra, C. Schwab (eds.), *Uncertainty Quantification in Computational Fluid Dynamics*.

93. M. Bader, H.-J. Bungartz, T. Weinzierl (eds.), *Advanced Computing*.

94. M. Ehrhardt, T. Koprucki (eds.), *Advanced Mathematical Models and Numerical Techniques for Multi-Band Effective Mass Approximations*.

95. M. Azaïez, H. El Fekih, J.S. Hesthaven (eds.), *Spectral and High Order Methods for Partial Differential Equations ICOSAHOM 2012*.

96. F. Graziani, M.P. Desjarlais, R. Redmer, S.B. Trickey (eds.), *Frontiers and Challenges in Warm Dense Matter*.

97. J. Garcke, D. Pflüger (eds.), *Sparse Grids and Applications – Munich 2012*.

98. J. Erhel, M. Gander, L. Halpern, G. Pichot, T. Sassi, O. Widlund (eds.), *Domain Decomposition Methods in Science and Engineering XXI*.

99. R. Abgrall, H. Beaugendre, P.M. Congedo, C. Dobrzynski, M. Ricchiuto, V. Perrier (eds.), *High Order Nonlinear Numerical Methods for Evolutionary PDEs - HONOM 2013*.

100. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations VII*.

101. R. Hoppe (ed.), *Optimization with PDE Constraints - OPTPDE 2014*.

For further information on these books please have a look at our mathematics catalogue at the following URL: www.springer.com/series/3527

Monographs in Computational Science and Engineering

1. J. Sundnes, G.T. Lines, X. Cai, B.F. Nielsen, K.-A. Mardal, A. Tveito, *Computing the Electrical Activity in the Heart.*

For further information on this book, please have a look at our mathematics catalogue at the following URL: www.springer.com/series/7417

Texts in Computational Science and Engineering

1. H. P. Langtangen, *Computational Partial Differential Equations.* Numerical Methods and Diffpack Programming. 2nd Edition
2. A. Quarteroni, F. Saleri, P. Gervasio, *Scientific Computing with MATLAB and Octave.* 4th Edition
3. H. P. Langtangen, *Python Scripting for Computational Science*. 3rd Edition
4. H. Gardner, G. Manduchi, *Design Patterns for e-Science*.
5. M. Griebel, S. Knapek, G. Zumbusch, *Numerical Simulation in Molecular Dynamics*.
6. H. P. Langtangen, *A Primer on Scientific Programming with Python.* 4th Edition
7. A. Tveito, H. P. Langtangen, B. F. Nielsen, X. Cai, *Elements of Scientific Computing.*
8. B. Gustafsson, *Fundamentals of Scientific Computing.*
9. M. Bader, *Space-Filling Curves.*
10. M. Larson, F. Bengzon, *The Finite Element Method: Theory, Implementation and Applications.*
11. W. Gander, M. Gander, F. Kwok, *Scientific Computing: An Introduction using Maple and MATLAB.*

For further information on these books please have a look at our mathematics catalogue at the following URL: www.springer.com/series/5151