Lecture Notes in Pure and Applied Mathematics

Free and Moving Boundaries

Analysis, Simulation and Control

Volume 252

Edited by

Roland Glowinski

Jean-Paul Zolésio

Chapman & Hall/CRC
Taylor & Francis Group

Free and Moving Boundaries

Analysis, Simulation and Control

PURE AND APPLIED MATHEMATICS

A Program of Monographs, Textbooks, and Lecture Notes

Free and Moving Boundaries

Analysis, Simulation and Control

Edited by

Roland Glowinski
University of Houston
Texas, U.S.A.

Jean-Paul Zolésio
CNRS & INRIA
Sophia Antipolis, France

Chapman & Hall/CRC
Taylor & Francis Group
Boca Raton London New York

Chapman & Hall/CRC is an imprint of the
Taylor & Francis Group, an informa business

Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

No claim to original U.S. Government works
Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-1-58488-606-8 (Softcover)

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

Free and moving boundries : analysis, simulation and control / editors, Roland Glowinski and
Jean-Paul Zolésio.
p. cm. -- (Lecture notes in pure and applied mathematics ; v. 252)
Includes bibliographical references.
ISBN-13: 978-1-58488-606-8 (alk. paper)
ISBN-10: 1-58488-606-4 (alk. paper)
1. Boundary element methods. I. Glowinski, R. II. Zolésio, J. P. III. Title. IV. Series.

TA347.B69F68 2007
620.001’51535--dc22 2006102259

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Preface

This volume comprises selected papers from the International Federation of Processing (IFIP) 7.2 conference Free and Moving Boundary Analysis, Simulation and Control, which took place in Houston, Texas, in December 2004.

These papers pertain to moving boundaries and boundary control in systems described by partial differential equations and are relevant to Working Group 7.2 of Technical Group TC7 of the IFIP.

Optimal control theory was applied to distributed physical systems through boundary control theory.

In physical systems, boundary control can be understood as a moving boundary control depending on the Lagrangian or Eulerian viewpoint in the modeling. In free boundary control theory we have the concept of *arbitrary Euler–Lagrange* representation because we must deal with the two modelings. In numerical analysis the concept of fictitious domains introduced by Glowinski is a rigorous treatment of the moving domains and moving boundary conditions that must be taken into account in sensitive analysis. The prodigious progress in numerical analysis relying on the combination of finite element approximation with time discretization by operator-splitting and the Lagrange multiplier–based fictitious domains method allows flow calculation for nonacademic problems. The popular level set technique for moving geometry is a valuable tool for handling topological changes. Its application to systems represented by partial differential equations with boundary conditions is investigated together with boundary control, moving domains, topological derivatives, and coupled systems with dynamical free boundaries. This volume also discusses several applications to electromagnetic devices, flow, control, computing, image analysis, topological changes, and free boundaries.

Roland Glowinski and Jean-Paul Zolésio

About the Editors

Roland Glowinski, Ph.D., is Professor of Mathematics and Mechanical engineering at the University of Houston.

Dr. Glowinski obtained his Ph.D. in Paris in 1970. He was elected a corresponding member of the French National Academy of Sciences. He has received many awards, including the International Federation of Processing (IFIP) Silver Core Award (1985). He has published more than 300 scientific papers and 7 books, and he is coeditor of several conference proceedings. He was scientific director at INRIA and a professor at the University of Paris VI before he became an adjunct Professor of Computational and Applied Mathematics at Rice University. He is a member of the Board of Regents of the University Leonardo da Vinci, Paris; Professor Emeritus of the University Pierre and Marie Curie (Paris VI); and a Docent Professor of Computational and Applied Mathematics at the University of Jyvaskyla (Finland). He has been a member of the Scientific Council of Electricité de France (1990–1996), and Director of CERFACS, Toulouse, France (1992–1994).

Jean-Paul Zolésio, Ph.D., is Research Director at the French National Center for Research (CNRS, mathematics). Since 2000 he has been associated with the French National Institute for Research in Computer Science and Control.

Dr. Zolésio is an associate member of Centre de Recherches Mathématiques (CRM) at the University of Montreal. He obtained his *doctorat d'état* in Nice (1979). He received the IFIP Silver Core Award in 1992. He has been a visiting professor at the University of California, Los Angeles, and at the Scolla Normale di Pisa, Italy. He is the author or coauthor of more than 200 scientific articles and the coauthor or coeditor of 14 books mainly devoted to the effect of geometry variation in physical systems described by partial differential equations.

Contributors

Giles Auchmuty
Division of Mathematical Sciences
National Science Foundation
Arlington, Virginia

A.V. Balakrishnan
Flight Systems Research Center
University of California at
Los Angeles
Los Angeles, California

Louis Blanchard
INRIA
Sophia-Antipolis, France

Alexandre Caboussat
Department of Mathematics
University of Houston
Houston, Texas

Sunčica Čanić
Department of Mathematics
University of Houston
Houston, Texas

Claude Dedeban
France Telecom R et D ANT
and OpRaTel
Fort de la Tête de Chien
La Turbie, France

Mathieu Dehaes
Department of Mathematics
and Statistics
University of Montreal
Montreal, Quebec, Canada

Michel C. Delfour
Center for Mathematics
Research
Department of Mathematics
and Statistics
University of Montreal
Montreal, Quebec, Canada

Pierre Dubois
France Telecom RD ANT
and
INRIA
Fort de la Téte de Chien
La Turbie, France

Raja Dziri
Department of Mathematics
University of Tunis
Tunis, Tunisia

Karsten Eppler
Weierstraß Institute for Applied
Analysis und Stochastics
Berlin, Germany

Giorgio Fabbri
Department of Mathematics
Universitá "La Sapienza"
Rome, Italy

Tao Feng
Department of Engineering, Physics
and Mathematics
Mid Sweden University
Sundsvall, Sweden

Giovanni P. Galdi
Department of Mechanical Engineering
University of Pittsburgh
Pittsburgh, Pennsylvania

Roland Glowinski
Department of Mathematics
University of Houston
Houston, Texas
and
Université P. et M. Curie
Paris, France

Giovanna Guidoboni
Department of Mathematics
University of Houston
Houston, Texas

Mårten Gulliksson
Department of Engineering, Physics and Mathematics
Mid Sweden University
Sundsvall, Sweden

Cavit Hafizoglu
Department of Mathematics
University of Virginia
Charlottesville, Virginia

Helmut Harbrech
Institute for Computer Science and Practical Mathematics
University of Kiel
Kiel, Germany

J. Henry
INRIA-Futurs
University of Bordeaux
Talence, France

Vincent Heuveline
Institute of Applied Mathematics
University of Karlsruhe
Karlsruhe, Germany

M. Hintermüller
Department of Mathematics and Scientific Computing
University of Graz
Graz, Austria

Ronald H.W. Hoppe
Department of Mathematics
University of Houston
Houston, Texas

Katarina Jegdić
Department of Mathematics
University of Houston
Houston, Texas

Barbara Lee Keyfitz
Fields Institute
Toronto, Ontario, Canada
and
University of Houston
Houston, Texas

P.I. Kogut
Department of Differential Equations
National University of Dnipropetrovsk
Dnipropetrovsk, Ukraine

V.A. Kovtunenko
Lavrent'ev Institute of Hydrodynamics
Novosibirsk, Russia

K. Kunisch
Department of Mathematics and Scientific Computing
University of Graz
Graz, Austria

Irena Lasiecka
Department of Mathematics
University of Virginia
Charlottesville, Virginia

G. Leugering
Institute of Applied Mathematics
University of Erlangen–Nuremberg
Erlangen, Germany

Wenbin Liu
Institute of Mathematics
and Statistics
University of Kent
Canterbury, U.K.

Tsorng-Whay Pan
Department of Mathematics
University of Houston
Houston, Texas

Svetozara I. Petrova
Institute of Mathematics
University of Augsburg
Augsburg, Germany

Marco Picasso
Institute of Analysis and Scientific
Computing
École Polytechnique Fédérale
de Lausanne
Lausanne, Switzerland

Jacques Rappaz
Institute of Analysis and Scientific
Computing
École Polytechnique Fédérale
de Lausanne
Lausanne, Switzerland

Jan Sokolowski
Institut Élie Cartan, Laboratoire de
Mathématiques
Université Henri Poincaré
Nancy I
Vandoeuvre lés Nancy Cedex, France
and
Systems Research Institute
of the Polish Academy
of Sciences
Warsaw, Poland

Daniel Toundykov
Department of Mathematics
University of Virginia
Charlottesville, Virginia

Amjad Tuffaha
Department of Mathematics
University of Virginia
Charlottesville, Virginia

Antoni Zochowski
Systems Research Institute
of the Polish Academy
of Sciences
Warsaw, Poland

Jean-Paul Zolésio
CNRS and INRIA, Sophia-Antipolis
Sophia Antipolis, France

Contents

CHAPTER 1

Optimal tubes: Geodesic metric, Euler flow, and moving domain

Jean-Paul Zolésio
CNRS and INRIA, Sophia-Antipolis, France

1 Introduction to moving domain

Study of shape optimization began in the 1970s, and the first results concerned the *shape derivatives*, the *shape gradient* and the associated *shape differential equation*, which furnishes existence results for asymptotic evolution of a shape-gradient flow (see, for example, [9] and the historical introduction of [27]). Shape analysis became a branch of geometry in the 1990s and several results concerning topology, compactness and *intrinsic analysis* (mainly through the oriented distance function b_Ω), and density perimeter are found in [10], [25], and the associated literature. Recently *moving domains* appeared in several settings such as fluid dynamics, optimal control theory, large deformations, fluid-structure coupling, moving front, images (traveling), dynamical antennas, and modelings of many industrial devices. A popular version of the shape differential equation is the level set method, which emphasizes a specific parameterization of a moving domain Ω_t, its boundary being the level set zero of a one-parameter function $\phi(t, .)$. It is classical ([8], [25]) that a speed vector (whose flow mapping carries that moving domain) is

$$V^\phi(t,x) = -\frac{\partial}{\partial t}\phi(t,x)\frac{\nabla_x \phi(t,x)}{||\nabla_x \phi(t,x)||^2} \tag{1.1}$$

Any other field building this tube is in the form $W = V^\phi + Z$ with $\langle Z(t), n_t \rangle = 0$ on Γ_t.

We can see that the presence in the denominator of the term $||\nabla_x \phi||$ will not help define correctly the flow mapping of that vector field V^ϕ. At the very least we will have to control the term $||V^\phi(t,x)||_{R^N} = |\frac{\partial \phi}{\partial t}|/||\nabla_x \phi||$ in order to use the shape differential equation technique. We can bypass that difficulty as follows. Consider any shape gradient descent method for minimizing a shape functional $J(\Omega)$, whose shape gradient $G(\Omega) \in \mathcal{D}'(R^N, R^N)$

is a vector distribution with compact support included in the boundary of the domain (which, in a *smooth* situation, takes the form $G(\Omega) = \gamma^*_{\partial\Omega} \cdot (gn)$ where g, the *shape gradient density*, is a scalar distribution on the boundary with zero transverse order, n being the normal field). The shape differential equation consists of solving the nonlinear problem (see [7], [8], . . . , [25]):

$$\boxed{\forall t, \quad 0 \le t \le \tau, \qquad V(t,.) = -A^{-1}.G(\Omega_t(V))} \tag{1.2}$$

leading to the decrease of the functional:

$$J(\Omega_t(V) \le J(\Omega_0) - \alpha \int_0^t ||V(s,.)||^2_{\mathcal{D}(A^{1/2})} ds \tag{1.3}$$

Let V^* be a solution to (1.2). The problem is then to find a function ϕ such that the associated vector speed V^ϕ builds the same tube $Q^* = Q_{V^*}$. The necessary and sufficient condition (under some smoothness) is that the normal components of the two vector fields are equal *on the lateral boundary* Σ_V, that is,

$$-\frac{\partial}{\partial t}\phi/|\nabla_x\phi| = \langle V^*(t), n_t\rangle \quad \text{on } \partial\Omega_t(V^*) \tag{1.4}$$

Assume that ϕ solves Equation (1.4) on the lateral boundary Σ_{V^*}. By "multiplying" that equation by the nonnegative term $|\nabla\phi(t)|$ we find that ϕ solves the problem:

$$\frac{\partial}{\partial t}\phi(t,x) + \langle \nabla_x\phi(t,.), V^*(t)\rangle = 0 \quad \text{on } \partial\Omega_t(V^*) \tag{1.5}$$

An obvious way to solve Equation (1.5) is to consider the *global* convection problem:

$$\frac{\partial}{\partial t}\phi(t,x) + \langle \nabla_x\phi(t,.), \bar{V}^*(t)\rangle = 0 \quad \text{in } D \tag{1.6}$$

where $\bar{V}^*$ is *any admissible extension* of the cylinder $(0,\tau)\times D$ of the vector field $V^*|_{\Sigma_{V^*}}$ (the restriction of V^* to the lateral boundary of the tube Q_{V^*}). Possible choices of the vector field $\bar{V}^*$ are V itself, but there are many other examples, one of which is $\bar{V}^* = V^* op^*$, where p^* stands for the projection mapping or the local (or *narrow*) $(p^*)^h$ projection onto the boundary $\partial\Omega_t(V^*)$. We provide existence and uniqueness results for that convection Equation (1.6) when $\bar{V}^*, div\bar{V}^*$ are in $L^1(0,\tau,L^2(D))$, while the initial condition is $\phi_0 \in L^\infty(D)$.

Let V^* and ϕ be solutions, respectively, to (1.2) and (1.6). Then we get

$$\frac{\partial}{\partial t}\phi(t,x)/|\nabla\phi| = \langle \nabla_x\phi(t,.)/|\nabla\phi|, \bar{V}^*\rangle$$

so that

$$\boxed{\left|\frac{\partial}{\partial t}\phi(t,x)/|\nabla\phi\right|\,| \leq ||\bar{V}^*(t,x)||_{R^N}}. \tag{1.7}$$

Assume that $\bar{V}^* = V^*$. Then if $V^* \in E = L^1(0,\tau,L^2(D,R^N))$, with ϕ being the solution to Equation (1.6), we get $V^\phi \in E$. If we assume that $divV^\phi \in E$, then for any given $\phi_0 \in L^\infty(D)$ we get the existence of a solution to (1.6).

In the classical setting (developed in [7] and [8]) the shape gradient G of the shape functional J is bounded (in some "negative" Sobolev space of distributions over the universe D) and continuous (with respect to the *Courant metric*, see [25]), then $\forall k, k \geq 1$, the shape differential equation has a smooth solution $V^* \in \mathcal{C}^{0,k} = C^0([0,\tau], C^k(\bar{D},R^N)) \subset E$. Then the flow mapping $T_t(V^*)$ is classically defined and the unique solution to the convection problem (1.6), if $\bar{V}^*$ is also chosen in $\mathcal{C}^{0,k}$, is given by

$$\boxed{\phi(t) = \phi_0 o T_t(\bar{V}^*)^{-1}} \tag{1.8}$$

As $t \rightarrow T_t(\bar{V}^*)^{-1}(.) \in \mathcal{C}^{1,k}$ (see [5]) assuming the initial data $\phi_0 \in C^k(\bar{D})$, we get $\phi \in \mathcal{C}^{0,k-1}(D\backslash K_\phi)$, where the compact set $K_\phi = \{x \in D\ s.t.\ \nabla\phi(x) = 0\}$.

As $\phi(t) = \phi_0 o T_t^{-1}$ we get $\nabla\phi(t) = ((DT_t)^{-1} \cdot \nabla\phi_0) o T_t^{-1}$ so that $K_{\phi_0} =$ empty set implies that $\forall t, K_{\phi(t)}$ is empty. Then $V^\phi \in \mathcal{C}^{0,k-1}$ (assuming now that $k \geq 2$) and $\Omega_t(V^\phi = \Omega_t(\bar{V}^*)\Omega_t(V^*)$. (In other words, the tree vector fields $V^*, \bar{V}^*, V^\phi$ built the same tube Q_V as they have the same normal speed v on the lateral boundary Σ_V.) From Equation (1.2) we get

$$-\frac{\partial}{\partial t}\phi(t,x) + \langle \nabla_x\phi(t,.), A^{-1}.G(\Omega_t(V^\phi))\rangle = 0, \tag{1.9}$$

which is a Hamilton-Jacobi–like equation for the function ϕ. From 1.3, we get

$$J(\Omega_t(V) \leq J(\Omega_0) - \alpha \int_0^t \left\| \frac{\partial\phi(s,.)}{\partial t} / \nabla_x\phi(s,.) \right\|^2_{\mathcal{D}(A_k^{1/2})} ds \tag{1.10}$$

Briefly, we could say that the shape differential Equation (1.2) is solved by the fixed point method (see [7], [27]) in a classical setting that does not permit topological changes in the moving domain. We introduce here the weak mathematical setting that permits handling of that equation with possible topological changes by avoiding any homeomorphism.

Topological evolution requires zero capacity connecting vanishing subsets: in order to treat mathematically a complete example with a shape functional governed by a *boundary PDE problem* with conditions such as the Dirichlet, Neumann, or Robin types, one has to be very careful because any topological change, such as the separation of two components, should respect these conditions in the sense that the vanishing connecting set should be with *zero capacity* in order not to force a boundary condition on an artificial boundary. As far as the H^1 Sobolev space type is concerned, it is necessary to go to dimension 3 in order to handle separation with a connecting vanishing line (with codimension 2) having zero capacity. This short introduction is designed to assist in understanding the need for a mathematical setting for weak evolution of geometry. The classical geometry flow mapping is replaced by the more general tube evolution concept. The *characteristic function* $\zeta(t,x) = \zeta(t,x)^2$ is convected by Fields V^* and $\bar{V}^*$ so that

$$\zeta(t)\phi(t) = \phi(t)^+ \tag{1.11}$$

A tube is the $N+1$ dimensional graph of the n dimensional moving domain. In fact it is just an $N+1$ dimensional domain, but it is interesting to decouple the time and space variable roles in the evolution.

2 Tubes by product spaces

We consider a bounded open domain D in R^N and the smooth vector field

$$V \in L^1(0,\tau,W^{1,\infty}(D,R^N))$$

verifying a "D-bilateral viability" condition, say $\langle V,n\rangle = 0$ on ∂D. Then the flow mapping $T_t(V)$ associated with V smoothly maps the set D onto itself. For any measurable subset $\Omega \subset D$ we defined the transported (or perturbed) domain $\Omega_t = T_t(V)(\Omega)$ whose characteristic function $\zeta(t) = \zeta_\Omega o T_t^{-1}(V)$ solves, in a distribution sense, the classical convection Equation (2.16). Given initial data ϕ_0 and ψ_0 in $L^2(\Omega)$ and right-hand sides f and g in $L^1(0,\tau,L^2(D))$, we consider the following problems

$$\phi(0) = \phi_0, \quad \frac{\partial}{\partial t}\phi + \nabla\phi \cdot V = f \tag{2.12}$$

$$\psi(0) = \psi_0, \quad \frac{\partial}{\partial t}\psi + div(\psi V) = g \tag{2.13}$$

Given two real numbers (p,q), $1 < p \le \infty$, $1 \le q < \infty$, we consider the linear space for speed vector fields:

$$\begin{aligned} E^{p,q} &= \{V \in L^p(0,\tau,L^q(D,R^N)) \; s.t. \; divV \in L^p(0,\tau,L^q(D))\} \\ \mathcal{E}^{p,q} &= \{V \in E^{p,q}, \; s.t. \; V \cdot n = 0 \text{ in } W^{-1,1}(\partial D)\} \end{aligned} \tag{2.14}$$

The null condition on the normal component of the vector field at the boundary can be weakly written as

$$\forall \phi \in L^2(I,C^1(\bar{D})), \qquad \int_I \int_D (divV\phi + \nabla\phi.V)\; dtdx = 0 \tag{2.15}$$

Proposition 2.1

([13]) *Assume* $V \in \mathcal{E}^{2,2}$. *If* $(divV)^+ \in L^1(0,\tau,L^\infty(D))$, *problem* (*2.17*) *has solutions*

$$\phi \in L^\infty(0,\tau,L^2(D)) \cap H^1(0,\tau,H^{-1}(D)) \subset C^0([0,\tau],H^{-1/2}(D))$$

If $(divV)^- \in L^1(0,\tau,L^\infty(D))$ *problem 2.18 has solutions*

$$\psi \in L^\infty(0,\tau,L^2(D)) \cap H^1(0,\tau,H^{-1}(D)) \subset C^0([0,\tau],H^{-1/2}(D))$$

The first idea would be to consider $divV \in L^1(0,\tau,L^\infty(D))$. Then both problems have solutions. We might be tempted to conclude that each problem has a unique solution. That argument does not apply as one of the two solutions ϕ or ψ should be smooth in order to be “put in duality.” Then under previous poor regularity on V we will not get existence or uniqueness for the shape convection problem 2.16:

$$\zeta(0) = \zeta_{\Omega_0}, \quad \frac{\partial}{\partial t}\zeta + \nabla\zeta \cdot V = 0, \quad \zeta = \zeta^2 \tag{2.16}$$

Functional setting: To give meaning to the product $\nabla\zeta \cdot V$ in (2.16), we write it as

$$\nabla\zeta \cdot V = div(\zeta V) - \zeta divV$$

Then, as soon as $\zeta \in L^\infty((0,\tau) \times D)$, that term makes sense in $L^1(0,\tau, W^{-1,1}(D))$ when V and its divergence $divV$ are in $L^1(0,\tau,L^1(D))$. Through Equation (2.16), any solution ζ, $\zeta \in L^\infty \subset L^1(0,\tau,W^{-1,1}(D))$, is then an element of $W^{1,1}(0,\tau,W^{-1,1}(D))$, then $\zeta \in C^0([0,\tau],W^{-1,1}(D))$. Note that in the tube elements the continuity for the characteristic function ζ is stronger. As $\zeta = \zeta^2$ we get the continuity in $L^p(D)$ for all finite p:

Proposition 2.2

Let $V \in \mathcal{E}^{1,1}$. *If we consider any solution* ζ *to the convection problem* (*2.16*), *then*

$$\zeta \in C^0([0,\tau],L^1(D))$$

An important point is that when $f = 0$ and the initial condition ϕ_0 lies between 0 and 1, among the solutions to problem (2.17), there exists a solution verifying $0 \leq \zeta \leq 1$:

Proposition 2.3

Let $V \in \mathcal{E}^{1,1}$, $f = 0$ *and* $c \leq \Phi_0(t,x) \leq d$, *a.e.* (t,x), *then there exists a solution* $\zeta \in L^\infty((0,\tau) \times D)$ *to problem* (*2.17*) *such that* $c \leq \zeta(t,x) \leq d, a.e.$ (t,x).

The proof is immediate with the use of a smooth sequence of vector field $V_n \to V$ in $\mathcal{E}^{1,1}$. Let $\zeta_n = \Phi_0 o(T_t(V_n)^{-1})$ be the classical solution to (2.17) associated with speed vector V_n. As evidence we get $c \leq \zeta_n \leq d$. Considering a subsequence that weakly converges to an element ζ in $L^\infty(I \times D)$, the conclusion derived in the limit is the following weak formulation of 2.17: $\forall \psi \in C^1(I \times D), \psi(\tau, .) = 0$,

$$\int_0^\tau \int_d \zeta_n(-\psi_t - div(\psi V_n)\ dxdt = \int_D \Phi_0(x)\psi(0,x)dx.$$

By the same technique we derive Corollary 2.4:

Corollary 2.4

Given $V \in \mathcal{E}^{1,1}, \Phi_0^1$ and Φ_0^2 in $L^\infty(D)$ with $\Phi_0^1(x) \leq \Phi_0^2(x)$ a.e. $x \in D$. Then there exist two solutions ζ^1 and ζ^2 in $L^\infty(I \times D)$ to (2.17) with the right-hand side $f = 0$ verifying the same order $\zeta^1(t,x) \leq \zeta^2(t,x)$, a.e. $(t,x) \in I \times D$.

It is important to note that as far as Equation (2.17) is considered, with right-hand side $f = 0$, for bounded initial conditions with the existence of the solution monotonically depending on that initial condition, that result requires no extra regularity on the divergence of the vector fields. We consider a bounded open domain D in R^N and a smooth vector field $V \in L^1(0, \tau, W^{1,\infty}(D, R^N))$ verifying a "D-bilateral viability" condition, say $\langle V, n \rangle = 0$, on ∂D. Then the flow mapping $T_t(V)$ associated with V maps smoothly the set D onto itself. For any measurable subset $\Omega \subset D$ we defined the transported (or perturbed) domain $\Omega_t = T_t(V)(\Omega)$ whose characteristic function $\zeta(t) = \zeta_\Omega o T_t^{-1}(V)$ solves, in a distribution sense, the classical convection equation 2.16. Given initial data ϕ_0 and ψ_0 in $L^2(\Omega)$ and right-hand sides f and g in $L^1(0, \tau, L^2(D))$, we consider the following problems:

$$\phi(0) = \phi_0, \quad \frac{\partial}{\partial t}\phi + \nabla\phi.V = f \tag{2.17}$$

$$\psi(0) = \psi_0, \quad \frac{\partial}{\partial t}\psi + div(\psi V) = g \tag{2.18}$$

Given two real numbers (p,q), $1 < p \leq \infty$, $1 \leq q < \infty$, we consider the linear space for speed vector fields:

$$E^{p,q} = \{V \in L^p(0, \tau, L^q(D, R^N))\ s.t.\ div\ V \in L^p(0, \tau, L^q(D))\}$$

$$\mathcal{E}^{p,q} = \{V \in E^{p,q},\ s.t.\ V \cdot n = 0 \text{ in } W^{-1,1}(\partial D)\} \tag{2.19}$$

The null condition on the normal component of the vector field at the boundary can be weakly written as

$$\forall \phi \in L^2(I, C^1(\bar{D})), \quad \int_I \int_D (divV\phi + \nabla\phi \cdot V)\ dtdx = 0 \tag{2.20}$$

Proposition 2.5

(*Quoted from* [13]) *Assume* $V \in \mathcal{E}^{2,2}$. *If* $(divV)^+ \in L^1(0,\tau,L^\infty(D))$ *problem 2.17 has solutions*

$$\phi \in L^\infty(0,\tau,L^2(D)) \cap H^1(0,\tau,H^{-1}(D)) \subset C^0([0,\tau],H^{-1/2}(D))$$

If $(divV)^- \in L^1(0,\tau,L^\infty(D))$ *problem (2.18) has solutions*

$$\psi \in L^\infty(0,\tau,L^2(D)) \cap H^1(0,\tau,H^{-1}(D)) \subset C^0([0,\tau],H^{-1/2}(D))$$

2.1 Tube definition

Being given a measurable subset $\Omega_0 \subset D$, we consider the tubes built by V with bottom Ω_0 as being: Let

$$\mathcal{H} = C^0([0,\tau],L^1(D)) \cap L^1(0,\tau,BV(D))$$

Then

$$\mathcal{T}^{p,q}_{\Omega_0} = \{(\zeta,V) \in \mathcal{H} \times \mathcal{E}^{p,q},\ \text{verifying 2.16}\} \tag{2.21}$$

The limit case $q = 1$ will be of great interest in this study while the ideal situation $p = q = 1$ is still out of the scope of this paper. For several variational situations we shall consider $p = q$ and then we shall write $\mathcal{E}^p$. As soon as ζ verifies $\zeta = \zeta^2$ it can be written as the characteristic function of a measurable subset Q in $(0,\tau) \times D$ (defined up to a $N+1$ zero measure subset), that is $\zeta = \chi_Q$; also for $a.e.t$ we shall write $\zeta(t) = \chi_{\Omega_t}$. The condition $\zeta \in L^1(0,\tau,BV(D))$ is then equivalent to the more usual one

$$\int_0^\tau P_D(\Omega_t)\ dt < \infty \tag{2.22}$$

2.2 $\mathcal{T}^{p,q}_{\Omega_0}$ is weakly closed

We define the topological locally convex linear space $L^\infty(0,\tau,C^0_{comp}(D,R^N))$ as being the inductive limit of the Banach spaces $L^\infty(0,\tau,C^0(K,R^N))$ when K ranges among the compact subsets of D. A sequence g_n converges to zero in that space if there exists a compact set K such that

$$\forall n, \quad g_n \in L^\infty(0,\tau,C^0(K,R^N)), \quad ||max_{x\in\bar{D}}|g_n(.,x)|\,||_{L^\infty(0,\tau,R^N)} \to 0$$

We consider the Banach vector spaces of bounded (vector) measures $\mathcal{M}^1(D,R^N)$ and the Banach space $L^1(0,\tau,\mathcal{M}^1(D,R^N))$, which turns out to be in duality with $L^\infty(0,\tau,C^0_{comp}(D,R^N))$ through the obvious bilinear form

$$\langle\mu,g\rangle := \int_0^1 \langle\mu(t),g(t)\rangle dt$$

We recall here the following result of [26]:

Theorem 2.6

Assume $p > 1, q > 1$. Let $(V_n, \zeta_n) \in \mathcal{T}^{p,q}_{\Omega_0}$ verifying the following three weak convergences (w.c.):

$$i)\ V_n,\ divV_n\ w.c.to\ V(resp.\ divV)\ in\ L^p(0,\tau,L^q(D)) \tag{2.23}$$

The previous weak convergences refer to $\sigma(L^p(0,\tau,L^q(D)), L^{p^*}(0,\tau,L^{q^*}(D)))$ topology.

$$ii)\ \forall g \in L^\infty(0,\tau,C^0_c(D,R^N)), \quad \int_0^\tau \langle\nabla\zeta_n(t), g(t)\rangle dt \to \int_0^\tau \langle\nabla\zeta(t), g(t)\rangle dt \tag{2.24}$$

Then the element (V,ζ) belongs to $\mathcal{T}^{p,q}_{\Omega_0}$.

We investigate now the very important limiting case where p and/or q are equal to 1 (see [32]):

Theorem 2.7

Assume $p > 1$ and $q = 1$. Let $(V_n, \zeta_n) \in \mathcal{T}^{p,1}_{\Omega_0}$ and assume that:

i) V_n (resp. $divV_n$) is bounded in $L^p(0,\tau,L^1(D,R^N))$ (resp. $L^p(0,\tau,L^1(D))$) and V_n (resp. $divV_n$) is weakly convergent in $L^p(0,\tau,\mathcal{M}^1(D,R))$ to some bounded measure $\mu \in L^p(0,\tau,\mathcal{M}^1(D,R^N))$ (resp. $div\mu$)

$$ii)\ \forall g \in L^\infty(0,\tau,C^0_c(D,R^N)), \int_0^\tau \langle\nabla\zeta_n(t), g(t)\rangle dt \to \int_0^\tau \langle\nabla\zeta(t), g(t)\rangle dt$$
$$\to \int_0^\tau \langle\nabla\zeta(t), g(t)\rangle dt$$

Then the sequence ζ_n strongly converges in $L^1(0,\tau,L^1(D))$ to some element $\zeta = \zeta^2 \in l^\infty((0,\tau \times D))$

The weaker situation concerns the limiting case $p = 1$, $q > 1$ and the pointwise constraint as follows:

Theorem 2.8

Let $q \geq 1$, and let $(V_n, \zeta_n) \in \mathcal{T}^{1,q}_{\Omega_0}$ and assume that there exists an element $\theta \in L^1(0,\tau)$ such that

$$a.e.\ t, 0 < t < \tau, \quad ||V_n(t)||_{L^1(D,R^N)} + |divV_n|_{L^1(D)} \leq \theta(t) \tag{2.25}$$

so that V_n (resp. $divV_n$)are bounded in $L^1((0,\tau)\times D, R^N)$ (resp. in $L^1((0,\tau)\times D)$) as $||V_n||_{L^1((0,\tau)\times D,R^N)} \leq |\theta|_{L^1(0,\tau)}$. Moreover, assume:

i) V_n *(resp.* $divV_n$*) is weakly convergent in* $\mathcal{M}^1(]0,\tau[\times D, R^N))$ *to some bounded measure* $\mu \in \mathcal{M}^1(]0,\tau[\times D, R^N))$ *(resp.* $div\mu$*)*

ii) $\forall g \in L^\infty(0,\tau, C^0_c(D,R^N)), \int_0^\tau \langle \nabla\zeta_n(t), g(t)\rangle dt \to \int_0^\tau \langle \nabla\zeta(t), g(t)\rangle dt$

Then the sequence ζ_n *strongly converges in* $L^1(0,\tau,L^1(D))$ *to some element* $\zeta = \zeta^2$*.*

2.3 Tubes associated with BV vector fields

Let $V \in \mathcal{E}^p$, Equipped with the graph norm $\mathcal{E}^p$ is a Banach space.

Proposition 2.9

Let $1 \le p < \infty$; *then* $H^1_0(D)$ *is a dense subspace in* $\mathcal{E}^p(D)$*.*

The proof is made in three steps.

Lemma 2.10

For any $V \in \mathcal{E}^p(D)$*, let* V^0 *designate the extension out of* D *by zero. Then we have* $divV^0 = (divV)^0$ *and* $V^0 \in \mathcal{E}^p := \{V \in L^p(R^N,R^N), divV \in L^p(R^N)\}$*.*

Let b designate the oriented distance function to the bounded domain D. For a given $h\rangle 0$, small enough parameter, consider the cut-oriented distance function with support in the narrow band h, $b_h := \rho o b$ where the cutting function ρ is smooth, positive, with support in the interval $[-2h,+2h]$ and taking the value $\rho = 1$ on the subinterval $[-h,+h]$. We introduce the flow mapping $T_h := T_h(\nabla b_h)$ and consider the following mapping:

$$\mathcal{T}_h : V \in \mathcal{E}^p(D) \longrightarrow V_h := det(DT_h) DT_h V^0 o T_h$$

Lemma 2.11

$div(V_h) = detDT_h div(V^0) o T_h$

It derives that $divV_h = detDT_h((divV)^0) o T_h \in L^2(R^N)$. Moreover, as $h\rangle 0$ we have $V_h \cdot n = 0$ so that

$$V_h \in \mathcal{E}^p(D)$$

We consider a mollifier $r_h \to \delta_0$ as $h \to 0$ with support of $r_h \subset B(0,h/2)$ so that the support verifies (for $h\rangle 0$) $supt\{r_h * (\mathcal{T}_h \cdot V)\} \subset D$ so that

$$V^h := r_h * (\mathcal{T}_h \cdot V) \in \mathcal{D}(D,R^N) \subset H^1_0(D,R^N).$$

We can now verify that $V^h \to V$ strongly in $\mathcal{E}^p(D)$, which establishes the density result without any geometrical assumption on the domain D. Then we may restrict to the bounded domain D the results from [22]:

Theorem 2.12

Let $V \in \mathcal{E}^{1,1} \cap L^1(0, \tau, BV(D, R^N))$. Assume that $(divV)^+$ (resp. $(divV)^-$) is in $L^1(0, \tau, L^\infty(D, R^N))$, then problem 2.16 (resp. 2.18) has a unique solution ζ such that $(\zeta, V) \in \mathcal{T}^1_{\Omega_0}$.

Then we see that some regularity on V implies that for a given V, the characteristic solution is unique. We denote it by ζ_V. The converse is false: for a given ζ the set of V such that $(\zeta, V) \in \mathcal{T}^{p,q}_\Omega$ is a closed convex set that we denote by $\mathcal{K}(\zeta)$. For the convex set $1 < p, q < \infty$, we can find a *minimal* element V_ζ (which minimizes the associated norms in $\mathcal{K}(\zeta)$).

2.4 Boundedness of the density perimeter

Following [3] and [4], we consider for any closed set A in D the density perimeter associated to any $\gamma > 0$ by the following.

$$P_\gamma(A) = sup_{\epsilon \in (0,\gamma)} \left[\frac{meas(A^\epsilon)}{2\epsilon} \right] \tag{2.26}$$

where A^ϵ is the dilation $A^\epsilon = \cup_{x \in A} B(x, \epsilon)$. We recall some main properties:

1. The mapping $\Omega \to P_\gamma(\partial\Omega)$ is lower-semicontinuous in the H^c-topology,
2. The property $P_\gamma(\partial\Omega) < \infty$ implies that meas $(\partial\Omega) = 0$ and $\Omega \setminus \partial\Omega$ is open in D.
3. If $P_\gamma(\partial\Omega_n) \leq m$ and Ω_n converges in the H^c-topology to some open subset
4. $\Omega \subset D$, then the convergence holds in the $L^2(D)$ topology.

The parabolic situation: whenever $V \in C^\infty(I \times \bar{D})$, the mapping $t \to P_\gamma(\partial\Omega_t)$ is not continuous in that the mapping cannot be an element of $H^1(0, \tau)$. For any smooth vector field, $V \in C^0([0, \tau], W^{1,\infty}_0(D, R^N))$, we consider

$$\Theta_\gamma(V, \Omega_0) = Min \left\{ \int_0^\tau \left(\frac{\partial}{\partial t} \mu \right)^2 dt | \mu \in \mathcal{M}_\gamma(V, \Omega_0) \right\} \tag{2.27}$$

where

$$\mathcal{M}_\gamma(V, \Omega_0) = \{\mu \in H^1(0, \tau), P_\gamma(\partial\Omega_t(V)) \leq \mu(t) \; a.e.t, \; \mu(0) \leq (1+\gamma) P_\gamma(\partial\Omega_0)\}$$

In general that set is nonempty. When that set is empty, we put $\Theta_\gamma(V, \Omega_0) = +\infty$. Notice that even when the mapping $p = (t \longrightarrow P_\gamma(\Omega_t(V)))$ is an element of $H^1(0, \tau)$ (then $p \in \mathcal{M}_\gamma(V, \Omega_0)\}$), we may have: $\Theta(V, \Omega_0) < ||p'||^2_{L^2(0,\tau)}$ as the minimizer will escape to a possible variation of the function p.

Proposition 2.13

Let $V \in C^0([0,\tau], W_0^{1,\infty}(D,R^N)), divV = 0,$ *we have:*

$$P_\gamma(\partial\Omega_t(V)) \leq 2P_\gamma(\partial\Omega_0) + \sqrt{\tau}\Theta(V,\Omega_0)^{1/2} \tag{2.28}$$

Moreover, if $V_n \in C^0([0,\tau], W_0^{1,\infty}(D,R^N))$, *verifies* $V_n \longrightarrow V$ *in* $L^2((0,\tau)\times D, R^N)$ *and the uniform boundedness:* $\exists M \rangle 0, \Theta(V_n,\Omega_0) \leq M$ *then*

$$\Theta(V,\Omega_0) \leq liminf\Theta(V_n,\Omega_0)$$

An *alternative* approach to consider is

$$\tilde{\Theta}_\gamma(V,\Omega_0) := ||p_\gamma||_{BV(0,\tau)}, p_\gamma(t) := P_\gamma(\partial\Omega_t(V))$$

We would derive the same kind of estimates.

2.5 Tube energy — variational problem

For any given positive constants $a > 0, \sigma \geq 0, \mu \geq 0$ and $\nu \geq 0$, we shall consider the minimization associated with the following functionals:

$$J^a(V) = 1/2\int_I\int_D (a+\xi_V)(|V(t,x)|^2 + (divV(t,x))^2)\ dxdt \tag{2.29}$$

$$J^a_{\sigma,\mu,\nu}(V) = J^a(V)$$

$$+\sigma\int_0^\tau ||\nabla(\xi_V(t))||_{M^1(D)}dt + \mu\Theta(V,\Omega_0) + \nu\int_0^\tau\int_D DV..DV\ dxdt \tag{2.30}$$

We shall consider the three situations associated with $\sigma+\mu+\nu > 0$ and $\sigma\mu\nu = 0$. When ν is zero, the terms σ and μ will play a surface tension role at the dynamical interface, while the case $\sigma + \mu = 0$ should be considered a mathematical regularization as in the nonusual variational interpretation developed in the previous section, $\nu > 0$ does not lead to the usual viscosity term (i.e., does not lead to the Navier-Stokes equations).

Theorem 2.14

Assuming $V \in \mathcal{E}^2$, $a > 0$, $\sigma.\eta.\nu > 0$, *there exists* $V \in \mathcal{E}^2$ *such that* $\forall W \in \mathcal{E}^2$*:*

$$J^a_{\sigma,\eta,\nu}(V) \leq J^a_{\sigma,\eta,\nu}(W)$$

Let us consider a minimizing sequence V_n of $J^a_{\sigma,\eta,\nu}$ in $\mathcal{E}^2$.

Then V_n is bounded in $L^2(0,\tau,L^2(D,R^N))$, we consider a subsequence weakly converging to V. By the compactness results, we get the strong convergence of the characteristic functions: ξ_{V_n} strongly converges to ξ_{V_n}. For any $\Phi \in L^2(0,\tau,L^2(D))$ we get $\xi_{V_n}\Phi$, which strongly converges to $\xi_{V_n}\Phi$ (as ξ_{V_n} converges almost everywhere and is dominated by 1). Then $\xi_{V_n}V_n$

weakly converges to $\xi_V V$ in $L^2(0,\tau,L^2(D))$:

$$\forall\Phi \in L^2(0,\tau,L^2(D)), \int_0^\tau \int_D \xi_{V_n} V_n \Phi \; dxdt$$
$$= \int_0^\tau \int_D (\xi_{V_n}\Phi) V_n \; dxdt \to \int_0^\tau \int_D (\xi_V \Phi) V \; dxdt$$

We set $\xi_{V_n}(V_n)^2 = (\xi_{V_n} V_n)^2$, then

$$\int_0^\tau \int_D (\xi_V V)^2 \; dxdt \leq liminf \int_0^\tau \int_D (\xi_{V_n} V_n)^2 \; dxdt.$$

It follows that the limiting element V realizes the minimum of the functional $J^a_{\sigma,\eta,\nu}$ over the linear space $\mathcal{E}^2$.

2.6 Saddle point

Let us consider the following more general situation: we consider a vector field V in $\mathcal{E}^{2,2}$ and any function $G \in L^\infty(D)$ verifying $G \geq \alpha > 0$.

We consider the Lagrangian expression for the functional

$$g(V) = Inf_{\zeta\in\mathcal{U}_V} Sup_{\phi\in\mathcal{H}_V} \mathcal{L}_V(\zeta,\phi)$$

with

$$\mathcal{L}_V(\zeta,\phi) = \int_0^\tau \int_D \left\{ 1/2\zeta^2 G + \zeta\left(\frac{\partial}{\partial t}\phi + div(\phi V)\right)\right\} dxdt - \int_{\Omega_0} \phi(0)dx,$$

$$\mathcal{H}_V = \left\{\phi \in L^2(0,\tau,L^2(D)) \, s.t. \, \frac{\partial}{\partial t}\phi + div(\phi V) \in L^2(I\times D), \phi(\tau) = 0\right\},$$

$$\mathcal{U}_V = \left\{\zeta \in L^2(I\times D), \; s.t. \; \frac{\partial}{\partial t}\zeta + \nabla\zeta V) \in L^2(I\times D), \zeta(0) = \chi_{\Omega_0}\right\}.$$

Note that the elements of $\mathcal{H}_V$ are continuous $\phi \in C^0([0,\tau], W^{-1/2,1}(D)$, so that $\phi(\tau)$ makes sense.

The Lagrangian $\mathcal{L}_V$ is concave-convex on $\mathcal{U}_V \times \mathcal{H}_V$. $L^2(I\times D)\times\mathcal{E}^{2,2}$.

Saddle points (ξ,λ) are the solution to the system composed of Equation (2.17) (with $\Phi_0 = \chi_{\Omega_0}, f = 0$) and the following backward adjoint equation

$$\frac{\partial}{\partial t}\lambda + div(\lambda V) = -\xi_V G, \quad \lambda(\tau) = 0 \qquad (2.31)$$

The converse is true when we have an extra density condition on V and $divV$:

Assumption on V : $\{\phi \in C^\infty(I\times D)\cap\mathcal{E}^{2,2}\}$ is dense in $\mathcal{H}_V$ (2.32)

That weakly coupled system (2.17), (2.31) possesses solutions when $(divV)^+ \in L^1(I, L^\infty(D))$. We derive the following uniqueness results for the convection problem (2.17):

Proposition 2.15

Assume $V \in \mathcal{E}^{2,2}$, verifying 2.32 and $(divV)^+ \in L^1(I, L^\infty(D))$.

Then, with $f = 0$ and $\Phi_0 = \chi_{\Omega_0}$ (convection problem), or more generally $\Phi_0 \in L^\infty(D)$, problem (2.17) possesses a unique solution ζ_V verifying

$$0 \le \zeta_V \le 1 \ a.e.(t, x) \in I \times D$$

or (in the more general setting)

$$Infess\ \Phi_0 \le \zeta_V \le Supess\ \Phi_0.$$

We have the monotony: $\Omega_0^1 \subset \Omega_0^2$ (or, in the more general setting $\Phi_0^1 \le \Phi_0^2$) implies $\zeta_V^1 \le \zeta_V^2$.

Proof

From the strong assumption (2.32), the set of saddle points is not empty and is completely characterized by the system of Equations (2.17) through (2.31). let us denote by $\mathcal{S}_V$ the set of saddle points. We know that it can be written as $\mathcal{S}_V = A_V \times B_V$, which means that if (ζ^i, λ^i), $i = 1, 2$ are saddle points, then ζ^1, λ^2 and ζ^2, λ^1 are also saddle points. We derive that Equation (2.31), with right-hand side $G\zeta^i$, has solutions, and we derive uniqueness of ζ_V (from the fact that $G > 0$), single element in A_V (in other words A_V is reduced to a single element ζ_V). From uniqueness we know that $0 \le \zeta \le 1$ (see Proposition 2.3).

2.6.1 Level sets considerations

Parameterization of moving domains by level sets is a classical geometrical viewpoint, considering the boundary as manifold defined through implicit definition as $\partial\Omega_t = \{x \in D, s.t. \Phi(t, x) = 0\}$. Since 1980 the derivative of *shape* functional $J(\Omega)$ has been considered in [8] (see also [25]) with domains $\Omega_t = \{x \in D,\ s.t.\ \Phi(t, x)\rangle 0\}$ (or the specific case $\Omega_t = \{x \in D,\ s.t.\ \Psi(x)\rangle t\}$). When the function Φ is smooth enough, it turns out that $\Omega_t = T_t(V^\Phi)(\Omega_0)$ where the speed vector field is given by $V^\Phi(t, x) := -\frac{\partial}{\partial t}\Phi(t, x)||\nabla\Phi(t, x)||^{-2}\nabla\Phi(t, x)$. Of course the flow mapping is defined when Φ is such that this field is Lipschitzian continuous in a neighborhood of $\partial\Omega_t$. The weaker tube considerations permit enlargement of the definition of the domain convection when $V^\Phi \in \mathcal{E}^{2,2}$ with the constraint on the positive part of the divergence.

$$0 < a_0,\ s.t.\ a_0||\nabla\Phi(t,.)|| \leq \left|\frac{\partial}{\partial t}\Phi(t,.)\right| \leq m\,||\nabla\Phi(t,.)|| \tag{2.33}$$

The fact that $divV^\Phi \in L^\infty(D)$ is a much more difficult context:

$$||\nabla\Phi||^2 divV^\Phi = \nabla(\Phi_t)\cdot\nabla\Phi - \Phi_t||\nabla\Phi||^{-2}\nabla(||\nabla\Phi||^2) + \Phi_t\Delta\Phi$$

2.6.2 Derivative with respect to the speed field V

Functional J in the form of min max has a well-known Gateaux derivative onto the parameter V. Now it is important to notice that in the present saddle point formulation the linear vector spaces depend on the parameter V.

With $div \in L^1(I, L^\infty(D))$ the Ambrosio results derive as we have $H^1(D) \subset W^{1,1}(D) \subset BV(D)$.

Applying the results from [2] and [18] for differentiation under uniqueness of the saddle point, assuming

$$V \in B := L^2(0,\tau, H_0^1(D, R^N)), \quad \mathcal{L}_V^\nu := \mathcal{L}_V + \nu/2\int_0^\tau\int_D DV\,..\,DV\ dxdt$$

we have

$$J'(V,W) = \langle (a+\zeta)V - \nabla(\zeta divV) - \nu\Delta V - \lambda\nabla\zeta, W\rangle_{F'\times F}$$

We shall now explain the term $\lambda\nabla\zeta$, using the concept of *transverse field* Z, which was introduced in [12] and developed in [5], [11], [13], [17], [26], and [27]. Indeed, we have

$$\lambda\nabla\zeta = \nabla(\zeta\lambda) - \zeta\nabla\lambda$$

We shall explain that last term $\zeta\nabla\lambda$ through the *transverse field* problem and its adjoint.

2.7 Transverse derivative

For two given vector fields V and W, the *transverse* field Z is the solution to the Lie bracket evolution

$$H_V\cdot Z = W, H_V\cdot Z := \frac{\partial}{\partial t}Z + [Z,V], [Z,V] := DZ\cdot V - DV\cdot Z \tag{2.34}$$

To *avoid some technicalities* we assume now (as will be the case in the first example associated with arteris modeling) that V and W are free divergence vector fields: $divV = divW = 0$. The adjoint operator is then:

$$H_V^*\cdot\Lambda := -\frac{\partial}{\partial t}\Lambda - D\Lambda\cdot V - D^*V\cdot\Lambda \tag{2.35}$$

Lemma 2.16

$$H_V \cdot (\zeta \vec{e}) = \left(\frac{\partial}{\partial t}\zeta + \nabla\zeta \cdot V \right) \vec{e} + \zeta H_V \cdot \vec{e} \tag{2.36}$$

$$H_V^*(\zeta \vec{f}) = -\left(\frac{\partial}{\partial t}\zeta + \nabla\zeta \cdot V \right) \vec{f} - \zeta H_V^* \cdot \vec{f} \tag{2.37}$$

We see that if (ζ, V) is a tube, then the two previous expressions simplify as the first term vanishes from (2.16). Specifically, we have

$$H_V^* \cdot (\zeta\nabla\lambda) = -\zeta \left(\frac{\partial}{\partial t}\nabla\lambda + D\nabla\lambda \cdot V + D^*V \cdot \nabla\lambda \right) = -\zeta\nabla \left(\frac{\partial}{\partial t}\lambda + \nabla\lambda \cdot V \right)$$

Then if we set the scalar operator h_V as

$$h_V \cdot \psi := \frac{\partial}{\partial t}\psi + \nabla\psi \cdot V$$

we get (ζ, V), being a tube:

$$H_V^*(\zeta\nabla\lambda) = -\zeta\nabla(h_V \cdot \lambda) \tag{2.38}$$

Note that here as $divV = 0$ the adjoint operator verifies $h_V^* = -h_V$. By reversing the time, that operator h_V is then self-adjoint. Nevertheless it is *not* maximally defined in the admissible setting, so that it is not possible to use the same classical abstract setting of evolution linear system in order to solve (2.17) and/or (2.18).

An important issue in this study is the existence of the derivative of the characteristic function ξ with respect to the vector field V, in case of uniqueness of $\xi := \zeta_V$ associated with V (for example, when $V \in L^1(I, BV(D, R^N))$. We give an existence result for that derivative. Indeed $a_s = \frac{\xi^s - \xi}{s}$ does converge in the space of measures to the element $\dot{\xi}$. From the weak formulation of (2.17) we get

$$\int_0^\tau \int_D \left(a_s \left(\frac{\partial}{\partial t}\phi + \langle V, \nabla\phi \rangle + \phi divV \right) + \xi^s \langle \nabla\phi, W \rangle \right) \, dxdt = 0 \tag{2.39}$$

Lemma 2.17

As $s \to 0$ we have $\xi^s \to \xi$ in $L^2((0, \tau) \times D))$

We set

$$f = \frac{\partial}{\partial t}\phi + \langle V, \nabla\phi \rangle$$

If we assume the field to be *time smooth*, $V \in W^{1,\infty}([0, \tau], L^3(D))$, then for any $f \in W^{1,\infty}([0, \tau],\ L^{6/5}(D))$, there exists a ϕ with $\phi(\tau) = 0$ solving the previous equation, that is:

Proposition 2.18

Let $V \in W^{1,\infty}([0,\tau], L^3(D))$, the sequence $\frac{\xi^s - \xi}{s}$ converges in

$$W^{1,\infty}([0,\tau], L^{6/5}(D))'$$

to an element $\dot{\xi}$.

3 Variational principle in the Euler problem

The two-fluid functional: In the moving domain Ω_t we assume the density $\rho_i = 1 + a$ while in the exterior domain $\Omega_t^c = \bar{D} \setminus \Omega_t$ the density is $\rho_e = a$. The one-fluid configuration will correspond to $a = 0$. Then we assume $a \geq 0$. The Euler functional is

$$J(V) = \int_I \int_D [(a + \xi_V) 1/2 |V(t,x)|^2 - \xi_V g(t,x) - f(t,x) \cdot V(t,x)]\, dxdt \tag{3.40}$$

where ξ_V (which we shall denote by ξ when no confusion is possible) is the solution to the following problem:

$$\xi(0) = \xi_{\Omega_0}, \quad \frac{\partial}{\partial t}\xi + \nabla\xi \cdot V = 0, \quad \xi = \xi^2 \tag{3.41}$$

In that section we assume that V is a divergence-free vector field so that we get:

Theorem 3.1

i) The functional J is Gateaux differentiable on $E^\infty = E \cap C^0([0,\tau], C^\infty(\bar{D}, R^N))$.

ii) If we assume that $V \in E^\infty$ is such that $\forall W \in E^\infty, J'(V,W) = 0$, then there exists $P \in \mathcal{D}'(D, R^N)$ such that:

$$\frac{\partial}{\partial t}((a + \xi_V)V) + D((a + \xi_V)V) \cdot V + \nabla P = g \nabla \xi_V \tag{3.42}$$

The proof follows the classical control theory approach: field V can be considered a control parameter, while the ξ solution of 2.16 is the state variable. In that approach the derivative of Equation (3.41) leads to the derivative ξ' solution of Equation (3.43) below. That equation is not completely studied here, but as field V is assumed to be smooth, ξ is obtained by the flow mapping $T_t(V)$ so that it is easily verified that the mapping $V \longrightarrow \xi$ is Gateaux differentiable in $L^2(0,\tau, H^2(D) \cap H_0^1(D))$. By direct calculus, as V is smooth we have:

$$J'(V,W) = \int_I \int_D (\xi'(1/2|V|^2 - g) + [(a + \xi)V - f]W)\, dxdt$$

where ξ' is the Gateaux derivative of ξ at V in the direction W, given by:

$$\xi'(0) = 0, \quad \frac{\partial}{\partial t}\xi' + \nabla\xi' \cdot V = -\nabla\xi \cdot W \tag{3.43}$$

We introduce the adjoint problem:

$$\lambda(\tau) = 0, \quad -\frac{\partial}{\partial t}\lambda - \nabla\lambda \cdot V = 1/2|V|^2 - g \tag{3.44}$$

$$\begin{aligned} J'(V,W) &= \int_I \int_D \left(\left(-\frac{\partial}{\partial t}\lambda - \nabla\lambda \cdot V \right) \xi' + [(a+\xi)V - f]W \right) \, dxdt \\ &= \int_0^\tau \int_D \left(\left(\frac{\partial}{\partial t}\xi' + \nabla\xi' \cdot V \right) \lambda + [(a+\xi)V - f]W \right) \, dxdt \\ &= \int_0^\tau \int_D ((-\nabla\xi \cdot W)\lambda + ((a+\xi)V - f)W) \; dxdt \end{aligned}$$

But as

$$\int_D ((-\nabla\xi \cdot W)\lambda dx = \int_D \xi \cdot W(\nabla\lambda) \; dx$$

we get

$$J'(V,W) = \int_0^\tau \int_D (\xi\nabla\lambda + (a+\xi)V - f)W \; dxdt \tag{3.45}$$

We consider the vector field Λ defined as follows:

$$\Lambda = \xi_V \nabla\lambda$$

and Λ solves the following problem.

Proposition 3.2

The variable Λ solves the backward problem

$$\Lambda(\tau) = 0, -\frac{\partial}{\partial t}\Lambda - D\Lambda \cdot V - D^*V \cdot \Lambda = \xi\nabla(1/2|V|^2) \tag{3.46}$$

Proof

With (2.16) we get

$$-\frac{\partial}{\partial t}\xi\nabla\lambda - \nabla\xi \cdot V\nabla\lambda = 0 \tag{3.47}$$

Also from (3.44) we get

$$\frac{\partial}{\partial t}\Lambda = \frac{\partial}{\partial t}\xi\nabla\lambda + \xi\nabla\left(\frac{\partial}{\partial t}\lambda\right)$$

But from the equation verified by λ we get:

$$-\nabla\left(\frac{\partial}{\partial t}\lambda\right)-\nabla(\nabla\lambda\cdot V)=1/2\nabla(|V|^2)-\nabla g$$

After multiplying by ξ and with $\nabla(A\cdot B)=D^*A\cdot B+D^*B\cdot A$ we get:

$$-\xi\nabla\left(\frac{\partial}{\partial t}\lambda\right)-\xi(D^*(\nabla\lambda)\cdot V+D^*V\cdot\nabla\lambda)=1/2\xi\nabla(|V|^2)-\xi\nabla g \quad (3.48)$$

By adding (3.47) and (3.48) we get:

$$\begin{aligned}-\xi\nabla\left(\frac{\partial}{\partial t}\lambda\right)-\frac{\partial}{\partial t}\xi\nabla\lambda-\nabla\xi\cdot V\nabla\lambda-\xi D^*(\nabla\lambda)\cdot V-D^*V\cdot(\xi\nabla\lambda)\\ =1/2\xi\nabla(|V|^2)-\xi\nabla g.\end{aligned}$$

Note that

$$\begin{aligned}[\nabla\xi\cdot V\nabla\lambda+\xi D^*(\nabla\lambda)\cdot V]_i&=\frac{\partial}{\partial x_j}\xi V_j\frac{\partial}{\partial x_i}\lambda+\xi\frac{\partial}{\partial x_i}\left(\frac{\partial}{\partial x_j}\lambda\right)V_j\\ &=\left(\frac{\partial}{\partial x_j}\xi\frac{\partial}{\partial x_i}\lambda+\xi\frac{\partial}{\partial x_j}\left(\frac{\partial}{\partial x_i}\lambda\right)\right)V_j\\ &=\frac{\partial}{\partial x_j}\left(\xi\frac{\partial}{\partial x_i}\lambda\right)V_j=\frac{\partial}{\partial x_j}(\Lambda_i)V_j=(D\Lambda\cdot V)_i\end{aligned}$$

and then $\nabla\xi\cdot V\nabla\lambda+\xi D^*(\nabla\lambda)\cdot V=D\Lambda\cdot V$.

3.1 Necessary optimality condition

The extremes of J at a smooth vector field V lead to the existence of a distribution π such that:

$$\xi_V\nabla\lambda+(a+\xi_V)V-f=-\nabla\pi, \quad (3.49)$$

that is, we have the condition:

$$\Lambda=f-(a+\xi_V)V-\nabla\pi \quad (3.50)$$

Plugging the necessary condition (3.50) in (3.46), we get:

$$\begin{aligned}-\frac{\partial}{\partial t}(f-(a+\xi_V)V-\nabla\pi)-D(f-(a+\xi_V)V-\nabla\pi)\cdot V\\ -D^*V\cdot(f-(a+\xi_V)V-\nabla\pi)=\xi\nabla(1/2|V|^2-g)\end{aligned}$$

which can be rewritten as follows:

$$\begin{aligned}
&\frac{\partial}{\partial t}((a+\xi_V)V)+D((a+\xi_V)V)\cdot V+\frac{\partial}{\partial t}(\nabla\pi-f)+D(\nabla\pi)\cdot V+D^*V\cdot(\nabla\pi)\\
&\qquad -(D(f)\cdot V+D^*V\cdot f)\\
&\qquad =-D^*V\cdot((a+\xi_V)V)+\xi\nabla(1/2|V|^2-g)\\
&\qquad =-aD^*V\cdot V-\xi_V D^*V\cdot V+1/2\xi_V\nabla(|V|^2)-\xi\nabla g
\end{aligned}$$

Here we have an interesting simplification through the classical Lemma 3.3.

Lemma 3.3

$$D^*V\cdot V=1/2\nabla(|V|^2) \tag{3.51}$$

Then we get

$$=-a/2\nabla(|V|^2)-\xi\nabla g$$

That simplication must be underlined here. The previous lemma applies as the integrand function in the very definition of the functional J is chosen as the kinetic energy itself. If the functional J was in the form

$$J(V)=\int_0^\tau\int_D(a+\xi_V)\mathcal{R}(V)+\cdots$$

we would then have the term

$$\xi_V(-D^*V\cdot V+1/2\nabla(\mathcal{R}(V))$$

which would not cancel.

Finally, as $D(\nabla\pi)\cdot V+D^*V\cdot(\nabla\pi)=\nabla(\nabla\pi\cdot V)$ we get:

$$\begin{aligned}
&\frac{\partial}{\partial t}((a+\xi_V)V)+D((a+\xi_V)V)\cdot V)\nabla\left(\frac{\partial}{\partial t}(\pi)+\nabla\pi\cdot V+a/2|V|^2\right)\\
&\qquad =\frac{\partial}{\partial t}f+(D(f)\cdot V+D^*V\cdot f)-\xi\nabla g\\
&\qquad =(D^*(f)\cdot V+D^*V\cdot f)+(D(f)-D^*f)\cdot V\xi\nabla g\\
&\qquad =\nabla(f\cdot V)+(D(f)-D^*f)\cdot V-\xi\nabla g
\end{aligned}$$

then we get V as a solution to the problem

$$\frac{\partial}{\partial t}(\rho_V V)+D(\rho_V V)\cdot V+\nabla P=\frac{\partial}{\partial t}f+(D(f)-D^*f)\cdot V-\xi\nabla g \tag{3.52}$$

where the density is

$$\rho_V=(a+\xi_V)$$

and the pressure is given by

$$P = \frac{\partial}{\partial t}\pi + \nabla\pi \cdot V + a/2|V|^2 - f \cdot V$$

If we assume that $\sigma(f) = Df - D^*f$ is zero, then $curlf = 0$ (as $\sigma..\sigma = |curlf|^2$) and f derives from a potential, for example in the following form:

$$f(t,x) = \int_0^t \nabla F(\sigma, x) d\sigma$$

Then V solves the following Euler equation:

$$\frac{\partial}{\partial t}(\rho_V V) + D(\rho_V V) \cdot V + \nabla P = \nabla F - \xi \nabla g$$

In fact F is of no use in the functional, as any additive gradient term can be absorbed by the pressure term P as follows:

$$\frac{\partial}{\partial t}(\rho_V V) + D(\rho_V V) \cdot V + \nabla P = g \nabla \xi \tag{3.53}$$

with

$$P = \frac{\partial}{\partial t}\pi + \nabla\pi \cdot V + a/2|V|^2 - f \cdot V + \xi g - F. \tag{3.54}$$

4 Metric and geodesic

The final time is $\tau = 1$. The interval is $I =]0,1[$.

We consider the distance between two sets Ω_i, $i = 1,2$ such that there exists a tube (ζ, V) *connecting* them:

$$\zeta(0) = \chi_{\Omega_1}, \quad z(1) = \chi_{\Omega_2} \tag{4.55}$$

$$\mathcal{T}^1_{\Omega_1,\Omega_2} := \{(\zeta, V) \in \mathcal{T}^{1,1}_{\Omega_1}, \text{verifying } 4.55\,\} \tag{4.56}$$

We set the distance

$$d(\Omega_1, \Omega_2) := Inf \left\{ \left\| \frac{\partial}{\partial t}\zeta \right\|_{L^1(I, W^{-1,1}(D))}, (\zeta, V) \in \mathcal{T}^1_{\Omega_1,\Omega_2} \right\} \tag{4.57}$$

Notice that $||\frac{\partial}{\partial t}\zeta||_{L^1(I,W^{-1,1}(D))} = ||\nabla\zeta \cdot V||_{L^1(I,W^{-1,1}(D))}$. So that, in smooth situations we would have:

$$d(\Omega_1, \Omega_2) = sup_{\mu \in L^\infty(I, W^{1,\infty}_0(D))}$$

$$\int_0^1 \int_{\partial\Omega_t} \mu V(t) \cdot n_t d\Gamma_t dt / ||\mu||_{L^\infty(I, W^{1,\infty}_0(D))}$$

so that

$$d(\Omega_1, \Omega_2) = \int_0^1 \int_{\partial\Omega_t} ||V(t) \cdot n_t| d\Gamma_i \, dt$$

(here we assume the boundary of D far enough from Γ_t).

Obviously the two first axioms for a distance are verified. We consider the triangle property. Given three sets Ω_i we consider any $\epsilon > 0$. There exists $(\zeta^{1,\epsilon}, V^{1,\epsilon})$ (resp. $(\zeta^{2,\epsilon}, V^{2,\epsilon})$) connecting Ω_1 and Ω_2 (resp. Ω_2 and Ω_3) and achieving the infimum up to ϵ in the definition of the distances $d(\Omega_1, \Omega_2)$ (resp. $d(\Omega_2, \Omega_3)$). Also there exists $\phi^{i,\epsilon} \in L^1(I, W_0^{1,\infty}(D))$ realizing, up to ϵ, the norms $||\frac{\partial}{\partial t}\zeta^{i,\epsilon}||_{L^1(I,W^{-1,1}(D))}$. We consider the vector field V defined as

$$\begin{aligned} V^\epsilon &= 2V^{1,\epsilon}(2t, x), \;\; 0 < t < 1/2, \\ &= 2V^{2,\epsilon}(2t-1, x), 1/2 < t < 1 \end{aligned} \tag{4.58}$$

We consider

$$\begin{aligned} \zeta^\epsilon(t, x) &= \zeta^{1,\epsilon}(2t, x), \;\; 0 < t < 1/2, \\ &= \zeta^{2,\epsilon}(2t-1, x), 1/2 < t < 1 \end{aligned} \tag{4.59}$$

Then $(\zeta^\epsilon, V^\epsilon)$ is a tube that connects the sets Ω_1 and Ω_3 in the sense of (4.55): $(\zeta^\epsilon, V^\epsilon) \in \mathcal{T}^1_{\Omega_1,\Omega_3}$. Then from the distance definition we have

$$d(\Omega_1, \Omega_3) \leq \left\| \frac{\partial}{\partial t}\zeta^\epsilon \right\|_{L^1(I,W^{-1,1}(D))}$$

which by definition of these norms and using (2.16), leads to

$$\text{there exists } \phi^\epsilon = \phi, \int_0^1 (||\nabla\phi(t)||_{L^\infty(D,R^N)} dt \leq 1$$

such that

$$d(\Omega_1, \Omega_3) \leq \langle \zeta_t^\epsilon, \phi \rangle + \epsilon = \int_0^1 \int_D \zeta^\epsilon div(\phi V^\epsilon) \; dxdt + \epsilon$$

We consider the following specific test element:

$$\begin{aligned} \phi^\epsilon &:= \phi^{1,\epsilon}(2t, x), \quad 0 < t < 1/2, \\ &= \phi^{2,\epsilon}(2t-1, x), \quad 1/2 < t < 1 \end{aligned}$$

We get for that specific admissible element

$$d(\Omega_1, \Omega_3) \leq \int_0^1 \int_D \zeta^\epsilon div(\phi^\epsilon V^\epsilon) \; dxdt$$

That last time the integral can be decomposed into two additive time integrals on $[0,1/2]$ and $[1/2,1]$. After a change of variables in both we get:

$$d(\Omega_1,\Omega_3) \leq \Sigma_{\{i=1,2\}} \int_0^1 \int_D \zeta^{i,\epsilon} div(\phi^{i,\epsilon} V^{i,\epsilon})\ dxdt$$

But

$$\left| \int_0^1 \int_D \zeta^{i,\epsilon} div(\phi^{i,\epsilon} V^{i,\epsilon})\ dxdt - \left\| \frac{\partial}{\partial t} \zeta^{i,\epsilon} \right\|_{L^1(I,W^{-1,1}(D))} \right| \leq \epsilon$$

and

$$\left| d(\Omega_1,\Omega_2) - \left\| \frac{\partial}{\partial t} \zeta^{i,\epsilon} \right\|_{L^1(I,W^{-1,1}(D))} \right| \leq \epsilon$$

so that

$$d(\Omega_1,\Omega_3) \leq d(\Omega_1,\Omega_2) + d(\Omega_2,\Omega_3) + 4\epsilon$$

But $\epsilon > 0$ is arbitrarily small, and the triangle inequality derived is:

Proposition 4.2

Given any measurable subset $\Omega \subset D$ the family

$$\mathcal{O}_\Omega^{1,d} := \{\Omega' \subset D,\ s.t.\ \mathcal{T}^1_{\Omega,\Omega'} is\ not\ empty\},$$

equipped with the distance d defined at (4.57), is a metric space.

In order to get a *complete* metric space, we modify the distance as follows, with $G = L^1(I, W^{-1,1}(D))$ and $\mathcal{L}^1 = \{V \in L^1(D, R^N),\ s.t.\ divV = L^1, V \cdot n = 0\}$ (we restrict this to the family of sets with *given volume*). We set

$$p(t) := ||\nabla\zeta||_{M^1(D,R^N)}, |p'|_{M^1(I)}$$

$$\delta(\Omega_1,\Omega_2) = Inf\left\{ \left\| \frac{\partial}{\partial t}\zeta \right\|_G + a||p'||_{M^1(I)} + b||V||_{\mathcal{L}^1} + c||V||_{L^1(I,BV(D,R^N))}, \right.$$
$$\left. (\zeta, V) \in \mathcal{T}^1_{\Omega_1,\Omega_2} \right\} \tag{4.60}$$

Theorem 4.2

Let $a > 0, b \geq 0, c \geq 0$. The family

$$\mathcal{O}_\Omega^{1,\delta} := \{\Omega' \subset D,\ s.t.\ \mathcal{T}^1_{\Omega,\Omega'}\ contains\ elements\ such\ that$$

$$\nabla\zeta \in L^1(I, M^1(D,R^N)), V \in L^1(I, BV(D)) divV = 0\},$$

equipped with the distance δ defined by (4.60), is a complete metric space.

The term $||p||_{L^1(0,1)}$ could be replaced by the easier one: $\int_0^1 p(t)dt$. This would be perfect as far as the completion of the metric would be concerned, but it fails on the first metric axiom. In [32] one finds three examples of "pseudo distances" built around this idea: introducing corrections such as $\int_0^1 |p(t) - \int_0^1 p(s)ds|dt$, also by taking $c = 0$ and $p > 1$ in the space definitions.

Proof

Let Ω_n be a Cauchy sequence. Then $\delta(\Omega_1, \Omega_n) \leq M$.

As a consequence there exist tubes $(\zeta^{\epsilon,n} V_n^\epsilon)$ verifying the minimum in $\delta(\Omega_1, \Omega_n)$ up to ϵ and verifying $||\zeta_t|| + ||\nabla\zeta||_{L^1(I,M1(D))} + ||V|| \leq 2M$. Then there exists a subsequence such that ζ^n weakly converges and V^n strongly converges to V in L^1. From the closure results for the tubes we may conclude that (ζ, V) is a tube and

$$\int_D \zeta^n(1)\psi dx = \int_{\Omega_1} \psi dx + \int_0^1 \int_{\Omega_t^n} V^n \cdot \nabla\psi \; dxdt$$

which passes to the limit so that $\Omega_n \to \Omega^*$ in the δ metric where $\chi_{\Omega^*} := \zeta(1)$.

We see also in that proof that there exists an element, the geodesic of which realizes the minimum in the definition of the metric δ, at least for $c > 0$. In view of the closure results for the family of tubes under $a > 0$, $b > 0$, $c = 0$ we address here the question of extending that metric to a larger setting. For example, in ([32]), using (2.8), we are able to extend that theorem with $c = 0$ but with the vector fields V in the following closed convex set associated with a given element $\theta \in L^1(0,1)$

$$||V(t,.)||_{L^1(D,R^N)} + ||divV||_{L^1(D)} \leq \theta(t), a.e.t$$

Many considerations underlie this question. The main point is that everything is easier if we replace the L^1 structures by L^p, $p > 1$, but we lose the triangle property of the metric. Then it is necessary to propose a large family of pseudo metrics (i.e., verifying $d(\Omega_1, \Omega_3) \leq 2^p(d(\Omega_1, \Omega_2) + d(\Omega_2, \Omega_3))$. Also it is easy to have open sets in the family by replacing the perimeter by the *density perimeter*, that is, replacing the perimeter term by the functional Θ previously introduced.

4.1 Geodesic tube characterization

An important *Algo* question is to derive a characterization for the geodesic tube. Applications are, for example, in image analysis (traveling, warping). This can be done using the previous transverse derivative leading to a backward adjoint state for solving the *shape geodesics.* In that direction we recall the basic calculus of [26]; see also [27]. In the metric definition or surface tension term for the free boundary in the fluid, the time integral of the perimeter can be generalized by the lateral surface of the tube (see the case $f = 1$).

The optimal tube in the metric definition (which is the geodesic) in the smooth situation is classically determined via the *transverse derivative*, which we have briefly summarized previously. It is characterized in terms of the transverse field and its *backward adjoint* λ, which leads to a usual backward principle. In the remainder of this section we briefly introduce that material. Assume we consider the distance built on the $L^2(\partial\Omega_t)$ norm, that is, V solves, for all admissible W

$$d(\Omega_1, \Omega_2) \leq R(V + sW)$$

with

$$R(V + sW) = \int_0^1 \left(\int_{\partial\Omega_t(V+sW)} |\langle V(t) + sW(t), n_t \rangle|^2 d\Gamma_t \right)^{1/2} dt$$

The first-order necessary condition is derived through

$$R'(V, W) = 1/2 \int_0^1 \left(\int_{\Gamma_t} (\langle V(t), W(t) \rangle)^2 \Gamma_t \right)^{-1/2} dt \int_0^1 R_1(t) dt$$

where

$$R_1(t) = \left[\frac{\partial}{\partial s} \int_{\partial\Omega_t(V+sW)} (\langle V(t) + sW(t), n_{t,s} \rangle)^2 d\Gamma_t(V + sW) \right]_{\{s=0\}}$$

We make use of

$$\int_{\partial\Omega_t(V+sW)} (\langle V(t) + sW(t), n_{t,s} \rangle)^2 d\Gamma_t(V + sW)$$
$$= \int_{\Omega_t(V+sW)} div\{\langle V(t) + sW(t), \nabla b_{t,s} \rangle (V(t) + sW(t)\} d\Omega_t(V + sW)$$

so that $R_1(t) = R_{1a} + R_{1b}(t) + R_{1c}(t)$ with:

$$R_{1a}(t) = \int_{\Gamma_t} 2\langle V(t), W(t) \rangle d\Gamma_t,$$

$$R_{1b}(t) = \int_{\Gamma_t} \{\langle 2DV \cdot V, n_t \rangle + \langle D^2 b_t \cdot V(t), V(t) \rangle$$
$$+ \langle V(t), n_t \rangle div V(t)\} \langle Z(t), n_t \rangle d\Gamma_t$$

$$R_{1c}(t) = - \int_{\Gamma_t} \langle D^2 b_t \cdot Z(t) + D^* Z(t) \cdot n_t, V(t) \rangle \langle V(t), n_t \rangle d\Gamma_t$$

Then we introduce the backward tangential problem, $\lambda^1 \in L^2(I, L^2(\partial\Omega_t))$,

$$\lambda^1(1) = 0, \lambda^1_t + div_{\partial\Omega_t}(\lambda^1 V) = g^1, \tag{4.61}$$

$$g^1 := \{\langle 2DV \cdot V, n_t\rangle + \langle D^2 b_t \cdot V(t), V(t)\rangle + \langle V(t), n_t\rangle div V(t)\}$$

and we get

$$\int_0^1 (R_{1a}(t) + R_{1b}(t))\, dt = \int_0^1 \int_{\partial\Omega_t} \langle 2V + \lambda^1 n_t, W(t)\rangle d\Gamma_t\, dt$$

The tool is in [27].

$$H^*\Lambda = g^1\nabla\zeta, \Lambda = \lambda^1\nabla\zeta, \int_0^1 R_{1b}(t)\, dt = \langle H^*\Lambda, Z\rangle = \langle \Lambda, H \cdot Z\rangle = \langle \Lambda, W\rangle$$
$$= \langle \lambda^1\nabla\zeta, W\rangle = \int_0^1 \int_{\Gamma_t} \lambda_1 \langle W, n_t\rangle d\Gamma_t\, dt)$$

The *elimination* of $Z(t)$ in the term $R_{1c}(t)$ is more delicate. Assume the vector field is searched in the form $V(t) = v(t)n_{\Gamma_t(V)}$ and $DV(t) \cdot n_{\Gamma_t(V)} = 0$ on $\Gamma_t(V)$. The previous expressions simplify: $g^1(t) = v(t) div V(t)$ on $\Gamma_t(V)$, also $\langle D^2 b_{\Gamma_t(V)} \cdot Z(t), n_{\Gamma_t(V)}\rangle = 0$ and $\langle DZ(t) \cdot V(t), n_{\Gamma_t(V)}\rangle = v(t)\langle DZ(t) \cdot n_{\Gamma_t(V)}, n_{\Gamma_t(V)}\rangle$, so that

$$R_{1c}(t) = -\int_{\Gamma_t(V)} \langle DZ(t) \cdot n_t, n_t\rangle(\langle V(t), n_t\rangle)^2 d\Gamma_t(V)$$

The transverse tube is governed by the normal component $Z(t) \cdot n_t$ so that the term $\langle DZ.n, n\rangle = \langle \nabla(\langle Z, \nabla b\rangle), n\rangle$ can be chosen equal to zero, so that R_{1c} vanishes and the optimal field realizing the distance would solve the following new equation:

$$V(t) = v(t)n_t, \quad DV(t) \cdot n_t = 0, \quad V(t) = -1/2\lambda(t)n_t$$

with

$$\lambda_t - 1/2 div_{\partial\Omega_t}(\lambda^2 n_t) = v(t) div V(t) \tag{4.62}$$

The difficult point is to determine which are the admissible perturbation speed vector field W. Here we assume that all field W are admissible. That point still has to be discussed. We refer to a forthcoming paper.

An important *alternative* to that tube analysis is to avoid many difficulties by considering (as was done in [6]) global constraints on the $BV(I \times D)$ norm of ζ. But necessary conditions would lead, for example, to the free boundary in fluid, to time-space curvature, which is not physical as far as I understand surface tension modeling. Also for the metric the closure and compactness arguments would be nice, but the first metric axiom would fail (as said before). Nevertheless, for many reasons it is an important issue to consider the measure of the lateral boundary $\Sigma := \cup\{t\} \times \Omega_t$ when $\zeta(t,.) = \chi_{\Omega_t}$.

We could completely avoid any field V considerations in this tube analysis by setting the following alternative tubes approach:

$$S := L^1(I \times D) \cap W^{1,1}(I, W^{-1,1}(D)) \subset C^0([0,1], W^{-1/2,1}(D))$$

$$\mathcal{T}_{\Omega_0}^{\mathbf{N+1},1} := \{\zeta \in S,\ s.t.\ \zeta = \zeta^2, \zeta(0) = \chi_{\Omega_0}\} \tag{4.63}$$

Then the *closure of tubes family* results are simply the following: Let such ζ_n be bounded in S with the additive *lateral boundary* boundedness:

$$P^{N+1}(Q_\zeta) := ||\nabla_{t,x}\zeta_n||_{M^1(]0,1[\times D)} \leq M \tag{4.64}$$

Then there exist subsequences weakly converging in S to some element $\zeta = \zeta^2$ as by the compact injection of $BV(I \times D)$ in $L^1(I \times D)$, which weakly converge in S, stands as a strong $L^1(I \times D)$ convergence.

An important question is to know if, being given an element $\zeta \in S$ with $\zeta(0) = \zeta(0)^2$, there exists a vector field V such that $(zeta, V)$ is a tube (i.e., verifying (2.16)), and under which conditions that vector field would derive from a potential: $V(t,x) = \nabla_x \Phi(t,x)$. The answer to these two questions are positives when the lateral boundary Σ of the tube is smooth enough. We consider the time-space normal field on Σ written as

$$\nu(t,x) = \frac{1}{\sqrt{(1 + v(t,x)^2)}}(-v(t,x), n_t(x)) \tag{4.65}$$

We call v the *normal speed of the boundary* as, following [7], any smooth vector field V whose flow $T_t(V)$ builds that Q (i.e., such that $\Omega_t = T_t(V)(\Omega_0)$) verifies $\langle V(t), n_t \rangle = v(t)$ on $\Gamma_t := \partial\Omega_t$. Conversely, if $v = \langle V, n_t \rangle$ then the field V built Q. When Q is smooth enough it can always be built by a gradient: there exists $V = \nabla A$, which builds Q (see [26]).

4.1.1 Smooth tubes are built by gradient flow

Consider a smooth tube $Q = \cup_{\{0<t<\tau\}}\{t\} \times \Omega_t$ with lateral boundary $\Sigma = \cup_{\{0<t<\tau\}}\{t\} \times \partial\Omega_t$ smooth enough in R^{N+1}. The unitary normal field outgoing to Q is $\nu = \sqrt{1 + v^2}(-v, n_t)$

Proposition 4.3

Consider the potential function: $\Phi(t,x) = v(t) op_t b^h_{\Omega_t}$, *then*

$$\nabla_x \Phi(t,.) = v(t) n_t \textit{ on } \partial\Omega_t$$

The proof consists of computing directly the gradient of that potential

$$\nabla_x \Phi = \nabla(vop_t) b^h_{\Omega_t} + v(t) op_t \nabla_x b^h_{\Omega_t}$$

as $b^h_{\Omega_t} = 0$ on $\partial\Omega_t$ and $\nabla b_{\Omega_t} = n_t$ derives the result.

The necessary conditions for optimality associated with the tube functional involving $P^{N+1}(\zeta)$ implies the transverse derivative of that term. As in smooth situations, it is a tube lateral boundary (i.e., a noncylindrical time–space boundary in R^{N+1}). We synthesize here some results from [26] (see also [27]) concerning the time–space mean curvature of the lateral boundary Σ and the $N+1$ shape derivatives. Also the oriented distance itself can be differentiated with respect to the field.

4.1.2 Transverse field and oriented distance function

From [1], [25], and [30], we know that the oriented distance function solves the convection problem

$$\frac{\partial}{\partial t} b_{\Omega_t(V)} + \nabla b_{\Omega_t(V)} . V(t, p_{\partial\Omega_t(V)}) = 0 \tag{4.66}$$

When the vector field V is smooth enough, then the projection mapping is in BV. Then from Theorem 2.7 (also [22]) we get the uniqueness of the solution b to that convection problem.

We introduce the perturbation of the vector field $V + sW$ and the derivative of the oriented distance function with respect to that perturbation parameter, that is

$$\dot{b}_{\Omega_t(V);W} := \frac{\partial}{\partial s} b_{\Omega_t(V+sW)}|_{\{s=0\}} \tag{4.67}$$

It solves the following linear problem:

$$\frac{\partial}{\partial t}\dot{b} + \nabla\dot{b} \cdot V(t) op(t) - \langle \nabla b(t), (DV(t)) op(t) \cdot \nabla(b\dot{b}) \rangle = -\nabla b(t) \cdot W \tag{4.68}$$

In order to deal with derivative, with respect to the field V in several such expressions, we give here some technical developments including the shape derivative of the oriented distance function introduced in [30] (see also [1]).

4.1.3 Mean curvature of the lateral time–space boundary

Assuming the moving domain is smooth enough, we consider the normal speed term v chosen as $v = \langle V(t), \nabla b_{\Omega_t(V)} \rangle$ and

$$\frac{\partial}{\partial t}\left(\frac{v}{\sqrt{1+v^2}}\right) = \frac{1}{(\sqrt{1+v^2})^3} \frac{\partial}{\partial t} v$$

but

$$\frac{\partial}{\partial t} v = \left\langle \frac{\partial}{\partial t} V, \nabla b \right\rangle + \left\langle \frac{\partial}{\partial t} \nabla b, V \right\rangle$$

Now we have

$$\frac{\partial}{\partial t} b_{\Omega_t(V)} = -\langle V(t), n_t \rangle o p_t \tag{4.69}$$

where p_t is the projection onto the boundary $\Gamma_t(V) = \partial\Omega_t(V)$. And

$$\frac{\partial}{\partial t} \nabla b_{\Omega_t(V)} = -(\nabla_{\Gamma_t} \langle V(t), n_t \rangle) o p_t$$

then we get

$$\frac{\partial}{\partial t} v = \left\langle \frac{\partial}{\partial t} V(t), n_t \right\rangle - \langle (\nabla_{\Gamma_t} \langle V(t), n_t \rangle) o p_t, V_{\Gamma_t} \rangle$$

$$\frac{\partial}{\partial t} \left(\frac{v}{\sqrt{1+v^2}} \right) = \frac{1}{\left(\sqrt{1+v^2}\right)^3} \left(\left\langle \frac{\partial}{\partial t} V(t), n_t \right\rangle \right) - \langle \nabla_{\Gamma_t} o p_t, V_{\Gamma_t} \rangle$$

On the other hand we have

$$div \left(\frac{1}{(\sqrt{1+v^2})} n \right) = - \left\langle \nabla \left(\frac{1}{(\sqrt{1+v^2}}, n \right\rangle \right) + \frac{1}{(\sqrt{1+v^2})^3} divn$$

so that we get

$$div \left(\frac{1}{\sqrt{1+v^2}} n \right) = - \frac{1}{(\sqrt{1+v^2})^3} \langle \epsilon(V) \cdot n_t, n_t \rangle + \frac{H_t}{\sqrt{1+v^2}}$$

where $\epsilon(V) = 1/2(DV + DV^*)$ is the deformation tensor.

We consider the situation in which the field V verifies the following property:

$$V(t) = V(t) o p_t \text{ in a neighborhood of } \Gamma_t \tag{4.70}$$

where p_t is the R^N projection mapping onto Γ_t (horizontal projection). Then we get

$$p_t = I_d - b_{\Omega_t(V)} \nabla b_{\Omega_t(V)}$$

and

$$\frac{\partial}{\partial t} p_t = - \frac{\partial}{\partial t} b_{\Omega_t(V)} \nabla b_{\Omega_t(V)} - b_{\Omega_t(V)} \nabla \left(\frac{\partial}{\partial t} b_{\Omega_t(V)} \right)$$

The restriction to the boundary Γ_t leads to the distance $b_{\Omega_t(V)} = 0$, so the expressions simplify as follows (also we shall now denote by b_t that distance function):

$$\frac{\partial}{\partial t} p_t |_{\Gamma_t} = \langle V(t), n_t \rangle n_t$$

and on the boundary $\Gamma_t(V)$ we get

$$DV(t) \cdot n_t = 0$$

Proposition 4.4

Assume that the field V verifies for each t:

$$V(t) = V(t) o p_t$$

Then on the boundary $\Gamma_t(V)$ we have

$$\mathcal{D}iv_{t,x}\nu = -\frac{1}{(\sqrt{1+v^2})^3}\left(\left\langle \frac{\partial}{\partial t}V, n_t\right\rangle - \langle \nabla_{\Gamma_t}(\langle V(t), n_t\rangle), V(t)_{\Gamma_t}\rangle\right) + \frac{1}{\sqrt{1+v^2}}H_t$$

The time–space mean curvature of the lateral boundary Σ is given by

$$\mathcal{H} = -\frac{1}{(\sqrt{1+v^2})^3}\left(\left\langle \frac{\partial}{\partial t}V, n_t\right\rangle - \langle \nabla_{\Gamma_t}(\langle V(t), n_t\rangle), V(t)_{\Gamma_t}\rangle\right) + \frac{1}{\sqrt{1+v^2}}H_t \tag{4.71}$$

The normal component of the horizontal field is given by

$$\langle \tilde{Z}, \nu\rangle = \frac{1}{\sqrt{1+v^2}}\langle Z, n_t\rangle$$

If $f(\Sigma)$ is the restriction to the lateral boundary Σ of a function $F(t,x)$ defined over R^{N+1}, we get the (lateral) shape boundary derivative $f'_\Sigma(\tilde{Z})$ in the direction of the horizontal field $\tilde{Z}$ as follows:

$$f'_\Sigma(\tilde{Z}) = \frac{\partial}{\partial \nu}F.$$

In a general setting we recall that

$$f'_\Sigma(\tilde{Z}) = \left(\frac{d}{ds}(f(\Sigma_s) o \mathcal{T}_s(\tilde{Z})\right)_{s=0} - \langle \nabla_\Sigma f(\Sigma), \tilde{Z}_\Sigma\rangle$$

Notice that the operator ∇_Σ, as a tangential differential operator of the space–time surface Σ is itself a time–space manifold, and we get

$$f'_\Sigma(\tilde{Z}) = \dot{f}(\Sigma, \tilde{Z}) - \frac{vz}{1+v^2}\frac{\partial}{\partial t}f - \langle Z - \frac{z}{1+v^2}n_t, \nabla f\rangle$$

4.2 Lateral boundary derivative

Consider a given function $F \in C^1([0,\tau]\times\bar{D})$. In a first step we assume that F is zero in the neighborhood of $t=\tau$, so that the following derivative of the lateral boundary integral could be considered as a derivative of the integral on the total boundary of the tube (as it will generate no term on the top $t=\tau$ of the tube). Then the usual derivative expressions apply: we consider the derivative of the lateral integral.

$$\Sigma^s = \{(t, T_t(V+sW)(x)) | x \in \partial\Omega_0\}$$

$$\frac{\partial}{\partial s}_{s=0}\left(\int_{\Sigma^s} F\ d\Sigma^s\right) = \int_\Sigma \left(\left(\frac{\partial}{\partial\nu}F + \mathcal{H}_\Sigma F\right)\langle \mathcal{Z}, \nu\rangle_{R^{N+1}}\right) d\Sigma$$

where $\mathcal{H}_\Sigma$ is the mean curvature of the lateral boundary of the tube.

At each point $(t,x)\in\Sigma$ we have:

$$\langle \mathcal{Z}(t,x), \nu(t,x)\rangle_{R^{N+1}} = \frac{1}{\sqrt{1+\langle V(t), n_t\rangle^2}}\langle Z(t), n_t\rangle$$

Moreover

$$\frac{\partial}{\partial\nu}F = \frac{1}{\sqrt{1+(\langle V(t), n_t\rangle)^2}}\left(-\langle V(t), n_t\rangle\frac{\partial}{\partial t}F + \frac{\partial}{\partial n_t}F\right)$$

Then

$$\frac{\partial}{\partial s}_{s=0}\left(\int_{\Sigma^s} F\ d\Sigma^s\right) = \int_\Sigma \left[\frac{1}{\sqrt{1+v^2}}\left(-v\frac{\partial}{\partial t}F + \frac{\partial}{\partial n_t}F\right)\right.$$
$$\left(-\frac{1}{(\sqrt{1+v^2})^3}\left(\left\langle \frac{\partial}{\partial t}V, n_t\right\rangle - \langle\nabla_{\Gamma_t}v, V(t)_{\Gamma_t}\rangle\right)\right.$$
$$\left.\left.+\frac{1}{\sqrt{1+v^2}}H_t\right)F\right]\frac{1}{\sqrt{1+v^2}}\langle Z, n\rangle_{R^N} d\Sigma \tag{4.72}$$

Proposition 4.5

Assume the vector field V in the canonical form $V(t) = V(t)op_t$ in a neighborhood of the lateral boundary Σ, and let $v = \langle V(t), n_t\rangle$ on Γ_t. Then we have

$$\frac{\partial}{\partial s}_{s=0}\left(\int_{\Sigma^s} F\ d\Sigma^s\right) = \int_0^\tau\int_{\Gamma_t}\left[\frac{1}{\sqrt{1+v^2}}\left(-v\frac{\partial}{\partial t}F + \frac{\partial}{\partial n_t}F\right)\right. \tag{4.73}$$
$$+ F\left(-\frac{1}{(\sqrt{1+v^2})^3}\left(\left\langle\frac{\partial}{\partial t}V, n_t\right\rangle - \langle\nabla_{\Gamma_t}v, V(t)_{\Gamma_t}\rangle\right)\right.$$
$$\left.\left.+\ \frac{1}{\sqrt{1+v^2}}H_t\right)\right]\langle Z, n\rangle_{R^N} d\Gamma_t\ dt$$

In the specific case where $F = 1$ all the derivatives of F cancel and we have the derivative of the lateral surface of the tube:

$$\frac{\partial}{\partial s}_{s=0}\left(\int_{\Sigma^s} d\Sigma^s\right) = \int_\Sigma \left[-\frac{1}{(1+v^2)^2}\left(\left\langle\frac{\partial}{\partial t}V, n_t\right\rangle - \langle\nabla_{\Gamma_t} v, V(t)_{\Gamma_t}\rangle\right) + \left(\frac{1}{1+v^2}H_t\right)\right]\langle Z, n\rangle_{R^N} d\Sigma \tag{4.74}$$

The optimality condition for a minimal surface tube is easily obtained via the adjoint problem solution λ as

$$\frac{\partial}{\partial s}_{s=0}\left(\int_{\Sigma^s} d\Sigma^s\right) = \int_\Sigma \lambda\langle W, n_t\rangle d\Sigma \tag{4.75}$$

where λ solves:

$$\begin{aligned}\lambda(\tau) = 0, -\frac{\partial}{\partial t}\lambda - div(\lambda V) \\ = -\frac{1}{(1+v^2)^2}\left(\left\langle\frac{\partial}{\partial t}V, n_t\right\rangle - \langle\nabla_{\Gamma_t} v, V(t)_{\Gamma_t}\rangle\right) + \frac{1}{1+v^2}H_t\end{aligned} \tag{4.76}$$

The optimal condition for a tube with minimal lateral surface would be

$$-\frac{1}{(1+v^2)}\left(\left\langle\frac{\partial}{\partial t}V, n_t\right\rangle - \langle\nabla_{\Gamma_t} v, V(t)_{\Gamma_t}\rangle\right) + H_t = 0 \tag{4.77}$$

5 Heat equation with insulated boundary

We now consider the noncylindrical situation: the boundary Σ is insulated or adiabatic. As the domain moves it is not the usual Neumann boundary condition, but the one described below.

Noncylindrical evolution problems, such as the Navier-Stokes equation for moving boundaries in a fluid (see [5]) is a challenging optimal control issue. In the case of linear problems we deal with easier situations. Nevertheless, a difficult issue is that we need to handle such problems with nonsmooth geometry. The study of the noncylindrical heat equation is an old story. Far from being exhaustive, let us paraphrase the works of P. Acquistapace [21], and more recently in [24]. In these works, the boundary of the moving domain should be smooth enough. The obvious technique was based on the transport into a cylindrical problem which, in terms of an abstract setting, leads to a dynamical system with a nonautonomous operator with a moving domain. Here we revisit that analysis in the scope of optimal control of the moving domain Ω_t. As classically in shape analysis, the control parameter will be the speed vector field $V(t,x)$, whose flow mapping $T_t(V)$ builds the noncylindrical evolution domain $Q_V = \cup_{0\langle t\langle\tau}\{t\} \times \Omega_t$. Let $V \in C^0([0,\tau], C^1(D, R^N))$ with $V \cdot n = 0$ on ∂D. The moving domain is $\Omega_t := T_t(V)(\Omega_0)$ and its characteristic function is $\zeta = \zeta_0 oT_t(V)^{-1}$. We consider the unique solution u to the parabolic problem:

$$\frac{\partial}{\partial t}u - \Delta u = 0 \text{ in } Q_V, \frac{\partial}{\partial n_t}u + vu = 0 \text{ on the moving boundary} \tag{5.78}$$

$\Gamma_t, u(0) = u_0$

This boundary condition cannot be written as $\frac{\partial}{\partial \nu}u = 0$ on the lateral time–space boundary Σ.

5.1 The weak formulation

$\forall \psi \in C^1([0,\tau] \times R^n)$ with $\psi(\tau) = 0$

$$\int_0^\tau \int_{\Omega_t} \left(-u\frac{\partial}{\partial t}\psi + \nabla u \cdot \nabla \psi \right) \, dtdx = \int_{\Omega_0} \psi(0)(x) \, dx \tag{5.79}$$

Introducing $U(t,x) = u(t)oT_t(V)(x)$, the transported solution on the cylindrical domain, we get U as the solution to the parabolic boundary value problem:

$$\begin{aligned} U_t + U(lnJ)_t - J^{-1}div(UJDT^{-1} \cdot V(t)oT_t(V)) \\ -J^{-1}div(JDT^{-1} \cdot (DT^*)^{-1} \cdot \nabla U) = 0 \end{aligned} \tag{5.80}$$

with the boundary condition

$$\langle DT_t(V)^{-1} \cdot (DT_t(V)^{-1})^* \cdot \nabla U, n \rangle + \langle DT_t(V)^{-1} \cdot V(t)oT_t(V), n \rangle U = 0 \tag{5.81}$$

Given some element $U_d \in L^2(D)$ and $\sigma > 0$, we introduce the functional in the following form

$$j(V) = 1/2 \int_0^\tau \int_D \zeta((u - U_d)^2 + |\nabla u|^2) \, dxdt + \sigma/2 \int_0^\tau |V(t)|^2 dt \tag{5.82}$$

We consider the minimization problem

$$Inf\{j(V,\zeta) | (V,\zeta) \in \mathcal{T}_\infty\} \tag{5.83}$$

5.2 The dual problem

$$-\frac{\partial}{\partial t}p - \Delta p = 0 \text{ in } Q_V \tag{5.84}$$

$$p(\tau) = u_\tau \tag{5.85}$$

$$\frac{\partial}{\partial n_t}p = 0 \text{ on the moving boundary} \Gamma_t \tag{5.86}$$

The adjoint weak formulation is as follows:

$$\forall \psi \in C^1([0,\tau] \times R^n) \text{ with } \psi(0) = 0,$$

$$\int_0^\tau \int_{\Omega_t} \left(p \frac{\partial}{\partial t} \psi + \nabla u \cdot \nabla \psi \right) dt dx + \int_0^\tau \int_{\partial \Omega_t} p \psi v \; d\Gamma_t \; dt = \int_{\Omega_\tau} \psi(\tau)(x) \; dx \tag{5.87}$$

Setting $P = poT_t(V)$ we get the same equation as 5.80, but with the final condition at $t = \tau$, and the following boundary condition:

$$\begin{aligned}(DT_t(V)^{-1} \cdot (DT_t(V)^{-1})^* \cdot \nabla P, n\rangle + \langle DT_t(V)^{-1} \cdot V(t) o T_t(V), n\rangle P \\ + \langle V(t) o T_t(V), (DT_t(V)^{-1})^* \cdot n \rangle P = 0\end{aligned} \tag{5.88}$$

For the elements (V, ζ) in $\mathcal{T}^{1,lip}_{\Omega_0}$, of course we have $\zeta = \zeta_{\Omega_0} o T_t(V)^{-1}$; then the element ζ is uniquely associated with the vector field V.

Stability of weak relaxed solution: Given a tube $(V, \zeta) \in \mathcal{T}^{p,\infty}_{\Omega_0}$, we consider the solution u to. The weak parabolic problem:

$$\forall \psi \in C^1([0,\tau] \times R^n) \text{ with } \psi(\tau) = 0, , \mathcal{H} = \nabla u$$

$$\int_0^\tau \int_D \zeta \left(-u \frac{\partial}{\partial t} \psi + \mathcal{H} \cdot \nabla \psi \right) \; dt dx = \int_{\Omega_0} \psi(0)(x) \; dx \tag{5.89}$$

For optimal control purposes we introduce the adjoint problem. The weak relaxed dual problem is the following one:

$$\forall \psi \in C^1([0,\tau] \times R^n) \text{ with } \psi(0) = 0,$$

$$\int_0^\tau \int_D \zeta \left(p \frac{\partial}{\partial t} \psi + \mathcal{H} \cdot \nabla \psi \right) dt dx + \int_0^\tau \int_{\Omega_t} p \psi v \; d\Gamma_t \; dt = \int_{\Omega_\tau} \psi(\tau)(x) \; dx \tag{5.90}$$

Solution stability: In order to get $\mathcal{H}^0 = (\nabla u)^0$ in 5.89, we need the tube to be an open set or at least such property for a.e. concerning the set Ω_t such that $\zeta(t) = \chi_{\Omega_t}$, a.e. A technique for that approach is to replace the perimeter with the density perimeter $P_\gamma \partial \Omega_t$, which we recall below.

Proposition 5.1

Assume that (ζ_v, V_n) is a sequence of smooth tubes, $\zeta_n = \chi_{Q_n}$. We say that Q_n is built by smooth speed vector fields V_n. For each n we classically have a solution u_n. Assume that (V_n, ζ_n) converges to (V, ζ) in $\sigma(L^2, L^2) \times L^1$ topology, with $\zeta_n = \zeta_{\Omega_0} o T_t^{-1}(V_n)$. Also assume that u_n^0 (the extension by zero) weakly converges in $L^2(0, \tau, D)$ to a limit element u as well as $(\nabla u_n)^0$ to some element $\mathcal{H}$. Then $(u, \mathcal{H}) \in L^2(Q)^{N+1}$ and is the solution to problem (5.89) in Q.

The optimal control problem is: Let $V \in C^0([0,\tau], C^1(D, R^N))$ with $V \cdot n = 0$ on ∂D and $\zeta = \zeta_0 o T_t(V)^{-1}$

$$j(V,\zeta) = 1/2 \int_0^\tau \int_D \zeta((u - U_d)^2 + |\nabla u|^2)\ dxdt \tag{5.91}$$

$$j_\sigma(V,\zeta) = j(V,\zeta) + \sigma \left(\int_0^\tau \int_D (||V||^2 + (divV)^2)\ dxdt + ||P_\gamma(\partial\Omega_t)||_{BV(0,\tau)} \right) \tag{5.92}$$

5.3 Optimization problem

We consider the minimization problem:

$$Inf\{j(V,\zeta) | (V,\zeta) \in \mathcal{T}_\infty\} \tag{5.93}$$

5.4 Minimizing sequences

Let (V_n, ζ_n) be a minimizing sequence. The tube Q_n is smooth and the heat equation solution u_n is classically defined. Obviously the null extensions to the cylinder $[0,\tau] \times D$ of both u_n and the gradients ∇u_n are bounded in $L^2([0,\tau] \times D)$. We consider a weakly converging subsequence, still denoted u_n, weakly converging to u and ∇u_n, weakly converging to some vector field Z. On the other hand, as the $BV(0,\tau)$ norm of $P_\gamma(\partial\Omega_t^n)$ is bounded there exists a subsequence, still denoted Ω^n, such that $P_\gamma(\partial\Omega_t^n)$ converges in $L^1(0,\tau)$ to some integrable function f. Then for almost every t, $P_\gamma(\partial\Omega_t^n) \to f(t)$. As a result $a.e.t, P_\gamma(\partial\Omega_t^n) \leq M(t)$ then for almost every time t, the open set Ω_t^n converges to some open set Ω_t, both in H^c and L^p topologies and $meas(\partial\Omega_t) = 0$ and $P_\gamma(\partial\Omega_t) \leq liminf_{n-\rangle\infty} P_\gamma(\partial\Omega_t^n) = f(t)$.

Let $\phi \in \mathcal{D}([0,\tau] \times D)$ such that a.e. t, $\phi(t) \in \mathcal{D}(\Omega_t)$. For $n \geq N_t$ we have $\phi(t) \in \mathcal{D}(\Omega_t^n)$ so that, a.e., we have:

$$\int_{\Omega_t^n} \langle \nabla u_n(t), \phi(t) \rangle\ dx = -\int_{\Omega_t^n} u_n(t) div\phi(t)\ dx$$

Obviously

$$\int_D \langle \nabla u_n(t), \phi(t) \rangle dx = \int_{\Omega_t^n} \langle \nabla u_n(t), \phi(t) \rangle\ dx$$

and the same concerning u, so that in the limit we get $Z = \nabla u$.

References

[1] M.C. Delfour and J.-P. Zolésio. Oriented distance function and its evolution equation for initial sets with thin boundary. *SIAM J. Control Optim.* 42 (2004), no. 6, 2286–2304.

[2] M. Cuer and J.-P. Zolésio. Control of singular problem via differentiation of a min-max. *Systems Control Lett.* 11 (1988), no. 2, 151–158.

[3] D. Bucur and J.-P. Zolésio. Free boundary problems and density perimeter. *J. Differential Equations* 126 (1996), 224–243.

[4] D. Bucur and J.-P. Zolésio. Boundary optimization under pseudo curvature constraint. *Annali dela Scuola Normale Superiore di Pisa*, IV, XXIII (1996), 48.4, 681–699.

[5] R. Dziri and J.-P. Zolésio. Dynamical shape control in non-cylindrical Navier-Stokes equations. *J. Convex Analysis* 6 (1999) (2), 293–318.

[6] N. Gomez and J.-P. Zolésio. Shape sensitivity and large deformation of the domain for Norton-Hoff flow. In G. Leugering, ed., *Proceedings of the IFIP-WG7.2 Conference, Chemnitz*, volume 133 of Int. Series of Num. Math., 167–176, 1999.

[7] J.-P. Zolésio. Identification de domaine par déformations. *Thèse de doctorat d'état*, Université de Nice, 1979.

[8] J.-P. Zolésio. In *Optimization of Distributed Parameter Structures*, vol. II (E. Haug and J. Céa, eds.), Adv. Study Inst. Ser. E: Appl. Sci., 50, Sijthoff and Nordhoff, Alphen aan den Rijn, 1981: i) The speed method for shape optimization, 1089–1151; ii) Domain variational formulation for free boundary problems, 1152–1194; iii) Semiderivative of repeated eigenvalues, 1457–1473.

[9] J. Sokolowski and J.-P. Zolésio. *Introduction to Shape Optimization*, Springer-Verlag, Heidelberg, 1991.

[10] B. Kawohl, O. Pironneau, L. Tartar, and J.-P. Zolésio. *Optimal Shape Design*, Springer-Verlag, Heidelberg, 1998.

[11] J.-P. Zolésio. Variational principle in the Euler flow. In G. Leugering, ed., *Proceedings of the IFIP-WG7.2 Conference, Chemnitz*, volume 133 of Int. Series of Num. Math., 1999.

[12] J.-P. Zolésio. Shape differential with non smooth field. In *Computational Methods for Optimal Design and Control*, J. Borggard, J. Burns, E. Cliff, and S. Schreck, eds., vol. 24 of Progress in Systems and Control Theory, 426–460, Birkhäuser, Washington, D.C., 1998.

[13] J.-P. Zolésio. Weak set evolution and variational applications. In *Shape Optimization and Optimal Design*, J.-P. Zolésio, ed., Lecture Notes in Pure and Applied Mathematics, vol. 216, 415–442, Marcel Dekker, New York, 2001.

[14] M.C. Delfour and J.-P. Zolésio. Structure of shape derivatives for non smooth domains, *Journal of Functional Analysis* (1992), 104.

[15] C. Cannarsa, G. Da Prato, and J.-P. Zolésio. The damped wave equation in a moving domain, *Journal of Differential Equations* 85 (1990), 1–16.

[16] M.C. Delfour and J.-P. Zolésio. Shape analysis via oriented distance functions, *J. Funct. Anal.* 123 (1994), 1–56.

[17] R. Dziri and J.-P. Zolésio. Dynamical shape control in non-cylindrical hydrodynamics, *Inverse Problems* 15 (1999), no. 1, 113–122.

[18] M.C. Delfour and J.-P. Zolésio. Shape sensitivity analysis via min max differentiability, *SIAM J. Control Optim.* 26 (1988), no. 4, 834–862.

[19] G. Da Prato and J.-P. Zolésio. Dynamical programming for non cylindrical parabolic equations, *Sys. Control Lett.* 11 (1988), 121–139.

[20] G. Da Prato and J.-P. Zolésio. *Existence and control for wave equation in moving domain*, New York, Lecture Notes in Control and Information Science, 144, Springer-Verlag, New York, 1988.

[21] P. Acquistapace. Boundary control for non-autonomous parabolic equations in non-cylindrical domains. In *Boundary Control and Variation* (Sophia Antipolis, 1992),

Lecture Notes in Pure and Applied Mathematics, 163, Marcel Dekker, New York, 1994, 1–12.

[22] L. Ambrosio. Lecture notes on optimal transport problems. Mathematical aspects of evolving interfaces (Funchal, 2000), Lecture Notes in Mathematics, 1812, Springer, Berlin, 2003, 1–52.

[23] S. Boisgérault and J.-P. Zolésio. Shape derivative of sharp functionals governed by Navier-Stokes flow. In O. John, K. Najzar, W. Jäger, J. Necas, and J. Stará, eds., *Partial Differential Equations, Theory and Numerical Simulation*, CRC Research Notes in Mathematics, Chapman & Hall, Boca Raton, FL, 2000, 49–63.

[24] K. Burdzy, Z. Chen, and J. Sylvester. The heat equation and reflected Brownian motion in time-dependent domains. *Ann. Probab.* 32 (2004), no. 1B, 775–804.

[25] M. Delfour and J.-P. Zolésio. *Shapes and Geometries*, Advances in Design and Control, 4, SIAM, Philadelphia, 2001.

[26] J.-P. Zolésio. Set weak evolution and transverse field, *Variational Applications and Shape Differential Equation*, INRIA report RR-464, 2002 (http://www-sop.inria.fr/rapports/sophia/RR-464).

[27] M. Moubachir and J.-P. Zolésio, *Moving Shape Analysis and Control: Application to Fluid Structure Interaction*, Pure and Applied Mathematics, CRC Press, Boca Raton, FL, 2006.

[28] M.C. Delfour and J.-P. Zolésio. Structure of shape derivatives for nonsmooth domains. *Journal of Functional Analysis*, 104, 1992.

[29] M.C. Delfour and J.-P. Zolésio. Shape analysis via oriented distance functions. *Journal of Functional Analysis*, 123, 1994.

[30] F.R. Desaint and J.-P. Zolésio. Manifold derivative in the Laplace-Beltrami equation. *Journal of Functional Analysis* 151 (1997), no. 1, 234, 269.

[31] J.-P. Zolésio. Introduction to shape optimization and free boundary problems. In M.C. Delfour, ed., *Shape Optimization and Free Boundaries*, vol. 380 of NATO ASI, Series C: Mathematical and Physical Sciences, Montreal, 1992, 397, 457.

[32] J.-P. Zolésio. Shape topology by tube geodesic. In *Information Processing: Recent Mathematical Advances in Optimization and Control.* Presses de l'Ecole des Mines de Paris, 2004, 185–204.

CHAPTER 2

Numerical simulation of pattern formation in a rotating suspension of non-Brownian settling particles

Tsorng-Whay Pan
Department of Mathematics, University of Houston, Houston, Texas

Roland Glowinski
Department of Mathematics, University of Houston, Houston, Texas
and Université P. et M. Curie, Paris, France

1 Introduction

Nonequilibrium systems often organize into interesting spatiotemporal structures or patterns. Examples include the patterns in pure fluid flow systems, such as Rayleigh-Bénard convection in a vertical temperature gradient, the Taylor-Couette flow between two concentric rotating cylinders, and those concerning particulate flow systems in a partially or fully filled horizontal rotating cylinder, such as well-defined periodic clusters, normal to the axis of rotation. Particulate flows exhibiting axial clusters along the horizontal axis in a partially filled horizontal rotating cylinder were observed in [1–3], which are in part attributed to the presence of the free surface caused by the partial filling of the cylinder. Similar cluster and other pattern formations were also found for a settling suspension of uniform non-Brownian particles in a *fully* filled horizontal rotating cylinder [4–7]. In [4] Lipson used a horizontal rotating cylinder filled with oversaturated solution to grow crystal without any interaction with a substrate, and found that crystals accumulate in well-defined periodic clusters, normal to the axis of rotation. Lipson and Seiden have suggested that it could be caused by the interaction between particles and the flow in the tube [5]. A variety of patterns and structures were encountered for a settling suspension of uniform non-Brownian particles in a fully filled horizontal rotating cylinder [7], in which Matson et al. explained that those patterns and dynamics result from the interplay between the viscous drag and the gravitational and centrifugal forces.

In this article, we would like to extend a distributed Lagrange multiplier–based fictitious domain (also called domain embedding) method with operator splitting [8,9,11–13] to simulate particulate flows in a completely filled horizontal rotating cylinder. We have successfully obtained well-defined periodic clusters of 160 balls in a horizontal rotating cylinder. At least from flow field projection on planes passing through the central axis of the rotation cylinder, we can conclude that the velocity field of fluid does not have a strong influence on the formation of particle clusters. The content of this article is as follows: In Section 2, we provide a mathematical model for incompressible viscous fluid–rigid body interaction, including a distributed Lagrange multiplier–based fictitious domain variational formulation. In Section 3, we discuss the time discretization by operator splitting, starting from the above fictitious domain formulation, and then the finite element approximation of the problem. The results of numerical experiments are presented and discussed in Section 4.

2 A model problem and its fictitious domain formulation

To perform the *direct numerical simulation* of the interaction between rigid bodies and fluid, we have developed a methodology that combines a distributed Lagrange multiplier–based fictitious domain (also called domain embedding) method with operator splitting [8,9,11–13]. This approach (or closely related ones derived from it) has been used by other investigators (e.g., [10,14–16]). First we are going to recall the basic ideas of the above methodology by considering the motion of a single particle B of general shape in a Newtonian viscous incompressible fluid (of density ρ_f and viscosity μ_f) contained in a horizontal truncated cylinder $\mathbf{C}$ under the effect of gravity. For the situation depicted in Figure 2.1, the flow is modeled by the Navier-Stokes equations, namely (with obvious notation),

$$\rho_f \left[\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u}\right] - \mu_f \Delta \mathbf{u} + \nabla p = \rho_f\, \mathbf{g} \quad in \quad \{(\mathbf{x},t)|\mathbf{x}\in \mathbf{C}\setminus\overline{B(t)},$$

$$t \in (0,T)\}, \tag{2.1}$$

$$\nabla\cdot\mathbf{u}(t) = 0 \quad in \quad \mathbf{C}\setminus\overline{B(t)},\ 0<t<T, \tag{2.2}$$

$$\mathbf{u}(0) = \mathbf{u}_0(\mathbf{x}), \quad (with\ \nabla\cdot\mathbf{u}_0 = 0), \tag{2.3}$$

$$\mathbf{u} = \mathbf{g}_0 \quad on \quad \Gamma_0\times(0,T), \left(with \int_{\Gamma_0} \mathbf{g}_0\cdot\mathbf{n}\, d\Gamma = 0\right), \tag{2.4}$$

where Γ_0 is the boundary of the truncated cylinder $\mathbf{C}$, $\mathbf{g}$ denotes gravity, and $\mathbf{n}$ is the unit normal vector pointing outward to the flow region. We assume a *no-slip condition* on $\gamma(=\partial B)$. The motion of the rigid body B

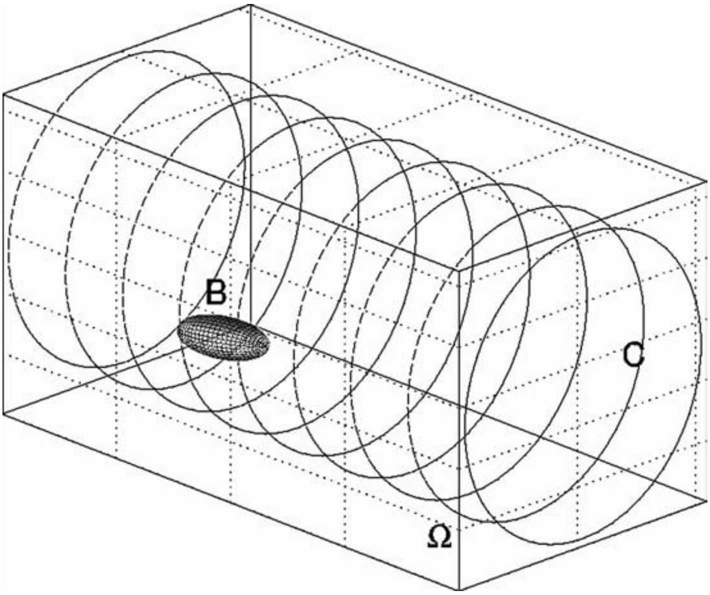


FIGURE 2.1 The flow region with an ellipsoid in a horizontal cylinder.

satisfies the Euler-Newton equations, namely

$$\mathbf{v}(\mathbf{x},t) = \mathbf{V}(t) + \overrightarrow{\omega}(t) \times \overrightarrow{\mathbf{G}(t)\mathbf{x}},\ \forall \mathbf{x} \in \overline{B(t)},\ \forall t \in (0,T), \tag{2.5}$$

$$\frac{d\mathbf{G}}{dt} = \mathbf{V}, \tag{2.6}$$

$$M_p \frac{d\mathbf{V}}{dt} = M_p\,\mathbf{g} + \mathbf{F}_H + \mathbf{F}^r, \tag{2.7}$$

$$\frac{d(\mathbf{I}_p\,\omega)}{dt} = \mathbf{T}_H + \overrightarrow{\mathbf{G}\mathbf{x}_\mathbf{r}} \times \mathbf{F}^r, \tag{2.8}$$

with the resultant and torque of the hydrodynamical forces given by, respectively,

$$\mathbf{F}_H = -\int_\gamma \sigma \mathbf{n}\, d\gamma, \quad \mathbf{T}_H = -\int_\gamma \overrightarrow{\mathbf{G}\mathbf{x}} \times \sigma \mathbf{n}\, d\gamma \tag{2.9}$$

with $\sigma = \mu_f(\nabla\mathbf{u} + \nabla\mathbf{u}^t) - p\mathbf{I}$. Relations (2.1) through (2.9) are completed by the following initial conditions

$$\mathbf{G}(0) = \mathbf{G}_0, \quad \mathbf{V}(0) = \mathbf{V}_0, \quad \omega(0) = \omega_0, \quad B(0) = B_0. \tag{2.10}$$

Above, M_p, $\mathbf{I}_p$, $\mathbf{G}$, $\mathbf{V}$ and ω are the mass, inertia, center of mass, velocity of the center of mass, and angular velocity of the rigid body B, respectively. In (2.8) we found it preferable to deal with the *kinematic angular momentum*

$\mathbf{I}_p\,\omega$ making the formulation more conservative. In order to avoid the overlapping of the rigid bodies and rigid body-wall penetration, which can happen in the numerical simulation, we have introduced an artificial short-range repulsion force $\mathbf{F}^r$ in (2.7), which becomes active when the shortest distance between two (convex) rigid bodies or between a (convex) rigid body and wall is less than a prechosen distance (for more details, see, e.g., [8] and [9]; see also [17] for another approach) and then a torque in (2.8) acts on the point $\mathbf{x}_r$ where $\mathbf{F}^r$ applies on B. For nonconvex particles, we can apply a similar approach to activate the short-range repulsion force $\mathbf{F}^r$.

To solve the system (2.1) through (2.10) we can use, for example, *Arbitrary Lagrange-Euler* (*ALE*) methods as in [18–20], or *fictitious domain methods*, which allow the flow calculation on a fixed grid, as in [8,9,11–13]. The fictitious domain methods that we advocate share some common features with the *immersed boundary method* of Ch. Peskin (see, e.g., refs. [21–23]), but also some significant differences in the sense that we take systematic advantage of *distributed Lagrange multipliers* to force the rigid body motion inside the particle. As with the methods in [21–23], our approach takes advantage of the fact that the flow can be computed on a grid that does not have to vary in time, a substantial simplification indeed. In order to take full advantage of the fictitious domain approach we will embed the truncated cylinder $\mathbf{C}$ in a rectangular parallelepiped (denoted by Ω) with a square cross-section whose edge length is slightly larger than the diameter of the cylinder $\mathbf{C}$ as shown in Figure 2.1. The region outside $\mathbf{C}$ is denoted by $\mathbf{A} = \Omega \setminus \overline{\mathbf{C}}$, and the boundary of Ω is denoted by Γ. Also we assume that $\mathbf{g}_0$ defined on Γ_0 is nothing but the velocity field on the surface of a horizontal rotating cylinder; hence we can easily extend it on $\overline{\mathbf{A}}$ according to the angular velocity of the cylinder. For extended value on Γ, we still use $\mathbf{g}_0$ in the following.

The principle of fictitious domain methods is simple. It consists of

- Filling the rigid bodies with a fluid having the same density and viscosity as the surrounding one.
- Compensating the above step by introducing, in some sense, an *antiparticle* of mass $(-1)M_p\,\dfrac{\rho_f}{\rho_s}$ and inertia $(-1)\mathbf{I}_p\,\dfrac{\rho_f}{\rho_s}$, taking into account the fact that any rigid body motion $\mathbf{v}(\mathbf{x},t)$ verifies $\nabla\cdot\mathbf{v} = 0$ and $\mathbf{D}(\mathbf{v}) = \mathbf{0}$ (ρ_s : rigid body density).
- Finally, imposing the rigid body velocity on $\overline{B(t)}$, namely,

$$\mathbf{v}(\mathbf{x},t) = \mathbf{V}(t) + \overrightarrow{\omega(t)} \times \overrightarrow{\mathbf{G}(t)\mathbf{x}}, \quad \forall \mathbf{x} \in \overline{B(t)}, \quad \forall t \in (0,T), \tag{2.11}$$

  via a Lagrange multiplier λ supported by $\overline{B(t)}$. Vector λ forces rigidity in $B(t)$ in the same way that ∇p forces $\nabla\cdot\mathbf{v} = 0$ for incompressible fluids.

We obtain then an equivalent formulation of (2.1) through (2.10) defined on the whole domain Ω, namely

For a.e. $t > 0$, *find* $\mathbf{u}(t) \in \mathbf{W}_{\mathbf{g}_0}(t), p(t) \in L^2_0(\Omega), \mathbf{V}(t) \in \mathbb{R}^3, \mathbf{G}(t) \in \mathbb{R}^3$, $\omega(t) \in \mathbb{R}^3,\ \lambda(t) \in \Lambda(t),\ \lambda_A \in \Lambda_A$ *such that*

$$\left\{\begin{aligned}
&\rho_f \int_\Omega \left[\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u}\right]\cdot \mathbf{v}\, d\mathbf{x} - \int_\Omega p\nabla\cdot\mathbf{v}\, d\mathbf{x}\\
&+\mu_f \int_\Omega \nabla\mathbf{u}{:}\nabla\mathbf{v}\, d\mathbf{x} - \langle \lambda, \mathbf{v} - \mathbf{Y} - \theta \times \mathbf{Gx}\rangle_{\Lambda(t)}\\
&-\langle \lambda_A, \mathbf{v}\rangle_{\Lambda_A} + \left(1 - \frac{\rho_f}{\rho_s}\right)\left[M_p \frac{d\mathbf{V}}{dt}\cdot\mathbf{Y} + \frac{d(\mathbf{I}_p\,\omega)}{dt}\cdot\theta\right]\\
&-\mathbf{F}^r\cdot\mathbf{Y} - \overrightarrow{\mathbf{Gx_r}} \times \mathbf{F}^r\cdot\theta\\
&= \left(1 - \frac{\rho_f}{\rho_s}\right) M_p \mathbf{g}\cdot\mathbf{Y} + \rho_f \int_\Omega \mathbf{g}\cdot\mathbf{v}\, d\mathbf{x},\ \forall \mathbf{v} \in \mathbf{W}_0,\ \forall \mathbf{Y} \in \mathbb{R}^3,\ \forall\theta \in \mathbb{R}^3,
\end{aligned}\right. \tag{2.12}$$

$$\int_\Omega q\nabla\cdot\mathbf{u}(t) d\mathbf{x} = 0, \quad \forall q \in L^2(\Omega), \tag{2.13}$$

$$\langle \mu, \mathbf{u}(t) - \mathbf{V}(t) - \omega(t) \times \mathbf{G}(t)\mathbf{x}\rangle_{\Lambda(t)} = 0, \quad \forall \mu \in \Lambda(t), \tag{2.14}$$

$$\langle \mu_A, \mathbf{u}(t) - \mathbf{g}_0(t)\rangle_{\Lambda_A} = 0, \quad \forall \mu_A \in \Lambda_A, \tag{2.15}$$

$$\frac{d\mathbf{G}}{dt} = \mathbf{V}, \tag{2.16}$$

$$\mathbf{V}(0) = \mathbf{V}_0,\ \omega(0) = \omega_0,\ \mathbf{G}(0) = \mathbf{G}_0,\ B(0) = B_0, \tag{2.17}$$

$$\mathbf{u}(\mathbf{x}, 0) = \tilde{\mathbf{u}}_0(\mathbf{x}) = \left\{\begin{array}{l}\mathbf{u}_0(\mathbf{x}),\ \forall \mathbf{x} \in \mathbf{C} \setminus B(0),\\ \mathbf{V}_0 + \omega_0 \times \mathbf{G}_0\mathbf{x},\ \forall \mathbf{x} \in \overline{B(0)},\\ \mathbf{g}_0(0),\ \forall \mathbf{x} \in \overline{A},\end{array}\right. \tag{2.18}$$

with the following functional spaces

$$\begin{aligned}
&\mathbf{W} = (H^1(\Omega))^3,\ \mathbf{W}_0 = (H^1_0(\Omega))^3,\\
&\mathbf{W}_{\mathbf{g}_0}(t) = \{\mathbf{v} | \mathbf{v} \in \mathbf{W},\ \mathbf{v} = \mathbf{g}_0(t) \quad on\ \Gamma\},\\
&L^2_0(\Omega) = \left\{ q | q \in L^2(\Omega),\ \int_\Omega q\, d\mathbf{x} = 0 \right\},\\
&\Lambda(t) = (H^1(B(t)))^3,\ \Lambda_A = \{\mu | \mu \in (H^1(A))^3\}.
\end{aligned}$$

In (2.12), (2.14) and (2.15), $\langle\cdot,\cdot\rangle_{\Lambda(t)}$ and $\langle\cdot,\cdot\rangle_{\Lambda_A}$ are inner products on $\Lambda(t)$ and Λ_A, respectively. Various examples of this are given in [9] and [24] (Chapter 8). The velocity field inside $\mathbf{A}$ is enforced in (2.12) and (2.15) via the Lagrange multiplier λ_A supported by $\overline{\mathbf{A}}$.

Remark
The second gravity term in the righthand side of (2.12) can be combined with the pressure. Hence in the following we will not use this term anymore.

In (2.12) through (2.18), only the center of mass, the translation velocity of the center of mass, and the angular velocity of the particle are considered. Knowing these two velocities and the center of mass of the particle, one is able to translate and rotate the particle of general shape in space by tracking two extra points, $\mathbf{x}_1$ and $\mathbf{x}_2$, in each particle, which follow the rigid body motion

$$\frac{d\mathbf{x}_i}{dt} = \mathbf{V}(t) + \vec{\omega}(t) \times \overrightarrow{\mathbf{G}(t)\mathbf{x}_i}, \quad \mathbf{x}_i(0) = \mathbf{x}_{i,0}, \ i = 1, 2. \tag{2.19}$$

In practice we shall track two orthogonal normalized vectors rigidly attached to the body B and originating from the center of mass $\mathbf{G}$.

Remark
In the simulations reported in this article, we have used spherical balls of constant density, so Equation (2.8) becomes

$$\mathbf{I}_p \frac{d\omega}{dt} = \mathbf{T}_H \tag{2.20}$$

and Equation (2.12) becomes

$$\begin{cases} \rho_f \displaystyle\int_\Omega \left[\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u}\right]\cdot \mathbf{v}\, d\mathbf{x} - \int_\Omega p\nabla\cdot\mathbf{v}\, d\mathbf{x} \\ +\mu_f \displaystyle\int_\Omega \nabla\mathbf{u} : \nabla\mathbf{v}\, d\mathbf{x} - \langle\lambda, \mathbf{v} - \mathbf{Y} - \theta \times \mathbf{G}\mathbf{x}\rangle_{\Lambda(t)} \\ -\langle\lambda_A, \mathbf{v}\rangle_{\Lambda_A} + \left(1 - \dfrac{\rho_f}{\rho_s}\right)\left[M_p \dfrac{d\mathbf{V}}{dt}\cdot\mathbf{Y} + \mathbf{I}_p \dfrac{d\omega}{dt}\cdot\theta\right] - \mathbf{F}^r\cdot\mathbf{Y} \\ = \left(1 - \dfrac{\rho_f}{\rho_s}\right) M_p\ \mathbf{g}\cdot\mathbf{Y} + \rho_f \displaystyle\int_\Omega \mathbf{g}\cdot\mathbf{v}\, d\mathbf{x}, \\ \forall \mathbf{v} \in \mathbf{W}_0,\ \forall \mathbf{Y} \in \mathbb{R}^3,\ \forall\theta \in \mathbb{R}^3. \end{cases} \tag{2.21}$$

Also, we do not need to track the motion of the two extra points described in Equation (2.19).

3 Time and space discretization

3.1 Lie's scheme: a first-order operator-splitting scheme

Many operator-splitting schemes can be applied to problems (2.12) through (2.19). One of the advantages of operator–splitting schemes is that we can decouple difficulties such as (i) the incompressibility condition, (ii) the nonlinear advection term, and (iii) the rigid body motion, so that each one of them can be handled separately and, in principle, optimally. Let $\triangle t$ be a time discretization step and $t^{n+s} = (n+s)\triangle t$. Lie's scheme is a *first-order* operator-splitting scheme [25], which, when applied to problems (2.12) through (2.19), yields:

$$\mathbf{u}^0 = \tilde{\mathbf{u}}_0, \mathbf{G}^0 = \mathbf{G}_0, \quad \mathbf{V}^0 = \mathbf{V}_0, \quad \omega^0 = \omega_0 \textit{ given}; \tag{3.22}$$

for $n \geq 0, \mathbf{u}^n (\simeq \mathbf{u}(t^n))$, $\mathbf{G}^n$, $\mathbf{V}^n$ and ω^n *being known, we first compute* $\mathbf{u}^{n+\frac{1}{6}}$, $p^{n+\frac{1}{6}}$ *via the solution of*

$$\begin{cases} \rho_f \displaystyle\int_\Omega \frac{\partial \mathbf{u}}{\partial t} \cdot \mathbf{v}\, d\mathbf{x} - \int_\Omega p \nabla \cdot \mathbf{v}\, d\mathbf{x} = 0, \quad \forall \mathbf{v} \in \mathbf{W}_0, \\ \textit{a.e. on } (t^n, t^{n+1}), \\ \displaystyle\int_\Omega q \nabla \cdot \mathbf{u}\, d\mathbf{x} = 0, \quad \forall q \in L^2(\Omega), \\ \mathbf{u}(t^n) = \mathbf{u}^n, \\ \mathbf{u}(t) \in \mathbf{W},\ \mathbf{u}(t) = \mathbf{g}(t^{n+1}) \quad \textit{on} \quad \Gamma \times (t^n,\ t^{n+1}), \\ p(t) \in L_0^2(\Omega), \end{cases} \tag{3.23}$$

and set $\mathbf{u}^{n+\frac{1}{6}} = \mathbf{u}(t^{n+1}),\ p^{n+\frac{1}{6}} = p(t^{n+1})$.

Next, compute $\mathbf{u}^{n+\frac{2}{6}}$ *via the solution of*

$$\begin{cases} \displaystyle\int_\Omega \frac{\partial \mathbf{u}}{\partial t} \cdot \mathbf{v}\, d\mathbf{x} + \int_\Omega (\mathbf{u}^{n+\frac{1}{6}} \cdot \nabla)\mathbf{u} \cdot \mathbf{v}\, d\mathbf{x} = 0, \quad \forall \mathbf{v} \in \mathbf{W}_0^{n+1,-}, \\ \textit{a.e. on } (t^n, t^{n+1}), \\ \mathbf{u}(t^n) = \mathbf{u}^{n+\frac{1}{6}}, \\ \mathbf{u}(t) \in \mathbf{W}, \\ \mathbf{u}(t) = \mathbf{g}(t^{n+1}) \quad \textit{on } \Gamma_-^{n+1} \times (t^n, t^{n+1}), \end{cases} \tag{3.24}$$

and set $\mathbf{u}^{n+\frac{2}{6}} = \mathbf{u}(t^{n+1})$.

Then, compute $\mathbf{u}^{n+\frac{3}{6}}$ *via the solution of*

$$\begin{cases} \rho_f \displaystyle\int_\Omega \frac{\partial \mathbf{u}}{\partial t} \cdot \mathbf{v}\, d\mathbf{x} + \alpha \mu_f \int_\Omega \nabla \mathbf{u} : \nabla \mathbf{v}\, d\mathbf{x} \\ \quad = \langle \lambda_A, \mathbf{v} \rangle_{\Lambda_A},\ \forall \mathbf{v} \in \mathbf{W}_0, \textit{a.e. on } (t^n, t^{n+1}), \\ \langle \mu_A, \mathbf{u} - \mathbf{g}(t^{n+1}) \rangle_{\Lambda_A} = 0,\ \forall \mu_A \in \Lambda_A, \\ \mathbf{u}(t^n) = \mathbf{u}^{n+\frac{2}{6}},\ \mathbf{u}(t) \in \mathbf{W}, \end{cases} \tag{3.25}$$

and set $\mathbf{u}^{n+\frac{3}{6}} = \mathbf{u}(t^{n+1})$.

Now predict the motion of the center of mass of the particle via

$$\frac{d\mathbf{G}}{dt} = \mathbf{V}(t)/2, \tag{3.26}$$

$$\left(1 - \frac{\rho_f}{\rho_s}\right) M_p \frac{d\mathbf{V}}{dt} = \mathbf{F}_r/2, \tag{3.27}$$

$$\mathbf{G}(t^n) = \mathbf{G}^n,\ \mathbf{V}(t^n) = \mathbf{V}^n, \tag{3.28}$$

for $t^n < t < t^{n+1}$. *Then set* $\mathbf{G}^{n+\frac{4}{6}} = \mathbf{G}(t^{n+1})$ *and* $\mathbf{V}^{n+\frac{4}{6}} = \mathbf{V}(t^{n+1})$.

Using $\mathbf{G}^{n+\frac{4}{6}}$ *obtained in the above step, we enforce the rigid body motion in the region* $B^{n+\frac{4}{6}}$ *occupied by the particle*

$$\begin{cases} \rho_f \displaystyle\int_\Omega \frac{\partial \mathbf{u}}{\partial t} \cdot \mathbf{v}\, d\mathbf{x} + \beta\mu_f \int_\Omega \nabla\mathbf{u} : \nabla\mathbf{v}\, d\mathbf{x}\ + \left(1 - \frac{\rho_f}{\rho_s}\right) M_p \frac{d\mathbf{V}}{dt} \cdot \mathbf{Y} \\ \quad + \left(1 - \frac{\rho_f}{\rho_s}\right) \mathbf{I}_p \frac{d\omega}{dt} \cdot \theta = \left(1 - \frac{\rho_f}{\rho_s}\right) M_p \mathbf{g} \cdot \mathbf{Y} \\ +\langle \lambda,\ \mathbf{v} - \mathbf{Y} - \theta \times \overrightarrow{\mathbf{G}^{n+\frac{4}{6}}\mathbf{x}} \rangle_{\Lambda^{n+\frac{4}{6}}}, \\ \forall \mathbf{v} \in \mathbf{W}_0,\ \mathbf{Y} \in \mathbb{R}^3, \theta \in \mathbb{R}^3,\ \textit{a.e. on}\ (t^n, t^{n+1}), \\ \mathbf{u}(t^n) = \mathbf{u}^{n+\frac{3}{6}},\ \mathbf{V}(t^n) = \mathbf{V}^{n+\frac{4}{6}},\ \omega(t^n) = \omega^n, \\ \mathbf{u} \in \mathbf{W}, \mathbf{u}(t) = \mathbf{g}_0(t^{n+1}) \quad \textit{on}\ \Gamma \times (t^n, t^{n+1}), \\ \lambda \in \Lambda^{n+\frac{4}{6}},\ \mathbf{V} \in \mathbb{R}^3,\ \omega \in \mathbb{R}^3, \end{cases} \tag{3.29}$$

$$\langle \mu, \mathbf{u} - \mathbf{V} - \omega \times \overrightarrow{\mathbf{G}^{n+\frac{4}{6}}\mathbf{x}} \rangle_{\Lambda^{n+\frac{4}{6}}} = 0, \quad \forall \mu \in \Lambda^{n+\frac{4}{6}}, \tag{3.30}$$

and set $\mathbf{u}^{n+1} = \mathbf{u}(t^{n+1}), \mathbf{V}^{n+\frac{5}{6}} = \mathbf{V}(t^{n+1}), \omega^{n+1} = \omega(t^{n+1})$.

Correct the motion of the center of mass of the particle via

$$\frac{d\mathbf{G}}{dt} = \mathbf{V}(t)/2, \tag{3.31}$$

$$\left(1 - \frac{\rho_f}{\rho_s}\right) M_p \frac{d\mathbf{V}}{dt} = \mathbf{F}_r/2, \tag{3.32}$$

$$\mathbf{G}(t^n) = \mathbf{G}^{n+\frac{4}{6}}, \quad \mathbf{V}(t^n) = \mathbf{V}^{n+\frac{5}{6}}, \tag{3.33}$$

for $t^n < t < t^{n+1}$. *Then set* $\mathbf{G}^{n+1} = \mathbf{G}(t^{n+1})$ *and* $\mathbf{V}^{n+1} = \mathbf{V}(t^{n+1})$.

In (3.22) through (3.33), $\Gamma_-^{n+1} = \{\mathbf{x} | \mathbf{x} \in \Gamma, \mathbf{g}_0(t^{n+1})(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) < 0\}$ and $\mathbf{W}_0^{n+1,-} = \{\mathbf{v} | \mathbf{v} \in \mathbf{W}, \mathbf{v} = \mathbf{0}\ \textit{on}\ \Gamma_-^{n+1}\}$, $\Lambda^{n+\frac{4}{6}} = (H^1(B^{n+\frac{4}{6}}))^3$, $B^{n+\frac{4}{6}}$ is the region occupied by the particle B according to $\mathbf{G}^{n+\frac{4}{6}}$, and $\alpha + \beta = 1$. In the numerical simulation, we usually choose $\alpha = 1$ and $\beta = 0$.

3.2 Space discretization

We assume that $\Omega \subset \mathbb{R}^3$ and is a rectangular parallelepiped. Concerning the *finite element approximation* of problems (2.12) through (2.19), we have

$$\mathbf{W}_h = \{\mathbf{v}_h | \mathbf{v}_h \in (C^0(\overline{\Omega}))^3, \quad \mathbf{v}_h|_T \in (P_1)^3, \quad \forall T \in \mathcal{T}_h, \}, \tag{3.34}$$

$$\mathbf{W}_{0h} = \{\mathbf{v}_h | \mathbf{v}_h \in \mathbf{W}_h, \quad \mathbf{v}_h = \mathbf{0} \quad on\ \Gamma\}, \tag{3.35}$$

$$L^2_h = \{q_h | q_h \in C^0(\overline{\Omega}), \quad q_h|_T \in P_1, \quad \forall T \in \mathcal{T}_{2h}\}, \tag{3.36}$$

$$L^2_{0h} = \left\{ q_h | q_h \in L^2_h, \int_\Omega q_h \, d\mathbf{x} = 0 \right\} \tag{3.37}$$

where $\mathcal{T}_h$ is a tetrahedrization of Ω, $\mathcal{T}_{2h}$ is twice coarser than $\mathcal{T}_h$, and P_1 is the space of the polynomials in three variables of degree ≤ 1. A finite dimensional space approximating $\Lambda(t)$ is as follows: let $\{\xi_i\}_{i=1}^N$ be a set of points from $\overline{B(t)}$ that cover $\overline{B(t)}$ (uniformly, for example); we define then

$$\Lambda_h(t) = \{\mu_h | \mu_h = \sum\nolimits_{i=1}^N \mu_i \delta(\mathbf{x} - \xi_i), \ \mu_i \in \mathbb{R}^3, \ \forall i = 1, \ldots, N\}, \tag{3.38}$$

where $\delta(\cdot)$ is the Dirac measure at $\mathbf{x} = \mathbf{0}$. Then we shall use $\langle \cdot, \cdot \rangle_{\Lambda_h(t)}$ defined by

$$\langle \mu_h, \mathbf{v}_h \rangle_{\Lambda_h(t)} = \sum\nolimits_{i=1}^N \mu_i \cdot \mathbf{v}_h(\xi_i), \forall \mu_h \in \Lambda_h(t), \mathbf{v}_h \in \mathbf{W}_h. \tag{3.39}$$

A typical choice of points for defining (3.38) is to take the grid points of the velocity mesh internal to the particle B and whose distance to the boundary of B is greater than, for example, $h/2$ (used in the simulation), and to complete with selected points from the boundary of $B(t)$ (e.g., see Figure 2.2 for an example of selected points on the boundary of $B(t)$). As we did for $\Lambda_h(t)$ and $\langle \cdot, \cdot \rangle_{\Lambda_h(t)}$, we define the finite dimensional space Λ_{A_h} and the inner product $\langle \cdot, \cdot \rangle_{\Lambda_{A_h}}$ via a set of points of the velocity mesh internal to the region $\overline{\mathbf{A}}$ and whose distance to the surface of the cylinder $\mathbf{C}$ is greater than, for example, h, and a set of the points chosen from the surface of the cylinder $\mathbf{C}$. In practice, we have chosen Ω with its square cross-section slightly larger than the cross-section of the cylinder in order to have collocation points between the surface of cylinder $\mathbf{C}$ and Γ so that the enforcement of the constraint over $\overline{\mathbf{A}}$ can be done much more easily.

Remark

The inner product-like bracket $\langle \cdot, \cdot \rangle_{\Lambda_h(t)}$ in (3.39) makes little sense for the continuous problem, but it is meaningful for the discrete problem; it amounts to forcing the rigid body motion of $B(t)$ via a *collocation method.* A similar technique has been used to enforce Dirichlet boundary conditions by F. Bertrand et al. in [26].

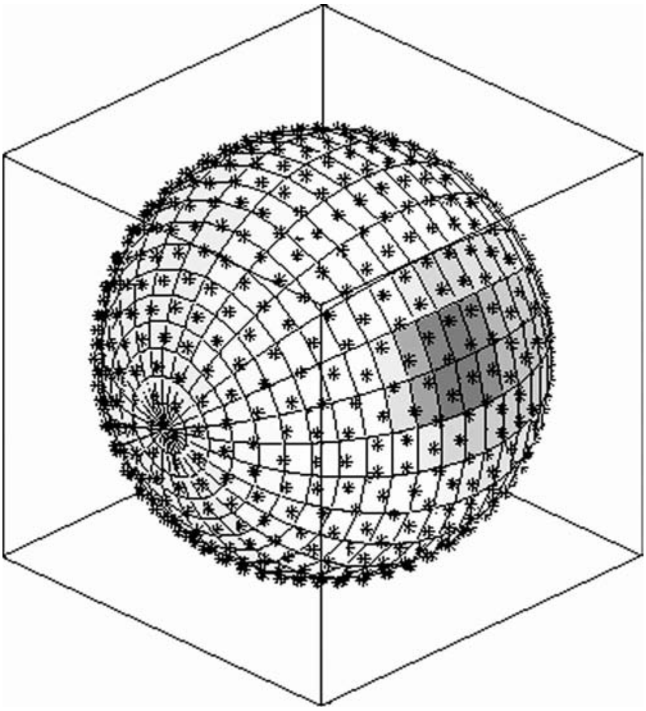


FIGURE 2.2 An example of selected points on the boundary of the rigid body.

Using the above finite dimensional spaces and the backward Euler's method for most of the subproblems in scheme (3.22) through (3.33), we obtain the following scheme after dropping some of the subscripts h (similar ones are discussed in [8,9,11–13]):

$$\mathbf{u}^0 = \tilde{\mathbf{u}}_0,\ \mathbf{G}^0 = \mathbf{G}_0,\ \mathbf{V}^0 = \mathbf{V}_0,\ \omega^0 = \omega_0 \ \textit{given}; \tag{3.40}$$

for $n \geq 0$, $\mathbf{u}^n (\simeq \mathbf{u}(t^n))$, $\mathbf{G}^n$, $\mathbf{V}^n$, and ω^n *being known, we compute* $\mathbf{u}^{n+\frac{1}{6}}$, $p^{n+\frac{1}{6}}$ *via the solution of*

$$\begin{cases} \rho_f \displaystyle\int_\Omega \frac{\mathbf{u}^{n+\frac{1}{6}} - \mathbf{u}^n}{\triangle t} \cdot \mathbf{v}\, d\mathbf{x} - \int_\Omega p^{n+\frac{1}{6}} \nabla \cdot \mathbf{v}\, d\mathbf{x} = 0, \quad \forall \mathbf{v} \in \mathbf{W}_{0h}, \\ \displaystyle\int_\Omega q \nabla \cdot \mathbf{u}^{n+\frac{1}{6}}\ d\mathbf{x} = 0, \quad \forall q \in L^2_h, \\ \mathbf{u}^{n+\frac{1}{6}} \in \mathbf{W}_h, \quad \mathbf{u}^{n+\frac{1}{6}} = \mathbf{g}_{0h}^{n+1} \quad \textit{on } \Gamma, p^{n+\frac{1}{6}} \in L^2_{0h}. \end{cases} \tag{3.41}$$

Next, compute $\mathbf{u}^{n+\frac{2}{6}}$ *via the solution of*

$$\begin{cases} \displaystyle\int_\Omega \frac{\partial \mathbf{u}}{\partial t}\cdot \mathbf{v}\,d\mathbf{x} + \int_\Omega (\mathbf{u}^{n+\frac{1}{6}}\cdot\nabla)\mathbf{u}\cdot\mathbf{v}\,d\mathbf{x} = 0, \quad \forall \mathbf{v}\in \mathbf{W}_{0h}^{n+1,-}, \\ a.e.\ on\ (t^n, t^{n+1}), \\ \mathbf{u}(t^n) = \mathbf{u}^{n+\frac{1}{6}}, \\ \mathbf{u}(t)\in \mathbf{W}_h, \quad \mathbf{u}(t) = \mathbf{g}_{0h}^{n+1} \quad on\ \Gamma_-^{n+1}\times(t^n, t^{n+1}), \end{cases} \tag{3.42}$$

and set $\mathbf{u}^{n+\frac{2}{6}} = \mathbf{u}(t^{n+1})$.

Then, compute $\mathbf{u}^{n+\frac{3}{6}}$ *and* $\lambda_{A_h}^{n+\frac{3}{6}}$ *via the solution of*

$$\begin{cases} \displaystyle\rho_f\int_\Omega \frac{\mathbf{u}^{n+\frac{3}{6}} - \mathbf{u}^{n+\frac{2}{6}}}{\triangle t}\cdot\mathbf{v}\,d\mathbf{x} + \alpha\mu_f\int_\Omega \nabla\mathbf{u}^{n+\frac{3}{6}} : \nabla\mathbf{v}\,d\mathbf{x} \\ \quad = \left\langle \lambda_{A_h}^{n+\frac{3}{6}}, \mathbf{v}\right\rangle_{\Lambda_{A_h}}, \quad \forall\mathbf{v}\in\mathbf{W}_{0h}, \\ \left\langle \mu_A, \mathbf{u}^{n+\frac{3}{6}} - \mathbf{g}_{0h}^{n+1}\right\rangle_{\Lambda_{A_h}} = 0, \quad \forall \mu_A\in\Lambda_{A_h}; \\ \mathbf{u}^{n+\frac{3}{6}}\in\mathbf{W}_h, \quad \lambda_{A_h}^{n+\frac{3}{6}}\in\Lambda_{A_h}. \end{cases} \tag{3.43}$$

Now predict the motion of the center of mass of the particle via

$$\frac{d\mathbf{G}}{dt} = \mathbf{V}(t)/2, \tag{3.44}$$

$$\left(1-\frac{\rho_f}{\rho_s}\right)M_p\frac{d\mathbf{V}}{dt} = \mathbf{F}_r/2, \tag{3.45}$$

$$\mathbf{G}(t^n) = \mathbf{G}^n, \quad \mathbf{V}(t^n) = \mathbf{V}^n, \tag{3.46}$$

for $t^n < t < t^{n+1}$. *Then set* $\mathbf{G}^{n+\frac{4}{6}} = \mathbf{G}(t^{n+1})$ *and* $\mathbf{V}^{n+\frac{4}{6}} = \mathbf{V}(t^{n+1})$.

With the center $\mathbf{G}^{n+\frac{4}{6}}$ *obtained in the above step, we enforce the rigid body motion in the region* $B(t^{n+\frac{4}{6}})$ *occupied by the particle*

$$\begin{cases} \displaystyle\rho_f\int_\Omega \frac{\mathbf{u}^{n+1} - \mathbf{u}^{n+\frac{4}{6}}}{\triangle t}\cdot\mathbf{v}\,d\mathbf{x} + \beta\mu_f\int_\Omega\nabla\mathbf{u}^{n+1} : \nabla\mathbf{v}\,d\mathbf{x} \\ \displaystyle + \left(1-\frac{\rho_f}{\rho_s}\right)M_p\frac{\mathbf{V}^{n+\frac{5}{6}} - \mathbf{V}^{n+\frac{4}{6}}}{\triangle t}\cdot\mathbf{Y} + \left(1-\frac{\rho_f}{\rho_s}\right)\mathbf{I}_p\frac{\omega^{n+1}-\omega^n}{\triangle t}\cdot\theta \\ \displaystyle = \left(1-\frac{\rho_f}{\rho_s}\right)M_p\mathbf{g}\cdot\mathbf{Y} + \left\langle \lambda^{n+\frac{4}{6}}, \quad \mathbf{v} - \mathbf{Y} - \theta\times\overrightarrow{\mathbf{G}^{n+\frac{4}{6}}\mathbf{x}}\right\rangle_{\Lambda_h^{n+\frac{4}{6}}}, \\ \forall\mathbf{v}\in\mathbf{W}_{0h}, \quad \mathbf{Y}\in\mathbb{R}^3, \quad \theta\in\mathbb{R}^3; \\ \mathbf{u}^{n+1}\in\mathbf{W}_h, \ \ \mathbf{u}^{n+1} = \mathbf{g}_{0h}^{n+1} \ \ on\ \Gamma, \lambda^{n+\frac{4}{6}}\in\Lambda_h^{n+\frac{4}{6}}, \ \ \mathbf{V}^{n+\frac{5}{6}}\in\mathbb{R}^3, \\ \omega^{n+1}\in\mathbb{R}^3, \end{cases} \tag{3.47}$$

$$\left\langle \mu, \mathbf{u}^{n+1} - \mathbf{V}^{n+\frac{5}{6}} - \omega^{n+1} \times \overrightarrow{\mathbf{G}_j^{n+\frac{4}{6}}\mathbf{x}} \right\rangle_{\Lambda_h^{n+\frac{4}{6}}} = 0, \quad \forall \mu \in \Lambda_h^{n+\frac{4}{6}}. \tag{3.48}$$

Correct the motion of the center of mass of the particle via

$$\frac{d\mathbf{G}}{dt} = \mathbf{V}(t)/2, \tag{3.49}$$

$$\left(1 - \frac{\rho_f}{\rho_s}\right) M_p \frac{d\mathbf{V}}{dt} = \mathbf{F}_r/2, \tag{3.50}$$

$$\mathbf{G}(t^n) = \mathbf{G}^{n+\frac{4}{6}}, \quad \mathbf{V}(t^n) = \mathbf{V}^{n+\frac{5}{6}}, \tag{3.51}$$

for $t^n < t < t^{n+1}$. Then set $\mathbf{G}^{n+1} = \mathbf{G}(t^{n+1})$ and $\mathbf{V}^{n+1} = \mathbf{V}(t^{n+1})$.

In (3.40) through (3.51), $\Gamma_-^{n+1} = \{\mathbf{x} | \mathbf{x} \in \Gamma, \mathbf{g}_{0h}^{n+1}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) < 0\}$ and $\mathbf{W}_{0h}^{n+1,-} = \{\mathbf{v} | \mathbf{v} \in \mathbf{W}_h, \mathbf{v} = \mathbf{0} \textit{ on } \Gamma_-^{n+1}\}$, $\Lambda_h^{n+s} = \Lambda_h(t^{n+s})$, $\mathbf{g}_{0h}^{n+1}$ is an approximation of $\mathbf{g}_0^{n+1}$ belonging to

$$\gamma \mathbf{W}_h = \{\mathbf{z}_h | \mathbf{z}_h \in (C^0(\Gamma))^3, \quad \mathbf{z}_h = \tilde{\mathbf{z}}_h|_\Gamma \quad \textit{with } \tilde{\mathbf{z}}_h \in \mathbf{W}_h\}$$

and verifying $\int_\Gamma \mathbf{g}_{0h}^{n+1} \cdot \mathbf{n} \, d\Gamma = 0$.

3.3 On the solution of subproblems (3.41–3.48)

The degenerated quasi-Stokes problem (3.41) is solved by an Uzawa preconditioned conjugate gradient algorithm as in [24,27], where the discrete elliptic problems used for preconditioning are solved by a matrix-free fast solver from FISHPAK created by Adams et al. in [28]. The advection problem (3.42) for the velocity field is solved by a wavelike equation method as in [24,29,30].

Systems (3.44) through (3.46) and (3.49) through (3.51) are systems of ordinary differential equations thanks to operator splitting. For its solution one can use a time-marching scheme with a time step smaller than $\triangle t$ (i.e., we can divide $\triangle t$ into smaller steps) to predict the translation velocity of the center of mass and the position of the center of mass, and then the regions occupied by each particle so that the repulsion forces can be effective in preventing particle–particle and particle–wall overlapping (e.g., see [9]).

Remark

For those cases involving particles of general shape, keeping the distance constant between points $\mathbf{x}_1$ and $\mathbf{x}_2$ described in (2.19) in each particle is important because we are dealing with rigid particles. To satisfy this constraint we have developed a distance-preserving scheme [31] to move points $\mathbf{x}_1$ and $\mathbf{x}_2$.

Problem (3.43) is a classical *saddle-point problem*, which can be solved by conjugate gradient algorithms. Actually, problem (3.43) is a particular case of

$$\begin{cases} \alpha\displaystyle\int_\Omega \mathbf{u}\cdot\mathbf{v}\,d\mathbf{x}+\mu\int_\Omega \nabla\mathbf{u}{:}\nabla\mathbf{v}\,d\mathbf{x}=\int_\Omega \mathbf{f}\cdot\mathbf{v}\,d\mathbf{x}+\langle\lambda,\mathbf{v}\rangle, \quad \forall\mathbf{v}\in\mathbf{W}_{0h}, \\ \langle\mu,\mathbf{u}-\mathbf{g}\rangle=0, \quad \forall\mu\in\Lambda;\ \mathbf{u}\in\mathbf{W}_h,\ \mathbf{u}=\mathbf{g} \quad on\ \Gamma,\lambda\in\Lambda. \end{cases} \tag{3.52}$$

A conjugate gradient method for the solution of problem (3.52) reads as follows:

$$\lambda^0\in\Lambda \textit{ is given}, \tag{3.53}$$

solve

$$\begin{cases} \alpha\displaystyle\int_\Omega \mathbf{u}^0\cdot\mathbf{v}\,d\mathbf{x}+\mu\int_\Omega \nabla\mathbf{u}^0{:}\nabla\mathbf{v}\,d\mathbf{x}=\int_\Omega \mathbf{f}\cdot\mathbf{v}\,d\mathbf{x}+\langle\lambda^0,\mathbf{v}\rangle, \\ \forall\mathbf{v}\in\mathbf{W}_{0h};\ \mathbf{u}^0\in\mathbf{W}_h,\ \mathbf{u}=\mathbf{g} \quad on\ \Gamma, \end{cases} \tag{3.54}$$

then solve

$$\langle\mathbf{g}^0,\mu\rangle=\langle\mu,\mathbf{u}^0-\mathbf{g}\rangle, \quad \forall\mu\in\Lambda;\ \mathbf{g}^0\in\Lambda, \tag{3.55}$$

and set

$$\mathbf{w}^0=\mathbf{g}^0. \tag{3.56}$$

For $m\geq 0$, assuming that $\lambda^m,\mathbf{u}^m$, $\mathbf{w}^m$, $\mathbf{g}^m$ are known, compute λ^{m+1}, $\mathbf{u}^{m+1}$, $\mathbf{w}^{m+1}$, $\mathbf{g}^{m+1}$ as follows:

Solve

$$\begin{cases} \alpha\displaystyle\int_\Omega \overline{\mathbf{u}}^m\cdot\mathbf{v}\,d\mathbf{x}+\mu\int_\Omega \nabla\overline{\mathbf{u}}^m{:}\nabla\mathbf{v}\,d\mathbf{x}=\langle\mathbf{w}^m,\mathbf{v}\rangle, \\ \forall\mathbf{v}\in\mathbf{W}_{0h};\ \overline{\mathbf{u}}^m\in\mathbf{W}_{0h}, \end{cases} \tag{3.57}$$

and set

$$\langle\overline{\mathbf{g}}^m,\mu\rangle=\langle\mu,\overline{\mathbf{u}}^m\rangle, \quad \forall\mu\in\Lambda;\ \overline{\mathbf{g}}^m\in\Lambda. \tag{3.58}$$

Then compute

$$\rho_m=\langle\mathbf{g}^m,\mathbf{g}^m\rangle/\langle\overline{\mathbf{g}}^m,\mathbf{w}^m\rangle, \tag{3.59}$$

and set

$$\lambda^{m+1}=\lambda^m-\rho_m\mathbf{w}^m,\ \mathbf{u}^{m+1}=\mathbf{u}^m-\rho_m\overline{\mathbf{u}}^m,\ \mathbf{g}^{m+1}=\mathbf{g}^m-\rho_m\overline{\mathbf{g}}^m. \tag{3.60}$$

If $\langle\mathbf{g}^{m+1},\mathbf{g}^{m+1}\rangle/\langle\mathbf{g}^0,\mathbf{g}^0\rangle\leq\epsilon$, then take $\mathbf{u}=\mathbf{u}^{m+1}$. If not, compute

$$\gamma_m=\langle\mathbf{g}^{m+1},\mathbf{g}^{m+1}\rangle/\langle\mathbf{g}^m,\mathbf{g}^m\rangle, \tag{3.61}$$

and set

$$\mathbf{w}^{m+1}=\mathbf{g}^{m+1}+\gamma_m\mathbf{w}^m. \tag{3.62}$$

Do $m=m+1$ and go back to (3.57).

Remark

The above conjugate gradient algorithm is similar to the one discussed in [32,33]; here a distributed Lagrange multiplier has been used instead of the boundary Lagrange multiplier used in [32,33].

4 Numerical experiments: A rotating suspension of 160 non-Brownian settling balls

In this test case, we consider the simulation of the motion of 160 balls in a horizontal rotating cylinder. The computational domain is $\Omega = (0, 1+4h_v) \times (0,4) \times (0, 1+4h_v)$. The fluid density is $\rho_f = 1$ and the fluid viscosity is $\mu_f = 0.15$. The flow field initial condition is $\mathbf{u} = \mathbf{0}$. The density of the balls is $\rho_s = 1.25$. The diameters of the balls are 0.15. The initial position of the balls and the initial translation and angular velocities of the balls are obtained by letting them settle in a horizontal cylinder with no rotation during one time unit. All balls are at the bottom of the horizontal cylinder at $t = 1$. Then we start to rotate the cylinder via the following boundary condition of the flow field

$$\mathbf{g}_0(\mathbf{x}, t) = \omega(t) \begin{pmatrix} (x_3 - x_{3c}) \\ 0 \\ -(x_1 - x_{1c}) \end{pmatrix}, if\ \mathbf{x} = (x_1, x_2, x_3)^t \in \mathbf{A},$$

with $\omega(t) = \omega_0 + (1 - e^{-20(t-1)})(\omega_1 - \omega_0)$, and $x_1 = x_{1c} = 0.5 + 2h_v$ and $x_3 = x_{3c} = 0.5 + 2h_v$ for $t \geq 1$ in order to have a smooth transition from angular velocity $\omega_0 = 0$ to $\omega_1 = 4$ with respect to the central axis of the cylinder. The pair of mesh size and time step used in the simulation presented in the following is $\{h_v, \triangle t\}$=$\{1/64, 0.001\}$. The snapshots of the positions of the 160 balls at $t = 0$ and $t = 7$ are shown in Figure 2.3. These balls behave first like a granular flow of particles inside a horizontal rotating cylinder without fluid, and later a couple balls did get lift-off. The averaged particle Reynolds number for $6 \leq t \leq 7$ is 0.244 with the diameter of the balls as characteristic length.

Then we used the result obtained with $\omega = 4$ as the initial condition and followed the same approach as above to increase the angular velocity from $\omega = 4$ to $\omega = 8$ and run over the time interval $7 \leq t \leq 16$. The snapshots of the 160 ball positions at $t = 16$ are shown in Figure 2.4. We still have particles moving against the cylinder wall. The averaged particle Reynolds number for $15 \leq t \leq 16$ is 0.711. Then again we increased ω from 8 to 12 and ran over the time interval $16 \leq t \leq 36$. The snapshots of the 160 ball positions at $t = 36$ are shown in Figure 2.4. We found that the 160 balls did start flying inside the rotating cylinder but were constrained against the cylinder wall as shown in the lower left picture in Figure 2.4 with three

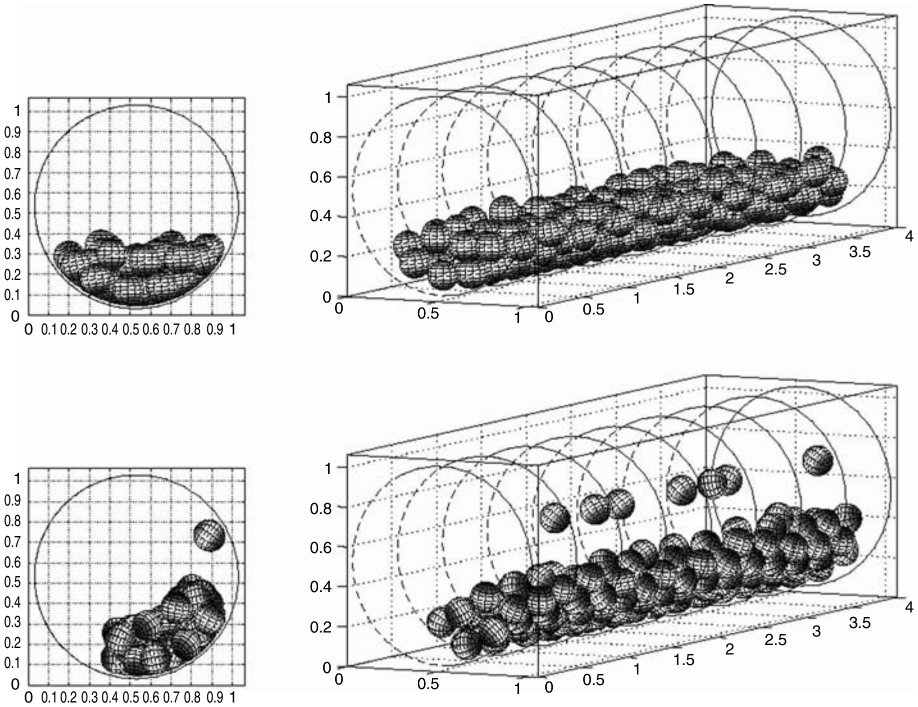


FIGURE 2.3 The position of the 160 balls at $t = 1$ with $\omega = 0$ (top) and that of the 160 balls at $t = 7$ (bottom) with $\omega = 4$.

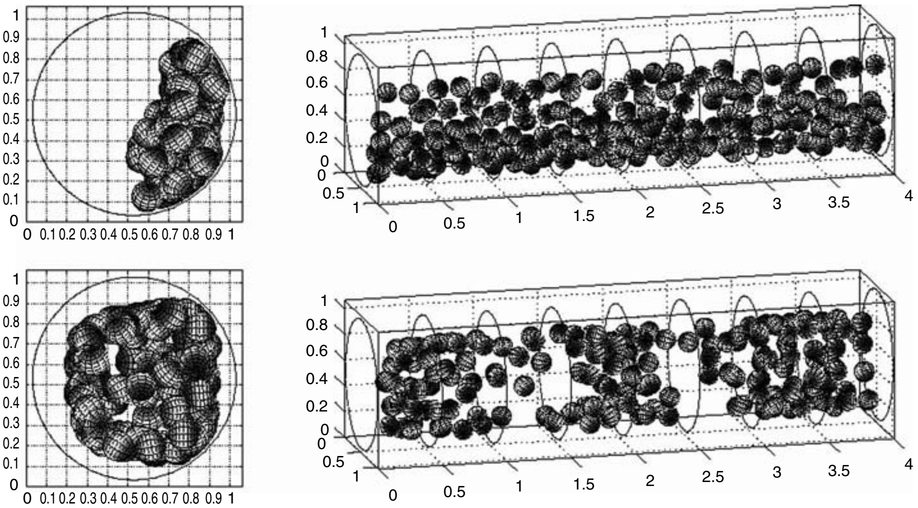


FIGURE 2.4 The position of the 160 balls at $t = 16$ with $\omega = 8$ (top) and that of the 160 balls at $t = 36$ with $\omega = 12$ (bottom).

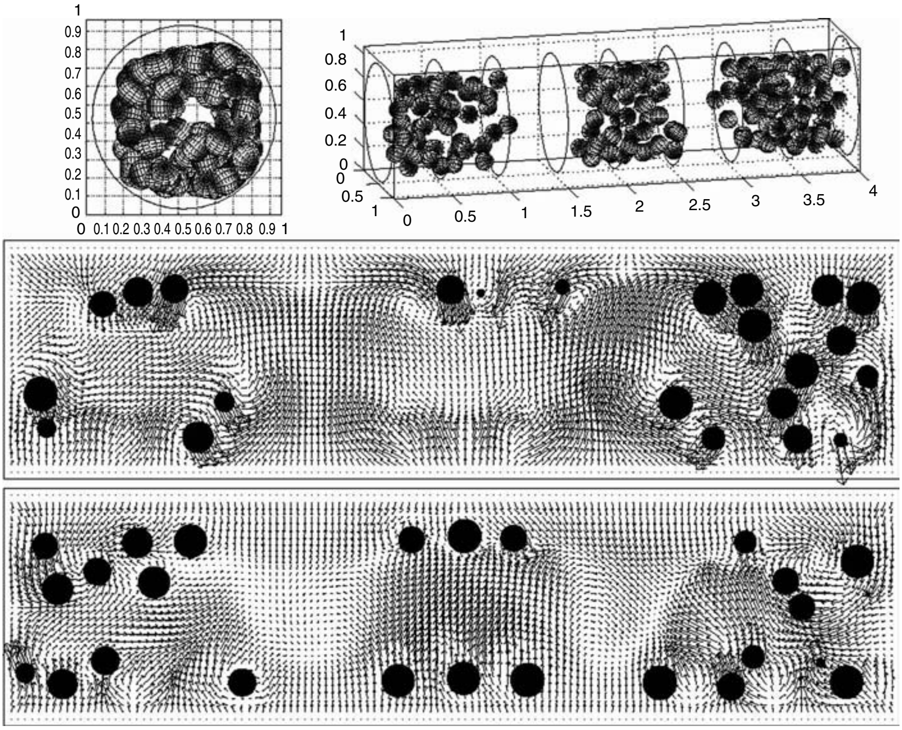


FIGURE 2.5 The position of the 160 balls at $t = 55$ obtained with $\omega = 16$ (top), the projected velocity field on a plane parallel to x_2x_3-plane and passing through the central axis of the cylinder (middle) and the projected velocity field on a plane parallel to x_1x_2-plane and passing through the central axis of the cylinder (bottom) for $\{h_v, \triangle t\}$=$\{1/30, 0.001\}$.

clusters forming in the horizontal rotating cylinder. The averaged particle Reynolds number for $35 \leq t \leq 36$ is 3.218.

When increasing ω from 12 to 16, we did obtain three clusters with clear separation as shown in Figure 2.5. The averaged particle Reynolds number for $54 \leq t \leq 55$ is 4.734. The projections of velocity field onto planes parallel to the coordinate planes and passing through the central axis of the cylinder are also shown in Figure 2.5. We can see that the fluid velocity field does not form any pattern that has a strong influence on the formation of particle clusters, but the particle motion has influence on the flow field.

5 Acknowledgments

We acknowledge the helpful comments and suggestions of R. Bai, E.J. Dean, J. He, H.H. Hu, P.Y. Huang, G.P. Galdi, D.D. Joseph, Y. Kuznetsov, and V. Paulsen. We also acknowledge the support of the National Science Foundation (grants ECS-9527123, CTS-9873236, DMS-9973318, CCR-9902035, DMS-0209066, DMS-0443826) and DOE/LASCI (grant R71700K-292-000-99).

References

[1] Tirumkudulu, M., Tripathi, A., and Acrivos, A., Particle segregation in monodisperse sheared suspensions, *Phys. Fluids*, 11, S13, 1999.

[2] Tirumkudulu, M., Mileo, A., and Acrivos, A., Particle segregation in monodisperse sheared suspensions in a partially filled rotating horizontal cylinder, *Phys. Fluids*, 12, 1615, 2000.

[3] Joseph, D.D., Wang, J., Bai, R., Yang, B.H., and Hu, H.H., Particle motion in a liquid film rimming the inside of a partially filled rotating cylinder, *J. Fluid Mech.*, 496, 139, 2003.

[4] Lipson, S.G., Periodic banding in crystallization from rotating supersaturated solutions, *J. Phys.: Condens. Matters*, 13, 5001, 2001.

[5] Lipson, S.G. and Seiden, G., Particles banding in rotating fluids: A new pattern-forming system, *Physica A: Statistical Mechanics and Its Application*, 314, 272, 2002.

[6] Breu, A.P.J., Kruelle, C.A., and Rehberg, I., Pattern formation in a rotating aqueous suspension, *Europhys. Lett.*, 62, 491, 2003.

[7] Matson, W.R., Ackerson, B.J., and Tong, P., Pattern formation in a rotating suspension of non-Brownian settling particles, *Phy. Rev. E*, 67, 050301 (R), 2003.

[8] Glowinski, R., Pan, T.-W., Hesla, T., and Joseph, D.D., A distributed Lagrange multiplier/fictitious domain method for flows around moving rigid bodies: Application to particulate flows, *Int. J. Multiphase Flow*, 25, 755, 1999.

[9] Glowinski, R., Pan, T.-W., Hesla, T., Joseph, D.D., and Périaux, J., A fictitious domain approach to the direct numerical simulation of incompressible viscous flow past moving rigid bodies: Application to particulate flow, *J. Comput. Phys.*, 169, 363, 2001.

[10] Dong S., Liu, D., Maxey, M., and Karniadakis, G.E., Spectral distributed multiplier (DLM) method: Algorithm and benchmark test, *J. Comp. Phys.*, 195, 695, 2004.

[11] Glowinski, R., Hesla, T., Joseph, D.D., Pan, T.-W., and Périaux, J., Distributed Lagrange multiplier methods for particulate flows, in *Computational Science for the 21st Century*, Bristeau, M.O., Etgen, G., Fitzgibbon, W., Lions, J.-L., Périaux, J., and Wheeler, M.F., Eds., John Wiley, Chichester, 1997, 270.

[12] Glowinski, R., Pan, T.-W., Hesla, T., Joseph, D.D., and Périaux, J., A fictitious domain method with distributed Lagrange multipliers for the numerical simulation of particulate flow, in *Domain Decomposition Methods 10*, Mandel, J., Farhat, C., and Cai, X.C., Eds., American Mathematical Society, Providence, RI, 1998, 121.

[13] Glowinski, R., Pan, T.-W., Hesla, T., Joseph, D.D., and Périaux, J., A distributed Lagrange multiplier/fictitious domain method for flow around moving rigid bodies: Application to particulate flow, *Int. J. Num. Meth. Fluids*, 30, 1043, 1999.

[14] Baaijens, F.P.T., A fictitious domain/mortar element method for fluid-structure interaction, *Int. J. Numer. Meth. Fluids*, 35, 743, 2001.

[15] Wagner, G.J., Moes, N., Liu, W.K., and Belytschko, T., The extended finite element method for rigid particles in Stokes flow, *Int. J. Numer. Meth. Eng.*, 51, 293, 2001.

[16] Yu, Z., Phan-Thien, N., Fan, Y., and Tanner, R.I., Viscoelastic mobility problem of a system of particles, *J. Non-Newtonian Fluid Mech.*, 104, 87, 2002.

[17] Maury, B., A many-body lubrication model, *C. R. Acad. Sci. Paris, Série 1*, 325, 1053, 1997.

[18] Hu, H.H., Direct simulation of flows of solid-liquid mixtures, *Int. J. Multiphase Flow*, 22, 335, 1996.

[19] Johnson, A. and Tezduyar, T., 3-D simulation of fluid-rigid body interactions with the number of rigid bodies reaching 100, *Comp. Meth. Appl. Mech. Eng.*, 145, 301, 1997.

[20] Maury, B. and Glowinski, R., Fluid-particle flow: A symmetric formulation, *C.R. Acad. Sci. Paris, Série 1*, 324, 1079, 1997.

[21] Peskin, C.S., Numerical analysis of blood flow in the heart, *J. Comp. Phys.*, 25, 220, 1977.

[22] Peskin, C.S. and McQueen, D.M., Modeling prosthetic heart valves for numerical analysis of blood flow in the heart, *J. Comp. Phys.*, 37, 113, 1980.

[23] Peskin, C.S., Lectures on mathematical aspects of physiology, *Lectures in Applied Math.*, 19, 69, 1981.

[24] Glowinski, R., Finite element methods for incompressible viscous flow, in *Handbook of Numerical Analysis, Vol. IX*, Ciarlet, P.G. and Lions, J.-L. Eds., North-Holland, Amsterdam, 2003, 3.

[25] Chorin, A.J., Hughes, T.J.R., Marsden, J.E., and McCracken, M., Product formulas and numerical algorithms, *Comm. Pure Appl. Math.*, 31, 205, 1978.

[26] Bertrand, T., Tanguy, P.A., and Thibault, F., A three-dimensional fictitious domain method for incompressible fluid flow problems, *Int. J. Num. Meth. Fluids*, 25, 719, 1997.

[27] Glowinski, R., Pan, T.-W., and Périaux, J., Distributed Lagrange multiplier methods for incompressible flow around moving rigid bodies. *Comp. Meth. Appl. Mech. Eng.*, 151, 181, 1998.

[28] Adams, J., Swarztrauber, P., and Sweet, R., FISHPAK: A package of Fortran subprograms for the solution of separable elliptic partial differential equations, The National Center for Atmospheric Research, Boulder, CO, 1980.

[29] Dean, E.J. and Glowinski, R., A wave equation approach to the numerical solution of the Navier-Stokes equations for incompressible viscous flow, *C.R. Acad. Sci. Paris, Série 1*, 325, 783, 1997.

[30] Dean, E.J., Glowinski, R., and Pan, T.-W., A wave equation approach to the numerical simulation of incompressible viscous fluid flow modeled by the Navier-Stokes equations, in *Mathematical and Numerical Aspects of Wave Propagation*, De Santo, J.A., Ed., SIAM, Philadelphia, 1998, 65.

[31] Pan, T.-W., Joseph, D.D., and Glowinski, R., Simulating the dynamics of fluid-ellipsoid interactions, *Computers & Structures*, 83, 463, 2005.

[32] Glowinski, R., Pan, T.-W., and Périaux, J., A fictitious domain method for Dirichlet problem and applications, *Comp. Meth. Appl. Mech. Eng.*, 111, 283, 1994.

[33] Glowinski, R., Pan, T.-W., and Périaux, J., A fictitious domain method for external incompressible viscous flow modeled by Navier-Stokes equations, *Comp. Meth. Appl. Mech. Eng.*, 112, 133, 1994.

CHAPTER 3

On the homogenization of optimal control problems on periodic graphs

P.I. Kogut
Department of Differential Equations, National University of Dnipropetrovsk,
Dnipropetrovsk, Ukraine

G. Leugering
Institute of Applied Mathematics, University of Erlangen–Nuremberg,
Erlangen, Germany

1 Introduction

Problems of transmission, transport and flow in networked systems are currently very much under consideration in various industrial and medical applications, such as bloodflow in arteries, the flow of gas in networks of pipelines, the flow of water in sewer or irrigational systems, heat flow in problems of hyperthermia, and so forth. It is typically not reasonable, and moreover numerically prohibitive, to study such problems on the level of each subdomain, taking into account the local representations of the flow in terms of partial differential equations. More precisely, in such networks there are mushy regions, where a local resolution is not called for, and where instead an average model should be imposed to describe the dynamics. The underlying mathematical procedure behind this point of view is the theory of homogenization of partial differential equations. There is a vast literature on homogenization results available dealing with linear elliptic, parabolic and hyperbolic equations as such. Far less is known in the context of nonlinear equations. It is not possible to provide a proper account of the activities in this field. See, for example, Zhikov [33] and Benssoussan, and Lions and Papanicolau [1] as general references. The theory of homogenization is most complete and well established for periodic structures, where typically the results concern equations on perforated domains or so-called reticulated structures. See Cioranescu and Saint-Jean

Paulin [28] as an exemplary reference. One of the main reasons for the number of papers and textbooks on *periodic homogenization* is that the homogenized limits can be computed explicitly, at least in principle. This is generally not true for nonperiodic problems. It is, however, only very recently that results have been reported in the context of homogenization of partial differential equations on thin networks, like graphs. See Bouchittee and Fragala [2–4] and Zhikov [29–31], Panasenko [25,26] and Chechkin and Zhikov [7] and Lenczner [21], where linear elliptic problems have been considered.

Homogenization theory has very recently been utilized in problems concerning *micro-flows*, that is in problems mentioned at the beginning of this section. We mention the work of Canic et al. [5]. However, homogenization concerns a single artery, and no network problems are considered. Control problems for hyperbolic and elliptic equations on networks without any homogenization have been discussed in a series of papers by Leugering and Lagnese; see, for example, the recent monograph by Lagnese and Leugering [22], where domain decomposition of optimal control problems on networks, among other topics, is discussed in depth. See also Leugering and Schmidt [23] for nonlinear flow problems on graphs.

In general, very few authors deal with problems of optimal control in the context of homogenization. We mention the work of Saint-Jean Paulin, Kesavan, Cioransecu, and Murat [8,11], and Hoppe and Petrova [10]. The common approach is to first derive optimality conditions that turn the (typically unconstrained) optimization problem into a partial differential equations (PDE) problem parametrized by the homogenization parameter. That problem is subjected to an asymptotic analysis, and the homogenized limiting problem is viewed as an optimality condition of *some* limiting problem. The authors of these notes took the more natural and mathematically more involved route of variational calculus, the so-called direct method, which consists of directly analyzing the optimization problem parametrized by the homogenization parameter, and without resorting to the corresponding optimality conditions. See Kogut and Leugering [12–18,20].

In these notes we consider a linear-quadratic optimal control problem for 1-D partial differential elliptic equations defined on a periodic planar graph. We study the asymptotic behavior of this problem when the ε-period of the graph tends to zero, and look for the limiting homogenized optimal control problem. In particular, we require that an optimal solution, and the minimum of the cost functional for the homogenized problem, are the limit values (in some reasonable sense) of the corresponding quantities for original problems. Detailed proofs can be found in [19]. The elliptic homogenization is supposed to be extended to parabolic and hyperbolic problems in a forthcoming publication.

2 ε-Periodic graph-like structures in $\mathbf{R}^2$

We begin with some basic definitions. Let $\square$ and $\square^S \subset \square$ be the following sets

$$\square = [0,1)^2 = [0,1) \times [0,1),$$
$$\square^S = \{(x_1, x_2) : [x_1 \in [0,1) \wedge x_2 = 0] \vee [x_2 \in [0,1) \wedge x_1 = 0]\}$$

Definition 2.1
We say that the set $\square$ is the cell of periodicity for some graph $\mathcal{F}$ on $\mathbf{R}^2$ if $\square$ contains a star structure with the following properties

- *all edges of this structure have a common point $M \in int\, \square$*
- *each edge is a line segment and all end-points of these edges belong to the boundary of $\square$;*
- *in the set of end-points (vertices) there exist pairs $\underline{(M_i; M_k)}$ such that $x_1^{M_i} = x_1^{M_k}$ or $x_2^{M_i} = x_2^{M_k}$*

Let $\varepsilon \in E = (0, \varepsilon^0]$ be a small parameter. We call $\mathcal{F}_\varepsilon$ an ε-periodic graph on $\mathbf{R}^2$ if

$$\mathcal{F}_\varepsilon = \varepsilon \mathcal{F}(x) = \{\varepsilon x : x \in \mathcal{F}\}.$$

By analogy we define the unbounded grid G on $\mathbf{R}^2$ with $\square^S$ as the cell of periodicity and the ε-grid G_ε as the corresponding ε-scale of the original set G.

Let

$$I^{ed} = \{I_j\ ,\ j = 1, 2, \ldots, K\} \tag{2.1}$$

be the set of all edges on $\square$ and let

$$\mathcal{M} = \{M_i\ ,\ i = 1, 2. \ldots, L\} \tag{2.2}$$

be the set of all vertices on $\square$ that belong to $\square^S$.

Let Ω be an open bounded domain in $\mathbf{R}^2$ with Lipschitz boundary $\partial\Omega = \Gamma_1 \cap \Gamma_2$ such that

$$\Gamma_1 \cap \Gamma_2 = \emptyset, \quad \Gamma_1 \subset G.$$

Definition 2.2
We say that Ω_ε has an ε-periodic graph-like structure of

$$\Omega_\varepsilon = \Omega \cap \mathcal{F}_\varepsilon, \quad \textit{and } \varepsilon = 1/n,$$

where n is an integer.

It is easy to see that the cell of periodicity for Ω_ε is $\varepsilon\square$ and $\Gamma_1 \cap (\cup \varepsilon \mathcal{M}) \neq \emptyset$. In what follows, we will always suppose that $\varepsilon \in E = \left\{1, \frac{1}{2}, \frac{1}{3} \cdots \right\}$.

3 Description of Ω_ε in terms of singular measures

Following Zhikov's approach (see [29] through [31]) we will describe the geometry of the sets $\mathcal{F}_\varepsilon$ and G_ε in the terms of singular measures in $\mathbf{R}^2$. We note that these measures do not satisfy the regularity property with respect to the corresponding Lebesgue measures.

For every segment $I_i \in I^{ed}, i = 1, 2, \ldots, K$ we denote by μ_i its corresponding Lebesgue measure. Now we define the $\square$-periodic Borel measure μ in $\mathbf{R}^2$ as follows

$$\mu = \sum_{i=1}^{K} g_i \cdot \mu_i \qquad \text{on } \square, \tag{3.3}$$

where $g_1, g_2, \ldots, g_K$ are nonnegative weights such that $\int_\square d\mu = 1$. Thus the support of the measure μ is the union of all edges $I_i \in I^{ed}$, each of which is a 1-dimensional manifold in $\mathbf{R}^2$. We introduce the ε-periodic measure μ_ε by

$$\mu_\varepsilon(B) = \varepsilon^2 \mu(\varepsilon^{-1} B) \text{ for every Borel set } B \subset \mathbf{R}^2. \tag{3.4}$$

Then

$$\int\limits_{\varepsilon\square} d\mu_\varepsilon = \varepsilon^2 \int\limits_{\square} d\mu = \varepsilon^2.$$

As follows from [29], such defined measure μ_ε is weakly convergent to the Lebesgue measure, that is,

$$d\mu_\varepsilon \rightharpoonup dx \Leftrightarrow \lim_{\varepsilon \to 0} \int\limits_{\mathbf{R}^2} \varphi d\mu_\varepsilon = \int\limits_{\mathbf{R}^2} \varphi dx \tag{3.5}$$

for every $\varphi \in C_0^\infty(\mathbf{R}^2)$. Let us denote by $\gamma_i,\ i = 1, 2$ the following sets:

$$\gamma_1 := \{x \in \square^S | 0 \leq x_1 < 1,\ x_2 = 0\},\ \gamma_2 := \{x \in \square^S | 0 \leq x_2 < 1,\ x_1 = 0\}$$

We decompose $\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2$, where $\mathcal{M}_1 = \{M_j\}_{j=1,\ldots L_1}$ and $\mathcal{M}_2 = \{M_j\}_{j=1,\ldots L_2}$, $L = L_1 + L_2$ are the sets of vertices belonging to γ_1, γ_2, respectively. Let μ_1, μ_2 be measures in $\mathbf{R}^1$ such that

$$\mu_1^S = \sum_{j=1}^{L_1} \rho_j^1 \delta_{M_j^1}, \quad \mu_2^S = \sum_{j=1}^{L_2} \rho_j^2 \delta_{M_j^2}, \tag{3.6}$$

where the nonnegative weights ρ_j^i satisfy $\sum\limits_{j=1}^{L_i} \rho_j^i = 1,\ i = 1, 2$ and $\delta_{M_j^i}$ are the Dirac measures located at M_j^i. It is easy to see that μ_i^S are the Radon measures satisfying $\int_{\gamma_i} d\mu_i^S = 1,\ i = 1, 2$. We define

$$\mu^S = \mu_1^S + \mu_2^S,$$

which is a $\square^S$-periodic measure in the graph G. We may then define the corresponding scaling measure

$$\mu_\varepsilon^S(B) = \varepsilon\mu^S\left(\frac{1}{\varepsilon}B\right). \tag{3.7}$$

We observe that

$$\int\limits_{\varepsilon\square^S} d\mu_\varepsilon^S = \varepsilon\int\limits_{\gamma_1} d\mu_1^S + \varepsilon\int\limits_{\gamma_2} d\mu_2^S = 2\varepsilon.$$

Proposition 3.1

Let ℓ be the Hansdorff-Lebesgue measure in $\mathbf{R}^1$ and let Γ be a "line" in G. Then μ_ε^S converges weakly to ℓ,

$$\int\limits_\Gamma \Psi(x)d\mu_\varepsilon^S \longrightarrow \int\limits_\Gamma \Psi(x)d\ell \qquad \textit{for each } \Psi \in C_0^\infty(\Gamma) \tag{3.8}$$

as $\varepsilon \to 0$.

4 Sobolev spaces on ε-periodic graphs

For the sake of brevity, in the present section we refer the reader to the notations and results in [29] through [31] and introduce some additional spaces associated with boundary optimal control problems on ε-periodic graphs. We define the spaces $L^2(\Omega,\ d\mu_\varepsilon)$ and $L^2(\Gamma_1,\ d\mu_\varepsilon^S)$ in the usual way. The natural definition of weak and strong convergence with respect to the sequence of measures is taken from [29] through [31]. The important point is that we have compactness:

Proposition 4.1

Every bounded sequence $\{h_\varepsilon \in L^2(\Gamma_1, d\mu_\varepsilon^S)\}$ is compact with respect to the μ^S-weak convergence. Furthermore, the μ^S-weak convergence of $h_\varepsilon \to h$ and the relation

$$\lim_{\varepsilon\to 0}\int\limits_{\Gamma_1} h_\varepsilon^2\, d\mu_\varepsilon^S = \int\limits_{\Gamma_1} h^2 dl \tag{4.9}$$

imply the strong μ^S-convergence of $h_\varepsilon \to h$.

Let $C^\infty(\Omega, \Gamma_2)$ be the class of smooth functions $\varphi \in C^\infty(\bar{\Omega})$ such that $\varphi|_{\Gamma_2} = 0$. Here Γ_2 is the second "part" of the boundary $\partial\Omega = \Gamma_1 \cup \Gamma_1$.

Definition 4.2
A function $y(x)$ belongs to $V(\Omega,\Gamma_2,d\mu)$ if there exists a vector $z\in(L^2(\Omega,d\mu))^2$ and a sequence $\{y_m\in C^\infty(\Omega,\Gamma_2)\}_{m\in N}$ such that

$$\lim_{m\to\infty}\int\limits_\Omega (y_m-y)^2\,d\mu=0, \tag{4.10}$$

$$\lim_{m\to\infty}\int\limits_\Omega |\nabla y_m-z|^2\,d\mu=0. \tag{4.11}$$

It is important to note that locally the Sobolev spaces of Definition 4.2 are identical to the classical spaces. Moreover, and more importantly, the continuity condition reduces to the classical nodal condition in the network formulation. This is the content of the following:

Proposition 4.3

Let $\mathcal{F}$ be a $\Box$-periodic unbounded graph on $\mathbf{R}^2$, let μ be the $\Box$-periodic Borel measure in $\mathbf{R}^2$ defined by (3.4), and let y be any function of $V(\Omega,\Gamma_2,d\mu)$. Then:

(*i*) *$y|_{I_i}\in H^1(I_i)$ for any edge $I_i\in\Omega\cap\mathcal{F}$;*

(*ii*) *the restriction y on the set $\Omega\cap\mathcal{F}$ is a continous function.*

We define on the cell of periodicity $\Box$ (or the period torus) the sets of so-called potential and solenoidal vectors as it was done in [32]. Let $C^\infty_{per}=C^\infty_{per}(\Box)$ be the space of infinitely differentiable periodic functions, and let $[L^2_{per}(\Box,d\mu)]^2$ be the space of μ-measurable periodic functions $f=[f_1,f_2]$ such that $\int_\Omega(f_1^2(x)+f_2^2(x))\,d\mu<\infty$. We adopt the notion of solenoidal and potential vectors from Zhikov [7].

5 Two-scale convergence with respect to the measures μ and μ^S

We recollect the main results concerning the extension of the well-known method of two-scale convergence that was independently obtained in [29] and [2]. Also we give the definition and main properties of two-scale limits with respect to the scaling measure μ^S_ε. This measure is connected with boundary conditions on $\partial(\Omega\backslash\mathcal{F}_\varepsilon)$ and this one is singular with respect to the measure μ.

Definition 5.1 (*See [29]*).
A sequence $\{y_\varepsilon\in L^2(\Omega,\Box)$ is called weakly two-scale convergent to a function $y\in L^2(\Omega,\Box)$ (here $y=y(x,z)$, $L^2(\Omega,\Box)=L^2(\Omega\times\Box,\,dx\times d\mu)$) if

$$\lim_{\varepsilon\to 0}\int\limits_\Omega \Phi(x,\varepsilon^{-1}x)y_\varepsilon(x)\,d\mu_\varepsilon=\int\limits_\Omega\int\limits_\Box \Phi(x,z)y(x,z)\,d\mu(z)\,dx \tag{5.12}$$

for every test function $\Phi(x,z) = \varphi(x) \cdot b(z)$ *such that* $\varphi \in C(\bar{\Omega}), b \in L^2_{per}(\square)$. *We will always denote this type of convergence as* $y_\varepsilon \overset{2}{\rightharpoonup} y(x,z)$.

An analogous definition is then easily derived for sequences $\{h_\varepsilon \in L^2(\Gamma_1,\, d\mu_\varepsilon^S)\}$.

Again, the important point is that we enjoy compactness.

Proposition 5.2

If sequences $\{y_\varepsilon \in L^2(\Omega,\, d\mu_\varepsilon)\}$ *and* $\{h_\varepsilon \in L^2(\Gamma_1, d\mu_\varepsilon^S)\}$ *are bounded, then these sequences are compact in the sense of the weak two-scale convergence.*

6 Variational convergence of constrained minimization problems on varying graphs

We consider the family

$$\left\langle \inf_{(y,u,h)\in\Xi_\varepsilon} I_\varepsilon(y,u,h) \right\rangle \tag{6.13}$$

of the constrained minimization problem on the ϵ-graph $\Omega_\varepsilon = \Omega \cap \varepsilon\mathcal{F}$ where

(i) Ξ_ε is a subset of $H^1(\Omega,\, d\mu_\varepsilon) \times (L^2(\Omega,\, d\mu_\varepsilon) \times L^2(\Gamma_1,\, d\mu_\varepsilon^S)$

(ii) $I_\varepsilon : \Xi_\varepsilon \to \bar{\mathbf{R}} = \mathbf{R} \cup \{+\infty\}$.

Here $u \in L^2(\Omega,\, d\mu_\varepsilon)$ is a distributed control, $h \in L^2(\Gamma_1,\, d\mu_\varepsilon^s)$ is a boundary control, and $y \in H^1(\Omega,\, d\mu_\varepsilon)$ is a "state" variable. We also say that a minimal sequence $\{(y_m, u_m, h_m)\}_{m\in\mathbf{N}}$ for I_ε, is weakly convergent to a triplet (y^0, u^0, h^0) if the sequences converge weakly in their mutual spaces and there exist gradients $\{\nabla y_m \in [L^2(\Omega,\, d\mu_\varepsilon)]^2\}$ such that the sequence $\{\nabla y_m\}$ is compact with respect to the weak convergence in $[L^2(\Omega,\, d\mu_\varepsilon)]^2$. It is a classical result that boundedness, sequential weak closedness of the admissible sets, properness, and sequential lower semicontinuity in the right topology guarantees existence of an optimal solution and the canonical properties of minimizing sequences.

In order to proceed with this program we consider two-scale convergence of the triplets within the admissible sets.

Definition 6.1

We say that the sequence of triplets $\{(y_\varepsilon, u_\varepsilon, h_\varepsilon)\}_{\varepsilon\in E}$ *is weakly convergent in the scale of spaces* $\{H^1(\Omega,\, d\mu_\varepsilon) \times L^2(\Omega,\, d\mu_\varepsilon) \times L^2(\Gamma_1,\, d\mu_\varepsilon^S)\}_{\varepsilon\in E}$, *or shortly is* w_ε*-convergent, if there are functions*

$$y \in H^1(\Omega),\ u \in L^2(\Omega),\ h \in L^2(\Gamma_1),\ p \in L^2(\Omega, \square),$$

and there exists a sequence of gradients $\{\nabla y_\varepsilon \in [L^2(\Omega,\, d\mu_\varepsilon)]^2\}_{\varepsilon\in E}$ *such that*

$$u_\varepsilon \rightharpoonup u \;\; \mu\text{-weakly},\; h_\varepsilon \rightharpoonup h \;\; \mu^S\text{-weakly} \tag{6.14}$$

$$y_\varepsilon \overset{2}{\rightharpoonup} y(x), \tag{6.15}$$

$$\nabla y_\varepsilon \overset{2}{\rightharpoonup} p(x,z), \tag{6.16}$$

$$p(x,y) - \nabla y(x) \in L^2(\Omega, V_{pot}). \tag{6.17}$$

In the following theorem we state sufficient conditions for w_ε-compactness of sequences.

Theorem 6.2

Let $\{(y_\varepsilon, u_\varepsilon, h_\varepsilon)\}_{\varepsilon\in E}$ *be any sequence in the scale* $\{H^1(\Omega,\, d\mu_\varepsilon)\times L^2(\Omega,\, d\mu_\varepsilon)\times L^2(\Gamma_1,\, d\mu_\varepsilon^S)\}_{\varepsilon\in E}$ *for which the following conditions hold:*

(*i*) *the sequence* $\{(y_\varepsilon, u_\varepsilon, h_\varepsilon)\}_{\varepsilon\in E}$ *is bounded;*

(*ii*) *there exists a bounded sequence of gradients* $\{\nabla y_\varepsilon\}$.

Then the original sequence has a w_ε*-convergent subsequence.*

Given this concept of convergence of triplets, we may now introduce the notion of convergence of sets in spaces with varying measures in the methods of Kuratowski.

Definition 6.3
We say that a set

$$\Xi_0 \subset H^1(\Omega)\times L^2(\Omega)\times L^2(\Gamma_1)$$

is the sequential two-scale limit in the sense of Kuratowski or K-limit, of the sequence

$$\{\Xi_\varepsilon \subset H^1(\Omega, d\mu_\varepsilon)\times L^2(\Omega, d\mu_\varepsilon)\times L^2(\Gamma_1, d\mu_\varepsilon^S)\}_{\varepsilon\in E} \tag{6.18}$$

if the following conditions are satisfied:

(*i*) *for every triplet* $(y,u,h)\in\Xi_0$ *there exists a constant* $\varepsilon^0\in E$ *and a sequence* $\{(y_\varepsilon, u_\varepsilon, h_\varepsilon)\}$ *w-converging to* (y,u,h) *such that* $(y_\varepsilon, u_\varepsilon, h_\varepsilon)\in\Xi_\varepsilon$ *for every* $\varepsilon\le\varepsilon^0$*;*

(*ii*) *if* $\{\Xi_{\varepsilon_k}\}$ *is a subsequence of* $\{\Xi_\varepsilon\}_{\varepsilon\in E}$ *and* $\{(y_k, u_k, h_k)\}$ *is a sequence w-converging to* (y,u,h) *such that* $(y_k, u_k, h_k)\in\Xi_{\varepsilon_k}$ *for every* $k\in\mathbf{N}$, *then* $(y,u,h)\in\Xi_0$.

Obviously, this notion coincides with the classical Kuratowski convergence if the topology is the standard one. Let us return to the main object of this section, namely the sequence of constrained minimization problems (6.13). We now introduce the notion of "limit" minimization problems as ε tends to zero.

Definition 6.4
We say that a minimization problem on $H^1(\Omega) \times L^2(\Omega) \times L^2(\Gamma_1)$

$$\left\langle \inf_{(y,u,h)\in\Xi_0} I_0(y,u,h) \right\rangle \tag{6.19}$$

is the variational limit of the sequence (6.13) with respect to the w-convergence if the following conditions are satisfied:

- (*i*) Ξ_0 *is a nonempty two-scale* K*-limit of the sets* $\{\Xi_\varepsilon\}_{\varepsilon\in E}$*;*
- (*ii*) *for every triplet* $(y,u,h) \in \Xi_0$ *and for every sequence* $\{(y_k,u_k,h_k)\}_{k\in\mathbf{N}}$ w*-converging to* (y,u,h) *such that* $(y_k,u_k,h_k) \in \Xi_{\varepsilon_k}$ *for some* $\varepsilon_k \to 0$ *as* $k \to \infty$ *it is*

$$I_0(y,u,h) \leq \liminf_{k\to\infty} I_{\varepsilon_k}(y_k,u_k,h_k); \tag{6.20}$$

- (*iii*) *for every triplet* $(y,u,h) \in \Xi_0$ *there exists a sequence* $\{(y_\varepsilon,u_\varepsilon,h_\varepsilon)\}_{\varepsilon\in E}$ *such that*
  $(y_\varepsilon,u_\varepsilon,h_\varepsilon) \in \Xi_\varepsilon$ *for every* $\varepsilon \in E$*,*
  $(y_\varepsilon,u_\varepsilon,h_\varepsilon) \to (y,u,h)$ *in the sense of* w*-convergence,*

$$I_0(y,u,h) \geq \limsup_{\varepsilon\to 0} I_\varepsilon(y_\varepsilon,u_\varepsilon,h_\varepsilon). \tag{6.21}$$

Remark
Note that this definition can be interpreted as some extension of the notion of Γ-convergence. Indeed, in the case when $d\mu_\varepsilon = dx$ and $d\mu_\varepsilon^S = dl$ the limit functional $I_0 : \Xi_0 \to \mathbf{R}$ coincides with the sequential Γ-limit of the sequence

$$\left\{ P_{\Xi_\varepsilon} I_\varepsilon : \left(H^1(\Omega) \times L^2(\Omega) \times L^2(\Gamma_1) \right) \to \bar{\mathbf{R}} \right\}$$

with respect to the product of corresponding weak topologies where

$$P_{\Xi_\varepsilon} I_\varepsilon(y,u,h) = \begin{cases} I_\varepsilon(u,y,h), \ (y,u,h) \in \Xi_\varepsilon, \\ +\infty, \qquad \text{otherwise.} \end{cases}$$

In the following theorem we will establish that under some natural assumptions the variational convergence of a sequence (6.13) to a problem (6.19) implies the convergence of the minimum values of I_ε on Ξ_ε to the minimum value of I_0 on Ξ_0, and moreover we will prove that in this case every w-cluster point of the sequence of the minimizers of I_ε is a minimizer of I_0.

Theorem 6.5

Assume that the constrained minimization problem (6.19) is the variational limit of the sequence (6.13) in the sense of Definition 5.4 and this problem *has a nonempty set of solutions. For every $\varepsilon \in E$, let $(y^0_\varepsilon, u^0_\varepsilon, h^0_\varepsilon) \in \Xi_\varepsilon$ be a solution of the corresponding problem (6.13) (i.e., $(y^0_\varepsilon, u^0_\varepsilon, h^0_\varepsilon)$ is a minimizer). If the sequence $\{(y^0_\varepsilon, u^0_\varepsilon, h^0_\varepsilon)\}_{\varepsilon \in E}$ is relatively w-compact, then every w-cluster point (y^0, u^0, h^0) of this sequence is a minimizer of I_0 on Ξ_0 and*

$$I_0(y^0, u^0, h^0) = \lim_{k\to\infty} I_{\varepsilon_k}(y^0_{\varepsilon_k}, u^0_{\varepsilon_k}, h^0_{\varepsilon_k}), \tag{6.22}$$

$$\min_{(y,u,h)\in\Xi_0} I_0(y, u, h) = \lim_{\varepsilon\to 0} \min_{(y,u,h)\in\Xi_\varepsilon} I_\varepsilon(y, u, h). \tag{6.23}$$

7 K-convergence for optimal control problems

We proceed to apply the notions above to optimal control problems on networks such as $\Omega_\varepsilon = \Omega \cap \varepsilon\mathcal{F}$. Let $\{A_\varepsilon(x) \in \mathcal{L}(\mathbf{R}^2, \mathbf{R}^2)\}_{\varepsilon\in\mathbf{N}}$ be a family of μ_ε-measurable square matrix such that for every $\varepsilon \in E$

$$\alpha \parallel \xi \parallel^2 \leq (A_\varepsilon(x)\xi, \xi) \leq \alpha^{-1} \parallel \xi \parallel^2 \qquad \mu_\varepsilon - \text{ a.e. in } \Omega, \tag{7.24}$$

where $\alpha > 0$ is some constant that is dependent on ε.

Hereinafter we suppose that $\partial\Omega = \Gamma_1 \cup \Gamma_2$ and Γ_1-part of the boundary belongs to the grid G for every $\varepsilon \in E$. We consider the following elliptic problem on a graph and its variational interpretation:

For every $u \in L^2(\Omega,\, d\mu_\varepsilon)$ and $h \in L^2(\Omega,\, d\mu^S_\varepsilon)$, find a state $y \in V(\Omega, \Gamma_2, d\mu_\varepsilon)$ such that

$$\begin{cases} -\text{div}(A_\varepsilon(x)\nabla y) + \alpha \cdot y = u, \\ y = 0 \qquad \text{on } \Gamma_2, \qquad \alpha > 0, \\ \frac{\partial}{\partial\nu_A} y = h \qquad \text{on } \Gamma_1, \end{cases} \tag{7.25}$$

where by $\frac{\partial}{\partial\nu_A} y = h$ we mean the so-called Neumann condition.

We say that $y \in V(\Omega, \Gamma_2, d\mu_\varepsilon)$ is a solution of the problem (7.25) if the following integral identity holds

$$\int_\Omega (A_\varepsilon(x)\nabla y, \nabla\varphi)\, d\mu_\varepsilon + \int_\Omega \lambda\, y\, \varphi\, d\mu_\varepsilon = \int_\Omega u\, \varphi\, d\mu_\varepsilon + \int_{\Gamma_1} h\, \varphi\, d\mu^S_\varepsilon \tag{7.26}$$

for every $\varphi \in C^\infty(\Omega, \Gamma_2)$. Here ∇y is some gradient of y.

First of all we show that for every $u \in L^2(\Omega, d\mu_\varepsilon), h \in L^2(\Gamma_1, d\mu^S_\varepsilon)$ and $\varepsilon \in E$ there exists a unique pair $(y_\varepsilon, \nabla y_\varepsilon)$ that satisfies identity (7.26).

Lemma 7.1

Under the standing assumption with respect to the measures μ and μ^S and with respect to the matrix A_ε there exists a unique function $y_\varepsilon \in V(\Omega, \Gamma_2, d\mu_\varepsilon)$ and a unique gradient of this function $\nabla y_\varepsilon \in \Gamma(y_\varepsilon)$ that satisfy (7.26).

Proof

As follows from Definition 3.2, the set $C^\infty(\Omega, \Gamma_2)$ is dense in the class $V(\Omega, \Gamma_2, d\mu_\varepsilon)$. Therefore the left-hand side of (7.26) induces a new scalar product on

$$L^2(\Omega, d\mu_\varepsilon) \times \left[L^2(\Omega, d\mu_\varepsilon)\right]^2,$$

and the corresponding norm is equivalent to the usual norm in this space. Thus the existence and uniqueness of the solution regarded as the pair $(y_\varepsilon, \nabla y_\varepsilon)$ is an easy consequence of the Lax-Milgram lemma. However, the uniqueness is twofold here: there exists a unique function y_ε of the Sobolev space $V(\Omega, \Gamma_2, d\mu_\varepsilon)$ such that only one of its gradients satisfies the identity (7.26). The uniqueness and existence of such gradient was proved in [29]. It is interesting to note that the gradient ∇y_ε in this identity is defined only by matrix A_ε, and it is not related to the equation itself.

Remark
It is easy to see that the solution of (7.26) satisfies the following estimate

$$\| y_\varepsilon \|_{L^2(\Omega, d\mu_\varepsilon)} + \| \nabla y_\varepsilon \|_{(L^2(\Omega, d\mu_\varepsilon))^2} \leq \alpha^{-1} (\| u \|_{L^2(\Omega, d\mu_\varepsilon)} + \| h \|_{L^2(\Gamma_1, d\mu_\varepsilon^S)}) \tag{7.27}$$

for every $\varepsilon \in E, u \in L^2(\Omega, d\mu_\varepsilon)$ and $h \in L^2(\Gamma_1, d\mu_\varepsilon^S)$. Indeed if we take $\varphi = y_\varepsilon$ as a test function in (7.26) and use Young's inequality we immediately obtain (7.27).

Definition 7.2
We say that the subset Graph (A_ε) *of the space* $V(\Omega, \Gamma_2, d\mu_\varepsilon) \times L^2(\Omega, d\mu_\varepsilon) \times L^2(\Gamma_1, d\mu_\varepsilon^S)$ *is the graph of the control object* (*7.25*) *if*

$$\mathit{Graph}(A_\varepsilon) = \Big\{ (y, u, h) \in V(\Omega, \Gamma_2, d\mu_\varepsilon) \times L^2(\Omega, d\mu_\varepsilon) \times L^2(\Gamma_1, d\mu_\varepsilon^S) \ : \\ \textit{such that } (7.26) \textit{ holds}$$

We need to study the limiting behavior of the sequence of control object (7.25) as $\varepsilon \to 0$.

Definition 7.3
We say that a control object

$$\begin{cases} -div(A^{hom}(x)\nabla y) + \alpha y = u & in\ \mathcal{D}'(\Omega) \\ y = 0 \quad on\ \Gamma_2 \\ \frac{\partial}{\partial \nu_{A^{hom}}} y = u \quad on\ \Gamma_1 \end{cases} \tag{7.28}$$

is the corresponding homogenized control object with respect to the family (7.25) if its graph Graph(A^{hom}) where

$$Graph(A^{hom}) = \Bigg\{ (y,u,h) \in H^1(\Omega) \times L^2(\Omega) \times L^2(\Gamma_1),\ y|_{\Gamma_2} = 0 :$$

$$\int\limits_{\Omega} \Big[(A^{hom}\nabla y, \nabla\varphi) + \alpha y\varphi\Big]\ dx = \int\limits_{\Omega} u\varphi\, dx + \int\limits_{\Gamma_1} h\varphi\, dl,\ \ \forall\, C^\infty(\Omega,\Gamma_2) \Bigg\}$$

is the sequential two-scale K-limit of the sequence $\{Graph(A_\varepsilon); \varepsilon \in E\}$.

K-convergence of the sequence $\{\mathrm{Graph}(A_\varepsilon)\}_{\varepsilon\in E}$ and the recovery of its K-limit follow.

Theorem 7.4

Suppose that the matrix $A_\varepsilon(x)$ in (7.25) is defined as $A_\varepsilon(x) = A(\varepsilon^{-1}x)$, where $A(z)$ is a periodic μ-measurable matrix satisfying condition (7.24). Then for the sequence $\{Graph(A_\varepsilon)\}_{\varepsilon\in E}$ there exists the sequential two-scale K-limit Λ such that

$$\Lambda = Graph(A^{hom}), \tag{7.29}$$

where the limiting matrix A^{hom} is defined as

$$A^{hom}\xi = \int\limits_{\square} A(z)\Big(\xi + v^0(x,z)\Big)\ d\mu(z). \tag{7.30}$$

Here $v^0 \in L^2(\Omega, V_{pot})$ is the solution of the problem

$$\min_{v\in V_{pot}} \int\limits_{\square} (\xi + v, A(\xi + v))\ d\mu = \int\limits_{\square} (\xi + v^0, A(\xi + v^0))\ d\mu. \tag{7.31}$$

Let $\{A_\varepsilon(x) \in \mathcal{L}(\mathbf{R}^2, \mathbf{R}^2)\}_{\varepsilon\in\mathbf{N}}$ be a family of square μ_ε-measurable matrices satisfying the inequality (7.24) for every $\varepsilon \in E$. We also have a K-compactness result for the sequence of graphs of control objects (7.25). As a result, we find that for every $\varepsilon \in E$ the graph of the control object (7.25) is sequentially closed with respect to the product of corresponding weak topologies.

8 Homogenization of the optimal control problems of ε periodic graphs.

We define the optimal control problem on the ε-periodic graph $\Omega_\varepsilon = \Omega \cap \varepsilon\mathcal{F}$ as follows:

$$\min \left\{ I_\varepsilon(y,u,h) := k_1 \int\limits_\Omega (y - z_d)^2 \, d\mu_\varepsilon + k_2 \int\limits_\Omega u^2 \, d\mu_\varepsilon + k_3 \int\limits_{\Gamma_1} h^2 \, d\mu_\varepsilon^S \right\} \tag{8.32}$$

$$y \in V(\Omega, \Gamma_2, d\mu_\varepsilon), \tag{8.33}$$

$$-\mathrm{div}(A_\varepsilon(x)\nabla y) + \alpha y = u, \alpha > 0 \tag{8.34}$$

$$\frac{\partial}{\partial \mu_{A_\varepsilon}} y = h \quad \mu_\varepsilon^S\text{-almost everywhere on } \Gamma_1, \tag{8.35}$$

$$|u| \leq c_u \ \mu_\varepsilon - \text{a.e. in } \Omega, \quad |h| \leq c_h \ \mu_\varepsilon^S - \text{a.e.} \quad \text{on } \Gamma_1, \tag{8.36}$$

Here Ω is an open bounded domain in $\mathbf{R}^2$ with Lipschitz boundary $\partial\Omega = \Gamma_1 \cup \Gamma_2$, Γ_1 belongs to the grid G, c_u and c_h are some positive constants, $z_d \in \mathbf{C}(\bar{\Omega})$ is a given function, μ_ε and μ_ε^S are the *scaling* measures defined by (3.4) and (3.7) respectively and $A_\varepsilon(x) \in \mathcal{L}(\mathbf{R}^2, \mathbf{R}^2)$ is a μ_ε-measurable symmetric matrix satisfying the inequality (7.24); $k_i > 0$.

By $h \in L^2(\Gamma_1, d\mu_\varepsilon^S)$ and $u \in L^2(\Omega, d\mu_\varepsilon)$ we mean *boundary* and *distributed* control functions, respectively. For example, the inclusion $h \in L^2(\Omega_1, d\mu_\varepsilon^S)$ implies that this function is uniquely defined by the respective set of values $\{\eta_1\}$ at the points $K_\varepsilon = \Gamma_1 \cap (\cup \varepsilon\mathcal{M})$. Here $\mathcal{M} = \{M_i, i = 1, 2, \ldots, L\}$ is the set of all vertices on $\Box$ that belong to $\Box^S$. This fact immediately follows from the following property of the measure μ_ε^S: by definition of μ^S (see (3.6)) $\mu_\varepsilon^S(\Gamma_1 \backslash K_\varepsilon) = 0$.

Besides, for every fixed pair $(u,h) \in L^2(\Omega, d\mu_\varepsilon) \times L^2(\Gamma_1, d\mu_\varepsilon^S)$ by a solution y of the boundary value problems (8.33) through (8.35) we mean a function $y \in V(\Omega, \Gamma_2, d\mu_\varepsilon)$ satisfying the integral identity (7.26) for each $\varphi \in C^\infty(\Omega, \Gamma_2)$. In the sequel we will call the set Ξ_ε the set of admissible triplets for each $\varepsilon \in E$. Each of the sets Ξ_ε is sequentially closed with respect to the product of the weak topologies in $L^2(\Omega, d\mu_\varepsilon) \times L^2(\Omega, d\mu_\varepsilon) \times L^2(\Gamma_1, d\mu_\varepsilon^S)$. We thus conclude that under our standing assumptions, we have existence and uniqueness of solutions:

Lemma 8.1

For each $\varepsilon \in E$, the optimal control problems (8.33) through (8.32) has a unique solution—there exists a unique triplet $(y_\varepsilon^0, u_\varepsilon^0, h_\varepsilon^0) \in \Xi_\varepsilon$ such that

$$I_\varepsilon(y_\varepsilon^0, u_\varepsilon^0, h_\varepsilon^0) = \inf_{(y,u,h) \in \Xi_\varepsilon} I_\varepsilon(y,u,h).$$

Note that the uniqueness of this solution is a consequence of the convexity property for Ξ_ε and strictly convexity of the cost functional I_ε.

We are now in the position to state the result:

Theorem 8.2

Under supposition of Lemma 8.1 there exists a unique homogenized optimal control problem, which has the following representation:

$$\begin{cases} -div(A^{hom}(x)\nabla y) + \alpha y = u & in\ \Omega, \\ y = 0 & on\ \Gamma_2 \\ \partial y/\partial \nu_{A^{hom}} = h & on\ \Gamma_1 \\ |u| \le c_u,\ |h| \le c_h & a.e., \\ I_0(y,u,h) = k_1 \int\limits_\Omega (y - z_\partial)^2\, dx + k_2 \int\limits_\Omega u^2\, dx & \\ \qquad\qquad + k_3 \int\limits_{\Gamma_1} h^2\, dl \to min, & \end{cases} \tag{8.37}$$

where

$$\partial y/\partial \nu_{A^{hom}} = \sum_{i,j=1}^{2} a_{ij}^{hom}(x) \frac{\partial y}{\partial x_{ij}} \cos(n, x_i),$$

$\cos(n, x_i)$ *= $i-th$ direction cosine of n, n being the normal at Γ_1 exterior to Ω.*

Moreover, the sequence of optimal solutions for the original problems (8.32) through (8.36) $\{(y^0_\varepsilon, u^0_\varepsilon, h^0_\varepsilon) \in \Xi_\varepsilon\}$ and corresponding minimal values of the cost functional (8.32) satisfy the following variational properties:

$$\lim_{\varepsilon\to 0} \inf_{(y,u,h)\in\Xi_\varepsilon} I_\varepsilon(y,u,h) = \inf_{(y,u,h)\in\Xi_0} (y,u,h), \tag{8.38}$$

$$\begin{cases} y^0_\varepsilon \overset{2}{\to} y^0\ strongly, \\ u^0_\varepsilon \to u^0\ \mu\text{-}strongly, \\ h^0_\varepsilon \to h^0\ \mu^S\text{-}strongly, \end{cases} \tag{8.39}$$

where (y^0, u^0, h^0) is a unique solution of the homogenized problem (8.37), that is,

$$I_0(y^0, u^0, h^0) = \inf_{(y,u,h)\in\Xi_0} I_0(y,u,h). \tag{8.40}$$

9 An example of the homogenization of an optimal control problem on an ε-periodic square grid.

On the domain Ω of $\mathbf{R}^2$ with

$$\Omega = \{x \in \mathbf{R}^2 | 0 < x_1 < a,\ x < x_2 < \gamma(x_1)\},$$

$$\gamma \in C^\infty([0,a]),\ 0 < \gamma_0 = \min_{y_1\in[0,a]} \gamma(x_1),$$

we consider the ε-periodic square grid $\varepsilon\mathcal{F}$ with the cell of periodicity $\varepsilon\square$. Here the set $\square = [0,1)^2$ contains the cross-structure. Let $\partial\Omega = \Gamma_1 \cup \Gamma_2$, where $\Gamma_1 = \{x \in \bar{\Omega} | x_2 = 0,\ 0 < x_1 < a\}$. We begin with some standard notations on graphs (see [22]). Let $V^\varepsilon = \{v_J : J \in \mathcal{J}_\varepsilon\}$ be the set of vertices of our ε-periodic graph (grid) Ω_ε, let $E^\varepsilon = \{e_i : i \in \dot{\mathcal{I}}_\varepsilon\}$ be the index set of corresponding edges. Here by $\mathcal{J}_\varepsilon$ and $\dot{\mathcal{I}}_\varepsilon$ we denote the index sets for vertices and edges, respectively. For the given vertex v_J we consider the set of edges that are incident at v_J. The corresponding set of edge indices is denoted by

$$\dot{\mathcal{I}}_J = \{i \in \dot{\mathcal{I}}_\varepsilon : e_i \text{ is incident at } v_J\}.$$

The cardinality of $\dot{\mathcal{I}}_J$ is the edge degree at v_J, that is, $d_J = |\dot{\mathcal{I}}_J|$. It is easy to see that $d_J \leq 4$ in our case.

Note that every edge e_i on the graph Ω_ε can be parameterized by $x \in [0, l_i]$ where $l_i \leq \varepsilon/2$ denotes the length of the edge e_i. With every edge e_i we will associate a so-called *state-function*

$$y_i : [0, l_i] \to \mathbf{R}^1,\ i \in \dot{\mathcal{I}}_\varepsilon.$$

Because we are going to consider an optimal control problem on a bounded periodic graph, it is necessary to specify boundary and so-called transmission conditions at the vertices V^ε of Ω_ε. To this end we subdivide the set of vertices (nodes) as follows:

$$V^\varepsilon = V_S^\varepsilon \cup V_M^\varepsilon,$$

where V_S^ε denotes the set of simple modes such that $d_J = 1$ and V_M^ε signifies the set of multiple nodes where $4 \geq d_J > 1$. The set of simple nodes V_S^ε will be divided as

$$V^\varepsilon = V_{\Gamma_1}^\varepsilon \cup V_{\Gamma_2}^\varepsilon,$$

where $V_{\Gamma_1}^\varepsilon$ represents the set of simple nodes belonging to the Γ_1-boundary and $V_{\Gamma_2}^\varepsilon$ signifies those simple nodes that belong to the Γ_2-boundary. It is easy to see that in the case of the ε-periodic grid on Ω, there is not any simple node lying in the interior of the domain Ω. Further, we will look at the set $V_{\Gamma_1}^\varepsilon$ as the set of control-active Neumann nodes and at $V_{\Gamma_2}^\varepsilon$ as the set of nodes with zero Dirichlet conditions.

On all edges e_i we consider the differential operator L_i of the following form

$$L_i y_i = -R_i y_i'' + \alpha y_i,$$

where $R_i \geq \alpha > 0$. Moreover, using the ε-periodic structure of Ω_ε, we will always suppose that for every ε-cell $\varepsilon\square_j$ we have

$$R_i = \beta, \quad R_{i+1} = \gamma, \quad R_{i+2} = \beta, \quad R_{i+3} = \gamma, \tag{9.41}$$

where $\alpha^{-1} \geq \beta, \gamma \geq \alpha > 0$.

We now define the classes of admissible controls U^ε and H^ε where

$$U^\varepsilon = \left\{ \begin{array}{r} u : \Omega_\varepsilon \to \mathbf{R}^1 \ |u|_{e_i} \in L^2(0, l_i); \ |u(x)| \le c_u \\ \text{for almost every } x \in \Omega_\varepsilon, \end{array} \right\} \tag{9.42}$$

$$H^\varepsilon = \left\{ \begin{array}{l} h = (h_1, h_2, \dots, h_{L_\varepsilon}) \in \mathbf{R}^{L_\varepsilon} | L_\varepsilon = |V^\varepsilon_{\Gamma_1}|, \\ \sum\limits_{K=1}^{L_\varepsilon} h_K^2 < \infty, |h_K| \le c_h \end{array} \right\}. \tag{9.43}$$

Here c_u, c_h are some positive constant, by $|V^\varepsilon_{\Gamma_1}|$ we denote that amount of all simple nodes belonging to Γ_1.

Let k_1, k_2, k_3 $(k_i > 0)$ be penalty terms, and let $z_d \in C(\bar{\Omega})$ be a given function. Then we consider the following optimal control problem on the grid Ω_ε

$$-R_i y_i'' + \alpha y_i = u_i, \quad x \in (0, l_i), \quad i \in \dot{\mathcal{I}}_\varepsilon, \tag{9.44}$$

$$y_i(v_J) = 0, \quad i \in \dot{\mathcal{I}}^\varepsilon_J, \quad v_J \in V^\varepsilon_{\Gamma_2}, \tag{9.45}$$

$$\sum_{i \in \mathcal{I}^\varepsilon_J} R_i y_i'(v_J) = 0, \quad v_J \in V^\varepsilon_M, \tag{9.46}$$

$$y_i(v_J) = h_{k(J)}, \quad \forall\, v_J \in V^\varepsilon_{\Gamma_1}, \quad i \in \mathcal{I}^\varepsilon_J; \quad k(J) \in \{1, 2, \dots, L_\varepsilon\}, \tag{9.47}$$

$$u \in U^\varepsilon, \quad h \in H^\varepsilon, \tag{9.48}$$

$$I_\varepsilon(y, u, h) = \sum_{i \in \dot{I}_\varepsilon} k_1 \int\limits_0^{l_i} (y_i - z_d|_{e_i})^2\, dx + k_2 \sum_{i \in \dot{\mathcal{I}}_\varepsilon} \int\limits_0^{l_\varepsilon} (u_i)^2\, dx + k_3 \sum_{k=1}^{L_\varepsilon} h_K^2 \to \min \tag{9.49}$$

Using the property of the sets U^ε and H^ε and invoking the standard arguments it is easy to prove that for every $\varepsilon \in E$ problems (9.44) through (9.49) admit a unique optimal triplet $(y^0_\varepsilon, u^0_\varepsilon, h^0_\varepsilon)$, which can be characterized by some adjoint system.

Our aim is to study the asymptotic behavior of this problem as ε tends to 0. For this we reformulate problems (9.44) through (9.49) in the term of some variational control problem defined on spaces with singular measures.

We introduce the $\square$-periodic Borel measure μ in $\mathbf{R}^2$ as follows.

$$\mu = \frac{1}{2}(\mu_1 + \mu_2 + \mu_3 + \mu_4),$$

where μ_i are the 1-dimensional Lebesgue measures on the corresponding line segments (edges) I_i. Also we define the $\square^S$-periodic Radon measure μ^S in $\mathbf{R}^1$ as $\mu^S = \delta_{(\frac{1}{2}, 0)}$, where $\delta_{(\frac{1}{2}, 0)}$ is the Dirac measure at the point $(\frac{1}{2}, 0) \in \square^S$. It is easy to see that

$$\int\limits_{\square} d\mu = 1 \quad \text{and} \quad \int\limits_{\square^S} d\mu^S = 1.$$

Therefore we may define the scaling measures $\mu_\varepsilon(B) = \varepsilon^2\mu(\varepsilon^{-1}B)$, $\mu^S_\varepsilon(B_1) = \varepsilon\mu^S(\varepsilon^{-1}B_1)$, where B, B_1 are corresponding Borel sets in $\mathbf{R}^2$ and $\mathbf{R}^1$, respectively, each of which converges weakly to the correpondng Lebesgue measure $d\mu_\varepsilon \rightarrow dx$ weakly, $d\mu^S_\varepsilon \rightarrow dl$ weakly. Here dx is the Lebesgue measure in $\mathbf{R}^2$ and dl is the Lebesgue measure on $\mathbf{R}^1$.

Now we define the matrix $A_\varepsilon(x) = A(\varepsilon^{-1}x)$ as follows

$$A(z) = \begin{bmatrix} a_{11}(z) & 0 \\ 0 & a_{22}(z) \end{bmatrix},$$

where

$$a_{11}(z) = \beta \qquad \text{and} \qquad a_{22}(z) = \gamma.$$

It is easy to see that such a defined matrix is symmetric, μ-measurable, and satisfies the property (7.24).

As a result the original optimal control problem can be presented in the form

$$\begin{cases} \int\limits_\Omega [(A(\varepsilon^{-1}x)\nabla y, \nabla\varphi) + \alpha y\varphi]\ d\mu_\varepsilon = \int\limits_\Omega u\varphi\ d\mu_\varepsilon \\ + \int\limits_{\Gamma_1} h\varphi\ d\mu^S_\varepsilon,\ \forall\ \varphi \in C^\infty(\Omega, \Gamma_2), \\ y \in V(\Omega, \Gamma_2, d\mu_\varepsilon), \\ |u| \leq c_u \qquad \mu_\varepsilon - \text{a.e. in } \Omega, \\ |h| \leq c_h \qquad \mu^S_\varepsilon - \text{a.e. on } \Gamma_1, \end{cases} \tag{9.50}$$

$$I_\varepsilon(y,u,h) = k_1 \int\limits_\Omega (y - z_d)^2\, d\mu_\varepsilon + k_2 \int\limits_\Omega u^2\, d\mu_\varepsilon + k_3 \int\limits_{\Gamma_1} h^2\, d\mu^S_\varepsilon \rightarrow \inf$$

Then, due to Theorem 8.2, the control problem (9.50) admits a homogenization, and the limit problem can be recovered in the form

$$\begin{cases} -\text{div}(A^{hom}\nabla y) + \alpha y = u & \text{in } \Omega \\ y = 0 & \text{on } \Gamma_2 \\ \partial y/\partial\nu_{A^{hom}} = h & \text{on } \Gamma_1 \end{cases} \tag{9.51}$$

$$\begin{cases} |u| \leq c_u \text{ a.e. in } \Omega, \quad |h| \leq c_h \text{ a.e. on } \Gamma_1 \\ I_0(y,u,h) = k_1 \int\limits_\Omega (y - z_d)^2\, dx + k_2 \int\limits_\Omega u^2\, dx + k_3 \int\limits_{\Gamma_1} h^2\, dl \rightarrow \min \end{cases} \tag{9.52}$$

The identification of the matrix A^{hom} can be done in different ways, either by definition, that is as the solution of the minimum problem (see [29])

$$(A^{hom}\xi, \xi) = \min_{p \in V_{pot}} (A(\xi + p), \xi + p),$$

or using a more classical method: First, we homogenize the problem on the grid with nonzero thickness in the usual way (see [33], [28]) and then pass to the limit as the thickness goes to zero. As for the second approach, it was shown in ([7]) that in this case the corresponding diagram of homogenization is commutative. We can reproduce this result by our method. However, due to space limitations we refer the reader to [19], and to Mazja and Slutsckij [24] for the direct asymptotic analysis of similar problems.

References

[1] Bensossan, A., Lions, J.L., and Papanicolaou, G., *Asymptotic Analysis for Periodic Structures*, North-Holland, Amsterdam, 1978.

[2] Bouchitté, G. and Fragalà, I. Homogenization of thin structures by two-scale method with respect to measures, *SIAM J. Math. Anal.*, 32/6 (2001), 1198–1226.

[3] Bouchitté, G. and Fragalà, I., Homogenization of elastic thin structures: A measure-fattening approach, *J. Convex Anal.*, 9/2 (2002), 339–362.

[4] Bouchitté, G. and Bellieud, M., Homogenization of a soft elastic material reinforced by fibers, *Asymptotic Anal.*, 32/2 (2002), 153–183.

[5] Canic, S., Lamponi, D., Mikelic, A., and Tambaca, J., *Self-consistent effective equations modeling blood flow in medium-to-large compliant arteries*, preprint 2004.

[6] D'Apice, C., De Maio, U., Kogut, P.I., and Mel'nyk, T.A., *On Homogenization of an Optimal Control Problem of Linear Elliptic Equation in Thick Multi-Structures with Dirichlet and Neumann Boundary Controls*, preprint No. 3, University of Salerno (Italy), DIIMA, 2005, 37 pp.

[7] Chechkin, G., Zhikov, V., Lukkassen, D., and Piatnitski, A., On homogenization of networks and junctions, *J. Asymp. Anal.*, 30/1 (2000), 61–80.

[8] Cioranescu, D., and Murat, F., A strange term coming from nowhere, in the book "Topics in the Math. Modelling of Composite Materials," *Prog. Non-linear Diff. Equ. Appl.*, 31 (1997), 49–93.

[9] Dal Maso, G., *An Introduction to* Γ*-Convergence*, Birkhäuser, Boston, 1993.

[10] Hoppe, R. and Petrova, I., Optimal shape design in biomimetics based on homogenization and adaptivity, *Math. Comput. Simul.*, 65/3 (2004), 257–272.

[11] Kesavan, S. and Saint Jean Paulin, J., Quelques problèmes de contrôle bon marché, *C.R. Acad. Sci. Paris*, t. 332, Série 1 (2001), 67–72.

[12] Kogut, P.I. and Mel'nyk, T.A., Limit analysis of a class of optimal control problems in thick multi-structures, *Problemi Upravlenia & Informatiki*, 2 (2005), 13–37 (in Russian); English transl. in *J. Automation and Information Sciences* (forthcoming).

[13] Kogut, P.I. and Mel'nyk, T.A., Asymptotic analysis of optimal control problems in thick multi-structures. In *Generalized Solutions in Control Problems*, Proceedings of the IFAC Workshop GSCP–2004, Pereslavl-Zalessky, Russia, September 21–29, 2004, General Theory: Distributed Parameter Systems, 265–275.

[14] Kogut, P. and Leugering, G., Homogenization of optimal control problems in variable domains. Principle of the fictitious homogenization, *Asymptotic Anal.*, 26/1 (2001), 37–72.

[15] Kogut, P. and Leugering, G., On S-homogenization of an optimal control problem with control and state constraints, *Z. Anal. Anwend.*, 20/2 (2001), 395–429.

[16] Kogut, P. and Leugering, G., S-homogenization of optimal control problems in Banach spaces, *Math. Nach.*, 233–234 (2002), 141–169.

[17] Kogut, P. and Leugering, G., Asymptotic analysis of state constrained semi-linear optimal control problems, *JOTA*, submitted.

[18] Kogut, P. and Leugering, G., Homogenization of Dirichlet optimal control problems with exact partial controllability constraints, *Asymptotic Analysis*, submitted.

[19] Kogut, P. and Leugering, G., Homogenization of optimal control problems for one-dimensional elliptic equations on periodic graphs, ESAIM COCV, submitted.

[20] Kogut, P. and Leugering, G., Homogenization of optimal control problems on periodic structures, *Proceedings of the IFIP* WG 7.2 Workshop, Houston, Texas, December 2004, submitted.

[21] Lenczner, M. and Senouci-Bereski, G., Homogenization of electrical networks including voltage to voltage amplifiers, *Mathematical Methods in Appl. Science*, 9/6 (1999), 899–932.

[22] Lagnese, J.E. and Leugering, G., *Domain Decomposition Method in Optimal Control of Partial Differential Equations.* LSNM Vol. 148, Birkhäuser, Basel, 2004.

[23] Leugering, G. and Schmidt, E.J.P.G., On the modelling and stabilization of flows in networks of open canals, *SIAM J. on Control and Optimization*, 41 (2002), 164–180.

[24] Mazja, V. and Slutsckij, A., Averayence of a differential operator on thick periodic grid, *Math. Nachr.*, 133 (1987), 107–133.

[25] Panasenko, G.P., Asymptotic solution of the elasticity theory system of equations for lattice and skeletal structures, *Russian Academy of Sci., Sbornik Mathematics*, 75/1 (1993), 85–110.

[26] Panasenko, G.P., Homogenization of Lattice-Like Domains. L-Convergence. Reprint No. 178, Analyse numerique Lyon Saint-Etienne, 1994.

[27] Tartar, L., Quelques remarques sur l'homognisation. *In Functional Analysis and Numerical Analysis.* Proceedings Japan-France Seminar, Ed. Fujiat, Soc. for the Promotion of Science, Japan, 1978, 469–482.

[28] Saint Jean Paulin, J. and Cioranescu, D., Homogenization of reticulated structures, *Applied Mathematical Sciences*, Vol. 136, Springer-Verlag, New York, 1999.

[29] Zhikov, V.V., On an extension of the method of two-scale convergence and its applications, *Sbornik: Mathematics*, 191/7 (2000), 973–1014.

[30] Zhikov, V.V., Weighted Sobolev spaces, *Sb. Mat.*, 189/8 (1998), 27–58.

[31] Zhikov, V.V., Homogenization of elastitic problems on singular structures, *Izvestija: Math.*, 66/2 (2002), 299–365.

[32] Zhikov, V.V., Connectedness and homogenization. Examples of fractal conductivity, *Sbovnik: Math.*, 187/8 (1996), 1109–1147.

[33] Zhikov, V.V., Kozlov, S.M., and Oleinik, O.A., *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, Berlin, 1994.

[34] Zhikov, V.V., On the homogenization technique for variational problems, *Functional Analysis and Its Applications*, 33/1 (1991), 11–24.

CHAPTER 4

Lift and sedimentation of particles in the flow of a viscoelastic liquid in a channel

Giovanni P. Galdi
Department of Mechanical Engineering, University of Pittsburgh, Pittsburgh, Pennsylvania

Vincent Heuveline
Institute of Applied Mathematics, University of Karlsruhe, Germany

1 Introduction

As is well known, the motion of small particles in a viscous liquid represents one of the main focuses of engineering research (see, e.g., [30,40] and the references cited therein). Studies on particle–liquid interaction cover a wide range of applications, including manufacturing of short-fiber composites [2,35], separation of macromolecules by electrophoresis [17,18,48,49], flow-induced microstructures [30], models of blood flow [44], and particle-laden materials [8]. The presence of the particles affects the flow of the liquid, and this, in turn, affects the motion of the particles, so that the problem of determining the flow characteristics is highly coupled. It is just this latter feature that makes any fundamental problem related to liquid–particle interaction a particularly challenging one.

The objective of this paper is to furnish some contributions to two problems related to the motion of symmetric particles in a viscoelastic fluid bounded by two parallel walls. Specifically, in the first problem (Problem A) a sphere is moving under the gradient of shear generated by a unidirectional two-dimensional Poiseuille flow of a viscoelastic liquid in a horizontal channel. In the second problem (Problem B) a cylinder is sedimenting in a vertical channel under the action of gravity. In both problems the setting is two-dimensional and the viscoelastic liquid is taken to be a second-order liquid model [28]. Our main goal is to evaluate, at first order in a suitable Reynolds number and in the Weissenberg number, the equilibrium position of the sphere and of the cylinder with respect to one of the walls, and its translational and angular velocities. Moreover, unlike previous contributions to the problem (see below in this Introduction), in Problem A the sphere is allowed to have a nonzero (negative) buoyancy and we investigate the dependence of the equilibria on the (effective) mass of the sphere.

In order to better describe our results and methods, we would like to give the main motivation and the relevant literature related to both problems. We begin with Problem A.

The study of particle migration and lift was greatly influenced by the experiments of Segrè and Sibelberg [46,47] in a purely Newtonian liquid. These authors studied the migration of neutrally buoyant spheres in pipe flows at a Reynolds number between 2 and 700, and found that the spheres migrate away from the wall and centerline and cluster at about 0.6 of a pipe radius. Since the publication of this paper, several investigations have been made to explain inertial lift of particles in viscous shear flow, and in particular the Segrè-Silberberg effect. The first significant contribution is due to Saffmann [43]. In this article a sphere is moving by *prescribed* translational, $\boldsymbol{U}$, and angular, $\boldsymbol{\omega}$, velocities in an *unbounded* Navier-Stokes pure shear flow, when $|\boldsymbol{U}|$, $|\boldsymbol{\omega}|$ and the magnitude of the gradient of shear are sufficiently "small." However, the results are not satisfactory for two main reasons. First, because of absence of walls, they are not able to describe the Segrè-Silberberg effect; second, the slip translational velocity and slip angular velocity of the particle, which are in principle functionals of the solutions, are prescribed quantities. These restrictions were removed by Ho and Leal [23] who, in the wake of the work of Cox and Brenner [10], considered the problem of calculating the equilibrium height of a *neutrally buoyant* sphere in a two-dimensional unidirectional (Poiseuille and Couette) flow of a Navier-Stokes liquid in a channel, without prescribing the motion of the sphere. However, their results require, again, that the channel Reynolds number $Re \ll 1$ and $Re \ll (a/d)^2$, where a is the radius of the sphere and d the channel width. The equilibrium height predicted by Ho and Leal are in agreement with the experiments of [46] and [47]. Their results are based upon the assumption that the solution (if it exists at all) can be expanded around a Stokes solution. The same problem for a *purely viscoelastic* liquid, that is, in absence of inertia, was successively studied in [24]. In [24] it is shown that, for a "small" Weissenberg number, the sphere tends to move toward the center of the channel, in agreement with the experimental result of [32]. More refined results along the same lines of Ho and Leal were given by Vasseur and Cox [51], again by similar formal expansions. More recently, Schonberg and Hinch [45] analyzed the lift on a neutrally buoyant small sphere in a plane Poiseuille flow for the case when the channel Reynolds number is of order unity. The equilibrium position is found to move toward the wall as the Reynolds number increases. The same problem for neutrally buoyant and non-neutrally buoyant small spheres has been studied by Asmolov [3], for large channel Reynolds numbers, but the motion of the sphere is prescribed.

Recent 3-D studies by Kurose and Komori [34] and Cherukat et al. [9] focus on lift and drag on a stationary sphere in a linear (not Poiseuille) shear flow. Issues related to equilibrium positions do not arise in these studies. Finally, direct numerical simulation of the lift-off of a sphere in a viscoelastic liquid near a horizontal wall was given by Patankar et al. [39].

Even though the above results represent a first significant attempt toward understanding the mathematical theory of lift, they are incomplete regarding the following issues. (i) In [3], [43] and [45], it is required that the slip translational velocity and the angular slip velocity of the particle, which are in principle functionals of the solutions, are *prescribed* quantities; (ii) Equilibrium when the motion of the sphere is *not* prescribed and is analyzed only for *neutrally buoyant* particles and for *indefinitely small* channel Reynolds number [23]. (iii) There are *no* results available for the important situation of the equilibrium of a heavier-than-fluid sphere, when the motion of the sphere is *not* prescribed. Finally, (iv) *no* results are available in the case of a viscoelastic liquid in the presence of inertia.

In the first problem treated in this paper (Problem A, see Section 3) we address all the above issues and furnish answers, at least at the first order in a suitable (gravity) Reynolds number, $\mathcal{R}$, and in the Weissenberg number, We. The strategy we pursue is the following. We *fix* the position, h, of the center of the sphere and find a corresponding solution to the steady motion of the system sphere-liquid. Due to the fact that we keep h fixed, the translational velocity of the sphere that we thus obtain need not be directed parallel to the wall of the channel and may have, in principle, a nonzero component, U_2, directed perpendicular to the wall. By using the Lorentz reciprocal theorem [23], we show that $U_2 = F(\boldsymbol{v}, \mathcal{R}, We, Fr, h, \alpha)$, where F is an appropriate function of the velocity field of the liquid, $\boldsymbol{v}$, of the Reynolds, Weissenberg, and Froude number, Fr, of h and of the buoyancy, α. The equilibrium position $h = h_e$ will be the one that makes $F = 0$. We next show that $U_2 = \mathcal{R}(1+E)U_2^{(1)} + O(\mathcal{R}^2 + We^2)$ and that

$$F = \mathcal{R}\frac{Fr^2}{\mathcal{T}_2(h)}[\mathcal{G}(h) + E\mathcal{G}_V(h) - K] + O(\mathcal{R}^2 + We^2),$$

where $\mathcal{T}_2(h)(> 0)$, $\mathcal{G}(h)$, and $\mathcal{G}_V(h)$ are functions of h only, $E = \frac{We}{\mathcal{R}}$ is the elasticity number, and $K = \frac{\alpha}{Fr^2}$. Therefore, at first order in $\mathcal{R}$, We, we have

$$U_2^{(1)} = \frac{Fr^2}{(1+E)\mathcal{T}_2(h)}[\mathcal{G}(h) + E\mathcal{G}_V(h) - K]. \tag{1.1}$$

The functions $\mathcal{G}$ and $\mathcal{G}_V$ represent the contribution to the lift due to the inertia and the elasticity of the liquid, respectively. The equilibrium position h_e is obtained by imposing $U_2^{(1)} = 0$ and it is found to be a function of

E and K. The function $\mathcal{G}(h) + E\mathcal{G}_V(h) - K$ is evaluated numerically and we find the following results (Section 3). Consider first the purely Newtonian case $E = 0$. Then, if $\alpha = 0$ (zero buoyancy), there are only two stable equilibrium positions and they are located symmetrically with respect to the middle of the channel. The location l of these positions, evaluated from the middle of the channel, depends on the (scaled to the thickness of the channel) radius R of the sphere, and it ranges from l ∼0.46 for $R = 0.005$ to l ∼4.1 for $R = 0.1$. This result is in good agreement with the experiments of Segrè and Sibelberg [46], [47], which show $l \sim 0.6$. In fact, the trend of our findings gives $l \to 0.6$ as $R \to 0$. If $\alpha > 0$ (heavier-than-liquid spheres) we found one locally stable equilibrium in the lower half of the channel, $h^{(1)}$, and another locally stable equilibrium in the upper half, $h^{(3)}$. These positions are no longer symmetric with respect to the middle of the channel. Moreover, while $h^{(1)}$ always exists, $h^{(3)}$ exists only if

$$\alpha/Fr^2 < 0.00134. \tag{1.2}$$

The dynamics of the spheres is then studied by solving numerically the differential Equation (1.1) with $U_2^{(1)} = dh/dt$. We found that the position $h^{(3)}$ exists and it is locally stable. However, if (1.2) holds, then both $h^{(1)}$ and $h^{(3)}$ are local attractors. However, if

$$\alpha/Fr^2 > 0.00134, \tag{1.3}$$

the particle will jump to the position $h^{(1)}$ on the lower half of the channel, and that becomes a global attractor. This fact has a very simple interpretation from the physical point of view. In fact, for a given flow rate, a particle can stay in equilibrium in the upper half of the channel if it is not too heavy, that is, if there is enough lift. Otherwise, the particle will fall down to the lower equilibrium. If $E > 0$, so that the liquid is viscoelastic, what we find is, essentially, that as we increase the elasticity number, the stable equilibrium positions $h^{(1)}$ and $h^{(3)}$ will both move toward the center of the channel and eventually, when E reaches a critical value depending on α and Fr, they will both collapse into the position $h^{(2)}$ in the middle of the channel, which will then become the only stable equilibrium and global attractor. So in other words, the elasticity of the liquid tends to move the sphere toward the center of the channel, no matter if the buoyancy is zero or not.

Next we shall discuss Problem B. This problem concerns some of the fundamental aspects of particle sedimentation in the presence of walls. Understanding the motion of particles settling near walls in a liquid is not only of fundamental theoretical interest, but it is also of importance in many industrial processes involving particle-laden materials [8]. Characteristic examples include falling-ball viscometry, flow of slurries [30], and coating processes for thin films [41]. We shall be concerned with sedimentation of spheres. When a homogeneous sphere settles near a rigid wall, its terminal state depends on the physical properties of the sphere and of the liquid. In particular, if the liquid is Newtonian, and the inertial effects are not too large, the sphere will move *away* from the wall to reach a steady state characterized by the following parameters: translational velocity, $\boldsymbol{U}$, angular velocity, $\boldsymbol{\omega}$, and distance, h_{eq} from the wall. On the other extreme, for a viscoelastic liquid with negligible inertia, the sphere will move *toward* the wall, with some other values of the characteristic parameters; see, for example, [30]. Theoretical studies on the motion of spheres in the presence of walls, under different flow conditions, have been the object of many papers; see [5,9–11,23,24,37,51] and the references cited therein. The specific case of a sedimenting sphere, with the objective of giving a quantitative explanation of the above-mentioned phenomenon, has been investigated both for Newtonian [51], and non-Newtonian [5], liquids at small Reynolds and Weissenberg numbers. Even though these results represent a first significant attempt toward understanding the mathematical theory of the wall effect on a sedimenting sphere, they are partially incomplete for the following reasons. (i) They are focused only on *some* aspects of the phenomenon, like the evaluation of the lateral lift force on the sphere [51], or the terminal velocity of the sphere [5]. Actually, in [51] the velocity of the sphere is *prescribed*, while in [5] the distance from the sphere to the wall is *prescribed*; (ii) In the viscoelastic case, they are obtained by assuming that the sphere moves in the presence of only one wall [5]; (iii) They are all based on (formal) expansions of the velocity and pressure fields, like *inner-outer expansion* in the case of Navier-Stokes liquids, [51], and also expansion in the Weissenberg number, in the case of a viscoelastic liquid [5].

Another goal of this chapter is to address the above issues and to give answers at the first order in Reynolds and Weissenberg numbers. Specifically, we study the equilibrium terminal states of a sphere sedimenting in a viscoelastic liquid (second order) bounded by two vertical planes. The method we use is essentially the one that we have previously described for Problem A. Specifically, we eventually show (Section 4.2) that, at first order, the velocity of the sphere, $U_2^{(1)}$, orthogonal to the walls, is given by the following formula

$$U_2^{(1)} = \frac{\alpha^2}{(1+E)\mathcal{T}_2(h)}[\mathcal{G}^{(v)}(h) + E\mathcal{G}_V^{(v)}(h)], \tag{1.4}$$

where $\mathcal{G}^{(v)}(h)$ and $\mathcal{G}_V^{(v)}(h)$ are the contributions to the lateral lift due to inertial and elastic effects of the liquid, respectively. The equilibrium position h_{eq} is obtained by imposing $U_2^{(1)} = 0$, and we find it to be a function of E.

The study of equilibrium and stability is performed in Section 4. We find, in particular, that if $E = 0$ (purely Newtonian liquid), then there is only one $h_{eq} = h^{(1)}$ and it is situated in the middle of the channel. A study of the trajectories shows that this position is a global attractor. As soon as we increase E from zero to a positive value, we find a critical value E_c, say, below which inertial effects are predominant and $h^{(2)}$ is still a global attractor. However, if we pass E_c, two (symmetric) new locally stable positions appear, $h^{(1)}$ and $h^{(3)}$, while $h^{(2)}$ becomes unstable. As in Problem A, in Problem B it is also found that the drag at first order is zero. This implies that the translational and angular velocity of the sphere coincide with the analogous quantities in the Stokes approximation.

2 Problem formulation and analytical preliminaries

Our objective in this paper is to investigate equilibrium positions and the velocity of a disk moving in the flow of a viscoelastic liquid under two different physical situations. The liquid is modeled by a second-order liquid model [28]. In the first problem, Problem A (see Figure 4.1), the disk is subject to a shear (Poiseuille) flow in a horizontal channel. In the second problem, Problem B (see Figure 4.2), the disk is falling under the action of gravity in a vertical channel. In both problems we are interested in *steady motions* of the system liquid–disk, that is, the translational velocity $\boldsymbol{U}$ and the angular velocity $\boldsymbol{\omega}$ of the disk Σ are constant in time, and the motion of the fluid as seen from a frame $\mathcal{I}$ attached to Σ and moving with velocity $\boldsymbol{U}$ is *independent of time.* Thus, for a steady motion to occur, it is clear that the (unknown) velocity $\boldsymbol{U}$ of Σ must be directed along the channel walls, which we will take, without loss of generality, parallel to the x_1-axis of the frame $\mathcal{I}$. We also take the origin of $\mathcal{I}$ coinciding with the center of Σ and use the thickness d of the layer as a length scale. In Problem A it is convenient

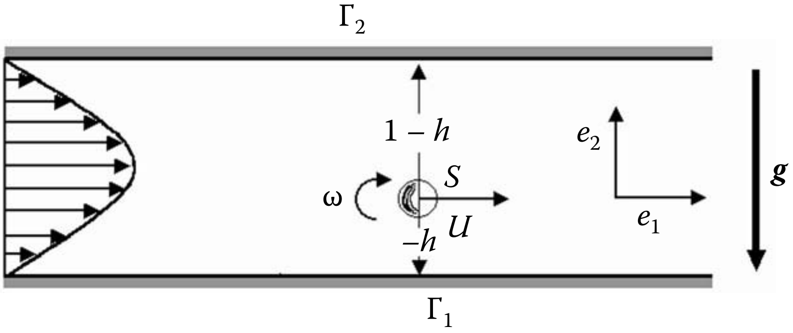


FIGURE 4.1 Schematic view of the system for Problem A.

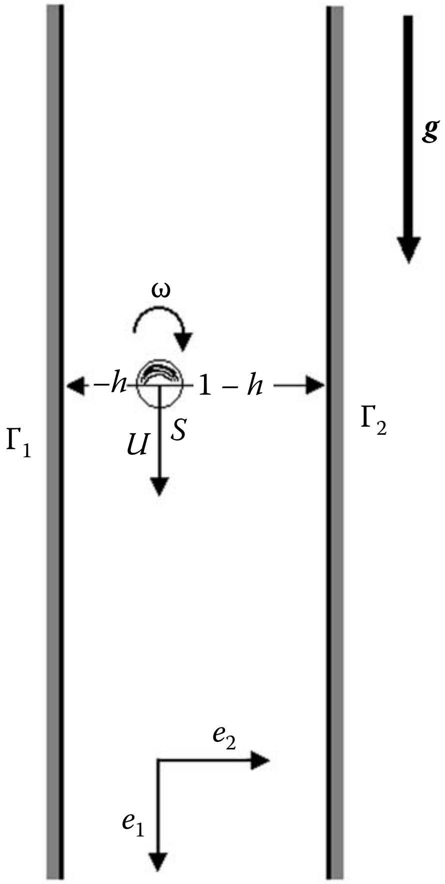


FIGURE 4.2 Schematic view of the system for Problem B.

to define $V \equiv \sqrt{gd}$ as velocity scale, with g acceleration of gravity. We then find that the Poiseuille flow $(\boldsymbol{v}_0, p_0)$, as seen from $\mathcal{I}$, assumes the following dimensionless form:

$$\boldsymbol{v}_0(x_2; h) = -6Fr[h^2 + h(2x_2 - 1) + x_2(x_2 - 1)]\boldsymbol{e}_1 \equiv Fr\, f(x_2, h)\boldsymbol{e}_1$$

$$p_0 = -12\frac{\Phi}{\mu d^2}x_1, \tag{2.5}$$

where $Fr = \frac{\Phi}{Vd}$, is the Froude number, Φ is the given flow rate (which, without loss, we assume to be positive), μ is the shear viscosity, and $-h$ and $1 - h$ are the x_2-coordinates of the walls Γ_1 and Γ_2 of the channel (see Figure 4.1). In Problem B we may take as V any scale velocity. Problems A and B can then be formulated, in a combined way, as follows. Find

$\{\boldsymbol{v}, p, \omega, U, h\}$, satisfying the following dimensionless equations:

$$\left.\begin{aligned} &\operatorname{div} \boldsymbol{T}(\boldsymbol{v}, P) = \mathcal{R}\boldsymbol{v} \cdot \operatorname{grad} \boldsymbol{v} \\ &\operatorname{div} \boldsymbol{v} = 0 \end{aligned}\right\} \quad \text{in } \Omega$$

$$\begin{gathered} \boldsymbol{v}\,|_S = \omega \boldsymbol{e}_3 \times \boldsymbol{x}, \quad \boldsymbol{v}\,|_{\Gamma_1} = \boldsymbol{v}\,|_{\Gamma_2} = -\boldsymbol{U} \\ \lim_{|x_1| \to \infty} (\boldsymbol{v}(x_1, x_2) - \boldsymbol{v}_0(x_2; h) + \boldsymbol{U}) = 0 \\ \int_{-h}^{1-h} v_1(x_1, x_2) dx_2 = \mathit{Fr} - \boldsymbol{U} \cdot \boldsymbol{e}_1 \\ \int_S \boldsymbol{T}(\boldsymbol{v}, P) \cdot \boldsymbol{n} = \boldsymbol{G}, \quad \int_S \boldsymbol{x} \times \boldsymbol{T}(\boldsymbol{v}, P) \cdot \boldsymbol{n} = \boldsymbol{0}. \end{gathered} \tag{2.6}$$

Here Ω is the region occupied by the fluid, S is the surface of Σ, and $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{v}, P)$ is the Cauchy stress tensor given by

$$\boldsymbol{T}(\boldsymbol{v}, P) = \boldsymbol{T}_N(\boldsymbol{v}, P) - \mathit{We}\boldsymbol{S}(\boldsymbol{v}) \tag{2.7}$$

where

$$\boldsymbol{T}_N(\boldsymbol{v}, P) = 2\boldsymbol{D} - \operatorname{grad} P, \; \boldsymbol{D} = \boldsymbol{D}(\boldsymbol{v}) := \tfrac{1}{2}(\boldsymbol{L} + \boldsymbol{L}^\top), \; \boldsymbol{L} = \boldsymbol{L}(\boldsymbol{v}) = \operatorname{grad} \boldsymbol{v},$$

$$\boldsymbol{S}(\boldsymbol{v}) = 2\boldsymbol{u} \cdot \operatorname{grad} \boldsymbol{D} + 2\boldsymbol{D} \cdot \boldsymbol{L}^\mathsf{T} + 2\boldsymbol{L} \cdot \boldsymbol{D} + 4\varepsilon \boldsymbol{D} \cdot \boldsymbol{D}.$$

Moreover, $P = p - \mathcal{R}x_2$, $\mathcal{R} = \frac{\rho V d}{\mu}$, and ρ is fluid density, $\mathit{We} = \frac{-\alpha_1 V}{d\mu}$, $\varepsilon = \alpha_2/\alpha_1$ where $\alpha_1 (< 0)$ and α_2 are the so-called *quadratic constants* related to the normal stress coefficients Ψ_1 and Ψ_2 by the formulas $\alpha_1 = -\frac{1}{2}\Psi_1$, $\alpha_2 = \Psi_1 + \Psi_2$; see [28], Chapter 17. In Problem A, $\mathit{Fr} > 0$ and $\boldsymbol{G} = -\mathcal{R}\alpha \boldsymbol{e}_2$ where $\alpha = \pi(R/d)^2(\rho_s/\rho - 1)$ and R, ρ_s are the radius and the density of the disk, respectively. In Problem B, $\mathit{Fr} = 0$ and $\boldsymbol{G} = \alpha \boldsymbol{e}_1$.[1]

Our first objective is to prove the existence and uniqueness of a solution to Problem (2.6) for any *given* $h \in (a, 1-a)$, where $a := R/d$ $(< 1/2)$. To this end, we denote by $W^{m,q}(\Omega)$, $m \geq 0$, $q \geq 1$, the usual Sobolev space of functions in Ω having all derivatives up to the order m inclusive that are summable to the q-th power in Ω; see, e.g., [1]. We denote by $\|\cdot\|_{m,p}$ the associated norm ($\|\cdot\|_{0,q} \equiv \|\cdot\|_q$).

The following theorem holds.

Theorem 2.1

Let $h \in (a, 1-a)$ be given and let $\mathit{Fr} \geq 0$, $\alpha \geq 0$. Then there exists $\mathcal{R}_0 > 0$ and $\mathit{We} > 0$ such that if $\mathcal{R} < \mathcal{R}_0$ and $\mathit{We} < \mathit{We}_0$, problem (2.6) admits one

[1] We shall assume throughout that $\rho_s \geq \rho$ so that $\alpha \geq 0$.

and only one solution $\boldsymbol{v}$, P, $\boldsymbol{U}$, ω *such that*

$$(\boldsymbol{v}-\boldsymbol{v}_0+\boldsymbol{U})\in W^{3,2}(\Omega),\quad (P-p_0)\in W^{1,2}(\Omega).$$

Moreover, let $\boldsymbol{v}_s, P_s, \boldsymbol{U}_s, \omega_s$ *be the solution to (2.6) corresponding to* $\mathcal{R}=We=0$*. Then*

$$\begin{aligned}\boldsymbol{v}&=\boldsymbol{v}_s+(\mathcal{R}+We)\boldsymbol{v}(1)+(\mathcal{R}^2+We^2)v^{(2)},\quad P=P_s+(\mathcal{R}+We)P^{(1)}\\&\quad+(\mathcal{R}^2+We^2)P^{(2)}\\\boldsymbol{U}&=\boldsymbol{U}_s+(\mathcal{R}+We)\boldsymbol{U}^{(1)}+(\mathcal{R}^2+We^2)\boldsymbol{U}^{(2)},\quad \omega=\omega_s+(\mathcal{R}+We)\omega^{(1)}\\&\quad+(\mathcal{R}^2+We^2)\omega^{(2)}\end{aligned}\tag{2.8}$$

$$\sum_{i=1}^{2}\left(|\boldsymbol{U}^{(i)}|+|\omega^{(i)}|+|\sup_{x\in\Omega}|\boldsymbol{v}^{(i)}(x)|+\|\operatorname{grad}\boldsymbol{v}^{(i)}\|_{1,2}+\|\operatorname{grad}P^{(i)}\|_{1,2}\right)\le C,\tag{2.9}$$

with C *depending only on* Ω, Fr, $\mathcal{R}_0$ *and* We_0*. Finally, if* $We=0$ *(purely Navier-Stokes liquid), then* $\boldsymbol{v}$, P, $\boldsymbol{U}$ *and* ω *are real-analytic in* $\mathcal{R}$ *and their series are absolutely convergent in* $W^{2,2}(\Omega)$, $W^{1,2}(\Omega)$, $\mathbb{R}^2$ *and* $\mathbb{R}$, *respectively.*

The proof of the first part of this theorem, including the estimate (2.9), can be obtained by following exactly the methods of proof used in [16] and, therefore, it will be omitted. However, in the appendix to this chapter, we shall furnish a proof of the analyticity property.

Remark
It is worth emphasizing that the *translational velocity* $\boldsymbol{U}$ of solutions given in Theorem 2.1 need *not* be directed along the walls Γ_1, Γ_2; that is, these solutions may have $U_2\neq 0$. Our next objective is to find conditions under which $U_2=0$. Such conditions will give precisely the possible values for h.

To reach this goal, we rewrite the last two equations in (2.6) in an equivalent form. We introduce the *auxiliary fields* $\{\boldsymbol{w}^{(i)},\pi^{(i)}\}$ defined as solutions to the following linear problems ($i=1,2,3$)

$$\begin{gathered}\left.\begin{aligned}&\operatorname{div}\boldsymbol{T}_N(\boldsymbol{w}^{(i)},\pi^{(i)})=0\\&\operatorname{div}\boldsymbol{w}^{(i)}=0\end{aligned}\right\}\text{ in }\Omega\\\boldsymbol{w}^{(i)}|_S=\boldsymbol{\beta}_i,\quad \boldsymbol{w}^{(i)}|_{\Gamma_1}=\boldsymbol{w}^{(i)}|_{\Gamma_2}=0\\\lim_{|x_1|\to\infty}\boldsymbol{w}^{(i)}(x_1,x_2)=0.\end{gathered}\tag{2.10}$$

where $\boldsymbol{\beta}_i = \boldsymbol{e}_i$ for $i = 1, 2$, and $\boldsymbol{\beta}_3 = \boldsymbol{e}_3 \times \boldsymbol{x}$. Note that all fields $\boldsymbol{w}, \pi$ depend only on h. Furthermore, $\boldsymbol{w}^{(i)}, p^{(i)}$ and all corresponding derivatives decay exponentially fast to zero as $|x| \to \infty$ [14, Chapter XI]. In view of the symmetry $x_1 \to -x_1$ of problems (2.10), it is easily shown that the following relations hold:

$$\begin{gathered}\int_S \boldsymbol{T}_N(\boldsymbol{w}^{(2)}, \pi^{(2)}) \cdot \boldsymbol{n} = \mathcal{T}_2(h)\boldsymbol{e}_2, \quad \int_S \boldsymbol{x} \times \boldsymbol{T}_N(\boldsymbol{w}^{(2)}, \pi^{(2)}) \cdot \boldsymbol{n} = 0,\\ \int_S \boldsymbol{T}_N(\boldsymbol{w}^{(3)}, \pi^{(3)}) \cdot \boldsymbol{n} = \mathcal{T}_3(h)\boldsymbol{e}_1, \quad \int_S \boldsymbol{T}_N(\boldsymbol{w}^{(1)}, \pi^{(1)}) \cdot \boldsymbol{n} = \mathcal{T}_1(h)\boldsymbol{e}_1,\\ \int_S \boldsymbol{v}_0 \cdot \boldsymbol{T}_N(\boldsymbol{w}^{(1)}, \pi^{(1)}) \cdot \boldsymbol{n} = Fr\mathcal{F}_1(h)\boldsymbol{e}_1, \quad \int_S \boldsymbol{v}_0 \cdot \boldsymbol{T}_N(\boldsymbol{w}^{(2)}, \pi^{(2)}) \cdot \boldsymbol{n} = 0,\\ \int_S \boldsymbol{v}_0 \cdot \boldsymbol{T}_N(\boldsymbol{w}^{(3)}, \pi^{(3)}) \cdot \boldsymbol{n} = Fr\mathcal{F}_3(h)\boldsymbol{e}_1,\end{gathered} \tag{2.11}$$

where $\mathcal{T}_i(h)$, $i = 1, 2, 3$, and $\mathcal{F}_i(h)$, $i = 1, 3$, depend only on h. Moreover, obviously,

$$\int_S \boldsymbol{x} \times \boldsymbol{T}_N(\boldsymbol{w}^{(1)}, \pi^{(1)}) \cdot \boldsymbol{n} = \mathcal{R}_1(h)\boldsymbol{e}_3, \quad \int_S \boldsymbol{x} \times \boldsymbol{T}_N(\boldsymbol{w}^{(3)}, \pi^{(3)}) \cdot \boldsymbol{n} = \mathcal{R}_2(h)\boldsymbol{e}_3, \tag{2.12}$$

where, again, $\mathcal{R}_i(h)$, $i = 1, 3$, depend only on the location of the disk. It is easy to see that

$$\mathcal{T}_3(h) = \mathcal{R}_1(h), \quad \text{for all} > 1 - a > h > a. \tag{2.13}$$

Actually, if we multiply $(2.10)_1$ with $i = 3$ by $\boldsymbol{w}^{(1)}$, integrate by parts over Ω and take into account (2.12), we obtain

$$\mathcal{T}_3(h) = 2\int_\Omega \boldsymbol{D}(\boldsymbol{w}^{(1)}) : \boldsymbol{D}(\boldsymbol{w}^{(3)}).$$

Likewise, if we multiply $(2.10)_1$ with $i = 1$ by $\boldsymbol{w}^{(3)}$, integrate by parts over Ω, and take into account (2.12), we obtain

$$\mathcal{R}_1(h) = 2\int_\Omega \boldsymbol{D}(\boldsymbol{w}^{(3)}) : \boldsymbol{D}(\boldsymbol{w}^{(1)}),$$

and (2.13) follows. Moreover, we have that

$$\begin{gathered}\mathcal{T}_1(h) > 0, \mathcal{T}_2(h) > 0, \mathcal{R}_2(h) > 0, \mathcal{T}_1(h)\mathcal{R}_2(h) - \mathcal{R}_1(h)\mathcal{T}_3(h) > 0,\\ \text{for all } 1 - a > h > a.\end{gathered} \tag{2.14}$$

In fact, let $\boldsymbol{w}=\sum_{i=1}^{3}\lambda_i\boldsymbol{w}^{(i)}$, $\Pi=\sum_{i=1}^{3}\lambda_i\pi^{(i)}$, $\lambda_i\in\mathbb{R}$, $i=1,2,3$. From (2.10) we find

$$\left.\begin{aligned}&\operatorname{div}\boldsymbol{T}_N(\boldsymbol{w},\Pi)=0\\&\operatorname{div}\boldsymbol{w}=0\end{aligned}\right\}\ \text{in }\Omega$$

$$\boldsymbol{w}\,|_S=\sum_{i=1}^{3}\lambda_i\boldsymbol{\beta}_i,\quad \boldsymbol{w}\,|_{\Gamma_1}=\boldsymbol{w}\,|_{\Gamma_2}=0 \tag{2.15}$$

$$\lim_{|x_1|\to\infty}\boldsymbol{w}(x_1,x_2)=0.$$

Multiplying $(2.15)_1$ by $\boldsymbol{w}$, integrating by parts over Ω, and taking into account (2.13) and (2.12) we get

$$\lambda_1^2\mathcal{T}_1+\lambda_2^2\mathcal{T}_2+\lambda_3^2\mathcal{R}_2+2\lambda_1\lambda_3\mathcal{T}_3=2\int_\Omega|\boldsymbol{D}(\boldsymbol{w})|^2.$$

Since the right-hand side of this equation is always positive, unless $\lambda_1=\lambda_2=\lambda_3=0$, and the λs are arbitrary, the property (2.14) follows. We now multiply $(2.6)_1$ by $\boldsymbol{w}^{(i)}$, $i=1,2,3$, integrate by parts over Ω and use the asympotic properties of $\boldsymbol{w}^{(i)}$ to obtain ($i=1,2,3$)

$$\begin{aligned}\int_S\boldsymbol{\beta}_i\cdot\boldsymbol{T}(\boldsymbol{v},P)\cdot\boldsymbol{n}&=\mathcal{R}\int_\Omega\boldsymbol{v}\cdot\operatorname{grad}\boldsymbol{v}\cdot\boldsymbol{w}^{(i)}-We\int_\Omega\boldsymbol{S}(\boldsymbol{v}):\boldsymbol{D}(\boldsymbol{w}^{(i)})\\&\quad+2\int_\Omega\boldsymbol{D}(\boldsymbol{v}):\boldsymbol{D}(\boldsymbol{w}^{(i)}).\end{aligned} \tag{2.16}$$

Likewise, multiplying $(2.10)_1$ by $\boldsymbol{v}-\boldsymbol{v}_0+\boldsymbol{U}$ and integrating by parts over Ω we find ($i=1,2,3$)

$$\int_S(\boldsymbol{U}+\omega\boldsymbol{e}_3\times\boldsymbol{x}-\boldsymbol{v}_0)\cdot\boldsymbol{T}_N(\boldsymbol{w}^{(i)},\pi^{(i)})\cdot\boldsymbol{n}=2\int_\Omega\boldsymbol{D}(\boldsymbol{v}):\boldsymbol{D}(\boldsymbol{w}^{(i)}). \tag{2.17}$$

From (2.12), (2.16), and (2.17), one deduces that the last two equations in (2.6) are equivalent to the following ones:

$$\begin{aligned}\omega\mathcal{R}_1(h)+U_1\mathcal{T}_1(h)&=Fr\,\mathcal{F}_1(h)+G_1-\mathcal{R}\int_\Omega\boldsymbol{v}\cdot\operatorname{grad}\boldsymbol{v}\cdot\boldsymbol{w}^{(1)}\\&\quad+We\int_\Omega\boldsymbol{S}(\boldsymbol{v}):\boldsymbol{D}(\boldsymbol{w}^{(1)})\\U_2\mathcal{T}_2(h)&=-\mathcal{R}\int_\Omega\boldsymbol{v}\cdot\operatorname{grad}\boldsymbol{v}\cdot\boldsymbol{w}^{(2)}+We\int_\Omega\boldsymbol{S}(\boldsymbol{v}):\boldsymbol{D}(\boldsymbol{w}^{(2)})\\&\quad+G_2\\\omega\mathcal{R}_2(h)+U_1\mathcal{T}_3(h)&=Fr\,\mathcal{F}_2(h)-\mathcal{R}\int_D\boldsymbol{v}\cdot\operatorname{grad}\boldsymbol{v}\cdot\boldsymbol{w}^{(3)}\\&\quad+We\int_\Omega\boldsymbol{S}(\boldsymbol{v}):\boldsymbol{D}(\boldsymbol{w}^{(3)}).\end{aligned} \tag{2.18}$$

From these equations it is possible to draw a number of conclusions. Consider first Problem A. In such a case, $G_1 = 0$ and $G_2 = -\mathcal{R}\alpha$. Let $\boldsymbol{v}_s$, $\boldsymbol{U}_s$ and ω_s be as in Theorem 2.1. Recalling that $\mathcal{T}_2(h) > 0$, we have $U_{s2} = 0$, and so $\boldsymbol{v}_s = U_{s1}(\boldsymbol{w}^{(1)} - \boldsymbol{e}_1) + \omega_s \boldsymbol{w}^{(3)} + Fr\,\boldsymbol{w}^{(4)}$, where U_{s1} and ω_s solve $(2.19)_{1,3}$ with $\mathcal{R} = We = 0$, namely,

$$\begin{aligned} \omega_s \mathcal{R}_1(h) + U_{s1}\mathcal{T}_1(h) &= Fr\,\mathcal{F}_1(h) \\ \omega_s \mathcal{R}_2(h) + U_{s1}\mathcal{T}_3(h) &= Fr\,\mathcal{F}_2(h) \end{aligned} \tag{2.19}$$

and

$$\begin{gathered} \left.\begin{aligned} &\operatorname{div} \boldsymbol{T}(\boldsymbol{w}^{(4)}, \pi^{(4)}) = 0 \\ &\operatorname{div} \boldsymbol{w}^{(4)} = 0 \end{aligned}\right\} \text{ in } \Omega \\ \boldsymbol{w}^{(4)}\,|_S = 0, \quad \boldsymbol{w}^{(4)}\,|_{\Gamma_1} = \boldsymbol{w}^{(4)}\,|_{\Gamma_2} = 0 \\ \lim_{|x_1|\to\infty} \Big(\boldsymbol{w}^{(4)}(x_1, x_2) - f(x_2, h)\boldsymbol{e}_1\Big) = 0 \end{gathered} \tag{2.20}$$

Notice that by (2.14), Equation (2.19) has one and only one solution that can be expressed as

$$U_{s1} = Fr\,A(h), \quad \omega_s = Fr\,B(h)$$

where A and B are functions of h only. Therefore, the Stokes velocity field $\boldsymbol{v}_s$ can be rewritten as

$$\boldsymbol{v}_s = Fr\,[A(h)(\boldsymbol{w}^{(1)} - \boldsymbol{e}_1) + B(h)\boldsymbol{w}^{(3)} + \boldsymbol{w}^{(4)}] \equiv Fr\,\overline{\boldsymbol{v}}_s.$$

In view of (2.8) and (2.9), from $(2.19)_2$ it follows that

$$(\mathcal{R} + We)U_2^{(1)}\mathcal{T}_2(h) = Fr^2\mathcal{R}\mathcal{G}(h) + Fr^2\,We\,\mathcal{G}_V(h) - \mathcal{R}\alpha + \Lambda \tag{2.21}$$

where

$$\mathcal{G}(h) := \int_\Omega \overline{\boldsymbol{v}}_s \cdot \operatorname{grad}\overline{\boldsymbol{v}}_s \cdot \boldsymbol{w}^{(2)}, \quad \mathcal{G}_V(h) := \int_\Omega \boldsymbol{S}(\overline{\boldsymbol{v}}_s) : \boldsymbol{D}(\boldsymbol{w}^{(2)}) \tag{2.22}$$

and

$$|\Lambda| \le C\,(\mathcal{R} + We)^2, \quad C = C(\Omega, Fr) > 0.$$

Thus, at first order in $\mathcal{R}$ and We, we find that $U_2^{(1)}$ is given by

$$U_2^{(1)} = \frac{Fr^2}{(1+E)\mathcal{T}_2(h)}\,[\mathcal{G}(h) + E\mathcal{G}_V(h) - K] \tag{2.23}$$

where $K = \frac{\alpha}{Fr^2}$ and $E = \frac{We}{\mathcal{R}}$ is the *elasticity number*. Because the equilibrium heights h_{eq} are those at which $U_2^{(1)} = 0$, from (2.23) we deduce that, at first order in $\mathcal{R}$ and We, h_{eq} is the solution to the equation

$$\mathcal{G}(h) + E\mathcal{G}_V(h) - K = 0. \tag{2.24}$$

Notice that the quantities $\mathcal{R}Fr^2\mathcal{G}(h)$ and $WeFr^2\mathcal{G}_V(h)$ represent, at first order in $\mathcal{R}$ and We, the Newtonian and purely viscoelastic lifts, namely, they are the components of the force exerted by the liquid in the direction orthogonal to the translational velocity of the disk. Solutions to Equation (2.24) were computed numerically and they will be discussed in Section 3. Once the values for h_{eq} have been obtained, from $(2.19)_{1,3}$ and with the help of Theorem 2.1, we may calculate the translational and angular velocities of the disk, $U_1^{(1)}$ and $\omega^{(1)}$, at first order in $\mathcal{R}$ and We. In fact, we have

$$\begin{aligned}\omega^{(1)}\mathcal{R}_1(h) + U_1^{(1)}\mathcal{T}_1(h) &= \frac{Fr^2}{1+E}\left[-\int_\Omega \overline{\boldsymbol{v}}_s\cdot\operatorname{grad}\overline{\boldsymbol{v}}_s\cdot\boldsymbol{w}^{(1)}\right.\\ &\qquad\left.+\,E\int_\Omega \boldsymbol{S}(\overline{\boldsymbol{v}}_s):\boldsymbol{D}(\boldsymbol{w}^{(1)})\right]\\ \omega^{(1)}\mathcal{R}_2(h) + U_1^{(1)}\mathcal{T}_3(h) &= \frac{Fr^2}{1+E}\left[-\mathcal{R}\int_D \overline{\boldsymbol{v}}_s\cdot\operatorname{grad}\overline{\boldsymbol{v}}_s\cdot\boldsymbol{w}^{(3)}\right.\\ &\qquad\left.+\,E\int_\Omega \boldsymbol{S}(\overline{\boldsymbol{v}}_s):\boldsymbol{D}(\boldsymbol{w}^{(3)})\right].\end{aligned} \tag{2.25}$$

Let us now consider Problem B. In such a case, we have that $G_1 = \alpha$ and $G_2 = Fr = 0$ in (2.19). Therefore, we still find $U_{s2} = 0$, while $\boldsymbol{v}_s = U_{s1}(\boldsymbol{w}^{(1)} - \boldsymbol{e}_1) + \omega_s\boldsymbol{w}^{(3)}$, with U_{s1} and ω_s solving the following system

$$\begin{aligned}\omega_s\mathcal{R}_1(h) + U_{s1}\mathcal{T}_1(h) &= \alpha\\ \omega_s\mathcal{R}_2(h) + U_{s1}\mathcal{T}_3(h) &= 0.\end{aligned} \tag{2.26}$$

In view of (2.14), Equation (2.26) has one and only one solution, which can be expressed as

$$U_{s1} = \alpha\, A_1(h), \quad \omega_s = \alpha\, B_1(h)$$

where A_1 and B_1 are functions of h only. Therefore, the Stokes velocity field $\boldsymbol{v}_s$ can be rewritten as

$$\boldsymbol{v}_s = \alpha\left[A_1(h)(\boldsymbol{w}^{(1)} - \boldsymbol{e}_1) + B_1(h)\boldsymbol{w}^{(3)}\right] \equiv \alpha\overline{\overline{\boldsymbol{v}}}_s.$$

Employing (2.8) and (2.9), from $(2.19)_2$ we have that

$$(\mathcal{R} + We)U_2^{(1)}\mathcal{T}_2(h) = \alpha^2\mathcal{R}\mathcal{G}^{(v)}(h) + \alpha^2\, We\mathcal{G}_V^{(v)}(h) + \Lambda_1 \tag{2.27}$$

where

$$\mathcal{G}^{(v)}(h) := \int_\Omega \overline{\overline{\boldsymbol{v}}}_s \cdot \operatorname{grad} \overline{\overline{\boldsymbol{v}}}_s \cdot \boldsymbol{w}^{(2)}, \quad \mathcal{G}_V^{(v)}(h) := \int_\Omega \boldsymbol{S}(\overline{\overline{\boldsymbol{v}}}_s) : \boldsymbol{D}(\boldsymbol{w}^{(2)}) \tag{2.28}$$

and

$$|\Lambda_1| \leq C\,(\mathcal{R} + \mathit{We})^2, \quad C = C(\Omega, \mathit{Fr}) > 0.$$

Thus, at first order in $\mathcal{R}$ and We, it follows that U_2' is given by

$$U_2^{(1)} = \frac{\alpha^2}{(1+E)\mathcal{T}_2(h)}\Big[\mathcal{G}^{(v)}(h) + E\mathcal{G}_V^{(v)}(h)\Big]. \tag{2.29}$$

Because the equilibrium positions h_{eq} are those at which $U_2^{(1)} = 0$, from (2.23) we deduce that, at first order in $\mathcal{R}$ and We, h_{eq} is the solution to the equation

$$\mathcal{G}^{(v)}(h) + E\mathcal{G}_V^{(v)}(h) = 0. \tag{2.30}$$

Solutions to Equation (2.30) will be computed numerically and discussed in Section 4. The first-order contributions to translational and angular velocities can be computed from the analog of (2.26), which in such a case becomes

$$\begin{aligned}
\omega^{(1)}\mathcal{R}_1(h) + U_1^{(1)}\mathcal{T}_1(h) &= \frac{\mathit{Fr}^2}{1+E}\Big[-\int_\Omega \overline{\overline{\boldsymbol{v}}}_s \cdot \operatorname{grad} \overline{\overline{\boldsymbol{v}}}_s \cdot \boldsymbol{w}^{(1)} \\
&\qquad + E\int_\Omega \boldsymbol{S}(\overline{\overline{\boldsymbol{v}}}_s) : \boldsymbol{D}(\boldsymbol{w}^{(1)})\Big] \\
\omega^{(1)}\mathcal{R}_2(h) + U_1^{(1)}\mathcal{T}_3(h) &= \frac{\mathit{Fr}^2}{1+E}\Big[-\mathcal{R}\int_D \overline{\overline{\boldsymbol{v}}}_s \cdot \operatorname{grad} \overline{\overline{\boldsymbol{v}}}_s \cdot \boldsymbol{w}^{(3)} \\
&\qquad + E\int_\Omega \boldsymbol{S}(\overline{\overline{\boldsymbol{v}}}_s) : \boldsymbol{D}(\boldsymbol{w}^{(3)})\Big].
\end{aligned} \tag{2.31}$$

3 Numerical solution process

From the numerical point of view, the most important step in our method is the solution of the auxiliary Stokes problems (2.10) and (2.20). The remaining steps are essentially standard postprocessing procedures, which encompass numerical integrations on the disk as well as the solution of scalar ordinary differential equations in order to determine the dynamics of the disk. The main difficulty for the numerical treatment of the Stokes problems is related to the needed setup of unbounded channel. In order to obtain accurate results, it is further important to notice that these Stokes problems must be solved for a large number of disk positions (typically 50 to 100 height positions).

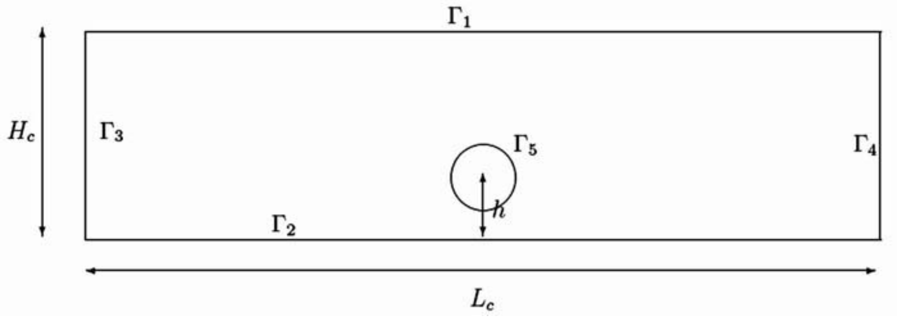


FIGURE 4.3 Description of the computational domain and related needed notations.

In order to solve problems (2.10) and (2.20), we consider a discretization based on conforming mixed finite elements with continuous pressure. In that context, the unbounded channel is replaced by a truncated bounded channel, Ω, which is assumed to be large enough to mimic the properties of the unbounded channel by means of adequate boundary conditions. In our numerical experiments the length of the truncated channel is set to be $L_C = 10$ assuming a scaled channel height $H_C = 1$ (see Figure 4.3).

The considered discretization starts from a variational formulation of Equations $(2.10)_1$ and $(2.20)_1$. Let $W = [H^1(\Omega)]^2 \times L^2(\Omega)$. For $\boldsymbol{w} = \{\boldsymbol{v}, p\} \in W$ and $\phi = \{\varphi, q\} \in W$, we define the semilinear form

$$\mathcal{A}(\boldsymbol{w}; \phi) = \mu(\operatorname{grad} \boldsymbol{v}, \operatorname{grad} \varphi)_\Omega - (p, \operatorname{div} \varphi)_\Omega - (\operatorname{div} \boldsymbol{v}, q)_\Omega, \tag{3.32}$$

which is obtained by testing Equations $(2.10)_1$ and $(2.20)_1$ by $\phi \in W$ and by partial integration of the diffusive term and the pressure gradient. A general weak form of Equations $(2.10)_1$ and $(2.20)_1$ can be formulated as follows. Find $\boldsymbol{w} = \{\boldsymbol{v}, p\} \in \boldsymbol{w}_b + W$, such that

$$\mathcal{A}(\boldsymbol{w}; \phi) = 0, \quad \forall \phi \in W. \tag{3.33}$$

Here $\boldsymbol{w}_b$ describes the prescribed homogeneous Dirichlet boundary conditions on Γ_1 and Γ_2 as well as the adequate Dirichlet boundary conditions on the sphere. For the computation of $\{\boldsymbol{w}^{(i)}, p^{(i)}\}$ for $i = 1, 2, 3$ we consider the so-called do-nothing (i.e., natural boundary) conditions on the inlet Γ_3 and on the the outlet Γ_4 (see, e.g., [22] for more details). For the auxiliary field $\{\boldsymbol{w}^{(4)}, p^{(4)}\}$ we impose the homogeneous Dirichlet boundary conditions on Γ_3 and Γ_4.

The discretization uses a conforming finite element space $W_h \subset W$ defined on quasiuniform triangulations $\mathcal{T}_h = \{K\}$ consisting of quadrilateral cells K covering the domain Ω. In order to resolve accurately the sphere geometry, we further consider local refined grids by means of hanging nodes

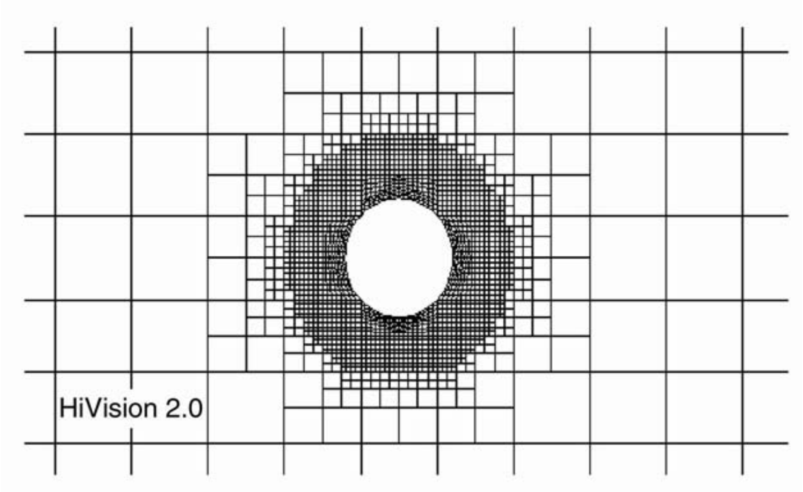


FIGURE 4.4 Local refined grid with hanging nodes in order accurately resolve the disk geometry.

(see Figure 4.4). For the trial and test spaces W_h we consider the standard Hood-Taylor finite element [25], that is, we define

$$W_h = \{(v,p) \in [C(\overline{D})]^3; \quad v|_K \in [Q_2]^2, p|_K \in Q_1\},$$

where Q_r describes the space of isoparametric tensor-product polynomials of degree r (for a detailed description of this standard construction process see, for example, [6]). This choice for the trial and test functions has the advantage that it guarantees a stable approximation of the pressure because the uniform *Babuska-Brezzi* inf-sup stability condition is satisfied uniformly in Ω (see [7,21] and references therein). The advantage, when compared to equal-order function spaces for the pressure and the velocity is that no additional stabilization terms are needed. The discrete counterpart of problem (3.33) then reads:

Find $\boldsymbol{w}_h = \{\boldsymbol{v}_h, p_h\} \in \boldsymbol{w}_{b,h} + W_h$, such that

$$\mathcal{A}(\boldsymbol{w}_h; \phi_h) = 0, \quad \forall \phi_h \in W_h, \tag{3.34}$$

where $\boldsymbol{w}_{b,h}$ describes the prescribed Dirichlet boundary data.

In Figure 4.6 the translational and angular velocities (scaled to Fr^2) in the Stokes approximation are given; see (2.19). In both figures, the radius of the disk is 0.05.

The linear algebraic system (3.34) is solved by the the *Generalized Minimal Residual Method* (*GMRES*) (see [42]) preconditioned by means of multigrid iteration (see [52,53] and references therein for the description of the

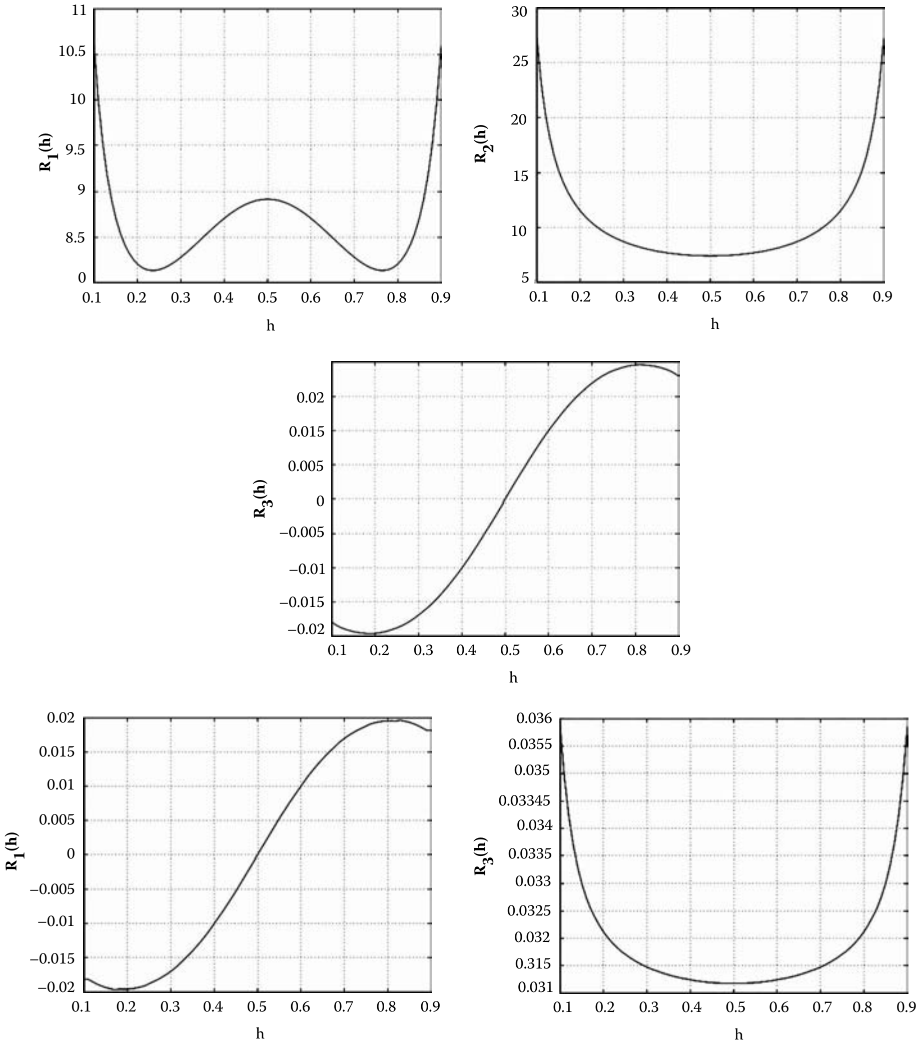


FIGURE 4.5a

different multigrid techniques for flow simulations). This preconditioner, based on a multigrid scheme oriented toward conformal higher-order FEM, is a key ingredient of the overall solution process. Two specific features characterize the proposed scheme: varying orders of the FEM ansatz on the mesh hierarchy, and a Vanka-type smoother [50] adapted to higher-order discretization. This somewhat technical part of the solver is described in full detail in [20]. Its implementation is part of the HiFlow project (see [19]).

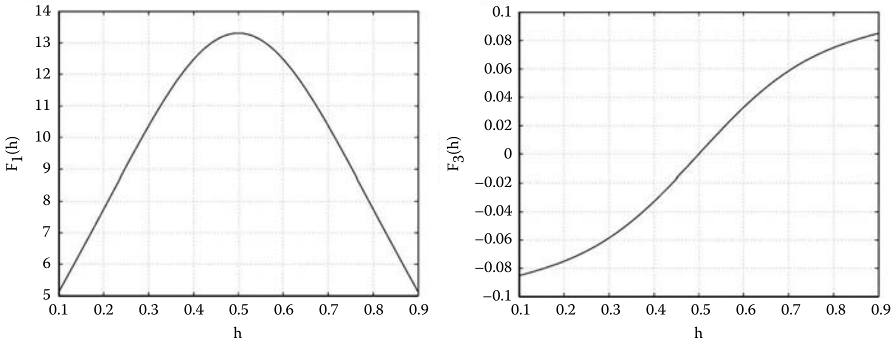


FIGURE 4.5b

4 Results for problem A

In this section we shall furnish the results obtained for the shear flow problem in the horizontal channel. Figures 4.5a and 4.5b show the graphs of the functions $\mathcal{T}$, $\mathcal{F}$ and $\mathcal{R}$ defined in (2.12) and (2.13). We begin to present our result for the case $E = 0$, that is, the liquid is purely Newtonian, and successively we will consider the influence of viscoelasticity. From (2.24), we then find that the equation determining the equilibrium heights becomes

$$\mathcal{G}(h) - K = 0, \quad K = \frac{\alpha}{Fr^2}. \tag{3.35}$$

Let us now study in detail this latter equation, starting with the case $\alpha = 0$. From the physical point of view, this means that the sphere has zero buoyancy. In such a case, the equilibrium heights h_{eq} are given by the solutions to the equation $\mathcal{G}(h) = 0$. With a view to Figure 4.7, we then find

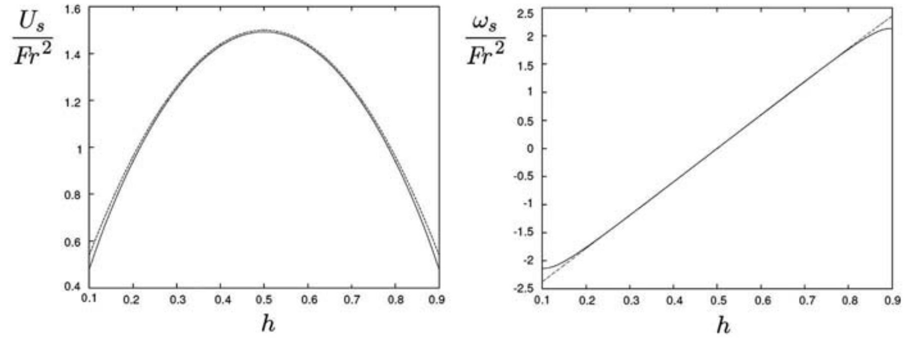


FIGURE 4.6 Translational velocity (full line, $U_s \equiv U_{s1}$) and angular velocity (full line, ω_s) of the disk in the Stokes approximation, as a function of h. Both quantities are normalized to the square of the Froude number. They are compared to the local velocity of the Poiseuille field (dotted line). The radius of the disk is 0.05.

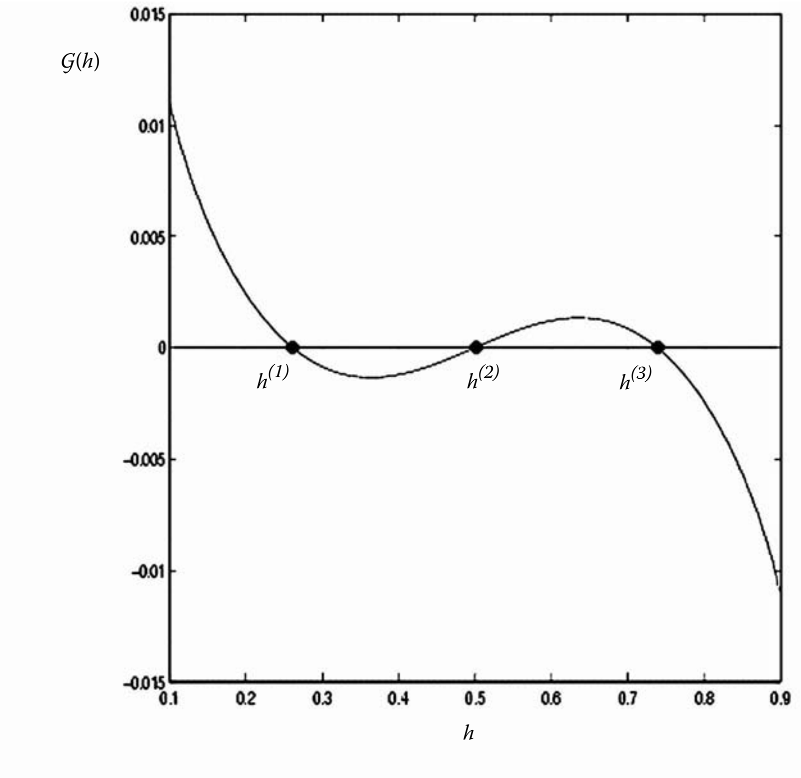


FIGURE 4.7 Equilibrium for zero buoyancy.

three possible solutions:

$$h^{(1)} = 0.261, \quad h^{(2)} = 0, \quad h^{(3)} = 0.738.$$

Notice that $h^{(1)}$ and $h^{(3)}$ are *stable*, while $h^{(2)}$ is *unstable.* These conclusions come from the following argument. Consider a small variation in h at the equilibrium position $h^{(1)}$. If it is at the left of $h^{(1)}$, that is, the sphere is pushed downward in the channel, $\mathcal{G}(h)$ is positive there and, therefore, the lift is positive. This means that the fluid will exert a force in the upward direction that will try to bring the particle back to $h^{(1)}$. Analogously, if the perturbation is at the right of $h^{(1)}$, that is, the sphere is pushed upward in the channel, $\mathcal{G}(h)$ is negative there and the lift is negative. This means that the liquid will exert a force in the downward direction that will bring the particle back to $h^{(1)}$.

For the same reasons, $h^{(3)}$ is stable and $h^{(2)}$ is unstable. We may summarize these results by saying that a given equilibrium position h_{eq} is *stable* if the slope of $\mathcal{G}(h)$ is *negative* at h_{eq} and it is *unstable* if it is *positive.*

Let us next consider the case of a *negative buoyancy*, that is, the density of the particle is larger than that of the liquid. We then have $\alpha > 0$. From

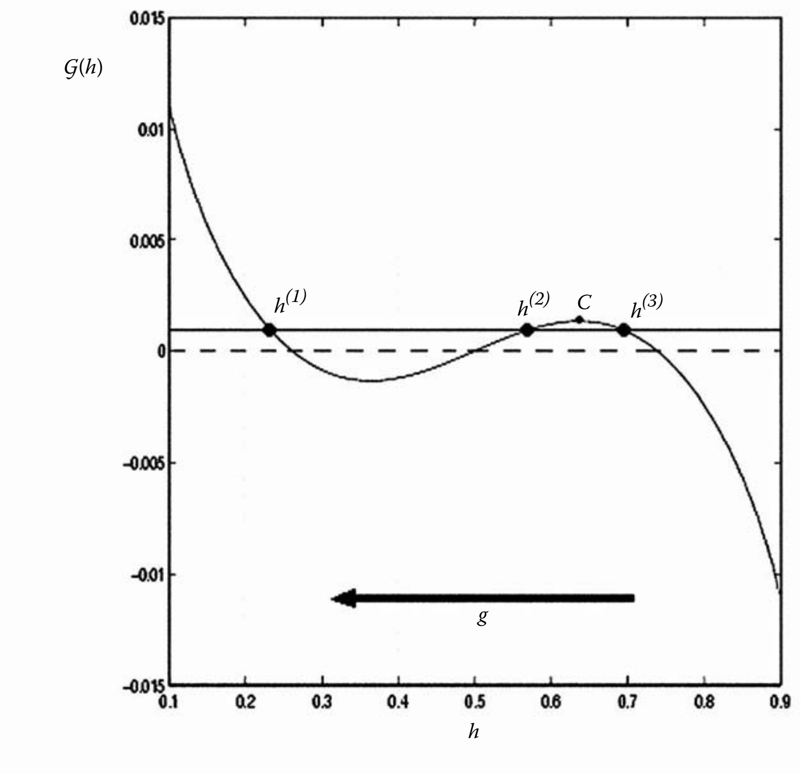


FIGURE 4.8 Equilibrium for nonzero buoyancy.

(3.35) we thus deduce that the equilibrium heights are given by the intersection of the straight line α/Fr^2 (represented by the solid straight line in Figure 4.8) with the curve $\mathcal{G}(h)$. With a view to Figure 4.2, we see that the equilibrium heights move upward in the curve. In particular, the stable ones, $h^{(1)}$ and $h^{(3)}$, become closer to the bottom plate. We also have the following interesting *jump phenomenon* in the equilibrium height. Consider a particle in the top half of the channel and set $\delta = \mathcal{G}(h_c)$, where h_c is the coordinate of the point C, the local maximum of $\mathcal{G}(h)$. We computed the value for δ and found $\delta = 0.00134$. Therefore, if

$$\alpha/Fr^2 < 0.00134, \tag{3.36}$$

the position $h^{(3)}$ exists and it is locally stable. However, if

$$\alpha/Fr^2 > 0.00134, \tag{3.37}$$

the particle will jump to the position $h^{(1)}$ on the lower half of the channel, which always exists and is always locally stable. This fact has a very simple

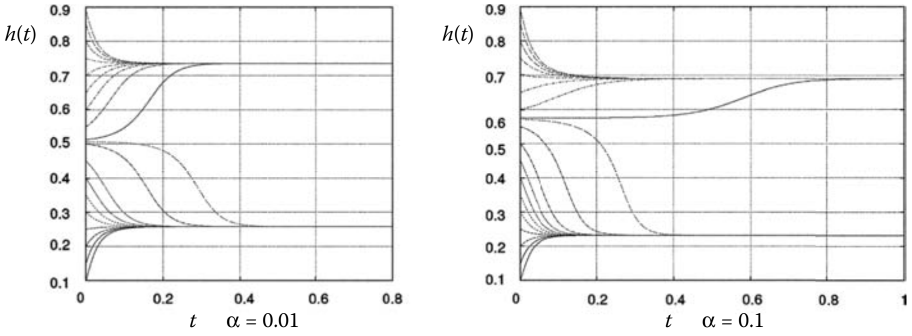


FIGURE 4.9 Trajectories of the disk for $Fr = 10$ and for $\alpha = 0.01, 0.1$.

interpretation from the physical point of view. In fact, for a given flow rate (Fr), a particle can stay in equilibrium in the top half of the channel if it is not too heavy, that is if there is enough lift. Otherwise, the particle will fall down. Notice that for a given α, we can always increase Fr (that is we can always increase the flow rate) in such a way that the particle stays in the equilibrium position $h^{(3)}$. The above result can also be interpreted in a different way. We *fix* the buoyancy, α, of the particle and take Fr sufficiently small in such a way that (3.37) holds, namely, there is only one stable equilibrium height ($h^{(1)}$) located on the branch of the curve $\mathcal{G}(h)$ close to the bottom plate. If we increase Fr, we will reach a critical value at which (3.36) is valid, and another stable equilibrium height $h^{(3)}$ appears. Using (2.23) with $E = 0$ we have computed the trajectories of disks that start at different heights in the channel. The results for $Fr = 10$ and for different values of α are shown in Figure 4.9 and Figure 4.10.

As we mentioned previously, in the above computations the (scaled) radius of the disk R was fixed to be 0.005. It is also interesting to see how the

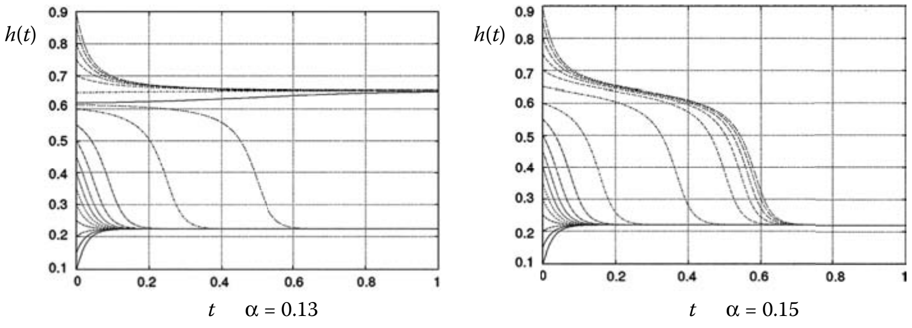


FIGURE 4.10 Trajectories of the disk for $Fr = 10$ and for $\alpha = 0.13, 0.15$.

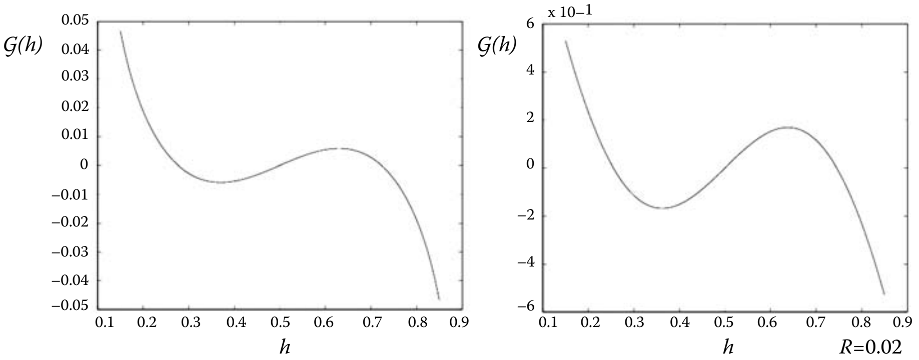


FIGURE 4.11 Graph of $\mathcal{G}(h)$ for a disk of scaled radius $R = 0.1$, 0.02.

lift ($\mathcal{G}(h)$, that is) changes with the radius. In Figure 4.11 we give $\mathcal{G}(h)$ for $R = 0.1$ Notice that the values of $\mathcal{G}(h)$ for $R = 0.1$ are one order of magnitude bigger than those for $R = 0.005$; see Figure 4.6.

We shall now take $E > 0$ and present the alteration of the equilibrium heights due to the viscoelastic property of the liquid. In accordance with classical experimental results on viscoelastic liquid (see [28], [36]) we have fixed the constant ε to be -1.8. We recall that if $E > 0$, the equation determining the possible equilibrium heights is (see (2.24))

$$\mathcal{G}(h) + E\mathcal{G}_V(h) - K = 0, \quad K = \frac{\alpha}{Fr^2}. \tag{3.38}$$

What we find is, basically, that as we increase the elasticity number, the stable equilibrium positions $h^{(1)}$ and $h^{(3)}$ (see Figure 4.7) will both move toward the center of the channel and eventually, when E reaches a critical value depending on α and Fr, they will both collapse into the position $h^{(2)}$ in the middle of the channel, which will then become the only stable equilibrium. Recall that apart from some multiplicative factor, $\mathcal{G}(h) + E\mathcal{G}_V(h)$ represents the total lift acting on the disk. Plots of $\mathcal{G}(h)+E\mathcal{G}_V(h)$ are given, for different values of E, in Figure 4.12 and Figure 4.13. Reasoning as in the case $E = 0$ and with the help of (3.38) and of Figures 4.12 and 4.13 we can establish the following. There is a critical elasticity number E_c such that if $E < E_c$, the curve $\mathcal{G}(h) + E\mathcal{G}_V(h)$ has one local maximum and one local minimum. Thus, under this condition, if $K = \dfrac{\alpha}{Fr^2} = 0$ there are two symmetric, locally stable equilibrium heights. Denote now by h_c the coordinate of the local maximum C and set $\delta_E = \mathcal{G}(h_c) + E\mathcal{G}_V(h_c)$. Then, if $K < \delta_E$ there are two stable equilibrium heights, one situated in the upper half of the channel and the other in the lower half and they are given by the coordinates of intersection of the curve $\mathcal{G}(h)+E\mathcal{G}_V(h)$ with the line $y(h) = K$. If $K > \delta_E$, then there is only one stable equilibrium height situated in the lower half.

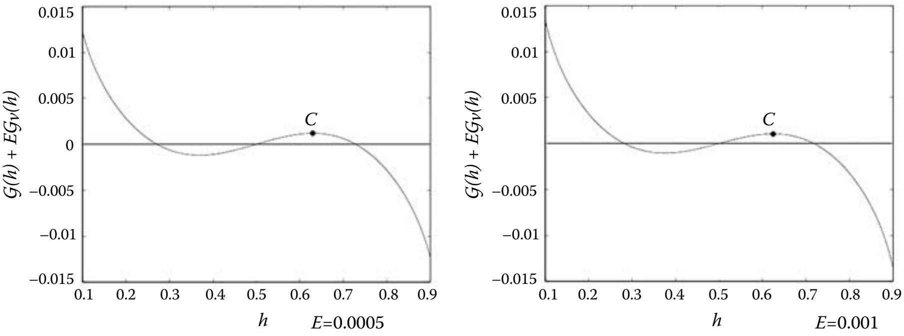


FIGURE 4.12 Plots of the curve $\mathcal{G}(h) + E\mathcal{G}_V(h)$ for $E = 0.0005, 0.001$.

If, however, $E > E_c$, then there is only one equilibrium and stable height given by the intersection of $\mathcal{G}(h) + E\mathcal{G}_V(h)$ with the line $y(h) = K$. If $K = 0$, this height is located at the center of the channel; see Figure 4.12. The computed value of E_c is $E_c = 0.0067$. So, the overall effect of viscoelasticity is to move the disk toward the center of the channel. This is in agreement with the results of [24] for the case of a purely viscoelastic liquid and for a disk of zero buoyancy.

Also for the viscoelastic case we have computed the trajectories of the disk, according to (2.23), for $Fr = 1$ and for different values of the buoyancy α and elasticity number E. The results are shown in Figures 4.14 through 4.16.

Concerning the components of the drag acting on the disk, that is, the integrals appearing on the right-hand side of (2.26), we found that they are, effectively, zero. As a consequence, the first-order contributions to the translational and angular velocities of the disk are zero, and therefore they coincide with the analogous quantities evaluated in the Stokes approximation.

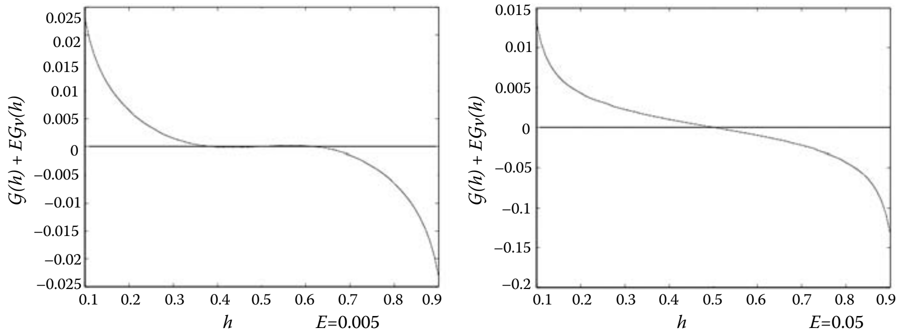


FIGURE 4.13 Plots of the curve $\mathcal{G}(h) + E\mathcal{G}_V(h)$ for $E = 0.005, 0.05$. The critical value E_c of E for which the only zero of $\mathcal{G}(h) + E\mathcal{G}_V(h)$ is $h = 0.5$ was computed to be $E = 0.0067$.

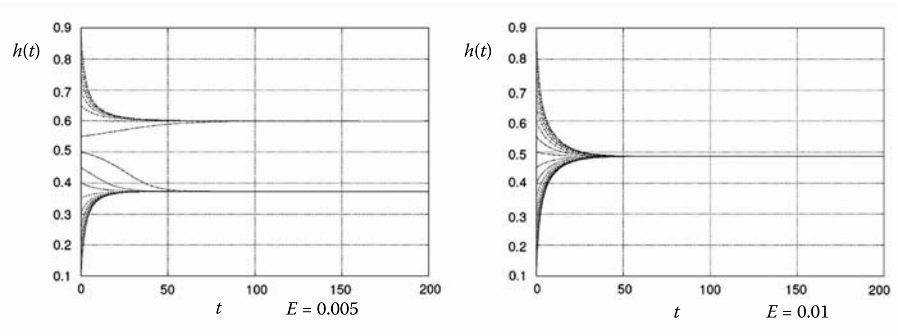


FIGURE 4.14 Trajectories of the disk in the viscoelastic case for $Fr = 1$, $\alpha = 0.0001$, and $E = 0.005, 0.001$.

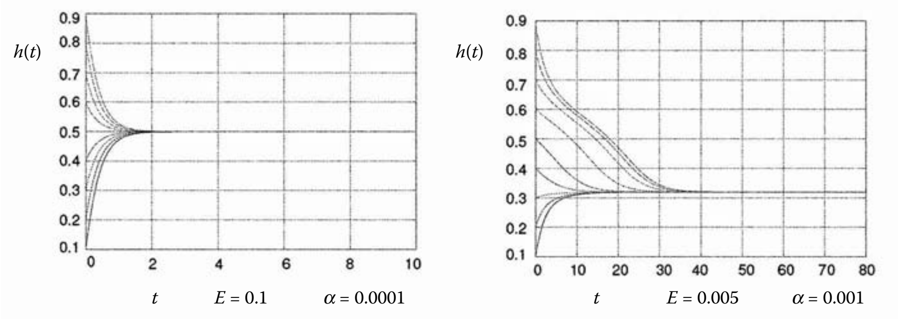


FIGURE 4.15 Trajectories of the disk in the viscoelastic case for $Fr = 1$, $\alpha = 0.0001, 0.001$, and $E = 0.1, 0.005$.

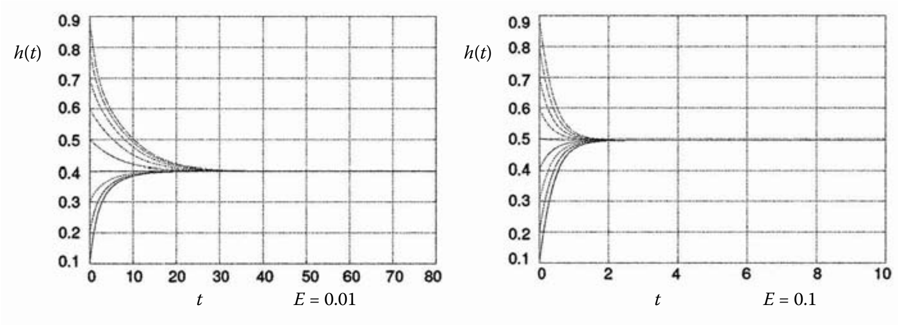


FIGURE 4.16 Trajectories of the disk in the viscoelastic case for $Fr = 1$, $\alpha = 0.001$, and $E = 0.01, 0.1$.

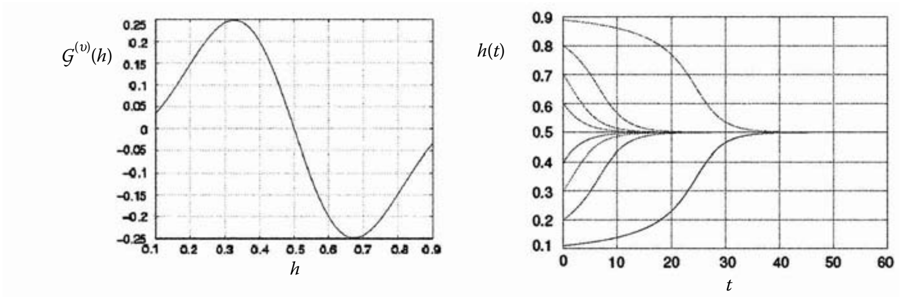


FIGURE 4.17 Plot of the curve $\mathcal{G}^{(v)}(h)$ and trajectories of the disk in the purely Newtonian case. The (scaled) radius of the disk is 0.05.

5 Results for Problem B

In this section we shall describe the results found for the fall of the disk in a vertical channel. As in the case of the shear problem in the horizontal channel, we shall first describe the results for the purely Newtonian case. The equilibrium positions are then given, at first order in $\mathcal{R}$ by the solutions to the equation (2.30) with $E = 0$, that is,

$$\mathcal{G}^{(v)}(h) = 0,$$

while the equations of the trajectories are given by (2.29), namely,

$$U_2^{(1)} = \frac{\alpha^2}{\mathcal{T}_2(h)} \mathcal{G}^{(v)}(h).$$

What we find is that the lateral lift on the disk ($\sim \mathcal{G}^{(v)}(h)$) is zero only in the middle of the channel, which is the only equilibrium position allowed. This position is also stable and is a global attractor. Such results are quantitatively described in Figure 4.17, which shows the graph of $\mathcal{G}^{(v)}(h)$ and the trajectories of the disk. We next consider the effect of viscoelasticity on these results. Basically, as E increases from zero to positive values, the stable equilibrium position moves from the center of the channel toward the walls. Plots of $\mathcal{G}^{(v)}(h) + E\,\mathcal{G}_V^{(v)}(h)$ for different values of E are furnished in Figures 4.18 and 4.19, while corresponding trajectories, computed by solving (2.29), are given in Figures 4.20 and 4.21 for $\alpha = 1$. As in the case of Problem A, the first-order contributions to the drag (the integrals on the righthand side of (2.32)), are essentially zero from a computational point of view. This implies that at first order in $\mathcal{R}$ and We, the translational and angular velocities of the disk coincide with those in the Stokes approximation given as solutions to the algebraic system (2.26). Plots of the latter are given in Figure 4.22.

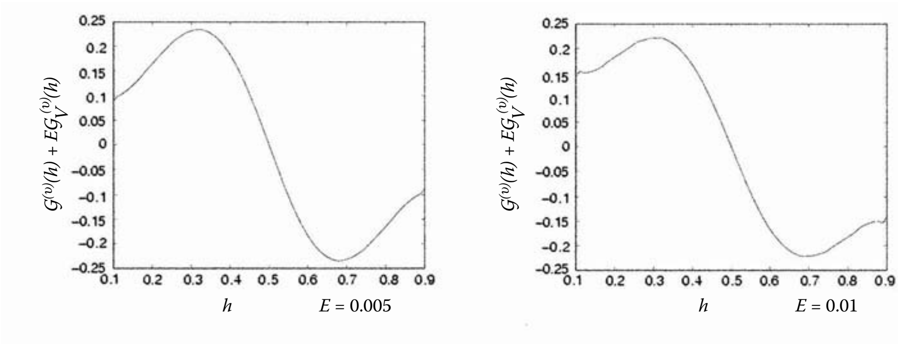


FIGURE 4.18 Plot of the curve $\mathcal{G}^{(v)}(h)+E\,\mathcal{G}_V(h)$ for $E = 0.005, 0.001$. Viscoelastic effects are still negligible. The only equilibrium position is in the middle of the channel ($h = 0.5$) and it is also stable.

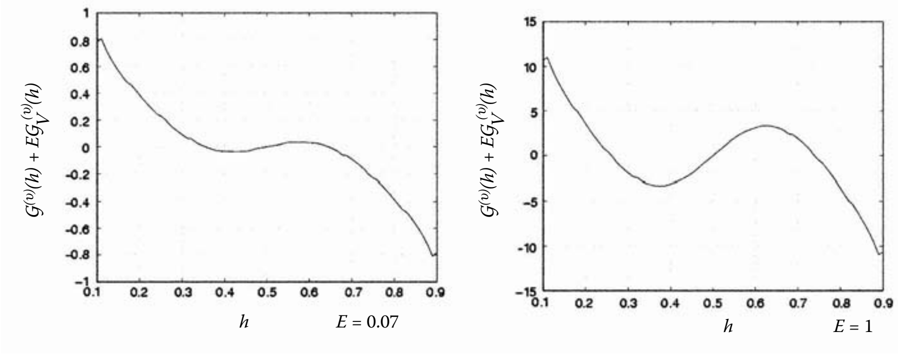


FIGURE 4.19 Plot of the curve $\mathcal{G}^{(v)}(h) + E\,\mathcal{G}_V(h)$ for $E = 0.07, 1$. The viscoelastic effects kick in. The equilibrium position in the middle of the channel loses its stability to two symmetric (locally) stable equilibrium positions.

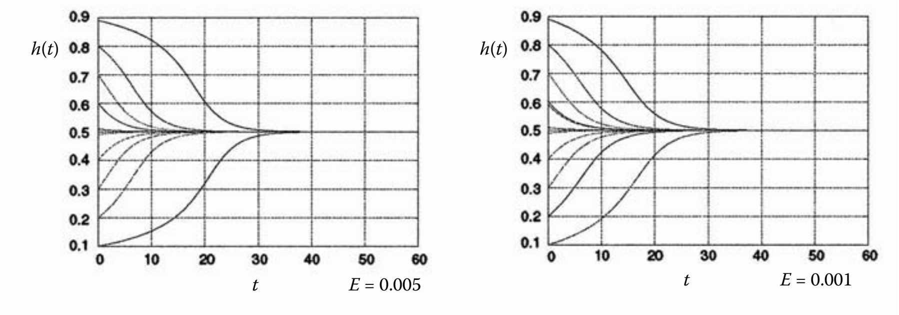


FIGURE 4.20 Trajectories of the disk for $E = 0.005, 0.001$. Viscoelastic effects are still negligible and the equilibrium position in the middle of the channel is a global attractor.

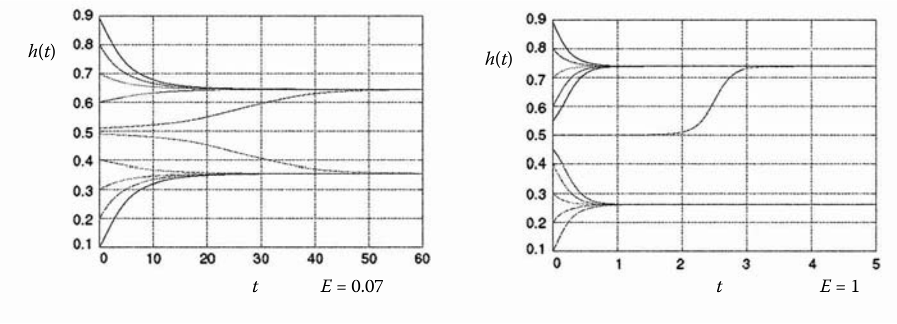


FIGURE 4.21 Trajectories of the disk for $E = 0.07, 1$. The viscoelastic effects produce two symmetric locally stable positions which are closer to the wall as E becomes larger.

Acknowledgments

The work of G.P. Galdi was partially supported by National Science Foundation (NSF) grants DMS-0103970 and DMS-0404834, and by the NSF US-Portugal-Switzerland Cooperative Research Program.

Appendix

The objective of this appendix is to show the following result.

Theorem 5.1

Let $h \in (a, 1-a)$, $Fr \geq 0$ *and* $\boldsymbol{G} \in \mathbb{R}^3$ *be given where* $\boldsymbol{G}$ *is (real) analytic in* $\mathcal{R}$. *Then, there exists* $\mathcal{R}_0 > 0$ *such that, if* $\mathcal{R} < \mathcal{R}_0$, *problems (2.6) through*

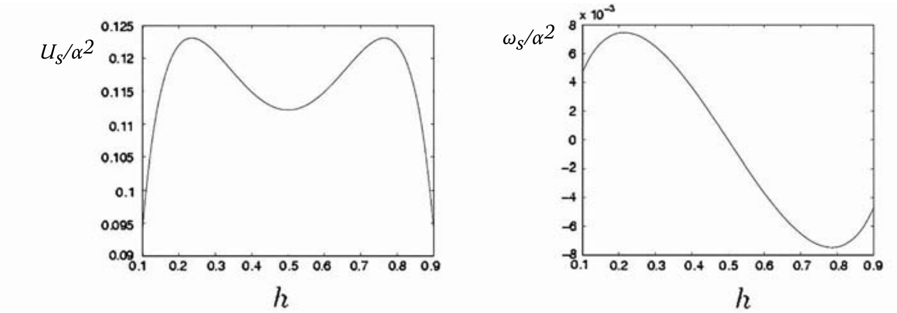


FIGURE 4.22 Translational velocity ($U_s \equiv U_{s1}$) and angular velocity (ω_s) of the disk in the Stokes approximation, as a function of h. Both quantities are normalized to α^2. The radius of the disk is 0.05.

(2.7) with $We = 0$ *has one and only one solution* $\{\boldsymbol{v}, P, \boldsymbol{U}, \omega\}$ *such that*

$$(\boldsymbol{v} - \boldsymbol{v}_0 + \boldsymbol{U}) \in W^{2,2}(\Omega), \quad P \in W^{1,2}(\Omega).$$

Moreover, the solution is (real) analytic in $\mathcal{R}$ *and the series*

$$\boldsymbol{v} = \boldsymbol{v}_s + \sum_{n=1}^{\infty} \boldsymbol{v}_n \mathcal{R}^n, \quad P = P_s + \sum_{n=1}^{\infty} P_n \mathcal{R}^n, \quad \boldsymbol{U} = \boldsymbol{U}_s + \sum_{n=1}^{\infty} \boldsymbol{U}_n \mathcal{R}^n,$$

$$\omega = \omega_s + \sum_{n=1}^{\infty} \omega_n \mathcal{R}^n,$$

are absolutely convergent in the norms of $W^{2,2}(\Omega)$*,* $W^{1,2}(\Omega)$*,* $\mathbb{R}^2$*, and* $\mathbb{R}$*, respectively.*

In order to prove this result, we need some preliminary considerations and preparatory lemmas. We first put problems (2.6) through (2.7) in an equivalent form. To this end, we begin to construct a suitable extension of $\boldsymbol{v}_0$. Let $\zeta = \zeta(x_1, x_2)$ be a smooth function such that (with $r = \sqrt{x_1^2 + x_2^2}$)

$$\zeta(x_1, x_2) = \begin{cases} 1 & \text{if } r > 2\delta \\ 0 & \text{if } r < \delta \end{cases}$$

and set

$$\boldsymbol{a}_h = (a_{h1}(x_1, x_2), a_{h2}(x_1, x_2))$$

where

$$a_{h1}(x_1, x_2) = \frac{\partial \zeta}{\partial x_2} \int_{-h}^{x_2} f(\eta, h) d\eta + \zeta(x_1, x_2) f(x_2, h),$$

$$a_{2h}(x_1, x_2) = -\frac{\partial \zeta}{\partial x_1} \int_{-h}^{x_2} f(\eta, h) d\eta,$$

where $f(x_2, h)$ is defined in (2.5). Taking into account that $\max f(x_2, h) = 3/2$, and that $\max |f'(x_2, h)| = 6$, by direct inspection one shows that $\boldsymbol{a}_h$ satisfies the following properties:

1. $\boldsymbol{a}_h \in C^\infty(\Omega)$;
2. $\operatorname{div} \boldsymbol{a}_h = 0$ in Ω;
3. $\boldsymbol{a}_h(x_1, x_2) = \boldsymbol{v}_0(x_2; h)$, $|x_1| > 2\delta$;
4. $|\operatorname{grad} \boldsymbol{a}_h(\boldsymbol{x})| \le M$, $\boldsymbol{x} \in \Omega$,
5. $\|\boldsymbol{a} \cdot \operatorname{grad} \boldsymbol{a}\|_q \le M$, $1 \le q \le \infty$,

6. $\displaystyle\int_{-h}^{1-h} a_{h1}(x_1,x_2)(\eta)d\eta = \int_{-h}^{1-h} f(\eta,h)d\eta = 1$

where M is independent of h. We next set

$$\boldsymbol{u} = \boldsymbol{v} - \mathit{Fr}\,\boldsymbol{a}_h + \boldsymbol{U}, \quad \Pi = P - \zeta p_0.$$

Using the properties of $\boldsymbol{a}_h$, we then find that the field $\boldsymbol{u}$ satisfies the following problem:

$$\left.\begin{aligned} &\operatorname{div}\boldsymbol{T}(\boldsymbol{u},\Pi) = \mathcal{R}(B(\boldsymbol{u},\boldsymbol{u}) - B(\boldsymbol{U},\boldsymbol{u}) + \mathit{Fr}\,(B(\boldsymbol{a}_h,\boldsymbol{u}) + B(\boldsymbol{u},\boldsymbol{a}_h) \\ &\qquad\qquad -B(\boldsymbol{U},\boldsymbol{a}_h))) + \boldsymbol{F} \\ &\operatorname{div}\boldsymbol{u} = 0 \end{aligned}\right\} \text{ in } \Omega$$

$$\begin{gathered} \boldsymbol{u}\,|_S = \omega\boldsymbol{e}_3\times\boldsymbol{x} + \boldsymbol{U}, \quad \boldsymbol{u}\,|_{\Gamma_1} = \boldsymbol{u}\,|_{\Gamma_2} = \boldsymbol{0} \\ \lim_{|x_1|\to\infty} \boldsymbol{u}(x_1,x_2) = \boldsymbol{0} \\ \int_{-h}^{1-h} u_1(x,x_2)dx_2 = 0 \\ \int_S \boldsymbol{T}(\boldsymbol{u},p)\cdot\boldsymbol{n} = \boldsymbol{G}, \quad \int_S \boldsymbol{x}\times\boldsymbol{T}(\boldsymbol{u},p)\cdot\boldsymbol{n} = \boldsymbol{0}, \end{gathered} \tag{5.39}$$

where $\boldsymbol{T}\equiv\boldsymbol{T}_N$, $B(\boldsymbol{a},\boldsymbol{b}) = \boldsymbol{a}\cdot\operatorname{grad}\boldsymbol{b}$ and $\boldsymbol{F} = \mathcal{R}\mathit{Fr}\;B(\boldsymbol{a}_h,\boldsymbol{a}_h) + \mathit{Fr}^2\boldsymbol{g}$, with $\boldsymbol{g}$ a function of bounded support. We shall now look (formally) for a solution to (5.39) of the form

$$\boldsymbol{u} = \sum_{n=0}^{\infty}\boldsymbol{u}_n\mathcal{R}^n,\ \Pi = \sum_{n=0}^{\infty}\Pi_n\mathcal{R}^n,\ \boldsymbol{U} = \sum_{n=0}^{\infty}\boldsymbol{U}_n\mathcal{R}^n,\ \omega = \sum_{n=0}^{\infty}\omega_n\mathcal{R}^n,$$

where the zero-th order terms are a solution to the following Stokes problem:

$$\begin{gathered} \left.\begin{aligned} &\operatorname{div}\boldsymbol{T}(\boldsymbol{u}_0,\Pi_0) = \mathit{Fr}\,\boldsymbol{g} \\ &\operatorname{div}\boldsymbol{u}_0 = 0 \end{aligned}\right\} \text{ in } \Omega \\ \boldsymbol{u}_0\,|_S = \omega_0\boldsymbol{e}_3\times\boldsymbol{x} + \boldsymbol{U}_0, \quad \boldsymbol{u}_0\,|_{\Gamma_1} = \boldsymbol{u}_0\,|_{\Gamma_2} = \boldsymbol{0} \\ \lim_{|x_1|\to\infty} \boldsymbol{u}_0(x_1,x_2) = \boldsymbol{0} \\ \int_{-h}^{1-h} u_{01}(x,x_2)dx_2 = 0 \\ \int_S \boldsymbol{T}(\boldsymbol{u}_0,\Pi_0)\cdot\boldsymbol{n} = \boldsymbol{G}_0, \int_S \boldsymbol{x}\times\boldsymbol{T}(\boldsymbol{u}_0,\Pi_0)\cdot\boldsymbol{n} = \boldsymbol{0}, \end{gathered} \tag{5.40}$$

while, for $n \geq 1$,

$$\left.\begin{aligned} \operatorname{div} \boldsymbol{T}(\boldsymbol{u}_{n+1}, \Pi_{n+1}) &= \sum_{k=0}^{n} (B(\boldsymbol{u}_{n-k}, \boldsymbol{u}_k) - B(\boldsymbol{U}_{n-k}, \boldsymbol{u}_k)) \\ &\quad + Fr(B(\boldsymbol{a}_h, \boldsymbol{u}_n) + B(\boldsymbol{u}_n, \boldsymbol{a}_h) - B(\boldsymbol{U}_n, \boldsymbol{a}_h)) + \widetilde{\boldsymbol{F}} \\ \operatorname{div} \boldsymbol{u}_{n+1} &= 0 \end{aligned}\right\} \text{ in } \Omega$$

$$\begin{gathered} \boldsymbol{u}_{n+1}\,|_S = \omega_{n+1}\boldsymbol{e}_3 \times \boldsymbol{x} + \boldsymbol{U}_{n+1}, \quad \boldsymbol{u}_{n+1}\,|_{\Gamma_1} = \boldsymbol{u}_{n+1}\,|_{\Gamma_2} = \boldsymbol{0} \\ \lim_{|x_1|\to\infty} \boldsymbol{u}(x_1, x_2) = \boldsymbol{0} \\ \int_{-h}^{1-h} \boldsymbol{u}_{n+1} \cdot \boldsymbol{e}_1(x, x_2) dx_2 = 0 \\ \int_S \boldsymbol{T}(\boldsymbol{u}_{n+1}, \Pi_{n+1}) \cdot \boldsymbol{n} = \boldsymbol{G}_{n+1}, \ \int_S \boldsymbol{x} \times \boldsymbol{T}(\boldsymbol{u}_{n+1}, \Pi_{n+1}) \cdot \boldsymbol{n} = \boldsymbol{0}. \end{gathered} \tag{5.41}$$

In (5.41) $\widetilde{\boldsymbol{F}} = Fr^2 B(\boldsymbol{h}, \boldsymbol{h})$ if $n = 0$, and it is zero otherwise. Moreover, $\boldsymbol{G}_k$ are the coefficients of the power series of $\boldsymbol{G}$. We want to show that problems (5.40) and (5.41) are solvable for all $n \geq 0$ with corresponding estimates. Let $\mathcal{H}(\Omega)$ be the class of functions $\boldsymbol{\varphi}$ such that

1. $\boldsymbol{\varphi} \in C_0^\infty(\overline{\Omega})$;
2. $\operatorname{div} \boldsymbol{\varphi} = 0$ in Ω;
3. $\boldsymbol{\varphi} \equiv \boldsymbol{0}$ in a neighborhood of Γ_1 and Γ_2; (5.42)
4. $\boldsymbol{\varphi} = \overline{\boldsymbol{\varphi}} \equiv \boldsymbol{\Phi}_1 + \Phi_2 \boldsymbol{e}_3 \times \boldsymbol{y}$, for some $\boldsymbol{\Phi}_1 \in \mathbb{R}^2$, $\Phi_2 \in \mathbb{R}$, in a neighborhood of S.

Reasoning as in [15], one can show the following Poincaré inequality:

$$\|\boldsymbol{\varphi}\|_2 \leq \gamma_0 \|\boldsymbol{D}(\boldsymbol{\varphi})\|_2, \tag{5.43}$$

where γ_0 is a constant independent of h. Furthermore, one shows that the *translational velocity* $\boldsymbol{\Phi}_1$ and the *spin* Φ_2 of a generic $\boldsymbol{\varphi} \in \mathcal{H}(\Omega)$ can be controlled by the L^2-norm of $\boldsymbol{D}$. Specifically, we have

$$|\boldsymbol{\Phi}_1| + |\Phi_2| \leq \gamma \|\boldsymbol{D}(\boldsymbol{\varphi})\|_2 \tag{5.44}$$

where γ depends only on R. We shall denote by $H(\Omega)$ the completion of $\mathcal{H}(\Omega)$ in the norm $\|\boldsymbol{D}(\cdot)\|_2$.

We have the following lemma.

Lemma A1
Let $\boldsymbol{F}_1 \in L^2(\Omega)$, $\boldsymbol{F}_2 \in \mathbb{R}^2$ be given. Then, the problem

$$\left.\begin{aligned} \operatorname{div} \boldsymbol{T}(\boldsymbol{u},\Pi) &= \boldsymbol{F}_1 \\ \operatorname{div} \boldsymbol{u} &= 0 \end{aligned}\right\} \text{ in } \Omega$$

$$\boldsymbol{u}\,|_S = \omega \boldsymbol{e}_3 \times \boldsymbol{x} + \boldsymbol{U}, \quad \boldsymbol{u}\,|_{\Gamma_1} = \boldsymbol{u}\,|_{\Gamma_2} = \boldsymbol{0}$$

$$\lim_{|x_1|\to\infty} \boldsymbol{u}(x_1,x_2) = \boldsymbol{0} \tag{5.45}$$

$$\int_{-h}^{1-h} u_1(x,x_2)dx_2 = 0$$

$$\int_S \boldsymbol{T}(\boldsymbol{u},\Pi)\cdot\boldsymbol{n} = \boldsymbol{F}_2, \ \int_S \boldsymbol{x}\times\boldsymbol{T}(\boldsymbol{u},\Pi)\cdot\boldsymbol{n} = \boldsymbol{0},$$

has one and only one solution $\{\boldsymbol{u} \in W^{2,2}(\Omega), \Pi \in W^{1,2}(\Omega), \boldsymbol{U}, \omega\}$. Moreover, this solution satisfies the estimate

$$|\boldsymbol{U}| + |\omega| + \|\boldsymbol{u}\|_{2,2} + \|\Pi\|_{1,2} \le C\left(\|\boldsymbol{F}_1\|_2 + |\boldsymbol{F}_2|\right), \tag{5.46}$$

where $C = C(\Omega) > 0$.

Proof

We give a weak formulation of the problem. Thus, multiplying $(5.45)_1$ by $\boldsymbol{\varphi} \in \mathcal{H}(\Omega)$, and integrating by parts over Ω, we find

$$\boldsymbol{\Phi}_1 \cdot \int_S \boldsymbol{T}(\boldsymbol{u},p)\cdot\boldsymbol{n} + \Phi_2\boldsymbol{e}_3 \cdot \int_S \boldsymbol{x}\times\boldsymbol{T}(\boldsymbol{u},p)\cdot\boldsymbol{n} - \int_\Omega \boldsymbol{D}(\boldsymbol{u}) : \boldsymbol{D}(\boldsymbol{\varphi}) = \int_\Omega \boldsymbol{F}_1\cdot\boldsymbol{\varphi},$$

where $\boldsymbol{\Phi}_1 + \Phi_2\boldsymbol{e}_3 \times \boldsymbol{x}$ is the trace of $\boldsymbol{\varphi}$ at S. Therefore, using the last two equations in $(5.45)_2$ it follows that

$$\int_\Omega \boldsymbol{D}(\boldsymbol{u}) : \boldsymbol{D}(\boldsymbol{\varphi}) = -\int_\Omega \boldsymbol{F}_1\cdot\boldsymbol{\varphi} + \boldsymbol{\Phi}_1\cdot\boldsymbol{F}_2. \tag{5.47}$$

We shall say that $\{\boldsymbol{u}, h, \omega, U\}$ is a *weak solution* to problem (5.45) if: (i) $\boldsymbol{u} \in H(\Omega)$, (ii) $\boldsymbol{u} = \omega\boldsymbol{e}_3 \times \boldsymbol{x} + U\boldsymbol{e}_1$, $\boldsymbol{x} \in S$, and (iii) $\boldsymbol{u}$ satisfies (5.47) for all $\boldsymbol{\varphi} \in \mathcal{H}(\Omega)$.

The existence of a field $\boldsymbol{u}$ satisfying requirements (i) and (iii) can be proved by the classical Galerkin method. As is known, the method furnishes existence provided we obtain a suitable a priori bound on the solution. This latter can be obtained as follows. Replacing, formally, $\boldsymbol{\varphi}$ in (5.47) with $\boldsymbol{u}$, we get

$$\|\boldsymbol{D}(\boldsymbol{u})\|_2^2 = -\int_\Omega \boldsymbol{F}_1\cdot\boldsymbol{u} + \boldsymbol{U}\cdot\boldsymbol{F}_2. \tag{5.48}$$

Using in (5.48) a Schwarz inequality along with (5.43) and (5.44) we find

$$\|\boldsymbol{D}(\boldsymbol{u})\|_2 \le \gamma_0 \, \|\boldsymbol{F}_1\|_2 + \gamma \, |\boldsymbol{F}_2|, \tag{5.49}$$

which furnishes the desired a priori estimate. By means of (5.49) and of the Galerkin method we thus establish the existence of a weak solution that, in addition, satisfies (5.48). From standard regularity theory (see for example [14], Lemma VI.1.2), we have that $\boldsymbol{u} \in W^{2,2}(\Omega)$ and that it satisfies $(5.45)_1$ for some $\Pi \in W^{1,2}(\Omega)$. Moreover, the following estimate holds

$$\|\boldsymbol{u}\|_{2,2} + \|\Pi\|_{1,2} \le c \, (\|\boldsymbol{F}_1\|_2 + \|\boldsymbol{D}(\boldsymbol{u})\|_2 + |\boldsymbol{U}| + |\omega|) \tag{5.50}$$

where we used the inequality

$$\|\mathrm{grad}\, \boldsymbol{u}\|_2 \le \sqrt{2}\|\boldsymbol{D}(\boldsymbol{u})\|_2 \, ; \tag{5.51}$$

(see [15]). The lemma then follows from (5.49), (5.50), and (5.44).

Lemma A2

Let $\boldsymbol{v}, \boldsymbol{w} \in W^{1,2}(\Omega)$. Then

$$\|B(\boldsymbol{v}, \boldsymbol{w})\|_2 \le c\|\boldsymbol{D}(\boldsymbol{v})\|_2\|\boldsymbol{D}(\boldsymbol{w})\|_2,$$

where $c = c(\Omega) > 0$.

Proof

By the Schwarz inequality, we have

$$\|B(\boldsymbol{v}, \boldsymbol{w})\|_2 \le \|\boldsymbol{v}\|_4\|\boldsymbol{w}\|_4.$$

Because

$$\|\boldsymbol{u}\|_4 \le c \, \|\mathrm{grad}\, \boldsymbol{u}\|_2 \ \ \boldsymbol{u} \in W^{1,2}(\Omega)$$

(see, e.g., [14], Lemma IX.2.1), the lemma follows from these last two displayed inequalities and from (5.51).

Set

$$V_n := \|\boldsymbol{u}_n\|_{2,2} + \|\Pi\|_{1,2} + |\boldsymbol{U}_n| + |\omega_n|$$

$$\begin{aligned}\boldsymbol{H}_n := \sum_{k=0}^{n} (B(\boldsymbol{u}_{n-k}, \boldsymbol{u}_k) - B(\boldsymbol{U}_{n-k}, \boldsymbol{u}_k)) + Fr\,(B(\boldsymbol{a}_h, \boldsymbol{u}_n) + B(\boldsymbol{u}_n, \boldsymbol{a}_h) \\ -B(\boldsymbol{U}_n, \boldsymbol{a}_h)) + \widetilde{\boldsymbol{F}_n},\end{aligned}$$

where $\widetilde{\boldsymbol{F}_n} = \widetilde{\boldsymbol{F}}$ if $n = 1$ and $\widetilde{\boldsymbol{F}_n} = 0$ otherwise. From Lemma A2 and from the properties 1 through 6 of the function $\boldsymbol{a}_h$ it easily follows that

$$\|\boldsymbol{H}_n\|_2 \leq c\left(\sum_{k=0}^{n} V_{n-k}V_k + \mathit{Fr}\, V_n + \mathit{Fr}^2\delta_{n1}\right), \quad n \geq 0, \tag{5.52}$$

We now apply the results of Lemma A1 and A2 to problems (5.40) and (5.41) and use (5.52). We thus get

$$\begin{aligned} &V_0 \leq c\,(\mathit{Fr} + |\boldsymbol{G}_0|) \\ &V_{n+1} \leq c\left(\sum_{k=0}^{n} V_{n-k}V_k + \mathit{Fr}\, V_n + \mathit{Fr}^2\delta_{n1} + |\boldsymbol{G}_{n+1}|\right), \quad n \geq 0. \end{aligned} \tag{5.53}$$

Let A_n be defined through the following recurrent relations:

$$\begin{aligned} &A_0 = c\,(\mathit{Fr} + |\boldsymbol{G}_0|) \\ &A_{n+1} = c\left(\sum_{k=0}^{n} A_{n-k}A_k + \mathit{Fr}\, A_n + \mathit{Fr}^2\delta_{n1} + |\boldsymbol{G}_{n+1}|\right), \quad n \geq 0. \end{aligned} \tag{5.54}$$

Clearly, $V_k \leq A_k$, for all $k \geq 0$. We shall now show that, provided $\mathcal{R}$ is sufficiently restricted, the series $Z := \sum_{n=0}^{\infty} A_n\mathcal{R}^n$ is converging, thus implying the convergence of $\sum_{n=0}^{\infty} V_n\mathcal{R}^n$, which will complete the existence part of the theorem. To reach our goal we observe that, multiplying both sides of $(5.54)_2$ by $\mathcal{R}^n$, using a Cauchy product formula and summing over n from 0 to ∞, we obtain

$$Z - A_0 = c[\mathcal{R}(Z^2 + \mathit{Fr}\,Z + \mathit{Fr}^2) + \mathcal{S}) \tag{5.55}$$

where $\mathcal{S} = \sum_{n=0}^{\infty} |G_n|\mathcal{R}^n - |G_0|$. The solution to (5.55), which reduces to a_0 at $x = 0$, is given by

$$Z(x) = \frac{1}{2c\mathcal{R}}\Big[(1 - c\mathcal{R}\mathit{Fr}) - \sqrt{(1 - c\mathcal{R}\mathit{Fr})^2 - 4(A_0 + c\mathit{Fr}^2 + \mathcal{S})\,c\mathcal{R}}\Big],$$

which is positive and has an analytic branch provided

$$1 > c\mathcal{R}\mathit{Fr}, \ (1 - c\mathcal{R}\mathit{Fr})^2 > 4(A_0 + c\mathit{Fr}^2 + \mathcal{S})\,c\mathcal{R}.$$

Let $B > 0$ be such that the series $\mathcal{S} + |\boldsymbol{G}_0|$ converges for all $\mathcal{R} \in (0, B]$ and let $G_M = G_M(B)$ denote an upper bound for $\mathcal{S} + |\boldsymbol{G}_0|$, uniformly in $\mathcal{R} \in (0, B]$. Taking into account that $A_0 = V_0$ and inequality $(5.53)_1$, we thus deduce that this latter condition is satisfied provided we choose $\mathcal{R} < \min\{B, \dfrac{C}{Fr + Fr^2 + G_M}\}$, where $C > 0$ depends only on Ω. The theorem is therefore proved.

References

[1] Adams, R., 1975, *Sobolev Spaces*, Academic Press, New York.

[2] Advani, A.S., 1994, *Flow and Rheology in Polymer Composites Manufacturing*, Elsevier, Amsterdam.

[3] Asmolov, E.S., 1999, The inertial lift on a spherical particle in a plane Poiseuille flow at large channel Reynolds number, *J. Fluid Mech.*, **381**, 63–87.

[4] Bagnold, R.A., 1974, Fluid forces on a body in shear-flow; experimental use of "stationary flow," *Proc. R. Soc. Lond. A*, **20**, 147–171.

[5] Becker, L.E., McKinley, G.H., and Stone, H.A., 1996, Sedimentation of a sphere near a plane wall: Weak non-Newtonian and inertial effects, *J. Non-Newtonian Fluid Mech.*, **63**, 201–233.

[6] Brenner, S.C. and Scott, R.L., 1994, *The Mathematical Theory of Finite Element Methods,* Springer, Berlin-Heidelberg-New York.

[7] Brezzi, F. and Falk, R., 1991, Stability of higher-order Hood-Taylor methods, *SIAM J. Numer. Anal.*, **28**(3), 581–590.

[8] Chhabra, R.P., 1993, *Bubbles, Drops and Particles in Non-Newtonian Fluids*, CRC Press, Boca Raton, FL.

[9] Cherukat, P., McLaughlin, J.B., and Graham, A.L., 1994, The inertial lift on a rigid sphere translating in a linear shear flow, *Int. J. Multiphase Flow*, **20**, 339–353.

[10] Cox, R.G. and Brenner, H., 1968, The lateral migration of solid particles in Poiseuille flow. I. Theory, *Chem Eng. Sci.*, **23**, 625–643.

[11] Cox, R.G. and Hsu, S.K., 1977, The lateral migration of solid particles in a laminar flow near a plane, *Int. J. Multiphase Flow*, **3**, 201–222.

[12] Dean, E.J. and Glowinski, R., 1997, A wave equation approach to the numerical solution of the Navier-Stokes equations for incompressible viscous flow, *C.R. Acad. Sci. Paris*, t. **325**, Série 1, 783–791.

[13] Eichhorn, R. and Small, S., 1964, Experiments on the lift and drag of spheres suspended in a Poiseuille flow, *J. Fluid Mech.*, **20**, 513–527.

[14] Galdi, G.P., 1998, *An Introduction to the Mathematical Theory of the Navier-Stokes Equations: Nonlinear Steady Problems*, Springer Tracts in Natural Philosophy, Vol. 38, Springer-Verlag, New York, 2nd Corrected Edition.

[15] Galdi, G.P., 2004, On the motion of a rigid body in a viscous liquid: A mathematical analysis with applications, *Handbook of Mathematical Fluid Mechanics*, Vol. 1: *Stationary Differential Equations*, North-Holland, Amsterdam, 71–156.

[16] Galdi, G.P., Pokorny, M., Vaidya, A., Joseph, D.D., and Feng, J., 2002, Orientation of symmetric bodies falling in a second-order liquid at non-zero Reynolds number, *Math. Mod. Meth. Appl. Sci.*, **12**(11), 1–39.

[17] Grossman, P.D. and Soane, D.S., 1990, Orientation effects on the electrophoretic mobility of rod-shaped molecules in free solution, *Anal. Chem.*, **62**, 1592–1596.

[18] Hames, B.D. and Rickwood, D., Eds., 1984, *Gel Electrophoresis of Proteins*, IRL Press, Washington, D.C.

[19] Heuveline, V., 2000, HiFlow a general finite element C++ package for 2D/3D flow simulation. *www.hiflow.de.*

[20] Heuveline, V., 2004, On higher-order mixed FEM for low Mach number flows: Application to a natural convection benchmark problem, *Int. J. Numer. Math. Fluids*, **17**, 125–148.

[21] Heuveline, V., and Schieweck, F., 2004, The Inf-Sup condition for higher order finite elements on meshes with hanging nodes. Technical report, University of Heidelberg, SFB 359.

[22] Heywood, J.G., Rannacher, R., and Turek, S., 1992. Artificial boundaries and flux and pressure conditions for the incompressible Navier-Stokes equations, *Int. J. Numer. Math. Fluids*, **22**, 325–352.

[23] Ho, B.P. and Leal, L.G., 1974, Inertial migration of rigid spheres in two-dimensional unidirectional flows, *J. Fluid Mech.*, **65**, 365–400.

[24] Ho, B.P. and Leal, L.G., 1976, Migration of rigid spheres in two-dimensional unidirectional shear flow of a second-order fluid, *J. Fluid Mech.*, **76**, 783–799.

[25] Hood, P. and Taylor, C., 1973, A numerical solution of the Navier-Stokes equations using the finite element techniques, *Comp. and Fluids*, **1**, 73–100.

[26] Hu, H. and Joseph, D.D., 1999, Lift on a sphere near a plane wall in a second-order fluid, *J. Non-Newtonian Fluid Mech.*, **88**, 173–184.

[27] Huang, P.Y. and Joseph, D.D., 1999, Effects of shear thinning on migration of neutrally buoyant particles in pressure driven flow of Newtonian and viscoelastic fluids, *J. Non-Newtonian Fluid Mech.*, **90**, 159–325.

[28] Joseph, D.D., 1990, *Fluid Dynamics of Viscoelastic Liquids*, Applied Mathematical Sciences, **84**, Springer-Verlag, Heidelberg.

[29] Joseph, D.D., 1996, Flow induced microstructure in Newtonian and viscoelastic fluids, in *Proceedings of the Fifth World Congress of Chemical Engineering, Particle Technology Track*, **6**, 3–16.

[30] Joseph, D.D., 2000, Interrogations of direct numerical simulation of solid-liquid flow, Web Site: *http://www.aem.umn.edu/people/faculty/joseph/interrogation.html.*

[31] Joseph, D.D., Ocando, D., and Huang, P.Y., 2000, Slip velocity and lift, accepted by *J. Fluid Mech.*

[32] Karnis, A. and Mason, S.G., 1966, Particle motions in sheared suspensions. XIX: Viscoelastic media, *Trans. Soc. Rheol.*, **10**, 571–592.

[33] King, M.R. and Leighton, D.T., 1997, Measurement of the inertial lift on a moving sphere in contact with a plane wall in shear flow, *Phys. Fluids*, **9**, 1248–1255.

[34] Kurose, R. and Komori, S., 1999, Drag and lift forces on a rotating sphere in a linear shear flow, *J. Fluid Mech.*, **384**, 183–206.

[35] Lee, S.C., Yang, D.Y., Ko, J., and You, J.R., 1997, Effect of compressibility on flow field and fiber orientation during the filling stage of injection molding, *J. Mater. Process. Tech.*, **70**, 83–92.

[36] Liu, Y.J. and Joseph, D.D., 1993, Sedimentation of particles in polymer solutions, *J. Fluid Mech.*, **255**, 565–595.

[37] McLaughlin, J.B., 1991, Inertial migration of a small sphere in linear shear flows, *J. Fluid Mech.*, **224**, 261–274.

[38] Milne-Thomson, L.M., 1952, *Theoretical Aerodynamics*, Van Nostrandt, New York.

[39] Patankar, N., Huang, Y., Ko, T., and Joseph, D.D., 2000, Lift-off of a single particle in Newtonian and viscoelastic fluids by direct numerical simulation, *J. Fluid Mech.*, **438**, 67–100.

[40] Roco, M.C., Ed., 1993, *Particulate Two-Phase Flow*, Series in Chemical Engineering, Butterworth-Heinemann, Boston.

[41] Ruschak, K.J., 1985, Coating flows, *Annual Review of Fluid Mechanics*, **17**, 65–89.

[42] Saad, Y., 1996, *Iterative Methods for Sparse Linear Systems*, Computer Science/ Numerical Methods, PWS Publishing Company, Boston.

[43] Saffman, P.G., 1965, The lift on a small sphere in a slow shear flow, *J. Fluid Mech.*, **22**, 385; and Corrigendum, *J. Fluid Mech*, **31**, 624 (1968).

[44] Schmid-Schonbein, H. and Wells, R., 1969, Fluid drop-like transition of erythrocytes under shear, *Science*, **165** (3890), 288–291.

[45] Schonberg, J.A. and Hinch, E.J., 1989, Inertial migration of a sphere in Poiseuille flow, *J. Fluid Mech.*, **203**, 517–524.

[46] Segrè, G. and Silberberg, A., 1961, Radial Poiseuille flow of suspensions, *Nature*, **189**, 209.

[47] Segrè, G. and Silberberg, A., 1962, Behaviour of macroscopic rigid spheres in Poiseuille flow, Part I, *J. Fluid Mech.*, **14**, 115.

[48] Tinland, B., Meistermann, L., and Weill, G., 2000, Simultaneous measurements of mobility, dispersion, and orientation of DNA during steady-field gel electrophoresis coupling a fluorescence recovery after photobleaching apparatus with a fluorescence detected linear dichroism setup, *Phys. Rev.* E, **61** (6), 6993–6998.

[49] Trainor, G.L., 1990, DNA sequencing, automation and human genome, *Anal. Chem.*, **62**, 418–426.

[50] Vanka, S., 1986, Block-implicit multigrid calculation of two-dimensional recirculating flows, *Comp. Meth. Appl. Mech. Eng.*, **59** (1), 29–48.

[51] Vasseur, P. and Cox, R.G., 1976, The lateral migration of a spherical particle in two-dimensional shear flows, *J. Fluid Mech.*, **78**, 385–414.

[52] Wesseling, P., 1992, *An Introduction to Multigrid Methods*, Wiley, Chichester.

[53] Wesseling, P. and Oosterlee, C.W., 2001, Geometric multigrid with applications to computational fluid dynamics. *J. Comp. Appl. Math.*, **128**, 311–334.

CHAPTER 5

Modeling and simulation of liquid-gas free surface flows

Alexandre Caboussat
Department of Mathematics, University of Houston,
Houston, Texas

Marco Picasso
Institute of Analysis and Scientific Computing,
École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Jacques Rappaz
Institute of Analysis and Scientific Computing,
École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

1 Introduction

Free surface flows are of interest in many engineering or mathematical problems. In the field of casting, injection or mold filling for instance, complex free surface phenomena in three space dimensions are involved, with topology changes and turbulence effects. These phenomena do not allow any regularity properties and require efficient numerical models for their treatment.

Complex flows with liquid-gas free surfaces have already been considered in the literature; see for instance [2,4,10,14–16]. The modeling of such free surface flows is then of great interest especially because of the implications from the computational point of view. Usually the models for two-phase free surface flows assume that both media are either compressible or incompressible. The model presented here assumes the presence of an incompressible liquid and a compressible gas. It is simplified in the gas domain, because the dynamical effects in the gas are not of interest.

Our model is as follows. The gas is assumed to be an ideal gas and the velocity of the gas is disregarded. A constant pressure is computed inside each connected component of the gas domain. A volume-of-fluid method is used to track the liquid domain and to compute the velocity and pressure fields in the liquid. The connected components of the gas domain are found by using an original numbering algorithm. Surface tension effects are also added by computing the curvature of the liquid-gas interface.

An implicit time-splitting algorithm is applied to decouple all the physical phenomena (see, e.g., [8]). Advection phenomena (including the motion of the volume fraction of liquid and the prediction of the liquid velocity) are solved first. Then the bubbles of gas are tracked and the pressure inside each bubble of gas is computed using the ideal gas law. The surface tension effects are computed on the liquid–gas interface. Finally a generalized Stokes problem is solved in order to update the velocity and pressure in the liquid.

Numerical results are presented to validate our model in the framework of mold filling and bubble simulations.

2 Mathematical flow model

The model presented in this section is extracted from [5,6]. Let Λ be a cavity of $\mathbb{R}^d$, $d = 2, 3$ in which the fluid must be confined, and let $T > 0$ be the final time of simulation. For any given time t, let Ω_t be the domain occupied by the fluid, let Γ_t be the free surface defined by $\partial\Omega_t \backslash \partial\Lambda$ and let Q_T be the space–time domain containing the liquid, that is, $Q_T = \{(x,t) : x \in \Omega_t, 0 < t < T\}$.

Some of the notations are reported in Figure 5.1 in a two-dimensional situation, namely the filling of an S-shaped channel. This situation corresponds

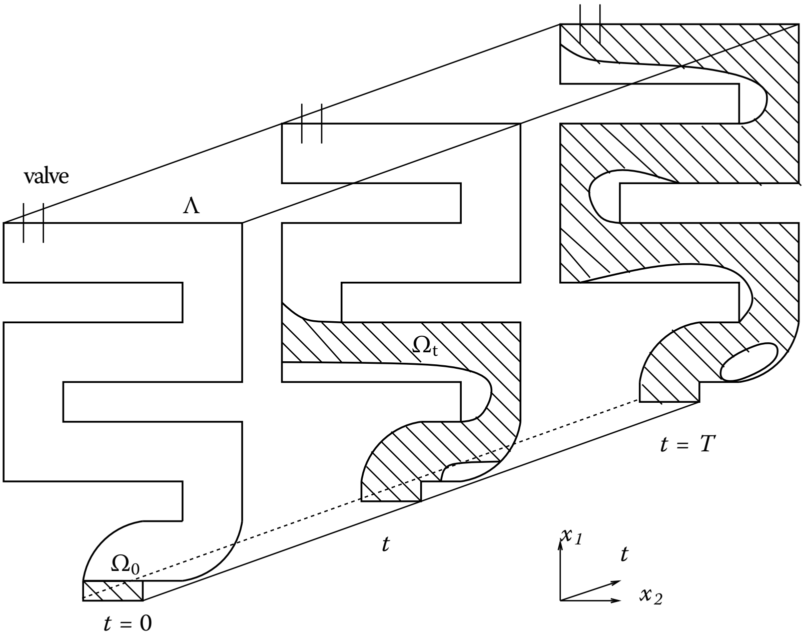


FIGURE 5.1 Computational domain for the filling of an S-shaped channel. At initial time, the channel Λ is empty. Then water enters from the bottom and fills the channel. Throughout the simulation, bubbles of gas may appear or disappear. A valve located on top of the channel allows the gas initially present in the channel to escape.

to water entering a thin S-shaped channel lying between two horizontal planes, thus gravity can be neglected. A valve is located at the end of the channel so that gas may escape.

In the liquid region, the velocity field $\mathbf{v} : Q_T \rightarrow \mathbb{R}^d$ and the pressure field $p : Q_T \rightarrow \mathbb{R}$ are assumed to satisfy the time-dependent, incompressible Navier-Stokes equations, that is,

$$\rho \frac{\partial \mathbf{v}}{\partial t} + \rho(\mathbf{v} \cdot \nabla)\mathbf{v} - 2\mathrm{div}\,(\mu \mathbf{D}(\mathbf{v})) + \nabla p = \mathbf{f} \qquad \text{in } Q_T, \tag{2.1}$$

$$\mathrm{div}\ \mathbf{v} = 0 \qquad \text{in } Q_T. \tag{2.2}$$

Here $\mathbf{D}(\mathbf{v}) = \frac{1}{2}(\nabla\mathbf{v} + \nabla\mathbf{v}^T)$ is the rate of deformation tensor, ρ the constant density and $\mathbf{f}$ the external forces.

Let $\varphi : \Lambda \times (0,T) \rightarrow \mathbb{R}$ be the characteristic function of the liquid domain Q_T. The function φ equals one if liquid is present, zero if it is not. In order to describe the kinematics of the free surface, φ must satisfy (in a weak sense):

$$\frac{\partial \varphi}{\partial t} + \mathbf{v} \cdot \nabla\varphi = 0 \qquad \text{in } Q_T. \tag{2.3}$$

The initial conditions are the following. At initial time, the characteristic function of the liquid domain φ is given, which defines the liquid region at initial time: $\Omega_0 = \{x \in \Lambda : \varphi(x,0) = 1\}$. The initial velocity field $\mathbf{v}$ is then prescribed in Ω_0.

The boundary conditions for the velocity field are the following. On the boundary of the liquid region in contact with the walls, inflow or slip boundary conditions are enforced. The forces acting on the free surface Γ_t are the normal forces due to the gas pressure and the normal forces due to the surface tension effects. The following equilibrium relation is then assumed to be satisfied on the liquid–gas interface:

$$-p\mathbf{n} + 2\mu\mathbf{D}(\mathbf{v})\mathbf{n} = -P\mathbf{n} + \sigma\kappa\mathbf{n} \qquad \text{on } \Gamma_t, \quad t \in (0,T), \tag{2.4}$$

where $\mathbf{n}$ is the unit normal of the liquid-gas free surface oriented toward the gas, P is the pressure in the gas, κ is the local curvature (or mean curvature in the three-dimensional case) of the interface and σ is a constant surface tension coefficient, which depends on the liquid and the gas. Therefore, the *continuum surface force* model (see e.g., [4,15,19]) is considered for the modeling of surface tension effects.

Gas may be trapped by the surrounding liquid and may therefore be compressed. In our model, the velocity in the gas is disregarded, because modeling the gas velocity would require solving the Euler compressible equations, which is with respect to CPU time.

The pressure P in the gas is assumed to be constant in each bubble of gas, that is to say in each connected component of the gas domain. Let $k(t)$ be the number of bubbles of gas at time t and let $B_i(t)$ denote the domain occupied by the bubble number i (the i-th connected component of the gas domain). Let $P_i(t)$ be the pressure in $B_i(t)$. The pressure in the gas P is then defined by $P(\mathbf{x}, t) = P_i(t)$, if $\mathbf{x} \in B_i(t)$. Moreover, the gas is assumed to be an ideal gas. The temperature is assumed to be constant. Let $V_i(t)$ be the volume of $B_i(t)$. At initial time, all the gas bubbles have a given pressure. At time t, the pressure in each bubble is computed by using the ideal gas law:

$$P_i(t)V_i(t) = \text{ constant} \qquad i = 1, \ldots, k(t). \tag{2.5}$$

In what follows, it is assumed that the total fraction number of molecules inside the set of bubbles that are not in contact with a valve (see Figure 5.1), is conserved between two time steps.

In most situations and when the time step is small enough, three situations may appear between two time steps at different locations in the process: first, a single bubble may stay a single bubble; then a bubble can split into two bubbles, and finally, two bubbles may merge into one. More complicated situations may appear but can be expressed as combinations of these three situations. Figure 5.2 illustrates these three situations.

The first column in Figure 5.2 shows the case of a single bubble. Assume that the pressure $P(t)$ in the bubble at time t and the volumes $V(t)$ and $V(t + \tau)$ are known. The fraction number of molecules inside the bubble is conserved, so that the gas pressure at time $t + \tau$ is easily computed from the relation $P(t + \tau)V(t + \tau) = P(t)V(t)$.

The second column corresponds to the merging of two bubbles. The pressure at time $t + \tau$ is computed by taking into account the conservation of

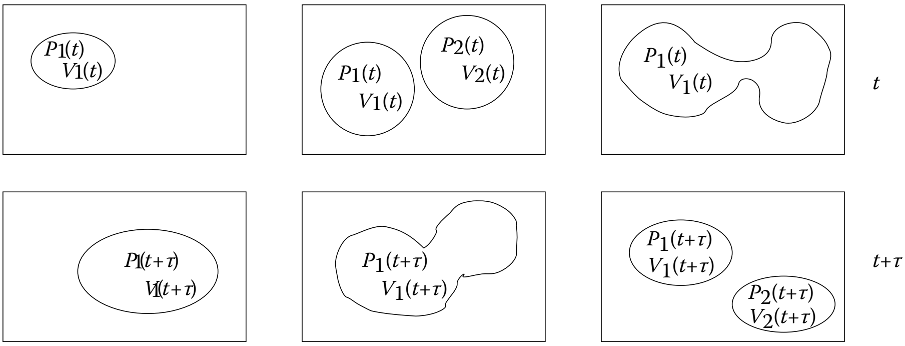


FIGURE 5.2 Computation of the pressure between times t and $t + \tau$. Left: a single bubble remains a single bubble; middle: two bubbles merge into one; right: a bubble splits into two bubbles.

the number of molecules in the bubbles, which yields $P_1(t+\tau)V_1(t+\tau) = P_1(t)V_1(t) + P_2(t)V_2(t)$.

The case where one bubble splits into two bubbles is finally illustrated in the third column. The number of molecules inside the gas domain is conserved between time steps t and $t+\tau$, that is $P_1(t+\tau)V_1(t+\tau) + P_2(t+\tau)V_2(t+\tau) = P_1(t)V_1(t)$. The relative fraction of molecules in bubble 1 at time t, which is in bubble 1 (respectively, 2) at time $t+\tau$, is determined from the computation of the subvolumes of bubble 1 at the exact time of splitting. Then the pressures $P_1(t+\tau)$ and $P_2(t+\tau)$ at time $t+\tau$ can be computed by taking into account the compression or decompression of each bubble separately (see [5] for details).

The mathematical description of our model is now complete. The model unknowns are the characteristic function φ in the whole cavity, the velocity $\mathbf{v}$ and pressure p in the liquid domain, the bubbles B_i (i.e., the connected components of the gas domain), the constant pressure P_i in each bubble of gas and the curvature κ and the normal vector $\mathbf{n}$ on the liquid–gas interface. These unknowns satisfy Equations (2.1) through (2.5).

3 Time-splitting scheme

An implicit splitting algorithm is proposed to solve (2.1) through (2.4) when the pressure in the gas, P, is computed with (2.5) and when the surface tension effects are taken into account.

Let $0 = t^0 < t^1 < t^2 < \ldots < t^N = T$ be a subdivision of the time interval $[0, T]$; define $\tau^n = t^n - t^{n-1}$ the n-th time step, $n = 1, 2, \ldots, N$, τ the largest time step.

Let φ^n, $\mathbf{v}^n$, Ω^n, k^n, P^n, B_i^n, $i = 1, 2, \ldots, k^n$ and κ^n, $\mathbf{n}^n$ be approximations of φ, $\mathbf{v}$, Ω_{t^n}, k, P, B_i, $i = 1, 2, \ldots, k$ and κ, $\mathbf{n}$ respectively at time t^n. Then the approximations φ^{n+1}, $\mathbf{v}^{n+1}$, Ω^{n+1}, k^{n+1}, P^{n+1}, B_i^{n+1}, $i = 1, 2, \ldots, k^{n+1}$ and κ^{n+1}, $\mathbf{n}^{n+1}$ at time t^{n+1} are computed by means of an implicit splitting algorithm, as illustrated in Figure 5.3.

First two advection problems are solved, leading to a prediction of the new velocity $\mathbf{v}^{n+1/2}$ together with the new approximation of the characteristic function φ^{n+1} at time t^{n+1}, which allows us to determine the new fluid domain Ω^{n+1}, the new gas domain $\Lambda\backslash\bar{\Omega}^{n+1}$, and the new liquid–gas interface, $\Gamma^{n+1} = \partial\Omega^{n+1}\backslash\partial\Lambda$. Then, the connected components of gas (bubbles) B_i^{n+1}, $i = 1, \ldots, k^{n+1}$ are tracked with a procedure explained in the following, and the pressure P_i^{n+1} in each bubble B_i^{n+1} is computed. Then an approximation of the curvature κ^{n+1} is obtained on the interface Γ^{n+1} together with a normal vector $\mathbf{n}^{n+1}$. Finally, a generalized Stokes problem is solved on Ω^{n+1} with boundary condition (2.4) on the liquid–gas interface, and essential boundary conditions on the boundary of the liquid domain in contact with the boundary of the cavity Λ and the velocity $\mathbf{v}^{n+1}$ and pressure p^{n+1} in the liquid are obtained.

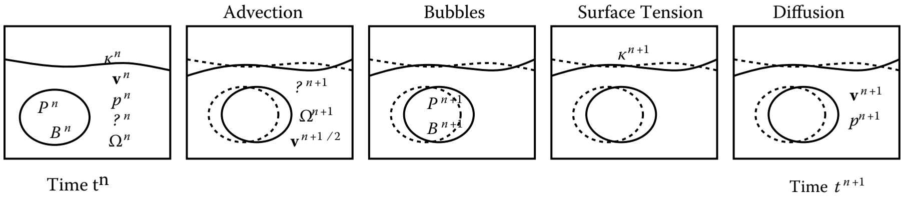


FIGURE 5.3 The splitting algorithm (from left to right). Two advection problems are solved to determine the new approximation of the characteristic function φ^{n+1}, the new liquid domain Ω^{n+1} and the predicted velocity $\mathbf{v}^{n+1/2}$. Then a constant pressure P_i^{n+1} is computed in each bubble B_i^{n+1}. The curvature κ^{n+1} and the normal vector $\mathbf{n}^{n+1}$ are then obtained on the liquid–gas interface. Finally, a generalized Stokes problem is solved to obtain the velocity $\mathbf{v}^{n+1}$ and the pressure p^{n+1} in the new liquid domain Ω^{n+1}, taking into account the pressure P^{n+1} and curvature κ^{n+1} on the liquid–gas interface.

The advection step consists of solving, between the times t^n and t^{n+1}, the two advection problems:

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v} = 0, \qquad \frac{\partial \varphi}{\partial t} + \mathbf{v} \cdot \nabla \varphi = 0, \tag{3.6}$$

with initial conditions given by the values of the functions $\mathbf{v}$ and φ at time t^n. This step is solved exactly by the method of characteristics [9,12] and yields a prediction of the velocity $\mathbf{v}^{n+1/2}$ and an approximation of the characteristic function of the liquid domain φ^{n+1} at time t^{n+1}, that is, $\mathbf{v}^{n+1/2}(\mathbf{x} + \tau^n \mathbf{v}^n(\mathbf{x})) = \mathbf{v}^n(\mathbf{x})$, and $\varphi^{n+1}(\mathbf{x} + \tau^n \mathbf{v}^n(\mathbf{x})) = \varphi^n(\mathbf{x})$, for all $\mathbf{x}$ belonging to Ω^n. The domain Ω^{n+1} is then defined as the set of points such that φ^{n+1} equals one.

Given the new liquid domain Ω^{n+1}, the next task consists of finding the gas bubbles B_i^{n+1}, $i = 1, \ldots, k^{n+1}$ (that is to say the connected components of the gas domain $\Lambda \backslash \Omega^{n+1}$). The algorithm for detecting a first connected component in the gas domain is the following. Given a point P in the gas domain $\Lambda \backslash \Omega^{n+1}$, we first search for a function u such that $-\Delta u = \delta_P$ in $\Lambda \backslash \Omega^{n+1}$, with $u = 0$ on Ω^{n+1} and u continuous. Because the solution u to this problem is strictly positive in the connected component containing point P and vanishes outside, the first bubble is detected. The physical interpretation in two space dimensions is the following. An elastic membrane is placed over the cavity Λ, deformation being impossible in the liquid domain, a point force being applied at point P.

The above procedure is repeated by initializing a point P in the subdomain of $\Lambda \backslash \Omega^{n+1}$ consisting of the bubbles that are not numbered yet, until all the bubbles are detected, which yields k^{n+1} and B_i^{n+1}, $i = 1, \ldots, k^{n+1}$.

Once the connected components of gas are numbered, an approximation P_i^{n+1} of the pressure in bubble i at time t^{n+1} has to be computed.

The pressure is constant inside each bubble of gas and is computed with the ideal gas law (2.5), except for bubbles in contact with a valve that has atmospheric pressure (see Figure 5.1).

In the case of a single bubble traveling in the liquid (see Figure 5.2, left), the law of ideal gas yields $P_1^{n+1}V_1^{n+1} = P_1^n V_1^n$, which means that the number of molecules inside the bubble is conserved between time t^n and t^{n+1}. In the case where two bubbles merge (see Figure 5.2, middle), this relation becomes $P_1^{n+1}V_1^{n+1} = P_1^n V_1^n + P_2^n V_2^n$. The third case is when a bubble splits into two (see Figure 5.2, right). Each of the parts of bubble 1 at time t^n contributes to bubbles B_1^{n+1} and B_2^{n+1}. The volume fraction of bubble B_1^n, which contributes to bubble B_j^{n+1}, is noted $V_{1,j}^{n+1/2}$, $j = 1, 2$. The computation of the pressure is then decomposed in two steps. First the volume fraction contributions $V_{1,j}^{n+1/2}$ are computed for $j = 1, 2$. Then the pressure in bubble B_j^{n+1} is computed by taking into account the compression or decompression of each of these bubbles, that is:

$$P_j^{n+1} = P_1^n \frac{V_{1,j}^{n+1/2}}{V_j^{n+1}}, \quad j = 1, 2. \tag{3.7}$$

The next step of the splitting algorithm is to compute the curvature of the free surface and the normal vector $\mathbf{n}^{n+1}$. Because the characteristic function φ^{n+1} is not smooth, it is first smoothed (see [19] for instance) in order to obtain a smooth function $\tilde{\varphi}^{n+1}$ such that the liquid–gas interface is given by the level line $\{x \in \Lambda : \tilde{\varphi}^{n+1}(x) = 1/2\}$, with $\tilde{\varphi}^{n+1} < 1/2$ in the gas domain, and $\tilde{\varphi}^{n+1} > 1/2$ in the liquid domain. Then the normal vector $\mathbf{n}^{n+1}$, directed outside the liquid domain toward the gas domain and the curvature κ^{n+1} on the liquid–gas interface are then given by the usual *level set* formulation (see e.g. [11]):

$$\mathbf{n}^{n+1} = -\frac{\nabla\tilde{\varphi}^{n+1}}{||\nabla\tilde{\varphi}^{n+1}||}, \qquad \kappa^{n+1} = \mathrm{div}\mathbf{n}^{n+1} = -\mathrm{div}\frac{\nabla\tilde{\varphi}^{n+1}}{||\nabla\tilde{\varphi}^{n+1}||}, \tag{3.8}$$

where $||\cdot||$ denotes the Euclidean norm in $\mathbb{R}^d$, $d = 2, 3$.

The smoothed function $\tilde{\varphi}^{n+1}$ is obtained from the characteristic function φ^{n+1} as follows. Let $K_\varepsilon(x)$ be a *kernel* function: the function $K_\varepsilon(x)$ has a compact support and is radially symmetric, monotonically decreasing with respect to $r = ||x||$ and is normalized. The kernel used in our algorithm is the one proposed in [19]. The convolution of φ^{n+1} with K_ε leads to a smoothed volume fraction of liquid $\tilde{\varphi}^{n+1}$ defined by:

$$\tilde{\varphi}^{n+1}(x) = \int_\Lambda \varphi^{n+1}(y)K_\varepsilon(x-y)dy, \quad \forall x \in \Lambda. \tag{3.9}$$

The final step of the splitting algorithm is the diffusion step, which consists of solving a generalized Stokes problem on the domain Ω^{n+1} using the predicted velocity $\mathbf{v}^{n+1/2}$ and the boundary condition (2.4). The following backward Euler scheme is used:

$$\rho \frac{\mathbf{v}^{n+1} - \mathbf{v}^{n+1/2}}{\tau^n} - 2\mathrm{div}\left(\mu \mathbf{D}(\mathbf{v}^{n+1})\right) + \nabla p^{n+1} = \mathbf{f} \quad \text{in } \Omega^{n+1}, \tag{3.10}$$

$$\mathrm{div}\ \mathbf{v}^{n+1} = 0 \qquad \text{in } \Omega^{n+1}, \tag{3.11}$$

where $\mathbf{v}^{n+1/2}$ is the prediction of the velocity obtained after the advection step. The boundary conditions on the free surface between the fluid and the bubble number i depend on the gas pressure P_i^{n+1} and the curvature κ^{n+1} and are given by (2.4).

4 Space discretization

Advection and diffusion phenomena now decoupled, a two-grids method is used, which combines a structured mesh of small cubic cells and a finite element unstructured mesh. Equations (3.6) are first solved using the method of characteristics on the fine structured mesh of small cells in order to reduce numerical diffusion and have a more accurate approximation of the liquid region. Then the diffusion step and the bubbles treatment are solved on a finite element unstructured mesh.

Assume that the grid is made of cubic cells of size h, each cell labeled by indices (ijk). Let φ_{ijk}^n and $\mathbf{v}_{ijk}^n$ be the approximate value of φ and $\mathbf{v}$ at the center of cell number (ijk) at time t^n. The unknown φ_{ijk}^n is the volume fraction of liquid in the cell (ijk) (piecewise constant approximation of φ^n on each cell of the structured grid; see [2] for a general presentation of VOF-like methods). The advection step on cell number (ijk) consists of advecting φ_{ijk}^n and $\mathbf{v}_{ijk}^n$ by $\tau^n \mathbf{v}_{ijk}^n$ and then projecting the values on the structured grid. An example of cell advection and projection is presented in Figure 5.4 in two space dimensions.

In order to enhance the quality of the volume fraction of liquid, postprocessing procedures have been implemented; see [9] for details.

Once values φ_{ijk}^{n+1} and $\mathbf{v}_{ijk}^{n+1/2}$ have been computed on the cells, values of the fraction of liquid φ_P^{n+1} and of the velocity field $\mathbf{v}_P^{n+1/2}$ are computed at the nodes P of the finite element mesh as follows: for any vertex P of the finite element mesh, let ψ_P be the corresponding basis function (i.e., the continuous, piecewise linear function having value one at P, zero at the other vertices). All the tetrahedrons K containing vertex P are considered, and all the cells (ijk) having center of mass C_{ijk} contained in these tetrahedrons are taken into account to compute the value of the field at vertex P. Then, φ_P^{n+1}, the volume fraction of liquid at vertex P and time t^{n+1} is computed using

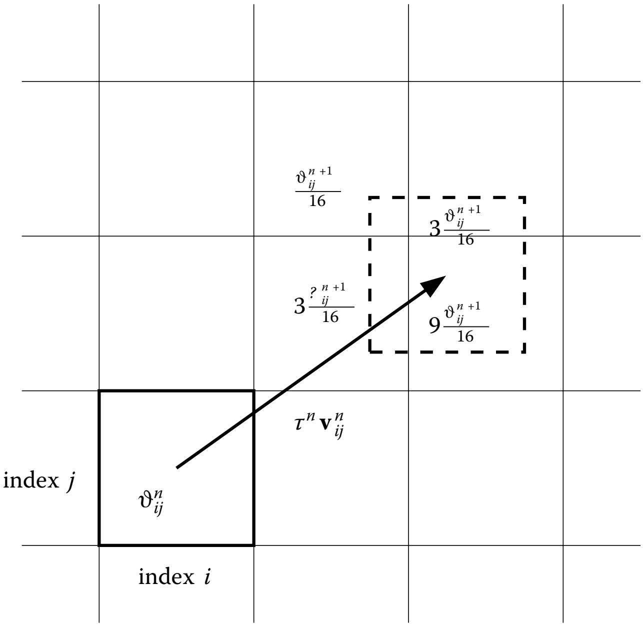

$\frac{\vartheta_{ij}^{n+1}}{16}$
$3\frac{\vartheta_{ij}^{n+1}}{16}$
$3\frac{?_{ij}^{n+1}}{16}$
$9\frac{\vartheta_{ij}^{n+1}}{16}$
$\tau^n \mathbf{v}_{ij}^n$
index j
ϑ_{ij}^n
index i

FIGURE 5.4 An example of two-dimensional advection of φ_{ij}^n by $\tau^n \mathbf{v}_{ij}^n$, and projection on the grid. The advected cell is represented by the dashed lines. The four cells containing the advected cell receive a fraction of φ_{ij}^n, according to the position of the advected cell.

the following formula:

$$\varphi_P^{n+1} = \frac{\sum_{P\in K}^{K} \sum_{C_{ijk}\in K}^{ijk} \psi_P(C_{ijk})\varphi_{ijk}^{n+1}}{\sum_{P\in K}^{K} \sum_{C_{ijk}\in K}^{ijk} \psi_P(C_{ijk})}.$$

The same kind of formula is used to obtain the predicted velocity $\mathbf{v}^{n+1/2}$ at the vertices of the finite element mesh. When these values are available at the vertices of the finite element mesh, the liquid region is defined as follows. An element of the mesh is said to be liquid if (at least) one of its vertices P has a value $\varphi_P^{n+1} > 0.5$. The computational domain Ω_h^{n+1} used for solving the diffusion problems (3.10) and (3.11) is then defined to be the union of all liquid elements.

Numerical experiments reported in [9] have shown that choosing a structured mesh cell size approximately 5 to 10 times smaller than the size of the finite elements is a good trade-off between reduction of the numerical diffusion and computational cost. Furthermore, because the characteristics method is used, the time step is not restricted by the CFL number.

Nevertheless the choice of CFL numbers ranging from 1 to 5 have been proved to be efficient numerically.

The numbering of the bubbles of gas requires solving several Poisson problems. The Poisson problems are solved on the finite element unstructured mesh, using piecewise linear finite elements.

The pressure inside each bubble of gas is computed with (3.7) and the approximations of the fractions of volumes $V_{i,j}^{n+1/2}$ are computed on the finite element mesh. Details may be found in [6].

The next step of the splitting algorithm consists of the computation of the curvature on the liquid–gas interface. Let $\mathcal{T}_h$ be the triangulation of the cavity Λ, Ω_h^{n+1} be the approximation of the liquid domain composed by elements K in $\mathcal{T}_h$, and Γ_h^{n+1} the approximation of Γ^{n+1}. Let ψ_{P_j} be the basis functions of the piecewise linear finite element space associated with each node P_j, $j = 1, \ldots, N$ in the cavity. Finally, let the piecewise linear finite elements space be denoted by $X_h^1(\Lambda)$, and $\tilde{\varphi}_h^{n+1}$ the approximation of $\tilde{\varphi}^{n+1}$ in $X_h^1(\Lambda)$. The value of $\tilde{\varphi}_h^{n+1}$ at a point P of $\mathcal{T}_h$ is explicitly given by

$$\tilde{\varphi}_P^{n+1} = \sum_{K \in \mathcal{T}_h} \frac{1}{d+1} |K| \sum_{P_j \in K} \varphi_h^{n+1}(P_j) K_\varepsilon(P_j - P), \quad i = 1, \ldots, N, \tag{4.12}$$

where $|K|$ denotes the surface (resp. volume) of element K. The approximation of the curvature κ^{n+1} is then approximated by the L^2-projection of κ^{n+1} given by (3.8) on $X_h^1(\Lambda)$ with *mass lumping*. The value of the approximated curvature at a point P of $\mathcal{T}_h$ is explicitly given, after integration by parts, by

$$\kappa_P^{n+1} = \frac{d+1}{|\Omega_j|} \left[\sum_{K \in \mathcal{T}_h} |K| \left(\frac{1}{d+1} \sum_{P_i \in K} \frac{\nabla \tilde{\varphi}_h^{n+1}(P_i)}{\left\| \nabla \tilde{\varphi}_h^{n+1} \right\|} \right) \nabla \psi_{P_j} \Big|_K \right.$$
$$\left. - \sum_{\substack{\partial K \subset \partial \Lambda \\ P_j \in K}} \frac{1}{d} \frac{\nabla \tilde{\varphi}_h^{n+1}(P_j)}{\left\| \nabla \tilde{\varphi}_h^{n+1} \right\|} \mathbf{n}_{\Lambda,h}(P_j) \, |\partial K| \right], \tag{4.13}$$

where $|\Omega_j| = \sum_{K, P_j \in K} |K|$ and $\mathbf{n}_{\Lambda,h}$ denotes the approximation of the external normal vector to the cavity Λ. The approximation of the normal vector to the free surface in (3.8) is given by the L^2-projection in $X_h^1(\Lambda)$ of the piecewise constant function $-\nabla \tilde{\varphi}_h^{n+1} / \| \nabla \tilde{\varphi}_h^{n+1} \|$.

Finally the diffusion step consists of solving the Stokes problems (3.10) and (3.11). Let $\mathbf{v}_h^{n+1}$, p_h^{n+1} be the piecewise linear approximation of $\mathbf{v}^{n+1}$, p^{n+1}. The Stokes problem is solved with stabilized $\mathbb{P}_1 - \mathbb{P}_1$ finite elements (Galerkin Least Squares method [7]) and consists of finding the velocity

$\mathbf{v}_h^{n+1}$ and pressure p_h^{n+1} such that:

$$\begin{aligned}
&\int_{\Omega_h^{n+1}} \frac{\mathbf{v}_h^{n+1} - \mathbf{v}_h^{n+1/2}}{\tau^n} \mathbf{w} dx + 2\mu \int_{\Omega_h^{n+1}} D(\mathbf{v}_h^{n+1}) : D(\mathbf{w}) dx \\
&- \int_{\Omega_h^{n+1}} \mathbf{f}\mathbf{w} dx - \int_{\Omega_h^{n+1}} p_h^{n+1} \operatorname{div} \mathbf{w} dx - \int_{\Omega_h^{n+1}} \operatorname{div} \mathbf{v}_h^{n+1} q dx \qquad (4.14) \\
&+ \int_{\Gamma_h^{n+1}} (P^{n+1} - \sigma\kappa_h^{n+1}) \mathbf{n}_h^{n+1} \mathbf{w} dS \\
&- \sum_{K \subset \Omega_h^{n+1}} \alpha_K \int_K \left(\frac{\mathbf{v}_h^{n+1} - \mathbf{v}_h^{n+1/2}}{\tau^n} + \nabla p_h^{n+1} - \mathbf{f} \right) \cdot \nabla q dx = 0,
\end{aligned}$$

for all $\mathbf{w}$ and q the velocity and pressure test functions, compatible with the boundary conditions on the boundary of the cavity Λ.

Once $\mathbf{v}_h^{n+1}$ is computed on the finite element mesh, the values $\mathbf{v}_{ijk}^{n+1}$ on each cell (ijk) of the structured grid are interpolated by restriction of $\mathbf{v}_h^{n+1}$ at the center of each cell C_{ijk}.

5 Numerical results

Numerical results in three space dimensions are presented to validate our model, both in the frame of mold filling and in the frame of bubbles and droplets simulations. All the computations were performed on computers with single processor Pentium Xeon 2.8 GHz CPU, 3 Gb memory, and running under Linux operating systems. The results are postprocessed with Calcosoft^{TM} or Ensight^{TM} softwares.

An S-shaped channel lying between two horizontal plates is filled. Results are compared with experiment [17]. The channel is contained in a 0.17 m× 0.24 m × 0.08 m rectangle. Water is injected with constant velocity 8.7 m/s, which corresponds to the experimental value reported in [17]. A valve is located at the top of the channel, as in Figure 5.1, allowing gas to escape. Density and viscosity are taken to be respectively $\rho = 1000$ kg/m^3 and $\mu = 0.01$ kg/(ms), and initial pressure in the gas is $P_{\text{atmo}} = 101300.0$ Pa. Surface tension effects can be neglected because the ratio between the capillary number and Reynolds number is very small ($Ca \simeq 1.5$); see, e.g., [14].

Numerical results are presented in the following. The final time is $T = 0.00532$ s and the time step is $\tau = 0.0001$ s. In Figure 5.5, the experiment is compared to 3D computations when the influence of the surrounding gas is taken into account. Notice that if the gas is not taken into account, the bubbles of trapped gas inside the cavity vanish instantaneously. The CPU time for the simulations is approximately 344 minutes and most of the CPU time is spent in solving the Stokes problem in the liquid domain.

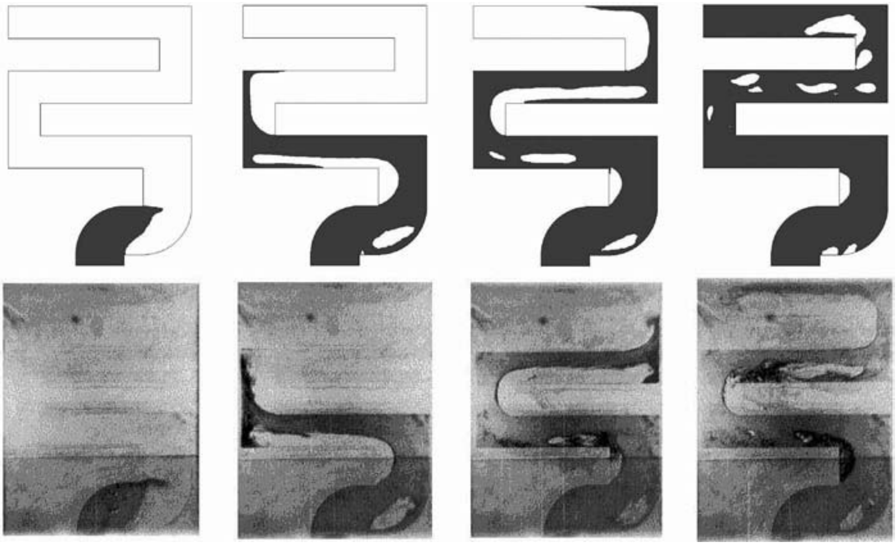


FIGURE 5.5 S-shaped channel: influence of gas bubbles. Computations with coarse mesh and $\alpha_T = 4h^2$. First row: 3D results with bubbles in the middle plane and second row: experimental results [17]. First column: time equals 7.15 ms, second column: 25.3 ms, third column: 39.3 ms and fourth column: 53.6 ms. Reprinted from *J. Comp. Phys*n 203, A. Caboussat, M. Picasso, and J. Rappaz, 626–649. Copyright (2005), with permission from Elsevier.

The computed liquid flow goes a little bit too fast compared to the experiment. This is mainly due to the slip boundary conditions on the walls of the cavity and can be corrected by adding a turbulent viscosity model.

Deformed droplets have been widely treated in the literature [4,10,16]. The rising of a bubble is considered here. A bubble of air is placed initially at the bottom of a cylinder filled with water. Under gravity forces, the bubble rises until reaching the top of the cylinder. Results in three space dimensions are illustrated in Figure 5.6 for a mesh made out of $115'200$ tetrahedrons. The CPU time for this computation is approximately 20 hours to achieve 1000 time steps.

6 Perspectives

The modeling of the free surface flows presented here has been extended to viscoelastic free surface flows [3,18]. Another important issue is the mold casting. In this case, the solidification model (see, for instance, [1,13]) must be coupled to the fluid flow equations. Examples are numerous in the field of metallurgy for the simulation of the filling of molds with liquid metal.

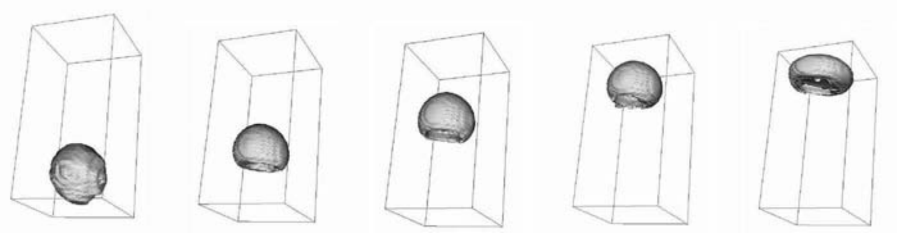


FIGURE 5.6 Rising bubble: three-dimensional results for $\sigma = 0.0738$ Nm^{-1}. Representation of the gas domain at times $t = 0.0, 0.25, 0.5, 0.75$ and 1 s. (left to right, top to bottom). Reprinted from *Computers and Fluids*, in press, A. Caboussat, A numerical method for the simulation of the surface flows with surface tension. Copyright (2006), with permission from Elsevier.

The authors wish to thank Dr. Vincent Maronnier, Calcom-ESI, Parc Scientifique, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland for implementation support. Calcom-ESI is acknowledged for kindly providing the Calcosoft^{TM} Pre- and Post-Processors.

Alexandre Caboussat's research was supported by the Swiss National Science Foundation, Grant PBEL2-103152.

References

[1] N. Ahmad, H. Combeau, J.-L. Desbiolles, T. Jalanti, G. Lesoult, J. Rappaz, M. Rappaz, and C. Stomp. Numerical Simulation of Macrosegregation: A Comparison between Finite Volume Method and Finite Element Method Predictions and a Confrontation with Experiments. *Metallurgical and Materials Transactions A*, 29A:617–630, 1998.

[2] E. Aulisa, S. Manservisi, and R. Scardovelli. A Mixed Markers and Volume-of-Fluid Method for the Reconstruction and Advection of Interfaces in Two-Phase and Free-Boundary Flows. *J. Comput. Phys.*, 188:611–639, 2003.

[3] A. Bonito, M. Picasso, and M. Laso. Numerical Simulation of 3D Viscoelastic Flows with Free Surfaces. XIIIth Workshop on Numerical Methods for Non-Newtonian Flows, 2005. Submitted to *J. Non-Newtonian Fluid Mechanics.*

[4] J. U. Brackbill, D. B. Kothe, and C. Zemach. A Continuum Method for Modeling Surface Tension. *J. Comput. Phys.*, 100:335–354, 1992.

[5] A. Caboussat. *Analysis and Numerical Simulation of Free Surface Flows.* Ph.D. thesis 2893, École Polytechnique Fédérale de Lausanne, 2003, available at `http://library.epfl.ch/theses`.

[6] A. Caboussat, M. Picasso, and J. Rappaz. Numerical Simulation of Free Surface Incompressible Liquid Flows Surrounded by Compressible Gas. *J. Comput. Phys.*, 203:626–649, 2005.

[7] L. P. Franca and S. L. Frey. Stabilized Finite Element Method: II. The Incompressible Navier-Stokes Equations. *Comp. Meth. Appl. Mech. Eng.*, 99:209–233, 1992.

[8] R. Glowinski and L.H. Juarez. Finite Element Method and Operator-Splitting for a Time-Dependent Viscous Incompressible Free-Surface Flow. *Comp. Fluid Dynam. J.*, 12(3):459–468, 2003.

[9] V. Maronnier, M. Picasso, and J. Rappaz. Numerical Simulation of Three Dimensional Free Surface Flows. *Int. J. Num. Meth. Fluids*, 42(7):697–716, 2003.

[10] M. Meier, G. Yadigaroglu, and B. L. Smith. A Novel Technique for Including Surface Tension in PLIC-VOF Methods. *European Journal of Mechanics B—Fluids*, 21:61–73, 2002.

[11] S. Osher and R. P. Fedkiw. Level Set Methods: An Overview and Some Recent Results. *J. Comput. Phys.*, 169:463–502, 2001.

[12] O. Pironneau, J. Liou, and T. Tezduyar. Characteristic-Galerkin and Galerkin/Least-Squares Space-Time Formulations for the Advection-Diffusion Equation with Time-Dependent Domain. *Comput. Methods Appl. Mech. Eng.*, 100:117–141, 1992.

[13] M. Rappaz and V. Voller. Modeling of Micro-Segregation in Solidification Processes. *Metallurgical and Materials Transactions A*, 21A:749–753, 1990.

[14] Y. Renardy and M. Renardy. PROST: A Parabolic Reconstruction of Surface Tension for the Volume-of-Fluid Method. *J. Comput. Phys.*, 183:400–421, 2002.

[15] W.J. Rider and D.B. Kothe. Reconstructing Volume Tracking. *J. Comput. Phys.*, 141:112–152, 1998.

[16] R. Scardovelli and S. Zaleski. Direct Numerical Simulation of Free Surface and Interfacial Flows. *Annual Review of Fluid Mechanics*, 31:567–603, 1999.

[17] M. Schmid and F. Klein. Einflüss der Wandreibung auf das Füllverhalten Dünner Platten. Preprint, Steinbeis Transferzentrum, Fachhochschule Aachen, 1996.

[18] M. J. Shelley, F.-R. Tian, and K. Wlodarski. Hele-Shaw Flow and Pattern Formation in a Time-Dependent Gap. *Nonlinearity*, 10:1471–1495, 1997.

[19] M.W. Williams, D.B. Kothe, and E.G. Puckett. Accuracy and Convergence of Continuum Surface Tension Models. In *Fluid Dynamics at Interfaces*, W. Shyy and R. Narayanan, Eds., Cambridge University Press, New York, 1999, 294–305.

CHAPTER 6

Transonic regular reflection for the unsteady transonic small disturbance equation—details of the subsonic solution

Katarina Jegdić
Department of Mathematics, University of Houston, Houston, Texas

Barbara Lee Keyfitz
Fields Institute, Toronto, Ontario, Canada and
Department of Mathematics, University of Houston, Houston, Texas

Sunčica Čanić
Department of Mathematics, University of Houston, Houston, Texas

1 Introduction

We revisit and give a more detailed presentation of [3] by Čanić, Keyfitz, and Kim on a solution to a special Riemann problem for the unsteady transonic small disturbance (UTSD) equation. The Riemann initial data consist of two states in the upper half plane $\{(x, y) : x \in \mathbb{R},\ y \geq 0\}$ separated by an incident shock, and results in a regular reflection where the flow behind the reflected shock is subsonic. Written in self-similar coordinates $\xi = x/t$ and $\eta = y/t$, this configuration leads to a system that changes type. We find a solution in the hyperbolic part of the domain using the standard theory of one-dimensional conservation laws and the notion of quasi-one-dimensional Riemann problems developed in [1]. Solution in the elliptic part of the domain is described by a free boundary value problem. The free boundary is given by the position of the reflected shock which is, through the Rankine-Hugoniot relations, coupled to the subsonic state behind the shock. The main idea in solving this free boundary value problem is to fix the position of the reflected shock within some bounded set of admissible curves, solve the fixed boundary value problem, and then update the position of the reflected shock using the Rankine-Hugoniot relations.

The novelty of our paper is in the study of the fixed boundary value problem. We consider a more general class of fixed boundary value problems for which the operators in the domain and on the boundary satisfy certain structural conditions. The main tool is the theory developed in Gilbarg and Hormander [7], Gilbarg and Trudinger [8], Lieberman [10] through [13], and Lieberman and Trudinger [14].

1.1 Related work

This approach to solving Riemann problems for two-dimensional systems of hyperbolic conservation laws was first developed by Čanić, Keyfitz, and Lieberman [2] in a study of nonlinear stability of transonic shocks for the steady transonic small disturbance equation. The ideas have been extended to the cases of regular reflection for the UTSD equation: with a subsonic state behind the reflected shock in [3] and with a supersonic state immediately behind the reflected shock in [4]. A Riemann problem for the nonlinear wave system (NLWS), which leads to Mach reflection, is studied in [5].

The main features of this method in studying two-dimensional Riemann problems for a special class of systems of hyperbolic conservation laws (including the UTSD equation, the NLWS, the isentropic compressible gas dynamics equations, etc.) have been presented in Keyfitz [9]. We also mention the earlier work of Chang and Chen [6] in stating the free boundary value problems modeling regular reflection for the adiabatic gas dynamics equations.

1.2 Summary of the paper

In Section 2 we formulate a Riemann problem for the UTSD equation leading to a transonic regular reflection. We write the problem in self-similar coordinates (ξ, η) and obtain a system that changes type. We find a solution in the hyperbolic part of the domain and the equation of the reflected shock. The free boundary value problem is stated in Theorem 2.1 and the rest of the paper is devoted to finding its solution.

In Section 3 we change coordinates to $(\rho = \xi + \eta^2/4, \eta)$. We introduce several cutoff functions to ensure that the free boundary value problem in the elliptic part of the domain is well posed and suitable for applying the theory of second-order elliptic equations developed by Lieberman. This modified free boundary value problem is stated in Theorem 3.1. The main idea in finding its solution is to fix the position of the reflected shock within some set $\mathcal{K}$ of admissible curves, to solve the fixed boundary value problem, and to update the position of the reflected shock. This gives a mapping $J : \mathcal{K} \to \mathcal{K}$.

The fixed boundary value problem is studied in Section 4. We use the results in [7], [8] and [10] through [14], which are valid for a more general class of second-order boundary value problems as long as the operators in the domain and on the boundary have some desired properties. This observation motivates our study in Section 4.4 where, instead of considering only the fixed boundary value problem resulting from the transonic regular reflection for the UTSD equation, we consider a class of fixed boundary value problems satisfying certain structural conditions given in Section 4.3. This more general fixed boundary value problem is stated in Theorem 4.2. This is a nonlinear problem and first we find a solution to its linearized version in Section 4.4.1. Using the fixed-point theory we solve the nonlinear problem in Section 4.4.2.

In Section 5 we use the Schauder fixed-point theorem to show that the map J, defined on the set $\mathcal{K}$, has a fixed point. This completes the proof of Theorem 3.1.

Finally, the conditions under which the solution to the modified problem in Theorem 3.1 solves the original free boundary value problem of Theorem 2.1 are discussed in Section 6. Because not all of the cutoffs could be removed entirely, a solution to the original free boundary value problem is found only in a neighborhood of the reflection point.

2 The statement of the free boundary value problem

In this section we formulate, for the UTSD equation, a Riemann problem resulting in a transonic regular reflection. We study this phenomenon in self-similar coordinates, which yields a system of mixed type. Using the standard one-dimensional theory of hyperbolic conservation laws and the results on quasi-one-dimensional Riemann problems [1], we find a solution in the hyperbolic part of the region in Section 2.1. We formulate the free boundary problem in Theorem 2.1 and give the outline of its proof in Section 2.2.

Consider the UTSD equation

$$\begin{aligned} u_t + u\,u_x + v_y &= 0, \\ -v_x + u_y &= 0, \end{aligned} \tag{2.1}$$

where $U := (u,v) : [0,\infty)\times\mathbb{R}\times\mathbb{R} \to \mathbb{R}^2$. The Riemann initial data (Figure 6.1a) is given in the upper half plane $\{(x,y) : y \geq 0\}$ and consists of two states

$$U_0 = (0,0) \quad \text{and} \quad U_1 = (1,-a), \tag{2.2}$$

separated by a half line $x = a\,y$, $y \geq 0$, with a parameter

$$a > \sqrt{2} \tag{2.3}$$

fixed. We impose symmetry across the x-axis, meaning

$$u_y = v = 0 \quad \text{along} \quad y = 0. \tag{2.4}$$

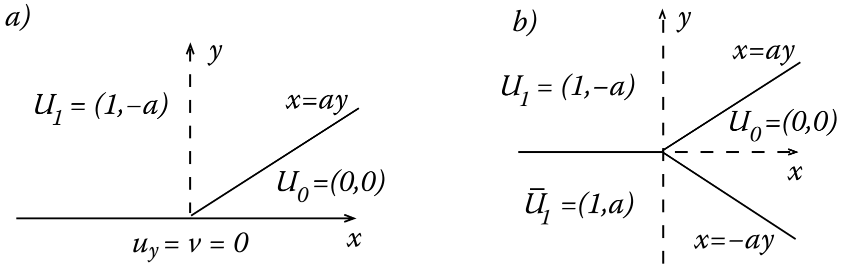


FIGURE 6.1 The Riemann initial data.

We note that the symmetric Riemann data (2.2), (2.4) posed in the upper half-plane is equivalent to the initial data given in three sectors in the full plane, as depicted in Figure 6.1b. In some parts of this study, it will be more convenient to consider the Riemann problem in the full plane with states $U_0 = (0,0)$, $U_1 = (1,-a)$ and $\overline{U}_1 = (1,a)$, instead of the original problems (2.1), (2.2), and (2.4) in the half-plane.

We study the initial-boundary value problems (2.1), (2.2) and (2.4) in self-similar coordinates $\xi = x/t$ and $\eta = y/t$. From (2.1) we get

$$\begin{aligned}(u-\xi)\,u_\xi - \eta\,u_\eta + v_\eta &= 0,\\ -v_\xi + u_\eta \qquad\qquad &= 0.\end{aligned} \tag{2.5}$$

It is clear that when the system (2.5) is linearized about a constant state $U = (u,v)$, the system is hyperbolic outside and elliptic inside the sonic parabola

$$P_U: \quad \xi + \frac{\eta^2}{4} = u. \tag{2.6}$$

Using the Rankine-Hugoniot conditions, the initial discontinuity $x = a\,y$ propagates as a shock given, in (ξ,η)-coordinates, by the equation $\xi = a\,\eta + a^2 + 1/2$. The initial-boundary value problems (2.1), (2.2) and (2.4) can be replaced by the system (2.5) with the following boundary conditions

$$\begin{aligned}&U(\xi,\eta) = U_0 \quad \text{on} \quad \{(\xi,\eta) : \xi + \eta^2/4 = C, \xi > a\,\eta + a^2 + 1/2, \eta > 0\},\\ &U(\xi,\eta) = U_1 \quad \text{on} \quad \{(\xi,\eta) : \xi + \eta^2/4 = C, \xi < a\,\eta + a^2 + 1/2, \eta > 0\},\\ &u_\eta = v = 0 \qquad \text{on} \quad \eta = 0,\end{aligned} \tag{2.7}$$

where C is a large positive constant. In the full plane, the equivalent boundary conditions are

$$\begin{aligned}U(\xi,\eta) = U_0 \text{ on } &\{(\xi,\eta) : \xi + \eta^2/4 = C,\ (\xi > a\,\eta + a^2 + 1/2,\ \eta \geq 0) \text{ or}\\ &\qquad\qquad\qquad\qquad (\xi < -a\eta + a^2 + 1/2,\ \eta \leq 0)\},\\ U(\xi,\eta) = U_1 \text{ on } &\{(\xi,\eta) : \xi + \eta^2/4 = C,\ \xi < a\,\eta + a^2 + 1/2,\ \eta > 0\},\\ U(\xi,\eta) = \overline{U}_1 \text{ on } &\{(\xi,\eta) : \xi + \eta^2/4 = C,\ \xi > -a\,\eta + a^2 + 1/2, \eta < 0\}.\end{aligned}$$

Let us further denote by P_0 and P_1 the sonic parabolas corresponding to the states U_0 and U_1, respectively. Notice that the sonic parabola for $\overline{U}_1$ coincides with P_1.

2.1 The solution in the hyperbolic region. The position of the reflected shock

In this part of the paper we briefly sketch the solution to the initial-boundary value problems (2.5) and (2.7) in the hyperbolic region using the notion of quasi-one-dimensional Riemann problems and we derive the position of the reflected shock. More details on how to solve a quasi-one-dimensional Riemann problem for the UTSD equation can be found in [1].

Let us denote the incident shock separating states U_0 and U_1 by

$$S : \xi = a\,\eta + a^2 + \frac{1}{2},\ \eta \geq 0.$$

By the choice of the parameter a (see (2.3)), the shock S does not interact with the parabola P_1. Let us denote by

$$\Xi_a := (\xi_a, 0) = \left(a^2 + \frac{1}{2}, 0\right) \tag{2.8}$$

the point where the shock hits the ξ-axis. We solve the quasi-one-dimensional Riemann problem at Ξ_a (for details, see [1]). The states on the left and on the right in this Riemann problem are $\overline{U}_1 = (1, a)$ and $U_1 = (1, -a)$, respectively. Because $a > \sqrt{2}$ there are two solutions to this problem, known as *weak* and *strong regular reflection*, each consisting of two shocks, one below and one above the ξ-axis. The intermediate states for the two solutions are given by

$$U_R = (1 + a^2 - a\sqrt{a^2 - 2}, 0) \quad \text{and} \quad U_F = (1 + a^2 + a\sqrt{a^2 - 2}, 0). \tag{2.9}$$

Here, the subscripts R and F stand for *reflected* and *fast reflected.* Let P_R and P_F denote the sonic parabolas for the states U_R and U_F, respectively. It is clear that the point Ξ_a is inside P_F for any choice of parameter $a > \sqrt{2}$. However, Ξ_a is inside P_R only if $a \in \left(\sqrt{2}, \sqrt{1 + \sqrt{5}/2}\right)$. In this paper we are interested in the case where the point of interaction of the shock S with the ξ-axis is inside the sonic parabola for the solution U at this point; namely, we study a *transonic regular reflection.* We denote the value of U at Ξ_a by $U_* = (u_*, v_*)$, and we choose

$$U_* = U(\Xi_a) := \begin{cases} U_R \text{ or } U_F, & \sqrt{2} < a < \sqrt{1 + \sqrt{5}/2} \\ U_F, & a \geq \sqrt{1 + \sqrt{5}/2}. \end{cases} \tag{2.10}$$

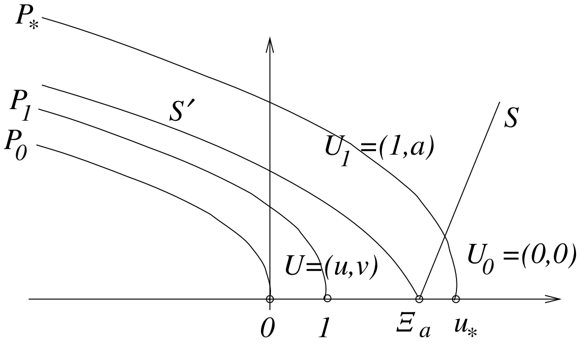


FIGURE 6.2 The position of the incident shock S and the reflected shock S' in (ξ, η)-coordinates.

Further, we denote the reflected shock by S' and the sonic parabola for the state U_* by P_*. Because the point Ξ_a is within the subsonic region determined by P_*, the shock S' is transonic (Figure 6.2). By causality, S' cannot cross P_*. The asymptotic analysis in [3] shows that S' approaches the sonic parabola P_1 as $\xi \to -\infty$.

Remark
If the reflected shock S' were rectilinear, its equation would be given by $\xi = k_* \eta + a^2 + 1/2$. Here, $k_* = k_R$ in the case of the solution with the intermediate state U_R, and $k_* = k_F$ when the intermediate state is U_F, with

$$k_R = -\frac{1}{a - \sqrt{a^2 - 2}} \quad \text{and} \quad k_F = -\frac{1}{a + \sqrt{a^2 - 2}}. \tag{2.11}$$

Let us assume that the reflected transonic shock S' is given by the following equation

$$\xi = \xi(\eta), \; \eta \geq 0. \tag{2.12}$$

We denote a solution of the system (2.5) behind the shock S' by $U = (u, v)$. Hence, the curve (2.12) satisfies the Rankine-Hugoniot condition with states U and U_1:

$$\frac{d\xi}{d\eta} = -\frac{[v]}{[u]} \quad \text{and} \quad \frac{d\xi}{d\eta} = \frac{[\frac{1}{2}u^2 - \xi u]}{[v - \eta u]}, \tag{2.13}$$

where $[\cdot]$ denotes the jumps across the shock. We eliminate v in (2.13) and obtain

$$\frac{d\xi}{d\eta} = -\frac{\eta}{2} - \sqrt{\xi + \frac{\eta^2}{4} - \frac{u+1}{2}}. \tag{2.14}$$

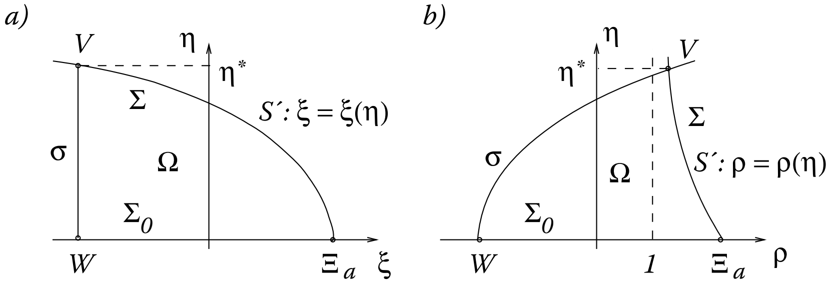


FIGURE 6.3 The domain Ω and its boundary.

The negative sign is appropriate here. Furthermore, by eliminating $d\xi/d\eta$ in (2.13), we obtain the following relation between u and v along the shock S'

$$v = -a + (u-1)\left(\frac{\eta}{2} + \sqrt{\xi + \frac{\eta^2}{4} - \frac{u+1}{2}}\right). \tag{2.15}$$

2.2 The statement of the main result and the outline of its proof

In this section we give the formulation of the free boundary value problem arising in the transonic regular reflection for the UTSD equation presented above.

First, we restrict the unbounded domain behind the reflected shock S'. More precisely, we introduce a cutoff parameter $\eta^* > 0$, which is fixed throughout the paper. We define $V := (\xi(\eta^*), \eta^*)$ and $W := (\xi(\eta^*), 0)$, the closed vertical line segment $\sigma := [V, W]$, the open horizontal line segment $\Sigma_0 := (W, \Xi_a)$ and the set $\Sigma := \{(\xi(\eta), \eta) : \eta \in (0, \eta^*)\}$, where $\xi(\eta)$, $\eta \geq 0$, is the unknown curve describing the position of the reflected shock S' (recall (2.12)). Further, we denote by Ω the domain whose boundary is $\partial\Omega = \Xi_a \cup \Sigma \cup \sigma \cup \Sigma_0$ (Figure 6.3a).

Next, we impose a Dirichlet condition $u = f$ along the vertical boundary σ. We assume that $f : \mathcal{R} \to \mathbb{R}$ is in the Holder space H_γ for a parameter $\gamma \in (0,1)$ to be determined later (for the definitions of Holder spaces see Section 4.1), where $\mathcal{R}$ is an open set containing σ, defined in Section 4.2. Moreover, we impose the following two conditions

$$\begin{aligned} &1 + \epsilon_* \leq f(\xi, \eta) \leq u_*, \qquad (\xi, \eta) \in \mathcal{R}, \\ &f(\xi(\eta^*), \eta) > \xi(\eta^*) + \tfrac{\eta^2}{4}, \eta \in [0, \eta^*], \end{aligned} \tag{2.16}$$

for an arbitrary parameter $\epsilon_* \in (0, u_* - 1)$, which is fixed throughout the paper.

Remark

Note that Σ, Σ_0, σ, the domain Ω, the size of the angles at the corners V and Ξ_a, and the boundary data f along σ depend on the unknown position of the reflected shock $S' : \xi = \xi(\eta)$. However, we will find $\xi(\eta)$ within a certain bounded set $\mathcal{K}$ (whose bounds depend only on $a > \sqrt{2}$, $\eta^* > 0$ and $\epsilon_* \in (0, u_* - 1)$) giving a priori bounds on Σ, and therefore also on σ, Σ_0, Ω and the angles at V and Ξ_a. In particular, the second condition in (2.16) will make sense. For the definition of the set $\mathcal{K}$, see Section 4.2.

With this notation we prove

Theorem 2.1

(*Free boundary value problem*)
Let $a > \sqrt{2}$, $\eta^* > 0$ and $\epsilon_* \in (0, u_* - 1)$, with u_* specified by (2.10), be given. Let f be any function in H_γ such that the inequalities (2.16) hold. There exists $\gamma_0 > 0$ depending on the parameters a, η_* and ϵ_*, such that for any $\gamma \in (0, \min\{\gamma_0, 1\})$ and $\alpha_{\mathcal{K}} = \gamma/2$, the problem

$$\left.\begin{aligned} (u-\xi)u_\xi - \eta u_\eta + v_\eta &= 0 \\ -v_\xi + u_\eta &= 0 \end{aligned}\right\} \quad \textit{in} \quad \Omega, \tag{2.17}$$

$$\left.\begin{aligned} \frac{d\xi}{d\eta} &= -\frac{\eta}{2} - \sqrt{\xi + \frac{\eta^2}{4} - \frac{u+1}{2}} \\ \frac{d\xi}{d\eta} &= -\frac{v+a}{u-1} \end{aligned}\right\} \quad \textit{on} \quad \Sigma, \tag{2.18}$$

$$\xi(0) = \xi_a, \tag{2.19}$$

$$v = u_\eta = 0 \quad \textit{on} \quad \Sigma_0, \tag{2.20}$$

$$u = f \quad \textit{on} \quad \sigma, \tag{2.21}$$

$$u(\Xi_a) = u_*, \tag{2.22}$$

has a solution $u, v \in H^{(-\gamma)}_{1+\alpha_*}$ in a finite neighborhood of Ξ_a, for all $\alpha_* \in (0, \alpha_{\mathcal{K}}]$. Moreover, the curve $\xi = \xi(\eta)$, $\eta \in (0, \eta^*)$, giving the location of the free boundary Σ, satisfies $\xi \in H_{1+\alpha_{\mathcal{K}}}$.

The Holder spaces are defined in Section 4.1. The outline of the proof of this theorem and the format of rest of the paper is as follows.

First, we change coordinates and consider the problems (2.17) through (2.22) in the more convenient (ρ, η)-coordinate system in Section 3.1. In order to use the elliptic theory by Gilbarg, Lieberman and Trudinger, we reformulate the problem using a second-order free boundary value problem for u and an equation for v in terms of u. We modify the problem so that it is strictly elliptic and well defined by introducing several cut-off functions in Section 3.3.

The main idea in solving this modified second-order free boundary value problem for $u(\rho,\eta)$ is: (1) fix the position of the reflected shock within a certain bounded set $\mathcal{K}$ of admissible curves, (2) find a solution of the fixed modified boundary value problem, and (3) update the position of the shock curve. This defines a mapping $J:\mathcal{K}\to\mathcal{K}$ for which we show there is a fixed point in Section 5.

Given a shock curve within the set $\mathcal{K}$, finding a solution to the fixed modified boundary value problem for $u(\rho,\eta)$ is a challenging task completed in Section 4. As already mentioned in the introduction, this part of our paper does not depend on the specific form of the fixed boundary value problem arising from the study of the UTSD equation, that is, the results in Section 4 apply to a more general class of operators satisfying certain structural conditions.

In Section 6 we discuss whether and how we could remove the cut-off functions introduced in Section 3.3 and we complete the proof of Theorem 2.1.

3 The modified problem

In this section we reformulate the free boundary value problem stated in Theorem 2.1 so that we can solve it using the techniques developed by Gilbarg, Lieberman and Trudinger. We write the problems (2.17) through (2.22) as a second-order problem for u and an equation for v. Instead of imposing the Rankine-Hugoniot conditions along the free boundary, we derive an oblique derivative boundary condition for u along Σ and a shock evolution equation. To make sure that the second-order problem for u is strictly elliptic, that the operator describing the boundary condition along Σ is strictly oblique, and that the shock evolution equation is well defined, we introduce several auxiliary cutoff functions. This modified problem is stated in Theorem 3.1.

3.1 The (ρ,η)-coordinate system

We define a new variable $\rho=\xi+\frac{\eta^2}{4}$ and in the rest of the paper we work in the (ρ,η)-coordinate system. For simplicity, we use the same notation for the domain Ω and its boundary in the (ρ,η)-coordinates as we do in Section 2.2 in the (ξ,η)-coordinates (Figure 6.3b). Under this change of variables, the system (2.5) becomes

$$\begin{aligned}(u-\rho)u_\rho-\tfrac{\eta}{2}u_\eta+v_\eta&=0,\\ \tfrac{\eta}{2}u_\rho-v_\rho+u_\eta\qquad&=0,\end{aligned}\tag{3.23}$$

Equation (2.14) implies

$$\frac{d\rho}{d\eta}=-\sqrt{\rho-\frac{u+1}{2}},\tag{3.24}$$

and from (2.15) we have

$$v = -a + (u-1)\left(\frac{\eta}{2} + \sqrt{\rho - \frac{u+1}{2}}\right). \tag{3.25}$$

On the other hand, by eliminating v in (3.23) we obtain the second-order equation for u

$$\left((u-\rho)u_\rho + \frac{u}{2}\right)_\rho + u_{\eta\eta} = 0, \tag{3.26}$$

and, from the first equation in (3.23), we can recover v in terms of u as

$$v(\rho,\eta) = \int_0^\eta \left\{\frac{y}{2}u_y - (u-\rho)u_\rho\right\} dy. \tag{3.27}$$

3.2 The oblique derivative boundary condition along the reflected shock

A condition of the form

$$\beta \cdot \nabla u = 0 \tag{3.28}$$

holds along the reflected shock S'. Here, $\nabla u = (u_\rho, u_\eta)$, $\beta = \beta(u,\rho,\rho') \in \mathbb{R}^2$ and $\rho' = d\rho/d\eta$. To obtain this, we differentiate Equation (3.25) along the shock S' and use Equations (3.23) to express the derivatives v_ρ and v_η in terms of u_ρ and u_η (for details of this calculation, see [3]). We obtain

$$\beta = \left(\rho'\left\{\frac{7u+1}{8} - \rho\right\}, \frac{5u+3}{8} - \rho\right). \tag{3.29}$$

3.3 Formulation of the modified free boundary value problem

In this section we reformulate the free boundary value problems (2.17) through (2.22) in (ρ,η) coordinates as a second-order elliptic free boundary value problem for $u(\rho,\eta)$ (using Equation (3.26)).

In Section 2.1, 3.1 and 3.2 we have shown that if $U = (u,v)$ is a solution to (3.23) in Ω, then the Rankine-Hugoniot condition (2.13) along the reflected shock S' implies Equation (3.24) for the position of S' and the oblique derivative relation (3.28) with β given by (3.29). The operations under which we derived (3.24) and (3.28) from (2.13) can be reversed up to a constant and, hence, if $U = (u,v)$ satisfies (3.24) and (3.28), and if (2.13) holds at one point on the reflected shock S', then the Rankine-Hugoniot condition (2.13) holds at each point along S'. Our idea here is that instead of imposing the condition (2.13) along the reflected shock, we require that the shock curve $\rho(\eta)$, $\eta \geq 0$, satisfies the differential Equation (3.24) with an initial condition $\rho(0) = \xi_a$, and that the solution $u(\rho,\eta)$ satisfies the oblique derivative boundary condition (3.28) along the free boundary $\Sigma = \{(\rho(\eta),\eta) : \eta \in (0,\eta^*)\}$.

To study the second-order Equation (3.26) in the domain Ω, we introduce three cut-off functions: a function ϕ to ensure that (3.26) is strictly elliptic, ψ to ensure that the shock evolution Equation (3.24) is well defined, and a function χ to ensure that the vector β in (3.28) is nowhere tangential to Σ.

Let us introduce the operator Q by

$$Q(u) := \left((u-\rho)u_\rho + \frac{u}{2}\right)_\rho + u_{\eta\eta} = (u-\rho)u_{\rho\rho} + u_{\eta\eta} - \frac{u_\rho}{2} + u_\rho^2.$$

To ensure strict ellipticity, we replace Q by the operator

$$\begin{aligned}\tilde{Q}(u) &:= \left(\phi(u-\rho)u_\rho + \frac{u}{2}\right)_\rho + u_{\eta\eta} \\ &= \phi(u-\rho)u_{\rho\rho} + u_{\eta\eta} + \left(\frac{1}{2} - \phi'(u-\rho)\right)u_\rho + \phi'(u-\rho)u_\rho^2. \qquad (3.30)\end{aligned}$$

Here, ϕ is a function given by

$$\phi(x) = \begin{cases} \delta, & x < \delta \\ x, & x \geq \delta, \end{cases} \qquad (3.31)$$

for some positive δ to be specified in Section 6. Because we will need ϕ' to be continuous in our study, we modify ϕ in a neighborhood of $x = \delta$ to be smooth and such that $\phi'(x) \in [0,1]$, for all $x \in \mathbb{R}$. Note that the operator $\tilde{Q}$ is strictly elliptic because

$$\lambda := \min\{\phi(u-\rho), 1\} \geq \min\{\delta, 1\} > 0. \qquad (3.32)$$

After we derive a priori bounds on a solution u to the problem $\tilde{Q}(u) = 0$ in Ω (see Lemma 4.2), we will show that the operator $\tilde{Q}$ is also uniformly elliptic, that is, the ellipticity ratio of $\tilde{Q}$ is bounded from above uniformly in u and $(\rho,\eta) \in \Omega$ (see Proposition 4.1).

To ensure that the nonlinear shock evolution Equation (3.24) is well defined, we replace it by

$$\frac{d\rho}{d\eta} = -\sqrt{\psi\left(\rho - \frac{u+1}{2}\right)}, \qquad (3.33)$$

with a function ψ given by

$$\psi(x) = \begin{cases} \delta_*, & x < \delta_* \\ x, & x \geq \delta_*. \end{cases} \qquad (3.34)$$

Here, $\delta_* > 0$ is a parameter to be chosen in Section 5. We will need ψ' to be continuous, and for that we modify ψ to be smooth in a neighborhood of $x = \delta_*$.

Finally, we define the operator N by

$$N(u) := \beta \cdot \nabla u, \tag{3.35}$$

where $\beta = \beta(u, \rho, \rho')$ is given by (3.29). Let $\nu := (-1, \rho')/\sqrt{1 + (\rho')^2}$ denote the unit inner normal to the boundary Σ. We compute

$$\beta \cdot \nu = \frac{-\rho'(\eta)\,(u-1)}{4\sqrt{1+(\rho')^2}}.$$

The definition of ψ implies $\rho'(\eta) \le -\sqrt{\delta_*} < 0$, for all $\eta \in [0, \eta_*]$, and therefore $\beta \cdot \nu = 0$ holds only if $u = 1$. Let us introduce a function $\chi : \mathbb{R}^3 \to \mathbb{R}^2$ by

$$\chi(u, \rho, \rho') = \begin{cases} (\rho' \{1 + 7\epsilon_*/8 - \rho\}, 1 + 5\epsilon_*/8 - \rho), & u < 1 + \epsilon_* \\ \beta(u, \rho, \rho'), & u \ge 1 + \epsilon_*, \end{cases} \tag{3.36}$$

where ϵ_* is the same positive parameter as in (2.16). As mentioned in Remark 2.2, we will assume that the reflected shock curve belongs to a certain admissible set of curves $\mathcal{K}$ (see Section 4.2), imposing a priori bounds on both $\rho(\eta)$ and $\rho'(\eta)$, $\eta \in (0, \eta^*)$, in terms of fixed parameters a, η^* and ϵ_*. This implies

$$\chi \cdot \nu \ge \min\left\{ \frac{1 + \sqrt{\delta_*}}{\sqrt{\xi_a}}, \frac{\sqrt{\delta_*}\,\epsilon_*}{4\sqrt{\xi_a}} \right\} > 0, \qquad \text{for all } u \in \mathbb{R} \text{ and } \rho \in \mathcal{K}. \tag{3.37}$$

We define the modified operator

$$\tilde{N}(u) := \chi \cdot \nabla u, \tag{3.38}$$

which is by (3.37) strictly oblique. After we show uniform a priori bounds on a solution u to the problem $\tilde{Q}(u) = 0$ in the domain Ω (see Lemma 4.2), we will show that in fact $\chi = \beta$, that is, the cutoff function χ can be removed and that we have $\tilde{N} = N$. Moreover, we will find a uniform lower bound on the obliqueness constant for the operator N, which will imply that the operator N is uniformly oblique (see Proposition 4.1).

We prove the following theorem for the modified free boundary problem and in Section 6 we discuss the removal of the remaining cutoff functions ϕ and ψ and we deduce Theorem 2.1.

Theorem 3.1

(*Modified free boundary value problem*)
Let $a > \sqrt{2}$, $\eta^* > 0$ and $\epsilon_* \in (0, u_* - 1)$ be given, and let $\delta > 0$ be arbitrary. Let f be any function in H_γ such that inequalities (2.16) hold. There exists

$\gamma_0 > 0$, depending on a, η_*, ϵ_* and δ, such that for any $\gamma \in (0, \min\{\gamma_0, 1\})$ and $\alpha_{\mathcal{K}} = \gamma/2$, the problems

$$\tilde{Q}(u) = \left(\phi(u-\rho)u_\rho + \frac{u}{2}\right)_\rho + u_{\eta\eta} = 0 \qquad in \quad \Omega, \tag{3.39}$$

$$\frac{d\rho}{d\eta} = -\sqrt{\psi\left(\rho - \frac{u+1}{2}\right)} \qquad on \quad \Sigma, \tag{3.40}$$

$$\rho(0) = \xi_a, \tag{3.41}$$

$$N(u) = \beta \cdot \nabla u = 0 \qquad on \quad \Sigma, \tag{3.42}$$

$$u_\eta = 0 \qquad on \quad \Sigma_0, \tag{3.43}$$

$$u = f \qquad on \quad \sigma, \tag{3.44}$$

$$u(\Xi_a) = u_*, \tag{3.45}$$

have a solution $u \in H^{(-\gamma)}_{1+\alpha_*}$, for all $\alpha_* \in (0, \alpha_{\mathcal{K}}]$. The function $\rho(\eta)$, $\eta \in (0, \eta^*)$, describing the position of the reflected shock satisfies $\rho \in H_{1+\alpha_{\mathcal{K}}}$.

4 The fixed boundary value problem

The goal of this section is to fix the function $\rho = \rho(\eta)$, $\eta \in (0, \eta^*)$, describing the free boundary Σ, within a certain set of admissible functions, and to solve the nonlinear fixed boundary problems (3.39), and (3.42) through (3.45).

In Section 4.1 we recall the definitions of Holder norms and spaces that we use in this paper. More details can be found in [7] and in Section 4 of [8]. We define the set $\mathcal{K}$ of admissible curves in Section 4.2, and in Section 4.4 we fix $\rho \in \mathcal{K}$ and consider the fixed boundary value problem. We remark that the results in section Section 4.4 rely heavily on the study of the second-order elliptic mixed boundary value problems (*mixed* meaning that we impose different types of boundary conditions—Dirichlet and oblique derivative) in Gilbarg & Trudinger [8], Lieberman [10] through [13] and Lieberman & Trudinger [14]. However, their results do not depend on the particular form of the elliptic operator $\tilde{Q}$ and the oblique derivative boundary operator $\tilde{N}$, defined in Equations (3.30) and (3.38), as long as these operators are strictly elliptic and strictly oblique, respectively. With this in mind, we consider a more general class of boundary value problems satisfying certain structural conditions. These conditions are given in Section 4.3. We solve the linearized version of the problem in Section 4.4.1 and then we use a fixed-point theorem to solve the nonlinear problem in Section 4.4.2.

4.1 Holder norms and Holder spaces

Let $S \subseteq \mathbb{R}^2$ be an open set and let $u: S \to \mathbb{R}$. We define the *supremum norm* for u on the set S to be

$$|u|_{0;S} := \sup_{x \in S} |u(x)|.$$

For $\alpha \in (0,1)$ we define the *Holder seminorm with exponent* α as

$$[u]_{\alpha;S} := \sup_{x,y \in S,\, x \neq y} \frac{|u(x)-u(y)|}{|x-y|^{\alpha}},$$

and the *Holder norm with exponent* α as

$$|u|_{\alpha;S} := |u|_{0;S} + [u]_{\alpha;S}.$$

Let k be a positive integer and $\alpha \in (0,1)$. We define the $(k+\alpha)$-*Holder norm* to be

$$|u|_{k+\alpha;S} := \sum_{j=0}^{k} |D^j u|_{0;S} + [D^k u]_{\alpha;S},$$

where D^j denotes the j-th order derivatives

$$\left\{ \frac{\partial^j}{\partial^{j_1} x\, \partial^{j_2} y} : j = j_1 + j_2,\ j_1, j_2 \geq 0 \right\}.$$

The space of functions for which the $(k+\alpha)$-Holder norm on the set S is finite is denoted by $H_{k+\alpha;S}$.

Remark
For the boundary condition on σ, we assume $u = f \in H_{\gamma;\mathcal{R}}$, where $\gamma \in (0,1)$. To simplify our notation, we write H_γ. Further, we show in Theorem 2.1 (also in Theorem 3.1) that the function $\xi(\eta)$ (or, equivalently, $\rho(\eta)$) is in the Holder space $H_{1+\alpha_{\mathcal{K}};(0,\eta^*)}$, where $\alpha_{\mathcal{K}} \in (0,1)$. For simplicity, we write $H_{1+\alpha_{\mathcal{K}}}$.

Further, let $T \subseteq \partial S$. For a fixed $\delta > 0$ we define the set

$$S_{\delta;T} := \{x \in S : \operatorname{dist}(x,T) > \delta\},$$

and for $a > 0$ and b such that $a - b \geq 0$, we define the *weighted interior norm* by

$$|u|^{(-b)}_{a;\overline{S}\setminus T} := \sup_{\delta > 0} \delta^{a-b} |u|_{a;S_{\delta;T}}. \tag{4.46}$$

The space of functions on the set S for which the weighted interior norm (4.46) is finite is denoted by $H^{(-b)}_{a;\overline{S}\setminus T}$.

Remark

In our study, the domain of interest is Ω and the distinguished part of the boundary is $\mathbf{V} := \{V, W, \Xi_a\}$, where V, W and Ξ_a are the corners introduced in Section 2.2. To simplify our notation instead of $H^{(-\gamma)}_{1+\alpha;\overline{\Omega}\setminus\mathbf{V}}$ we write $H^{(-\gamma)}_{1+\alpha}$.

Remark

- If $0 < a' < a$, it is easy to show that $[u]_{a'} \le C\,[u]_a$, for a constant C depending on a', a and the diameter of the domain S.
- If $0 < a' < a$, $0 < b' < b$, $a - b \ge 0$ and $a' - b' \ge 0$, we have ([7]): a bounded sequence in $H^{(-b)}_a$ is precompact in $H^{(-b')}_{a'}$, and there exists a constant C, independent of u, such that $|u|^{(-b')}_{a'} \le C\,|u|^{(-b)}_a$.

4.2 Definition of the set $\mathcal{K}$ of admissible curves

We consider the Banach space $H_{1+\alpha_{\mathcal{K}}}$, as in Remark 4.1, where $\alpha_{\mathcal{K}} \in (0,1)$ is a parameter that will be specified in Section 5. The admissible set $\mathcal{K}$ is defined so that the curve $\rho(\eta)$, $\eta \in [0,\eta^*]$, is in $\mathcal{K}$ if and only if the following four conditions hold

- *smoothness*: $\rho \in H_{1+\alpha_{\mathcal{K}}}$,
- *initial conditions*:
$$\rho(0) = \xi_a \quad \text{and} \quad \rho'(0) = k_*, \tag{4.47}$$
where ξ_a and k_* are given by (2.8) and (2.11), respectively,
- *monotonicity*:
$$-\sqrt{\xi_a - 1} \le \rho'(\eta) \le -\sqrt{\delta_*}, \qquad \text{for all } \eta \in (0,\eta^*), \tag{4.48}$$
with δ_* (see also (3.34)) to be specified in Section 5,
- *boundedness*:
$$\rho_L(\eta) \le \rho(\eta) \le \rho_R(\eta), \qquad \text{for all } \eta \in [0,\eta^*], \tag{4.49}$$
where the functions ρ_L and ρ_R will be also given in Section 5.

Remark

The parameter δ_* and the curves ρ_L and ρ_R will be given in terms of $a > \sqrt{2}$, $\eta^* > 0$ and $\epsilon_* \in (0, u_* - 1)$. By the definition (5.101) of δ_*, we have $\sqrt{\delta_*} < 1 < \sqrt{\xi_a - 1}$, for $a > \sqrt{2}$, and so the condition (4.48) makes sense.

Remark

Note that Ω, σ and Σ_0 depend on the choice of the curve $\rho \in \mathcal{K}$ describing the boundary Σ. Hence, the Holder estimates we derive in Section 4.4, which depend on the size of the domain Ω and its boundary, also depend on ρ. However, the set $\mathcal{K}$ is bounded in terms of the fixed parameters a, η^* and ϵ_*, implying a priori bounds on Σ, Ω, σ and Σ_0. Therefore, our estimates, which depend on the size of Ω or the parts of $\partial\Omega$, will be uniform in $\rho \in \mathcal{K}$. Furthermore, the monotonicity property (4.48) implies that the domain Ω satisfies the exterior cone condition defined in [8], (page 203), and that the angles of Ω at the corners V and Ξ_a are bounded both from below and from above uniformly in $\rho \in \mathcal{K}$.

Remark
We may also define the set $\mathcal{R}$ in terms of the bounds on ρ_L and ρ_R.

4.3 Structural conditions

As already remarked, once the curve $\rho \in \mathcal{K}$ describing the boundary Σ is fixed, the techniques for solving the nonlinear fixed boundary problems (3.39), and (3.42) through (3.45) do not depend on the specific definition of the operator $\tilde{Q}$ in Ω nor on the specific definition of the boundary conditions along $\partial\Omega$. In this section we define a more general class of fixed boundary value problems, which we will solve in Section 4.4.

Let $\Omega \subset \mathbb{R}^2$ be a bounded, open and connected set as in Figure 6.3b, so that $\partial\Omega = \Sigma \cup \Xi_a \cup \Sigma_0 \cup \sigma$, where Σ_0 is an open line segment aligned with the ξ-axis, Σ is given by an arbitrary curve $\rho \in \mathcal{K}$, where $\mathcal{K}$ is defined as in Section 4.2, Ξ_a is a corner and σ is a closed set. We assume that $\tilde{\Sigma} := \Sigma \cup \Sigma_0$ has an inner normal ν at each point, and that $\tilde{\Sigma}$ and $\partial\Omega \setminus \tilde{\Sigma} = \sigma \cup \Xi_a$ meet at the set of corners $\mathbf{V}$.

We consider the boundary value problem

$$\begin{array}{lll} \tilde{Q}(u) = 0 & \text{in} & \Omega, \\ \tilde{N}(u) = 0 & \text{on} & \tilde{\Sigma} = \Sigma \cup \Sigma_0, \\ u = \tilde{f} & \text{on} & \partial\Omega \setminus \tilde{\Sigma} = \sigma \cup \Xi_a. \end{array} \tag{4.50}$$

The operators $\tilde{Q}$ and $\tilde{N}$ are given by

$$\tilde{Q}(u) := \sum_{i,j} a_{ij}(u,\rho,\eta) D^{ij}u + \sum_i b_i(u,\rho,\eta) D^i u + \sum_{i,j} c_{ij}(u,\rho,\eta) D^i u D^j u, \tag{4.51}$$

and

$$\tilde{N}(u) = \chi(u,\rho',\rho,\eta) \cdot \nabla u. \tag{4.52}$$

The function $\tilde{f}$ is defined on $\mathcal{R} \cup \Xi_a$, where $\mathcal{R}$ is an open set containing σ. We assume that $\tilde{f}$ is in the Holder space $H_{\gamma;\mathcal{R}}$, for a parameter $\gamma \in (0,1)$ to be determined later, and that

$$m_1 \le \tilde{f} \le m_2 \quad \text{and} \quad \tilde{f} > \rho \text{ on } \sigma \tag{4.53}$$

hold for constants $m_1, m_2 \geq 0$, independent of $\rho \in \mathcal{K}$, $\tilde{f}$ and u. Further, we impose the following structural conditions.

- The coefficients a_{ij}, b_i and c_{ij} are in C^1, and for a fixed curve $\rho \in \mathcal{K}$ we have $\chi_i \in C^2$.
- The operator $\tilde{Q}$ is strictly elliptic, meaning
$$\lambda \geq C_1 > 0, \qquad \text{for all } u \text{ and } \rho \in \mathcal{K}, \tag{4.54}$$
where λ denotes the smallest eigenvalue of the operator $\tilde{Q}$. Moreover, we assume a bound on the ellipticity ratio of the form
$$\frac{\Lambda}{\lambda} \leq C_2(|u|_0), \qquad \text{for all } u \text{ and } \rho \in \mathcal{K}, \tag{4.55}$$
where $C_2(|u|_0)$ is a continuous function on $\mathbb{R}^+$. Here, Λ denotes the maximum eigenvalue of $\tilde{Q}$.
- The operator $\tilde{N}$ is strictly oblique, that is,
$$\chi \cdot \nu \geq C_3 > 0, \qquad \text{for all } u \text{ and } \rho \in \mathcal{K}. \tag{4.56}$$
Also,
$$|\chi| \leq C_4(|u|_0), \qquad \text{for all } u \text{ and } \rho \in \mathcal{K}, \tag{4.57}$$
holds, where $C_4(|u|_0)$ is a continuous function on $\mathbb{R}^+$.
- For any solution u to the equation $\tilde{Q}(u) = 0$ in Ω we have
$$0 \leq \sum_{i,j} c_{ij}(u, \rho, \eta) D^i u D^j u, \tag{4.58}$$
and there exist $\mu_0, \Phi \in \mathbb{R}$, independent of u, such that
$$|\sum_{i,j} a_{ij}(u, \rho, \eta) D^{ij} u| \leq \lambda \left(\mu_0 \sum_i |D^i u|^2 + \Phi \right). \tag{4.59}$$

Remark

Suppose that there is a uniform bound on the supremum norm $|u|_0$, where u is any solution to the equation $\tilde{Q}(u) = 0$ in Ω. Then

- the sup-norms $|a_{ij}|_0$, $|b_i|_0$, $|c_{ij}|_0$ and $|\chi_i|_0$ are uniformly bounded in u and $\rho \in \mathcal{K}$, and a uniform bound on the α-Holder seminorm $[u]_\alpha$ implies that $[a_{ij}]_\alpha$, $[b_i]_\alpha$, $[c_{ij}]_\alpha$ and $[\chi_i]_\alpha$ are uniformly bounded in u and $\rho \in \mathcal{K}$ (here, $\alpha \in (0, 1)$ is arbitrary),
- the operator $\tilde{Q}$ is uniformly elliptic, by (4.55),

- the inequality (4.57) implies a uniform upper bound on $|\chi|$, and using (4.56) we have

$$\frac{\chi \cdot \nu}{|\chi|} \geq \frac{C_3}{C_4(|u|_0)} > 0, \qquad \text{uniformly in } u \text{ and } \rho \in \mathcal{K},$$

  so that the operator $\tilde{N}$ is uniformly oblique with an obliqueness constant $C_3/C_4(|u|_0)$, and
- because the matrix $[a_{ij}(u, \rho, \eta)]$ is uniformly positive definite, and the coefficients $c_{ij}(u, \rho, \eta)$ are uniformly bounded, there exists $k > 0$, independent of u and $\rho \in \mathcal{K}$, such that

$$\sum_{i,j} c_{ij}(u, \rho, \eta) D^i u D^j u \leq k \sum_{i,j} a_{ij}(u, \rho, \eta) D^i u D^j u. \tag{4.60}$$

Before stating the general boundary value we will solve, Theorem 4.2, we verify that the problem for the UTSD equation satisfies these conditions.

Proposition 4.1

For any $\rho \in \mathcal{K}$ fixed, the boundary value problems (3.39) and (3.42) through (3.45) for the UTSD equation satisfy the structural conditions (4.53) through (4.58). Moreover, for any $k > 1/\delta$, where δ is a positive parameter in the definition (3.31) of the cutoff function ϕ, the inequality (4.60) holds.

Proof

The condition (4.53) holds with $m_1 := 1 + \epsilon_*$ and $m_2 := u_*$.

Recall the inequalities (3.32) and (3.37), and note that the operators $\tilde{Q}$ and $\tilde{N}$, defined by Equations (3.30) and (3.38), satisfy (4.54) and (4.56) with constants C_1 and C_3 depending on the parameters $a > \sqrt{2}$, $\eta^* > 0$ and $\epsilon_* \in (0, u_* - 1)$, which are fixed throughout the paper, and on δ_* and δ, which will be specified in Sections 5 and 6, respectively, also in terms of a, η^* and ϵ_*. The Neumann condition $(0, 1) \cdot \nabla u = 0$ on Σ_0 is obviously both strictly and uniformly oblique. Therefore, if u is a solution to (3.39), (3.42) through (3.45), Lemma (4.2) implies the uniform bounds $1 + \epsilon^* \leq u \leq u_*$. Hence, the definition (3.36) gives that $\chi = \beta$, for all u, and the operators N and $\tilde{N}$ are identical. Moreover, we have the following uniform bound on the ellipticity ratio for the operator $\tilde{Q}$

$$\begin{aligned}\frac{\Lambda}{\lambda} = \frac{\max\{\phi(u - \rho), 1\}}{\min\{\phi(u - \rho), 1\}} &\leq \frac{\max\{\delta, |u|_0 + |\rho|_0, 1\}}{\min\{\delta, 1\}} \\ &\leq \frac{\max\{\delta, u_* + \xi_a, 1\}}{\min\{\delta, 1\}},\end{aligned} \tag{4.61}$$

using a priori bounds on both u and $\rho \in \mathcal{K}$ (the left bound ρ_L in (4.49) will be chosen so that $\rho_L \geq 1$). Hence, the operator $\tilde{Q}$ is uniformly elliptic. Note that $|\beta(u)| \leq C\sqrt{1+(\rho')^2} \leq C\sqrt{\xi_a}$, for a constant C independent of u and $\rho \in \mathcal{K}$, using the definition (3.29) of β and a priori L^∞ bounds on u, ρ and ρ'. Therefore,

$$\frac{\beta \cdot \nu}{|\beta|} \geq \frac{\sqrt{\delta_*}\,\epsilon_*}{4\,C\,\sqrt{\xi_a}} > 0, \tag{4.62}$$

giving a lower bound for the obliqueness constant of the operator N uniformly in u and $\rho \in \mathcal{K}$.

Clearly, the coefficients a_{ij}, b_i and c_{ij} of the operator $\tilde{Q}$ and the coefficients β_i of the operator N have the desired smoothness and their sup-norms (or α-seminorms) are bounded using the uniform bounds on $|u|_0$ (or $[u]_\alpha$), $|\rho|_0$ and $|\rho'|_0$.

From (3.30) we have $c_{11}(u) = \phi'(u-\rho)$ and $c_{12} = c_{21} = c_{22} = 0$. Hence, $\phi'(u-\rho)u_\rho^2 \geq 0$, and, therefore, (4.58) holds.

Further, for a solution u to Equation (3.39) we have

$$\begin{aligned}
|\phi(u-\rho)u_{\rho\rho} + u_{\eta\eta}| &\leq |\phi'(u-\rho)||u_\rho|^2 + \left|\frac{1}{2} - \phi'(u-\rho)\right| |u_\rho| \\
&\leq |u_\rho|^2 + \frac{1}{2}|u_\rho| \leq \frac{3}{2}|u_\rho|^2 + \frac{1}{2} \\
&\leq \lambda\left(\frac{3}{2\min\{1,\delta\}}|u_\rho|^2 + \frac{1}{2\min\{1,\delta\}}\right),
\end{aligned}$$

implying that (4.59) holds.

Finally, for $k > 1/\delta$, where δ is a positive parameter in the definition of the cutoff function ϕ (see (3.31)) to be determined in Section 6, we have $k \geq \phi'/\phi$, and therefore

$$\phi'(u-\rho)u_\rho^2 \leq k\phi(u-\rho)u_\rho^2 \leq k\left\{\phi(u-\rho)u_\rho^2 + u_\eta^2\right\}.$$

Hence, (4.60) holds.

4.4 Solution to the fixed boundary value problem

In this section we prove Theorem 4.2.

Theorem 4.2

(Fixed boundary value problem)
Suppose that the domain Ω and the operators $\tilde{Q}$ and $\tilde{N}$ satisfy the structural conditions of Section 4.3. There exists $\gamma_0 > 0$, depending on the size of the opening angles of the domain Ω at the set of corners $\mathbf{V}$ and on the ellipticity

ratio of $\tilde{Q}$, such that for every $\gamma \in (0, \min\{\gamma_0, 1\})$, $\alpha_{\mathcal{K}} \in (0, \min\{1, 2\gamma\})$, $\rho \in \mathcal{K}$ and any function $\tilde{f}$, which is in H_γ on an open set containing σ and satisfies inequalities (4.53), there exists a solution u to the fixed boundary value problem (4.50). Moreover, $u \in H_{1+\alpha_*}^{(-\gamma)}$, for all $\alpha_* \in (0, \alpha_{\mathcal{K}}]$.

The proof of Theorem 4.2 is organized as follows. In Lemma 4.2 we assume that a solution to (4.50) exists in $C^1(\Omega)$, and we find its lower and upper bounds using the Maximum Principle and Hopf's Lemma. We solve the linearized problem in Section 4.4.1 and we use a fixed-point theorem to solve the nonlinear problem (4.50) in Section 4.4.2.

Lemma 4.2

Suppose that $u \in C^1(\Omega)$ solves the fixed boundary value problem

$$\begin{array}{lll} \tilde{Q}(u) = 0 & \mathit{in} & \Omega, \\ \tilde{N}(u) := \chi \cdot \nabla u = 0 & \mathit{on} & \tilde{\Sigma}, \\ u = \tilde{f} & \mathit{on} & \partial\Omega \setminus \tilde{\Sigma}, \end{array}$$

where $\Omega \subset \mathbb{R}^2$ is a bounded, open and connected set, $\tilde{\Sigma}$ is a finite disjoint union of relatively open sets with an inner normal at every point, and the operators $\tilde{Q}$ and $\tilde{N}$ are strictly elliptic and oblique, respectively. Then

$$\min_{\partial\Omega\setminus\tilde{\Sigma}} \tilde{f} \leq u(\rho, \eta) \leq \max_{\partial\Omega\setminus\tilde{\Sigma}} \tilde{f} \tag{4.63}$$

holds for all $(\rho, \eta) \in \Omega$.

Proof

Because the operator $\tilde{Q}$ is strictly elliptic, by the Maximum Principle, if u has an extremum at the point X, then $X \in \partial\Omega$. To show (4.63), it suffices to show $X \notin \tilde{\Sigma}$.

Suppose $X \in \tilde{\Sigma}$. Then the tangential derivative of u along this part of the boundary must be zero, because X is also an extremum of the function restricted to the boundary. On the other hand, the derivative $\chi \cdot \nabla u$ is zero along $\tilde{\Sigma}$. Because the operator $\tilde{N}$ is oblique, the vector χ is not tangential to $\tilde{\Sigma}$. Therefore, if $X \in \tilde{\Sigma}$, the derivative of u at X is zero in two different directions. This yields $\nabla u(X) = 0$ and finally contradicts Hopf's Lemma (Lemma 3.4 in [8]). Hence, $X \in \partial\Omega \setminus \tilde{\Sigma}$.

4.4.1 The linear problem

In this part of our study we solve the linearized version of the fixed boundary value problem (4.50), under conditions (4.53) through (4.59), using Theorem 2.1 in [11], and we further derive estimates on its solution using Theorem 2.1 in [12].

Let $\alpha_{\mathcal{K}} \in (0,1)$ and $\rho \in \mathcal{K}$ be fixed. Let $\tilde{f}$ be in H_γ on an open set containing σ, for an arbitrary $\gamma \in (0,1)$, and suppose that inequalities (4.53) hold. Let $\gamma_1 \in (0,1)$ and $\epsilon \in (0,\alpha_{\mathcal{K}}]$ also be arbitrary, and let $z \in H_{1+\epsilon}^{(-\gamma_1)}$ be any function such that $m_1 \le z \le m_2$. The role of parameters γ_1 and ϵ is to establish the compactness needed in the study of the nonlinear problem in the next section (see Lemma 4.5). We define the linear operators

$$\begin{aligned} Lu &:= \sum_{i,j} a_{ij}(z,\rho,\eta)D^{ij}u + \sum_i b_i(z,\rho,\eta)D^i u + \sum_{i,j} c_{ij}(z,\rho,\eta)D^i z D^j u \\ &= \sum_{i,j} a_{ij}(z,\rho,\eta)D^{ij}u + \sum_i \left\{ b_i(z,\rho,\eta) + \sum_j c_{ji}(z,\rho,\eta)D^j z \right\} D^i u \qquad (4.64) \end{aligned}$$

in Ω, (4.65)

and

$$Mu := \chi(z,\rho',\rho,\eta)\cdot\nabla u \quad \text{on} \quad \tilde{\Sigma}. \qquad (4.66)$$

For convenience, we introduce a new function

$$\tilde{u}(\rho,\eta) := u(\rho,\eta) - \tilde{f}(\Xi_a), \quad (\rho,\eta)\in\Omega, \qquad (4.67)$$

and consider the linear fixed boundary value problem

$$\begin{array}{lll} L\tilde{u} = 0 & \text{in} & \Omega, \\ M\tilde{u} = 0 & \text{on} & \tilde{\Sigma} = \Sigma \cup \Sigma_0, \\ \tilde{u} = \tilde{f} - \tilde{f}(\Xi_a) & \text{on} & \partial\Omega \setminus \tilde{\Sigma} = \sigma \cup \Xi_a. \end{array} \qquad (4.68)$$

Theorem 4.3

(Linear problem)
Suppose that the domain Ω and the operators $\tilde{Q}$ and $\tilde{N}$ satisfy the conditions of Section 4.3. Let $\alpha_{\mathcal{K}} \in (0,1)$ and $\rho \in \mathcal{K}$ be fixed, and let $\tilde{f}$ be any function that is in H_γ on an open set containing σ and satisfies (4.53). Suppose that $z \in H_{1+\epsilon}^{(-\gamma_1)}$, for arbitrary parameters $\gamma_1 \in (0,1)$ and $\epsilon \in (0,\alpha_{\mathcal{K}}]$, is any function such that

$$m_1 \le z(\rho,\eta) \le m_2, \quad (\rho,\eta)\in\Omega, \qquad (4.69)$$

where m_1 and m_2 are as in (4.53), and there exists a constant m so that

$$|b_i(z,\rho,\eta) + \sum_j c_{ji}(z,\rho,\eta)D^j z| \le m\, d_{\mathbf{V}}^{\gamma_1 - 1}(\rho,\eta). \qquad (4.70)$$

Here $\mathbf{V} := \{V_1, V_2, V_3\}$ denotes the set of corners of Ω, $d_{V_i}(X) := |X - V_i|$ and $d_{\mathbf{V}}(X) := \min_i d_{V_i}(X)$.

Then there exists $\gamma_0 > 0$, depending on the geometry of Ω and the supremum norm $|z|_0$, so that for any $\gamma \in (0, \min\{\gamma_0, 1\})$, there exists a unique solution $\tilde{u} \in H^{(-\gamma)}_{1+\alpha_{\mathcal{K}}}$ of the linear problem (4.68).

Moreover, the following two estimates hold

$$|\tilde{u}|^{(-\gamma)}_{1+\alpha_{\mathcal{K}}} \le C \left\{ |\tilde{f} - \tilde{f}(\Xi_a)|_\gamma + \sup_{i,(\rho,\eta)\in\Omega} d^{-\gamma}_{V_i}(\rho,\eta)|\tilde{u}(\rho,\eta) - \tilde{u}(V_i)| \right\} \qquad (4.71)$$

and

$$|\tilde{u}|^{(-\gamma)}_{1+\alpha_{\mathcal{K}}} \le C_1 \, |\tilde{f} - \tilde{f}(\Xi_a)|_\gamma, \qquad (4.72)$$

where C and C_1 depend on $\alpha_{\mathcal{K}}$, $[z]_{\alpha_{\mathcal{K}}}$, $[\chi_i(z,\rho,\rho')]_{\alpha_{\mathcal{K}}}$, $|\rho|_{1+\alpha_{\mathcal{K}}}$, m, $|z|_0$ and the size of the domain Ω.

Remark

- We are assuming that the function ρ describing the boundary $\Sigma \subset \tilde{\Sigma}$ is such that $\rho \in H_{1+\alpha_{\mathcal{K}}}$, where $\alpha_{\mathcal{K}} \in (0,1)$. To show existence of a solution, we will use Theorem 2.1 of [11] which assumes more smoothness on ρ. More precisely, this theorem requires that the boundary $\tilde{\Sigma}$, along which we pose the oblique derivative boundary condition, is described by a curve in $H_{2+\alpha}$, with $\alpha \in (0,1)$. This is satisfied for $\Sigma_0 \subset \tilde{\Sigma}$, and our idea is to approximate Σ with a sequence of boundaries Σ_k described by smooth curves $\{\rho_k\} \subset \mathcal{K}$ and to solve the linear problem (4.68) as a limit of problems on regularized domains Ω_k.
- The parameter γ_0 (and therefore γ) in the statement of the theorem depends on the size of angles of the domain Ω at the set of corners $\mathbf{V}$ and on the ellipticity ratio of the linear operator L. By the choice of set $\mathcal{K}$, these angles are bounded uniformly in $\rho \in \mathcal{K}$ (recall Remark 4.2). Also, the ellipticity ratio for L is uniformly bounded, with respect to z, using the assumption (4.55) for $\tilde{Q}$ and condition (4.69) for the choice of z. Hence, γ_0 (and also γ) can be taken independent of both the domain Ω and the function z.
- To find the parameter γ_0 we will use Theorem 1 in [12]. This theorem assumes that the operator M is uniformly oblique. By assumptions (4.56) and (4.57) on the operator $\tilde{N}$, the lower bound on the obliqueness constant for M depends on $|z|_0$, which is uniformly bounded by the condition (4.69).

Proof

(of Theorem 4.3)

We divide the proof into four steps.

Step 1. Let $\tilde{u} \in H^{(-\gamma)}_{1+\alpha_{\mathcal{K}}}$ be a solution to (4.68). Using the standard elliptic theory (for example, Theorem 6.2 in [8]), we have $u \in C^2(\Omega)$, and by Lemma 4.2 we obtain the L^∞ estimate

$$|\tilde{u}|_{0;\Omega} \leq |\tilde{f} - \tilde{f}(\Xi_a)|_0. \tag{4.73}$$

Step 2. In this step we prove that if $\tilde{u} \in H^{(-\gamma)}_{1+\alpha_{\mathcal{K}}}$, for an arbitrary $\gamma \in (0,1)$ and an arbitrary $\tilde{f} \in H_\gamma$ on an open set containing σ satisfying (4.53), is a solution to the linear problem (4.68) with boundary Σ described by a smooth curve $\rho \in \mathcal{K}$, then the estimate (4.71) holds. First, we derive weighted local estimates on a particular seminorm of the first derivatives of $\tilde{u}$ inside the domain Ω and on the boundary $\partial\Omega \setminus \mathbf{V}$, where we pose different types of boundary conditions (the oblique derivative boundary condition on $\tilde{\Sigma} = \Sigma \cup \Sigma_0$ and the Dirichlet condition on σ). These local estimates follow from Gilbarg and Trudinger [8], and together with interpolation inequalities establish (4.71). The estimate (4.72) is deduced from (4.71) for the parameter γ sufficiently small using Theorem 1 of [12].

We claim that there exists a constant C, independent of $\tilde{u}$ and the choice of $\rho \in \mathcal{K}$, such that the auxiliary inequality

$$R^{1+\alpha_{\mathcal{K}}}[D\tilde{u}]_{\alpha_{\mathcal{K}};B_R(x_0)\cap\Omega} \leq C\left\{R^\gamma|\tilde{f} - \tilde{f}(\Xi_a)|_\gamma \right.$$
$$\left. + \sup_{i,(\rho,\eta)\in\Omega} |\tilde{u} - \tilde{u}(V_i)|_{0;B_{2R}(x_0)\cap\Omega}\right\} \tag{4.74}$$

holds in the following three cases

1. $x_0 \in \sigma$ and $B_{2R}(x_0) \cap \tilde{\Sigma} = \emptyset$,
2. $x_0 \in \tilde{\Sigma}$ and $B_{2R}(x_0) \cap \sigma = \emptyset$, and
3. $B_{2R}(x_0) \subseteq \Omega$.

To show (4.74) in case (1) we use the discussion on page 139 in [8] for elliptic problems with Dirichlet boundary conditions. It implies the estimate

$$R^{1+\alpha_{\mathcal{K}}-\gamma}[D\tilde{u}]_{\alpha_{\mathcal{K}};B_R(x_0)\cap\Omega} \leq C\{|\tilde{f} - \tilde{f}(\Xi_a)|_\gamma + |\tilde{u}|_{0;B_R(x_0)\cap\Omega}\}, \tag{4.75}$$

with a constant C depending on $\alpha_{\mathcal{K}}$, the domain Ω and the norms of the coefficients of the operator L defined in (4.64). Using the a priori bounds on the set $\mathcal{K}$ and conditions (4.69) and (4.70), we have that the constant C in (4.75) does not depend on the solution $\tilde{u}$ nor on the choice of $\rho \in \mathcal{K}$ describing the boundary Σ. The estimate (4.75) together with the L^∞ bound (4.73) gives

$$R^{1+\alpha_{\mathcal{K}}-\gamma}[D\tilde{u}]_{\alpha_{\mathcal{K}};B_R(x_0)\cap\Omega} \leq C\,|\tilde{f} - \tilde{f}(\Xi_a)|_\gamma,$$

and, clearly, (4.74) follows.

In case (2) we use Theorem 6.26 in [8] for the oblique derivative boundary value problems. For convenience we consider this theorem for the functions $\tilde{u} - \tilde{u}(V_i)$, $i \in \{1, 2, 3\}$, also satisfying the linear differential equation in Ω and the linear oblique derivative boundary condition on $\tilde{\Sigma}$. For each i we obtain the estimate

$$\begin{aligned} R^{1+\alpha_{\mathcal{K}}}[D\tilde{u}]_{\alpha_{\mathcal{K}};B_R(x_0)\cap\Omega} &= R^{1+\alpha_{\mathcal{K}}}[D(\tilde{u} - \tilde{u}(V_i))]_{\alpha_{\mathcal{K}};B_R(x_0)\cap\Omega} \\ &\leq C\,|\tilde{u} - \tilde{u}(V_i)|_{0;B_{2R}(x_0)\cap\Omega}, \end{aligned}$$

with a constant C depending on $\alpha_{\mathcal{K}}$ and the bound on the obliqueness constant of the linear oblique derivative operator M. Therefore, (4.74) holds with C depending on the parameter $\alpha_{\mathcal{K}}$ and the supremum norm $|z|_0$.

In case (3) we use Theorem 8.32 in [8]. Again we use this theorem for functions $\tilde{u} - \tilde{u}(V_i)$, $i \in \{1, 2, 3\}$, and solutions to the differential equation $L\tilde{u} = 0$ in Ω. For each i, we obtain

$$\begin{aligned} [D\tilde{u}]_{\alpha_{\mathcal{K}};B_R(x_0)\cap\Omega} &= [D(\tilde{u} - \tilde{u}(V_i))]_{\alpha_{\mathcal{K}};B_R(x_0)\cap\Omega} \\ &\leq C\,|\tilde{u} - \tilde{u}(V_i)|_{0;B_{2R}(x_0)}, \end{aligned}$$

and (4.74) follows for $R \in (0, 1]$. Here, the constant C depends on supremum norms of the coefficients of L, $\alpha_{\mathcal{K}}$-seminorms of the coefficients of L and M, and the size of the domain Ω.

Note that cases (1) through (3) cover all points $x_0 \in \overline{\Omega} \setminus \mathbf{V}$. Let $R := \tan\left(\frac{\theta}{4}\right) \frac{d_{\mathbf{V}}(x_0)}{\text{diam}\Omega}$, where θ stands for the corner angle. We multiply the estimate (4.74) by $R^{-\gamma}$ and use the interpolation inequalities (6.8) through (6.9) in [8] to obtain (4.71).

Next, we derive the estimate (4.72) from (4.71). Note that for each $i \in \{1, 2, 3\}$ we have

$$\sup_{(\rho,\eta)} d_{V_i}(\rho, \eta)^{-\gamma}|\tilde{u}(\rho, \eta) - \tilde{u}(V_i)| \leq |\tilde{u}|_{\gamma;\Omega\setminus\mathbf{V}}. \tag{4.76}$$

On the other hand, Theorem 3.1 of [12] indicates that there exist positive constants γ_0 and a_0, depending on the size of angles of the domain Ω at the set of corners $\mathbf{V}$ and on the ellipticity ratio for the operator L, such that for all $\gamma \in (0, \gamma_0)$ and all $a \in (1, 1 + a_0)$ we have

$$|\tilde{u}|_a^{(-\gamma)} \leq C\{|\tilde{f} - \tilde{f}(\Xi_a)|_\gamma + |\tilde{u}|_0\}.$$

We fix $\gamma \in (0, \min\{\gamma_0, 1\})$. Because

$$|\tilde{u}|_{\gamma;\Omega\setminus\mathbf{V}} = |\tilde{u}|_\gamma^{(-\gamma)} \leq C(a, \gamma, \text{diam}(\Omega))|\tilde{u}|_a^{(-\gamma)},$$

holds for any $a > \gamma$, we obtain

$$|\tilde{u}|_{\gamma;\Omega\setminus\mathbf{V}} \leq C\{|\tilde{f} - \tilde{f}(\Xi_a)|_\gamma + |\tilde{u}|_0\} \leq C|\tilde{f} - \tilde{f}(\Xi_a)|_\gamma,$$

using the L^∞ bound (4.73). Together with (4.71) and (4.76), this establishes the estimate (4.72).

Step 3. We approximate the boundary Σ, specified by $\rho \in \mathcal{K}$, by a sequence of boundaries $\{\Sigma_k\}$ given by smooth curves $\rho_k \in \mathcal{K}$. This leads to a sequence $\{\Omega_k\}$ of the domains approximating the domain Ω, and the sequences $\{\sigma_k\}$ and $\{\Sigma_{0,k}\}$ approximating boundaries σ and Σ_0, respectively. Let $\tilde{f}$ be in H_γ on an open set $\mathcal{R}$ containing σ, for $\gamma \in (0,1)$ arbitrary, such that the inequalities (4.53) hold. Because $\tilde{f}$ satisfies the second inequality in (4.53) for ρ and σ, by continuity we have that $\tilde{f}$ satisfies the second condition in (4.53) for ρ_k and σ_k, where $k \geq k_0$. We use Theorem 3.1 of [11] for the linear problem (4.68) in Ω_k and get a unique solution $\tilde{u}_k \in C^2(\Omega_k \cup \Sigma_k) \cap C(\overline{\Omega}_k)$. By step 2 we have that $\tilde{u}_k \in H^{(-\gamma)}_{1+\alpha_\mathcal{K}}$ and that the sequence $\{\tilde{u}_k\}$ satisfies the estimates (4.71) and (4.72) uniformly in k. (Recall Remark 4.4.1 and that because we have uniform bounds on the geometry of the domains Ω_k, we can take both parameters γ_0 and γ independent of k.)

Step 4. In this step we show that the sequence $\{\tilde{u}_k\}$ has a convergent subsequence and that its limit is a unique solution to the linear boundary value problem (4.68) in the domain Ω.

As $k \to \infty$, we have $\Sigma_k \to \Sigma$, $\Omega_k \to \Omega$ and $\sigma_k \to \sigma$. Because the estimate (4.72) holds uniformly in k, the sequence $\{\tilde{u}_k\}$ is uniformly bounded in $H^{(-\gamma)}_{1+\alpha_\mathcal{K}}$ and by the Arzela-Ascoli Theorem, it contains a subsequence $\{\tilde{u}_{j_k}\}$ that converges uniformly to a function $\tilde{u} \in H^{(-\gamma)}_{1+\alpha_\mathcal{K}}$. It is clear that the estimates (4.71) and (4.72) also hold for $\tilde{u}$ with the same constants C and C_1.

Next, we show that $\tilde{u}$ is a solution of the linear boundary value problem (4.68). Because $|\tilde{u}_k|^{(-\gamma)}_{1+\alpha_\mathcal{K}}$ is uniformly bounded, $\tilde{u}_k$ and $D\tilde{u}_k$ are equicontinuous on compact subsets of Ω, implying that $\tilde{u}$ satisfies the differential equation $L\tilde{u} = 0$ weakly in the domain Ω. Further, let $x_0 \in \Sigma$ and let $x_k \in \Sigma_k$ be such that $x_k \to x_0$. By the uniform convergence of equicontinuous sequences in $H^{(-\gamma)}_{1+\alpha_\mathcal{K}}$ we obtain that $\chi(z, \rho_{j_k}, \rho'_{j_k}) \cdot \nabla\tilde{u}_{j_k}(x_{j_k}) \to \chi(z, \rho, \rho')\cdot\nabla\tilde{u}(x_0)$. Hence, the oblique derivative boundary condition $M\tilde{u} = 0$ holds on Σ. Similarly, the oblique derivative boundary condition on Σ_0 holds. The Dirichlet condition at the corner point Ξ_a is clearly satisfied, and to show the Dirichlet condition on σ we also use continuity of $\tilde{f}$ in an open set containing σ and σ_k, for $k \geq k_0$ and k_0 is sufficiently large. Therefore, the function $\tilde{u}$ solves the linear problem (4.68) in the domain Ω.

Because $\tilde{u} \in C^2(\Omega)$, we use Lemma 4.2 and the linearity of the operators L and M to conclude that $\tilde{u}$ is the unique solution of the linear problem (4.68).

We note that uniqueness of the solution $\tilde{u}$ implies that the whole sequence $\{\tilde{u}_k\}$ in the previous proof converges.

4.4.2 The nonlinear problem

Let $\alpha_{\mathcal{K}} \in (0,1)$ and $\rho \in \mathcal{K}$ be given. Let $\gamma_1 \in (0,1)$ and $\epsilon \in (0, \alpha_{\mathcal{K}}]$ be arbitrary. By Theorem 4.3 and Remark 4.4.1, there exists a parameter $\gamma_0 > 0$ such that for any fixed $\gamma \in (0, \min\{\gamma_0, 1\})$, any function $z \in H_{1+\epsilon}^{(-\gamma_1)}$ satisfying conditions (4.69) and (4.70) and any function $\tilde{f} \in H_\gamma$ on an open set containing σ and satisfying (4.53), there exists a unique solution $\tilde{u} \in H_{1+\alpha_{\mathcal{K}}}^{(-\gamma)}$ to the linear problem (4.68). Let us define a mapping T so that

$$Tz := \tilde{u} + \tilde{f}(\Xi_a). \tag{4.77}$$

In this section we show that we can choose the parameter $\alpha_{\mathcal{K}} \in (0,1)$, depending on γ, so that the mapping T has a fixed point. This will complete the proof of Theorem 4.2.

The fixed point result we use is the following.

Theorem 4.4

(Theorem 11.3 in [8])
Let T be a compact mapping of a Banach space $\mathcal{B}$ into itself and suppose that there exists a constant M such that

$$\|u\|_{\mathcal{B}} \le M, \text{ for all } u \in \mathcal{B} \text{ and } \tau \in [0,1] \text{ satisfying } u = \tau T u. \tag{4.78}$$

Then T has a fixed point.

The verification of the conditions of this fixed point theorem consists of two parts. In Lemma 4.5 we select an appropriate Banach space $\mathcal{B}$ so that $T(\mathcal{B}) \subseteq \mathcal{B}$ and so that the mapping T is compact. Using Lemma 4.6 we choose the parameter $\alpha_{\mathcal{K}} \in (0,1)$, in terms of γ, so that there exists M, independent of $\tilde{u}$, for which the inequality (4.78) holds.

Lemma 4.5

Let $\gamma \in (0, \min\{\gamma_0, 1\})$, where $\gamma_0 > 0$ as in Theorem 4.3, and let $\alpha_{\mathcal{K}} \in (0,1)$ be arbitrary. If

$$\epsilon = \frac{\alpha_{\mathcal{K}}}{2} \quad \text{and} \quad \gamma_1 = \frac{\gamma}{2}, \tag{4.79}$$

then for $\mathcal{B} := H_{1+\epsilon}^{(-\gamma_1)}$, the mapping T given by (4.77) is precompact and $T(\mathcal{B}) \subseteq \mathcal{B}$.

Proof

Theorem 4.3 implies $T(H_{1+\epsilon}^{(-\gamma_1)}) \subseteq H_{1+\alpha_{\mathcal{K}}}^{(-\gamma)}$. To ensure $T(\mathcal{B}) \subseteq \mathcal{B}$, we choose γ_1 and ϵ, so that $0 < \epsilon \le \alpha_{\mathcal{K}}$ and $0 < \gamma_1 \le \gamma$. In order for the map T to be compact we need these inequalities to be strict (see Remark 4.1), and in particular the choices (4.79) suffice.

Lemma 4.6

Let $\gamma \in (0, \min\{\gamma_0, 1\})$, where $\gamma_0 > 0$ as in Theorem 4.3, and let $\alpha_{\mathcal{K}} \in (0,1)$ be arbitrary. Let ϵ and γ_1 be as in (4.79). There exists $M > 0$ such that if $\tilde{u} \in H^{(-\gamma_1)}_{1+\epsilon}$ and

$$\tilde{u} + \tilde{f}(\Xi_a) = \tau T(\tilde{u} + \tilde{f}(\Xi_a)), \tag{4.80}$$

for some $\tau \in [0,1]$, then

$$|\tilde{u} + \tilde{f}(\Xi_a)|^{(-\gamma)}_{1+\alpha_*} \le M, \tag{4.81}$$

where $\alpha_* := \min\{\alpha_{\mathcal{K}}, \gamma\}$. The constant M depends on the geometry of Ω, the bounds on the ellipticity ratio and the minimal eigenvalue of the operator $\tilde{Q}$, the bound on the obliqueness constant for $\tilde{N}$, the sup-norm $|\tilde{u}|_0$ and the Holder norm $|\tilde{f}|_{\gamma;\mathcal{R}}$.

Proof

We divide the proof into four steps.

Step 1. In this step we obtain an L^∞ bound on $\tilde{u} \in H^{(-\gamma_1)}_{1+\epsilon}$ satisfying (4.80).

Using the definition (4.77) of the map T, the assumption (4.80) implies that $\tilde{u}$ solves the following nonlinear fixed boundary problem

$$\begin{aligned} &\sum_{i,j} a_{ij}(\tilde{u} + \tilde{f}(\Xi_a))D^{ij}\tilde{u} + \sum_i b_i(\tilde{u} + \tilde{f}(\Xi_a))D^i\tilde{u} \\ &\qquad + \sum_{i,j} c_{ij}(\tilde{u} + \tilde{f}(\Xi_a))D^i(\tilde{u} + \tilde{f}(\Xi_a))D^j\tilde{u} = 0 \quad \text{in} \quad \Omega, \\ &\qquad \chi(\tilde{u} + \tilde{f}(\Xi_a)) \cdot \nabla\tilde{u} = 0 \quad \text{on} \quad \tilde{\Sigma}, \\ &\qquad \tilde{u} = \tau(\tilde{f} - \tilde{f}(\Xi_a)) \quad \text{on} \quad \partial\Omega \setminus \tilde{\Sigma}. \end{aligned} \tag{4.82}$$

Because we have $\tilde{u} \in C^2(\Omega)$, by Lemma 4.2 we obtain the L^∞ bound

$$|\tilde{u}|_0 \le \tau|\tilde{f} - \tilde{f}(\Xi_a)|_0 \le |\tilde{f} - \tilde{f}(\Xi_a)|_0. \tag{4.83}$$

Step 2. In this part of the proof we find an estimate for the term

$$\sup_{i,(\rho,\eta)\in\Omega} d^{-\gamma}_{V_i}(\rho,\eta)|\tilde{u}(\rho,\eta) - \tilde{u}(V_i)|$$

which is independent of $\tilde{u}$. The idea is to construct two linear problems independent of $\tilde{u}$ (with the same linear elliptic and linear oblique derivative boundary operators) and to show that we can bound $\tilde{u}$ both from below and from above using the solutions of these linear problems. Once these bounds are established, we use Lemma 4.1 in [12] giving corner barriers for the linear elliptic and linear oblique derivative boundary operators.

Let us define the linear operators $\overline{L}$ and $\overline{M}$ by

$$\overline{L}v := \sum_{i,j} a_{ij}(\tilde{u} + \tilde{f}(\Xi_a))D^{ij}v + \sum_i b_i(\tilde{u} + \tilde{f}(\Xi_a))D^i v \tag{4.84}$$

and

$$\overline{M}v := \beta(\tilde{u} + \tilde{f}(\Xi_a)) \cdot \nabla v. \tag{4.85}$$

First, we consider the linear problem

$$\begin{array}{lll} \overline{L}v = 0 & \text{in} & \Omega, \\ \overline{M}v = 0 & \text{on} & \tilde{\Sigma}, \\ v = \tau(\tilde{f} - \tilde{f}(\Xi_a)) & \text{on} & \partial\Omega \setminus \tilde{\Sigma}. \end{array} \tag{4.86}$$

Note that

$$\overline{L}\tilde{u} = \tilde{Q}(\tilde{u} + \tilde{f}(\Xi_a)) - \sum_{i,j} c_{ij}(\tilde{u} + \tilde{f}(\Xi_a))D^i\tilde{u}D^j\tilde{u} \leq 0,$$

because $\tilde{u}$ is a solution to the nonlinear problem (4.82) and the left inequality in (4.60) holds. Therefore, $\tilde{u}$ is a supersolution for (4.86), meaning, $v \leq \tilde{u}$. Further, consider the linear problem

$$\begin{array}{lll} \overline{L}w = 0 & \text{in} & \Omega, \\ \overline{M}w = 0 & \text{on} & \tilde{\Sigma}, \\ w = \frac{1}{k}\left(e^{k\tau(\tilde{f} - \tilde{f}(\Xi_a))} - 1\right) & \text{on} & \partial\Omega \setminus \tilde{\Sigma}, \end{array} \tag{4.87}$$

where $k \geq 0$ is such that (4.60) holds. Note that for $w_{sub} := \frac{1}{k}\left(e^{k\tilde{u}} - 1\right)$ we have

$$\begin{aligned} \overline{L}w_{sub} = e^{k\tilde{u}}(\tilde{Q}(\tilde{u} + \tilde{f}(\Xi_a)) + \sum_{i,j}(ka_{ij}(\tilde{u} + \tilde{f}(\Xi_a)) \\ - c_{ij}(\tilde{u} + \tilde{f}(\Xi_a)))D^i\tilde{u}D^j\tilde{u}) \\ \geq 0, \end{aligned}$$

because $\tilde{Q}(\tilde{u}+\tilde{f}(\Xi_a)) = 0$ and the right-hand inequality in (4.60) holds. This implies that w_{sub} is a subsolution for the problem (4.87), meaning, $w_{sub} \leq w$. On the other hand, from the definition of w_{sub}, clearly $w_{sub} > \tilde{u}$. Therefore, if $\tilde{u}$ solves the problem (4.82), then the inequalities

$$v \leq \tilde{u} \leq w \tag{4.88}$$

hold, where v and w denote arbitrary solutions of the linear problems (4.86) and (4.87), respectively.

Next we use Lemma 4.1 of [12], which gives a corner barrier function for the linear operators $\overline{L}$ and $\overline{M}$, defined in (4.84) and (4.85), respectively. By this lemma, there exist positive constants h_0 and γ_0 (depending on the size of the opening angles of the domain Ω at the set of corners $\mathbf{V}$, on the ellipticity ratio of the linear operator $\overline{L}$ and on the bounds on Σ and $\tilde{u}$) such that for every fixed parameter $\gamma \in (0, \gamma_0)$ there exist a constant $c_1 \in (0, 1]$ and a function $g \in C^2(\cup_i \overline{\Omega_i(h_0)} \setminus \mathbf{V}) \cap C(\cup_i \overline{\Omega_i(h_0)})$ (depending on the same parameters as γ_0 and also on γ) with the property that for each $i \in \{1, 2, 3\}$ we have

$$\begin{aligned}
&\overline{L}g \leq 0 && \text{in } \Omega_i(h_0),\\
&c_1 d_{V_i}^{\gamma} \leq g \leq d_{V_i}^{\gamma} && \text{in } \Omega_i(h_0),\\
&\overline{M}g \leq 0 && \text{on } \tilde{\Sigma}_i(h_0).
\end{aligned} \tag{4.89}$$

Here, $\Omega_i(h_0)$ and $\tilde{\Sigma}_i(h_0)$ denote the subsets of Ω and $\tilde{\Sigma}$, respectively, on which $d_{V_i} < h_0$. We remark that the parameter γ_0 provided by Lemma 4.1 of [12] is the same γ_0 as in Theorem 4.3, step 2, and, as before, we take $\gamma \in (0, \min\{\gamma_0, 1\})$. Further, (4.89) also implies

$$\begin{aligned}
&\overline{L}(-g) \geq 0 && \text{in } \Omega_i(h_0),\\
&\overline{M}(-g) \geq 0 && \text{on } \tilde{\Sigma}_i(h_0).
\end{aligned}$$

We multiply g by a positive constant C^* so that

$$C^* g + \tilde{u}(V_i) \geq \frac{1}{k}(e^{\tau(\tilde{f} - \tilde{f}(\Xi_a))} - 1) \quad \text{on} \quad \Omega_i(h_0) \setminus \tilde{\Sigma}_i(h_0),$$

for each $i \in \{1, 2, 3\}$. Note that the constant C^* does not depend on $\tilde{u}$. This gives

$$C^* g + \tilde{u}(V_i) \geq w \quad \text{in} \quad \Omega_i(h_0),$$

where w is a solution to the linear problem (4.87). With the right-hand inequality in (4.88) and (4.89), we get

$$\tilde{u} - \tilde{u}(V_i) \leq w - \tilde{u}(V_i) \leq C^* g \leq C^* d_{V_i}^{\gamma} \quad \text{in} \quad \Omega_i(h_0),$$

and, hence, for each i we have

$$d_{V_i}^{-\gamma}(\tilde{u} - \tilde{u}(V_i)) \leq C^* \quad \text{in} \quad \Omega_i(h_0). \tag{4.90}$$

Similarly, we multiply $-g$ by a positive constant C_* so that

$$-C_* g + \tilde{u}(V_i) \leq \tau(\tilde{f} - \tilde{f}(\Xi_a)) \quad \text{on} \quad \Omega_i(h_0) \setminus \tilde{\Sigma}_i(h_0),$$

for each $i \in \{1, 2, 3\}$. Again, the constant C_* is independent of $\tilde{u}$. This yields

$$-C_* g + \tilde{u}(V_i) \leq v \quad \text{in} \quad \Omega_i(h_0),$$

for a solution v of (4.86). Recalling the left-hand inequality in (4.88) and (4.89) we obtain

$$-C_* d_{V_i}^{\gamma} \leq -C_* g \leq v - \tilde{u}(V_i) \leq \tilde{u} - \tilde{u}(V_i) \quad \text{in} \quad \Omega_i(h_0),$$

and, therefore, for each i we have

$$d_{V_i}^{-\gamma}(\tilde{u} - \tilde{u}(V_i)) \geq -C_* \quad \text{in} \quad \Omega_i(h_0). \tag{4.91}$$

Note that on $\Omega \setminus (\cup_i \Omega_i(h_0))$, we have that for each i

$$d_{V_i}^{-\gamma} |\tilde{u} - \tilde{u}(V_i)| \leq \, h_0^{-\gamma} \, |\tilde{u} - \tilde{u}(V_i)|_0 \leq 2h_0^{-\gamma} \, |\tilde{f} - \tilde{f}(\Xi_a)|_0,$$

using the L^∞ bound (4.83) for $\tilde{u}$. Together with inequalities (4.90) and (4.91), this gives

$$\sup_{i,(\rho,\eta)\in\Omega} d_{V_i}^{-\gamma}(\rho, \eta) |\tilde{u}(\rho, \eta) - \tilde{u}(V_i)| \leq C, \tag{4.92}$$

for a constant C independent of $\tilde{u}$, as desired.

Step 3. In this step we show that there exist positive constants δ^* and C, depending on the geometry of Ω, the bounds on the minimal eigenvalue and the ellipticity ratio of the operator $\tilde{Q}$, on the obliqueness constant for the operator $\tilde{N}$, and the bounds on $\tilde{u}$ and $\tilde{f}$ such that

$$|\tilde{u}|_{\delta^*} \leq C. \tag{4.93}$$

(By assumptions (4.53) through (4.57) and the inequality (4.83), all of these bounds are uniform in $\tilde{u}$ and $\rho \in \mathcal{K}$.)

First we quote several results from Gilbarg and Trudinger [8] and Lieberman and Trudinger [14] that give local Holder estimates on the parts of the boundary $\partial\Omega \backslash \mathbf{V}$ where we impose different types of boundary conditions. More precisely, we derive local Holder estimates on $\tilde{\Sigma}$, where we assume the oblique derivative condition, and on $\partial\Omega \backslash (\mathbf{V} \cup \tilde{\Sigma})$, where we have a Dirichlet condition. For the Holder estimate at the set of corners $\mathbf{V}$, we use the inequality (4.92) established in step 2.

For the estimate on $\tilde{\Sigma}$ we use Theorem 2.3 in [14]. This theorem is proved when the considered part of the boundary has smoothness C^2. However, the authors of [14] remark that it suffices that the boundary is $H_{1+\alpha}$, for $\alpha \in (0, 1)$, which is the case for $\tilde{\Sigma} = \Sigma \cup \Sigma_0$. The assumptions of Theorem 2.3 in [14] are that the operator $\tilde{Q}$ satisfies the structure condition (4.59), that the operators $\tilde{Q}$ and $\tilde{N}$ are uniformly elliptic and uniformly oblique, respectively, and that the supremum norm of $\tilde{u}$ is uniformly bounded. This theorem implies that there exist α_0 and C such that

$$[\tilde{u}]_{\alpha_0} \leq C, \tag{4.94}$$

in a neighborhood of $\tilde{\Sigma}$. Here, α_0 depends on the bounds for the ellipticity ratio of $\tilde{Q}$ and the obliqueness constant for $\tilde{N}$, and $\mu_0|\tilde{u}|_0$, where μ_0 is the constant from the structure condition (4.59). The constant C depends also on Ω.

The estimate on the Dirichlet part of the boundary $\partial\Omega\backslash(\tilde{\Sigma} \cup \mathbf{V}) = \sigma\backslash\mathbf{V}$ follows from the assumption that $\tilde{u} = \tau(\tilde{f} - \tilde{f}(\Xi_a))$ on σ and that $\tilde{f} \in H_\gamma$ on an open set $\mathcal{R}$ containing σ. Hence, $[\tilde{u}]_\gamma \leq [\tilde{f}]_{\gamma;\mathcal{R}}$ on $\sigma\backslash\mathbf{V}$.

For the local estimate at the set of corners $\mathbf{V}$ we use the inequality (4.92), which implies $[\tilde{u}]_\gamma \leq C$.

Next, we take $\overline{\alpha} := \min\{\alpha_0, \gamma\}$ and note that we have shown that

$$\operatorname{osc}_{\partial\Omega\cap B_R(x_0)}(\tilde{u}) \leq KR^{\overline{\alpha}}, \qquad \text{for every } x_0 \in \partial\Omega \text{ and } R > 0, \tag{4.95}$$

where osc stands for the oscillation and $K = [\tilde{u}]_{\overline{\alpha}}$. Again, the parameter $\overline{\alpha}$ in (4.95) does not depend on $\tilde{u}$. More precisely, $\overline{\alpha}$ depends on the size of Ω (which can be estimated in terms of a priori bounds on $\rho \in \mathcal{K}$), the bounds on the ellipticity ratio for the operator $\tilde{Q}$ and the obliqueness constant of the operator $\tilde{N}$, and on the bounds on $|\tilde{u}|_0$ (and these bounds can be estimated uniformly in $\tilde{u}$ and $\rho \in \mathcal{K}$).

Finally, with inequality (4.95) being satisfied we have that the assumptions of Theorem 8.29 in [8] hold. This theorem implies that there exist positive δ^* and C such that the desired estimate (4.93) holds. Here, the parameter δ^* depends on the ellipticity ratio for $\tilde{Q}$, the minimal eigenvalue of $\tilde{Q}$ and the parameter $\overline{\alpha}$ in the inequality (4.95), while the constant C also depends on $|\tilde{u}|_0$. Notice that (4.93) also implies $\tilde{u} \in H_{\delta^*}$.

Step 4. Having $\tilde{u} \in H_{\delta^*}$ and the uniform estimate (4.93), we use Theorem 4.3 with z replaced by $\tilde{u}$ and $\alpha_{\mathcal{K}}$ replaced by $\min\{\delta^*, \alpha_{\mathcal{K}}, \gamma\}$. (Recall again that because we are not concerned with the existence of a solution to problem (4.82), we can treat (4.82) as a linear problem with $z := \tilde{u} + \tilde{f}(\Xi_a)$.) The estimate (4.71) of Theorem 4.3 gives

$$|\tilde{u}|^{(-\gamma)}_{1+\min\{\delta^*,\alpha_{\mathcal{K}},\gamma\}} \leq C. \tag{4.96}$$

Note that the seminorms $[\tilde{u}]_{\min\{\delta^*,\alpha_{\mathcal{K}},\gamma\}}$ and $[\chi_i(\tilde{u}, \rho, \rho')]_{\min\{\delta^*,\alpha_{\mathcal{K}},\gamma\}}$ are bounded independently of $\tilde{u}$ by (4.93), as well as the term

$$\sup_{i,(\rho,\eta)\in\Omega} d_{V_i}^{-\gamma}(\rho, \eta)|\tilde{u}(\rho, \eta) - \tilde{u}(V_i)|$$

by (4.92). Therefore, we have that the constant C in the estimate (4.96) does not depend on $\tilde{u}$. We use the inequality

$$|\tilde{u}|_{\gamma} = |\tilde{u}|_{\gamma}^{(-\gamma)} \le C(\delta^*, \alpha_{\mathcal{K}}, \gamma, \mathrm{diam}(\Omega))|\tilde{u}|_{1+\min\{\delta^*,\alpha_{\mathcal{K}},\gamma\}}^{(-\gamma)},$$

and the estimate (4.96) to get $\tilde{u} \in H_{\gamma}$.

To eliminate δ^* in (4.96), we repeat step 4 with δ^* replaced by γ and obtain the estimate analogous to (4.96); that is, we get

$$|\tilde{u}|_{1+\min\{\alpha_{\mathcal{K}},\gamma\}}^{(-\gamma)} \le C.$$

Therefore, (4.81) follows.

Finally, for $\gamma_0 > 0$ from Theorem 4.3, we take

$$\gamma \in (0, \min\{\gamma_0, 1\}) \quad \text{and} \quad \alpha_{\mathcal{K}} \in (0, \min\{1, 2\gamma\}),$$

and recall the choices (from Lemma 4.5):

$$\epsilon = \frac{\alpha_{\mathcal{K}}}{2} \quad \text{and} \quad \gamma_1 = \frac{\gamma}{2}.$$

With the notation $u := \tilde{u} + \tilde{f}(\Xi_a)$ and using Remark 4.1 and the estimate 4.81, we obtain

$$|u|_{1+\epsilon}^{(-\gamma_1)} = |u|_{1+\alpha_{\mathcal{K}}/2}^{(-\gamma/2)} \le C|u|_{1+\min\{\alpha_{\mathcal{K}},\gamma\}}^{(-\gamma)} \le C\,M, \tag{4.97}$$

for a constant C depending on $\alpha_{\mathcal{K}}$, γ and the diameter of Ω and the constant M is as in (4.81). Hence, the hypotheses of the quoted fixed-point theorem at the beginning of Section 4.4.2 (Theorem 11.3 in [8]) are satisfied. Therefore, the map T defined in (4.77) has a fixed point $u \in H_{1+\alpha_{\mathcal{K}}}^{(-\gamma)}$. This fixed point u solves the fixed boundary value problem (4.50) and because $H_{1+\alpha_{\mathcal{K}}}^{(-\gamma)} \subseteq H_{1+\alpha_*}^{(-\gamma)}$, for any $\alpha_* \in (0, \alpha_{\mathcal{K}}]$, the proof of Theorem 4.2 is completed.

Remark

Note that by (4.97) we also have a uniform estimate of the γ-Holder norm of u, a solution to the fixed boundary value problem (4.50), on $\Omega \cup \tilde{\Sigma}$. Namely, because

$$|u|_{\gamma} = |u|_{\gamma}^{(-\gamma)} \le C(\gamma, \mathrm{diam}(\Omega))|u|_{1+\min\{\alpha_k,\gamma\}}^{(-\gamma)},$$

we have

$$u \in H_{\gamma;\Omega\cup\Sigma} \quad \text{and} \quad |u|_{\gamma} \le C, \tag{4.98}$$

for a constant C depending on γ, the size of the domain Ω, bounds on the ellipticity ratio and on the minimal eigenvalue of the operator $\tilde{Q}$, the bounds on the obliqueness constant of the operator $\tilde{N}$ and on the supremum norm $|u|_0$, and the Holder seminorm $[\tilde{f}]_{\gamma;\mathcal{R}}$.

5 Solution to the modified free boundary value problem

In this section we prove Theorem 3.1. The main idea is to fix the function $\rho(\eta)$, $\eta \in [0, \eta^*]$, specifying the boundary $\Sigma = \{\rho(\eta), \eta) : \eta \in (0, \eta^*)\}$, find a solution $u(\rho, \eta)$ of the fixed boundary value problems (3.39) and (3.42) through (3.45) using Theorem 4.2, and then update the boundary Σ to $\{\tilde{\rho}(\eta), \eta) : \eta \in (0, \eta^*)\}$, using the shock evolution equation (3.33). More precisely, we find $\tilde{\rho}$ as a solution of the following initial value problem

$$\frac{d\tilde{\rho}}{d\eta} = -\sqrt{\psi\left(\tilde{\rho}(\eta) - \frac{u(\rho(\eta), \eta) + 1}{2}\right)}, \tag{5.99}$$

$$\tilde{\rho}(0) = \xi_a. \tag{5.100}$$

We define a map J so that $J(\rho) = \tilde{\rho}$. To prove Theorem 3.1, we show that the map J has a fixed point. We use the following:

Theorem 5.1

(Corollary 11.2 in [8]) Let $\mathcal{K}$ be a closed and convex subset of a Banach space $\mathcal{B}$ and let $J : \mathcal{K} \to \mathcal{K}$ be a continuous mapping so that $J(\mathcal{K})$ is precompact. Then J has a fixed point.

We choose the space $\mathcal{B} = H_{1+\alpha_{\mathcal{K}}}$, and we take the set $\mathcal{K} \subset \mathcal{B}$ as in Section 4.2. In this section we further specify the parameters γ and $\alpha_{\mathcal{K}}$, and δ_*, ρ_L and ρ_L in the definition (4.48) through (4.49) of the set $\mathcal{K}$ so that the hypothesis of this fixed-point theorem is satisfied.

Let $a > \sqrt{2}$, $\eta^* > 0$, $\epsilon_* \in (0, u_* - 1)$ and $\delta > 0$ be arbitrary. We make the following choices for δ_*, ρ_L and ρ_R:

- δ_* depends on the particular value u_* we consider so that

$$\begin{aligned} &\text{- if } U_* = U_R, \text{ then } 0 < \delta_* < \min\left\{\tfrac{a^2-1+a\sqrt{a^2-2}}{2}, \left(\tfrac{\epsilon_*}{2\eta^*}\right)^2\right\}, \\ &\text{- if } U_* = U_F, \text{ then } 0 < \delta_* < \min\left\{\tfrac{a^2-1-a\sqrt{a^2-2}}{2}, \left(\tfrac{\epsilon_*}{2\eta^*}\right)^2\right\}, \end{aligned} \tag{5.101}$$

- the definition of ρ_L depends on η^*, and

$$\text{- if } \eta^* \in \left(0, \sqrt{\xi_a - 1}\,\right], \text{ then } \rho_L(\eta) := \xi_a - \eta\sqrt{\xi_a - 1}, \eta \in [0, \eta^*],$$

$$\text{- if } \eta^* > \sqrt{\xi_a - 1}, \text{ then } \rho_L(\eta) := \begin{cases} \xi_a - \eta\sqrt{\xi_a - 1}, & \eta \in [0, \sqrt{\xi_a - 1}], \\ 1, & \eta \in (\sqrt{\xi_a - 1}, \eta^*], \end{cases} \tag{5.102}$$

- ρ_R is defined by

$$\rho_R(\eta) := \xi_a - \eta\sqrt{\delta_*}, \quad \eta \in [0, \eta^*]. \tag{5.103}$$

Clearly, the set $\mathcal{K}$ defined in Section 4.2 with the above specifications of the parameter δ_* and the curves $\rho_L(\eta)$ and $\rho_R(\eta)$, $\eta \in [0, \eta^*]$, is a well-defined, closed and convex subset of the Banach space $\mathcal{B}$.

Lemma 5.2

Let $a > \sqrt{2}$, $\eta^* > 0$ and $\epsilon_* \in (0, u_* - 1)$ be given. If δ_* is chosen as in (5.101), then the curve $\tilde{\rho}$ given by (5.99) and (5.100) is decreasing and satisfies the following lower bound

$$\tilde{\rho}(\eta) > 1, \quad \textit{for all } \eta \in [0, \eta^*]. \tag{5.104}$$

Proof

By the definition (3.34) of the function ψ and Equation 5.99 for $\tilde{\rho}'$, we have

$$\tilde{\rho}'(\eta) \leq -\sqrt{\delta_*} < 0, \qquad \text{for all } \eta \in (0, \eta^*), \tag{5.105}$$

implying the desired monotonicity of the curve $\tilde{\rho}$.

Next we show (5.104). If $\eta_0 \in [0, \eta^*]$ is such that

$$\tilde{\rho}(\eta_0) - \frac{u(\rho(\eta_0), \eta_0) + 1}{2} \geq \delta_*, \tag{5.106}$$

then certainly $\tilde{\rho}(\eta_0) - \frac{u(\rho(\eta_0),\eta_0)+1}{2} > 0$, implying

$$\tilde{\rho}(\eta_0) - 1 > \frac{u(\rho(\eta_0), \eta_0) - 1}{2} \geq \frac{\epsilon_*}{2}, \tag{5.107}$$

by Lemma 4.2. It is easy to check that (5.106) holds for $\eta_0 = 0$, using expression (2.9) for $u_* = u(\xi_a, 0)$ and the bounds of (5.101) on δ_*. Further, note that the change of $\tilde{\rho}$ over the values of η_0 for which (5.106) does not hold is $\sqrt{\delta_*}$ per unit interval in η. This implies that the total change of $\tilde{\rho}$ over the interval $[0, \eta^*]$ is bounded above by $\sqrt{\delta_*}\eta^*$. However, our choice (5.101) of δ_* implies

$$\sqrt{\delta_*}\eta^* < \frac{\epsilon_*}{2\eta^*}\eta^* = \frac{\epsilon_*}{2}. \tag{5.108}$$

From (5.107) and (5.108) we get that $\tilde{\rho}(\eta) - 1 > 0$, for all $\eta \in [0, \eta^*]$.

Remark

Note that for every $\eta \in [0, \eta^*]$ we have the following lower bound

$$\rho(\eta) - 1 > \frac{\epsilon_*}{2} - \sqrt{\delta_*}\eta^* > 0,$$

which can be estimated in terms of only a, η^* and ϵ_* using (5.101).

Lemma 5.3

Let $a > \sqrt{2}$, $\eta^* > 0$, $\epsilon_* \in (0, u_* - 1)$ and $\delta > 0$ be given and suppose that δ_*, ρ_L and ρ_R are chosen as in (5.101) through (5.103).

There exists a parameter $\gamma_0 > 0$, depending on a, η^*, ϵ_* and δ, such that for any $\gamma \in (0, \min\{\gamma_0, 1\})$ and $\alpha_{\mathcal{K}} = \frac{\gamma}{2}$ we have

(a) $J(\mathcal{K}) \subseteq \mathcal{K}$, and

(b) the set $J(\mathcal{K})$ is precompact in $H_{1+\alpha_{\mathcal{K}}}$.

Proof

Let the boundary Σ be given by a curve $\rho \in \mathcal{K}$, and let $u(\rho, \eta) \in H^{(-\gamma)}_{1+\alpha_{\mathcal{K}}}$ be a solution of the fixed boundary value problem found by Theorem 4.2. Recall that $\gamma \in (0, \min\{\gamma_0, 1\})$ is arbitrarily chosen, where γ_0 is a parameter depending on the size of the opening angles of the domain Ω at the corners, and on the boundary on the ellipticity ratio for $\tilde{Q}$. By Remark 4.4.1, γ_0 depends only on the fixed parameters a, η^*, ϵ_* and δ. Recall also that $\alpha_{\mathcal{K}} \in (0, \min\{1, 2\gamma\})$ is arbitrary. To prove this lemma we will take γ_0 smaller, still depending only on the a priori fixed parameters a, η^*, ϵ_* and δ, and we will specify $\alpha_{\mathcal{K}} = \gamma/2$.

Let $\tilde{\rho}(\eta)$, $\eta \in [0, \eta^*]$, be a solution of the initial-value problem (5.99), (5.100). To show part (a) we need to show that $\tilde{\rho} \in \mathcal{K}$.

Clearly, $\tilde{\rho}(0) = \xi_a$ and

$$\tilde{\rho}'(0) = -\sqrt{\psi\left(a^2 + \frac{1}{2} - \frac{u_* + 1}{2}\right)} = k_*,$$

by taking u_* and k_* as in (2.9) and (2.11), with respect to the particular value of the parameter a as in (2.10). This shows (4.47).

That the curve $\tilde{\rho}$ satisfies the right inequality of (4.48) has been shown in (5.105). To show the left side of (4.48), note that

$$\begin{aligned}\max_{\eta \in [0,\eta^*]} \left\{\tilde{\rho}(\eta) - \frac{u(\rho(\eta), \eta) + 1}{2}\right\} &\le \max_{\eta \in [0,\eta^*]} \tilde{\rho}(\eta) - \min_{\eta \in [0,\eta^*]} \frac{u(\rho(\eta), \eta) + 1}{2} \\ &\le \xi_a - \frac{2 + \epsilon_*}{2} < \xi_a - 1,\end{aligned}$$

by Lemma 4.2 and the choice of f (see (2.16)). Therefore, $\tilde{\rho}'(\eta) \ge -\sqrt{\xi_a - 1}$, for all $\eta \in (0, \eta^*)$.

Next we check that $\tilde{\rho}$ satisfies (4.49). Note that the left side of (4.48) implies $\tilde{\rho}(\eta) \ge \xi_a - \eta\sqrt{\xi_a - 1}$, $\eta \in [0, \eta^*]$. Because (5.104) also holds, we get $\tilde{\rho}(\eta) \ge \rho_L(\eta)$, $\eta \in [0, \eta^*]$, where ρ_L is given by (5.102). On the other hand, the inequality (5.105) implies $\tilde{\rho}(\eta) \le \xi_a - \eta\sqrt{\delta_*} = \rho_R(\eta)$, for all $\eta \in [0, \eta^*]$.

To complete the proof of (a), it is left to show

$$\tilde{\rho} \in H_{1+\alpha_{\mathcal{K}}}. \tag{5.109}$$

We recall the estimate (4.94), which is independent of u, and we replace the parameter γ_0 by $\min\{\gamma_0, \alpha_0\}$. Again we note that both γ_0 and α_0 depend only on the a priori fixed parameters a, η^*, ϵ_* and δ. Let

$$0 < \gamma < \min\{\gamma_0, 1\}, \tag{5.110}$$

Using the differential equation (5.99), it follows that $|\tilde{\rho}'|_\gamma \leq C$ and, hence,

$$|\tilde{\rho}|_{1+\gamma} \leq C\eta^*. \tag{5.111}$$

Therefore, $\tilde{\rho} \in H_{1+\gamma}$ and to ensure (5.109), we need to take $\alpha_{\mathcal{K}} \in (0, \gamma]$.

Moreover, because (5.111) holds uniformly in $\tilde{\rho}$, we have that $J(\mathcal{K})$ is contained in a bounded set in $H_{1+\gamma}$. To show (b) it suffices to choose $\alpha_{\mathcal{K}} \in (0, \gamma)$. We take $\alpha_{\mathcal{K}} := \gamma/2$.

We note that the map $J : \mathcal{K} \to \mathcal{K}$ given by (5.99) through (5.100) is also continuous. Therefore, by taking the parameter γ as in (5.110) and choosing $\alpha_{\mathcal{K}} = \gamma/2$, we have that the hypotheses of the fixed-point theorem from the beginning of this section (Corollary 11.2 in [8]) are satisfied. Hence, the map J has a fixed point $\rho \in \mathcal{K}$. We use this curve $\rho(\eta)$, $\eta \in [0, \eta^*]$, to specify the boundary Σ in Theorem (4.2) and we get a solution $u \in H^{(-\gamma)}_{1+\alpha_*}$, for all $\alpha_* \in (0, \alpha_{\mathcal{K}}]$. This completes the proof of Theorem 3.1.

6 The proof of Theorem 2.1

In order to derive Theorem 2.1 from Theorem 3.1, we need to see whether we can remove the cut-off functions we have introduced in Theorem 3.1. More precisely, we need to investigate under which conditions it is possible to replace the functions ϕ and ψ by the identity function and to replace χ by β.

Let $u \in H^{(-\gamma)}_{1+\alpha_*}$, for $\alpha_* \in (0, \alpha_{\mathcal{K}}]$, be a solution to the modified free boundary value problem (4.50) in Theorem 3.1, and suppose that $v \in H^{(-\gamma)}_{1+\alpha_*}$ is recovered using equation (3.27).

First, we recall from Proposition (4.1) that a priori bounds on u imply $\chi = \beta$, meaning that the operators N and $\tilde{N}$ are the same.

To remove the cut-off ϕ we prove the following:

Lemma 6.1

Suppose that $u \in C^1(\Omega)$ is a solution to the fixed boundary value problem (4.50) and define a function

$$w(\rho, \eta) := u(\rho, \eta) - \rho, \quad (\rho, \eta) \in \Omega. \tag{6.112}$$

Then

- *(a) w attains its minimum on $\sigma \cup \Sigma \cup \Xi_a$, and*
- *(b) w cannot attain a nonpositive minimum on Σ.*

Proof

First we show (a). Using the equation $\tilde{Q}(u) = 0$ in (4.50) and the definition (6.112), we obtain the following second-order uniformly elliptic equation for w

$$\phi(w)w_{\rho\rho} + w_{\eta\eta} + \left(\phi'(w) + \frac{1}{2}\right) w_\rho + \frac{1}{2} + \phi'(w)(w_\rho)^2 = 0.$$

Using the Minimum Principle (Theorem 3.5 in [8]), we have that w must attain its minimum on $\partial\Omega$. Suppose there is a minimum $X_0 \in \Sigma_0$. By (4.50) we have $w_\eta(X_0) = 0$, which contradicts Hopf's Lemma (Lemma 3.4 in [8]). Hence, the minimum must occur on $\partial\Omega \backslash \Sigma_0 = \sigma \cup \Sigma \cup \Xi_a$.

Next we show (b). Assume there is $X_0 \in \Sigma$ so that $w(X_0) = \min_\Omega w$. Because X_0 is the minimum of w, the tangential derivative of w along Σ at X_0 must be zero, that is,

$$0 = (w_\rho(X_0), w_\eta(X_0)) \cdot (\rho'(X_0), 1) = \rho'(X_0)(u_\rho(X_0) - 1) + u_\eta(X_0),$$

yielding

$$u_\eta(X_0) = -\rho'(X_0)(u_\rho(X_0) - 1). \tag{6.113}$$

On the other hand, Hopf's Lemma (Lemma 3.4 in [8]) implies that the derivative of w in the direction of an outward normal to Σ at X_0 must be negative, meaning

$$0 > (w_\rho(X_0), w_\eta(X_0)) \cdot (1, -\rho'(X_0)) = u_\rho(X_0) - 1 - \rho'(X_0)u_\eta(X_0). \tag{6.114}$$

We substitute (6.113) in (6.114) to find

$$u_\rho(X_0) < 1. \tag{6.115}$$

Next we use the oblique derivative boundary condition in (4.50) with β given by (3.29) and substitute (6.113) to obtain

$$u_\rho(X_0)\rho'(X_0)\frac{u(X_0) - 1}{4} + \rho'(X_0)\left(\frac{5u(X_0) + 3}{8} - \rho(X_0)\right) = 0.$$

Because $u > 1$ on Ω (see Proposition (4.1)), $\rho' < 0$ on Σ (see the definition of set $\mathcal{K}$ at the beginning of Section 4.2) and (6.115) holds, we have

$$\rho'(X_0)\frac{u(X_0) - 1}{4} + \rho'(X_0)\left(\frac{5u(X_0) + 3}{8} - \rho(X_0)\right) < 0,$$

and using that $u(X_0) = \rho(X_0) + w(X_0)$, this yields

$$\rho'(X_0)\frac{-\rho(X_0) + 7w(X_0) + 1}{8} < 0.$$

Because $\rho'(X_0) < 0$, we obtain

$$w(X_0) > \frac{\rho(X_0) - 1}{7}. \tag{6.116}$$

We recall that $\rho > 1$ on Σ (see (5.104)), and, hence, (6.116) implies $w(X_0) > 0$, as desired.

Note that the cut-off function ϕ differs from the identity function only if $u(\rho,\eta) - \rho < \delta$. In the previous lemma we showed that the function $u(\rho,\eta) - \rho$, $(\rho,\eta) \in \Omega$ attains its minimum on $\sigma \cup \Sigma \cup \Xi_a$. It is easy to calculate that at Ξ_a we have

$$u(\Xi_a) - \xi_a = u_* - \left(a^2 + \frac{1}{2}\right) =: m_1(a) > 0,$$

for both choices of u_* in (2.10). Further, on σ we have

$$u(\rho,\eta) - \rho = f(\eta) - \left(\xi(\eta^*) + \frac{\eta^2}{4}\right) =: m_2(a,\eta^*,\epsilon_*) > 0,$$

by definition (2.16) of f and the bounds on the closed set $\mathcal{K}$ (see Remark 2.2). Next, in the previous lemma we showed that if $u(\rho,\eta) - \rho$ attains its minimum at $X_0 \in \Sigma$, then this minimum must be positive. More precisely, using (6.116) and Remark 5 we have

$$u(X_0) - \rho(X_0) > \frac{\rho(X_0) - 1}{7} =: m_3(a,\eta^*,\epsilon_*) > 0.$$

We choose δ in the definition of ϕ so that

$$0 < \delta < \min\{m_1, m_2, m_3\},$$

with m_1, m_2 and m_3 as above and depending only on a, η^* and ϵ_*. Therefore, the function ϕ is the identity and the operators Q and $\tilde{Q}$ are the same.

Finally, we note that the cut-off function ψ is identity as long as

$$\rho \geq \frac{u_* + 1}{2} + \delta_*. \tag{6.117}$$

Using the values (2.9) for u_* and (5.101) for δ_* we obtain $\frac{u_*+1}{2} + \delta_* < \xi_a = \rho(\Xi_a)$. Because it is possible that for some choices of a and δ_* we have $\frac{u_*+1}{2} + \delta_* > 1$, the cut-off ψ can be removed only in a neighborhood of the reflection point Ξ_a.

This completes the proof of Theorem 2.1.

Acknowledgments

Research of the first two authors was partially supported by the Department of Energy, Grant DE-FG02-03ER25575. The third author was partially supported by the National Science Foundations under grants NSF DMS-0244343, NSF DMS-0225948 and NSF DMS-0245513.

We appreciate the hospitality of the Fields Institute during K.J.'s research visits, and we also thank the Focused Research Grant on Multi-dimensional Compressible Euler Equations for its encouragement and support.

References

[1] S. Čanić and B. L. Keyfitz, Riemann problems for the two-dimensional unsteady transonic small disturbance equation, *SIAM Journal on Applied Mathematics*, **58** (1998), 636–665.

[2] S. Čanić, B. L. Keyfitz, and G. M. Lieberman, A proof of existence of perturbed steady transonic shocks via a free boundary problem, *Communications on Pure and Applied Mathematics*, **LIII** (2000), 484–511.

[3] S. Čanić, B. L. Keyfitz, and E. H. Kim, Free boundary problems for the unsteady transonic small disturbance equation: Transonic regular reflection, *Methods and Applications of Analysis*, **7** (2000), 313–336.

[4] S. Čanić, B. L. Keyfitz, and E. H. Kim, A free boundary problem for a quasi-linear degenerate elliptic equation: Regular reflection of weak shocks, *Communications on Pure and Applied Mathematics*, **LV** (2002), 71–92.

[5] S. Čanić, B. L. Keyfitz, and E. H. Kim, Free boundary problems for nonlinear wave systems: Mach stems for interacting shocks, submitted.

[6] T. Chang and G. Q. Chen, Diffraction of planar shocks along compressive corner, *Acta Mathematica Scientia*, **3** (1986), 241–257.

[7] D. Gilbarg and L. Hormander, Intermediate Schauder estimates, *Arch. Rational Mech. Anal.*, **74** (1980), no. 4, 297–318.

[8] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York, 2nd edition, 1983.

[9] B. L. Keyfitz, Self-similar solutions of two-dimensional conservation laws, *Journal of Hyperbolic Differential Equations*, **1** (2004), 445–492.

[10] G. M. Lieberman, The Perron process applied to oblique derivative problems, *Advances in Mathematics*, **55** (1985), 161–172.

[11] G. M. Lieberman, Mixed boundary value problems for elliptic and parabolic differential equations of second order, *Journal of Mathematical Analysis and Applications*, **113** (1986), 422–440.

[12] G. M. Lieberman, Optimal Holder regularity for mixed boundary value problems, *Journal of Mathematical Analysis and Applications*, **143** (1989), 572–586.

[13] G. M. Lieberman, Local estimates for subsolutions and supersolutions of oblique derivative problems for general second order elliptic equations, *Transactions of the American Mathematical Society*, **304** (1987), no. 1, 343–353.

[14] G. M. Lieberman and N. S. Trudinger, Nonlinear oblique boundary value problems for nonlinear elliptic equations, *Transactions of the American Mathematical Society*, **295** (1986), no. 2, 509–546.

CHAPTER 7

Shape optimization for 3D electrical impedance tomography

Karsten Eppler
Weierstraß Institute for Applied Analysis and Stochastics,
Berlin, Germany

Helmut Harbrech
Institute for Computer Science and Practical Mathematics,
University of Kiel,
Kiel, Germany

Introduction

Let $D \subset \mathbb{R}^3$ denote a bounded domain with boundary $\partial D = \Sigma$ and assume the existence of a simply connected subdomain $S \subset D$, consisting of material with constant conductivity, essentially different from the likewise constant conductivity of the material in the subregion $\Omega = D \setminus \overline{S}$. We consider the identification problem of this inclusion if the Cauchy data of the electrical potential u are measured at the boundary Σ, that is, if a single pair $f = u|_\Sigma$ and $g = (\partial u/\partial \mathbf{n})|_\Sigma$ is known.

The problem under consideration is a special case of the general conductivity reconstruction problem and is severely ill posed. It has been intensively investigated as an inverse problem. We refer, for example, to Hettlich and Rundell [11] and Chapko and Kress [26] for numerical algorithms, and to Friedmann and Isakov [7] as well as Alessandrini, Isakov and Powell [6] for particular results concerning uniqueness. Moreover, we refer to Brühl and Hanke [8,9] for methods using the complete Dirichlet-to-Neumann operator at the outer boundary. We emphasize that we focus in the present paper on exact measurements and do not consider noisy data.

In [10], Roche and Sokolowski introduced a formulation as a shape optimization problem. However, we have proven in [14] that the shape Hessian degenerates at the optimal domain. Nevertheless, using second-order information in terms of a regularized Newton scheme yielded promising results in comparison to gradient-based methods. In particular, the method converges faster and provides higher accuracy. The present paper extends these results to three dimensions.

We employ boundary integral representations of the shape functional, its gradient, and its Hessian. After transforming the state equation to a boundary integral equation, we are able to perform all computations only on the boundary of the domain under consideration. To obtain a finite dimensional optimization problem, we assume the inclusion is starshaped and discretize its boundary by spherical harmonics. The boundary integral equations are solved efficiently by a fast wavelet Galerkin scheme, which computes the approximate solutions within linear complexity [15,16,19].

The present paper is organized as follows. In Section 1 we present the physical model and reformulate the identification problem as a shape optimization problem. We compute the gradient and the Hessian of the given shape functional and show how to use boundary integral equations to compute them. In Section 2 we discretize the boundary of the inclusion and replace the infinite dimensional optimization problem with a finite dimensional one. We also propose a wavelet-based fast boundary element method to compute the shape functional as well as its gradient and Hessian. In Section 3 we present a numerical experiment in which we compare the regularized Newton method with a quasi-Newton method.

1 Shape problem formulation

1.1 The physical model

Let $D \in \mathbb{R}^3$ be a simply connected domain with boundary $\Sigma = \partial D$ and assume that an unknown simply connected inclusion S with regular boundary $\Gamma = \partial S$ is located inside the domain D satisfying $\operatorname{dist}(\Sigma, \Gamma) > 0$ (see Figure 7.1). To determine the inclusion S we measure for a given current distribution $g \in H^{-1/2}(\Sigma)/\mathbb{R}$ the voltage distribution $f \in H^{1/2}(\Sigma)$ at the boundary Σ. Hence, we are seeking a domain $\Omega := D \setminus \overline{S}$ and an associated harmonic function u, satisfying the system of equations

$$\begin{aligned} \Delta u &= 0 && \text{in } \Omega, \\ u &= 0 && \text{on } \Gamma, \\ u &= f && \text{on } \Sigma, \\ \frac{\partial u}{\partial \mathbf{n}} &= g && \text{on } \Sigma. \end{aligned}$$

This system denotes an overdetermined boundary value problem that allows a solution only for the true inclusion S.

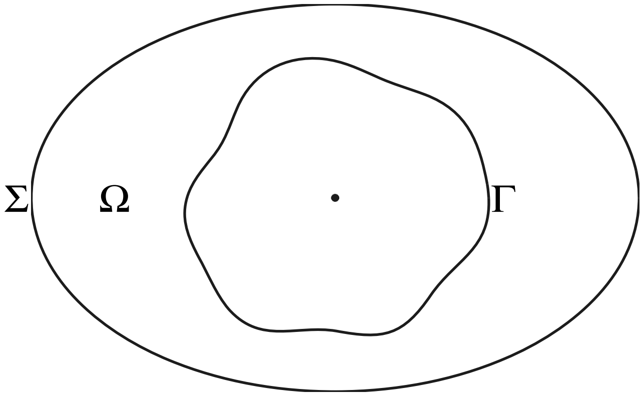


FIGURE 7.1 The domain Ω and its boundaries Γ and Σ.

Following Sokolowski and Roche [10], we introduce the auxiliary harmonical functions v and w satisfying

$$\begin{aligned} \Delta v &= 0 & \Delta w &= 0 & &\text{in } \Omega, \\ v &= 0 & w &= 0 & &\text{on } \Gamma, \\ \frac{\partial v}{\partial \mathbf{n}} &= g & w &= f & &\text{on } \Sigma, \end{aligned} \tag{1.1}$$

and consider the following shape-optimization problem

$$J(\Omega) = \int_{\Omega} \|\nabla(v - w)\|^2 d\mathbf{x} = \int_{\Sigma} \left(g - \frac{\partial w}{\partial \mathbf{n}}\right)(v - f) d\sigma \to \inf . \tag{1.2}$$

Herein, the infimum must be taken over all domains including a void with a sufficiently regular boundary. We refer to Roche and Sokolowski [10] for the existence of optimal solutions with respect to this shape-optimization problem.

1.2 Shape calculus

For sake of clarity in representation, we repeat the shape calculus concerning the problem under consideration by means of boundary variations. The shape calculus is in complete analogy to the two-dimensional one in [14]. For a survey on the shape calculus based on the material derivative concept, we refer the reader to Sokolowski and Zolesio [5] and Delfour and Zolesio [4] and the references therein.

Let the underlying variation fields $\mathbf{V}$ be sufficiently smooth such that $C^{2,\alpha}$-regularity is preserved for all perturbed domains. Moreover, for the sake of simplicity, we assume in addition that the outer boundary and its measurements are sufficiently regular such that the state functions $v = v(\Omega)$ and $w = w(\Omega)$ satisfy

$$v, w \in C^{2,\alpha}(\Omega). \tag{1.3}$$

Then, a formal differentiation of (1.2) in terms of local derivatives yields

$$dJ(\Omega)[\mathbf{V}] = \int_\Gamma \langle \mathbf{V}, \mathbf{n}\rangle \|\nabla(v-w)\|^2 d\sigma + 2\int_\Omega \langle \nabla(v-w), \nabla(dv - dw)\rangle d\mathbf{x},$$

where the local shape derivatives $dv = dv[\mathbf{V}]$ and $dw = dw[\mathbf{V}]$ satisfy

$$\begin{array}{lll} \Delta dv = 0 & \Delta dw = 0 & \text{in } \Omega, \\ dv \;\; = -\langle \mathbf{V}, \mathbf{n}\rangle \dfrac{\partial v}{\partial \mathbf{n}} & dw \;\; = -\langle \mathbf{V}, \mathbf{n}\rangle \dfrac{\partial w}{\partial \mathbf{n}} & \text{on } \Gamma, \\ \dfrac{\partial dv}{\partial \mathbf{n}} = 0 & dw \;\; = 0 & \text{on } \Sigma. \end{array} \tag{1.4}$$

The boundary integral representation of the shape gradient is now obtained via repeated integration by parts

$$\begin{aligned} dJ(\Omega)[\mathbf{V}] &= \int_\Gamma \langle \mathbf{V}, \mathbf{n}\rangle \{\|\nabla v\|^2 - \|\nabla w\|^2\} d\sigma \\ &= \int_\Gamma \langle \mathbf{V}, \mathbf{n}\rangle \left[\left(\frac{\partial v}{\partial \mathbf{n}}\right)^2 - \left(\frac{\partial w}{\partial \mathbf{n}}\right)^2\right] d\sigma, \end{aligned} \tag{1.5}$$

(cf. [10,14]). The identity $\nabla v|_\Gamma = \partial v/\partial \mathbf{n}$, and likewise for w, issues from the homogeneous Dirichlet boundary condition of the state equation (1.1). Moreover, note that as an immediate consequence of the shape calculus, (1.5) implies a simplified first-order necessary condition

$$\frac{\partial v}{\partial \mathbf{n}} = \frac{\partial w}{\partial \mathbf{n}} \qquad \text{on } \Gamma. \tag{1.6}$$

In the case of a hole S, which is starshaped with respect to a certain pole $\mathbf{p}$, the boundary $\Gamma = \partial S$ can be parametrized by a radial function r living on the sphere with a radius of one around the pole. Without loss of generality we assume throughout this paper this pole to be $\mathbf{0}$. Then, each point $\mathbf{x} \in \Gamma$ is represented uniquely by $\mathbf{x} = r(\widehat{\mathbf{x}}) \cdot \widehat{\mathbf{x}}$, where

$$\widehat{\mathbf{x}} := \frac{\mathbf{x}}{\|\mathbf{x}\|} \in \mathbb{S} := \{\mathbf{x} \in \mathbb{R}^3 : \|\mathbf{x}\| = 1\}.$$

As one can readily verify, the outer normal of Ω at the point $\mathbf{x} \in \Gamma$ is given by

$$\mathbf{n}(\mathbf{x}) = \frac{\nabla_{\mathbb{S}} r(\widehat{\mathbf{x}}) - r(\widehat{\mathbf{x}}) \cdot \widehat{\mathbf{x}}}{\sqrt{r^2(\widehat{\mathbf{x}}) + \|\nabla_{\mathbb{S}} r(\widehat{\mathbf{x}})\|^2}} \tag{1.7}$$

where the surface gradient $\nabla_{\mathbb{S}}$ with respect to the sphere is defined as

$$\nabla_{\mathbb{S}} r(\widehat{\mathbf{x}}) = \nabla r(\widehat{\mathbf{x}}) - \langle \widehat{\mathbf{x}}, \nabla r(\widehat{\mathbf{x}})\rangle \cdot \widehat{\mathbf{x}}.$$

Note that there holds in particular $\langle \nabla_{\mathbb{S}} r(\widehat{\mathbf{x}}), \widehat{\mathbf{x}}\rangle = 0$.

We choose the perturbation field $\mathbf{V}$ such that $\mathbf{V}(\mathbf{x}) = dr(\widehat{\mathbf{x}})\cdot\widehat{\mathbf{x}}$. Thus, the shape gradient (1.5) can be expressed equivalently in local coordinates as

$$dJ(\Omega)[dr] = \int_{\mathbb{S}} dr(\widehat{\mathbf{x}})\, r^2(\widehat{\mathbf{x}}) \left[\left(\frac{\partial w}{\partial \mathbf{n}}\right)^2 - \left(\frac{\partial v}{\partial \mathbf{n}}\right)^2\right] d\sigma, \tag{1.8}$$

where the minus sign issues from the fact that

$$\langle \widehat{\mathbf{x}}, \mathbf{n}\rangle = -r/\sqrt{r^2 + \|\nabla_{\mathbb{S}} r\|^2}$$

according to (1.7).

Lemma 1.1

The shape Hessian is given by

$$\begin{aligned} d^2 J(\Omega)[dr_1, dr_2] = \int_{\mathbb{S}} & 2r\, dr_1\, dr_2 \left[\left(\frac{\partial v}{\partial \mathbf{n}}\right)^2 - \left(\frac{\partial w}{\partial \mathbf{n}}\right)^2\right] \\ & + r^2\, dr_1\, dr_2 \frac{\partial}{\partial \widehat{\mathbf{x}}} \left\{\|\nabla w\|^2 - \|\nabla v\|^2\right\} \\ & + 2r^2\, dr_1 \left\{\frac{\partial w}{\partial \mathbf{n}} \frac{\partial dw[dr_2]}{\partial \mathbf{n}} - \frac{\partial v}{\partial \mathbf{n}} \frac{\partial dv[dr_2]}{\partial \mathbf{n}}\right\} d\sigma, \end{aligned} \tag{1.9}$$

where all data must be understood as traces on the boundary Γ.

Proof

The existence of a shape Hessian is provided by means of standard theory (cf. [4,5]). To derive the explicit structure, we proceed in a manner similar to [22,23] by differentiating the shape gradient (1.8). The domain Ω respective boundary Γ can be identified with its parametrization, that is, with the function $r : \mathbb{S} \to \Gamma$. Similarly, we can identify the perturbed domain Ω_ε respective boundary Γ_ε with the function $r_\varepsilon = r + \varepsilon dr_2$. Therefore, we find

$$\begin{aligned} & dJ(\Omega_\varepsilon)[dr_1] - dJ(\Omega_\varepsilon)[dr_1] \\ & = \int_{\mathbb{S}} dr_1 \left\{ r_\varepsilon^2 \left[\left(\frac{\partial w_\varepsilon}{\partial \mathbf{n}_\varepsilon}\right)^2 - \left(\frac{\partial v_\varepsilon}{\partial \mathbf{n}_\varepsilon}\right)^2\right] - r^2 \left[\left(\frac{\partial w}{\partial \mathbf{n}}\right)^2 - \left(\frac{\partial v}{\partial \mathbf{n}}\right)^2\right]\right\} d\sigma, \end{aligned}$$

where v_ε and w_ε are the solutions of the state equation with respect to the perturbed domain Ω_ε, and $\mathbf{n}_\varepsilon$ is the outer normal of Ω_ε at Γ_ε. Using Taylor's expansion

$$r_\varepsilon^2 = r^2 + 2\varepsilon\, r\, dr_2 + \mathcal{O}(\varepsilon^2)$$

yields

$$\begin{aligned} &dJ(\Omega_\varepsilon)[dr_1] - dJ(\Omega_\varepsilon)[dr_1] \\ &\quad = \int_{\mathbb{S}} dr_1 \left\{ 2r\varepsilon dr_2 \left[\left(\frac{\partial w_\varepsilon}{\partial \mathbf{n}_\varepsilon}\right)^2 - \left(\frac{\partial v_\varepsilon}{\partial \mathbf{n}_\varepsilon}\right)^2 \right] + \mathcal{O}(\varepsilon^2) \right\} d\sigma \\ &\quad + \int_{\mathbb{S}} dr_1 r^2 \left\{ \left[\left(\frac{\partial w_\varepsilon}{\partial \mathbf{n}_\varepsilon}\right)^2 - \left(\frac{\partial v_\varepsilon}{\partial \mathbf{n}_\varepsilon}\right)^2 \right] - r^2 \left[\left(\frac{\partial w}{\partial \mathbf{n}}\right)^2 - \left(\frac{\partial v}{\partial \mathbf{n}}\right)^2 \right] \right\} d\sigma. \end{aligned}$$

The first term in this expression will give the first term in (1.9). Hence, it remains to consider the difference

$$\left(\frac{\partial v_\varepsilon}{\partial \mathbf{n}_\varepsilon}\right)^2 - \left(\frac{\partial v}{\partial \mathbf{n}}\right)^2 = \langle \nabla v_\varepsilon|_{\Gamma_\varepsilon}, \nabla v_\varepsilon|_{\Gamma_\varepsilon} \rangle - \langle \nabla v|_\Gamma, \nabla v|_\Gamma \rangle,$$

because the corresponding term for w is treated in complete analogy. Observing $r_\varepsilon = r + \varepsilon dr_2$, we conclude with Taylor's expansion

$$\langle \nabla v_\varepsilon|_{\Gamma_\varepsilon}, \nabla v_\varepsilon|_{\Gamma_\varepsilon} \rangle = \langle \nabla v_\varepsilon|_\Gamma, \nabla v_\varepsilon|_\Gamma \rangle + 2\varepsilon dr_2 \frac{\partial}{\partial \widehat{\mathbf{x}}} \langle \nabla v_\varepsilon|_{\Gamma_\xi}, \nabla v_\varepsilon|_{\Gamma_\xi} \rangle,$$

where Γ_ξ is defined via the radial function $r_\xi = r + \xi dr_2$, $0 < \xi < \varepsilon$. Inserting the local shape derivative (1.4)

$$\nabla v_\varepsilon|_\Gamma = \nabla v|_\Gamma + \varepsilon dr_2 \nabla dv[dr_2]|_\Gamma + \mathcal{O}(\varepsilon^2),$$

we arrive at

$$\begin{aligned} \left(\frac{\partial v_\varepsilon}{\partial \mathbf{n}}\right)^2 - \left(\frac{\partial v}{\partial \mathbf{n}}\right)^2 &= 2\varepsilon dr_2 \frac{\partial}{\partial \widehat{\mathbf{x}}} \langle \nabla v_\varepsilon|_{\Gamma_\xi}, \nabla v_\varepsilon|_{\Gamma_\xi} \rangle \\ &\quad + 2\varepsilon dr_2 \langle \nabla dv[dr_2]|_\Gamma, \nabla v|_\Gamma \rangle + \mathcal{O}(\varepsilon^2). \end{aligned}$$

Computing $\lim_{\varepsilon\to 0}\{dJ(\Omega_\varepsilon)[dr_1] - dJ(\Omega_\varepsilon)[dr_1]\}/\varepsilon$ proves the assertion due to

$$\langle \nabla dv[dr_2]|_\Gamma, \nabla v|_\Gamma \rangle = \frac{\partial v}{\partial \mathbf{n}} \langle \nabla dv[dr_2]|_\Gamma, \mathbf{n} \rangle = \frac{\partial v}{\partial \mathbf{n}} \frac{\partial dv[dr_2]}{\partial \mathbf{n}}.$$

We like to stress that we have proven in [14] that the shape Hessian at the optimal domain $\Omega^\star$ is a compact mapping $H^{1/2}(\Gamma^\star) \to H^{-1/2}(\Gamma^\star)$, that is, in its natural energy space. This issues from the fact that it holds $\partial v/\partial \mathbf{n} = \partial w/\partial \mathbf{n}$ on $\Gamma^\star$ due to the necessary condition (1.6). Hence, the first two terms in (1.9) cancel out and only the third term remains containing the difference $\partial dv[dr]/\partial \mathbf{n} - \partial dw[dr]/\partial \mathbf{n}$. This difference yields the compactness because the local shape derivatives differ only from the boundary conditions on Σ; compare (1.4). As a main consequence, the poor presentation of the identification problem in EIT is strongly related to the poor presentation of the optimization problem (1.1), (1.2). We refer the reader to [14] for the details.

1.3 Reformulating the shape Hessian

This subsection is intended to transform the second term of the shape Hessian 1.9 so that it can be calculated. For the sake of brevity, we formulate the next results only with respect to v. But, of course, the equivalent results are valid also with respect to w.

Lemma 1.2

Let the normalized tangent $\mathbf{t}$ *in the point* $\mathbf{x} = r(\widehat{\mathbf{x}}) \cdot \widehat{\mathbf{x}} \in \Gamma$ *be defined by*

$$\mathbf{t} = \frac{\mathbf{n} \times (\widehat{\mathbf{x}} \times \mathbf{n})}{\|\mathbf{n} \times (\widehat{\mathbf{x}} \times \mathbf{n})\|} = \frac{\|\nabla_{\mathbb{S}} r(\widehat{\mathbf{x}})\|^2 \widehat{\mathbf{x}} + r \nabla_{\mathbb{S}} r(\widehat{\mathbf{x}})}{\|\nabla_{\mathbb{S}} r(\widehat{\mathbf{x}})\| \sqrt{r^2(\widehat{\mathbf{x}}) + \|\nabla_{\mathbb{S}} r(\widehat{\mathbf{x}})\|^2}}.$$

Then, on Γ *there holds the identity*

$$\frac{\partial}{\partial \widehat{\mathbf{x}}} \|\nabla v\|^2 = 2 \frac{\partial v}{\partial \mathbf{n}} \left\{ \frac{\|\nabla_{\mathbb{S}} r(\widehat{\mathbf{x}})\|}{\sqrt{r^2(\widehat{\mathbf{x}}) + \|\nabla_{\mathbb{S}} r(\widehat{\mathbf{x}})\|^2}} \frac{\partial^2 v}{\partial \mathbf{n} \partial \mathbf{t}} - \frac{r(\widehat{\mathbf{x}})}{\sqrt{r^2(\widehat{\mathbf{x}}) + \|\nabla_{\mathbb{S}} r(\widehat{\mathbf{x}})\|^2}} \frac{\partial^2 v}{\partial \mathbf{n}^2} \right\},$$

where $\partial^2 v/\partial \mathbf{n}^2 := \langle \nabla^2 v \cdot \mathbf{n}, \mathbf{n} \rangle$ *and* $\partial^2 v/(\partial \mathbf{n} \partial \mathbf{t}) := \langle \nabla^2 v \cdot \mathbf{n}, \mathbf{t} \rangle$.

Proof

We decompose the spatial directions into the normal $\mathbf{n}$ in the point $\mathbf{x} \in \Gamma$ and two orthonormal tangential directions

$$\mathbf{s} = \frac{\widehat{\mathbf{x}} \times \mathbf{n}}{\|\widehat{\mathbf{x}} \times \mathbf{n}\|}, \qquad \mathbf{t} = \frac{\mathbf{n} \times (\widehat{\mathbf{x}} \times \mathbf{n})}{\|\mathbf{n} \times (\widehat{\mathbf{x}} \times \mathbf{n})\|}.$$

Note that

$$\mathbf{n} \times (\widehat{\mathbf{x}} \times \mathbf{n}) = \widehat{\mathbf{x}} - \langle \widehat{\mathbf{x}}, \mathbf{n} \rangle \cdot \mathbf{n} = \frac{\|\nabla_{\mathbb{S}} r(\widehat{\mathbf{x}})\|^2 \widehat{\mathbf{x}} + r \nabla_{\mathbb{S}} r(\widehat{\mathbf{x}})}{r^2(\widehat{\mathbf{x}}) + \|\nabla_{\mathbb{S}} r(\widehat{\mathbf{x}})\|^2},$$

$$\|\mathbf{n} \times (\widehat{\mathbf{x}} \times \mathbf{n})\| = \frac{\|\nabla_{\mathbb{S}} r(\widehat{\mathbf{x}})\|}{\sqrt{r^2(\widehat{\mathbf{x}}) + \|\nabla_{\mathbb{S}} r(\widehat{\mathbf{x}})\|^2}},$$

and hence

$$\mathbf{t} = \frac{\|\nabla_{\mathbb{S}} r(\widehat{\mathbf{x}})\|^2 \widehat{\mathbf{x}} + r \nabla_{\mathbb{S}} r(\widehat{\mathbf{x}})}{\|\nabla_{\mathbb{S}} r(\widehat{\mathbf{x}})\| \sqrt{r^2(\widehat{\mathbf{x}}) + \|\nabla_{\mathbb{S}} r(\widehat{\mathbf{x}})\|^2}}.$$

The ansatz

$$\widehat{\mathbf{x}} = \alpha \mathbf{n} + \beta \mathbf{s} + \gamma \mathbf{t}$$

leads to

$$\alpha = -\frac{r(\widehat{\mathbf{x}})}{\sqrt{r^2(\widehat{\mathbf{x}}) + \|\nabla_{\mathbb{S}} r(\widehat{\mathbf{x}})\|^2}}, \qquad \gamma = \frac{\|\nabla_{\mathbb{S}} r(\widehat{\mathbf{x}})\|}{\sqrt{r^2(\widehat{\mathbf{x}}) + \|\nabla_{\mathbb{S}} r(\widehat{\mathbf{x}})\|^2}},$$

and $\beta = 0$ because $\widehat{\mathbf{x}} \perp \widehat{\mathbf{x}} \times \mathbf{n}$. Consequently, we find

$$\begin{aligned}\frac{\partial}{\partial \widehat{\mathbf{x}}} \|\nabla v\|^2 &= \langle \nabla \|\nabla v\|^2, \widehat{\mathbf{x}} \rangle = 2\langle \nabla^2 v \cdot \widehat{\mathbf{x}}, \nabla v \rangle = 2\frac{\partial v}{\partial \mathbf{n}} \langle \nabla^2 v \cdot \mathbf{n}, \widehat{\mathbf{x}} \rangle \\ &= 2\frac{\partial v}{\partial \mathbf{n}} \left\{ \frac{\|\nabla_{\mathbb{S}} r(\widehat{\mathbf{x}})\|}{\sqrt{r^2(\widehat{\mathbf{x}}) + \|\nabla_{\mathbb{S}} r(\widehat{\mathbf{x}})\|^2}} \frac{\partial^2 v}{\partial \mathbf{n} \partial \mathbf{t}} - \frac{r(\widehat{\mathbf{x}})}{\sqrt{r^2(\widehat{\mathbf{x}}) + \|\nabla_{\mathbb{S}} r(\widehat{\mathbf{x}})\|^2}} \frac{\partial^2 v}{\partial \mathbf{n}^2} \right\}.\end{aligned}$$

Hence, we have reduced the second term of the shape Hessian (1.9) to second-order derivatives of the states. The next lemma shows how to compute the second-order normal derivative.

Lemma 1.3

We denote by $\mathcal{H}$ the mean curvature. Then, on Γ there holds the identity

$$\frac{\partial^2 v}{\partial \mathbf{n}^2} = 2\mathcal{H} \frac{\partial v}{\partial \mathbf{n}} \tag{1.10}$$

provided that $v \in C^2(\overline{\Omega})$.

Proof

Because $v \in C^2(\overline{\Omega})$, the Laplace equation holds up to the boundary Γ. It might be written as (see [5], for example)

$$\Delta v = \frac{\partial^2 v}{\partial \mathbf{n}^2} - 2\mathcal{H} \frac{\partial v}{\partial \mathbf{n}} + \Delta_\Gamma v = 0,$$

where Δ_Γ denotes the Laplace-Beltrami operator with respect to Γ. The homogenous Dirichlet condition on Γ implies $\Delta_\Gamma v = 0$, which immediately yields the assertion.

Throughout the remainder of this paper we shall assume that the boundary manifold $\partial\Omega$ is given as a parametric surface consisting of smooth patches. More precisely, let $\square := [0,1]^2$ denote the unit square. The manifold $\partial\Omega = \Sigma \cup \Gamma \in \mathbb{R}^3$ is partitioned into a finite number of *patches*

$$\partial\Omega = \bigcup_{i=1}^{M} \Gamma_i, \qquad \Gamma_i = \gamma_i(\square), \qquad i = 1, 2, \ldots, M, \tag{1.11}$$

where each $\gamma_i : \square \to \Gamma_i$ defines a diffeomorphism of $\square$ onto Γ_i. The intersection $\Gamma_i \cap \Gamma_{i'}$, $i \neq i'$, of the patches Γ_i and $\Gamma_{i'}$ is supposed to be either $\emptyset$ or a common edge or vertex.

Abbreviating for $\mathbf{s} = [s_1, s_2]^T \in \square$

$$\gamma_{i,j}(\mathbf{s}) := \frac{\partial \gamma_i(\mathbf{s})}{\partial s_j}, \qquad \gamma_{i,j,k}(\mathbf{s}) := \frac{\partial^2 \gamma_i(\mathbf{s})}{\partial s_j \partial s_k}, \qquad j,k = 1,2,$$

the first and second fundamental tensors of differential geometry are given by

$$\mathbf{K}_i(\mathbf{s}) = [\langle \gamma_{i,j}(\mathbf{s}), \gamma_{i,k}(\mathbf{s})\rangle]_{j,k=1,2}, \qquad \mathbf{L}_i(\mathbf{s}) = [\langle \mathbf{n}, \gamma_{i,j,k}(\mathbf{s})\rangle]_{j,k=1,2}.$$

Using these definitions, the mean curvature involved in (1.10) reads as (cf. [1])

$$\mathcal{H}(\gamma_i(\mathbf{s})) = \frac{1}{2}\text{trace}(\mathbf{K}_i^{-1}(\mathbf{s})\mathbf{L}_i(\mathbf{s})).$$

Moreover, consider a function

$$u \in H^1(\partial\Omega)$$

which is defined via parametrization, that is, we have functions $\phi_i : \square \to \mathbb{R}$ satisfying $u \circ \gamma_i = \phi_i$, $i = 1, 2, \ldots, M$. Then, according to [1], the surface gradient $\nabla_{\partial\Omega} u$ is defined as follows

$$\nabla_{\partial\Omega} u(\gamma_i(\mathbf{s})) = [\gamma_{i,1}(\mathbf{s}), \gamma_{i,2}(\mathbf{s})]\mathbf{K}_i^{-1}(\mathbf{s}) \begin{bmatrix} \frac{\partial \phi_i(\mathbf{s})}{\partial s_1} \\ \frac{\partial \phi_i(\mathbf{s})}{\partial s_2} \end{bmatrix}. \tag{1.12}$$

With these preparations at hand, we are able to prove the next lemma, which makes use of the homogenous Dirichlet boundary conditions of v on Γ.

Lemma 1.4

On Γ there holds the identity

$$\frac{\partial^2 v}{\partial \mathbf{n} \partial \mathbf{t}} = \left\langle \nabla_\Gamma \frac{\partial v}{\partial \mathbf{n}}, \mathbf{t} \right\rangle.$$

Proof

Invoking the parametrization, we find

$$\frac{\partial}{\partial s_1}\frac{\partial v}{\partial \mathbf{n}} = \frac{\partial}{\partial s_1}\langle \nabla v, \mathbf{n}\rangle = \langle \nabla^2 v \cdot \mathbf{n}, \gamma_{i,1}\rangle + \left\langle \nabla v, \frac{\partial}{\partial s_1}\mathbf{n}\right\rangle.$$

From

$$\begin{aligned}\frac{\partial}{\partial s_1}\mathbf{n} &= \frac{\partial}{\partial s_1}\frac{\gamma_{i,1} \times \gamma_{i,2}}{\|\gamma_{i,1} \times \gamma_{i,2}\|} \\ &= \frac{1}{\|\gamma_{i,1} \times \gamma_{i,2}\|}\left\{\frac{\partial}{\partial s_1}(\gamma_{i,1} \times \gamma_{i,2}) - \left\langle \mathbf{n}, \frac{\partial}{\partial s_1}(\gamma_{i,1} \times \gamma_{i,2})\right\rangle \cdot \mathbf{n}\right\}\end{aligned}$$

and

$$\left\langle \nabla v, \frac{\partial}{\partial s_1}(\gamma_{i,1} \times \gamma_{i,2}) \right\rangle = \frac{\partial v}{\partial \mathbf{n}} \left\langle \mathbf{n}, \frac{\partial}{\partial s_1}(\gamma_{i,1} \times \gamma_{i,2}) \right\rangle$$

we conclude

$$\left\langle \nabla v, \frac{\partial}{\partial s_1} \mathbf{n} \right\rangle = 0.$$

One infers the analogous result with respect to the derivative $\partial/\partial s_2$ such that we arrive at

$$\frac{\partial}{\partial s_1} \frac{\partial v}{\partial \mathbf{n}} = \langle \nabla^2 v \cdot \mathbf{n}, \gamma_{i,1} \rangle, \qquad \frac{\partial}{\partial s_2} \frac{\partial v}{\partial \mathbf{n}} = \langle \nabla^2 v \cdot \mathbf{n}, \gamma_{i,2} \rangle.$$

Next, defining the two tangential vectors $\widetilde{\gamma}_{i,1}$ and $\widetilde{\gamma}_{i,2}$ via

$$[\widetilde{\gamma}_{i,1}, \widetilde{\gamma}_{i,2}] := [\gamma_{i,1}, \gamma_{i,2}] \mathbf{K}_i^{-1}$$

one readily verifies

$$\langle \gamma_{i,k}, \widetilde{\gamma}_{i,l} \rangle = \delta_{j,k}, \qquad k, l = 1, 2.$$

Hence, we can rewrite the tangent $\mathbf{t}$ by

$$\mathbf{t} = \langle \mathbf{t}, \widetilde{\gamma}_{i,1} \rangle \gamma_{i,1} + \langle \mathbf{t}, \widetilde{\gamma}_{i,2} \rangle \gamma_{i,2},$$

which implies

$$\begin{aligned} \frac{\partial^2 v}{\partial \mathbf{n} \partial \mathbf{t}} &= \langle \nabla^2 v \cdot \mathbf{n}, \mathbf{t} \rangle = \langle \nabla^2 v \cdot \mathbf{n}, \gamma_{i,1} \rangle \langle \mathbf{t}, \widetilde{\gamma}_{i,1} \rangle + \langle \nabla^2 v \cdot \mathbf{n}, \gamma_{i,2} \rangle \langle \mathbf{t}, \widetilde{\gamma}_{i,2} \rangle \\ &= \left\langle [\gamma_{i,1}, \gamma_{i,2}] \mathbf{K}_i^{-1} \begin{bmatrix} \langle \nabla^2 v \cdot \mathbf{n}, \gamma_{i,1} \rangle \\ \langle \nabla^2 v \cdot \mathbf{n}, \gamma_{i,2} \rangle \end{bmatrix}, \mathbf{t} \right\rangle = \left\langle \nabla_\Gamma \frac{\partial v}{\partial \mathbf{n}}, \mathbf{t} \right\rangle. \end{aligned}$$

We now combine the Lemmas 1.2, 1.3 and 1.4 and derive the final result.

Corollary 1.5

The shape Hessian 1.9 is equivalent to

$$\begin{aligned} d^2 J(\Omega)[dr_1, dr_2] = \int_{\mathbb{S}} & 2r \, dr_1 \, dr_2 \left[1 - \frac{2\mathcal{H} r^2}{\sqrt{r^2 + \|\nabla_{\mathbb{S}} r\|^2}} \right] \cdot \left[\left(\frac{\partial v}{\partial \mathbf{n}} \right)^2 - \left(\frac{\partial w}{\partial \mathbf{n}} \right)^2 \right] \\ & + 2r^2 \, dr_1 \, dr_2 \left[\frac{\partial v}{\partial \mathbf{n}} \left\langle \nabla_\Gamma \frac{\partial v}{\partial \mathbf{n}}, \mathbf{n} \times (\widehat{\mathbf{x}} \times \mathbf{n}) \right\rangle \right. \\ & \left. - \frac{\partial w}{\partial \mathbf{n}} \left\langle \nabla_\Gamma \frac{\partial w}{\partial \mathbf{n}}, \mathbf{n} \times (\widehat{\mathbf{x}} \times \mathbf{n}) \right\rangle \right] \\ & + 2r^2 \, dr_1 \left[\frac{\partial w}{\partial \mathbf{n}} \frac{\partial dw[dr_2]}{\partial \mathbf{n}} - \frac{\partial v}{\partial \mathbf{n}} \frac{\partial dv[dr_2]}{\partial \mathbf{n}} \right] d\sigma. \end{aligned} \tag{1.13}$$

1.4 Boundary integral equations

In this subsection we compute the unknown boundary data of the state functions v and w by boundary integral equations. We introduce the single-layer and the double-layer operators with respect to the boundaries $\Phi, \Psi \in \{\Gamma, \Sigma\}$ by

$$(V_{\Phi\Psi}u)(\mathbf{x}) := -\frac{1}{4\pi}\int_{\Phi}\frac{1}{\|\mathbf{x}-\mathbf{y}\|}u(\mathbf{y})d\sigma_{\mathbf{y}}, \qquad \mathbf{x}\in\Psi,$$

$$(K_{\Phi\Psi}u)(\mathbf{x}) := \frac{1}{4\pi}\int_{\Phi}\frac{\langle\mathbf{x}-\mathbf{y},\mathbf{n}_{\mathbf{y}}\rangle}{\|\mathbf{x}-\mathbf{y}\|^3}u(\mathbf{y})d\sigma_{\mathbf{y}}, \qquad \mathbf{x}\in\Psi.$$

Note that $V_{\Phi\Psi}$ denotes an operator of order -1 if $\Phi = \Psi$, that is $V_{\Phi\Phi} : H^{-1/2}(\Phi) \to H^{1/2}(\Phi)$, while it is an arbitrarily smoothing compact operator if $\Phi \neq \Psi$ because $\operatorname{dist}(\Gamma, \Sigma) > 0$. Likewise, if $\Sigma, \Gamma \in C^2$, the double-layer operator $K_{\Phi\Phi} : H^{1/2}(\Phi) \to H^{1/2}(\Phi)$ is compact while it smooths arbitrarily if $\Phi \neq \Psi$. We refer the reader to [1–3] for a detailed description of boundary integral equations.

The normal derivative of w is given by the Dirichlet-to-Neumann map

$$\begin{bmatrix} V_{\Gamma\Gamma} & V_{\Sigma\Gamma} \\ V_{\Gamma\Sigma} & V_{\Sigma\Sigma} \end{bmatrix} \begin{bmatrix} \frac{\partial w}{\partial \mathbf{n}}|_{\Gamma} \\ \frac{\partial w}{\partial \mathbf{n}}|_{\Sigma} \end{bmatrix} = \begin{bmatrix} 1/2 + K_{\Gamma\Gamma} & K_{\Sigma\Gamma} \\ K_{\Gamma\Sigma} & 1/2 + K_{\Sigma\Sigma} \end{bmatrix} \begin{bmatrix} 0 \\ f \end{bmatrix}; \tag{1.14}$$

compare (1.1). Likewise, the unknown boundary data of v are determined by

$$\begin{bmatrix} V_{\Gamma\Gamma} & -K_{\Sigma\Gamma} \\ -V_{\Gamma\Sigma} & 1/2 + K_{\Sigma\Sigma} \end{bmatrix} \begin{bmatrix} \frac{\partial v}{\partial \mathbf{n}}|_{\Gamma} \\ v|_{\Sigma} \end{bmatrix} = \begin{bmatrix} 1/2 + K_{\Gamma\Gamma} & -V_{\Sigma\Gamma} \\ -K_{\Gamma\Sigma} & V_{\Sigma\Sigma} \end{bmatrix} \begin{bmatrix} 0 \\ g \end{bmatrix}. \tag{1.15}$$

The unknown boundary data of the local shape derivatives $dv = dv[dr]$ and $dw = dw[dr]$ are derived from the boundary integral equations

$$\begin{bmatrix} V_{\Gamma\Gamma} & V_{\Sigma\Gamma} \\ V_{\Gamma\Sigma} & V_{\Sigma\Sigma} \end{bmatrix} \begin{bmatrix} \frac{\partial dw}{\partial \mathbf{n}}|_{\Gamma} \\ \frac{\partial dw}{\partial \mathbf{n}}|_{\Sigma} \end{bmatrix} = \begin{bmatrix} 1/2 + K_{\Gamma\Gamma} & K_{\Sigma\Gamma} \\ K_{\Gamma\Sigma} & 1/2 + K_{\Sigma\Sigma} \end{bmatrix} \begin{bmatrix} -\langle\mathbf{V},\mathbf{n}\rangle\frac{\partial w}{\partial \mathbf{n}}|_{\Gamma} \\ 0 \end{bmatrix} \tag{1.16}$$

and

$$\begin{bmatrix} V_{\Gamma\Gamma} & -K_{\Sigma\Gamma} \\ -V_{\Gamma\Sigma} & 1/2 + K_{\Sigma\Sigma} \end{bmatrix} \begin{bmatrix} \frac{\partial dv}{\partial \mathbf{n}}|_{\Gamma} \\ dv|_{\Sigma} \end{bmatrix} = \begin{bmatrix} 1/2 + K_{\Gamma\Gamma} & -V_{\Sigma\Gamma} \\ -K_{\Gamma\Sigma} & V_{\Sigma\Sigma} \end{bmatrix} \begin{bmatrix} -\langle\mathbf{V},\mathbf{n}\rangle\frac{\partial v}{\partial \mathbf{n}}|_{\Gamma} \\ 0 \end{bmatrix}. \tag{1.17}$$

2 Discretization

2.1 Finite dimensional approximation of boundaries

Because the infinite dimensional optimization problem cannot be solved directly, we replace it with a finite dimensional problem. Recall that the boundary Γ allows a unique representation

$$\Gamma = \big\{ r(\widehat{\mathbf{x}}) \cdot \widehat{\mathbf{x}} \in \mathbb{R}^3 : \widehat{\mathbf{x}} \in \mathbb{S} \big\}, \tag{2.18}$$

and the regularity $\Gamma \in C^{2,\alpha}$ is directly associated with $r \in C^{2,\alpha}(\mathbb{S})$. We now introduce the spherical harmonics.

For $n \in \mathbb{N}_0$ and $|m| \leq n$, consider the Legendre polynomials

$$P_n(t) := \frac{1}{2^n n!} \Big(\frac{d}{dt} \Big)^n (t^2 - 1)^n, \quad t \in \mathbb{R},$$

and the associated Legendre functions

$$P_n^{|m|}(t) := (1 - t^2)^{|m|/2} \Big(\frac{d}{dt} \Big)^{|m|} P_n(t), \quad t \in \mathbb{R}.$$

Then, the spherical harmonics $Y_n^m : \mathbb{S} \to \mathbb{R}$ are given by

$$Y_n^m(\widehat{\mathbf{x}}) := \sqrt{\frac{2n+1}{4\pi} \frac{(n-|m|)!}{(n+|m|)!}} P_n^{|m|}(\widehat{x}_3) \begin{cases} \operatorname{Re}((\widehat{x}_1 + i\widehat{x}_2)^m), & m \geq 0, \\ \operatorname{Im}((\widehat{x}_1 + i\widehat{x}_2)^m), & m < 0. \end{cases}$$

Because the spherical harmonics are the restriction of homogeneous harmonical polynomials to the unit sphere, the radial function r in (2.18) admits a unique representation

$$r(\widehat{\mathbf{x}}) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \alpha_{m,n} Y_n^m(\widehat{\mathbf{x}}), \quad \widehat{\mathbf{x}} \in \mathbb{S},$$

with certain numbers $\alpha_{m,n} \in \mathbb{R}$. Hence, it is reasonable to take a truncated series

$$r_N(\widehat{\mathbf{x}}) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \alpha_{m,n} Y_n^m(\widehat{\mathbf{x}}), \quad \widehat{\mathbf{x}} \in \mathbb{S}, \tag{2.19}$$

as an approximation of r. Other boundary representations like B-splines can be considered as well. The advantage of our approach is an exponential convergence $r_N \to r$ if the shape is analytical.

Because r_N has the $(N+1)^2$ degrees of freedom $a_{0,0}, a_{-1,1}, \ldots, a_{N,N}$, we arrive at a finite dimensional optimization problem in the open set

$$\Upsilon_N := \{ (a_{0,0}, a_{-1,1}, \ldots, a_{N,N}) : r_N > 0 \text{ on } \mathbb{S} \text{ and } \operatorname{dist}(\Sigma, \Gamma) > 0 \},$$

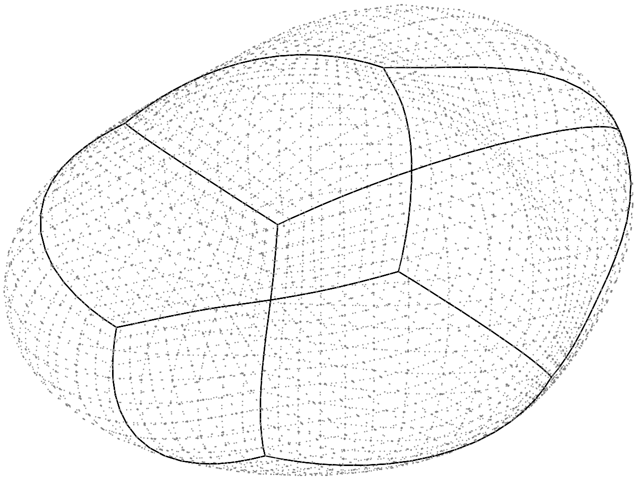


FIGURE 7.2 Parametric representation of Ω.

which is a subset of $\mathbb{R}^{(N+1)^2}$. Then, via the identification $r_N \Leftrightarrow \Omega_N$, the finite dimensional approximation of problem (1.2) reads as

$$J(\Omega_N) \to \min. \tag{2.20}$$

The associated gradient $dJ(\Omega_N)[dr]$ and Hessian $d^2J(\Omega_N)[dr_1, dr_2]$ must be computed with respect to all directions $dr, dr_1, dr_2 = Y_n^m(\mathbf{x})\mathbf{x}$, $m = -n, \ldots, n$, and $n = 0, \ldots, N$.

We would like to point out that a parametric representation in accordance with Subsection 1.3 can be constructed as follows. The cube $[-0.5, 0.5]^3$ consists of six patches. Each point $\mathbf{x} \in [-0.5, 0.5]^3$ can be placed on the boundary Γ via the operation

$$\mathbf{y}(\mathbf{x}) = r_N\left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right) \cdot \frac{\mathbf{x}}{\|\mathbf{x}\|} \in \Gamma. \tag{2.21}$$

That way, the surface is subdivided into six patches. The parametric representations $\gamma_i : \Gamma_i \to \Gamma$ can be easily derived from (2.21). In Figure 7.2 one finds an illustration of the proposed parametric representation.

2.2 Treating the optimization problem

The minimization problem defined by (2.20) implies to find its stationary points $\Omega_{N_r}^\star$

$$dJ(\Omega_{N_r}^\star)[dr] = 0 \tag{2.22}$$

for all directions $dr = Y_n^m(\mathbf{x})\mathbf{x}$, $m = -n, \ldots, n$, and $n = 0, \ldots, N$. To solve (2.22), we consider on the one hand a method that is based only on first-order information, namely a quasi-Newton method updated by the inverse BFGS-rule without damping; see [24,25] for the details.

On the other hand, we perform a Newton method, which we regularize since the shape Hessian is compact at the optimal domain $\Omega^\star$. Abbreviating the discrete gradient by $\mathbf{G}_n$ and the associated Hessian by $\mathbf{H}_n$, we consider in the n-th iteration step the descent direction

$$\mathbf{h}_n := -(\mathbf{H}_n^2 + \alpha_n \mathbf{I})^{-1}\mathbf{H}_n\mathbf{G}_n,$$

where $\alpha_n > 0$ is an appropriately chosen regularization parameter. This descent direction $\mathbf{h}_n$ solves the minimization problem

$$\|\mathbf{H}_n\mathbf{h}_n - \mathbf{G}_n\|^2 + \alpha_n\|\mathbf{h}_n\|^2 \to \min$$

and corresponds to a Tikhinov regularization of Equation (2.22). Moreover, note that we employ in both methods a quadratic line search with respect to the functional (1.2) based on the information of the actual value of the cost functional and its gradient, and on the value of the cost functional with respect to the new domain.

2.3 Numerical method to compute the state

We want to employ a boundary element method to compute the required boundary data of the state equations. Recall that in Subsection 1.3 we introduced a parametric representation of boundary $\partial\Omega = \Gamma \cup \Sigma$ by quadrilateral patches. A mesh of level j on $\partial\Omega$ is then induced by dyadic subdivisions of depth j of the reference square $\square$ into 4^j squares. This generates $4^j M$ *elements* (or elementary domains). On the given mesh we consider on each boundary $\Phi \in \{\Gamma, \Sigma\}$ piecewise bilinear basis functions $\{\theta_{j,k}^\Phi : k \in \triangle_j^\Phi\}$, where $\triangle_j^\Phi$ denotes an appropriate index set.

For $\Phi, \Psi \in \{\Sigma, \Gamma\}$, we introduce the system matrices

$$\mathbf{V}_{\Phi\Psi} = \frac{1}{4\pi}\left[\int_\Psi \int_\Phi \frac{1}{\|\mathbf{x}-\mathbf{y}\|}\theta_i^\Phi(\mathbf{y})\theta_j^\Psi(\mathbf{x})d\sigma_\mathbf{y}d\sigma_\mathbf{x}\right]_{i\in\triangle_j^\Phi, j\in\triangle_j^\Psi},$$

$$\mathbf{K}_{\Phi\Psi} = \frac{1}{4\pi}\left[\int_\Psi \int_\Phi \frac{\langle \mathbf{x}-\mathbf{y}, \mathbf{n}_\mathbf{y}\rangle}{\|\mathbf{x}-\mathbf{y}\|^3}\theta_i^\Phi(\mathbf{y})\theta_j^\Psi(\mathbf{x})d\sigma_\mathbf{y}d\sigma_\mathbf{x}\right]_{i\in\triangle_j^\Phi, j\in\triangle_j^\Psi},$$

and the mass matrices

$$\mathbf{M}_\Phi = \left[\int_\Phi \theta_i^\Phi(\mathbf{x})\theta_j^\Phi(\mathbf{x})d\sigma_\mathbf{x}\right]_{i,j\in\triangle_j^\Phi},$$

and the load vectors of Dirichlet data f_Φ and Neumann data g_Φ

$$\mathbf{f}_\Phi = \left[\int_\Phi \theta_i^\Phi(\mathbf{x}) f(\mathbf{x}) d\sigma_{\mathbf{x}}\right]_{i\in\triangle_j^\Phi}, \qquad \mathbf{g}_\Phi = \left[\int_\Phi \theta_i^\Phi(\mathbf{x}) g(\mathbf{x}) d\sigma_{\mathbf{x}}\right]_{i\in\triangle_j^\Phi}.$$

Then, the linear system of equations

$$\begin{bmatrix} \mathbf{V}_{\Gamma\Gamma} & \mathbf{V}_{\Sigma\Gamma} \\ \mathbf{V}_{\Gamma\Sigma} & \mathbf{V}_{\Sigma\Sigma} \end{bmatrix} \begin{bmatrix} \mathbf{a}_\Gamma \\ \mathbf{a}_\Sigma \end{bmatrix} = \begin{bmatrix} 1/2\mathbf{M}_\Gamma + \mathbf{K}_{\Gamma\Gamma} & \mathbf{K}_{\Sigma\Gamma} \\ \mathbf{K}_{\Gamma\Sigma} & 1/2\mathbf{M}_\Sigma + \mathbf{K}_{\Sigma\Sigma} \end{bmatrix} \begin{bmatrix} \mathbf{M}_\Gamma^{-1}\mathbf{f}_\Gamma \\ \mathbf{M}_\Sigma^{-1}\mathbf{f}_\Sigma \end{bmatrix}, \tag{2.23}$$

gives us the Neumann data $a_\Gamma = \sum_{i\in\triangle_j^\Gamma} [\mathbf{a}_\Gamma]_i \theta_i^\Gamma$ on Γ and

$$a_\Sigma = \sum_{i\in\triangle_j^\Sigma} [\mathbf{a}_\Sigma]_i \theta_i^\Sigma$$

on Σ from the Dirichlet data on Γ and Σ. Likewise, the system

$$\begin{bmatrix} \mathbf{V}_{\Gamma\Gamma} & -\mathbf{K}_{\Sigma\Gamma} \\ -\mathbf{V}_{\Gamma\Sigma} & 1/2\mathbf{M}_\Sigma + \mathbf{K}_{\Sigma\Sigma} \end{bmatrix} \begin{bmatrix} \mathbf{b}_\Gamma \\ \mathbf{a}_\Gamma \end{bmatrix} = \begin{bmatrix} 1/2\mathbf{M}_\Gamma + \mathbf{K}_{\Gamma\Gamma} & -\mathbf{V}_{\Sigma\Gamma} \\ -\mathbf{K}_{\Gamma\Sigma} & \mathbf{V}_{\Sigma\Sigma} \end{bmatrix} \begin{bmatrix} \mathbf{M}_\Gamma^{-1}\mathbf{g}_\Gamma \\ \mathbf{M}_\Sigma^{-1}\mathbf{f}_\Sigma \end{bmatrix}, \tag{2.24}$$

yields the Dirichlet data $b_\Gamma = \sum_{i\in\triangle_j^\Gamma} [\mathbf{b}_\Gamma]_i \theta_i^\Gamma$ on Γ and the Neumann data $a_\Sigma = \sum_{i\in\triangle_j^\Sigma} [\mathbf{a}_\Sigma]_i \theta_i^\Sigma$ on Σ from the Neumann data $\mathbf{g}_\Gamma$ on Γ and the Dirichlet data $\mathbf{f}_\Sigma$ on Σ. Note that we plugged in the L^2-orthogonal projection involving $\mathbf{M}_\Phi^{-1}$ to decouple the data vectors from the boundary integral operators on the right-hand side; see also [12,13].

Using the traditional piecewise bilinear nodal basis functions leads to the traditional boundary element method. Then the system matrices are densely populated and we end up with an at least quadratic complexity for computing the approximate solution of (2.23) and (2.24), that is, computational work scales like $\mathcal{O}((|\triangle_j^\Gamma| + |\triangle_j^\Sigma|)^2) = \mathcal{O}(16^j)$.

We employ instead appropriate *biorthogonal spline wavelets* as constructed in several papers; see, for example, [15,18,21]. Then we obtain quasi-sparse system matrices having only $\mathcal{O}(|\triangle_j^\Gamma| + |\triangle_j^\Sigma|) = \mathcal{O}(4^j)$ relevant matrix coefficients. Applying the matrix compression strategy developed in [16,19] combined with an exponentially convergent *hp*–quadrature method [20], the wavelet Galerkin scheme produces the approximate solution of (2.23) and (2.24) within linear complexity. In particular, due to the norm equivalences of the wavelet bases, the diagonals of the system matrices define appropriate preconditioners [16,17].

We mention that the appearing system matrices have to be computed only once for each domain, while the systems (2.23) and (2.24) have to be solved $(N+1)^2$ times with different right-hand sides to obtain the Neumann

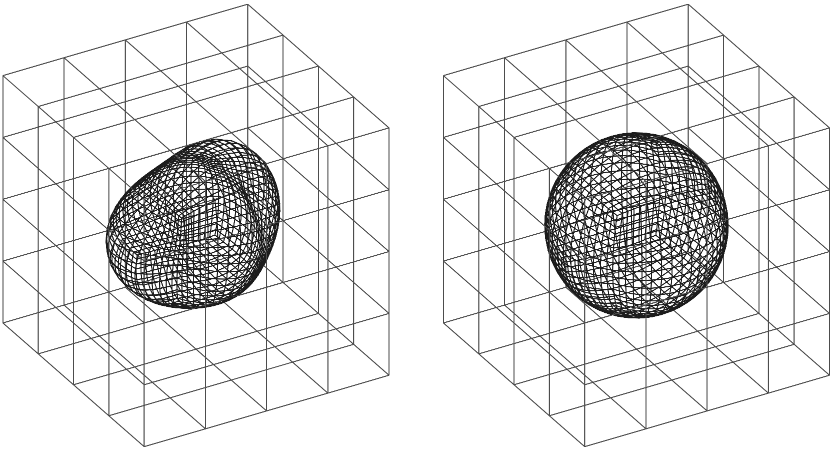


FIGURE 7.3 The exact inclusion (left) and the initial guess (right).

data of the local shape derivates. We emphasize that the iterative solution is much faster for the very sparsified system in wavelet coordinates compared to the dense system arising from the traditional boundary element method.

3 Numerical results

We choose D as the cube $[-1,1]^3$ and an inclusion S centered in $\mathbf{0}$ as shown in Figure 7.3. The Dirichlet data f are chosen as $(4x^2 - 3y^2 - z^2)|_\Sigma$, while the Neumann data g are computed numerically with appropriate accuracy. We use a sphere, also centered in $\mathbf{0}$, near the optimal domain as an initial guess, cf. Figure 7.3. The numerical setting is as follows. We choose $N = 5$, that is, 36 spherical harmonics to represent the boundary Γ. The cube is represented by six patches, which are twelve patches in all to represent the boundary $\partial\Omega$. The Galerkin discretization is performed on the mesh of level 4, which yields 3468 piecewise bilinear boundary elements. We follow [14] and choose $\alpha_n = 2^{-n}$ in the n-th step of the regularized Newton method. Thus, in each step we reduce the regularization parameter by a factor of 2. Again this choice turns out to be very efficient.

The left part of Figure 7.4 plots the history of the shape error coefficients, measured by the ℓ^2-norm of the coefficients associated with the spherical harmonics. The dashed line corresponds to the quasi-Newton method while the solid line belongs to the regularized Newton method. The regularized Newton method requires only 30 iteration steps to achieve the accuracy offered by the underlying discretization, which is indicated by stagnation of convergence about the shape error $5 \cdot 10^{-5}$; see Figure 7.4. In contrast, the quasi-Newton method does not compute the optimal shape as accurately even after 50 iterations. Its convergence is much slower compared to the

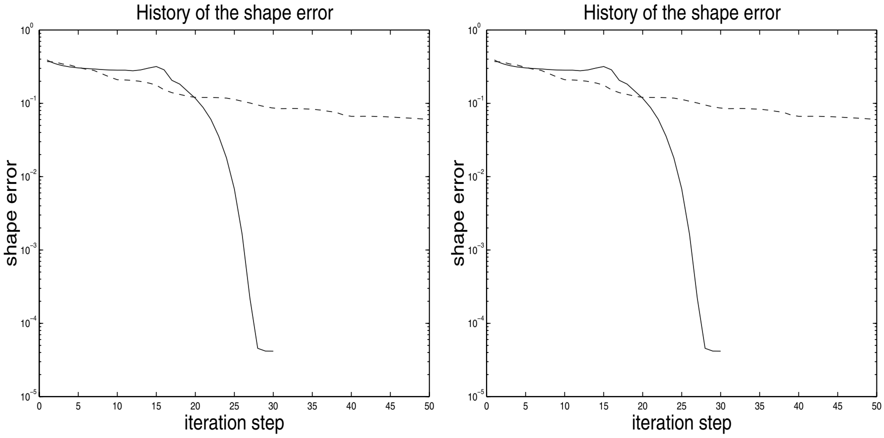


FIGURE 7.4 Shape error (left) and cost functional (right) versus iteration step.

regularized Newton method. Within 50 iterations it realizes only an approximation error of about $5 \cdot 10^{-2}$. Nearly the same behavior can be observed in the history of the cost functional, the left part of Figure 7.4. We emphasize that the regularized Newton scheme realizes a value of $5 \cdot 10^{-11}$ in constrast to $3 \cdot 10^{-5}$, which is achieved by the quasi-Newton method. The final approximations to the optimal domains can be found in Figure 7.5.

The Newton method consumes about 1.5 hours computing time at a standard personal computer, while the quasi-Newton method requires 10% more time. We mention that about 80 seconds are required to compute the system

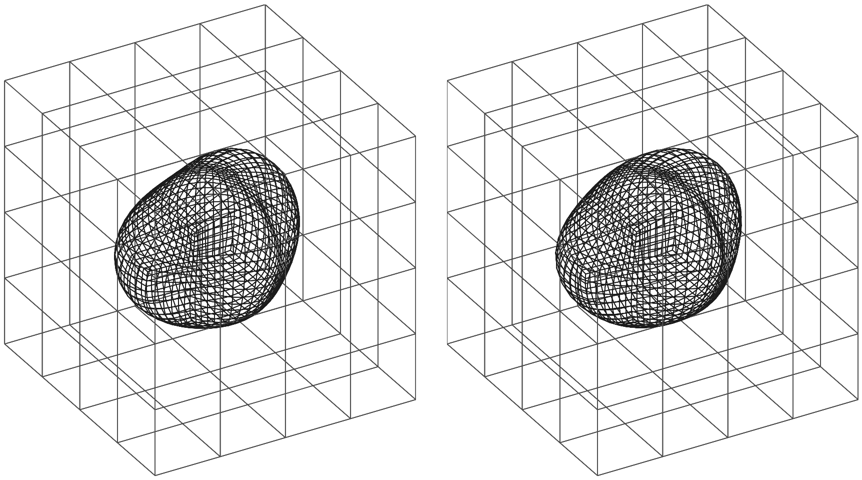


FIGURE 7.5 Approximate solutions produced by the regularized Newton method (left) and the quasi-Newton method (right).

matrices and to solve them with one right-hand side each. Therefore, one quasi-Newton step requires about 80 seconds if the line search becomes inactive, whereas a Newton step requires about twice that time, due mainly to the multiple iterative solution of the linear equation systems necessary to compute the local shape derivatives. But we emphasize that in the present example the regularized Newton scheme never requires the line search.

4 Conclusion

In this paper we considered second-order methods for the identification of voids or inclusions. The problem under consideration is known to be poorly presented. Because the shape Hessian is compact at the optimal domain, we propose a regularized Newton method for the resolution of the inclusion. The numerical example shows that the proposed regularized Newton method converges faster and yields a more accurate solution compared to a quasi-Newton scheme.

Acknowledgment

This research was carried out when the second author was a member of the Department of Mathematics, University of Utrecht, the Netherlands, supported by the EU-IHP project Nonlinear Approximation and Adaptivity: Breaking Complexity in Numerical Modelling and Data Representation.

References

[1] K. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, Pure and Applied Mathematics, Wiley, Chichester, 1983.

[2] R. Kress, *Linear Integral Equations*, Springer, Berlin-Heidelberg, 1989.

[3] W. Hackbusch, *Integralgleichungen*, B.G. Teubner, Stuttgart, 1989.

[4] M. Delfour and J.-P. Zolésio, *Shapes and Geometries*, SIAM, Philadelphia, 2001.

[5] J. Sokolowski and J.-P. Zolésio, *Introduction to Shape Optimization*, Springer, Berlin, 1992.

[6] G. Alessandrini, V. Isakov, and J. Powell, Local uniqueness in the inverse problem with one measurement, *Trans. Am. Math. Soc.*, 347, 1995, 3031–3041.

[7] A. Friedman and V. Isakov, On the uniqueness in the inverse conductivity problem with one measurement, *Indiana Univ. Math. J.*, 38, 1989, 563–579.

[8] M. Brühl, Explicit characterization of inclusions in electrical impedance tomography, *SIAM J. Math. Anal.*, 32, 2001, 1327–1341.

[9] M. Brühl and M. Hanke, Numerical implementation of two noniterative methods for locating inclusions by impedance tomography, *Inverse Problems*, 16, 2000, 1029–1042.

[10] J.-R. Roche and J. Sokolowski, Numerical methods for shape identification problems, *Control Cybern.*, 25, 1996, 867–894.

[11] F. Hettlich and W. Rundell, The determination of a discontinuity in a conductivity from a single boundary measurement, *Inverse Problems*, 14, 1998, 67–82.

[12] K. Eppler and H. Harbrecht, Numerical solution of elliptic shape optimization problems using wavelet-based BEM, *Optim. Methods Softw.*, 18, 2003, 105–123.

[13] K. Eppler and H. Harbrecht, 2nd Order Shape Optimization Using Wavelet BEM, 06-2003, TU Berlin, Institute of Mathematics, Berlin, 2003, submitted for publication.

[14] K. Eppler and H. Harbrecht, A regularized Newton method in electrical impedance tomography using shape Hessian information, 943, WIAS Berlin, 2004, submitted for publication.

[15] H. Harbrecht, Wavelet Galerkin Schemes for the Boundary Element Method in Three Dimensions, Technical University of Chemnitz, Chemnitz, 2001.

[16] R. Schneider, *Multiskalen- und Wavelet-Matrixkompression: Analysisbasierte Methoden zur Lösung großer vollbesetzter Gleichungssysteme*, B.G. Teubner, Stuttgart, 1998.

[17] W. Dahmen and A. Kunoth, Multilevel preconditioning, *Numer. Math.*, 63, 1992, 315–344.

[18] W. Dahmen, A. Kunoth, and K. Urban, Biorthogonal spline-wavelets on the interval—stability and moment conditions, *Appl. Comp. Harm. Anal.*, 6, 1999, 259–302.

[19] W. Dahmen, H. Harbrecht, and R. Schneider, Compression Techniques for Boundary Integral Equations—Optimal Complexity Estimates, SFB 393/02-06, Technical University of Chemnitz, Institute of Mathematics, Chemnitz, 2002.

[20] H. Harbrecht and R. Schneider, Wavelet Galerkin Schemes for Boundary Integral Equations—Implementation and Quadrature, SFB 393/02-21, Technical University of Chemnitz, Institute of Mathematics, D-09107 Chemnitz, 2002.

[21] H. Harbrecht and R. Schneider, Biorthogonal wavelet bases for the boundary element method, *Math. Nachr.*, 269–270, 2004, 167–188.

[22] K. Eppler, Optimal shape design for elliptic equations via BIE-methods, *J. of Applied Mathematics and Computer Science*, 10, 2000, 487–516.

[23] K. Eppler, Boundary integral representations of second derivatives in shape optimization, *Discussiones Mathematicae* (Differential Inclusion Control and Optimization), 20, 2000, 63–78.

[24] Ch. Grossmann and J. Terno, *Numerik der Optimierung*, B.G. Teubner, Stuttgart, 1993.

[25] P.E. Gill, W. Murray, and M.H. Wright, *Practical Optimization*, Academic Press, New York, 1981.

[26] R. Chapko and R. Kress, *A hybrid method for inverse boundary value problems in potential theory*, Preprint, University of Göttingen, Göttingen, Germany, 2003.

CHAPTER 8

Analysis for the shape gradient in inverse scattering

Pierre Dubois
France Telecom RD ANT, Fort de la Tête de Chien, La Turbie, France

Jean-Paul Zolésio
CNRS and INRIA, Sophia-Antipolis, France

1 Introduction

The *inverse scattering* problem in electromagnetic fields is studied through the identification or *reconstruction* of the obstacle. With respect to the measurement E_m of the scattered electric field in a *non-far* zone θ, we consider the classical minimization of a functional measuring the distance between E_m and the actual solution E over θ. We derive the expression for the shape derivative of the functional. The shape derivative techniques are those introduced in [1], [6], and [14]. The Maxwell solutions are developed in [3], [5], and [14]. Then we present the shape gradient calculus for a nonsmooth scattering surface, which could be a cylinder. We present numerical results of the shape gradient calculus on a metallic antenna's surface, using $SR3D$ software. These results are given in a 3D general setting instead of TM or TE.

2 Electromagnetic scattering

We present the analysis for the shape gradient in a 3D electromagnetic field in the presence of a scattering surface. B is a bounded body in $\mathcal{R}^3$, Γ its boundary, $\Gamma = \partial B$; Ω is the outer domain $\Omega = \mathcal{R}^3 \backslash \bar{B}$, $\bar{B} = B \cup \Gamma$ (Γ shall have no relative boundary). So we introduce the *technical boundary* S with radius R, it can be a sphere. We shall assume $O \in B$, so that for R large enough $\bar{B} \cup S = S$, $\Omega_{in} = \Omega \cap S$.

The electromagnetic field $(\vec{E}, \vec{H})$ is characterized by the Maxwell equation, in a homogenous domain Ω. Ω is defined by the following coefficients: $\varepsilon(\in C)$ the electrical permitivity, μ the magnetic permitivity, and $\sigma(\in R)$ its conductivity. The Maxwell equation follows:

$$\begin{cases} curl\vec{E} = -\frac{\partial}{\partial t}\epsilon\vec{E} - \vec{m}\partial_\Gamma \text{ on } \Omega \\ curl\vec{H} = \frac{\partial}{\partial t}\mu\vec{H} + \vec{j}\partial_\Gamma \\ div\vec{D} = \rho_e \\ div\vec{B} = \rho_m \end{cases} \tag{2.1}$$

where $\vec{j}$ and $\vec{m}$ are the electrical and magnetic current, respectively, that are the tangent field to the surface Γ. ρ_e and ρ_m are the volumic density respectively of electrical and magnetical quantity. Finally, we have $\vec{B}$ and $\vec{D}$ the density of, respectively, magnetical and electrical flow, which is defined by $\vec{D} = \varepsilon\vec{E}$ and $\vec{B} = \mu\vec{H}$. Thus, it follows that the boundary condition on a metallic surface Γ $(\sigma >> \omega\epsilon)$.

$$\begin{cases} \vec{E} \wedge \vec{n} = \vec{0} \qquad \text{on } \Gamma \\ \vec{H} \wedge \vec{n} = -\vec{j} \\ \vec{E} \cdot \vec{n} = \rho_e \\ \vec{H}.\vec{n} = 0 \end{cases} \tag{2.2}$$

where ρ_e is the volumic density of electrical quantity.

We use the following initial condition: $E(0) = e$ and $H(0) = h$ with $div\ e = 0$ and $div\ h = 0$, which implies $div\ E = div,\ H = 0$ in Ω (assuming ε and μ are constant).

The harmonic solutions of Maxwell equations are the complex valued fields E and H as follows:

$$\begin{cases} E(t,x) = \mathcal{R}(E(x).e^{i\omega t}) \\ H(t,x) = \mathcal{R}(H(x).e^{i\omega t}) \end{cases} \tag{2.3}$$

where E and H satisfy the Helmoltz system:

$$\begin{cases} curl\ E = i\omega\mu H - m\partial_\Omega \\ curlH = -i\omega\varepsilon E + j\partial_\Omega \end{cases} \tag{2.4}$$

which implies that $div\ E = div\ H = 0$ in Ω. So

$$curl\ curl\ E - k^2 E = -i\omega\mu j\partial_\Omega + grad\frac{\rho_e\partial_\Omega}{\varepsilon} - curlm\partial_\Omega \quad \text{in } \Omega \tag{2.5}$$

where $k = \sqrt{\mu\varepsilon}\omega$ is a wave number. The fundamental solution of the Helmholtz equation, which verifies the outgoing radiation condition, is given by Green's analytic kernel $G(X) = \frac{1}{4\pi}\frac{e^{ikr}}{r}$.

E_i is data that represent the incident electrical field. Notice that E_i is a physical field that verifies the Helmholtz equations. Find E_s ($\in H^1(\Omega)$) such that the following system

$$\begin{cases} curl\ curl\ \vec{E_s} - k^2\vec{E_s} = 0 & \text{on } \Omega \\ \vec{E_s} \wedge n = -\vec{E_i} \wedge n & \text{on } \Gamma \\ \lim_{r\to\infty} r(\partial_r \vec{E_s} + ik\vec{E_s}) = 0 \end{cases} \tag{2.6}$$

is a well-posed problem that has a unique solution over Ω [3]. We notice the charge conservation ($divE = 0$ for homogenous and isotropic domain) is assumed at the solution.

We set $\Omega_{in} = \mathcal{S}\backslash\overline{\mathcal{B}}$

$$\begin{cases} curl\ curl\ \vec{E_s} - k^2\vec{E_s} = 0 & \text{in } \Omega\backslash\overline{\Omega_{in}} \\ E_s \wedge n = p & \text{on } S \\ \lim_{r\to\infty} r(\partial_r \vec{E_s} + ik\vec{E_s}) = 0 \end{cases} \tag{2.7}$$

is well posed and has a unique solution (similar to (2.6)). We have the given value p $\in H^{1/2}(S)^3$ and $\exists\mathcal{A} \in \mathcal{L}(H^{1/2}(S)^3, H^{-1/2}(S)^3)$:

$$\mathcal{A}.E_s = \text{curl } \vec{E_s} \wedge n \quad \text{on } S. \tag{2.8}$$

We introduce the interior problem in Ω_{in} for $\vec{E_s} \in H^1(\Omega_{int})$

$$\begin{cases} curl\ curl\ \vec{E_s} - k^2\vec{E_s} = 0 & \text{in } \Omega_{in} \\ \vec{E_s} \wedge \vec{n} = -\vec{E_i} \wedge \vec{n} & \text{on } \Gamma \\ curl(\vec{E_s} \wedge n) = \mathcal{A}.\vec{E_s} & \text{on } S. \end{cases} \tag{2.9}$$

Proposition 2.1

Any solution to problem (2.9) can be extended over Ω in the unique radiating solution (2.6). The restriction to Ω_{in} of the radiating solution is the solution to the interior problem (2.6).

Corollary 2.2

The interior problem (2.9) has a unique solution.

We introduce the following decomposition $E_t(x) = E_i(x) + E_s(x)$, with $x \in \Omega$, where $E_t(x)$ is the total electrical field in a Ω vertex. $E_s(x)$ is the electrical scattered field by Γ. The boundary result (2.8) on S cannot be verified by the E_i vector field, so we introduce the given value g:

$$\mathcal{A}.E_s = g - \text{curl } \vec{E_s} \wedge n \quad \text{on } S. \tag{2.10}$$

Thus, the total electrical problem follows:

$$\begin{cases} curl\ curl\ \vec{E_t} - k^2\vec{E_t} = 0 & \text{in } \Omega_{int} \\ \vec{E_t} \wedge \vec{n} = 0 & \text{on } \Gamma \\ curl(\vec{E_t} \wedge n) + \mathcal{A}.E_t = g & \text{on } S. \end{cases} \tag{2.11}$$

In accordance with (2.2), the problem for the total field E_t has a unique solution. We repeat the Stokes formula for two vector fields A and B in Ω_{int}

$$\int_{\Omega_{int}} \langle curlA, B\rangle dx = \int_{\Omega_{in}} \langle curlB, A\rangle dx + \int_\Gamma \langle A \wedge n, B\rangle + \int_S \langle A \wedge n, B\rangle. \tag{2.12}$$

Using the boundary condition and the Helmholtz equation the weak formulation follows. From now on, we use $\vec{\varphi}$ for the Helmholtz solution and $\vec{\psi}$ for the function test

$$\int_{\Omega_{int}} curl\ \langle \varphi, curl\ \psi\rangle - k^2\langle \varphi, \psi\rangle = \int_\Gamma \langle curl\ \varphi \wedge n, \psi\rangle$$
$$+ \int_S \langle (g - \mathcal{A})(\varphi), \psi\rangle \quad \forall \psi \in H^1(\Omega_{int}). \tag{2.13}$$

3 Inverse problem

Now we focus on the inverse problem. The inverse problem is to determine the metallic antenna shape when we know an incident field E_i and the scattered field E_s in a given region θ. We found some work concerning the inverse problem solution in several books and articles ([11], [12], [13]) in several configurations (about incidencies and frequencies).

We consider the flow transformation Tr defined by:

$$Tr\ : \begin{cases} R^3 \longrightarrow R^3 \\ T_r(V)(x) = x + \int_0^r \vec{V}(s, T_s(V)(x))\, ds := x_r \\ \Omega_r(V) = T_r(V)(\Omega). \end{cases} \tag{3.14}$$

We define the cost functional in a fixed region θ for the moving domain $\Omega_{int\ r}$.

$$J(\Omega^r_{int}) = \int_\theta |E_m - \varphi|^2 d\gamma \tag{3.15}$$

where E_d is a given value that results in a scattered wave near an antenna's surface (Γ). (It can be used to optimize an infinity scattered diagram.) We introduce the following notation. The scalar product in C^3: $\langle a, b\rangle = \sum_i^3 a_i\overline{b_i} \in C$, which verifies $\langle a, a\rangle = \overline{\langle a, a\rangle} = |a|^2$ and $\langle \overline{a}, b\rangle = \overline{\langle a, \overline{b}\rangle}$.

Using the weak formulation, we define the following Lagrangian $\mathcal{L}$:

$$\mathcal{L}(r,\varphi,\psi) = \int_\theta |E_m - \varphi|^2 dx$$

$$+\Re\left(\int_{\Omega^r_{int}} \langle curl\ \psi,\ curl\ \varphi\rangle - k^2\langle\psi,\varphi\rangle\right) + \Re\left(\int_S \langle\mathcal{A}\varphi,\psi\rangle - \angle g,\psi\rangle\right) \quad (3.16)$$

with $(\varphi,\psi) \in F_r = \left\{\varphi \in H^1(\Omega^r_{int}); \varphi \wedge n_r = 0 \text{ on } \Gamma_r := T_r(V)(\Gamma)\right\}$.

We notice that the boundary S is not moved by Tr. In fact, the moving function Tr is zero when r is infinity. In order to bring back the problem on a min max over fixed functional spaces (not depending on r, we perform a change of functions. Using the min max derivative principle, we need to move $\mathcal{L}$ on a fixed domain Ω (not on Ω_r). Translating Ω, we need to transport the boundary condition $E_t \wedge n = 0$. In order to transport the boundary condition, we recall the Jacobian formula about the frame transformation. If $(X, \tau_1(x), \tau_2(x))$ is a base on Γ, $(X_r, DT_r(x).\tau_1(x), DT_r(x).\tau_2(x)))$ is a base on Γ_r in X_r, with the same properties. It follows that

$$\vec{n}(x_r) = \alpha \cdot (DT_r)^{-*}\vec{n}(x) \quad (3.17)$$

where α is a normalization coefficient. It holds that

$$\vec{n}\ o\ T_r(x) = \alpha \cdot (DT_r)^{-*}\vec{n}(x). \quad (3.18)$$

Inserting this into the boundary condition on S $\varphi \wedge n_r$, we have

$$\varphi \wedge ((DT^{-*}n)\ o\ T_r^-1) = 0, \quad (3.19)$$

which implies the function change:

$$\begin{aligned}\varphi &= DT^{-*}\ o\ T_r^{-1} \cdot u\ o\ T_r^{-1} \\ \psi &= DT^{-*}\ o\ T_r^{-1} \cdot p\ o\ T_r^{-1}\end{aligned} \quad (3.20)$$

Then we have a new Lagrangian expression for $(u,p) \in F = \{\varphi \in H^1(\Omega_{int});\ u \wedge n = 0 \text{ on } \Gamma\}$:

$$\begin{aligned} L(r,u,p) = &\int_\theta |E_m - (DT^{-*}oT^{-1}\cdot uoT^{-1})|^2 \\ &+\Re\int_{\Omega^r_{int}} \langle curl(DT^{-*}oT^{-1}\cdot uoT^{-1}), curl(DT^{-*}oT^{-1}\cdot poT^{-1})\rangle \\ &-\Re\int_{\Omega^r_{int}} k^2\ \langle(DT^{-*}oT^{-1}\cdot uoT^{-1}), (DT^{-*}oT^{-1}\cdot poT^{-1})\rangle \\ &-\Re\int_S \langle\mathcal{A}.(DT^{-*}oT^{-1}\cdot uoT^{-1}) - g, (DT^{-*}oT^{-1}\cdot poT^{-1})\rangle. \end{aligned} \quad (3.21)$$

In order to calculate the shape derivative, we recall the following properties:

$$\begin{aligned}&\forall p \in S^r, \forall r \in R \quad \mathcal{U} : u \to \mathcal{L}(r, u, p) \ \ is \ \ convex \\ &\forall u \in F, \forall r \in R \quad \mathcal{P} : p \to \mathcal{L}(r, u, p) \ \ is \ \ concave.\end{aligned} \tag{3.22}$$

Moreover,

$$\begin{aligned}&L_u(r, u^*, p) \cdot \delta u = 0 \ \ \forall p \in F \\ &L_p(r, u, p^*) \cdot \delta p = 0 \ \ \forall u \in F.\end{aligned} \tag{3.23}$$

Theorem 3.1

So (u^*, p^*) is the unique saddle point. The uniqueness is given by the direct and adjoint state unique solution. We can evaluate the shape derivative in r = 0.

$$L(r, u, p) = \begin{cases} R * S^r * R^r \to R \\ (r, u, p) \to L(r, u, p) \end{cases} \tag{3.24}$$

The existence of a unique saddle point (u^*, p^*) implies

$$\partial_r L(r, u^*, p^*)|_{r=0} = \min_{u \in F} \max_{p \in F} \partial_r L(r, u, p)|_{r=0}. \tag{3.25}$$

So we will differentiate L with respect to the parameter r. We recall the following:

$$\partial_r \int_{\Omega_{int}^r} F(r, x) dx|_{r=0} = \int_\Gamma F(0, x) v d\gamma + \int_{\Omega_{int}} (\partial_r F)|_{r=0} \ dx. \tag{3.26}$$

Thus, it holds that

$$\partial_r \varphi|_{r=0} = D^*(-v\vec{n}) \cdot u + D \cdot u(-v\vec{n}). \tag{3.27}$$

We will use this notation for the previous expression:

$$W(u, v) = D^*(-v\vec{n}) \cdot u + D \cdot u(-v\vec{n}). \tag{3.28}$$

Using complex scalar product properties:

$$\begin{aligned}\partial_r J(r, u, p) &= -\int_\theta (E_m - u).\overline{(W(u, v))} + \overline{(E_m - u)}.(W(u, v)) \ d\gamma \\ &= -\int_\theta 2\Re(E_m - u).\overline{(W(u, v)} d\gamma \\ &= -\int_\theta 2\Re(\overline{(E_m - u)}.(W(u, v)) d\gamma.\end{aligned} \tag{3.29}$$

Finally, we differentiate the Lagrangian as follows:

$$\begin{aligned}\partial_r L(r,u,p) = &-\sum_{i=1}^{N} 2\Re(\overline{(E_m(x_i)-u(x_i))}.(W(u(x_i),v(x_i)))\\ &+\Re\int_{\Omega_{int}}\langle(curlcurl(u),W(p,v)\rangle - k^2\langle(u),W(p,v)\rangle)|_{r=0}\\ &+\Re\int_{\Omega_{int}}\langle(curlcurl(p),W(u,v)\rangle - k^2\langle p,W(u,v)\rangle)|_{r=0}\\ &+\Re\int_{\Gamma} v\ \langle curl\ (u),\ curl\ (p)\langle -k^2\langle\ (u,p)\rangle\\ &+\Re\int_{\Gamma}\langle curlu\wedge n, W(p,v)\rangle|_{r=0} + \langle curlp\wedge n, W(u,v)\rangle|_{r=0}\\ &+\Re\int_{S}\langle curlu\wedge n, W(p,v)\rangle|_{r=0} + \langle curlp\wedge n, W(u,v)\rangle|_{r=0}\\ &+\Re\int_{S}\langle\mathcal{A}\cdot u\ - g, W(p,v)\rangle + \Re\int_{S}\langle\mathcal{A}W(u,v),p\rangle. \qquad (3.30)\end{aligned}$$

Using the complex $\langle,\rangle$ application, we have

$$\Re\int_{\mathcal{S}}\langle\mathcal{A}W(u,v),p\rangle = \Re\int_{\mathcal{S}}\langle\mathcal{A}p,W(u,v)\rangle. \qquad (3.31)$$

We focus on this part:

$$\begin{aligned}\int_{\Gamma}\langle curlu\wedge n, W(p,v)\rangle = &\int_{\Gamma}\langle curlu\wedge n, Dp\cdot(-v\vec{n})\rangle\\ &+\int_{\Gamma}\langle curlu\wedge n, D^*(-v\vec{n})\cdot p\rangle. \qquad (3.32)\end{aligned}$$

Moreover, we have

$$\int_{\Gamma}\langle curlu\wedge n, D^*(-v\vec{n})\cdot p\rangle = \int_{\Gamma}\langle curlu\wedge n, D^*_{\Gamma}(-v\vec{n})\cdot p\rangle \qquad (3.33)$$

We specify the following expression:

$$D^*_{\Gamma}(v\vec{n}) = (\vec{\nabla}_{\Gamma}(v\vec{n_1}),\vec{\nabla}_{\Gamma}(v\vec{n_2}),\vec{\nabla}_{\Gamma}(v\vec{n_3})), \qquad (3.34)$$

and

$$\int_{\Gamma}\vec{E}\cdot\nabla_{\Gamma}\varphi_i d\Gamma = -\int_{\Gamma}(div_{\Gamma}\vec{E})\varphi_i d\Gamma + \int_{\Gamma}H\varphi_i\vec{E}\cdot n, \qquad (3.35)$$

and such that

$$\begin{aligned}\int_{\Gamma}\langle curlu\wedge n, D^*_{\Gamma}(-v\vec{n})\cdot\vec{p}\rangle = &\int_{\Gamma} v\sum_{i=1}^{3} div_{\Gamma}((curlu\wedge n)p_i)n_i\\ &-\int_{\Gamma}H\sum_{i=1}^{3} vn_i\langle(curlu\wedge n)p_i,n\rangle. \qquad (3.36)\end{aligned}$$

The mixed product being zeros, it follows that

$$\int_\Gamma \langle curlu \wedge n, W(p,v)\rangle + \int_\Gamma \langle curlp \wedge n, W(u,v)\rangle$$

$$= \int_\Gamma \langle curlu \wedge n, Dp \cdot (-vn)\rangle + \int_\Gamma v \sum_{i=1}^{3} div_\Gamma((curlu \wedge n)p_i)n_i$$

$$\int_\Gamma \langle curlp \wedge n, Du \cdot (-vn)\rangle + \int_\Gamma v \sum_{i=1}^{3} div_\Gamma((curlp \wedge n)u_i)n_i, \qquad (3.37)$$

so the u^* field verifies the following problem:

$$\begin{cases} curlcurlu^* - k^2u^* = 0 & on\ \Omega_{int} \\ u^* \wedge n = 0 & on\ \Gamma \\ curlu^* \wedge n + \mathcal{A} \cdot u^* = g & on\ S \end{cases} \qquad (3.38)$$

and p^* is the unique solution ([10]) of the following restricted adjoint problem:

$$\begin{cases} curlcurl\ p^* - k^2p^* = -\sum_{i=1}^N 2(\overline{(E_m(x_i) - u(x_i))})\delta_{x_i} & on\ \Omega \\ p^* \wedge n = 0 & on\ \Gamma \\ curlp^* \wedge n + \mathcal{A}.p^* = 0 & on\ S. \end{cases} \qquad (3.39)$$

The solutions u^* and p^* are the restriction to Ω_r of the solution of the following problem:

$$\begin{cases} curlcurlu^* - k^2u^* = 0 & on\ \Omega \\ u^* \wedge n = 0 & on\ \Gamma \\ \lim_{r\to\infty} r(\partial_r \vec{u^*} + ik\vec{u^*}) = 0. \end{cases} \qquad (3.40)$$

The adjoint field p^* is the unique solution of the well-posed problem ([10]):

$$\begin{cases} curlcurl\ p^* - k^2p^* = -\sum_{i=1}^N 2(\overline{(E_m(x_i) - u^*(x_i))})\delta_{x_i} & on\ \Omega \\ p^* \wedge n = 0 & on\ \Gamma \\ \lim_{r\to\infty} r(\partial_r \vec{p^*} + ik\vec{p^*}) = 0. \end{cases} \qquad (3.41)$$

We know that E_d is a physical field and so it verifies the Maxwell equation. So $divE_d = divu = 0$, and $divp^* = 0$ on $(\Omega_{t_r} \cup S)$. We notice that is the knowledge of $(E_m - u^*)$ in the θ region that will be the incident data of the problem. So using the result of (3.37) and (u^*, p^*) being the solution of direct and adjoint problem, we have the Lagrangian derivative with respect to r, which takes the following form:

$$\begin{aligned}\partial_r L(r,u^*,p^*)|_{r=0} = \min_{u\in F}\max_{p\in F}\partial_r L(r,u,p)|_{r=0} &= \Re\int_\Gamma v(curl(u)curlp - k^2 u.p)\\ &+\Re\int_\Gamma v(div_\Gamma(curlu^*\wedge n)\langle p^*,n\rangle\\ &+div_\Gamma(curlp^*\wedge n)\langle u^*,n\rangle)\\ &+\Re\int_\Gamma 2v\langle curlu^*\wedge n, curlp^*\wedge n\rangle. \end{aligned}\qquad(3.42)$$

4 Parameterized geometries

4.1 Axisymmetrical geometries

Assume that the domain $\Omega\subset R^3$ is in the following form:

$$\Omega=\{(x,y,z)\in R^3 \ s.t.\ 0<z<Z^*, (x,y)=r(cos\theta, sin\theta), \text{ with}\\ 0<r<R(z)\}.$$

Then

$$\partial\Omega = S_0\cup\Gamma\cup S_{Z^*}$$

where

$$\Gamma=\{(x,y,z)\in R^3,\ s.t.\ 0<z<Z^*, (x,y)=R(z)(cos(\theta), sin(\theta)),\\ 0<\theta\le 2\pi\}.$$

We consider transformations T_t of R^3, which preserve that structure for the deformed domain

$$\Omega_t := T_t(\Omega)=\{(x,y,z)\in R^3 \ s.t.\ 0<z<Z^*,\\ (x,y)=r(cos(\theta),sin(\theta)),\quad 0<r<R_t(z)\}$$

where R_t is the graph of the transformed axisymmetrical domain. It is necessary that the speed vector field be chosen in the following *horizontal, radial* form:

$$V(t,x,y,z)=w\Big(t,\sqrt{x^2+y^2},z\Big)\frac{(x,y,0)}{\sqrt{x^2+y^2}}.$$

One vector field in R^3 tangent to the boundary Γ is

$$\tau(x,y,z)=\frac{1}{\sqrt{1+(R'(z))^2}}\left(R'(z)\frac{(x,y)}{r},1\right)\quad\text{where } r=\sqrt{x^2+y^2}.$$

The normal field in R^3 to the boundary Γ is then

$$n(x,y,z)=\frac{1}{\sqrt{1+(R'(z))^2}}\left(\frac{(x,y)}{r},-R'(z)\right).$$

We consider the normal speed

$$v(t, z, R_t(z)) := \langle V(t, x, y, z), n_\Gamma(x, y, z)\rangle_{R^3} \quad \text{on } \Gamma_t = T_t(\Gamma)$$

as the *control* parameter, while $v(t, 0, R_t(0)) = v(t, Z^*, R_t(Z^*)) = 0$, so that the bottom and top surfaces S_0 and S_{Z^*} remain invariants under the T_t transformation of R^3 (by setting: $w(t, 0, r) = w(t, Z^*, r) = 0$).

We get

$$\langle V(0), n\rangle = \frac{w(t, \sqrt{x^2 + y^2}, z)}{\sqrt{1 + (R'(z))^2}}.$$

4.1.1 Eulerian shape derivative of the functional under axisymmetrical geometries

The lateral boundary integration can be written as

$$\int_\Gamma f(x, y, z)\ d\Gamma(x, y, z) = \int_0^{2\pi}\int_0^{Z^*} f(R(z)(cos(\theta), sin(\theta)), z) \sqrt{1 + (R'(z))^2}\ d\theta\ dz.$$

Following the structure theorem for the derivative, there exists an (axisymmetrical) density measure on the lateral boundary $\Gamma = \partial\Omega$ such that

$$\begin{aligned} dJ(\Omega, V) &= \int_{\partial\Omega} gV(0) \cdot n\ d\Gamma \\ &= 2\pi \int_0^{Z^*} g(R(z), z)(w(0, R(z), z)\ dz. \end{aligned}$$

4.1.2 Derivative with respect to shape parameters

Assume now that the graph $\Gamma_R = \{(z, R(z)), 0 < z < Z^*\}$ is parametrized by some $\alpha \in R$ as follows: $R = R(\alpha; z)$. Then we understand that the derivative $\frac{\partial}{\partial\alpha}R(\alpha; z)$ will enter in the expression of the parameter derivative of the shape functional $\frac{\partial}{\partial\alpha}j(\Omega_\alpha)$. Indeed, we consider the transformation

$$\mathcal{T}_\alpha(X, Y, z) := (X, Y, z) \longrightarrow \left(\frac{R(\alpha; z)}{R(0, z)}(X, Y), z\right)$$

maps the graph Γ_R onto the graph $\Gamma_{R(\alpha;.)}$.

And we classically consider the speed vector, at $\alpha = 0$ (using the fact that $\mathcal{T}_0 = I_d$):

$$V(0; X, Y, z) = \frac{\partial}{\partial\alpha}\mathcal{T}_\alpha(X, Y, z) = \left(\frac{\frac{\partial}{\partial\alpha}R(0; z)}{R(0, z)}(X, Y), 0\right).$$

That is,

$$V(0;X,Y,z) = w(0,\sqrt{X^2+Y^2},z)\left(\frac{(X,Y)}{\sqrt{X^2+Y^2}},z\right)$$

with

$$w(0,\sqrt{X^2+Y^2},z) = \frac{\frac{\partial}{\partial\alpha}R(0;z)}{R(0,z)}\sqrt{X^2+Y^2}$$

In the previous integral we get:

$$\frac{\partial}{\partial\alpha}J(\Omega_\alpha)|_{\alpha=0} = dj(\Omega,\mathcal{V}(0)) = \int_0^{Z^*} g(0,z)\frac{\partial}{\partial\alpha}R(0,z)\ dz.$$

4.1.3 Example of the Bezier polynomial boundary

Assume, with $s = z/Z^*$, $m+1$ *control points* $Y_k(\alpha)$ are given and consider

$$R(\alpha;z) := \Sigma_{k=0,\ldots m}C_m^k s^k(1-s)^{(m-k)}Y_k(\alpha).$$

Then obviously we get

$$\frac{\partial}{\partial\alpha}R(0;z) = \Sigma_{k=0,\ldots m}C_m^k s^k(1-s)^{(m-k)}\frac{\partial}{\partial\alpha}Y_k(\alpha)|_{\alpha=0}.$$

4.2 Expression of the shape gradient for geometry depending on a finite number of parameters

Let us consider as a first example (used in the following validation) the case where $\Omega_r = R^3\backslash B(0,r)$. The idea is to introduce a transformation of R^3 that follows the shape evolution of the domain. In that case let us consider

$$T_r(x,y,z) := r(x,y,z)$$

Then obviously we have $T_r(\Omega_1) = \Omega_r$, so that introducing $j(r) := J(\Omega_r)$ we get $J'(r) = dj(\Omega_r, V^*(r))$ where the *specific speed vector* V^* is given by

$$V^*(r,x,y,z) := \left(\frac{\partial}{\partial r}T_r\right)o(T_r)^{-1}(x,y,z) \qquad (4.43)$$

So that $V^*(r,x,y,z) = \frac{1}{r}(x,y,z)$. The normal field on $\partial\Omega_r$ is given by $n_r(x,y,z) = \frac{1}{\sqrt{x^2+y^2+z^2}}(x,y,z)$, so that the normal speed is $\langle V(r), n_r\rangle = \frac{1}{r}$. Then we get:

$$j'(r) = \int_{\partial\Omega_r} g\langle V(r),n_r\rangle d\Gamma_r = \int_0^\pi d\phi\int_0^{2\pi} g(rcos(\theta)sin(\phi),$$
$$rsin(\theta)sin(\phi), rcos(\theta)sin(\phi))\ \frac{1}{r}\ r^2 sin(\phi)d\theta.$$

5 Singular case

We deal with a closed line singularity in the antenna. Ω is a bounded open in $\mathcal{R}^3$. $\Gamma = \partial\Omega$, and its boundary is a piecewise manifold of class C^2. We suppose it exists on a line γ lying in Γ, such that Γ is decomposed as follows:

$$\Gamma = \Gamma_1 \cup \Gamma_2 \cup \gamma \tag{5.44}$$

where Γ_1 and Γ_2 as smooth open sets in Γ and $\gamma = \bar{\Gamma_1} \cap \bar{\Gamma_2}$ is parameterized by:

$$\lambda \in C^2([0,1], \mathcal{R}^3) \tag{5.45}$$

Verifying:

$$\lambda(0) = \lambda(1), \quad |\lambda'(s)|\rangle 0 \quad and \quad \lambda(s) \in \gamma \tag{5.46}$$

for all $x \in \Gamma \backslash \gamma$ we denote $T_x\Gamma$, the tangent space, and $\vec{n(x)}$, the unitary normal field to $T_x\Gamma$ outgoing to Ω, at $x \in \Gamma$. At $x \in \gamma$, we assume the existence of two tangent spaces $T_x\Gamma_1$ and $T_x\Gamma_2$, respectively, to Γ and its associated normal fields $\vec{n_1}(x)$ and $\vec{n_2}(x)$, on both sides of the singularity line. For $x \in \gamma$ we denote $\tau(x)$, a unitary field that is tangent to γ, and $\vec{\nu_1}(x)$ and $\vec{\nu_2}(x)$, two unitary fields as follows:

$$\begin{cases} \vec{\nu_1}(x) \in T_x\Gamma_1 \\ \vec{\nu_2}(x) \in T_x\Gamma_2 \\ \vec{\nu_1}(x) \cdot \vec{\tau}(x) = \vec{\nu_2}(x) \cdot \vec{\tau}(x) = 0 \\ \vec{\nu_1}(x) \cdot \vec{n_1}(x) = \vec{\nu_2}(x) \cdot \vec{n_2}(x) = 0. \end{cases} \tag{5.47}$$

Let $\vec{E}$ and φ be defined on $\Gamma\backslash\gamma$ with respective boundaries E_1 and E_2, $(n_1, \varphi_1 \ and \ n_2, \varphi_2)$ of each part of γ

$$\begin{aligned} \int_\Gamma E.\nabla_\Gamma \varphi_i \, d\Gamma = &- \int_{\Gamma_1} (div_{\Gamma_1} E_1)\varphi_1 d\Gamma_1 - \int_{\Gamma_2} (div_{\Gamma_2} E_2)\varphi_2 d\Gamma_2 \\ &+ \int_{\Gamma_1} H\varphi_1 E_1.n_1 \\ &+ \int_{\Gamma_2} H_2\varphi_2 E_2.n_2 + \int_\gamma \varphi_1 \vec{E_1} \cdot \vec{\nu_1} + \varphi_2 \vec{E_2} \cdot \vec{\nu_2} d\gamma. \end{aligned} \tag{5.48}$$

We take two *parallel* curves γ_1^ε and γ_2^ε to γ, with $\gamma_k^\varepsilon \in \Gamma_k$, k = 1,2 and we decompose $\Gamma = \Gamma_1^\varepsilon \cup \Gamma_2^\varepsilon \cup b^\varepsilon$ with $\partial b^\varepsilon = \gamma_1^\varepsilon \cup \gamma_2^\varepsilon$.

$$\begin{gathered} For \ k = 1,2 \quad \int_\Gamma \langle curlu \wedge n, D^*(-V) \cdot u \rangle = \\ \int_{\Gamma_k^\varepsilon \cap b^\varepsilon} \sum_i^{1,3} V_i \nabla_\Gamma.[(curlu \wedge n)p_i] - \int_{\Gamma_k^\varepsilon} H \sum_i^{1,3} V_i \langle (curlu \wedge n)p_i, n \rangle \\ - \sum_i^{1,3} \int_{\gamma_k^\varepsilon} \langle (curlu \wedge n_1)p_i, \nu_1 \rangle V_i. \end{gathered} \tag{5.49}$$

Notice that $\langle (curlu \wedge n)p, n\rangle = 0,\ f = \langle curlu \wedge n, D^*(-V)\cdot u\rangle \in L^1(\Gamma)$

$$\int_\Gamma f = \sum_{i=1}^{2}\int_{\Gamma_i^\varepsilon} f + \int_{b^\varepsilon} f, \quad \int_{\gamma_i^\varepsilon} g^i d\gamma = \int_\gamma (g^i o\lambda_\varepsilon^i) D_\varepsilon^i \tag{5.50}$$

with $\lambda_\varepsilon^i : \gamma \to \gamma_i^\varepsilon$,

$$\int_\Gamma f - \int_{\Gamma_k^\varepsilon \cap b^\varepsilon} \sum_i^{1,3} V_i \nabla_\Gamma .[(curlu \wedge n)p_i] =$$

$$-\sum_i^{1,3}\int_{\gamma_k^\varepsilon} \langle (curlu \wedge n_1)p_i, \nu_1\rangle o\lambda_\varepsilon^i . V_i o\lambda_\varepsilon^i . D_\varepsilon^i d\gamma. \tag{5.51}$$

With $\varepsilon \to 0$, $D_\varepsilon^i \to 1$ and $\lambda_\varepsilon^i \to I$:

$$\int_\Gamma f - lim_{\varepsilon\to 0}\int_{\Gamma_k^\varepsilon \cap b^\varepsilon} \sum_i^{1,3} V_i \nabla_\Gamma .[(curlu \wedge n)p_i]$$

$$= -\sum_i^{1,3}\int_{\gamma_k^\varepsilon} \langle (curlu \wedge n_1)p_i, \nu_1\rangle o\lambda_\varepsilon^i \cdot V_i o\lambda_\varepsilon^i \cdot D_\varepsilon^i d\gamma. \tag{5.52}$$

The limits exist, then:

$$\int_\Gamma f - \int_\Gamma \sum_i^{1,3} V_i \nabla_\Gamma \cdot [(curlu \wedge n)p_i]$$

$$= -\sum_i^{1,3}\int_\gamma [\langle (curlu \wedge n_1)p_i, \nu_1\rangle + \langle (curlu \wedge n_2)p_i, \nu_2\rangle]\cdot V_i d\gamma \tag{5.53}$$

and

$$\int_\Gamma \langle curlu \wedge n, W(p,V)\rangle|_{r=0}$$

$$= \int_\Gamma \sum_i^{1,3} V_i \nabla_\Gamma \cdot [(curlu \wedge n)p_i] + \int_\Gamma \langle curlu \wedge n, D_\Gamma .(-V)\rangle$$

$$-\sum_i^{1,3}\int_\gamma [\langle (curlu \wedge n_1)p_i, \nu_1\rangle + \langle (curlu \wedge n_2)p_i, \nu_2\rangle].V_i d\gamma \tag{5.54}$$

$$\sum_i^{1,3}[\nabla_\Gamma .((curlu \wedge n)p_i)]V_i$$

$$= \nabla_\Gamma (curlu \wedge n)\langle p \cdot V\rangle + \langle curlu \wedge n, D_\Gamma^* p \cdot V\rangle. \tag{5.55}$$

We have $\vec{p}$, normal to the surface: $\langle p, V\rangle = v\langle p, n\rangle$ and so

$$\begin{aligned}\int_\Gamma \langle curlu \wedge n, W(p,V)\rangle|_{r=0} = &\int_\Gamma \nabla_\Gamma(curlu \wedge n)\langle p\cdot V\rangle\\ &+\int_\Gamma \langle curlu \wedge n, D^*_\Gamma p\cdot V\rangle\\ &+\int_\Gamma \langle curlu \wedge n, D_\Gamma p\cdot(-V)\rangle\\ &-\sum_i^{1,3}\int_\gamma [\langle (curlu \wedge n_1)p_i, \nu_1\rangle\\ &+\langle (curlu \wedge n_2)p_i, \nu_2\rangle]\cdot V_i d\gamma. \qquad (5.56)\end{aligned}$$

Introducing: $(D^*u - Du)\cdot V = curlu \wedge V$

$$\langle curlu \wedge n, D^*_\Gamma p\cdot V\rangle + \langle curlu \wedge n, D_\Gamma p\cdot(-V)\rangle = \langle curlu \wedge n, curlp \wedge V\rangle \qquad (5.57)$$

$\vec{u}$ and $\vec{p}$ are normals to Γ, so we have

$$\langle curlu \wedge n, curlp \wedge V\rangle = \langle curlu \wedge n, curlp \wedge n\rangle v. \qquad (5.58)$$

Applying this we have

$$dJ(\Omega, V) = \Re\int_\Gamma g\langle V, n\rangle + \Re\int_\gamma \langle \mathcal{H}, V\rangle \qquad (5.59)$$

with

$$\begin{aligned}g = &(\langle curlu.curlp\rangle - \langle k^2 u.p\rangle) + \nabla_\Gamma.(curlu \wedge n)\langle p, n\rangle\\ &+\nabla_\Gamma.(curlp \wedge n)\langle u, n\rangle + 2\langle curlu \wedge n, curlp \wedge n\rangle \qquad (5.60)\end{aligned}$$

and

$$\begin{aligned}\langle \mathcal{H}, V\rangle = &-\{\langle (curlu \wedge n_1), \nu_1\rangle + \langle (curlu \wedge n_2), \nu_2\rangle\}\langle p, V\rangle\\ &-\{\langle (curlp \wedge n_1), \nu_1\rangle + \langle (curlp \wedge n_2), \nu_2\rangle\}\langle u, V\rangle. \qquad (5.61)\end{aligned}$$

6 Numerical results

In this section we will validate the shape derivative with numerical examples using the finite differences method.

We use S.R.3D. software to solve the 3D harmonic Maxwell equation in isotropic and homogenous medium ($\varepsilon, \mu,$ and $\sigma \in \mathcal{R}$). This software, based on Rumsey's integral equations allows us to identify scattered electrical and magnetical fields and their tangential part on the scattering surface (the equivalent currents). The adjoint problem being similar to a direct state, we also use SR3D (with little modification, incidences are complex vectors) to solve it.

We choose two transformations. The first is a dilatation of a sphere where the radius r is the parameter of deformation, so we have $\vec{V}\cdot\vec{n} = 1$. The

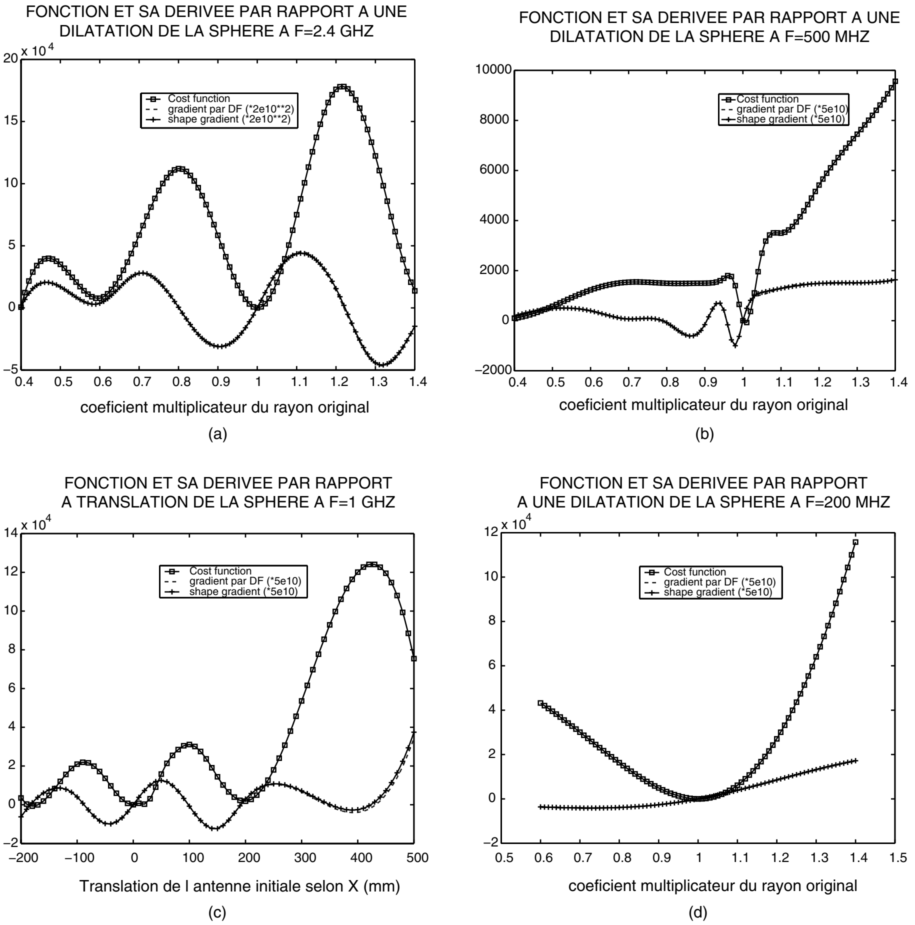


FIGURE 8.1

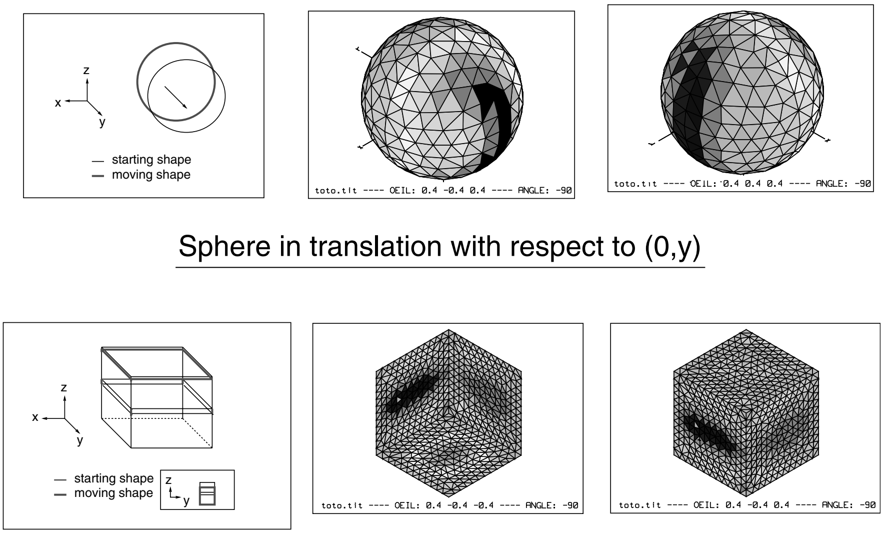
second is a translation of a unitary vector e_x ($\vec{V} \cdot \vec{n}$ is equal to the first component of vector $\vec{n}$).

In the following case, the E_d field is a scattered field near a sphere (radius = 170 mm, centered on 0). For the dilatation case, we compute the average shape derivative for $r = k * 170; 0.4 \leq k \leq 1.4$. The ideal radius is given by k = 1.

On the third figure, the ideal translation is given by k = 0. We compute the shape gradient for a translation that is between −200 mm and 500 mm.

The following examples take place on different frequencies, and different configurations of incidences and receptors.

When we go down in frequency (Figure 8.1a, b and d), the wavelength and the attractive wave of the cost function increase. In Figure 8.1c, (f = 500 MHz) when $k = 1.5$, we are still on the good *attraction curve.*


z
x
y
starting shape
moving shape
toto.tit ---- OEIL: 0.4 -0.4 0.4 ---- ANGLE: -90
toto.tit ---- OEIL: -0.4 0.4 0.4 ---- ANGLE: -90
z
x
y
starting shape
moving shape
z
y
toto.tit ---- OEIL: 0.4 -0.4 -0.4 ---- ANGLE: -90
toto.tit ---- OEIL: 0.4 -0.4 0.4 ---- ANGLE: -90

Sphere in translation with respect to (0,y)

A box we have cut in two and suppressed a part

FIGURE 8.2

Conversely, the wavelength decreases when frequencies increase. Consequently, there is a small attractive wave.

6.1 Visualization of shape derivative in 3D

In this section, we present a 3D visualization of the shape gradient results on the structure. In the following cases, E_d is given by the ideal structure scattering. Then we apply a small deformation. The incident field is created by 6 dipoles, which are tangential to the initial structure. See Figure 8.2 for the results. Yellow shading is for positive densities, green shading for zero, and blue shading for negative gradient densities.

In the three previous cases, the shape gradient is inverse to the applied deformation. We have taken examples in the frequency range where the moved structure boundary is still the attraction wave.

In the following, E_d is scattered by a box we suppressed, and we compute the shape gradient on a complete cube.

In this example, computed at increasing frequencies, the ideal frequencies appear for detecting the ideal structures. We can talk about ideal frequencies or a range of ideal frequencies to compute. If the frequency is bigger, the optimized structure is probably on the wrong attraction wave, and if the frequency is smaller, the length of the wave is bigger for the structure that doesn't scatter any field or scatter a numerical noise. Then the frequency must be tuned properly.

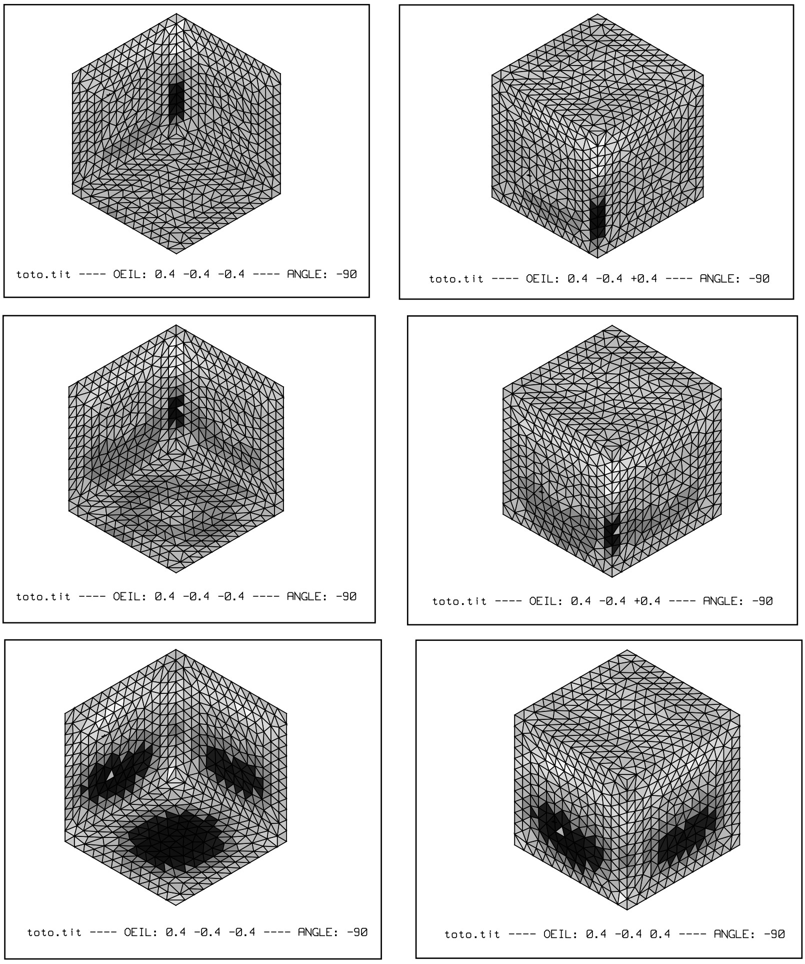

toto.tit ---- OEIL: 0.4 -0.4 -0.4 ---- ANGLE: -90
toto.tit ---- OEIL: 0.4 -0.4 +0.4 ---- ANGLE: -90
toto.tit ---- OEIL: 0.4 -0.4 -0.4 ---- ANGLE: -90
toto.tit ---- OEIL: 0.4 -0.4 +0.4 ---- ANGLE: -90
toto.tit ---- OEIL: 0.4 -0.4 -0.4 ---- ANGLE: -90
toto.tit ---- OEIL: 0.4 -0.4 0.4 ---- ANGLE: -90

FIGURE 8.3

7 Conclusion

In this paper we have calculated the shape derivative for a 3D Maxwell problem for metallic structures in harmonic regime with a min-max formulation. We have computed this shape derivative in 3D, where the incidence, the receptors(θ) are an unconstrained 3D electrical field, and where the antenna's shape is defined in 3D using and modifying the France Telecom software S.R.3D.

We have validated theoretical calculus by comparison with finite differences of direct state.

We have also discussed the importance of frequencies jump in reconstruction. This information is fundamental in order to use shape derivatives in the optimization method.

Finally, we presented the shape gradient calculus for structures that present *geometrical line* singularity.

References

[1] M. Moubachir and J.-P. Zolésio, *Moving Shape Analysis and Control*, Pure and Applied Math. 277, Chapman & Hall, Boca Raton, 2006.

[2] O. Dorn, H. Bertete-Aguirre, J.G. Berryman, and G.C. Papanicolaou, A Nonlinear Inversion Method for 3D. Electromagnetic Imaging Using Adjoint Fields N, to appear in *Inverse Problems.*

[3] J.-C. Nedelec, *Acoustic and Electromagnetic Equations: Integral Representation for Harmonic Problems*, Springer, New York, 2001.

[4] R. Dautray and J.L. Lions, *Analyse Mathématique et calcul numérique pour les sciences et les techniques*, Edition Masson, Paris, 1984.

[5] J.A. Kong, *Research Topics in Electromagnetic Wave Theory*, John Wiley & Sons, New York, 1981.

[6] J.-P. Zolésio and M.C. Delfour, *Shapes and Geometries: Analysis, Differential Calculus, and Optimisation*, Edition SIAM, Philadelphia, 1987.

[7] P.F. Combes, *Micro-ondes: Circuits passifs, Antennes et Propagations*, Edition Dunod, Paris, 1997.

[8] J.-P. Zolésio, *Weak Shape Formulation of Free Boundary Problem*, Scuola Normale Superiore Pise, Series IV, Vol. XXI, 1, 1994.

[9] P. Ratajczak, Applications du formalisme rigoureux des équations intégrales à létude de structure 3D.inhomogènes exitées par modes guidés, Doctorat de l'université de Nice Sophia-Antipolis Spécialité Electroniques, 1995.

[10] F. Millo, Conception optimale de structures rayonnantes, Doctorat de l'université de Nice Sophia-Antipolis, Spécialité Mathématiques, 1991.

[11] D. Colton and R. Kress, *Integral Equation Method in Scattering Theory*, John Wiley & Sons, Inc., New York, 1983.

[12] V. Isakov, *On Uniqueness in the Inverse Transmission Scattering Problem*, Commun. in Partial Differential Equations, 1567–1587, Marcel Dekker, New York, 1990.

[13] F. Hettlich, On the Uniqueness of the Inverse Conductive Scattering Problem for the Helmholtz Equation, *Inv. Prob.* 10, 129–144, 1994.

[14] J. Cagnol, M. P. Polis, and J.-P. Zolésio, *Shape Optimisation and Optimal Design*, Lecture Notes in Pure and Applied Mathematics, Marcel Dekker, New York, 2001.

CHAPTER 9

Array antenna optimization

Louis Blanchard
INRIA, Sophia-Antipolis, France

Jean-Paul Zolésio
CNRS and INRIA, Sophia-Antipolis, France

1 Introduction

An array antenna consists of a large family of *elementary antennas* located on a smooth surface in $\mathbb{R}^3$. The design problem consists of finding the best location of these sources on that surface in order to recover the same behavior as found in classical antennas. The objective functions are of two kinds: the feasible antenna should cover a precise area and the outside of this area should be minimized. The global synthesis of planar antenna arrays consists of optimizing, at the same time, positions and weights of the array elements in order to generate a desired radiation pattern. It is a highly nonlinear optimization problem.

Moreover, each source position is associated with a classical antenna resolution through the optimal source alimentation (*source weight*); indeed the main point is that this optimality should depend on the design (*positioning*) array parameter, say $r \in (\mathbb{R}^2)^N$, if the antenna is assumed to be planar as will be the case in the following experiments. However, it turns out that the sensibility of the optimal weights on the geometrical parameter r is numerically impossible when the source number N becomes very large. Consequently, we introduce a new model that completely bypasses this analysis, and present numerical validations taking advantage of the functional derivative expressed in terms of min or max; see [1] and [3–5].

In the literature, analytical methods have been applied primarily to solve the weight optimization problem. With respect to these methods, the majority use an error minimization between desired and synthesized beam patterns by *minimax* or *least squares* techniques. However, analytical methods are generally unable to optimize both positions and weights, and therefore only stochastic methods have been proposed for its solution.

This paper proposes a global method, based on a bi-criterion optimization method, which is able to synthesize a planar array with an optimization of position and weighting, in order to reduce the side lobes' peak and possibly grating lobes, according to different constraints, such as, for example, the minimal distance between the array elements or the width of the main lobe.

2 Mathematical formulation

We assume a radiation pattern of the antenna in the far field. Let us consider the planar array antennas shown in Figure 9.1, whose N elements are located on the (x, y) plane at $r_k = (x_k, y_k), \forall k \in [1..N]$. The beam pattern function of the array $d(r, \omega, u_0)$ is defined as follows

$$d(r, \omega, u_0) = \sum_{k=1}^{N} \omega_k g_k(u_0) e^{j\frac{2\pi}{\lambda}\langle u_0 . r_k\rangle_{\mathbb{R}^3}}$$

where

- λ is the background wavelength.
- ω_k is the weight coefficient of element k, with $\omega_k \in \mathbb{C}$ (i.e., amplitude and phase).
- $g_k(u_0)$ is the radiation pattern of element k.
- u_0 gives the angular position of the field point, with $u_0 \in S^2$ the unit sphere in $\mathbb{R}^3$.

By definition, the *directive gain* of an antenna in a given direction u_0 is the radiation pattern normalized by the corresponding *isotropic radiation*

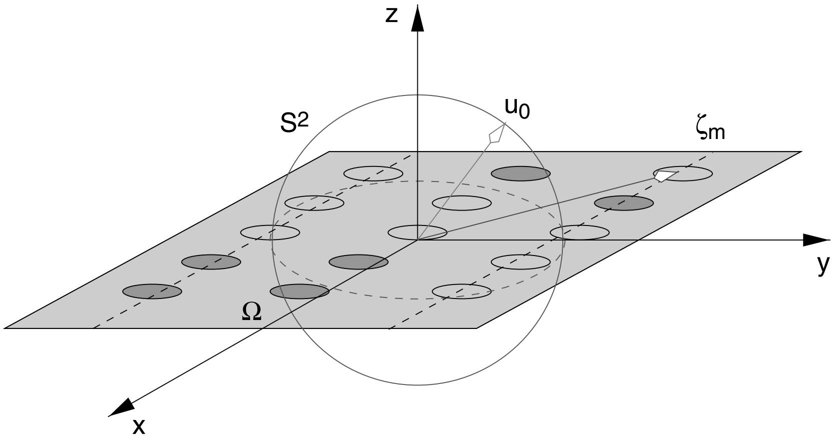


FIGURE 9.1 Planar array antenna Ω.

intensity:

$$\boxed{D(r,\omega,u_0) = |d(r,\omega,u_0)|^2 \left(\frac{1}{4\pi} \int_{S^2} |d(r,\omega,u)|^2 ds(u) \right)^{-1}}$$

2.1 The bi-criterion optimization

The design problem is to optimize geometrical positions (r) and weights (ω) of the array elements (at the same time) in order to satisfy the following two criteria:

- focusing the *directive gain* toward a desired angular sector $\mathcal{V}_\lambda(u_p)$ centered at $u_p \in S^2$. Therefore, to *maximize* the directive gain in the main lobe, we define the *fitness function*:

$$\mathcal{E}_{in}(r,\omega) = ||D(r,\omega)||^{-1}_{L^\infty(\mathcal{V}_\lambda(u_p))}$$

- decreasing the *directive gain* elsewhere, that is, in the angular sector $S^2 \backslash \mathcal{V}_\lambda(u_p)$. Therefore, to *minimize* the directive gain in the side lobes and in the grating lobes, we define the *fitness function*:

$$\mathcal{E}_{out}(r,\omega) = ||D(r,\omega)||_{L^\infty(S^2 \backslash \mathcal{V}_\lambda(u_p))}$$

Let us consider $\mathcal{W}$, a constraint manifold of the weighting vector ω. An approach to realize this bi-criterion optimization consists of finding a *Pareto* optimal vector $(r^*,\omega^*) \in \mathbb{R}^{2M} \times \mathcal{W}$, where (r^*) is an optimal geometry and (ω^*) its optimal weight, satisfying the proposition below.

Proposition 2.1

A vector $(r^*,\omega^*) \in \mathbb{R}^{2M} \times \mathcal{W}$ is Pareto optimal if and only if there exists no vector $(r,\omega) \in \mathbb{R}^{2M} \times \mathcal{W}$ such that

$$\mathcal{E}_{in}(r,\omega) \leq \mathcal{E}_{in}(r^*,\omega^*) \quad \textit{and} \quad \mathcal{E}_{out}(r,\omega) \leq \mathcal{E}_{out}(r^*,\omega^*) \tag{2.1}$$

where at least one of them is strict.

In fact, we consider the *Pareto* function for finding a stability point that simultaneously minimizes antagonistic criteria $\mathcal{E}_{out}(r,\omega)$ and $\mathcal{E}_{in}(r,\omega)$:

$$J(r,\omega) = \mathcal{E}_{in}(r,\omega)\mathcal{E}_{out}(r,\omega)$$

The optimization problem is

$$\boxed{(\mathcal{P}) \min_{(r,\omega)\in\mathbb{R}^{2M}\times\mathcal{W}} J(r,\omega) = \min_{r\in\mathbb{R}^{2M}} E(r)} \tag{2.2}$$

where $E(r)$ is the minimum functional value resulting from the problem:

$$(\mathcal{P}_1) \min_{\omega \in \mathcal{W}} J(r, \omega) \tag{2.3}$$

Note that for a fixed geometry of the antenna (r), the weight vector depends on this geometry, that is, $\omega(r)$: in order to solve the problem $(\mathcal{P})$, we need to compute the chain rule equation:

$$\frac{d}{dr} J(r, \omega(\zeta)) = \nabla_r J(r, \omega(r)) + \nabla_\omega J(r, \omega(r)). \frac{\partial}{\partial r} \omega(r)$$

However, by definition $\omega(r)$ is the optimal weighting, that is, $\nabla_\omega J(r, \omega(r)) = 0$ and the difficulty that exists in calculating $\frac{\partial}{\partial r}\omega(r)$ is avoided.

Numerically, the calculation of $\omega(r)$, the optimal weight, is done by a *Newton-Raphson* under a constraint method preceded by a projected conjugate gradient method (see, for example, Equation [2.3]).

2.2 Tangential Newton's method

The goal of this *Newton-Raphson* method is of course to find the optimal vector $\omega(r)$, the solution of problem $(\mathcal{P}_1)$ (Equation 2.3), but especially to solve the vector equation $\nabla_\omega J(r, \omega(r)) = 0$. Consequently we calculate the tangential gradient of the fitness functional $J(\omega)$:

$$F(\omega) := \nabla_{\mathcal{W}} J(\omega) = \Pi_\omega \nabla J(\omega)$$

where $J(\omega)$ is a shorthand notation for $J(r, \omega)$ and Π_ω is the orthogonal projector onto the orthogonal complement of $\vec{n}$, the normal vector of $\mathcal{W}$ on ω point: $\Pi_\omega = (Id - \vec{n} \otimes \vec{n})$. A necessary condition for solving the problem is to solve the vector equation

$$F(\omega) := \nabla_{\mathcal{W}} J(\omega) = 0. \tag{2.4}$$

This formulation of a zero finding problem uses the Newton iteration on the manifold $\mathcal{W}$, which consists of solving the Newton equation

$$F(\omega) + D_{\mathcal{W}} F(\omega) \Delta = 0 \tag{2.5}$$

where $D_{\mathcal{W}} F(\omega)\Delta$ denotes the directional tangential derivative of F at ω in the direction of Δ, which by definition is $D_{\mathcal{W}} F(\omega) := DF(\omega)\Pi_\omega$, and performing the update

$$\omega_+ = \omega + \Delta.$$

2.3 Tangential Newton's method on the unit sphere S in $\mathbb{C}^N$

We consider the unit sphere manifold $S := \{y \in \mathbb{C}^N \|y\|_{\mathbb{C}^N} = 1\}$. The orthogonal projector onto the tangent space at ω is:

$$\Pi_\omega = (Id_{2M} - \omega \otimes \omega).$$

We solve vector Equation (2.4) using the Newton-Raphson algorithm. However, the solutions of (2.4) are not isolated in $\mathbb{C}^N$; namely, if ω is a solution, then all the elements of the equivalence class $\{\lambda\omega \mid \forall\lambda \in \mathbb{C}\}$ are solutions, too. In fact, because F is homogeneous of degree one, that is, $\forall w \in \mathbb{C}^N \forall\lambda \in \mathbb{C} F(\lambda\omega) = \lambda F(\omega)$, the solution of the Newton equation (2.5), when unique, is $\Delta = -\omega$. So any point ω is mapped to $\omega_+ = 0$. This is clearly a solution of $F(\omega) = 0$, but it is a trivial solution. A remedy consists of constraining Δ to belong to the *horizontal space* orthogonal to ω, that is, $(\Delta \in H_\omega := \{y \in \mathbb{C}^N | y^t\omega = 0\})$. With this constraint on Δ, the solution Δ of $F(\omega + \Delta) = 0$ becomes isolated. Moreover, the Newton equation (2.5) has in general no solution in H_ω, so the Newton equation must be relaxed. That is why we project the Newton equation onto H_ω.

$$\begin{cases} \Pi_\omega(F(\omega) + D_S F(\omega)\Delta) = 0 \\ \omega^t\Delta = 0 \end{cases} \tag{2.6}$$

2.3.1 Practical implementation

The first equation of the system (2.6) $\Pi_\omega(F(\omega) + D_S F(\omega)\Delta) = 0$ can be rewritten

$$\exists\delta \in \mathbb{R} \text{ such as } F(\omega) + D_S F(\omega)\Delta = \delta\omega$$

and by definition $D_S F(\omega)\Delta = DF(\omega)\Pi_\omega\Delta = DF(\omega)\Delta - DF(\omega)\omega\omega^t\Delta$. The system (2.6) can be rewritten $\begin{bmatrix} DF(\omega) & -\omega \\ \omega^t & 0 \end{bmatrix} \begin{bmatrix} \Delta \\ \delta \end{bmatrix} = \begin{bmatrix} -F(\omega) \\ 0 \end{bmatrix}$. Moreover, if we suppose that the *Hessian* matrix $DF(\omega)$ is invertible and according to the homogeneity of degree one of the functional $F(\omega)$, this implies that $DF(\omega)\omega + F(\omega) = 0$, and we have the relation

$$\Delta = \delta[DF(\omega)]^{-1}\omega - [DF(\omega)]^{-1}F(\omega) = \delta[DF(\omega)]^{-1}\omega + \omega$$

and with the condition $\omega^t\Delta = 0$ we have

$$\delta = \frac{-1}{\omega^t[DF(\omega)]^{-1}\omega}\ , \qquad \Delta = \frac{-1}{\omega^t[DF(\omega)]^{-1}\omega}[DF(\omega)]^{-1}\omega + \omega$$

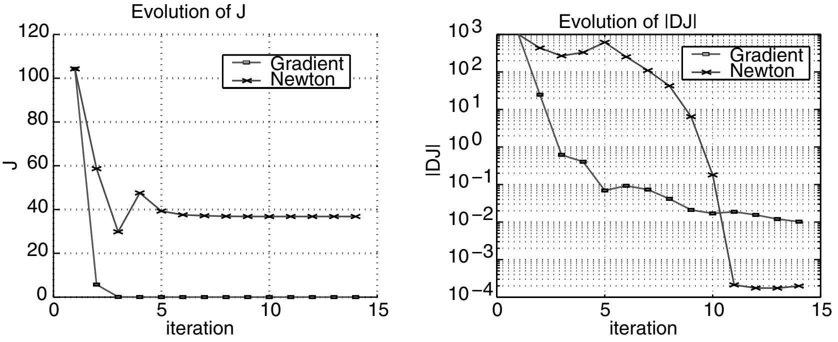


FIGURE 9.2 Convergence of $J(r, \omega(r))$ and $\|\nabla_\omega J(r, \omega(r))\|$.

2.3.2 Newton algorithm for minimizing $F(\omega)$ on the unit sphere S

1. Given ω such that $\omega^t \omega = 1$,
2. Compute Δ solution of the system (2.6):

$$\begin{cases} DF(\omega)\Delta_1 = \omega \\ \Delta = \frac{-1}{\omega^t \Delta_1}\Delta_1 + \omega \end{cases}$$

3. Compute the update $\qquad \omega_+ = \omega + \Delta.$

2.3.3 Comparison of gradient and Newton methods

In Figure 9.2, we compare the conjugate gradient and Newton's method applied to an initial vector ω_0 far from the optimal solution ω^*. We can see that Newton's method converges toward a local minimum.

In Figure 9.3 we compare the conjugate gradient and Newton's method applied to the tenth conjugate gradient iteration. We can see that Newton's

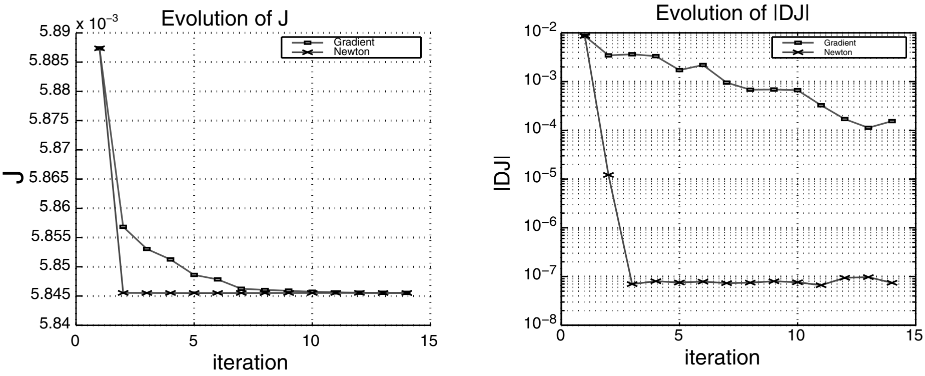


FIGURE 9.3 Convergence of $J(r, \omega(r))$ and $\|\nabla_\omega J(r, \omega(r))\|$.

method converges toward a global minimum more quickly than the gradient method, which results in a doubling of the accuracy for the solution of Equation 2.4 and demonstrates cubic convergence of Newton's method.

3 Numerical result

To assess the effectiveness of the proposed approach, we consider a planar antenna array where elements are divided into groups of subarrays, and each subarray is fed through a single weight (ω_k), and where (r_k) represents its center. These subarrays can be viewed as elements of a second array. In this example we assume that all subarrays are identical — that is, $(g_k(u_0) = g(u_0) \forall k \in [1 \dots N])$ — and refer to each one as a *primary array.* The array of the primary array is called the *secondary array.* The combined array factor will be equal to the product of these two array factors such that

$$d(r, \omega, u_0) = g(u_0). \sum_{k=1}^{N} \omega_k e^{j \frac{2\pi}{\lambda} \langle u_0 . r_k \rangle_{\mathbb{R}^3}}$$

with

$$g(u_0) = \frac{\sin\left(\frac{\pi}{\lambda} L_x (\sin(\theta_0) \cos(\varphi_0))\right)}{N_x \sin\left(\frac{\pi}{\lambda} \frac{L_x}{N_x} (\sin(\theta_0) \cos(\varphi_0))\right)} \; \frac{\sin\left(\frac{\pi}{\lambda} L_y (\sin(\theta_0) \sin(\varphi_0))\right)}{N_y \sin\left(\frac{\pi}{\lambda} \frac{L_y}{N_y} (\sin(\theta_0) \sin(\varphi_0))\right)}$$

where (θ_0, φ_0) is the parameterization of $u_0 \in S^2$, and (N_x, N_y) represents the number of elements in the primary array. (L_x, L_y) is the length of the primary array: $L_x = N_x . \frac{\lambda}{2}$, $L_y = N_y . \frac{\lambda}{2}$.

More precisely, we consider a rectangular array made of 60 subarrays, that is, 12 on axis x and 5 on axis y (Figure 9.4). Each subarray represents

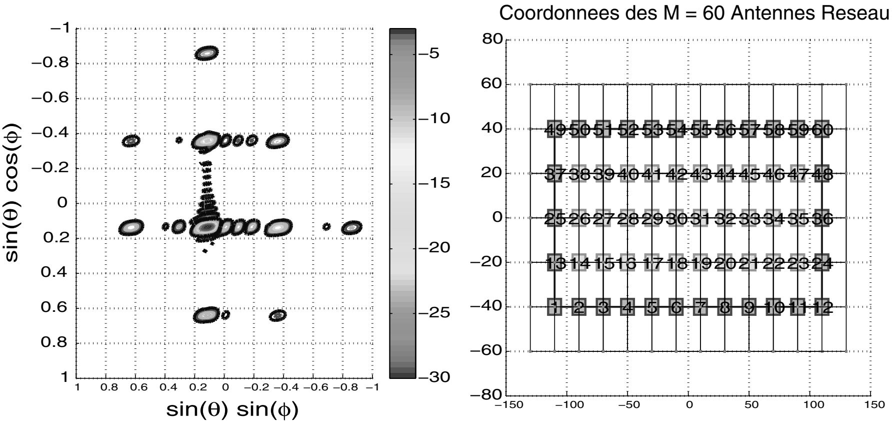


FIGURE 9.4 Isovalues of *directive gain* in *dB* (left) and geometry of the antenna array (right).

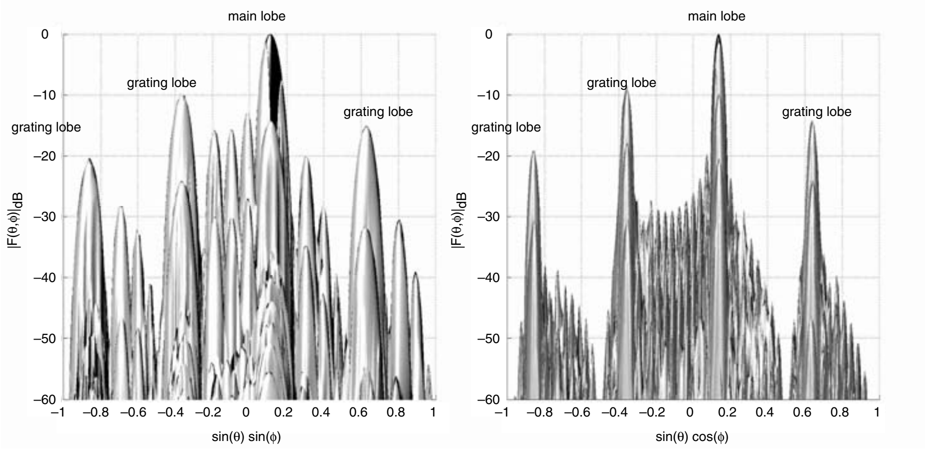


FIGURE 9.5 *Directive gain* in *dB* in the areas $\mathcal{V}_\lambda(u_p)$ (black) and $S^2\backslash\mathcal{V}_\lambda(u_p)$ (gray).

a square of 64 isotropic sources ($N_x = N_y = 8$). We decide to focus the *directive gain* in the angular sector $\mathcal{V}_\lambda(u_p)$ centered at $u_p = (\theta_p, \varphi_p) = (9°, 9°)$.

3.1 Shape optimization procedure

- First of all, with the initial and rectangular antenna array we realize a *weight optimization* in solving the problem ($\mathcal{P}_1$) (Equation 2.3). The result of this *weight optimization* method is discussed in Section 3.2. The uniformly spaced secondary array creates grating lobes in the visible region as we can see in Figure 9.5 because the spacing between the subarrays exceeds the maximal value $\frac{\lambda}{2}$. The grating lobes are undesirable because they represent unwanted *main lobes*. Indeed when grating lobes appear, *energy* is taken away from the main lobe, and placed into the grating lobe, resulting in a loss of *directive gain*.
- Then we realize a *shape optimization*, which optimizes both positions and weights of the array elements in solving problem ($\mathcal{P}$) (Equation 2.2). The result of the *shape optimization* method is discussed in Section 3.3.
- Finally, the optimal geometry found at the preceding stage (Figure 9.6) has a superposition of these subarrays between them. For that reason, we realize a *shape optimization under constraints* in solving the problem: $(\mathcal{P}_\mathcal{R}) \min_{r\in\mathcal{R}} \min_{\omega\in\mathcal{W}} J(r,\omega)$, where $\mathcal{R}$ is the feasible set for the geometrical parameter r. The result of the *shape optimization under constraints* method is discussed in Section 3.4.

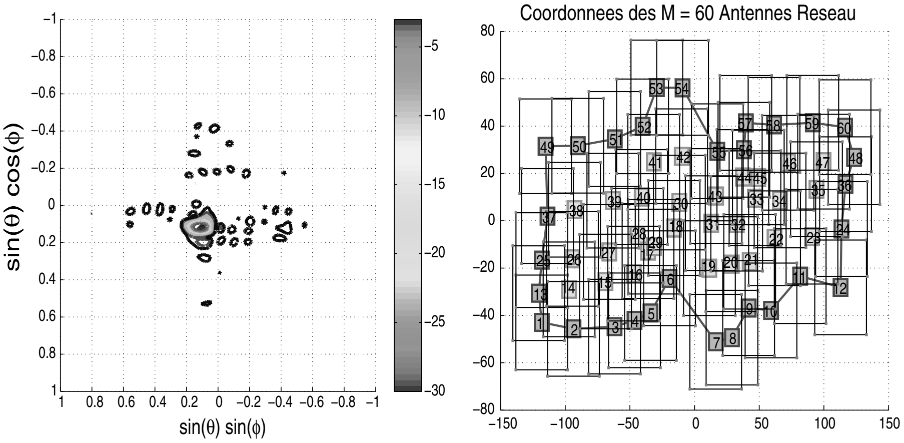


FIGURE 9.6 Isovalues of *directive gain* in *dB* (left) and geometry of the antenna array (right).

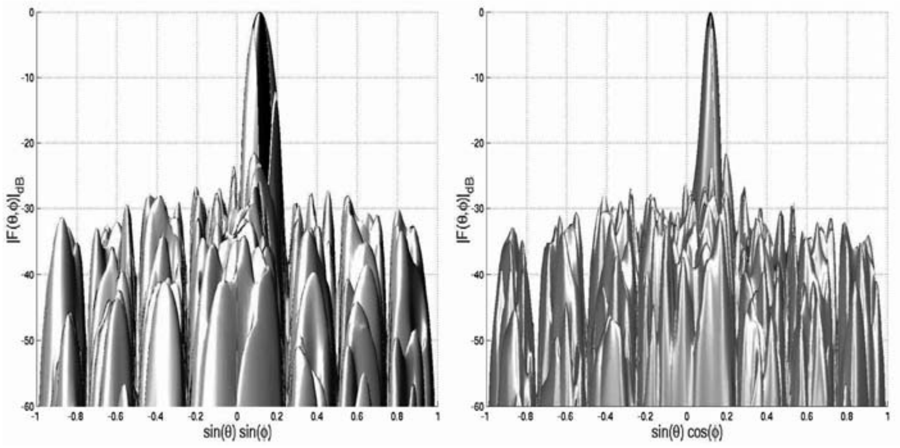


FIGURE 9.7 *Directive gain* in *dB* in the areas $\mathcal{V}_\lambda(u_p)$ (black) and $S^2\backslash\mathcal{V}_\lambda(u_p)$ (gray).

3.2 Optimization of weights

	Directive Gain Average	Directive Gain Maximum
In zone $\mathcal{V}_\lambda(u_p)$	31.7237 dB	34.93806 dB
In zone $S^2\backslash\mathcal{V}_\lambda(u_p)$	1.8319 dB	26.5047 dB

Note: Difference between the main lobe and the highest grating lobe: 8.43 dB.

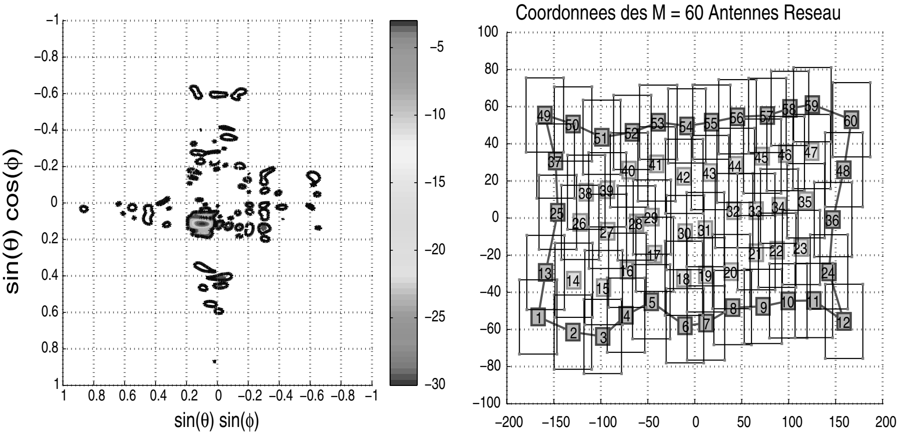


FIGURE 9.8 Isovalues of *directive gain* in *dB* (left) and geometry of the antenna array (right).

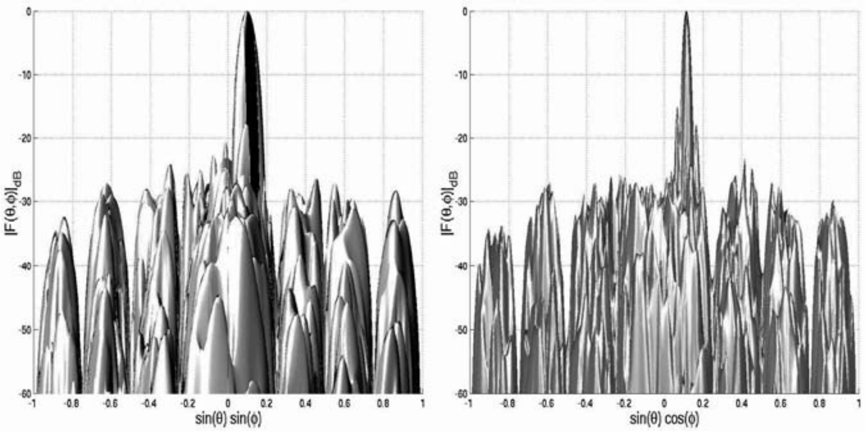


FIGURE 9.9 *Directive gain* in *dB* in the areas $\mathcal{V}_\lambda(u_p)$ (black) and $S^2\backslash\mathcal{V}_\lambda(u_p)$ (gray).

3.3 Optimization of element positions and weights

	Directive Gain Average	Directive Gain Maximum
In zone $\mathcal{V}_\lambda(u_p)$	33.6646 dB	37.3711 dB
In zone $S^2\backslash\mathcal{V}_\lambda(u_p)$	−1.5384 dB	12.5827 dB

Note: Difference between the main lobe and the highest side lobe: 24.79 dB.

3.4 Optimization of elements positions with constraints and weights

	Directive Gain Average	Directive Gain Maximum
In zone $\mathcal{V}_\lambda(u_p)$	33.3015 dB	38.5988 dB
In zone $S^2\backslash\mathcal{V}_\lambda(u_p)$	0.1552 dB	17.5115 dB

Note: Difference between the main lobe and the highest side lobe: 21.08 dB.

4 Conclusion

The shape optimization of the antenna's geometry prevents the creation of grating lobes, and reduces the side lobes, as we can see in Figures 9.7 and 9.9. Note that the main difficulty is solving problem $(\mathcal{P}_\mathcal{R})$ because we must apply the geometrical constraints for the antenna array.

References

[1] M.C. Delfour and J.-P. Zolésio. *Shapes and Geometries*, Advances in design and control, 4, SIAM, Philadelphia, 2001.

[2] M.C. Delfour and J.-P. Zolésio. Oriented distance function and its evolution equation for initial sets with thin boundary, *SIAM J. Control Optim.*, 2004, 42, 6, 2286–2304.

[3] M.C. Delfour and J.-P. Zolésio, Shape sensitivity analysis via min max differentiability, *SIAM J. Control Optim.*, 1988, 26, 4, 834–862.

[4] J.-P. Zolésio. In *Optimization of Distributed Parameter Structures*, vol. II (E. Haug and J. Céa, eds.), Adv. Study Inst. Ser. E: Appl. Sci., 50, Sijthoff and Nordhoff, Alphen aan den Rijn, 1981: i) The speed method for shape optimization, 1089–1151; ii) Domain variational formulation for free boundary problems, 1152–1194; iii) Semiderivative of repeated eigenvalues, 1457–1473.

[5] M. Cuer and J.-P. Zolésio. Control of singular problem via differentiation of a min-max, *Systems Control Lett.*, 1988, 11, 2, 151–158.

CHAPTER 10

The Stokes basis for three-dimensional incompressible flow fields

Giles Auchmuty
Division of Mathematical Sciences, National Science Foundation, Arlington, Virginia, on leave from Department of Mathematics, University of Houston, Houston, Texas

1 Introduction

This paper begins by developing the weak formulation of a spectral problem for the Stokes operator. A classical description of the Stokes eigenproblem is given in chapter 4 of Constantin and Foias [4]. The development there is done in the setting of a closed, densely defined linear operator on the Hilbert space of L^2-vector fields on Ω. Another version is described in section 2.6 of chapter 1 of Temam [8]. Here the Stokes eigenfields will be constructed using a direct variational characterization on a natural Sobolev-Hilbert space of vector fields. These eigenfields will be proven to be a basis of the space of H^1-incompressible flow fields obeying no-slip boundary conditions. It will be called the *Stokes basis*. Various properties of these base fields will be proved and spectral formulae for the energy and enstrophy will be derived. In particular, the helicity of an eigenfield and of its vorticity are related by a simple formula, and various formulae for the coefficients in spectral expansions of the field are derived and used.

2 Function spaces and notation

A *region* in $\mathbb{R}^3$ is a nonempty, connected, open subset of $\mathbb{R}^3$. Its closure is denoted by $\overline{\Omega}$ and its boundary is $\partial\Omega := \overline{\Omega} \setminus \Omega$. We generally require:

Condition B1

Ω is a bounded region in $\mathbb{R}^3$ and $\partial\Omega$ is the union of a finite number of disjoint closed $C^{1,1}$ surfaces, each surface having finite surface area.

A closed surface Σ in space is said to be $C^{1,1}$ if it has a unique unit outward normal ν at each point and ν is a Lipschitz continuous vector field on Σ. See Girault and Raviart [7], Section 1.1. for more details about this definition. The surface area measure on $\partial\Omega$ will be denoted by $d\sigma$.

When u, v are vectors in $\mathbb{R}^3$, their scalar product, the Euclidean norm, and vector product are denoted $u \cdot v$, $|u|$, and $u \wedge v$, respectively. In the following $\langle u, v\rangle$ denotes the L^2-inner product and $\|u\|$ denotes the L^2-norm of a function or vector field on Ω.

Let $H_0^1(\Omega)$ be the usual Sobolev-Hilbert space of all scalar-valued functions on Ω and $H_0^1(\Omega;\mathbb{R}^3)$ be the corresponding space of vector fields. Our results are based on using a special, natural inner product on this space, namely,

$$\langle u, v\rangle_{DC} := \int_\Omega [\operatorname{curl} u \cdot \operatorname{curl} v + \operatorname{div} u \cdot \operatorname{div} v]\; d^3x. \tag{2.1}$$

See [5], volume 3, chapter 9 for a proof that this inner product is equivalent to the usual H^1-inner product on $H_0^1(\Omega;\mathbb{R}^3)$.

When $v \in H_0^1(\Omega;\mathbb{R}^3)$, the *divergence of* v is the function $\rho \in L^2(\Omega)$ that satisfies

$$\int_\Omega \rho\, \varphi d^3x = \int_\Omega u \cdot \nabla\varphi d^3x \quad \text{for all } \varphi \in C_c^\infty(\Omega). \tag{2.2}$$

Here $C_c^\infty(\Omega)$ is the space of all C^∞ functions on Ω with compact support. A field $v \in H_0^1(\Omega;\mathbb{R}^3)$ is said to be *incompressible* or *divergence free* when $\rho \equiv 0$ in $L^2(\Omega)$. The subspace of all incompressible vector fields in $H_0^1(\Omega;\mathbb{R}^3)$ will be denoted $V_0^1(\Omega)$ and is a closed subspace of $H_0^1(\Omega;\mathbb{R}^3)$. It will be a real Hilbert space under the induced inner product

$$\langle v, w\rangle_C := \int_\Omega \operatorname{curl} v \cdot \operatorname{curl} w\; d^3x. \tag{2.3}$$

Fields in $V_0^1(\Omega)$ will be said to be C-orthonormal provided they are orthonormal with respect to this inner product.

When (B1) holds, the Gauss-Green theorem holds on Ω. Two consequences of the Gauss-Green theorem will be used repeatedly here; the following hold when each of the individual integrals is finite:

$$\int_\Omega u \cdot \nabla\varphi\; d^3x \;= \int_{\partial\Omega} \varphi\; (u \cdot \nu)\; d\sigma - \int_\Omega \varphi \operatorname{div} u\; d^3x, \tag{2.4}$$

$$\int_\Omega u \cdot \operatorname{curl} v\; d^3x \;= \int_{\partial\Omega} v \cdot (u \wedge \nu)\; d\sigma + \int_\Omega v \cdot \operatorname{curl} u\; d^3x. \tag{2.5}$$

We will need the following characterization of the L^2-orthogonal complement of $V_0^1(\Omega)$.

Theorem 2.1

Assume (B1) holds, $v \in H_0^1(\Omega;\mathbb{R}^3)$ *and*

$$\int_\Omega v \cdot w \, d^3x = 0 \quad \textit{for all } w \in V_0^1(\Omega). \tag{2.6}$$

Then $v = \nabla\varphi$ *where* $\varphi \in H_0^1(\Omega), \Delta\varphi \in L^2(\Omega)$ *and* $\partial\varphi/\partial\nu = 0$ *on* $\partial\Omega$.

Proof

This proof depends on the results of section 7 of Auchmuty [3] and the notation of that paper will be used. We first show that the space $C_0^1(\Omega)$ defined there is the same as $V_0^1(\Omega)$ defined here. When $v \in C_0^1(\Omega)$, then $v \in H_0^1(\Omega;\mathbb{R}^3)$ and $\operatorname{div} v = 0$ so $v \in V_0^1(\Omega)$.

Conversely, if $v \in V_0^1(\Omega)$, then Corollary 7.4 says that the projection of v onto $G_0^1(\Omega)$ is defined by $P_G v = \nabla\tilde{\varphi}$ where $\tilde{\varphi}$ minimizes

$$\int_\Omega |\Delta\varphi|^2 d^3x \quad \text{over } \varphi \in X_0.$$

The unique minimizer of this is $\tilde{\varphi} = 0$, so $P_G v = 0$ and thus $v \in C_0^1(\Omega)$. Hence the two spaces are equal.

Now use (iii) of theorem 7.2 in [3]. It implies that for any $v \in H_0^1(\Omega;\mathbb{R}^3)$, there is a $\varphi \in X_0$ such that

$$v = \nabla\varphi + u \quad \text{with } u \in V_0^1(\Omega).$$

Substitute this in (2.6) and take $w = u$; then the left-hand side is $\|u\|^2$. Hence $u = 0$ and then the theorem follows upon using the characterization of X_0 given there.

Note that (2.6) will hold provided it holds for all fields w that are C^∞, solenoidal, and have compact support in Ω.

3 The Stokes eigenproblem

The Stokes eigenproblem has been discussed in many texts including chapter 4 of [4] and chapter 1 of [8]. Here we shall describe a weak formulation in the setting of the Sobolev space $V_0^1(\Omega)$ of incompressible vector fields.

A vector field $v \in V_0^1(\Omega)$ is said to be an *eigenfield for the Stokes eigenproblem* on Ω corresponding to an eigenvalue λ, provided it is a nonzero solution of the equation

$$\int_\Omega (\operatorname{curl} v \cdot \operatorname{curl} w - \lambda v \cdot w) \, d^3x = 0 \quad \text{for all } w \in V_0^1(\Omega). \tag{3.7}$$

It is straightforward to verify that such a field is a weak version of the usual description of this eigenproblem. It may be considered as an eigenproblem for either the $\operatorname{curl}^2 := \operatorname{curl}(\operatorname{curl})$ operator, or the Laplacian vector, on this space of incompressible fields. Putting $w = v$ in (3.7) yields that any eigenvalue of this Stokes eigenproblem is positive.

The existence and properties of the eigenfields of this problem will be obtained using variational arguments. Consider the variational problem of minimizing the functional

$$\mathcal{C}(v) := \|v\|_C^2 := \int_\Omega |\operatorname{curl} v|^2 \, d^3x \tag{3.8}$$

on the set

$$B_1 := \left\{ v \in V_0^1(\Omega) : \int_\Omega |v|^2 d^3x = 1 \right\}. \tag{3.9}$$

The set B_1 is a weakly closed set in $V_0^1(\Omega)$ from Rellich's theorem, so standard arguments, including a Poincare inequality, lead to the following result.

Theorem 3.1

Assume that (B1) holds; then there are fields $\pm v^{(1)} \in V_0^1(\Omega)$, which minimize $\mathcal{C}$ on B_1. They satisfy (3.7) with $\lambda = \lambda_1$ being the least eigenvalue of this problem and $\lambda_1 = \mathcal{C}(v^{(1)}) > 0$.

Knowing the least eigenvalue λ_1 and a corresponding eigenfield $v^{(1)}$, the successive eigenvalues $\{\lambda_j : j \geq 1\}$ and the corresponding eigenfields $v^{(j)}$ may be characterized inductively. For $J \geq 1$, define

$$B_{J+1} := \{ v \in V_0^1(\Omega) : \int_\Omega |v|^2 d^3x = 1, \int_\Omega v \cdot v^{(j)} d^3x = 0 \quad \text{for } 1 \leq j \leq J \}. \tag{3.10}$$

Consider the problem of minimizing the functional $\mathcal{C}$ defined by (3.8) on B_{J+1}. The following result describes the solutions of this problem.

Theorem 3.2

Assume that (B1) holds, then there are fields $\pm v^{(J+1)} \in V_0^1(\Omega)$ that minimize $\mathcal{C}$ on B_{J+1}. They satisfy (3.7) with $\lambda = \lambda_{J+1} = \mathcal{C}(v^{(J+1)}).\lambda_{J+1}$ is the least eigenvalue of this problem greater than or equal to λ_J.

The proofs of these theorems parallel the usual proofs of similar properties for the eigenvalues and eigenfunctions of the Dirichlet Laplacian on $H_0^1(\Omega)$. In fact, the two-dimensional analog of this problem is precisely such a problem.

In this construction the fields $v^{(j)}$ are L^2-orthogonal from the definition of the B_J. Then Equation (3.7) yields that

$$\int_\Omega \operatorname{curl} v^{(j)} \cdot \operatorname{curl} v^{(k)} d^3x = \lambda_j \delta_{jk} \qquad \text{for} \quad j, k \geq 1. \tag{3.11}$$

That is, the corresponding vorticities $\omega^{(j)} := \operatorname{curl} v^{(j)}$ will be L^2-orthogonal on Ω. Write $\tilde{v}^{(j)} := {\lambda_j}^{-1/2} v^{(j)}$, then the set $\mathcal{B} := \{\tilde{v}^{(j)} : j \geq 1\}$ will be a C-orthonormal set in $V_0^1(\Omega)$.

This leads to the following result about the sequence of eigenvalues $\sigma_S(\Omega) := \{\lambda_j : j \geq 1\}$ and the corresponding eigenfields of this eigenproblem.

Theorem 3.3

Assume that (B1) holds, and the sequences $\sigma_S(\Omega), \mathcal{B}$ are defined as above. Then each eigenvalue λ_j has finite multiplicity and $\lambda_j \to \infty$ as $j \to \infty$. Moreover $\mathcal{B}$ is a maximal C-orthonormal set in $V_0^1(\Omega)$.

Proof

Suppose that there are infinitely many L^2-orthogonal eigenfields $v^{(j)}$ of this Stokes eigenproblem corresponding to eigenvalues λ_j with $\lambda_j < c$ for some number c. Then (3.7) implies that

$$\|v^{(j)}\|_C^2 = \int_\Omega |\operatorname{curl} v^{(j)}|^2 d^3x < c \qquad \text{for all} \quad j \geq 1.$$

Thus this sequence is bounded in $H_0^1(\Omega; \mathbb{R}^3)$ so it will be compact in $L^2(\Omega; \mathbb{R}^3)$ from Rellich's theorem. This is impossible if they are L^2-orthornormal, so the set must be finite. Thus the second sentence of the theorem holds.

Suppose $\mathcal{B}$ is not a maximal C-orthonormal set in $V_0^1(\Omega)$. Then there is a $w \in V_0^1(\Omega)$ with $\|w\|_C = 1$ and $\langle w, v^{(j)}\rangle_C = 0$ for all $j \geq 1$. Then $\tilde{w} := w/\|w\|$ will be in B_j for every $j \geq 1$. But $\mathcal{C}(\tilde{w})$ is finite, so this contradicts the definition of λ_j for j large enough. Thus the set $\mathcal{B}$ must be maximal as claimed.

This theorem shows that a class of Stokes eigenfields defined by the above method of successive optimization generates an orthonormal basis of the space $V_0^1(\Omega)$. The next two sections will describe some consequences of this.

This weak characterization of the eigenfields also yields the following interesting result. A vector field $v \in H^1(\Omega, \mathbb{R}^3)$ is said to be in $H_{\nu 0}^1(\Omega; \mathbb{R}^3)$ provided its normal boundary trace $v \cdot \nu = 0$ on $\partial\Omega$.

Proposition 3.4

Assume (B1) holds and $v^{(j)} in V_0^1(\Omega)$ is an eigenfield of (3.7). The $\omega^{(j)}$ is an incompressible field in the space $H_{\nu 0}^1(\Omega; \mathbb{R}^3)$.

Proof

The proof that $\omega^{(j)}$ is incompressible is straightforward. From (3.7) it will satisfy

$$\left| \int_\Omega \omega^{(j)} \cdot curl\, w\, d^3x \right| \leq \lambda_j \|w\|_2 \qquad \text{for all} \quad w \in V_0^1(\Omega).$$

This implies that $\operatorname{curl} \omega^{(j)} \in L^2(\Omega; \mathbb{R}^3)$. Evaluate

$$\int_\Omega \omega^{(j)} \cdot \nabla\varphi \; d^3x = \int_{\partial\Omega} \varphi \; (\omega^{(j)} \cdot \nu) \; d\sigma = 0$$

using (2.5) and the fact that $v^{(j)} = 0$ on $\partial\Omega$. This holds for all smooth φ on $\overline{\Omega}$ so the boundary trace $\omega^{(j)} \cdot \nu \equiv 0$ on $\partial\Omega$. When this holds, and the divergence and curl of $\omega^{(j)}$ are both L^2, it follows that $\omega^{(j)} \in H^1_{\nu 0}(\Omega; \mathbb{R}^3)$ from theorem 6.1, chapter 8 of Duvaut-Lions [6].

Let A be a smooth incompressible vector field with compact support in Ω. Substitute $w = \operatorname{curl} A$ in (3.7); then use of (2.5) leads to

$$\int_\Omega [\operatorname{curl} \omega^{(j)} \cdot \operatorname{curl} A - \lambda_j\, \omega^{(j)} \cdot A] d^3x = 0 \qquad \text{for all such A.} \tag{3.12}$$

This implies that $\omega^{(j)}$ is an eigenfield of the *zero-flux* curl^2 eigenproblem on Ω. This problem is of central importance in magnetostatics, and some of its properties are described in [1] and [2]. That is, the eigenvalues of the Stokes eigenproblem are a subset of those for the zero flux curl^2 eigenproblem. When Ω is not simply connected, the later problem has 0 as an eigenvalue of nonzero multiplicity, so these two problems will not be isospectral in this case.

4 Energy, enstrophy, and helicity formulae

The representation of an incompressible field with respect to this Stokes basis provides some useful spectral formulae for the energy and enstrophy of the incompressible fields obeying no-slip boundary conditions.

A vector field $v \in V_0^1(\Omega)$ will have a representation of the form

$$v = \sum_{j=1}^{\infty} c_j\, \tilde{v}^{(j)} \qquad \text{with} \quad c_j := [v, \tilde{v}^{(j)}]_C \tag{4.13}$$

because $\mathcal{B}$ is an orthonormal basis of $V_0^1(\Omega)$. Note that the coefficients are also given by

$$c_j = \sqrt{\lambda_j} \int_\Omega v \cdot v^{(j)} d^3x \tag{4.14}$$

upon using (3.7) and the definition of the $\tilde{v}^{(j)}$. Moreover, from Parseval's equality, an expression of the form (4.13) represents a field in $V_0^1(\Omega)$ if and only if

$$\| \operatorname{curl} v \|^2 = \|v\|_C^2 = \sum_{j=1}^{\infty} |c_j|^2 < \infty \tag{4.15}$$

This quantity is often called the *enstrophy* of the vector field v.

The *kinetic energy* of the field v will be

$$\|v\|^2 = \sum_{j=1}^{\infty} {\lambda_j}^{-1} |c_j|^2 \tag{4.16}$$

upon using (3.7) and the orthogonality relations.

The *helicity* of a field v is the functional

$$\mathcal{H}(v) := \int_\Omega v \cdot \operatorname{curl} v \, d^3x \tag{4.17}$$

and this will be finite for all $v \in V_0^1(\Omega)$. Moreover, use of (4.13) shows that this quantity can be expressed in terms of the values of

$$\langle v^{(j)}, \omega^{(k)} \rangle = \langle v^{(k)}, \omega^{(j)} \rangle \qquad \text{for} \quad j, k \geq 1.$$

The following relates the helicity of these basis fields and that of their vorticity.

Proposition 4.1

Assume (B1) holds, $v^{(j)}$ is a Stokes eigenfield corresponding to the eigenvalue λ_j and $\omega^{(j)} = \operatorname{curl} v^{(j)}$. Then $\mathcal{H}(\omega^{(j)}) = \lambda_j \, \mathcal{H}(v^{(j)})$.

Proof

When $v^{(j)}$ is a Stokes eigenfield corresponding to the eigenvalue λ_j, then Proposition 2.1 and (3.7) imply

$$\operatorname{curl}^2 v^{(j)} = \lambda_j v^{(j)} + \nabla p \quad \text{on}\, \Omega$$

for some p. Take scalar products with $\omega^{(j)}$ and integrate over Ω, then the result follows.

5 Reconstruction of the velocity from the vorticity

A well-known problem for fluid flows is the problem of determining the velocity of a flow field from the vorticity of a flow in a bounded region. For two-dimensional flows this is discussed in many texts. An analysis of this for incompressible flows in a bounded three-dimensional region may be found in Auchmuty [3].

Suppose $v \in V_0^1(\Omega)$ is given by (4.13), then the coefficients c_j are given by

$$c_j = \int_\Omega \operatorname{curl} v \cdot \tilde{\omega}^{(j)} d^3x. \tag{5.18}$$

In particular, these coefficients are determined in terms of the vorticity and the vorticities of the Stokes basis fields. That is, given the vorticity of a field, we need only evaluate these integrals to determine c_j and then, from (4.13),

$$v = \sum_{j=1}^{\infty} \lambda_j \, \langle \operatorname{curl} v, \omega^{(j)} \rangle v^{(j)} \quad \text{on}\, \Omega. \tag{5.19}$$

This is valid for any incompressible field satisfying no-slip boundary conditions and of finite enstrophy.

References

[1] G. Auchmuty and J.C. Alexander, L^2-well-posedness of 3d div-curl boundary value problems on bounded regions, *Q. of Applied Mathematics* (forthcoming).

[2] G. Auchmuty, Orthogonal decompositions and bases for three-dimensional vector fields, *Numer. Funct. Anal. and Optimiz.* **15** (1994), 445–488.

[3] G. Auchmuty, Reconstruction of the velocity from the vorticity in three-dimensional fluid flows, *Proc. Royal Soc. Lond.* A **454** (1998), 607–630.

[4] P. Constantin and C. Foias, *Navier-Stokes Equations*, University of Chicago Press, Chicago (1989).

[5] R. Dautray and J.L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, 6 volumes, Springer-Verlag, Berlin (1990).

[6] G. Duvaut and J.L. Lions, *Inequalities in Mechanics and Physics*, Springer-Verlag, Berlin (1976).

[7] V. Girault and P.A. Raviart, *Finite Element Methods for the Navier-Stokes Equations*, Springer-Verlag, Berlin (1986).

[8] R. Temam, *Navier-Stokes Equations: Theory and Numerical Analysis*, North-Holland, Amsterdam (1977).

CHAPTER 11

Nonlinear aeroelasticity, continuum theory, flutter/divergence speed, plate wing model

A.V. Balakrishnan
Flight Systems Research Center, University of California at Los Angeles,
Los Angeles, California

1 Introduction

An important safety consideration for aircraft in subsonic and transonic flight is an endemic instability known as wing-bending-torsion flutter, which occurs at a high enough speed. Almost all the current work is directed toward numerical computation, merging the FEM NASTRAN codes with the aerodynamic CFD *time-marching* codes to convert a time-domain waveform-approximating partial differential equations into ordinary differential equations. While this certainly allows “realistic” (nonlinear-complex geometry) wings, it does require numerical specification of parameters, thus limiting generality, and provides little or no insight into the phenomena at work, and, of course, is inadequate for any control design for stabilization (flutter suppression). Indeed, as noted in [10]: “despite computational and experimental research extending over more than 20 years, we do not have a good fundamental understanding of (transonic) flutter.”

It is possible, as we show, to retain the full continuum models in what is really a boundary value problem for a pair of coupled nonlinear partial differential equations, leading to an abstract time-domain model as a nonlinear convolution/evolution equation in a Hilbert space. This, in turn, allows us to use the Hopf bifurcation theory and characterize the flutter speed as a Hopf bifurcation point for the speed parameter, which is then completely characterized by the linearized equations about the equilibrium state. An essential tool for solving the linear equations is a singular integral equation discovered by Camillo Possio in 1938 and bearing his name, generalized for a nonzero angle of attack. The first results following this approach are given in [2,3], where the structure model is the Goland beam model.

In this paper we extend the beam model for the wing structure to a plate model, allowing in particular for nonzero wing thickness and camber bending. The emphasis is on problem formulation, although some preliminary results are included. It is believed that the problems should be of independent interest as well.

The plate wing model is described in Section 2 and the aerodynamics and the TSD potential equation used are discussed in Section 3. The aeroelastic boundary conditions, which play the crucial role, are developed in Section 4. The combined equations are captured for the first time in a time-domain formulation as a nonlinear convolution–evolution equation in a Hilbert space. Invoking the Hopf bifurcation theory on stability, the linearization procedure is described in Section 6, where the Possio equation provides the link between the structure and aerodynamics. The resulting aeroelastic equations in the Laplace domain are given in Section 7, which allows us to give a precise definition of flutter speed. The stationary equation and divergence speed are treated in Section 8. Concluding remarks are in Section 9.

2 The wing structure: Kirchoff thin plate model

As early as 1964, E. Dowell considered a Kirchoff thin plate model ([4], see also [18]), albeit two-dimensional (neglecting spanwise bending), and a highly nonlinear 3-D model for chordwise deflection and the Airly angle for flutter analysis in the supersonic case. The objective here is purely computational, as also in more recent papers [8,9], where no continuum models are considered of relevance to the present paper.

Here we consider a Kirchoff thin plate model (CFFF cantilever), following the classic Timoshenko work [22]; see also Leisa [18]. The bending deflection is both chordwise (so camber bending is allowed) and spanwise. In addition, we allow pitching about an axis parallel to the span axis, which is then coupled to the bending, but does not depend on the chord variable. Referring to Figure 11.1, where the wing axes and other relevant parameters are given, let $h(s, y, t)$ denote the bending normal to the wing and $\theta(y, t)$ the pitch angle in radians. Then the plate dynamics take the form:

$$m\ddot{h}(x, y, t) + D\Delta^2 h(x, y, t) + \frac{S}{2b}\ddot{\theta}(y, t) = \delta p, \tag{2.1}$$

where $0 < t,\ -b < x < b,\ 0 < y < l$ and $\Delta h = \frac{\partial^2 h}{\partial x^2} + \frac{\partial^2 h}{\partial y^2}$.

$m :$ mass per unit area of the plate

$\delta :$ plate thickness

$D = \frac{E\delta^3}{12(1-y^2)}$

$S :$ coupling constant $=$ (mass per unit length) $\cdot\, bx_\alpha$

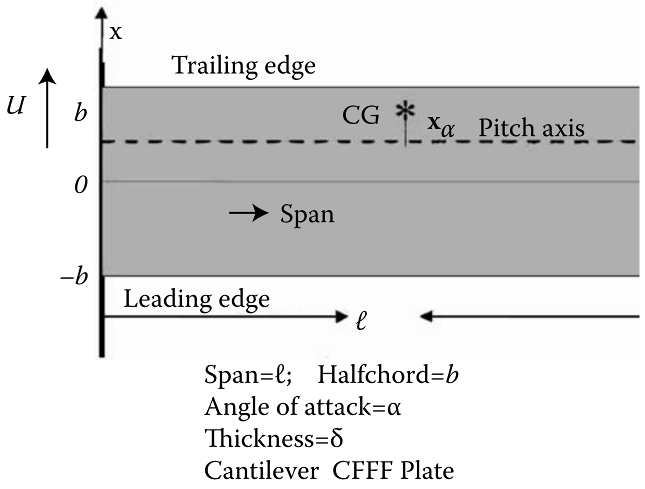


FIGURE 11.1 Wing structure.

bx_α = location of c.g. relative to the elastic axis
δp = aerodynamic pressure (see Section 4)

The CFFF plate is clamped at the root so that:

$$h(x,0,t) = 0, \qquad -b < x < b, \quad 0 < t$$

$$\frac{\partial}{\partial y}h(x,0,t) = 0, \qquad -b < x < b, \quad 0 < t,$$

and free at the other edges so that at $y = l$ (see p. 84, [6]):

$$\frac{\partial}{\partial y}\left(\Delta h + (1-\gamma)\frac{\partial^2 h}{\partial x^2}\right) = 0, \quad -b < x < b$$

$$\Delta h + (\gamma - 1)\frac{\partial^2 h}{\partial x^2} = 0, \quad -b < x < b,$$

and at $x = -b$ or $x = b$:

$$\frac{\partial}{\partial x}\left(\Delta h + (1-\gamma)\frac{\partial^2 h}{\partial y^2}\right) = 0, \quad 0 < y < l$$

$$\Delta h + (\gamma - 1)\frac{\partial^2 h}{\partial y^2} = 0, \quad 0 < y < l.$$

We have a cantilever CFFF plate.

The pitch angle $\theta(y,t)$ in radians satisfies:

$$I_\theta \ddot{\theta}(y,t) - aJ\frac{\partial^2\theta(y,t)}{\partial y^2} + S\ddot{h}(x,y,t)$$

$$= \text{aerodynamic moment per unit span} = \int\limits_{-b}^{b} (x - ab)\delta_p dx$$

the elastic axis being located at $x = ab$, $|a| < \frac{b}{2}$, with the end conditions

$$\theta(0,t) = 0; \quad \frac{\partial}{\partial y}\theta(y,t) = 0 \quad y = l.$$

Note that if we omit the chordwise dependence of the bending and set the damping constant S to zero in Equation (2.1), we obtain the 2-D plate equation of Dowell [16] omitting the *supersonic* term. It should be noted that the operator Δ^2 with the CFFF boundary condition (assumed from now on) is self-adjoint and nonnegative definite over the Hilbert space $H_h = L_2(\Omega)$ where

$$\Omega = \{-b < x < b;\ 0 < y < l\} \quad \text{with} \quad [\Delta^2 h, h] = [\Delta h, \Delta h], \quad \text{for } h \text{ in } \mathcal{D}(\Delta^2).$$

Similarly, over the Hilbert space $H_\theta = L_2[0,l]$, the operator A_θ, defined by

$$A_\theta\theta(\cdot) = -\ddot{\theta}(\cdot)$$
$$\mathcal{D}(A_\theta) = \{\theta(\cdot)|\theta(0) = \dot{\theta}(l) = 0; \quad \ddot{\theta}(\cdot) \in L_2(0,l)\},$$

is self-adjoint and nonnegative definite.

Moreover, the total potential energy is given by:

$$= \frac{1}{2}\int\limits_{-b}^{b}\int\limits_{0}^{l}(\Delta h)^2 dxdy + \frac{1}{2}\int\limits_{0}^{l}|\dot{\theta}(y)|^2 dy.$$

Note that neither Δ^2 nor A_θ has zero eigenvalues, but of course they have compact resolvents.

3 Aerodynamics: the TSD (transonic small disturbance) equation

We consider only inviscid flow; hence the flow is characterized by the velocity potential equation—the full potential equation of Euler (see [13], for example). Because our goal is flutter/divergence instability, and by the Hopf theory is determined by the linearized equations about the equilibrium, we work with the transonic small disturbance equation [10] (see [13] for a recent derivation). A numerical treatment is given in [8,9], again with the aim of *predicting* flutter and divergence. According to [10] it is capable of producing shocks. Following the notation in [10], let $\Phi(x, y, z, t)$ denote the velocity potential with $\Phi_\infty(x, y, z, t)$ the far-field potential with

$$\nabla\Phi_\infty(x, y, z, t) = q_\infty = iq_1 + jq_2 + kq_3 \tag{3.1}$$

and

$$|q_\infty| = U.$$

Let a_∞ denote the far-field speed of sound so that $M = \frac{U}{a_\infty} \leq 1$.

Let the disturbance potential ϕ be defined by

$$\phi = \frac{\Phi - \Phi_\infty}{U}.$$

Then the TSD is given by

$$\begin{aligned} &\frac{\partial^2\phi}{\partial t^2} + 2U\left(q_1\frac{\partial^2\phi}{\partial x\partial t} + q_2\frac{\partial^2\phi}{\partial y\partial t} + q_3\frac{\partial^2\phi}{\partial z\partial t}\right) \\ &\quad = a_\infty^2\left(1 - M^2q_1^2 - (1+\gamma)M^2q_1\frac{\partial\phi}{\partial x}\right)\frac{\partial^2\phi}{\partial x^2} \\ &\qquad + a_\infty^2\left(1 - M^2q_2^2 - (1+\gamma)M^2q_2\frac{\partial\phi}{\partial y}\right)\frac{\partial^2\phi}{\partial y^2} \\ &\qquad + a_\infty^2\left(1 - M^2q_3^3 - (1+\gamma)M^2q_3\frac{\partial\phi}{\partial z}\right)\frac{\partial^2\phi}{\partial z^2}, \\ &\qquad\qquad 0 < t, \quad -\infty < x, y, z < \infty \end{aligned} \tag{3.2}$$

where γ is the ratio of specific heats.

Typical Section Theory

We shall now exploit the high-aspect-ratio assumption on the wing as follows:

$$\frac{l}{b} \sim \infty$$

and omit the dependence on the y-coordinate in the TSD equation (3.2).

Then with α denoting the angle of attack, (3.2) becomes:

$$\begin{aligned} &\frac{\partial^2\phi}{\partial t^2} + 2U\left(\cos\alpha\frac{\partial^2\phi}{\partial x\partial t} + \sin\alpha\frac{\partial^2\phi}{\partial z\partial t}\right) \\ &\quad = a_\infty^2\left(1 - M^2\cos^2\alpha - (1+\gamma)M^2\cos\alpha\frac{\partial\phi}{\partial x}\right)\frac{\partial^2\phi}{\partial x^2} \\ &\qquad + a_\infty^2\left(1 - M^2\sin^2\alpha - (1+\gamma)M^2\sin\alpha\frac{\partial\phi}{\partial z}\right)\frac{\partial^2\phi}{\partial z^2}, \\ &\qquad 0 < t, \quad -\infty < x, \quad z < \infty. \end{aligned} \tag{3.3}$$

4 Aeroelastic boundary conditions

As is well known, the crucial part is the specification of the conditions at the structure interface—the aeroelastic boundary conditions for the TSD equation. In most of the work on aeroelasticity, being computational, field equations are approximated by ordinary differential equations and the specifications for the continuum model is never addressed. This is the case for the plate model we are considering.

There are two sets of conditions to be satisfied.

4.1 Flow tangency condition

With $h(\cdot)$ denoting *plunge* so that it is positive in a downward direction, we have that at the upper boundary

$$z = +\frac{\delta}{2}, \quad |x| < b$$

the wing z-coordinate $z(t)$ is given as

$$z(t) = \frac{\delta}{2} - h(x, y, t) - (x - ab)\theta(y, t)$$

and at

$$z = -\frac{\delta}{2}, \quad |x| < b.$$

Similarly:

$$z(t) = -\frac{\delta}{2} - h(x, y, t) - (x - ab)\theta(y, t).$$

The flow tangency condition requires that at $z = \pm\frac{\delta}{2}$ we must have

$$\left.\frac{\partial \Phi}{\partial z}\right|_{z=\pm\frac{\delta}{2}} = \frac{Dz(t)}{Dt} + \left.\frac{\partial \Phi_\infty}{\partial z}\right|_{z=\pm\frac{\delta}{2}}.$$

Hence

$$U\frac{\partial \phi}{\partial z} = \frac{\partial z(t)}{\partial t} + \nabla\Phi \cdot \nabla z \tag{4.1}$$

or,

$$U\frac{\partial \phi}{\partial z} = -\dot{h}(x, y, t) - (x - ab)\dot{\theta}(y, t) - \left(U\cos\alpha + U\frac{\partial \phi}{\partial x}\right) + \left(\frac{\partial h}{\partial x} + \theta(y, t)\right), \tag{4.2}$$

where we have also specialized to typical section theory.

On the sides $x = -b$ or $x = b$ we must have

$$\frac{\partial \phi}{\partial x} = \frac{\partial x}{\partial t} + \frac{\partial \phi}{\partial x} \cdot 1$$

which is automatically satisfied, because

$$\frac{\partial x}{\partial t} = 0.$$

Similarly, at $y = 0$ or $y = l$

$$\frac{\partial \phi}{\partial y} = \frac{\partial y}{\partial t} + \frac{\partial \phi}{\partial y} \cdot 1$$

is automatically satisfied, because

$$\frac{\partial y}{\partial t} = 0.$$

Hence, the only boundary condition to be satisfied is (4.2).

4.2 Kutta-Joukowsky conditions

It is well known that in the absence of viscosity, the Kutta-Joukowsky conditions have to be satisfied in order for a unique solution to exist of the potential equation—in our case the TSD equation. These conditions require "zero pressure jump at the trailing edge and off the wing."

Within the small disturbance assumption, the pressure $p(\cdot)$ can be expressed in terms of the linearized acceleration potential ψ (see equation (28) in [9]). In turn the pressure differential across the wing is:

$$\begin{aligned} \delta p &= p|_{z=+\frac{\delta}{2}} - p|_{z=-\frac{\delta}{2}}, \qquad |x| < b \\ &= -\rho_\infty \delta\psi \end{aligned}$$

where ρ_∞ is the free stream air density, and

$$\psi = U\left[\frac{\partial \phi}{\partial t} + U cos\alpha \frac{\partial \phi}{\partial x}\right) + U \sin\alpha \frac{\partial \phi}{\partial z}\right]$$

$$\delta\psi = \psi|_{z=\frac{\delta}{2}} - \psi|_{z=-\frac{\delta}{2}}, \qquad |x| < b.$$

Hence we have the conditions:

i) $\delta\psi = 0, \quad |x| > b$

ii) $\delta\psi = 0, \quad x \to b^-$.

5 Time-domain formulation

We can now combine the linear structure equations with the nonlinear aerodynamic equations with nonlinear boundary conditions, and look at the time-domain formulation of the total problem—this is new with this paper.

To the author's knowledge there has been no such formulation in the literature on aeroelasticity, even in linear cases, beginning with early texts. The point of departure here is to note that the TSD equation is a wave equation with an input on the boundary. Hence the classical Duhamel principle [20] applies, and we have the following form for the response:

$$\delta p(x,y,t) = \int\limits_{-b}^{b}\int\limits_{0}^{l}\int\limits_{0}^{t} P(x,y;\xi,\eta;t-\sigma;w_a(\xi,\eta,\sigma))d\xi\ d\eta\ d\sigma \tag{5.1}$$

where $w_a(\ldots)$ can be identified as the *downwash*

$$\begin{aligned} w_a(x,y,t) = -\dot{h}(x,y,t) - (x-ab)\dot{\theta}(y,t) \\ - \left(U\cos\alpha + U\frac{\partial\phi}{\partial x}\right)\left(\frac{\partial h}{\partial x} + \theta(y,t)\right). \end{aligned} \tag{5.2}$$

Let

$$\begin{vmatrix} h(\cdot,t) \\ \theta(\cdot,t) \\ \dot{h}(\cdot,\cdot,t) \\ \dot{\theta}(\cdot,t) \end{vmatrix}$$

denote the state vector for the structure.

Then we obtain the nonlinear convolution-evolution equation (extending the linear version developed in [2,3,17]):

$$\dot{x}(t) = Ax(t) + \int\limits_{0}^{t} F(x(\sigma),t-\sigma)d\sigma \tag{5.3}$$

where A is the infinitesimal generator of a semigroup as in [2] and the unique feature is the convolution term, in which the specification of the convolution operator $F(x,t)$ would be rather complex, especially to take care of the nonlinear boundary condition.

We omit the details of the precise version, but even this form is adequate to determine stability, or rather instability, flutter/divergence. For this purpose we fix M and consider the equation as a function of the parameter $U > 0$ (the convolution term is zero for $U = 0$), so that we can apply the Hopf bifurcation theory [12,19] of stability about the equilibrium state. The equilibrium solution is

$$x(t) = 0; \qquad \phi(\cdot,t) = 0$$

corresponding to the undisturbed far-stream aerodynamic variables and zero structure states. We shall do the linearization directly through the TSD equations and boundary conditions.

6 Linearization theory—the generalized Possio equation

In this section we proceed to linearize the aeroelastic equation—mainly the aerodynamic equation since the structure equations are linear already.

First we linearize the (quasilinear) TSD equation about $\phi(\cdot) = 0$. This yields

$$\begin{aligned}&\frac{\partial^2\phi}{\partial t^2} + 2U\left(\cos\alpha\frac{\partial^2\phi}{\partial x\partial t} + \sin\alpha\frac{\partial^2\phi}{\partial z\partial t}\right)\\ &\quad = a_\infty^2(1 - M^2\cos^2\alpha)\frac{\partial^2\phi}{\partial x^2} + a_\infty^2(1 - M^2\sin^2\alpha)\frac{\partial^2\phi}{\partial z^2},\\ &\quad 0 < t, -\infty < x, z < \infty, \quad \text{omitting the structure boundary.} \qquad (6.1)\end{aligned}$$

Next we linearize the boundary conditions (4.2) or (6.2).

This yields

$$\begin{aligned}&U\frac{\partial\phi}{\partial z} = -\dot{h}(x,y,t) - (x - ab)\dot{\theta}(y,t) - \left(U\cos\alpha\frac{\partial h}{\partial x} + \theta(y,t)\right),\\ &\quad z = \pm\frac{\delta}{2} \quad |x| < b. \qquad (6.2)\end{aligned}$$

In (6.1) and (6.2) we thus have a wave equation with a Neumann-type boundary condition, on part of the boundary.

The function space for the linear TSD equation will be $L_p(I\!R^2),\ 1 < p < 2$, and we note the strict inequality, < 2. As in [14], the main new feature will be the use of spatial (x, z) Fourier transforms, which have to be $L_p - L_q$ transforms and the use of the Mikhlin multiplier theory. We shall omit the technical details here.

Thus we proceed to take Laplace transforms in the time variable and Fourier transforms in x and z of the TSD equation.

Let the super hat denote the transforms. Thus

$$\hat{\phi}(x,z,\lambda) = \int_0^\infty e^{-\lambda t}\phi(x,z,t)dt, \qquad Re\lambda > \sigma_a$$

$$\hat{\hat{\phi}}(i\omega,z,\lambda) = \int_{-\infty}^\infty e^{i\omega x}\hat{\phi}(x,z,\lambda)dx, \qquad -\infty < \omega < \infty,\ |z| > \delta$$

Then we have, using the far-field vanishing conditions, that $\hat{\hat{\phi}}(\cdot)$ satisfies

$$\begin{aligned}&a_\infty^2(1 - M^2\sin^2\alpha)\frac{\partial^2\hat{\hat{\phi}}}{\partial z^2} - (2U\lambda\sin\alpha + 2U^2 i\omega\sin\alpha\cos\alpha)\frac{\partial\hat{\hat{\phi}}}{\partial z} - \cdots\\ &\quad -(\lambda^2 + \omega^2 a_\infty^2(1 - M^2\cos^2\alpha) + 2U\lambda i\omega\cos\alpha)\hat{\hat{\phi}} = 0 \qquad (6.3)\end{aligned}$$

for $|z| > \frac{\delta}{2}$, $-\infty < \omega < \infty$ with the boundary condition:

$$\begin{aligned} U\frac{\partial\hat{\phi}}{\partial z}\left(x, \frac{\delta}{2}, \lambda\right) &= U\frac{\partial\hat{\phi}}{\partial z}\left(x, -\frac{\delta}{2}, \lambda\right) \\ &= -\lambda\hat{h}(x,y,\lambda) - \lambda(x-ab)\hat{\theta}(y,\lambda) - \cdots \\ &\quad - U\cos\alpha\left(\frac{\partial\hat{h}}{\partial x} - \hat{\theta}(y,\lambda)\right), \quad |x| < b, \end{aligned} \tag{6.4}$$

where

$$\hat{h}(x,y,\lambda) = \int_0^\infty e^{-\lambda t} h(x,y,t)\,dt, \qquad Re\cdot\lambda > \sigma_a$$

$$\hat{\theta}(y,\lambda) = \int_0^\infty e^{-\lambda t}\theta(y,t)\,dt, \qquad Re\cdot\lambda > \sigma_a.$$

We go directly now to our main concern—calculation of δ_p,

$$\delta_p = -\rho_\infty \delta\psi$$

Defining

$$\hat{\psi}(x,z,\lambda) = \int_0^\infty e^{-\lambda t}\psi(x,z,t)dt$$

$$\hat{\hat{\psi}}(i\omega,z,\lambda) = \int_{-\infty}^\infty e^{i\omega x}\hat{\psi}(x,z,\lambda)dx$$

we have

$$\hat{\hat{\psi}}(i\omega,z,\lambda) = U\left[\lambda\hat{\hat{\phi}} + U\cos\alpha i\omega\hat{\hat{\phi}} + U\sin\alpha\frac{\partial\hat{\hat{\phi}}}{\partial z}\right]. \tag{6.5}$$

Now defining the *normalized frequency*

$$k = \frac{\lambda b}{U},$$

we may normalize to $b = 1$, and rewrite (6.3) as

$$a_1\frac{\partial^2\hat{\hat{\phi}}}{\partial z^2} - 2a_2\frac{\partial\hat{\phi}}{\partial z} - a_3\hat{\hat{\phi}} = 0 \tag{6.6}$$

where

$$\begin{aligned}
a_1 &= (1 - M^2 \sin^2 \alpha) \\
a_2 &= M^2 (k + i\omega \cos \alpha) \sin \alpha \\
a_3 &= k^2 M^2 + 2kM^2 i\omega \cos \alpha + \omega^2 (1 - M^2 \cos^2 \alpha).
\end{aligned}$$

Then

$$\hat{\hat{\phi}}(i\omega, z, \lambda) = A_+ e^{r_1 (z - \frac{\delta}{2})}, \quad z > \frac{\delta}{2}$$

where

$$r_1 = \frac{a_2 - \sqrt{a_2^2 + a_1 a_3}}{a_1}.$$

$Rer_1 < 0$. and

$$a_1 r_1^2 - 2a_2 r_1 - a_3 = 0$$

is a solution of (6.5) with

$$\hat{\hat{\phi}}(i\omega, z, \lambda) \to 0 \quad \text{as} \quad z \to \infty.$$

We now choose A_+ to satisfy the boundary condition:

$$\left. \frac{\partial \hat{\phi}}{\partial z} \right|_{z = \frac{\delta}{2}} = r_1 A_+$$

and hence

$$\frac{\partial \hat{\phi}}{\partial z} = r_1 \hat{\hat{\phi}}(i\omega, z, \lambda), \quad z > \frac{\delta}{2}.$$

Similarly, we can calculate that

$$\frac{\partial \hat{\phi}}{\partial z} = r_2 \hat{\hat{\phi}}(i\omega, z, \lambda), \quad z < \frac{-\delta}{2}$$

where

$$r_2 = \frac{a_2 + \sqrt{a_2^2 + a_1 a_2}}{a_1}, \quad \Re \cdot r_2 > 0.$$

Correspondingly, we have from (6.4):

$$\begin{aligned}
\hat{\hat{\phi}}(i\omega, z, \lambda) &= U[\lambda + Ui\omega \cos \alpha + r_1 U \sin \alpha] \hat{\hat{\phi}}(i\omega, z, \lambda), \quad z \geq \frac{\delta}{2} \\
&= U[\lambda + Ui\omega \cos \alpha + r_2 U \sin \alpha] \hat{\hat{\phi}}(i\omega, z, \lambda), \quad z \leq -\frac{\delta}{2}
\end{aligned}$$

Hence we have that

$$\delta\hat{\hat{\psi}}(i\omega,\lambda) = \hat{\hat{\psi}}\left(i\omega,\frac{\delta}{2},\lambda\right) - \hat{\hat{\psi}}\left(i\omega,-\frac{\delta}{2},\lambda\right).$$

Now

$$\hat{\hat{\phi}}\left(i\omega,\frac{\delta}{2},\lambda\right) = \frac{\frac{\partial\hat{\hat{\phi}}}{\partial z}\left(i\omega,\frac{\delta}{2},\lambda\right)}{r_1}$$

and hence

$$\hat{\hat{\psi}}\left(i\omega,\frac{\delta}{2},\lambda\right) = U\left(\frac{\lambda+i\omega\cos\alpha}{r_1}+U\sin\alpha\right)\frac{\partial\hat{\hat{\phi}}}{\partial z}\left(i\omega,\frac{\delta}{2},\lambda\right). \qquad (6.7)$$

Similarly,

$$\hat{\hat{\phi}}\left(i\omega,-\frac{\delta}{2},\lambda\right) = \frac{\frac{\partial}{\partial z}\hat{\hat{\phi}}\left(i\omega,-\frac{\delta}{2},\lambda\right)}{r_2}$$

and hence

$$\hat{\hat{\psi}}\left(i\omega,-\frac{\delta}{2},\lambda\right) = U\left(\frac{\lambda+i\omega\cos\alpha}{r_2}+U\sin\alpha\right)\frac{\partial\hat{\hat{\phi}}}{\partial z}\left(i\omega,-\frac{\delta}{2},\lambda\right). \qquad (6.8)$$

From (6.4) following the flow tangency condition we have that

$$\begin{aligned} U\frac{\partial\hat{\phi}}{\partial z}\left(x,\frac{\delta}{2},\lambda\right) &= U\frac{\partial\hat{\phi}}{\partial z}\left(x,-\frac{\delta}{2},\lambda\right), \quad |x|<1 \\ &= \hat{\omega}_a(x,y,\lambda). \end{aligned}$$

We now seek a solution to (6.5) such that

$$\frac{\partial\hat{\phi}}{\partial z}\left(x,\frac{\delta}{2},\lambda\right) = \frac{\partial\hat{\phi}}{\partial z}\left(x,-\frac{\delta}{2},\lambda\right)$$

for all x, $-\infty < x < \infty$, which holds automatically if $a_2 = 0$, corresponding to $\alpha = 0$, because of the resulting symmetry in z.

Let us use the notation

$$\hat{\nu}(i\omega,\lambda) = \frac{\partial\hat{\hat{\phi}}}{\partial z}\left(i\omega,+\frac{\delta}{2},\lambda\right) = \frac{\partial\hat{\hat{\phi}}}{\partial z}\left(i\omega,-\frac{\delta}{2},\lambda\right).$$

Then

$$\begin{aligned} \delta\hat{\hat{\psi}}(i\omega,\lambda) &= \hat{\hat{\psi}}\left(i\omega,+\frac{\delta}{2},\lambda\right) - \hat{\hat{\psi}}\left(i\omega,-\frac{\delta}{2},\lambda\right) \\ &= U(\lambda+i\omega\cos\alpha)\left(\frac{1}{r_2}-\frac{1}{r_1}\right)\hat{\nu}(i\omega,\lambda) \end{aligned}$$

or,

$$\hat{\nu}(i\omega,\lambda) = \frac{\delta\hat{\hat{\psi}}(i\omega,\lambda)}{U(\lambda+i\omega\cos\alpha)\Big(\frac{1}{r_2}-\frac{1}{r_1}\Big)}. \tag{6.9}$$

Next we define the Kussner doublet function

$$A(x,t) = -\frac{2}{U}\delta\psi(x,t), \quad |x|<1$$

and the Laplace transform:

$$\hat{A}(x,\lambda) = \int\limits_0^\infty e^{-\lambda t}A(x,t)dt, \quad Re\cdot\lambda > \sigma_a$$

and define the function $P(\cdot)$ by

$$\int\limits_{-\infty}^{\infty} P(x)e^{-i\omega x}dx = \frac{1}{k+i\omega\cos\alpha}\frac{1}{\left(\frac{1}{r_2}-\frac{1}{r_1}\right)} = m(M,k,i\omega)$$

$$= \frac{1}{2(k+i\omega\cos\alpha)}\cdot\frac{M^2k^2+2M^2ki\omega\cos\alpha+\omega^2(1-M^2\cos^2\alpha)}{\sqrt{M^2k^2+2M^2ki\omega\cos\alpha+\omega^2(1-M^2)}}. \tag{6.10}$$

Then (6.8) yields

$$\hat{\omega}_a(x,y,\lambda) = \int\limits_{-1}^{1} P(x-\xi)\hat{A}(\xi,\lambda)d\xi, \quad |x|<1, \tag{6.11}$$

where

$$\hat{\omega}_a(x,y,\lambda) = -\lambda\hat{h}(x,y,\lambda) - (x-a)\lambda\hat{\theta}(y,\lambda) + Ucos\alpha\left(-\frac{\partial\hat{h}}{\partial x} - \hat{\theta}(y,\lambda)\right), \tag{6.12}$$

which is recognized as a generalization for a nonzero angle of attack of the Possio integral equation [21].

We assume that (6.11) has a unique solution with $\hat{A}(x,\lambda)\to 0$ as $x\to 1$, with the function space formulation for $\alpha = 0$ given in [ ].

We note that the change from the beam model is exemplified by the appearance of the term $\frac{\partial\hat{h}}{\partial x}$ in (6.12), corresponding to the chordwise bending slope, or *camber bending.*

We note also that the Possio equation provides the link from the aerodynamics to the structure dynamics which is often the most mysterious part of computational aeroelasticity.

In the following we shall use $\hat{A}(x, y, \lambda)$ in place of $\hat{A}(x, \lambda)$ in view of the dependence on the span variable on the left-hand side of (6.11).

7 Linearized aeroelastic equations—flutter speed

We can now state the linearized aerostructure equations in the Laplace transform version where we set the initial condition of the structure to zero because our concern is only stability.

$$\lambda^2 m\hat{h}(x, y, \lambda) + D\Delta^2\hat{h}(x, y, \lambda) + \lambda^2 \frac{S}{2b}\hat{\theta}(y, \lambda) = \rho U \hat{A}(x, y, \lambda) \tag{7.1}$$

$$\lambda^2 I_\theta \hat{\theta}(y, \lambda) + \lambda^2 S\hat{h}(x, y, \lambda) - GJ\frac{d^2\hat{\theta}(y, \lambda)}{dy^2} = \rho U \int_{-1}^{1} (x - a)\hat{A}(x, y, \lambda)dx \tag{7.2}$$

with the CFFF boundary conditions for the plate and

$$\hat{\theta}(0, \lambda) = 0 = \hat{\theta}'(l, \lambda).$$

Also we shall use the notation

$$\hat{A}(\cdot, \cdot, \lambda) = \psi(M, k, \alpha)\hat{\omega}_a(\cdot, \cdot, \lambda)$$

to denote the solution of the Possio equation in the operator form.

Definition of Flutter Speed

Putting $U = 0$ in (7.1) and (7.2) we have the pure structure equations:

$$\lambda^2 m\hat{h}(x, y, \lambda) + D\Delta^2\hat{h}(x, y, \lambda) + \lambda^2 \frac{S}{2b}\hat{\theta}(y, \lambda) = 0$$
$$\lambda^2 I_\theta \hat{\theta}(y, \lambda) + \lambda^2 S\hat{h}(x, y, \lambda) + GJ\frac{d^2\hat{\theta}(y, \lambda)}{dy^2} = 0.$$

For $S = 0$ we would have uncoupled, undamped plate modes and pitch modes, and for small S there would be some mode coupling, which we neglect, as in the beam case [3].

As in [3] we study the *root locus*: take a particular mode, say λ_j, as a function of U, denoted $\lambda_j(U)$ with

$$\lambda_j(0) = i\omega_j.$$

Our first result is the following theorem.

Theorem 7.1

For each j we have

$$\left.\frac{\partial\lambda_j(0)}{\partial U}\right|_{U=0} \textit{ is real, negative}$$

extending the similar result for the beam.

Proof

Substituting

$$k = \frac{\lambda b}{U}$$

in (6.10) and letting $U \to 0$, we have

$$\lim_{U\to 0} = m\left(M, \frac{\lambda b}{U}, i\omega\right) \to \frac{M}{2}$$

and hence

$$\lim_{U\to 0} \hat{A}(x,y,\lambda) = \frac{2}{M}\omega_a(x,y,\lambda).$$

Hence, for the plate mode we have differentiating with respect to U in (7.1), where we neglect the change in mode shape,

$$2\lambda m\hat{h}(x,y,\lambda)\left.\frac{\partial\lambda}{\partial U}\right|_{U=0} = \rho\hat{A}(x,y,\lambda) = \frac{2\rho}{M}(-\lambda\hat{h}(x,y,\lambda)).$$

Hence

$$\left.\frac{\partial\lambda}{\partial U}\right|_{U=0} = -\frac{\rho}{mM}, \quad 0 < M \leq 1.$$

For the pitching mode, because we are taking $S = 0$, we may use the result derived for the uncoupled beam case [3]. The angle of attack plays no role in the limit $U \to 0$. Hence

$$\left.\frac{\partial\lambda}{\partial U}\right|_{U=0} = -2\rho\frac{(a^2+\frac{1}{3})}{MI_\theta}.$$

This enables us to define *flutter speed.*

The graph of $Re \cdot \lambda_j(U)$ versus U is known as the *stability curve*, and the lope at $U = 0$ is strictly negative.

Hence, the speed U_{F_j}, at which the *upward* curve crosses the U-axis,

$$Re \cdot \lambda_j(U) = 0 \quad U = U_{F_j}$$

$$Re \cdot \lambda_j(U) < 0 \quad U < U_{F_j}$$

$$Re \cdot \left(\frac{\partial\lambda_j(U)}{\partial U}\right) > 0 \quad U = U_{F_j}$$

is called the flutter speed corresponding to the j^{th} mode. It may not exist for every mode. By the flutter speed we mean

$$\min_j U_{F_j}$$

Of course we are interested only in the first few modes, the only ones to have any physical significance.

8 Stationary solution—divergence speed

By the stationary solution of the aeroelastic equation we mean the solution when all the time derivatives are set equal to zero. This problem has been studied a good deal in transonic aerodynamics [see 2,11,17], but not as much in transonic aeroelasticity with respect to aeroelastic boundary conditions see Section 4. For a given M, a solution may not exist, except perhaps for a sequence of values of U, the smallest of which is then called the *divergence speed.*

In terms of the linear or linearized equations, this means that we set $\lambda = 0$ in the Laplace transform version. This is what we shall consider first. Taking $\lambda = 0$ in Equations (7.1) and (7.2) we have

$$D\Delta^2 \hat{h}(x, y, 0) = \rho U \hat{A}(x, y, 0) \tag{8.1}$$

$$-GJ \frac{d^2\hat{\theta}(y, 0)}{dy^2} = \rho U \int_{-1}^{1} (x - a)\hat{A}(x, y, 0)dx \tag{8.2}$$

For simplicity of notation we shall now write

$$\begin{aligned} h(x, y) &\quad \text{for} \quad \hat{h}(x, y, 0) \\ \theta(y) &\quad \text{for} \quad \hat{\theta}(y, 0) \\ A(x, y) &\quad \text{for} \quad \hat{A}(x, y, 0) \end{aligned}$$

Then following [18] we have:

$$\rho U A(\cdot, \cdot) = \frac{2\sqrt{1 - M^2}}{1 - M^2 \cos^2 \alpha} U^2 \rho \left(-\mathcal{T} \frac{\partial h}{\partial x} - \theta(\cdot) g_1(\cdot) \right)$$

where

$$g_1(x) = \sqrt{\frac{1 - x}{1 + x}}, \quad |x| < 1$$

$$\mathcal{T} f = g; \quad g(x) = \frac{1}{\pi} \sqrt{\frac{1 - x}{1 + x}} \int_{-1}^{1} \sqrt{\frac{1 + \xi}{1 - \xi}} \frac{f(\xi)}{\xi - x} d\xi, \quad |x| < 1$$

$\mathcal{T}\frac{\partial h}{\partial x}$ is the function

$$\frac{1}{\pi}\sqrt{\frac{1-x}{1+x}}\int_{-1}^{1}\sqrt{\frac{1+\xi}{1-\xi}}\frac{1}{\xi-x}\frac{\partial}{\partial\xi}h(\xi,y)d\xi, \quad |x|<1,$$

which is convenient to denote by $L_h h$, where L_h is closed linear, mapping a dense domain of $L_2[\Omega]$, which includes the range of Δ^2, into $L_p[\Omega]$, $1 \leq p < 2$, where $\Omega = [-1,1] \times [0,l]$.

Let

$$c^2 = 2\rho\frac{\sqrt{1-M^2}}{1-M^2\cos^2\alpha}\cos^2\alpha \tag{8.3}$$

Then (8.1) can be expressed as

$$D\Delta^2 h = c^2U^2(L_h h - \theta(\cdot)g_1(\cdot)). \tag{8.4}$$

Similarly, let L_θ denote the closed linear mapping with a dense domain in $L_2[\Omega]$, which includes the range of Δ^2 and the range in $L_2[0,l]$ defined by

$$\begin{aligned} &L_\theta h = g; \\ &g(y) = \int_{-1}^{1}(s-a-1)\sqrt{\frac{1+s}{1-s}}\,\frac{\partial h}{\partial s}(s,y)ds, \quad 0<y<e \end{aligned}$$

Next

$$\int_{-1}^{1}(x-a)g_1(x)dx = -\frac{\pi}{2}(1+2a) < 0$$

while

$$\begin{aligned} &\int_{-1}^{1}(x-a)\frac{1}{\pi}\sqrt{\frac{1-x}{1+x}}dx\int_{-1}^{1}\sqrt{\frac{1+s}{1-s}}\frac{\partial}{\partial s}h(s,y)\frac{ds}{s-x} \\ &\qquad = \int_{-1}^{1}(s-a-1)\sqrt{\frac{1+s}{1-s}}\frac{\partial}{\partial s}h(s,y)ds \end{aligned}$$

which is recognized as

$$L_\theta h$$

Hence we have for the right side of (8.2):

$$\rho U\int_{-1}^{1}(x-a)A(x,\cdot)dx = c^2U^2\left(L_\theta h + \frac{\pi}{2}(1+2a)\theta\right)$$

or we have

$$A_\theta \theta = \frac{c^2U^2}{GJ}\left(L_\theta h + \frac{\pi}{2}(1+2a)\theta\right). \tag{8.5}$$

The problem is now to determine the values of U for which the coupled Equations (8.4) and (8.5) are rewritten as

$$\Delta^2 h - \frac{c^2U^2}{D}L_h h = -\frac{c^2U^2}{D}g_1(\cdot)\theta(\cdot) \tag{8.6}$$

$$A_\theta\theta - \frac{c^2U^2}{GJ}\frac{\pi}{2}(1+2a)\theta = \frac{c^2U^2}{GJ}L_\theta h \tag{8.7}$$

and can be solved.

First we can omit values of U for which

$$A_\theta\theta = \frac{c^2U^2}{GJ}\frac{\pi}{2}(1+2a)\theta.$$

This is a sequence of speeds for $\lambda = 0$ for the beam model of the wing; see [3].

Note that S plays no role in the determination of the divergence speed (see [10] for more on the dependence of the angle of attack α and the transonic dip on the divergence speed as a function of M).

The existence of the solution of the linearized equations may be shown to imply the same for the nonlinear equations following the "analyticity" arguments as in [14].

Similarly, we may omit values of U such that

$$\Delta^2 h = \frac{c^2U^2}{D}L_h h$$

or equivalently

$$\frac{D}{c^2U^2}h = \Delta^{-2}L_h h.$$

Here $\Delta^{-2}L_h$ is compact and denoting the sequence of eigenvalues by p_n, we see that

$$U_n = \frac{D}{c^2p_n}$$

yields the corresponding sequence of speeds and may be recognized as the sequence of plate divergence speeds, if in the plate model, the pitch angle dynamics is deleted.

If we omit these sequences, then

$$\left(\Delta^2 - \frac{c^2U^2}{D}L_h\right)$$

has a bounded inverse

$$\left(\Delta^2 - \frac{c^2U^2}{D}L_h\right)^{-1} = \left(I - \frac{c^2U^2}{D}\Delta^{-2}L_h\right)^{-1}\Delta^{-2}$$

and is compact. Similarly,

$$\left(A_\theta - \frac{c^2U^2}{GJ}\frac{\pi}{2}(1+2a)I\right)$$

has a bounded inverse

$$\left(A_\theta - \frac{c^2U^2}{GJ}\frac{\pi}{2}(1+2a)I\right)^{-1} = \left(I - \frac{c^2U^2}{GJ}\frac{\pi}{2}(1+2a)A_\theta^{-1}\right)^{-1}A_\theta^{-1}$$

and is compact. Hence we can rewrite (8.6) and (8.7) as

$$-\frac{D}{c^2U^2}h = \left(I - \frac{c^2U^2}{D}\Delta^{-2}L_h\right)^{-1}\Delta^{-2}(g_1(\cdot)\theta(\cdot)) + \cdots \tag{8.8}$$

$$+\frac{c^2U^2}{aJ}\theta = \left(I - \frac{c^2U^2}{GJ}\frac{\pi}{2}(1+2a)A_\theta^{-1}\right)^{-1}A_\theta^{-1}L_\theta h \cdots \tag{8.9}$$

The first simplification here is to note that for a small a—equivalently n large so that we are in the transonic range—we omit the term containing c^2 and approximate

$$\left(I - \frac{c^2U^2}{D}\Delta^{-2}L_h\right) \approx I$$

$$\left(I - \frac{c^2U^2}{aJ}\frac{\pi}{2}(1+2a)A_\theta^{-1}\right) \approx I$$

so that we have

$$-\frac{D}{c^2U^2}h = \Delta^{-2}(g_1(\cdot)\theta(\cdot))$$

$$\frac{aJ}{c^2U^2}\theta = A_\theta^{-1}L_\theta h$$

combining, which we have:

$$\frac{DaJ}{c^4U^4}\theta = A_\theta^{-1}L_\theta\Delta^{-2}(g_1(\cdot)\theta(\cdot)) \tag{8.10}$$

or

$$\frac{DaJ}{c^4U^4}h = \Delta^{-2}(g_1(\cdot)A_\theta^{-1}L_\theta h). \tag{8.11}$$

The operators on the right being compact, and not dependent on aeroelastic constants, this yields a sequence of values

$$U_n^4 = -\frac{DaJ}{c^4\gamma_n}$$

where

$$\gamma_n\theta_n = A_\theta^{-1}L_\theta\Delta^{-2}(g_1(\cdot)\theta_n(\cdot)).$$

Because $|\gamma_n| \to 0$, we see that U_n goes to infinity. For more on this and the transonic dip, as well as the case $M = 1$, see [20].

Returning to (8.8) and (8.9), we may combine them (or solve for θ, h) as:

$$-\frac{DGJ}{c^4U^4}h = \left(I - \frac{c^2U^2}{D}\Delta^{-2}L_h\right)^{-1}\Delta^{-2}(g_1(\cdot)$$
$$\times\left(I - \frac{c^2U^2}{aJ}\frac{\pi}{2}(1+2a)A_\theta^{-1}\right)^{-1}A_\theta^{-1}L_\theta h) \tag{8.12}$$

$$-\frac{DGJ}{c^4U^4}\theta = \left(I - \frac{c^2U^2}{aJ}\frac{\pi}{2}(1+2a)A_\theta^{-1}\right)^{-1}A_\theta^{-1}L_\theta$$
$$\times\left(I - \frac{c^2U^2}{D}\Delta^{-2}L_h\right)^{-1}\Delta^{-2}(g_1(\cdot)\theta(\cdot)) \tag{8.13}$$

We can express (8.12) as a function of U as

$$DGJh = U^4T(U)h$$

or because h is a function of U, as

$$T(U)h(U) - h(U) = 0, \quad \|h(U)\| \neq 0.$$

Now $T(U)$ can be extended to the analytic function in the complex plane, and it follows there can be at most a countable sequence $\{U_n\}$ of zero, and further that

$$|U_n| \to \infty \quad \text{as} \quad n \to \infty.$$

We are only interested in $U_n > 0$, and

$$U_n \to \infty \quad \text{as} \quad n \to \infty.$$

This is as far as we shall go into this problem in this paper; of course we will not go into techniques of calculating the sequence $\{U_n\}$. Finally, the nonlinear stationary problem will have a solution of the linearized version by the same kind of analycity assumption as in the beam case [10].

9 Conclusions

A time-domain continuum theory for a Kirchoff thin plate structure model and the transonic small difference potential equation for aerodynamics, with appropriate boundary interface conditions, has been presented. The flutter and divergence speeds are shown to be determined by the linearized equations, the solution of which involves the Possio integral equation generalized to allow a nonzero angle of attack. The emphasis is on problem formulation. The calculation of the flutter stability curve now involves the solution of a partial differential equation, in contrast to an ordinary differential equation in the beam model. The stationary solutions for calculating the divergence speed are more regular as M goes to 1 in contrast to the case of a zero angle of attack, but now the calculation involves an eigenvalue problem for linear operators.

Acknowledgment

I am indebted to Prof. Earl H. Dowell for helpful discussions on the plate model and the aeroelastic boundary conditions. Research was supported in part under NSF Grant No. ECS-0400730.

References

[1] P.J. Attar, E.H. Dowell, and J.R. White, *Modelling the LCO of a Delta Wing Using a High Fidelity Structural Model*, 1-19, AIAA, Preprint (2005).

[2] A.V. Balakrishnan, Subsonic flutter suppression using self-straining actuators, *Journal of the Franklin Institute*, vol. **138** (2001), 149–170.

[3] A.V. Balakrishnan, Possio integral equation of elasticity theory, *Journal of AeroSpace Engineering*, October (2003), 139–154.

[4] A.V. Balakrishnan, *A Convolution Evolution Equation in Aeroelasticity*, Progress in Nonlinear Differential Equations and Their Applications, vol. 55, Birkhäuser, Basel (2003), 61–82.

[5] A.V. Balakrishnan, The transonic small disturbance potential equation, *AIAA Journal*, vol. **42**, no. 6, June (2004), 1081–1088.

[6] A.V. Balakrishnan, On the transonic divergence speed for nonzero wing thickness, Preprint, unpublished.

[7] A.V. Balakrishnan and K. Iliff, A continuum aeroelastic model for inviscid subsonic bending-torsion wing flutter, Proceedings of the International Forum on Aeroelasticity and Structural Dynamics, Amsterdam (2003).

[8] J.T. Batina, Efficient algorithm for solution of the unsteady transonic small disturbance equation, *Journal of Aircraft*, vol. **25**, July (1988), 598–605.

[9] J.T. Batina, *A Finite Difference Approximate Factorization Algorithm for the Solution of the TSD Equation*, NASA TP 3129, January (1992).

[10] O. Bendiksen, Transonic Flutter Dynamical, AIAA paper 2002-1488, Denver, CO (April 2002).

[11] L. Bers, *Mathematical Aspects of Subsonic and Transonic Gas Dynamics*, Surveys in Applied Mathematics III, John Wiley & Sons, New York (1958).

[12] A.J. Chorin and J.E. Marsden, *A Mathematical Introduction to Fluid Mechanics*, 3rd edition, Springer-Verlag, Berlin (2000).

[13] J.D. Cole and J. Kevorkian, *Perturbation Methods in Applied Mathematics*, vol. 34, Springer-Verlag, Heidelberg (1981).

[14] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. 2, InterScience, New York (1966).

[15] E.H. Dowell, Nonlinear oscillations of a fluttering plate, *AIAA Journal*, vol. **4**, no. 7, July (1966), 1267–1275.

[16] E.H. Dowell, *Aeroelasticity of Plates and Shells*, Noordhoof International Publishing, Leyden (1975).

[17] C. Ferrari and F.G. Tricomi, *Transonic Aerodynamics*, Academic Press, New York (1968).

[18] A.W. Leissa, *Vibration of Plates*, Office of Technology Utilization, NASA, Washington, D.C. (1969).

[19] J.E. Marsden and M. McCracken, *The Hopf Bifurcation*, Springer Applied Mathematics Series, vol. 19, Springer-Verlag, Heidelberg (1976).

[20] D. Nixon, Basic equation for unsteady transonic flow, *Unsteady Transonic Aerodynamics*, edited by D. Nixon, vol. 120, Progress in Aeronautics and Astronautics, IAA, Washington, D.C. (1989), 57–73.

[21] D. Tang and E.H. Dowell, Effects of attack on nonlinear flutter of a delta wing, *AIAA Journal*, vol. **39**, no. 1, January (2001), 15–21.

[22] S. Timoshenko and S. Woinowsky-Kreiger, *Theory of Plates and Shells*, McGraw-Hill Book Company, New York (1959).

CHAPTER 12

Differential Riccati equations for the Bolza problem associated with point boundary control of singular estimate control systems

Irena Lasiecka
Department of Mathematics, University of Virginia,
Charlottesville, Virginia

Amjad Tuffaha
Department of Mathematics, University of Virginia,
Charlottesville, Virginia

1 Introduction

Recent developments in boundary and point control theory have provided strong motivation for studying Riccati equations with unbounded coefficients. By unbounded coefficients we mean the coefficients of the Riccati equations that are given in terms of unbounded, and even uncloseable, operators. These typically result from unbounded control actions or unbounded observations. The unboundedness of the coefficients is, of course, a source of mathematical difficulties. Standard methods for establishing well-posedness to these nonlinear equations are no longer applicable. The problem is particularly acute when the coefficients in the *nonlinear term of the Riccati equation are unbounded*. This latter case corresponds to the unbounded control action. In these cases, the well-posedness of Riccati equations may fail altogether, as evidenced in [21]. However, the situation is very different when the dynamics is generated by an analytic semigroup. This is to say that e^{At} is a generator of an analytic semigroup on a given Hilbert space H. For this class of dynamics the Riccati theory is by now well understood. The regularizing effect of analyticity compensates for the unboundedness of the coefficients, leading to a well-posed nonlinear problem.

On the other hand, analytic semigroups are rather special and are associated with parabolic or paraboliclike behavior. Instead, in many real-life and physically significant applications one must deal with both: nonanalytic dynamics and unbounded control–observation action. This raises the issue of feedback control synthesis and the corresponding Riccati theory available for the new classes of nonanalytic problems. A class of models that naturally emerges in this context is the so-called class of *Singular Estimate Control Systems* (SECS). A SECS system is driven by a nonanalytic semigroup $e^{At} \in \mathcal{L}(H)$ with unbounded control operator $B : U \to [D(A^*)]'$, which is assumed to satisfy the following singular estimate:

$$|e^{At}B|_{\mathcal{L}(U,H)} \leq \frac{c}{t^\gamma}; \quad t > 0; \quad 0 \leq \gamma < 1 \tag{1.1}$$

It turns out that SECS systems, being a natural generalization of analytic systems, arise naturally in the context of coupled PDE dynamics. The typical configuration is that of a coupled system that consists of an analytic (parabolic) component and a hyperbolic component interacting (often via boundary condition) at some lower dimensional manifold called the interface.

For such systems it is expected that analytic effects will diffuse from one region to another, providing some overall smoothing effect. This smoothing effect is expressed by the singular estimate (1.1). In fact, for many such models it is possible to show that the analytic effect is propagated into the hyperbolic region preserving the singular estimate feature for the entire (nonanalytic) structure.

The study of SECS systems has attracted considerable attention in recent years. Problems such as controllability, stabilization, and optimal control for this class of systems have been studied with many results (see [13,14,22] and references therein). In particular, a systematic Riccati theory for finite and infinite horizon problems has been developed. For this class of problems it has been shown that not only the associated Riccati equation with unbounded coefficients is well posed and admits a unique positive selfadjoint solution $P(t) \in \mathcal{L}(H)$, but also that the optimal feedback operator $B^*P(t)$ is bounded $H \to U$. This last property displays some smoothing effect of the dynamics. In the case of analytic semigroups, one has a much stronger smoothing effect propagated into solutions to the Riccati equation: $A^*P(t) \in \mathcal{L}(H)$. However, this latter property fails to be true for SECS systems.

The results reported above refer to Riccati equations associated with the *zero terminal* condition, which does not account for the final state penalization. On the other hand, it is well known that the presence of the nonsmooth terminal condition in unbounded control systems brings forward new phenomena and changes the Riccati theory drastically. This is true even in the case of analytic semigroups [1,10,15]. Just the existence alone of an optimal solution requires certain conditions of closeability, which do not appear at all in *standard* theories (see [15]) and references therein). In fact, it is known that these conditions are *necessary* when dealing with unbounded control actions and final state nonsmooth penalization (the so-called Bolza problem). It is thus expected that the Bolza problem within the context of SECS systems will display more singular behavior as well. It is a goal of this paper to provide a complete study of the Bolza problem for the SECS class of systems with unbounded coefficients.

We shall consider the Bolza problem defined for PDE dynamics with unbounded (point or boundary) controls. The assumption made on the system is that it is a SECS system that is, the pair A, B satisfies the singular estimate condition; see Section 2.3. The novelty of the work presented, with respect to the literature, is that the nonsmooth final state penalization problem will be considered in the context of SECS systems. We will show that although the Riccati equation is well posed for this class of systems, the gain operator exhibits integrable singularity. In addition, we will show that the finite time *transfer* function for the controlled system exhibits double singularity: one at the origin and another one at the terminal time. While the singularity of the *transfer* function at the origin has been known in SECS systems (see [13,14]), the singularity at the end point is new and caused by the effect of the terminal condition for the Riccati equation. This, in turn, propagates backward, with double intensity, the singularity arriving from the origin. The theory developed will be applied to a boundary-point control problem associated with a structural acoustic model.

2 Mathematical model and main results

2.1 Description of the problem

Let U, Y, Z and W be given Hilbert spaces. U and Y denote, respectively, control and state spaces while Z and W are observation spaces. We consider the dynamics governed by the state equation with a state $y(t) \in Y$ and control $u(t) \in U$:

$$y_t = Ay + Bu; \quad on\ [\mathcal{D}(A^\star)]'; \quad y(s) = y_s \in Y \tag{2.2}$$

Here the operator A is a generator of a strongly continuous semigroup on Y and the control operator B is assumed to satisfy $R(\lambda, A)^{-1}B \in \mathcal{L}(U, Y)$ for some $\lambda > 0$. With the above assumption, for each $t > 0$, the state equation is defined on a dual space $[\mathcal{D}(A^\star)]'$ ([13]–[15]). In particular, the operator $L_{T,s} : L_2([s, T]; U) \to Y$ defined by

$$L_{T,s}u \equiv \int_s^T e^{A(T-z)}Bu(z)dz \tag{2.3}$$

is densely defined and closeable. In addition, it is also bounded when considered as acting between the spaces: $L_{T,s} : L_2([s,T];U) \to [\mathcal{D}(A^\star)]'$. Next we introduce the observation operators $R \in \mathcal{L}(Y,W), G \in \mathcal{L}(Y,Z)$. With this notation we define the functional cost over the finite time interval $[s,T]$:

$$J(u,y,s,y_s) = \int_s^T |Ry(t)|_W^2 + |u(t)|_U^2 dt + |Gy(T)|_Z^2 \tag{2.4}$$

The control problem considered is to minimize $J(u,y,s,y_s)$ subject to the state Equation (2.2) over all $u \in L_2([s,T];U)$.

Because J is convex and $0 \in \mathcal{D}(J)$, in order to obtain the existence and uniqueness of an optimal solution, it suffices to establish lower semicontinuity of the functional $J(u,y(u),y_s,s)$ given by (3.4). This issue would be trivial if not for the finite time penalization. Indeed, because $Ry(u)$ is densely defined and closeable: $L_2([s,T];U) \to L_2([s,T],W)$, the corresponding term in the functional cost is lower-semicontinuous. This, however, may not be the case with $GL_{s,T}u$ term, unless GL_T is closeable: $L_2([s,T],U) \to Z$. Indeed, a series of counterexamples can be given (see [15]) showing that the *finite rank* operators G produce uncloseable operators $GL_{s,T}$, which in turn lead to the *nonexistence* of an optimal solution. In view of the above, a necessary and sufficient condition for the existence (and uniqueness) of an optimal solution is the requirement that GL_T be closeable. We note that this condition is trivially satisfied when B is bounded $U \to Y$ or when G has a bounded inverse: $Z \to Y$.

Also, special cases when G satisfies additional regularity properties (the so-called *smoothing case*, [15]) provide a natural class of operators G complying with the above hypothesis. However, we shall not dwell on this as our focus here is on singular problems.

Once the conditions for solvability of the optimization problem are set, our goal in this work is to express optimal control in a feedback form via solution of an appropriate Riccati equation. Thus, the solvability of associated Riccati equations is the major theme of this paper. We begin by formulating precise assumptions to be imposed on the problem studied.

2.2 Assumptions

The distinctive feature of the model is the fact that the control operator B is allowed to be unbounded. This brings forward an array of well-recognized technical difficulties, including the fact that the Riccati feedback operator may not necessarily be densely defined on $\mathcal{D}(A)$ ([7,8,15]). In this paper we focus on the Singular Estimate Control Systems (SECS) class of systems. In short, we shall study the problem under the following set of assumptions imposed on the dynamics:

2.3 Assumption

- A is a generator of a strongly continuous semigroup denoted by e^{At} on the Hilbert space Y.
- The control operator B is a linear operator from $U \to [\mathcal{D}(A^\star)]'$, satisfying the condition $R(\lambda, A)B \in \mathcal{L}(U, Y)$, for some $\lambda \in \rho(A)$. Without loss of generality, we can assume that $\lambda = 0$ and hence $A^{-1}B \in \mathcal{L}(U; Y)$, where $R(\lambda, A)$ is the resolvent of A and $\rho(A)$ is the resolvent set.
- The Singular Estimate Control (SEC) condition: There exists $\gamma < 1$ and a constant $C > 0$ such that $|e^{At}Bu|_Y \leq \frac{C}{t^\gamma}|u|_U$ for all $0 < t \leq 1$.
- Terminal time penalization operator: G is a bounded linear operator from Y to W such that the operator $GL_T : L_2([0,T]; U) \to Z$ is closeable.
- R is a bounded linear operator from Y to W.

Remark

We note that the present setup includes, as a special case, analytic semigroup models with unbounded control operators (see [8] and [15] and references therein). Indeed, in this latter case (mentioned above) the validity of the singular estimate follows from the assumption (imposed by the theory) that $R(\lambda, A)^{-\gamma}B \in \mathcal{L}(U, Y)$ and from analyticity of the semigroup, where the latter requirement implies $|A^\gamma e^{At}|_{\mathcal{L}(Y)} \leq Ct^{-\gamma}$. However, the class of SEC systems substantially extends the dynamical systems governed by analytic semigroups. It includes classes of models that may have dominant hyperbolic character. Typical prototypes are coupled systems exhibiting both parabolic and hyperbolic characteristics. These include structural acoustic systems, fluid structure interactions, composite plates, and thermal plates with boundary or point controls ([13]).

2.4 Main results

Theorem 2.1

Under the assumption in Section 2.3, for any initial state $y_s \in Y$ there exists a unique optimal control $u^0(t, s, y_s) \in L_2([s,T]; U)$ and optimal trajectory $y^0(t, s, y_s) \in L_2([s,T]; Y)$ such that $J(u^0, y^0, s, y_s) = min_{u \in L_2([s,T],U)}$ $J(u, y(u), s, y_s)$. Moreover, there exists a self-adjoint positive operator $P(t) \in \mathcal{L}(Y), t \in [0,T])$ such that $(P(t)x, x)_Y = J(u^0, y^0, t, x)$.

In addition, the following properties hold:

1. *The optimal control $u^0(t)$ is continuous on $[s,T])$ but has a singularity of order gamma at the terminal time. More specifically the following*

estimate holds:

$$|u^0(t,s,y_s)|_U \leq \frac{C}{(T-t)^\gamma}, \quad s \leq t < T$$

2. *The optimal output $y^0(t)$ is continuous on $[s,T]$ when $\gamma < 1/2$, but has a singularity of order $2\gamma - 1$ at the terminal time when $\gamma \geq 1/2$. The following estimate holds:*

$$|y^0(t,s,y_s)|_U \leq \frac{C}{(T-t)^{2\gamma-1+\epsilon}}, \quad s \leq t < T$$

3. *$P(t)$ is continuous on $[s,T]$ and $P(t) \in \mathcal{L}(Y, L_\infty([0,T];Y))$*
4. *$B^*P(t)$ exhibits the following singularity:*

$$|B^*P(t)x|_U \leq \frac{C|x|_Y}{(T-t)^\gamma}, \quad 0 \leq t < T$$

5. *$u^0(t,s,y_s) = -B^*P(t)y^0(t,s,y_s)$, $s \leq t < T$*
6. *$P(t)$ satisfies the Riccati differential equation with $t < T$ for all $x, y \in \mathcal{D}(A)$:*

$$\begin{aligned} &\langle P_t x, y\rangle_Y + \langle AP(t)x, y\rangle_Y + \langle P(t)Ax, y\rangle_Y + \langle Rx, Ry\rangle_Z \\ &= \langle B^*P(t)x, \quad B^*P(t)y\rangle_U \end{aligned}$$

Moreover $lim_{t\to T}P(t)x = G^*Gx$ for all $x \in Y$.

7. *The solution of the Riccati equation above is unique within the class of positive and self-adjoint operators for $\gamma < \frac{1}{2}$.*
8. *For $1 > \gamma \geq \frac{1}{2}$, the solution of the Riccati equation above is unique within the class of positive and self-adjoint operators $P(t) \in \mathcal{L}(Y)$, which satisfy the following condition: there exists $u \in L_2([t,T];U)$ such that if $y(u)$ solution of (2.2) with an initial condition $y_t \in \mathcal{D}(A)$, one has $B^\star P(t)y(\cdot) \in L_2([t,T],U)$.*

As seen above, the optimal control problem with the final penalization and unbounded controls displays different properties from standard Bolza problems. The main novelty of the problem is the singularity of the gain operator $B^*P(t)$ at the terminal point. In addition, the smoothing effect of the optimization is restricted to an open interval $(0,T)$. Both the control and the trajectory display pointwise singularity at the terminal point T (unlike problems without final penalization). However, the singular estimate condition does guarantee a good L_2 well-posedness of all trajectories. Indeed, the following lemmas are starting points of the analysis.

Lemma 2.2

(See [15].) Let A be a generator of a C_0 semigroup and let $B : U \to [D(A^*)]'$ satisfy the singular estimate, that is, $|e^{At}Bu|_H \leq \frac{C}{t^\gamma}|u|_U \; 0 \leq \gamma < 1$. Then the map $u \to Lu$ where $Lu(t) \equiv \int_0^t e^{A(t-s)}Bu(s)ds$ is bounded from $L_2([0,T];U) \to L_2([0,T];Y)$.

The L_2 estimate in the lemma above follows from Young's inequality. As a consequence of Lemma (2.2) we obtain:

Lemma 2.3

The optimal state trajectory L_2 is continuously dependent on the optimal control. More specifically, we have the following estimate:

$$\|y^0\|_{L_2([s,T];Y)} \leq K\|u^0\|_{L_2([s,T];U)}$$

The proof of Theorem 2.1 follows many technical steps and estimates. It is based on the theory of singular integrals and involves construction of special Banach spaces capturing singularity at the origin and at the terminal point. The details are given in [18].

At this point it may be instructive to say that in proving the theorem the main effort goes into the analysis of the transfer function corresponding to the optimal trajectory. That is to say, we are interested in the behavior of the transition operator $\Phi(t,s)$ defined as a solution to the following integral operator equation:

$$\Phi(t,s) = e^{A(t-s)} + \int_s^t e^{A(t-z)}BB^*P(z)\Phi(z,s)dz$$

and $P(t)$ defined as a realization of the value function. It is clear that the equation for Φ is singular (singular kernel); however, the iterated contraction fixed point argument applied on specially constructed weighted Banach spaces C_γ (see [8,10,15]) allows one to establish the unique solvability for Φ. In particular, the following singular estimate for the optimal transfer function $\Phi(t,s)B$ is proved in [18]:

Lemma 2.4

$$|\Phi(t,s)Bu|_Y \leq \frac{C}{(t-s)^\gamma(T-t)^\gamma}|u|_U, \quad s < t < T$$

where C depends only on $T - s > 0$.

2.5 Remark

The estimate above shows that the singularity at the origin, caused by the unboundedness of B, is propagated to the terminal point T. Thus, the optimal transfer function experiences a double singularity: one at the origin and the second one at T. This phenomenon should be contrasted with SECS without terminal penalization. In this latter case, the singularity takes place only at the origin $t \sim s$.

The singular estimate for the optimal transfer function, an issue of independent interest in system theory, at the same time constitutes a critical ingredient for the proof of solvability of the Riccati equation. Details are given in [18].

3 Optimal synthesis for Bolza problem associated with structural acoustic interactions

As said before, SECS typically arise in coupled PDE equations that exhibit both parabolic and hyperbolic dynamics. We shall consider one such system—structural acoustic interaction. Other examples such as fluid structure interaction models, composite plates, and thermal plates are discussed in [13,14].

We consider a bounded smooth convex domain Ω in $\mathbb{R}^n$, $n = 2, 3$ with boundary $\Gamma = \bar{\Gamma_0} \bigcup \bar{\Gamma_1}$ with $\Gamma_0 \bigcap \Gamma_1 = \phi$ and Γ_0 a flat wall. In applications this represents an acoustic chamber with two walls, a solid wall and a flexible flat wall where interaction with the structure takes place. The acoustic medium in the chamber is described by the wave equation in the variable z (where the quantity $\rho_1 z_t$ is acoustic pressure). The interaction between the acoustic medium and the structural medium on Γ_0 is described by the plate equation with coupling terms where w is the displacement of the plate (for details of modeling we refer to [4].

$$\begin{cases} z_{tt} = c^2 \Delta z & \text{in} \quad \Omega \times (0,T) \\ \frac{\partial}{\partial \nu} z + d_1 z = 0 & \text{in} \quad \Gamma_1 \times (0,T) \\ \frac{\partial}{\partial \nu} z + dDz_t = w_t & \text{in} \quad \Gamma_0 \times (0,T) \\ w_{tt} + \mathcal{A}w + \rho \mathcal{A}^\alpha w_t + \rho_1 z_t|_{\Gamma_0} = \mathcal{B}u & \text{in} \quad \Gamma_0 \times (0,T) \\ z(0,.) = z_0, \ z_t(0,.) = z_1 & \text{in} \quad \Omega \\ w(0,.) = w_0, \ w_t(0,.) = w_1 & \text{in} \quad \Gamma_0 \end{cases} \tag{3.5}$$

Here c denotes the speed of sound as usual, $d_1 > 0$, and the coefficient d represents potential boundary damping. The operator $D : L_2(\Gamma_0) \to L_2(\Gamma_0)$ is positive, self-adjoint and densely defined on $L_2(\Gamma_0)$ and subject to additional assumptions to be specified. The operator D represents boundary damping on Γ_0.

$\mathcal{A}$ denotes a positive, self-adjoint operator and is densely defined on $L_2(\Gamma_0)$, and $\rho, \rho_1 > 0$, where $\rho_1 z_t$ denotes the back pressure of the wall and $\rho \mathcal{A}^\alpha w_t$

denotes structural damping. When $\rho_1 = 0$ and $\frac{1}{2} \leq \alpha \leq 1$, the plate equation is known to generate an analytic semigroup on $\mathcal{D}(\mathcal{A}^{\frac{1}{2}}) \times L_2(\Gamma_0)$. Thus, system (3) exhibits coupling between the analytic component represented by the strongly damped elastic equation and the hyperbolic component represented by the wave equation.

The control operator $\mathcal{B}$ is an unbounded operator acting on a suitable control space U with the values in $[\mathcal{D}(\mathcal{A})]'$. The mathematical description of the operator $\mathcal{B}$ depends on specific application. In the context of smart materials and piezoceramic actuators these operators are represented by derivatives of *delta* functions supported either at some points ($n = 2$) or curves ($n = 3$) ([5] and [6]). Additional assumptions imposed on the control operators $\mathcal{B}$ will be formulated later.

The following assumptions are imposed on control $\mathcal{B}$ and damping operators $\mathcal{D}$.

3.1 Assumption

There exists a positive constant r, $0 < r < \frac{1}{2}$ such that $\mathcal{A}^{-r}\mathcal{B} \in \mathcal{L}(U, L_2(\Gamma_0))$; equivalently, $\mathcal{B}$ continuous: $U \to [\mathcal{D}(\mathcal{A}^r)]'$.

3.2 Assumption

$D : L_2(\Gamma_0) \supset \mathcal{D}(D) \to L_2(\Gamma_0)$ is a positive, self-adjoint operator, and there exists a constant r_0, $0 \leq r_0 \leq \frac{1}{4}$, and positive constants δ_1, δ_2 such that:

$$\delta_1 |z|_{\mathcal{D}(\mathcal{A}^{r_0})} \leq \langle Dz, z\rangle_{L_2(\Gamma_0)} \leq \delta_2 |z|_{\mathcal{D}(\mathcal{A}^{r_0})} \cdot \forall z \in \mathcal{D}(\mathcal{A}^{r_0}) \equiv \mathcal{D}(D^{\frac{1}{2}}).$$

3.3 Assumption

We assume the following relation between the damping parameters in the system:
i) either $r_0 + \frac{\alpha}{2} \geq r$ and $d > 0$, ii) or else $\alpha - 2r \geq 1/6$ and $H^{1/3}(\Gamma_0) = \mathcal{D}(\mathcal{A}^{1/12})$.

The control problem associated with this system is to minimize the following functional representing the pressure in an acoustic chamber:

$$J(u, z, w) = \int_0^T |u(t)|_U^2 dt + |\nabla z(T, .)|_{0,\Omega}^2 + |z_t(T, .)|_{0,\Omega}^2 \qquad (3.6)$$

We shall show that under the assumptions in Sections 3.1, 3.2, and 3.3 the optimal control problem (3.5) allows a unique optimal solution, and (ii) allows optimal feedback synthesis via a solution to the Riccati equation, which is well posed. This will be accomplished by putting the problem within the abstract framework of the previous section.

3.4 Semigroup framework

In order to adapt the model to the form of the control problem (2.2), we need to put the system into the semigroup form. To this end we introduce several spaces and operators.

We define $A_N : L_2(\Omega) \supset \mathcal{D}(A_N) \rightarrow L_2(\Omega)$ to be:

$$A_N h = -c^2 \Delta h, \quad \mathcal{D}(A_N) = \left\{ h \in H^2(\Omega) : \left(\frac{\partial}{\partial \nu} h + d_1 h \right) \Big|_{\Gamma} = 0 \right\}$$

We also define the Neumann map N from $L_2(\Gamma_0)$ to $L_2(\Omega)$, defined by:

$$\psi = Ng \Longleftrightarrow \left\{ \Delta\psi = 0 \ in \ \Omega; \quad \frac{\partial}{\partial \nu}\psi|_{\Gamma_0} = g, \quad \left(\frac{\partial}{\partial \nu}\psi + d_1 \psi \right) \Big|_{\Gamma_1} = 0 \right\}$$

N is continuous from $L_2(\Gamma_0) \rightarrow H^{3/2}(\Omega) \subset \mathcal{D}(A_N^{3/4-\epsilon})$ and computing $N^\star A_N$ we get: $N^\star A_N h = \{ h|_{\Gamma_0} \ on \ \Gamma_0, \ 0 \quad on \quad \Gamma_1 \}$

The above notation leads to the following abstract representation of structural acoustic interaction:

$$\left\{ \begin{aligned} z_{tt} + A_N z + dA_N N D N^\star A_N z_t - A_N N w_t = 0. \quad & on \ [\mathcal{D}(A_N)]' \\ w_{tt} + \mathcal{A} w + \rho \mathcal{A}^\alpha w_t + N^\star A_N z_t = \mathcal{B} u, \quad & on \ \mathcal{D}[(\mathcal{A})]' \end{aligned} \right. \tag{3.7}$$

The problem is considered on a state space $H = H_z \times H_w$, where $H_z \equiv \mathcal{D}(A_N^{1/2}) \times L_2(\Omega)$ and $H_w \equiv \mathcal{D}(\mathcal{A}^{1/2}) \times L_2(\Gamma_0)$.

We represent Equations (3.7) as a first-order abstract ODE by introducing the operator A:

$$A = \begin{pmatrix} A_z & C^\star \\ -C & A_w \end{pmatrix}, \quad C = \begin{pmatrix} 0 & 0 \\ 0 & N^* A_N \end{pmatrix} \tag{3.8}$$

where the operator A_z (resp A_w) is defined on H_z (resp. H_w) by:

$$A_z = \begin{pmatrix} 0 & I \\ -A_N & -dA_N N D N^\star A_N \end{pmatrix}; \quad A_w = \begin{pmatrix} 0 & I \\ -\mathcal{A} & -\rho\mathcal{A}^\alpha \end{pmatrix} \tag{3.9}$$

The operator $C : H_z \rightarrow H_w$ has the domain $\mathcal{D}(C) = \{ [z_1, z_2] \in H_z : N^\star A_N z_2 = z_2|_{\Gamma_0} \in L_2(\Gamma_0) \} \supset \mathcal{D}(A_N^{1/2}) \times \mathcal{D}(A_N^{1/4+\epsilon})$. The adjoint $C^\star$ is understood as $C^\star : H_w \rightarrow \mathcal{D}(A_N^{1/2}) \times [\mathcal{D}(A_N^{1/4+\epsilon})]'$. These operators represent the coupling within the system.

We can then conclude that $\mathcal{D}(A) = \{ [z_1, z_2, w_1, w_2] \in H_z \times H_w : z_2 \in \mathcal{D}(A_N^{1/2}), \ w_2 \in \mathcal{D}(\mathcal{A}^{1/2}), \ \mathcal{A}^{1-\alpha} w_1 + \rho w_2 \in \mathcal{D}(\mathcal{A}^\alpha), \ z_1 + dNDN^\star A_N z_2 - N w_2 \in \mathcal{D}(A_N) \}$.

A_z, A_w are maximally dissipative and thus generate strongly continuous semigroups of contractions by Lumer Philips (see Appendix A of [9]).

By exploiting m-dissipativity of A_z, A_w and the special structure of coupling operators C, $C^\star$, one shows that A is indeed maximally dissipative (see appendix A of [9]).

Define the control operator $B : U \to [\mathcal{D}(A^\star)]'$ to be: $B = [0, 0, 0, \mathcal{B}]^T$.

One can compute A^{-1} and verify that $A^{-1}B = [0, 0, -\mathcal{A}^{-1}\mathcal{B}, 0]^T$. Because, $-\mathcal{A}^{-1}\mathcal{B} = \mathcal{A}^{r-1}\mathcal{A}^{-r}\mathcal{B}$ is bounded from U to $\mathcal{D}(\mathcal{A}^{1/2})$ by the assumption in Section 3.1, $A^{-1}B$ is thus bounded from U to H. In other words, $B \in \mathcal{L}(U \to [\mathcal{D}(A^\star)]')$. This verifies the resolvent condition in the Assumption in Section 2.3.

With the above notation and $y(t) \equiv [z(t), z_t(t), w(t), w_t(t)]$, $y_0 = [z_0, z_1, w_0, w_1]$ we can express the model in Equations (3.7) as an abstract control system:

$$y_t = Ay + Bu, \ in \ [\mathcal{D}(A^\star)]', \quad y(0) = y_0 \in H \equiv H_z \times H_w \tag{3.10}$$

The associated Equation (3.6) is indeed a quadratic cost functional with terminal time penalization, so it already conforms to what is presented in the theory, with $R = 0$ on H in this case, $Z = H$ and G being the projection on the space H_z.

In order to apply the results of Theorem 2.1 we need to complete verification of the assumption in Section 2.3. For this, we must establish (i) a singular estimate and (ii) closeability of GL_T.

3.5 Singular estimate

The theorem below states that under the assumptions made on the model, the singular estimate does hold for the system defined by A and B.

Theorem 3.1

Under the stated assumptions in Sections 3.1, 3.2 and 3.3, the control system described in this model satisfies the singular estimate for $0 < t < 1$*:*

$$|e^{At}Bu|_H \leq C\frac{|u|_U}{t^\gamma}, \quad \gamma = \begin{cases} \frac{r}{\alpha}, \ r \leq \frac{1}{2}\alpha \\ \frac{1/2-\alpha+r}{1-\alpha}, \ r > \frac{1}{2}\alpha \end{cases}$$

The proof of this theorem is technical and is given in [9]. There are three main ingredients of the proof: (i) characterization of fractional powers of elastic operators, (ii) sharp regularity of traces to the wave equation with Neumann data (non-Lopatinski case), and (iii) theory of analytic semigroups associated with strongly damped elastic equations.

The parameter $1/2 \leq \alpha \leq 1$ represents the degree of structural damping on the plate while the parameter $r_0 \leq 1/4$ measures the regularity of the operator D, which represents boundary damping. The hypothesis states that the higher the analyticity of the system represented by the higher value of α, the less the amount of boundary damping needed (represented by the lower value of r_0 mentioned in the assumption in Section 3.2) to insure the singular estimate condition. To illustrate, we consider the following two canonical cases occurring often in applications. In the first case we have the maximal damping on the wall (Kelvin's Voigt damping) corresponding to $\alpha = 1$. In that case there is no need for any damping in the boundary conditions (i.e., $D = 0$). The second case corresponds to the minimal case, which is the so-called *square root damping*-$\alpha = 1/2$. In that case a strong structural damping in the boundary conditions seems necessary.

Case 1: $\alpha = 1$. This is the case of the Kelvin Voigt plate. If one considers a point control as in Equation (4.11) below with $r = 3/8 + \epsilon$, then $\alpha - 2r$ immediately satisfies assumption (ii) in Section 3.3 ($\alpha - 2r > 1/6$), in which case the system shall satisfy the singular estimate assumption without any need for boundary damping (i.e., take $d = 0$ so that r_0 is irrelevant).

Case 2: $\alpha = 1/2$. In this case, the plate equation demonstrates the minimum degree of analyticity. When $r = 3/8 + \epsilon$, it is sufficient to take boundary damping with $r_0 > 1/8$ for assumption (i) in Section 3.3 to hold. Using the assumption in Section 3.2, the norm $|\mathcal{A}^{r_0} z|^2$ is topologically equivalent to $\langle Dz, z\rangle_{L_2(\Gamma_0)}$, and thus D is equivalent to $\mathcal{A}^{2r_0}$. If $\mathcal{A}$ represents a plate operator (i.e., elliptic operator of the fourth order), one could take the operator D simply to be the Laplace Bertrami operator Δ_{Γ_0} so that $2r_0$ would be equal to $1/2$, corresponding to the second-order operator. This will be illustrated later.

On the other hand, if $r = 1/4$—which is the case of point control using the Dirac delta—no boundary damping is necessary (i.e., $r_0 = 0$) for the assumption (i) in Section 3.3 to hold and hence the singular estimate.

3.6 Closeability of GL_T

Because G is the projection on the space H_z, we compute $L_T|_{H_z}$ by restricting our attention to the first two coordinates z_1, z_2, so rewriting the system (3.7) we have:

$$\begin{pmatrix} z \\ z_t \end{pmatrix}_t = A_z \begin{pmatrix} z \\ z_t \end{pmatrix} + B_z w_t(u)$$

where $A_z \equiv \begin{pmatrix} 0 & I \\ -A_N & -dA_N N D N^{\star} A_N \end{pmatrix}$ and $B_z \equiv \begin{pmatrix} 0 \\ -A_N N \end{pmatrix}$.

Therefore, one can write the projection of L_T, defined in (2.3), onto H_z as:

$$GL_T u = L_T u|_{H_z} = \int_0^T e^{A_z(T-t)} \begin{pmatrix} 0 \\ -A_N N w_t(u) \end{pmatrix} dt$$

To show the closeability of $GL_T : L_2([0,T];U) \to H_z$, it suffices to show that there exists a closed operator $T : H_z \to H_z$ with a bounded inverse such that $T^{-1}GL_T$ is bounded from $L_2([0,T];U) \to H_z$. Because A_z is a generator of a semigroup of contractions, A_z is closed with a bounded inverse on H_z. This one shows that $A_z^{-1}GL_T$ is bounded from $L_2([0,T];U) \to H_z$. Indeed, because $A_z^{-1} = \begin{pmatrix} -dNDN^\star A_N & -A_N^{-1} \\ I & 0 \end{pmatrix}$, estimating the norm of $A_z^{-1}GL_T$ yields:

$$\|A_z^{-1}GL_Tu\|_{H_z} = \left\| \int_0^T e^{A_z(T-t)} \begin{pmatrix} Nw_t \\ 0 \end{pmatrix} dt \right\|_{\mathcal{D}(A_N^{1/2}) \times L_2(\Omega)}$$

$$\leq M \left[\int_0^T \|Nw_t\|_{\mathcal{D}(A_N^{1/2})} dt \right] \leq M \left[\int_0^T K \|Nw_t\|_{H^1(\Omega)} dt \right]$$

Next, we use the fact that the Neumann map N is bounded from $H^{-1/2}$ $(\Gamma_0) \to H^1(\Omega)$, so it follows that:

$$M \left[\int_0^T K \|Nw_t\|_{H^1(\Omega)} dt \right] \leq M \left[\int_0^T K \|w_t\|_{H^{-1/2}(\Gamma_0)} dt \right]$$

Because the $H^{-1/2}$ norm is controlled by the L_2 norm, we have:

$$M \left[\int_0^T K \|w_t\|_{H^{-1/2}(\Gamma_0)} dt \right] \leq M \int_0^T K \|w_t\|_{L_2(\Gamma_0)} dt$$

$$\leq MKT^{1/2} \left[\int_0^T \|w_t\|^2_{L_2(\Gamma_0)} dt \right]^{1/2} \leq C \left[\int_0^T \|y\|^2_H dt \right]^{1/2}$$

where y is used to denote the complete state trajectory $[z, z_t, w, w_t]$. We next use the lemma, which implies that the L_2 norm in time of the state variable is continuously dependent on the L_2 norm of the control, therefore we obtain:

$$\|A_z^{-1}GL_Tu\|_{H_z} \leq C \left[\int_0^T \|u\|^2_U dt \right]^{1/2}$$

This establishes that $A_z^{-1}GL_T$ is bounded from $L_2([0,T];U) \to H_z$, and hence the closeability of GL_T.

3.7 Main result for structural acoustic interaction model 3.5

We have shown that under the assumptions in Sections 3.1, 3.2, and 3.3, assumption 2.3 in Section 2.3 of the abstract Theorem 2.1 is verified. This leads to the following conclusion:

Theorem 3.2

Consider structural acoustic interaction in Section 3 under the assumptions 3.1, 3.2, and 3.3 along with functional cost in Equation (3.6). Then there exists a unique optimal control $u^0 \in L_2([0,T];U)$ and optimal synthesis such that all the conclusions of Theorem 3.1 are satisfied with $A, B, R,$ and G defined above in Section 3.1.

Remark
It is worth noting that the proof of closeability of the operator GL_T, a necessary and sufficient condition for solvability of the optimal control problem, depends on the singular estimate being satisfied for the overall system. Thus, in this case, the singular estimate is not only essential for the well-posedness of Riccati equations, but also for the very existence of an optimal solution.

In what follows we shall present concrete examples of structural interactions where the control operator and model for the wall are specified.

4 Structural acoustic interaction with piezoceramic actuators and flexible walls

In this section we shall describe concrete applications of the abstract model in Equation 3.5, which involve controlling pressure inside the acoustic chamber by actuating vibrations on the wall via piezoceramic patches. This model was introduced in [4].

We begin with a definition of the elastic operator $\mathcal{A}$, which describes a simply supported plate.

$$\mathcal{A} \equiv \Delta^2, \quad \mathcal{D}(\mathcal{A}) = \{w \in H^4(\Gamma_0), \quad w = \Delta w = 0, \quad on\ \partial\Gamma_0\}$$

We have $\mathcal{D}(\mathcal{A}^{1/2}) \sim H^2(\Gamma_0) \cap H_0^1(\Gamma_0)$.

The control operator $\mathcal{B} : U \to [\mathcal{D}(\mathcal{A})]'$ is given as follows:

$$\mathcal{B}u = \sum_{j=1}^{J} a_j u_j \delta'_{\xi_j}, \quad [u_1, ..., u_J] \in \mathbb{R}^J = \mathcal{U} \tag{4.11}$$

where if dim $\Omega = 2$, dim$\Gamma_0 = 1$, then ξ_j are points on Γ_0, a_j are constants, and δ'_{ξ_j} are (distributional) derivatives of the delta functional supported at ξ_j.

On the other hand, if dim $\Omega = 3$ (dim $\Gamma_0 = 2$), then ξ_j are closed regular curves on Γ_0, a_j are smooth functions, and each δ'_{ξ_j} denotes the normal (distributional) derivative supported at ξ_j. This control u_j represents voltage applied to patches enclosed by these closed curves ξ_j on the boundary Γ_0

causing a reduction in the vibrations, which is known as the piezoelectric control ([5] and [6]).

With this choice of the control operator $\mathcal{B}$ along with the bilaplacian $\mathcal{A}$, the pair $\mathcal{B}$, $\mathcal{A}$ indeed satisfies the assumption in Section 3.1. With the following value of r (see [2] for proof): $r = \frac{3}{8} + \epsilon$.

We shall also compute $\mathcal{B}^\star$. Consider the duality mapping on $\phi \in \mathcal{D}(\mathcal{A}^{1/2})$, the two-dimensional case ($n = 2$) $\mathcal{B}^\star : H^2(\Gamma_0) \to R^J$

$$\langle \mathcal{B}(u), \phi \rangle_{L_2(\Gamma_0)} = -\sum_{j=1}^{J} a_j u_j \phi^{'}(\xi_j), \quad \mathcal{B}^\star \phi = (-a_j \phi^{'}(\xi_j))_{j=1}^{J}, \tag{4.12}$$

and in three dimensions (n = 3) $\mathcal{B}^\star : H^2(\Gamma_0) \to [L_2(\xi_j)]^J$

$$\langle \mathcal{B}(u), \phi \rangle_{L_2(\Gamma_0)} = -\sum_{j=1}^{J} \int_{\xi_j} a_j u_j \frac{\partial}{\partial \nu} \phi \, d\xi_j, \quad \mathcal{B}^\star \phi = \left(- \int_{\xi_j} a_j \frac{\partial}{\partial \nu} \phi \, d\xi_j \right)_{j=1}^{J} \tag{4.13}$$

Having introduced the control operator, we complete the model by specifying damping imposed on the plate equation and on the boundary conditions of the wave equation. As before we shall consider two extreme cases: maximal and minimal structural damping corresponding to elastic structure.

Case 1

Here we take $\alpha = 1$ (Kelvin Voight damping). As discussed in the previous section, one does not need any damping in the boundary conditions of the wave. This results in the following model, which is a specialization of (3.5):

$$\begin{cases} z_{tt} = c^2 \Delta z & \text{in} \quad \Omega \times (0, T) \\ \frac{\partial}{\partial \nu} z + d_1 z = 0 & \text{in} \quad \Gamma_1 \times (0, T) \\ \frac{\partial}{\partial \nu} z = w_t & \text{in} \quad \Gamma_0 \times (0, T) \\ w_{tt} + \Delta^2 w + \rho \Delta^2 w_t + \rho_1 z_t|_{\Gamma_0} = \mathcal{B} u & \text{in} \quad \Gamma_0 \times (0, T) \\ w = \Delta w = 0 & \text{in} \quad \partial \Gamma_0 \end{cases} \tag{4.14}$$

We could replace simply supported boundary conditions by the clamped boundary condition $w = \frac{\partial}{\partial \nu} w = 0$, on $\partial \Gamma_0$. Applying Theorem 3.1 to this concrete model produces a value of $\gamma = \frac{3}{8} + \epsilon$. We can now sum up the results pertaining to system (4.14) with the target cost functional (3.6) by the following theorem.

Theorem 4.1

For all initial data in H; $z_0 \in H^1(\Omega), z_1 \in L_2(\Omega)$, $w_0 \in H^2(\Gamma_0) \bigcap H_0^1(\Gamma_0)$, $w_1 \in L_2(\Gamma_0)$ there exists a unique control $u^0 \in L_2([0, T]; \mathbb{R}^J)$ and the *corresponding trajectory* $(z^0, z_t^0) \in L_2([0,T]; H_z)$, $(w^0, w_t^0) \in L_2([0,T]; H_w)$ *is continuous on* $[0,T] \to H_z \times H_w$ *such that:*

1. $\|u^0(t)\|_{\mathbb{R}^J} \leq C \frac{1}{(T-t)^{\frac{3}{8}+\epsilon}}$
2. $[|z^0(t)|^2_{H_1(\Omega)} + |z^0(t)|^2_{L_2(\Omega)} + |w^0(t)|^2_{H^2(\Gamma_0)} + |w_t^0(t)|^2_{L_2(\Gamma_0)}]^{1/2} \leq C$
3. *$P(t)$ is a positive self-adjoint operator on $H = H_z \times H_w$ satisfying the Riccati equation given in property 6 of Theorem 2.1 with A as given in (3.8), $B^\star$ given in (4.12) and (4.13), $R = 0$ and $G^\star = G$.*
4. $u^0(t) = -B^\star[p_1(t,.), p_2(t,.), p_3(t,.), p_4(t,.)]^T = -\mathcal{B}^\star p_4(t,.)$ *where*

$$P(t)[z^0(t,.),\ z_t^0(t,.),\ w^0(t,.),\ w_t^0(t,.)] = [p_1(t,.),\ p_2(t,.),\ p_3(t,.),\ p_4(t,.)]$$

and $B^\star$ is defined in (4.12) and (4.13) when $n = 2$ and $n = 3$, respectively.

5. $|B^\star P(t)x|_U \leq \frac{C}{(T-t)^{\frac{3}{8}+\epsilon}} |x|_H$ *where* $x = (x_1, x_2, x_3, x_4) \in H \sim H^1(\Omega) \times L_2(\Omega) \times H^2(\Gamma_0) \bigcap H_0^1(\Gamma_0) \times L_2(\Gamma_0)$
6. *The minimum energy of system (4.14) is given by:*

$$J^0 = \int_\Omega \nabla p_1(0,.) \nabla z_0(.) d\Omega + \int_\Omega p_2(0,.) z_1(.) d\Omega + \int_{\Gamma_0} \Delta p_3(0,.) \Delta w_0(.) d\Gamma_0 + \int_{\Gamma_0} p_4(0,.) w_1(.) d\Gamma_0$$

Case 2

Here we consider $\alpha = 1/2$ (square root damping). As discussed in the previous section, one needs additional structural damping in the boundary conditions of the wave equation to insure the singular estimate condition. This results in the following structure:

$$\begin{cases} z_{tt} = c^2 \Delta z & \text{in} \quad \Omega \times (0,T) \\ \frac{\partial}{\partial \nu} z + d_1 z = 0 & \text{in} \quad \Gamma_1 \times (0,T) \\ \frac{\partial}{\partial \nu} z + D z_t = w_t & \text{in} \quad \Gamma_0 \times (0,T) \\ w_{tt} + \Delta^2 w + \rho \Delta w_t + \rho_1 z_t|_{\Gamma_0} = \mathcal{B} u & \text{in} \quad \Gamma_0 \times (0,T) \\ w = \Delta w = 0 & \text{in} \quad \partial \Gamma_0 \end{cases} \tag{4.15}$$

where $(Dv, w)_{L_2(\Gamma_0)} \equiv (\nabla v, \nabla w)_{L_2(\Gamma_0)}$, so D behaves like the Laplace Bertrami operator, which means the value of $r_0 = 1/4$ and this, as discussed earlier, is sufficient to establish the singular estimate.

Applying Theorem 3.1 to this concrete model results in a value of $\gamma = \frac{3}{4} + \epsilon$. Hence, we can now sum up the results pertaining to system (4.15) with the target cost functional (3.6) by the corresponding theorem below.

Theorem 4.2

For all initial data in H; $z_0 \in H^1(\Omega), z_1 \in L_2(\Omega)$, $w_0 \in H^2(\Gamma_0) \bigcap H^1_0(\Gamma_0)$, $w_1 \in L_2(\Gamma_0)$ there exists a unique control $u^0 \in L_2([0,T];\mathbb{R}^J)$ and the corresponding trajectory $(z^0, z^0_t) \in L_2([0,T]; H_z)$, $(w^0, w^0_t) \in L_2([0,T]; H_w)$ such that:

1. $\|u^0(t)\|_{\mathbb{R}^J} \le C \frac{1}{(T-t)^{\frac{3}{4}+\epsilon}}$
2. $[|z^0(t)|^2_{H_1(\Omega)} + |z^0(t)|^2_{L_2(\Omega)} + |w^0(t)|^2_{H^2(\Gamma_0)} + |w^0_t(t)|^2_{L_2(\Gamma_0)}]^{1/2} \le \frac{C}{(T-t)^{\frac{1}{2}+\epsilon}}$
3. $|B^\star P(t)x|_U \le \frac{C}{(T-t)^{\frac{3}{4}+\epsilon}} |x|_H$ *where* $x = (x_1, x_2, x_3, x_4) \in H \sim H^1(\Omega) \times L_2(\Omega) \times H^2(\Gamma_0) \bigcap H^1_0(\Gamma_0) \times L_2(\Gamma_0)$

   Results 3, 4, and 6 from Theorem 4.1 also apply here for system (4.15).

References

[1] Acquistapace, P. & Terreni, B., Classical solutions of non-autonomous Riccati equations arising in parabolic boundary control problems, *Applied Mathematics and Optimization*, 39 (1999), 361–410.

[2] Avalos, G. & Lasiecka, I., Differential Riccati equation for the active control of a problem in structural acoustics, *J. Optim. Theory Appl.*, 91, 3 (1996), 695–728.

[3] Balakrishnan, V., *Applied functional analysis*, Springer-Verlag, Heidelberg, 1985.

[4] Banks, H.T., Silcox, R.J., & Smith, R.C., The modeling and control of acoustic/ structure interaction problems via piezoceramic actuators: 2-d numerical examples, *ASME J. Vibration Acoustics*, 2 (1993), 343–390.

[5] Banks, H.T. & Smith, R.C., Well posedness of a model of for structural acoustic coupling in a cavity enclosed by a thin cylindrical shell, *Journal of Mathematical Analysis and Applications*, 191 (1995), 1–25.

[6] Banks, H.T., Smith, R.C., & Wang, Y., The modeling of piezoceramic patch interactions with shells, plates and beams, *Quart. Appl. Math.*, 53 (1995), 353–381.

[7] Barbu, V., Lasiecka, I., & Triggiani, R., Extended algebraic Riccati equations in the abstract hyperbolic case, *Nonlinear Analysis*, 40 (2000), 105–129. Invited paper for special issue in honor of Lakshmikantham.

[8] Bensoussan, A., Da Prato, G., Delfour, M.C., & Mitter, S.K., *Representation and control of infinite dimensional systems*, Birkhäuser, Basel, 1993.

[9] Bucci, F., Lasiecka, I., & Triggiani, R., Singular estimates and uniform stability of coupled systems of hyperbolic/parabolic PDEs, submitted.

[10] Flandoli, F., Riccati equations arising in boundary control problems with distributed parameters, *SIAM J. Control*, 22 (1984), 76–86.

[11] Kasevan, S., *Topics in functional analysis and applications*, New Age International, 1999.

[12] Frankowska, H. & Ochalm, A., On singularities of value function for Bolza optimal control problem, *Journal of Mathematical Analysis and Applications,* 306, 2 (June 15, 2005).

[13] Lasiecka, I., *Mathematical control theory of coupled PDE's*, NSF-CMBS Lecture Notes, SIAM, Philadelphia, 2002.

[14] Lasiecka, I., *Optimal control problems and Riccati equations for systems with unbounded controls and partially analytic generators: applications to boundary and point control problems*, Lecture Notes 1855, Springer-Verlag, New York, 2004.

[15] Lasiecka, I. and Triggiani, R., *Control theory for partial differential equations: continuous and approximations theories*, Vol. I, Cambridge University Press, Cambridge, 1998.

[16] Lasiecka, I. and Triggiani, R., Structural decomposition of thermoelastic: Semigroups with rotational forces, *Semigroup Forum*, 60 (2000), 16–66.

[17] Lasiecka, I. and Triggiani, R., Optimal control and differential Riccati equations under singular estimates for $e^{At}B$ in the absence of analyticity, *Advances in dynamics and control*, Special volume dedicated to A. V. Balakrishnan, Chapman & Hall/CRC Press, Boca Raton, 2004, 271–309.

[18] Lasiecka, I. & Tuffaha, A., The optimal quadratic cost control problem: The case of c_0-semigroup satisfying a singular estimate with terminal time penalization, submitted.

[19] Lions, J.L. & Magenes, E., *Non-homogeneous boundary value problems and applications*, Springer-Verlag, Heidelberg, 1972.

[20] Pazy, A., *Semigroups of linear operators and applications to partial differential equations*, Springer-Verlag, Heidelberg, 1983.

[21] Triggiani, R., The algebraic Riccati equations with unbounded control operator. The abstract hyperbolic case revisited, *Contemporary mathematics: Optimization methods in partial differential equations*, Vol. 209, AMS, Providence, RI, 315–339, 1997.

[22] Weiss, G. & Zwart, H., An example in lq optimal control, *Systems Control Lett.*, 8 (1998), 339–349.

CHAPTER 13

Energy decay rates for the semilinear wave equation with nonlinear localized damping and source terms—an intrinsic approach

Irena Lasiecka
Department of Mathematics, University of Virginia,
Charlottesville, Virginia

Daniel Toundykov
Department of Mathematics, University of Virginia,
Charlottesville, Virginia

1 Introduction

Let Ω be an open bounded connected domain in $\mathbb{R}^n$, with local Lipschitz boundary Γ. Define $Q_T \equiv [0,T] \times \Omega$, $\Sigma_T \equiv [0,T] \times \Gamma$, and let $\|\cdot\|$ stand for $L^2(\Omega)$ norm.

Consider the following model of the wave equation with localized damping $\chi g(u_t)$ and source term $f(u)$:

$$\begin{cases} u_{tt} - \Delta u + \chi g(u_t) = f(u) \qquad \text{in} \quad Q_T \\ \qquad\qquad\quad u(0) = u_0, \quad u_t(0) = u_1 \end{cases} \tag{1.1}$$

The functions g (resp. f) represent Nemytski operators associated with scalar, continuous, real-valued functions $g(s)$ (resp. $f(s)$). Map g, assumed monotone increasing, models dissipation. Instead, function f corresponds to a source. The dissipation acts on small subportion Ω_χ of Ω, and χ is the characteristic function of this subset. We postpone the description of Ω_χ, for now it suffices to say that Ω_χ covers a thin layer (a collar) near a portion of the boundary.

The aim of this paper is to study asymptotic behavior (as $t \to \infty$) and related decay rates for the corresponding solutions. We are predominantly interested in the Neumann type of boundary condition, which is the most challenging in the context of this problem.

1.1 Boundary conditions

A distinct feature of this paper is the analysis of dynamics under *Neumann* boundary conditions (BC) that do not satisfy the *Lopatinski* condition. Before specifying the boundary dynamics, we divide the boundary Γ into parts: Γ_0 and Γ_1 so that $\Gamma = \overline{\Gamma_0} \cup \overline{\Gamma_1}$. In addition, we define the interface set

$$J \equiv \overline{\Gamma}_0 \cap \overline{\Gamma}_1 \tag{1.2}$$

System (1.1) may be equipped with one of the following sets of *Neumann type* BC:

Neumann-Robin. Let constants $k_i \geq 0$, $i = 1, 2$, satisfy $k_0 + k_1 > 0$.

$$\left(\frac{\partial}{\partial \nu} u + k_0 u \right) \bigg|_{\Gamma_0} \equiv 0, \qquad \left(\frac{\partial}{\partial \nu} u + k_1 u \right) \bigg|_{\Gamma_1} \equiv 0 \tag{1.3}$$

Neumann-Dirichlet. For $k_1 \geq 0$

$$u|_{\Gamma_0} \equiv 0, \quad \Gamma_0 \neq \emptyset, \qquad \left(\frac{\partial}{\partial \nu} u + k_1 u \right) \bigg|_{\Gamma_1} \equiv 0 \tag{1.4}$$

Remark
Note that we do not assume that interface J is empty, thus solutions may develop singularities at the higher energy levels (see [9]). Indeed, this is the situation in the Neumann-Dirichlet case (1.4). Also, a weaker type of singularity may develop in the Neumann-Robin case (1.3) when $k_0 \neq k_1$ (see Part II of Theorem 3.1).

Remark
The results and the methods employed in this paper apply as well to the pure Dirichlet setting: $u \mid_\Gamma \equiv 0$. However, in that case the Lopatinski condition holds and the difficulties associated with the Neumann problem disappear. Hence we will not explicitly treat the Dirichlet scenario.

1.2 Source and dissipation terms

The only hypotheses imposed on the dissipation function $g(s)$ are that it is continuous and monotone increasing. The source term f, instead, is a piecewise $C^1(\mathbb{R})$ function such that the Nemytski map f is locally Lipschitz from $H^1(\Omega)$ to $L^2(\Omega)$. This assumption alone will guarantee local existence of solutions. To obtain the main result we will also require the following assumption on the growth of f:

[(Af-1)] Assume $|f'(s)| \leq C_f |s|^{p-1}$ for a constant C_f, where $1 < p \leq \frac{n}{n-2}$ if $n > 2$, and $1 < p < \infty$ if $n \leq 2$

To have global solutions, impose one of the following two assumptions:

[(Af-1)] The antiderivative F of f satisfies $|F(s)| \leq c_1 |s|^{p+1} + c_2 |s|^{2+\varepsilon}$, for $c_i \geq 0,\ i = 1, 2,$ and $0 < \varepsilon < p - 1$
where $1 < p \leq \frac{n}{n-2}$ if $n > 2$, and $1 < p < \infty$ if $n \leq 2$. Let λ_1 be the first eigenvalue of Δ on Ω, subject to any of the aforementioned BC (1.3 or 1.4). Assume: $\limsup_{|s| \to \infty} \frac{f(s)}{s} < -\lambda_1$

Remark
The bound on f' in (Af-1) may accommodate a small constant c, that is, $|f'(s)| < C_f |s|^{p-1} + c$. For clarity of presentation we omit this term. Also, if $f(0) = 0$ then (Af-1) implies a version of (Af-2).

With system (1.1) we associate the following energy functional

$$E(t) = E(u(t), u_t(t)) \equiv \frac{1}{2} \|u_t(t)\|^2 + \frac{1}{2} \|\nabla u(t)\|^2 - \int_\Omega F(u(t,x))dx + \mathcal{B}(\Gamma)(t)$$

where F is the antiderivative of f, and $\mathcal{B}(\Gamma)(t)$ denotes the boundary energy at time $t \geq 0$. For full Dirichlet or full Neumann BC, $\mathcal{B}(\Gamma) \equiv 0$, while in the Neumann-Robin case (1.3): $\mathcal{B}(\Gamma)(t) \equiv \frac{k_0}{2} \|u(t)\|^2_{L^2(\Gamma_0)} + \frac{k_1}{2} \|u(t)\|^2_{L^2(\Gamma_1)}$

1.3 Goals of the paper

We aim to derive an optimal asymptotic, as $t \to \infty$, energy decay rates driven by geometric and topological characteristics of the domain Ω, and nonlinear functions f, g. It is well known that the presence of the source not only destabilizes the otherwise dissipative system, but may lead to a blow-up of solutions in *finite time* ([7,26]). Thus, *our first aim* is to construct a framework that yields computable decay rates for the energy of *nondissipative* systems with source terms.

The second goal is to solve the problem with Neumann-type boundary conditions (see Section 1.1) (rather than Dirichlet), where the former do not satisfy Lopatinski conditions, which was used critically for stabilization of the Dirichlet problem. The method developed will also allow us to treat mixed problems with an interface of Neumann-Dirichlet BC.

The third objective of our work is to construct a general and complete theory that will make it possible to consider dissipation $g(s)$ *without any growth conditions*, neither at zero nor infinity. This is in contrast to the past literature results where, even for dissipative systems, linear bounds were assumed at infinity and polynomial bounds were prescribed at the origin. This third goal will be achieved by extending the uniform stability technique introduced in [18].

2 Existence, uniqueness and regularity of solutions

Let $\mathcal{H}(\Omega)$ denote the energy space $H^1_\Gamma(\Omega) \times L^2(\Omega)$, where $H^s_\Gamma(\Omega)$ denotes the completion, with respect to $H^s(\Omega)$-topology, of the $C^\infty(\Omega)$-functions that satisfy relevant boundary conditions—be it Neumann-Robin or Dirichlet-Neumann. For convenience we also introduce

$$\mathcal{H}^s(\Omega) \equiv H^{1+s}_\Gamma(\Omega) \times H^s_\Gamma(\Omega), \quad s \geq 0; \quad E_s(t) \equiv \|(u(t), u_t(t))\|^2_{\mathcal{H}^s(\Omega)} \qquad (2.5)$$

Theorem 2.1

Let $(u_0, u_1) \in \mathcal{H}(\Omega)$. In the case of Dirichlet-Neumann BC (1.4) we also adopt the compatibility condition $(u_0, u_1)|_{\Gamma_0} \equiv 0$.

Part I. *Existence and uniqueness*

(a) If (Af-3) holds, there exists a unique finite energy solution $(u, u_t) \in C([0,T]; \mathcal{H}(\Omega))$ for any $T < \infty$.

(b) Requirement (Af-3) in part (a) may be discarded, and the same conclusion follows by assuming condition (Af-2) and taking the initial condition (u_0, u_1) from a potential well:

$$W \equiv \{(u_0, u_1) \in \mathcal{H}(\Omega) : \|(-\Delta)^{1/2} u_0\| < s_0, \quad E(0) < d\} \qquad (2.6)$$

Constant s_0 is the first positive zero of $\frac{d}{ds}G(s)$, where

$$G(s) = \frac{1}{2}s^2 - c_1 P_\Omega^{p+1} s^{p+1} - c_2 R_\Omega^{2+\varepsilon} s^{2+\varepsilon}$$

and $R_\Omega \equiv \sup_{u \in H^1(\Omega)} \frac{\|u\|_{L^{2+\varepsilon}}}{\|(-\Delta)^{1/2}u\|}$, $P_\Omega \equiv \sup_{u \in H^1(\Omega)} \frac{\|u\|_{L^{p+1}(\Omega)}}{\|(-\Delta)^{1/2}u\|}$. In this setting, the corresponding solution stays in the well W for $t \geq 0$, and potential energy $\frac{1}{2}\|\nabla u(t)\|^2 + \mathcal{B}(\Gamma)(t) - \int_\Omega F(u(t)) d\Omega$ remains nonnegative for all $t \geq 0$.

Part II. *Regularity*

(a) Neumann-Robin BC. Assuming BC (1.3), dimension $n \leq 3$, and more regular initial data: $(u_0, u_1) \in \mathcal{H}^1(\Omega)$, the following improved regularity of the solution (u, u_t) takes place: $(u, u_t) \in C([0,T); H^{2-\varepsilon}(\Omega) \times H^1(\Omega))$. If $k_0 = k_1$, one can take $\varepsilon = 0$.

(b) Neumann-Dirichlet BC. Let the BC be given by (1.4). Write U(J) to denote some $\mathbb{R}^n$-neighborhood of the junction J defined in (1.2).

Assuming $u_0 \in H^2(\Omega) \cap H^1_{\Gamma_0}(\Omega)$, $u_1 \in H^1_{\Gamma_0}(\Omega)$ *we get* $u \in C([0,T);\, H^2(\Omega^*))$ *and* $u_t \in C([0,T);\, H^1(\Omega))$ *where* Ω^* *is the interior of* $\Omega \backslash U(J)$. *When* $J = \emptyset$ *(e.g., if* Ω *is an annulus), one can replace* Ω^* *with* Ω.

3 Decay rates for energy

In order to discuss the energy decay of solutions, one must be more specific about the geometry of the domain. Precise information is necessary because the damping in the equation acts only on a small subset of Ω.

3.1 Geometry of the domain

To formulate long-term energy behavior we require additional assumptions of both a geometric and topological nature. Sufficient support of the localizer χ in (1.1) is essential for stability, but on the other hand one would like to apply damping to as minimal an area as possible. The geometric optics argument indicates that in order to obtain uniform decay rates for energy, damping needs to cover a collar near a part of the boundary.

The section where the damping is absent will have to satisfy some geometric conditions. In the case when Dirichlet BC acts on the nondissipative part of the boundary, the traditional *star-shaped* condition suffices. However, in the Neumann case more is needed, with convexity being a sufficient requirement. Before we state these assumptions, consider yet another separation of Γ: split it into the controlled segment Γ_C (near which the damping is active), described indirectly via

$$\Omega_\chi = \{x \in \Omega : \operatorname{dist}(x, \Gamma_C) < \tau\}, \quad \text{(any fixed } \tau > 0)$$

(recall that Ω_χ is the set characterized by χ), and the remaining (partially) unobserved section Γ_U. Both Γ_C, Γ_U are connected and relatively open in Γ, with $\Gamma = \Gamma_C \cup \Gamma_U$.

Remark
There is no restriction on the size τ of the dissipative layer.

In the case of a Dirichlet-Neumann boundary (BC-1.4), in order to cope with singularities propagating from junction J, it is sufficient to assume that J falls into the controlled segment $J \subset \Gamma_C$.

Now, assume the following:
[(AD-1)] Let the (nonintersecting) curve Γ_U be given as a level set

$$\Gamma_U = \{y \in \mathbb{R}^n : \ell(y) = 0\}, \quad \ell \in C^3$$

where ℓ is defined on a suitable domain in $\mathbb{R}^n$, $|\nabla \ell| \neq 0$ and $\ell(x) \leq 0$ in some $\mathbb{R}^n$-neighborhood of Γ_U. Also (WLOG) $\nabla \ell$ points toward the exterior of Ω. The Hessian matrix of ℓ is nonnegative definite on Γ_U (see [17], p. 302). There exists a point $x_0 \in \mathbb{R}^n$, outside $\overline{\Omega}$, so that $(x - x_0) \cdot \nu(x) \leq 0$ on Γ_U with ν denoting the outward normal vector field on $\partial\Omega$ (pointing in the same direction as $\nabla \ell(x)$).

3.2 Uniform decay rates

Here we state the main theorem of the paper. The result is fairly general: it provides uniform energy decay rates assuming that the damping function $g(s)$ is monotone, continuous, and zero at the origin. More specific quantitative versions of the decay rates announced in Theorem 3.1 will be given later.

Let $p^* = 2n/(n-2)$ (if $n \leq 2$ take $p^* < \infty$).

Theorem 3.1 Energy Decay Rates

Suppose $(u_0, u_1) \in \mathcal{H}(\Omega)$, and assume the following:

1. *Geometrical assumptions on Ω: (AD-1), (AD-2), (AD-3).*
2. *The source term f is a piecewise $C^1(\mathbb{R})$ function, locally Lipschitz from $H^1(\Omega)$ to $L^2(\Omega)$, and satisfies (Af-1). In addition, assume the stability hypothesis:*
   - *either adopt a stronger version of condition (Af-3):*
   $$\frac{f(s)}{s} < -\lambda_1, \quad \textit{for} \quad s \neq 0 \tag{3.7}$$
   - *or, alternatively, assume (Af-2) and pick the initial condition from the potential well: $(u_0, u_1) \in W$ (part I-b of Theorem 2.1).*

Let $\hbar_\lambda$, and $\hbar_\mu$ be monotone, continuous functions defined on $\mathbb{R}^+$ and equal zero at the origin. If the following two estimates hold:

1. *Let $\lambda \in [0,1]$, $p \in [2, p^*]$. Suppose that for all $x \geq 1$*
$$x \leq \hbar_\lambda[g(x)x] \quad \textit{and} \quad x \mapsto [\hbar_\lambda(x)]^{\frac{2(1-\lambda)p}{(p-2\lambda)}} \textit{ is concave}$$
and $\|u_t(t)\|_{L^p(\Omega)}$ is uniformly bounded for $t \geq 0$. Let $h_1 = \hbar_\lambda^{2(1-\lambda)}$.
2. *Let $\mu \in [0,1]$, $p \in [2, p^*]$. Assume for all $x \geq 1$ we have*
$$g(x) \leq \hbar_\mu[g(x)x] \quad \textit{and} \quad x \mapsto [\hbar_\mu(x)]^{\frac{2p}{p(2-\mu)-2(1-\mu)}} \textit{ is concave}$$
and $\|\nabla u(t)\|_{L^p(\Omega)}$ is uniformly bounded for $t \geq 0$. Let $h_2 \equiv \hbar_\mu^{2/(2-\mu)}$.

Then there exist constants $K, C_{\text{dec}}, T > 0$ such that

$$E(t) \le KS\left(\tfrac{t}{T} - 1\right) \qquad \textit{for } \; t \ge T$$

where S satisfies the differential equation:

$$S_t + q(S(t)) = 0, \qquad S(0) = E(0) \tag{3.8}$$

$$q \equiv I - \left(I + \left[(h+I)^{-1}\frac{(\cdot)}{C_{\text{dec}}}\right]\right)^{-1}, \qquad h = h_0 + h_1 + h_2 \tag{3.9}$$

where h_1, h_2 are given above and the function h_0 is monotone increasing, concave, $h_0(0) = 0$ and satisfies

$$s^2 + g^2(s) \le h_0(sg(s)) \qquad \textit{for } \; |s| \le 1 \tag{3.10}$$

The constant C_{dec} satisfies the following inequality

$$\begin{aligned} C_{\text{dec}} \le C_0(E(0)) &+ \operatorname{sgn}^+(h_1)C_1\left(\|u_t\|_{L^\infty(\mathbb{R}^+, L^p(\Omega))}\right) \\ &+ \operatorname{sgn}^+(h_2)C_2\left(\|u_t\|_{L^\infty(\mathbb{R}^+, L^p(\Omega))}\right) \end{aligned}$$

where $C_i(x)$ are increasing, continuous functions, and

$$\operatorname{sgn}^+(h) = \begin{cases} 0 & \textit{if} \quad h \equiv 0 \\ 1 & \textit{otherwise} \end{cases}$$

Map q is strictly monotone increasing, by construction, so $\lim\limits_{t\to\infty} S(t) = 0$.

Remark
When functions $h_i(s), i = 0, 1, 2$ are linear, then q is linear as well, and the decay rates predicted by (3.8) are exponential. The more interesting cases are, of course, when functions h_i are sublinear. This slows down the decay of the energy. Detailed analysis of the relationship between dissipation and resulting decay rates will be given later.

Remark
Formulation of Theorem 3.1 allows for a very broad class of dissipative functions $g(x)$ *without any a priori assumptions imposed at the origin or infinity.* Indeed, behavior of g near zero is captured by map $h_0(x)$, which can always be constructed due to the monotonicity of $g(x)$ ([18]). As for functions $h_1(x), h_2(x)$, these quantify behavior of the damping at infinity (which can also be very rough). Maps h_1 (resp. h_2) can be removed if the damping $g(x)$ is bounded at infinity by a linear function from below (resp. above). Thus, in the special case when the damping is linearly bounded at infinity from above and below, the decay rates are driven by h_0 alone and all

the constants depend on finite energy only. Otherwise one needs to compute h_1 or h_2 and estimate the constant C_{dec} in terms of the initial condition. A constructive procedure that allows us to determine h_1, h_2 explicitly in terms of the damping is outlined in the Remark below. Thus Theorem 3.1 extends the corresponding result in [18], which only deals with linearly bounded damping at infinity (above and below).

Remark

Computing h_1 and h_2. Functions $\hbar_\lambda$ and $\hbar_\mu$ can be determined explicitly in terms of damping. Indeed, Theorem 3.1 implies that when $g(x)$ is bounded below (resp. above) by a linear function, the optimal choices for $\hbar_\lambda$ (resp $\hbar_\mu$) are just linear maps. So we focus on the cases where the damping is sublinear (resp. superlinear) at the infinity.

Because g is a continuous monotone map, there exists a convex function $\beta : \mathbb{R} \to \mathbb{R}$, $\beta(0) = 0$ so that $g = \beta'$; more precisely, g is the (single-valued) subgradient operator of β. In addition, the monotonicity of g ensures that β is invertible on $\mathbb{R}^+$ (so by β^{-1} we mean $(\beta\mid_{\mathbb{R}^+})^{-1}$). From the properties of subgradients we have

$$\beta(x) \leq g(x)x$$

Thus, if the dissipation is sublinear at large velocities: $|g(x)| \leq c|x|$ for $|x| \geq 1$ one can choose $\hbar_\lambda \equiv \beta^{-1}$ and $h_1(x) \equiv [\beta^{-1}(x)]^{2(1-\lambda)}$ where $\lambda \in [0,1)$ is picked so that $[\beta^{-1}]^{\frac{2(1-\lambda)p}{p-2\lambda}}$ is concave. Because β^{-1} is concave, the requirement on λ always holds if $p > 2$ (by letting $\lambda \to 1$). Map h_2 in this case can be just $h_2(x) = cx$, so for $p > 2$ we can pick $\lambda \in [0,1)$, which leads to decay rates driven by h_1. Optimal values of λ yield h_1 whose behavior is close to the identity map; however, these best values have a price: $x \mapsto [\beta^{-1}(x)]^{\frac{2(1-\lambda)p}{p-2\lambda}} = h_1(x)^{\frac{p}{p-2\lambda}}$ must remain concave, and if $h_1(x) \cong x$, then the exponent $\frac{p}{p-2\lambda}$ should be near 1, which only happens for large p, the ideal case being $p \to \infty$. This trade-off of the decay rates vs. the regularity of solutions will be illustrated in Sections 3.4 and 4.

A similar setup follows if g is superlinear at infinity: $|g(x)| \geq c|x|$ for $|x| \geq 1$. We let $\hbar_\mu \equiv g \circ \beta^{-1}$ (and define h_1 to be linear, e.g., $x \mapsto x/c$). Consequently, $h_2(x) = g(\beta^{-1}(x))^{2/(2-\mu)}$, where $\mu \in [0,1)$ ensures that $x \mapsto [\hbar_\mu(x)]^{\frac{2p}{p(2-\mu)-2(1-\mu)}} = h_2(x)^{\frac{p(2-\mu)}{p(2-\mu)-2(1-\mu)}}$ is concave. We would like to select μ so that $h_2(x) \cong x$, yet in this case the concavity requirement demands that $\frac{p(2-\mu)}{p(2-\mu)-2(1-\mu)}$ be close to 1. Thus if we fix μ that we like, the way to ensure that concavity requirement holds would be to select a larger p, if possible $p \to \infty$, because then $\frac{p(2-\mu)}{p(2-\mu)-2(1-\mu)} \to 1$.

The proof of Theorem 3.1 is technical and lengthy (see [27]). Among other things it depends on a construction of special multipliers that help us cope with the non-Lopatinski case (where there is no “hidden regularity”). The details are given in [27] and will not be reported here. Instead, in what follows we demonstrate how to apply the abstract ODE estimate (3.8) of Theorem 3.1 to effectively compute the decay rates for our model. As we shall see, the procedure depends on the behavior of the damping in the two regions: infinity and origin; these are analyzed next.

3.3 Dissipation

To apply Theorem 3.1 we need more information about function g. Two regions, in the domain of g, are of paramount importance: near the origin (at small velocities) and at infinity (large velocities). Quantitatively, the damping functions that are bounded above and below by linear maps (be it origin or infinity) lead to the fastest exponential decay rates. Any deviation from linear bounds leads to much weaker decay expressed by algebraic or logarithmic rates.

3.3.1 Damping at the origin

Consider damping functions whose behavior near zero may be divided into three cases: let $m_0, M_0 > 0$ be constants, then

1. [(Ag-O1)] Linear bounds at the origin: $m_0 x^2 \leq g(x)x \leq M_0 x^2$, $|x| < 1$.
2. Sublinear: let $0 < \theta_2 < \theta_1 < 1$; $m_0|x|^{1+\theta_1} \leq g(x)x \leq M_0|x|^{1+\theta_2}$, $|x| < 1$.
3. Superlinear: $g(x)x < M_0 x^2, \quad |x| < 1$.

We note that the upper bounds in the conditions (Ag-O 3.3.1, 3.3.1, 3.3.1) are satisfied automatically if g is locally Lipschitz at the origin. A typical illustration of superlinear damping at the origin would be any map g with the property $g'(0) = 0$. Following [18], our proofs utilize a concave function h_0, vanishing at zero, which provides an estimate

$$x^2 + g^2(x) \leq h_0[g(x)x], \quad |x| < 1 \tag{3.11}$$

Due to the monotonicity of g, such a map h_0 can always be constructed (see [18]), so the existence of h_0 is a property, not an assumption.

3.3.2 Damping at infinity

Let $m, M > 0$, and suppose that the growth of g at infinity falls into one of the following categories:

1. [(Ag-I1)] Linearly bounded growth: $mx^2 \leq g(x)x \leq Mx^2$, $|x| \geq 1$.
2. Sublinear growth: for $\theta \in [0, 1)$ let $m|x|^{\theta+1} \leq g(x)x \leq Mx^2$, $|x| \geq 1$. Note that we include saturated damping: $\theta = 0$.
3. Superlinear: $mx^2 \leq g(x)x \leq M|x|^{r+1}$, $|x| \geq 1$ with $1 < r$.

3.4 Corollaries of Theorem 3.1

In most cases, due to the complicated structure of q in (3.8), ODE (3.8) does not admit closed form solutions. However, if the damping is linearly bounded at the origin or at infinity, then equivalent energy estimates may be obtained via simpler differential equations.

3.4.1 Super- or sublinear dissipation at the origin

Corollary 3.2

Under the hypotheses of Theorem 3.1, suppose that damping is linearly bounded at infinity (Ag-I1).

1. ***Sublinear dissipation at the origin.*** *Assume (Ag-O3.3.1), set*
$$h_0(x) \equiv x^{2\theta_2/(\theta_1+1)}$$
which is concave (because $\theta_2 < \theta_1$). The ODE to solve becomes
$$S_t + CS^{(\theta_1+1)/(2\theta_2)} = 0, \quad S(0) = E(0)$$
Parameter $C > 0$ may depend on $E(0)$. Then the energy estimate $E(t) \leq KS\left(\frac{t}{T} - 1\right)$ holds for all $t \geq T$ and some constant $K > 0$.
2. ***Superlinear dissipation at the origin.*** *Suppose (Ag-O3.3.1) holds. When the function $\sqrt{x}\, g(\sqrt{x})$ is convex on $[0, \varepsilon)$, some $\varepsilon > 0$, the decay rates are given by $E(t) \leq KS\left(\frac{t}{T} - 1\right)$, all $t > T$ and $K > 0$. Parameter T comes from Theorem 3.1, and $S(t)$ satisfies*
$$\dot{S}(t) + CS(t)^{1/2} g\left[\left(\frac{S(t)}{C_{\text{dec}}}\right)^{1/2}\right] = 0, \quad S(0) = E(0) \tag{3.12}$$
where constant $C > 0$ may depend on $E(0)$.

Remark
When for a given constant $\alpha_1 > 0$, we have $g(\alpha_1 x) = \alpha_2 g(x)$ with α_2 independent of x (e.g., if g is scalar-homogeneous of a fixed degree), we will often omit C_{dec} from (3.12) assuming that C has been adjusted accordingly. Henceforth we will use similar parameter adjustments whenever applicable without explicit reminders.

3.4.2 Super- or sublinear dissipation at infinity

The results of preceding subsections relied on the assumption that the damping at infinity was bounded above and below by linear maps. It is known that with boundary or localized damping the *linear*-like behavior of dissipation at infinity is "ideal" and (assuming also linear-like damping at the origin) the decay rates are exponential. We consider below two "pathological" types of damping at infinity: sublinear and superlinear. The rates obtained in this setting require higher regularity of the initial data. Our analysis provides a complete description of the relation and trade-off between the super- (sub-) linearity of the damping versus decay rates and additional regularity of solutions that needs to be imposed.

Corollary 3.3

Assume (for simplicity) $f = 0$, adopt the hypotheses of Theorem 3.1, and suppose that damping is linearly bounded at the origin: (Ag-O3.3.1). Moreover, in the case of Dirichlet-Neumann BC (1.4), assume either $J = \emptyset$, or that the regularity of solutions is high enough to satisfy the regularity conditions below.

Assume higher energy of the initial data: $(u_0, u_1) \in \mathcal{H}^s(\Omega)$, some $s > 0$. Let $p = 2n/(n-2s)$, (if $n = 2s$ set $p = \infty$).

1. ***Sublinear damping at infinity.*** *Suppose (Ag-I2) is the case, then the conclusion of Theorem 3.1 holds with $h = h_0 + h_1$ in (3.8), where h_0 satisfies (3.11), while $h_1 : x \mapsto x^{(p-2)/(p-\theta-1)}$ (if $p = \infty$ take $h_1(x) = x$). Constant C_{dec} in (3.8) depends on T, and higher energy $E_s(0)$ (defined in (2.5)) of the initial data.*
2. ***Superlinear damping at infinity.*** *Assume (Ag-I3) and $p \geq r + 1$. Then we may apply Theorem 3.1 and ODE (3.8), with $h = h_0 + h_2$, where h_0 satisfies (3.11). Define $h_2 : x \mapsto x^{r(p-2)/[r(p-1)-1]}$ (if $p = \infty$ take $h_2(x) = x$). Constant C_{dec} in (3.8) depends on T, and energy $E_s(0)$.*

Remark

Note that in 3 dimensions with $(u_0, u_1) \in \mathcal{H}^1(\Omega)$, the restriction $p \geq r + 1$ implies that the damping exponent satisfies $r \leq 5$, which is optimal. Indeed, this is precisely the range of damping exponents for which the solutions to the forced wave equation, with $L^\infty(0, T; L_2(\Omega))$-forcing term, remain bounded in time (see [10]).

Remark

Optimization of decay rates. One may consider two extreme cases in the context of Corollary 3.3: (i) maximizing decay rates at the expense of the regularity of initial data, (ii) minimizing (additional) regularity of initial data while sacrificing the decay rates.

Scenario (ii) may be very important in the case of limited regularity of solutions. To illustrate this optimization we assume linear damping at the origin (Ag-O3.3.1), sublinear at infinity (Ag-I2), and let $n = 3$:

- **Maximal regularity of initial data**: Pick $(u_0, u_1) \in \mathcal{H}^1(\Omega)$. Then we have $p = 6$, and according to (1.) take $h_1 : x \mapsto x^{4/(5-\theta)}$. In the case of saturated damping $\theta = 0$ we obtain $h_1(t) = t^{4/5}$. So energy diminishes as the solution to $S_t(t) + CS(t)^{5/4} = 0$. Here we use all available regularity for the fastest possible energy decay.
- **Minimal regularity of initial data**. Select $(u_0, u_1) \in \mathcal{H}^s(\Omega)$ and let $s \to 0$. Then, p approaches 2, whence $h_1(x) = x^\alpha$ with $\alpha \to 0$. The overall energy decay will be driven by the equation $S_t(t)+CS(t)^{1/\alpha} = 0$, where the exponent $1/\alpha$ is very large, leading to poor algebraic decay rates.

3.5 On the results, techniques and relevant literature

The study of decay rates for dissipative wave equations and other hyperbolic-like structures (plates, shells) has attracted considerable attention in past years—see [10], [11], [13], [14] and [16] and the references therein.

Canonical types of the dissipation mechanism are: (i) full-interior dissipation, (ii) boundary dissipation, (iii) localized dissipation (the damping only affects a layer near the boundary). Both boundary and localized dissipation are much more challenging problems than full internal dissipation, so we will only focus on types (ii) and (iii).

Chronologically, boundary dissipation was considered in the literature first in [6], [14], [19], and [23]. Further refinements and extensions to nonlinear setups can be found in [13], [16], and [18] and the references therein.

Having boundary stabilization in place, certain problems with localized dissipation became a natural corollary via the following principle: *boundary stabilization + hidden regularity $\Longrightarrow$ localized stabilization*. By applying boundary stabilization techniques, one recovers total energy in terms of the boundary traces. Then one uses hidden regularity to express the contribution of the traces by the energy in the layer where the dissipation is active. Thus the problem with localized dissipation may be reduced, via hidden regularity, to the proof already available for boundary damping; see [1], [13], [21], and [24].

Of course, the fundamental premise is the hidden regularity; it is intimately connected with the Lopatinski condition, which depends on the boundary conditions in the original problem. With Dirichlet boundary data, the L_2-bound for normal derivatives on the boundary in terms of finite energy of solutions has been known since the publication of [15]. The situation is very different when *Neumann boundary conditions are imposed* because these do not accord the Lopatinski condition and hidden regularity does not hold (unless $\dim \Omega = 1$).

Our work places itself within the non-Lopatinski, Neumann framework. We consider configurations where the portion of the boundary that is not subject to the dissipation satisfies the Neumann condition. This unobserved section with Neumann data has been recognized already in [12] as the cause of serious geometric difficulties. To cope with the latter, special (nonradial) multipliers developed in [17] are used, and the resulting observability estimates derived in [27] provide the main tool for transfer of the energy.

Another focus of this paper is nonlinearity of the damping and the source. The damping-source problems for the wave equation were studied first in the context of fully internal damping ([7], [10], [20], and [26]). The results were obtained for dissipative sources (i.e., sinks) with a polynomial structure imposed on the dissipation ([10], [24] and references therein). For boundary or localized damping (see [13], [14] and references therein) sinks were assumed subcritical and structured, while damping had *linear bounds* at infinity, along with a polynomial structure at the origin.

The first paper dispensing with any bounds imposed on *damping at the origin* is [18]. Subsequently, [21] and [22] extended Lyapunov's technique of [13] to cover some subclasses of nonlinear C^1 dissipative functions. Most recently, [1] developed a weighted energy method that improves upon [21] and [22] and is closer in terms of the results to those in [18]; this method predicts rather explicit decay rates including good control of the constants. Optimality of the estimates is also asserted in [1].

The presence of the *sources* along with boundary dissipation was studied in [4]: the decay rates do not require polynomial growth conditions imposed on the damping at the origin, but these rates are suboptimal (as in [21]). Boundary damping with a linearly bounded forcing dependent on ∇u and with smallness restrictions imposed on the forcing term were studied in [8]; these restrictions make it possible to use standard multiplier methods for establishing energy decay.

All the works referenced above deal with damping that is *linearly bounded* at infinity from below and above. The bound from below was removed in [2] for the *internally* damped wave equations. In the case of boundary damping that is bounded at infinity [13] derives algebraic decay rates for *strong solutions* $H^2(\Omega) \times H^1(\Omega)$ only. In addition, the damping in [13] is subject to a polynomial growth at the origin. To the best of our knowledge there are no results on decay rates for boundary-localized damping that is not bounded linearly at infinity (except for [3], which deals with a different problem of considering smaller support of the damping for a pure wave equation (no source) and obtaining very weak logarithmic decay rate for very smooth (classical) solutions).

The goal set in this paper is to provide a unified treatment and methodology for derivation of optimal decay rates for a source-damping-driven wave equation. In short, the key novel features of our main result as stated in Theorem 3.1 can be summarized as follows: (i) Neumann versus Dirichlet problem with unobserved Neumann portion of the boundary, (ii) presence of unstructured sources with critical exponents, and (iii) unrestricted growth of dissipation both at the origin and infinity.

4 Computation of the decay rates

We conclude with several examples illustrating explicit decay rates obtained by our method.

4.1 Examples I: Linearly bounded damping at infinity

Throughout this section assume (Ag-I1).

Example 4.1: Linear damping at the origin
In this case we obtain exponential decay rates. Suppose $g(s)s = ks^2$, $|s| < 1$ for some $k > 0$. Take $(h_0 + h_1 + h_2)(s) = ks$, whence $E(t) \leq E(0)$ $\exp\left[-K\left(t/T - 1\right)\right]$, for a constant $K > 0$ and T given by Theorem 3.1.

Remark
We will not keep track of the constants; parameters c and c_0, below, need to be adjusted individually for each case.

Example 4.2: Superlinear exponential damping at the origin
Take $g(s) = s^3 e^{-1/s^2}$ for $0 < |s| < 1$. The damping at the origin is very weak because all derivatives of g vanish at 0. The decay rates we obtain are logarithmic. The map $t \mapsto \sqrt{t}\, g(\sqrt{t}) = \sqrt{t} e^{-1/t}$ is convex on $[0, \varepsilon)$, small $\varepsilon > 0$, and we solve $\dot{S}(t) + CS(t)^2 \exp\left(-S(t)^{-1}\right) = 0$, getting $S(t) = \left[\ln(ct + c_0)\right]^{-1}$.

This example was also considered in [1] for the problem without a source and with pure Dirichlet BC; [1] derives logarithmic decay rates using methods of weighted energy inequalities. This confirms optimality of our results obtained for non-Lopatinski boundary conditions and for the model with a source.

Example 4.3: Sublinear damping at the origin
Suppose $g(s)s = ms^{\theta+1}$, if $|s| < 1$ with $\theta \in (0, 1)$. Based on Corollary 3.21, we are led to analyzing $S_t + CS^{(\theta+1)/(2\theta)} = 0$, where constant C may depend on the initial energy $E(0)$. Solving the differential equation leads to $S(t) = \left(ct + c_0\right)^{-2\theta/(1-\theta)}$. Thus we have algebraic decay rates that converge to exponential as $\theta \nearrow 1$.

4.2 Examples II: Sublinear and superlinear dissipation at infinity.

Now consider sub- and superlinear damping at infinity and linearly bounded at the origin (Ag-O3.3.1). For simplicity, we assume $f \equiv 0$, so that regular initial data imply the corresponding regularity of the solution.

Example 4.4: Sublinear damping at infinity
Assume (Ag-I2), then the energy decays are algebraic and depend on the dimension of the space.

- **Case** $n = 2$. Applying Corollary 3.3.1 with $p \nearrow \infty$ we obtain $\frac{p-2}{p-\theta-1} \nearrow 1$. The differential equation to consider is $S_t + CS^{\frac{p-2}{p-\theta-1}} = 0$, where constant $C > 0$ depends on $E_s(0)$. Define $\omega = \frac{p-2}{p-\theta-1}$, we compute $S(t) = (ct + c_0)^{1/(1-\omega)}$ where $\omega \nearrow 1$ and, therefore, the decay exponent $1/(1-\omega)$ can be made arbitrarily large. More precisely, we obtain exponential energy decay if we use maximal available regularity $p = \infty$.
- **Case** $n = 3$. Assume the damping is saturated: $\theta = 0$, and suppose $(u_0, u_1) \in \mathcal{H}^1(\Omega)$. The differential equation takes the form $S_t + CS^{5/4} = 0$, with C dependent on the $\mathcal{H}^1(\Omega)$-norm of (u_0, u_1). Then $S(t) = (ct + c_0)^{-4}$.

  If we take the initial conditions only incrementally more regular (i.e., assume $E_s(0) = \|(u_0, u_1)\|_{\mathcal{H}^s(\Omega)} < \infty$ with $s > 0$) we obtain $S(t) = (ct + c_0)^{-\omega}$. Now $\omega = \omega(s) \searrow 0$ when $s \searrow 0$, and parameter C depends on $E_s(0)$.

Example 4.5: Superlinear damping at infinity
Assume (Ag-I3). Classify the results according to the dimension of the domain:

- **Case** $n = 2$. Let $(u_0, u_1) \in \mathcal{H}^s(\Omega)$, any $s > 0$. Based on Corollary 3.3.2, the ODE in question is $S_t + CS^\omega = 0$, where parameter $\omega > 1$ can be chosen arbitrarily close to 1 because regularity p in this case can go to ∞. Solving the equation yields $S(t) = (ct + c_0)^{1/(1-\omega)}$ where C depends on $E_s(0)$. Thus we can take an arbitrary damping exponent $r \in (0, \infty)$ (in (Ag-I3)), provided the initial conditions are bounded in $\mathcal{H}^s(\Omega)$, any $s > 0$.
- **Case** $n = 3$. In this instance, take the initial condition from $\mathcal{H}^1(\Omega)$. Energy decay rates are driven by ODE $S_t + CS^{6/5} = 0$, where C depends on $E_1(0)$. Consequently $S(t) = (ct + c_0)^{-5}$.

4.3 Examples III: Combining different types of damping

When combining dissipation with different characteristics at the origin and infinity, we get the slowest decay guaranteed by the damping.

Example 4.6: Exponential at the origin and saturated at infinity
Let $n = 3$ and assume absence of a source: $f = 0$. For best possible decay rates let $(u_0, u_1) \in \mathcal{H}^1(\Omega)$. Suppose dissipation is sublinear at infinity

(Ag-I2) and

$$g(s) = s^3 e^{-1/s^2} \quad \text{for } 0 < |s| < 1$$

From Theorem 3.1 it follows that we may take h_2 to be linear. Map h_0 can be chosen to satisfy $h_0[g(s)s] \geq s^2$ whenever $|s| < 1$, whereas, $h_1(t) = t^{4/5}$ by Corollary 3.31..

ODE (3.8) reduces to $S_t + Ch_0^{-1}\left(\frac{S}{C_{\rm dec}}\right) = 0$, which results in logarithmic decay.

Example 4.7: Polynomial at the origin and sublinear at infinity Take $g(s) = s^p$ for $|s| < 1$ some $p > 1$, and consider sublinear damping at infinity (Ag-I2) with exponent $0 \leq \theta < 1$ (whence h_2 can be linear). From Corollaries 3.2.2 and 3.3.1 we have $h_0(t) = t^{2/(p+1)}$, and $h_1(t) = t^{4/(5-\theta)}$. Because $S(t)$ in (3.8) tends to 0 as $t \to \infty$, the growth of $S(t)$ for large t is determined by the higher-order term among h_0, h_1. For example, choose h_0, if

$$2/(p+1) > 4/(5-\theta) \tag{4.13}$$

or h_1 otherwise. Either of the two implies algebraic energy decay; note, however, that we pick up the weakest rate. Suppose (4.13) holds; consider two equations ($i = 0, 1$): $\dot{S}_i(t) + Ch_i^{-1}(S_i(t)) = 0, \quad S_i(0) = E(0)$,

$$S_0(t) = (ct + c_0)^{-(p-1)/2}, \quad S_1(t) = (ct + c_0)^{-(1-\theta)/4} \tag{4.14}$$

Because we choose solution (4.14) corresponding to parameter p, we obtain from (4.13): $\frac{p-1}{2} < \frac{1-\theta}{2}$. Thus we end up with the slower decay among the two given in (4.14).

Example 4.8: Polynomial at the origin and superlinear at infinity Assume $n = 3$ with smooth initial data $(u_0, u_1) \in \mathcal{H}^1(\Omega)$, and $f \equiv 0$. If $|s| < 1$, let $|g(s)| = |s|^p$ for some $p > 1$. Whereas at infinity, when $|s| \geq 1$, assume $|g(s)| = s^5$. According to Corollaries 3.2.2 and 3.3.2, we may take $h_0(t) = t^{2/(p+1)}$, $h_2(t) = t^{5/6}$.

(i) First, suppose $p < 7/5$, that is $5/6 < 2/(p+1)$, then $h_0(t) \gg h_1(t)$ when t is small. Hence, as the proof of Corollary 3.2 indicates, map q in (3.8) may be replaced with h_0^{-1} and $S(t) = (ct + c_0)^{-5}$

(ii) If $p > \frac{7}{5}$, then we replace q in (3.8) by h_2^{-1} obtaining $S(t) = (ct + c_0)^{-\frac{2}{p-1}}$

Note that in case (i) we had $5 < \frac{2}{p-1}$, while in (ii) this inequality is reversed, that is the weakest rate dominates in each situation.

References

[1] Alabeau, F. On convexity and weighted integral inequalities for energy decay rates of nonlinear dissipative hyperbolic systems, *Applied Mathematics and Optimization*, to appear.

[2] Bucci, F. Uniform decay rates of solutions to a system of coupled PDE's with nonlinear internal dissipation. *Diff. and Integral Equations*, 16, 865–896 (2003).

[3] Bellassoued, M. Decay of solutions of the wave equation with arbitrary localized nonlinear damping, *J. Differential Equations*, 211, 305–332 (2005).

[4] Cavalcanti, M., Cavalcanti, V., & Martinez, P. Existence and decay rates for the wave equation with nonlinear boundary damping and source term, *J. Differential Equations*, 2004.

[5] Chueshov, I., Eller, M., & Lasiecka, I. On the attractor for a semilinear wave equation with critical exponent and nonlinear boundary dissipation, *Comm. PDE*, 27, 1901–1951 (2002).

[6] Chen, G. A note on boundary stabilization of the wave equation. *SIAM J. Control Optimization*, 19, 106–113 (1981).

[7] Georgiev, V. & Todorova, G. Existence of solutions of the wave equation with nonlinear damping and source term, *J. Differential Equations*, 109, 295–308 (1994).

[8] Guesmia, A. A new approach of stabilization of nondisipative distributed systems, *SIAM J. Control*, 42, 24–52 (2003).

[9] Grisvard, P. *Elliptic Problems in Nonsmooth Domains*, Pitman, 1985.

[10] Haraux, A. *Semilinear Hyperbolic Problems in Bounded Domains*, Mathematical Reports, vol. 3, Gordon & Breach, New York, 1987.

[11] Haraux, A. *Nonlinear Evolution Equations—Global Behaviour of Solutions*, Springer-Verlag, Heidelberg, 1981.

[12] Isakov, V. & Yamamoto, M. Carleman estimates with the Neumann boundary conditions and its applications to the observability inequality and inverse problems, *Differential Geometric Methods in the Control of PDE's*, Contemporary Mathematics, vol. 286, AMS, Providence, RI, 2000, 191–227.

[13] Komornik, V. *Exact Controllability and Stabilization, The Multiplier Method*, Collection RMA, Masson-John Wiley, Paris, 1994.

[14] Lagnese, J. Decay of the solution of the wave equation in a bounded region with boundary dissipation, *J. Diff. Equations*, 50, 163–182 (1983).

[15] Lasiecka, I., Lions, J. L. & Triggiani, R. Nonhomogenous boundary value problems for second order hyperbolic equations, *J. Math Pure et Appliques*, 65, 149–192 (1986).

[16] Lasiecka, I. *CBMS Lecture Notes on Mathematical Control Theory of Coupled PDE-s*, SIAM, Philadelphia, 2002.

[17] Lasiecka, I., Triggiani, R., & Zhang, X. Nonconservative wave equations with unobserved Neumann B.C.: Global uniqueness and observability in one shot, *Differential Geometric Methods in the Control of PDE's*, Contemporary Mathematics, vol. 268, AMS, Providence, RI, 2000, 227–325.

[18] Lasiecka, I. & Tataru, D. Uniform boundary stabilization of semilinear wave equation with nonlinear boundary dissipation. *Diff. and Integral Equations*, 6, 507–533 (1993).

[19] Lasiecka, I. & Triggiani, R. Uniform stabilization of the wave equation with Dirichlet or Neumann feedback control without geometrical conditions, *Appl. Math. Optim.*, 25, 189–224 (1992).

[20] Lions, J. L. & Strauss, W. Some nonlinear evolution equations, *Bull. Soc. Math. France*, 93, 43–96 (1965).

[21] Martinez, P. A new method to obtain decay estimates for dissipative systems with localized damping, *Rev. Mat. Complut.*, 12, 251–283 (1999).

[22] Martinez, P. A new method to obtain decay estimates for dissipative systems, *ESAIM Contr. Opt. Calc. Var.*, 4, 419–444 (1999).

[23] Morawetz, C., Ralston, J., & Strauss, W. Decay of solutions of the wave equation outside nontrapping obstacles, *Comm. Pure Appl. Math.*, 30, 447–508 (1977).

[24] Nakao, M. Decay of solutions of the wave equation with a local nonlinear dissipation, *Math. Ann.*, 305, 403–417 (1996).

[25] Ralston, J. Solutions of the wave equation with localized energy, *Comm. Pure Appl. Math.*, 22, 807–823 (1969).

[26] Serrin, J., Todorova, G., & Vitillaro, E. Existence for a nonlinear wave equation with damping and source term, *Integral and Differential Equations*, 16, 13–50 (2003).

[27] Toundykov, D. Decay rates for solutions to semilinear wave equation with localized Neumann-type nonlinear damping and critical exponents source terms. Preprint (2005).

[28] Vancostenoble, J. & Martinez, P. Optimality of energy estimates for the wave equation with nonlinear boundary velocity damping, *SIAM J. Control*, 39, 776–797 (2000).

CHAPTER 14

Electromagnetic $3D$ reconstruction by level set with zero capacity connecting sets

Claude Dedeban
France Telecom R et D ANT and OpRaTel, Fort de la Tête de Chien,
La Turbie, France

Pierre Dubois
France Telecom R et D ANT and INRIA, Fort de la Tête de Chien,
La Turbie, France

Jean-Paul Zolésio
CNRS and INRIA, Sophia-Antipolis, France

1 Introduction

The *inverse scattering* problem in electromagnetics is studied through the identification or *reconstruction* of the obstacle considered as a *nonsmooth surface* in R^3. Through measurements of E_m of the scattered electric field in a *non-far* zone θ, we consider the classical minimization of a functional measuring the distance between E_m and the actual solution E over θ. Using the shape derivative of the functional, we introduced the level-set method in $3D$. Using the $SR3D$ (scattering structures $3D$) software, based on the Rumsey principle, and several graphics algorithms, we construct the optimization of the level-set method for a $3D$ surface.

2 Electromagnetic scattering

We recall the analysis for the shape gradient in a $3D$ electromagnetic field in the presence of a scattering surface B is a bounded body in $\mathcal{R}^3$. Γ is its boundary, $\Gamma = \partial B$.

Ω is the outer domain $\Omega = \mathcal{R}^3 \setminus \bar{B}$, $\bar{B} = B \cup \Gamma$ (Γ shall have no relative boundary). The electromagnetic field$(\vec{E}, \vec{H}\ )$ is characterized by the Maxwell equation (2.1), in a homogenous domain Ω, where are defined the following coefficients: $\varepsilon \in C$ the electrical permitivity, μ the magnetic permitivity and σ its conductivity.

E_i is data that represent the incident electrical field. Notice that E_i is a physical field that verifies the Helmholtz equations. The radiating problem $E_s \in H^1(\Omega)$ such that

$$\begin{cases} curl\ curl\ \vec{E_s} - k^2\vec{E_s} = 0 & \text{on } \Omega \\ \vec{E_s} \wedge n = -\vec{E_i} \wedge n & \text{on } \Gamma \\ \lim_{r\to\infty} r(\partial_r \vec{E_s} + ik\vec{E_s}) = 0 & \end{cases} \tag{2.1}$$

is a well-posed problem that has a unique solution $\in \Omega$; see [7]. Charge conservation ($divE = 0$ for homogenous and isotropic domain) is assumed for this solution.

3 Inverse problem

We focus on the inverse problem, which is to determine the metallic antenna shape when we know the incident field E_i and the scattered field E_s in a given region θ. References [11], [12], and [13] cover the inverse problem solution in several configurations involving incidencies and frequencies.

We introduce the following continuous transformation T_r defined by:

$$T_r\ :\begin{cases} R^3 \longrightarrow R^3 \\ T_r(V)(x) = x + \int_0^r \vec{V}(s, T_s(V)(x))ds := x_r \\ Tr(\Omega) = \Omega_r \end{cases} \tag{3.2}$$

We define the cost function in a fixed region θ.

$$J(\Omega_r) = \int_\theta |E_m - E_d|^2 d\gamma \tag{3.3}$$

where E_d is a given value that results in a scattered wave near an antenna's surface (Γ). (It can be used to optimize an infty scattered diagram.)

The analysis of the shape gradient ([13]) gives us the following direct problem to solve u:

$$\begin{cases} curlcurlu - k^2u = 0 & on\ \Omega \\ u \wedge n = 0 & on\ \Gamma \\ \lim_{r\to\infty} r(\partial_r \vec{u} + ik\vec{u}) = 0 & \end{cases} \tag{3.4}$$

The adjoint field v is the unique solution of the well-posed problem ([6]):

$$\begin{cases} curlcurl\ p - k^2 p = -2\chi_\theta(\overline{E_m - u}) \ over\ \ \Omega \\ p \wedge n = 0 \qquad\qquad\qquad\qquad\quad over\ \ \Gamma \\ \lim_{r\to\infty} r(\partial_r \vec{p} + ik\vec{p}) = 0 \end{cases} \tag{3.5}$$

If we let (u, p) be solutions of the direct and adjoint problems, we have the Lagrangian derivative with respect to r, which takes the following form:

$$\partial_r \mathcal{L}(r, u, p)|_{r=0} = \Re \int_\Gamma v(curl(u)curlp - k^2 u.p) \tag{3.6}$$

$$+\Re \int_\Gamma v(div_\Gamma(curlu \wedge n)\langle p, n\rangle + div_\Gamma(curlp \wedge n)\langle u, n\rangle) \tag{3.7}$$

$$+\Re \int_\Gamma 2v\langle curlu \wedge n, curlp \wedge n\rangle \tag{3.8}$$

3.1 Computation of direct and adjoint problems by harmonic integral equations (SR3D software)

We use the $SR3D$ software to solve the $3D$ harmonic Maxwell equation (2.1) in isotropic and homogenous medium($\varepsilon, \mu,$ and $\sigma \in \mathcal{R}$). From the Huygens principle, this software is based on the electric relations integral equations and the Rumsey reaction principle:

$$\sum_{l=1}^{N} \mu_{rl} Q_{S_l}(S_l, \vec{j_l}, \vec{j_l^t}) + \frac{k_l^2}{\mu_{rl}} Q_{S_l}(S_l, \vec{p_l}, \vec{p_l^t}) - P_{S_l}(S_l, \vec{j_l}, \vec{p_l^t}) - P_{S_l}(S_l, \vec{p_l}, \vec{j_l^t}) \tag{3.9}$$

$$= -\sum_{l=1}^{N} \oint_{S_l} (\vec{E_l^i}(x).\vec{j_l^t} - \vec{H_l^i}(x).\vec{p_l^t})\ ds(x) \tag{3.10}$$

with

$$Q_{S_l}(S_l, \vec{j_l}, \vec{j_l^t}) = \oint_{S^t} \oint_S G(x, y)(\vec{j}(y).\vec{j^t}(x)$$

$$-\frac{1}{k^2} div_S \vec{j}(y).div_{S^t} \vec{j^t}(x))ds(y)ds^t(x) \tag{3.11}$$

and

$$P_{S_l}(S_l, \vec{j_l}, \vec{p_l^t}) = \oint_{S^t} \oint_S (grad_x G(x, y) * \vec{j}(y)).\vec{p^t}(x)ds(y)ds^t(x) \tag{3.12}$$

where $\vec{p_l}, \vec{p_l^t}, \vec{j_l}$ and $\vec{j_l^t}$ are the equivalent currents on the S_l surface interface. These formulations allow us to determine the values of unknown electric and magnetic currents based on knowledge of incident excitation. We have to determine the incident field on the surface before to solve the Rumsey reaction.

3.2 Solution of radiating dipole in free space

The incident field E_i generated by a dipole, which is characterized by its moment $\vec{p}$ and its application point (x_0) (the data), is the solution of the following well-posed problem in free space:

$$\begin{cases} curl\ E_i - i\omega\mu H_i = 0 & \in \mathcal{R}^3 \\ curl H_i + i\omega\varepsilon E_i = \vec{p}\chi_{x_0} & \\ \text{Radiation condition at infty} & \end{cases} \tag{3.13}$$

where χ_0 is dirac mass at x_0. Dismissing $\vec{h}$ from the system, and introducing, Sommerfield condition, we have the analytic expression of incident field E_i radiated by a dipole:

$$\vec{E_i} = i\omega\mu k \frac{e^{ikR}}{4\pi kR} \left\{ \left(-1 - \frac{3i}{kR} + \frac{3}{(kR)^2}\right) \left(\vec{p}.\frac{\vec{R}}{R}\right) \frac{\vec{R}}{R} + \left(1 + \frac{i}{kR} - \frac{1}{(kR)^2}\right) \right\} \tag{3.14}$$

with $\vec{R} = \vec{r} - \vec{r_0}$; $\vec{r_o}$ is the radius field at point x_0; and $R = |\vec{R}|$.

This expression of the solution of the radiating dipole enables us to introduce the calculus in integrals based on Rumsey's reaction in order to solve Equation (2.1) when the incident field E_i is generated by one or several dipoles.

3.3 Solution of adjoints sources in free space

We introduce $\vec{h} \in H^1(\Omega)$ verifying $curl\ p = i\omega\mu\vec{h}$. It implies $curl\ (i\omega\mu\vec{h}) - k^2 p = -2\chi_\theta(\overline{E_m - u})$.

Finally we have the following adjoint problem:

$$\begin{cases} curl\ p - i\omega\mu h = 0 & over\ \ \Omega \\ curl\ h + i\omega\varepsilon p = -2\chi_\theta(\overline{E_m - u}) & over\ \ \Omega \\ p \wedge n = 0 & over\ \ \Gamma \\ \lim_{r\to\infty} r(\partial_r \vec{p} + ik\vec{p}) = 0 & \end{cases} \tag{3.15}$$

This problem is equivalent to (3.13) with: $\vec{p} = -2(\overline{E_m - u})$.

So the problem adjoint can be seen as a scattering classical problem with a complex source.

The two problems can be solved with $SR3D$ numerical software based on Rumsey's principle (with second member modification). Then we will integrate the gradient density on the surface (Γ) to evaluate the shape gradient.

3.4 Level-set equation

Let t be the evolution parameter Ω_t, an open bounded $\mathcal{R}^3$. Ω_0 is given. $\vec{n}_t$ is the unitary normal to $\partial\Gamma_t$. Ω_t is constructed by the flow of the field V and we denote $\Omega_t = \Omega_t(V) := T_t(V)(\Omega)$.

$$\Omega_t^{\phi} = \{x \in \mathcal{R}^3 | \quad \Phi(t,x) < 0\} \tag{3.16}$$

The representation of domains by the Eulerian evolution is very easily adaptable to any parameterization. It will suffice to build, for any specific parameterization, the speed vector field V whose flow mapping follows the exact geometrical perturbations induced by the evolution of the parameter (see several examples in [14]). Among all examples of domain parameterizations, the level-set has a good feature with respect to the topological changes. The speed vector

$$V(t,x) = -\frac{\partial}{\partial t}\Phi \quad \frac{\nabla_x \Phi}{||\nabla\Phi||^2} \tag{3.17}$$

follows the evolution of that domain in the sense that its flow mapping does the job:

$$\Omega_t = T_t(V)(\Omega_0)$$

It is well known that the so-called *gradient method* decreases the functional.

Let $J(\Omega)$ be a shape functional. We denote by $dJ(\Omega, V)$ its Eulerian derivative (when it exists). When the mapping $V \to dj(\Omega, V)$ is linear and continuous there exists a gradient $G(\Omega)$, a vector distribution supported by the boundary $\partial\Omega$. The speed method consists of the choice of the speed vector field $V(t)$ opposite to the distribution $G(\Omega_t)$ in the following sense

$$V(t) = -A^{-1} \cdot G(\Omega_t) \tag{3.18}$$

where A is an appropriate duality operator. As we have $j(\Omega_t) = j(\Omega_0) + \int_0^t d(\Omega_s, V(s))\, ds$, we get

$$j(\Omega_t) = j(\Omega_0) - \int_0^t \langle A.V(s), V(s)\rangle ds \leq J(\Omega_0) - \alpha \int_0^t |V(s)|^2 ds \tag{3.19}$$

Obviously $t \to J(\Omega_t)$ is a decreasing function.

The level-set method consists of the following particular case of Equation 3.17 in Equation 3.18, that is,

$$-\frac{\partial}{\partial t}\Phi \quad \frac{\nabla_x \Phi}{||\nabla\Phi||^2} = -A^{-1} \cdot G(\Omega_t) \tag{3.20}$$

for which 3.20 becomes

$$\boxed{j(\Omega_t) \leq \ j(\Omega_0) - \alpha \int_0^t \left(\frac{\partial}{\partial t}\Phi \ ||V(s)||^{-1}\right)^2 \ ds} \tag{3.21}$$

Notice that Equation 3.20 is a vector equation. Its normal component is obtained in multiplying by $\nabla\phi$, and we get

$$-\frac{\partial}{\partial t}\Phi = -\langle A^{-1}\cdot G(\Omega_t),\ \nabla\Phi\rangle \tag{3.22}$$

Of course 3.22 dœs not imply 3.20; we must complete the tangential component

$$[A^{-1}\cdot G(\Omega_t)]_{\Gamma_t} = 0.$$

3.5 3D algorithm

The technique of *level-set* allows us to move mesh without moving vertex. In fact we will use a new voxel $3D$ grid to express the level-set equation solution on a corner of the vox. When the level-set function solution is known on each vertex of the new grid, we will interpolate the new isosurface of level 0 of ϕ by a triangular mesh. To interpolate the new isosurface we use the marching cube algorithm.

3.6 Narrow-band construction

In order to evaluate the level-set function $\phi_{t+1}\in R^3$ we create a large box that contains Γ. We determine a discretization step (we use the same step as the original triangular mesh $\lambda/5$), and we discretize the box and have a grid composed of voxels. After that, we dismiss all of the voxels that are not on the tubular nearness ((3.23) with $h=\lambda/3$).

3.7 Signed (or oriented) distance evaluation

The solution of the level-set equation is initialized with the oriented distance function to the antenna surface $\phi_{t=0}(X(t=0)) = b(\Gamma,X); X\in R^3$, introduced in [2].

The initial antenna surface is the isosurface of level 0 of the function b. $b(\Gamma,X)=0$ if $X\in\Gamma$

To determine the b function on each grid vertex, we must know if the vertex is inside or outside of the antenna and know the smallest distance between the vertex and Γ. There are some techniques to determine adherence to the volume, such as solid angulus method. We have used a volume cutting by frame and calculated the intersection between antenna and each frame (z=0) in order to transform the 3D problem into 2D. Then we do an angulus calculation to determine adherence to the 2D curve of the vertex.

3.8 Gradient on narrow-band

Given that $h>0$, we notice $u_h(\Gamma)$ the tubular neighborhood of $\Gamma\in h$: $u_h(\Gamma)=\bigcup_{x\in\Gamma}B_h(x)$, where $B_h(x)$ is the open sphere, x its center and h its radius. When h is small enough, $\exists$ is a unique p defined by $U_h(\Gamma)$, which is the projection operator.

$$\forall x \in U_h(\Gamma)\|p(x) - x\| \leq \|y - x\| \quad \text{and} \quad y \in \Gamma \tag{3.23}$$

Let g be defined on Γ; we denote by $g \circ p$ its extension to the tubular neighborhood.

From a numerical point of view, we choose $\lambda/3 > h > \lambda/5$. When h is bigger than $\lambda/3$, some parasite shape is generated on the end of the narrow band. The extension used on our method is

$$(gop)(X) = g(M) \text{ with } X \text{ and } M \in R^3$$

where M is the projection of X on Γ $\ MX = min_{Non\Gamma}NX$

3.9 Interpolation by marching cube method

When the value of the ϕ_{t+1} is known on each vertex, we must interpolate the isosurface 0 with a triangular mesh to optimize the criterion. We use the marching cube algorithm (Figure 14.1). The principle is the following: we scan all voxels of the narrow band. For each voxel, we observe all ϕ_{t+1} values on the vertex. If the values are all positives or all negatives, the isosurface is exterior to the voxel; otherwise there exists an intersection on the concerned voxel. We calculate the intersection of the isosurface and each edge of the voxel by linear interpolation from the vertex value knowledge. When we have found the different points of intersection, the marching cube algorithm permits us to construct a triangular mesh in the voxel according to the number and position of intersection points.

In this, scaning all voxels, we reduce the homogenous triangular mesh little by little. We assume the new isosurface is included on the narrow band because we control the maximum deformation by the normalization coefficient α.

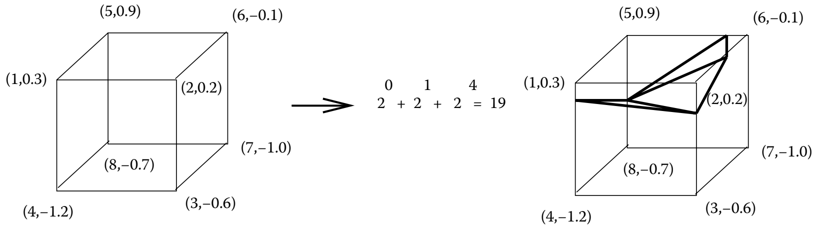


FIGURE 14.1 The marching cube algorithm.

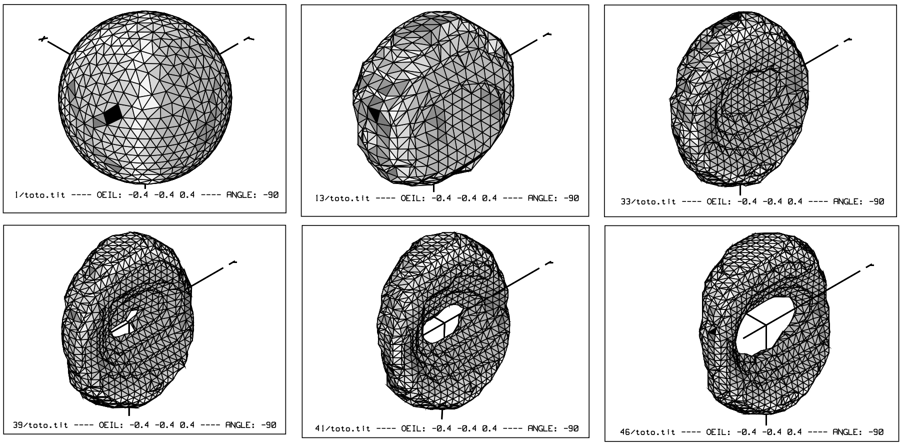

1/toto.tlt ---- OEIL: -0.4 -0.4 0.4 ---- ANGLE: -90
13/toto.tlt ---- OEIL: -0.4 -0.4 0.4 ---- ANGLE: -90
33/toto.tlt ---- OEIL: -0.4 -0.4 0.4 ---- ANGLE: -90
39/toto.tlt ---- OEIL: -0.4 -0.4 0.4 ---- ANGLE: -90
41/toto.tlt ---- OEIL: -0.4 -0.4 0.4 ---- ANGLE: -90
46/toto.tlt ---- OEIL: -0.4 -0.4 0.4 ---- ANGLE: -90

FIGURE 14.2 Convergence and separation.

3.10 3D reconstruction

In this section we present some cases of reconstruction. The incident field is composed of six dipoles placed at infinity (in the two senses of the three axes). The receptors (or adjoint sources) are placed around the structure and their numbers depend on the case.

We present here some cases, varying parameters as incidencies, receptors sources and topological congurations of the structures, observing, and making some comments, to explain a perfect optimization technique.

3.10.1 Topological changes: separation

The initial structure is a sphere (radius 160 mm) and the target is also a sphere (320 mm) (see Figure 14.2). We optimize the structure at 200, 500, and 800 MHz.

In this case, we can see the apparition of a structure with a topological change; however, the optimization method did not allow convergence to the ideal shape. In two directions we have a good attractive wave, and the shape converges to a good size in that direction. But in the third direction we are on the wave of a local minimum because the method converges to a local minimum. This case is very interesting because it proves that convergence is never certain and we must look for strategies of optimization in order to increase the probability of convergence.

3.10.2 Four spheres reunited from a star

We optimize the structure (see Figure 14.3) at 100, 200, 500, and 800 MHz. The initial structure is composed of four spheres placed as a star. Its radius

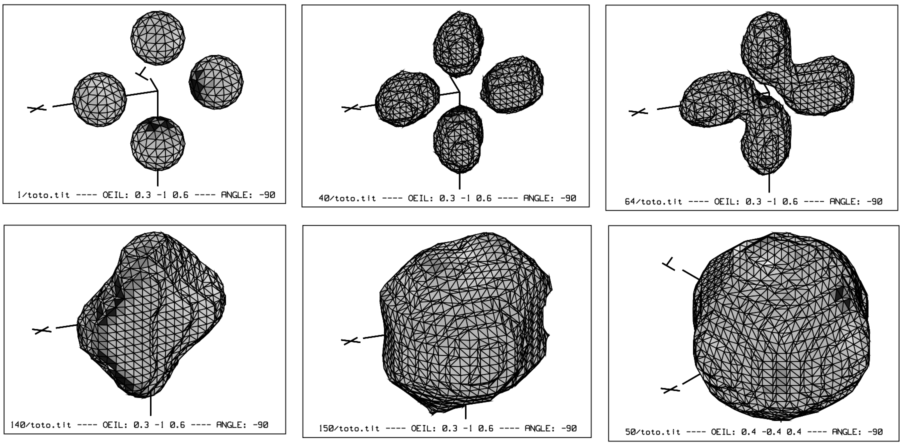

1/toto.tlt ---- OEIL: 0.3 -1 0.6 ---- ANGLE: -90
40/toto.tlt ---- OEIL: 0.3 -1 0.6 ---- ANGLE: -90
64/toto.tlt ---- OEIL: 0.3 -1 0.6 ---- ANGLE: -90
140/toto.tlt ---- OEIL: 0.3 -1 0.6 ---- ANGLE: -90
150/toto.tlt ---- OEIL: 0.3 -1 0.6 ---- ANGLE: -90
50/toto.tlt ---- OEIL: 0.4 -0.4 0.4 ---- ANGLE: -90

FIGURE 14.3 Four spheres reunited from a star.

is 100 mm. The distance between the spheres is 160 mm. The ideal structure is a sphere of 350 mm centered.

The joining up of the four spheres is complete after 600 iterations. We notice attraction takes more time than separation. This is the consequence of superposition of the gradient when two surfaces meet.

3.10.3 Separation of three spheres with simultaneous optimization

We work at 500, 800, 1100 and 1300 MHz. The initial structure is a parallelpip (Figure 14.4) and the ideal structure is composed of three spheres. Its radius is 100 mm.

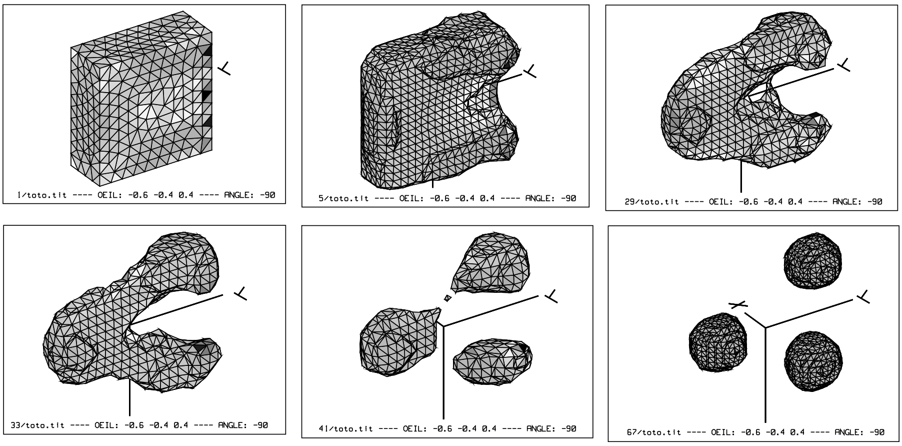

1/toto.tlt ---- OEIL: -0.6 -0.4 0.4 ---- ANGLE: -90
5/toto.tlt ---- OEIL: -0.6 -0.4 0.4 ---- ANGLE: -90
29/toto.tlt ---- OEIL: -0.6 -0.4 0.4 ---- ANGLE: -90
33/toto.tlt ---- OEIL: -0.6 -0.4 0.4 ---- ANGLE: -90
41/toto.tlt ---- OEIL: -0.6 -0.4 0.4 ---- ANGLE: -90
67/toto.tlt ---- OEIL: -0.6 -0.4 0.4 ---- ANGLE: -90

FIGURE 14.4 Separation of three spheres with simultaneous optimization.

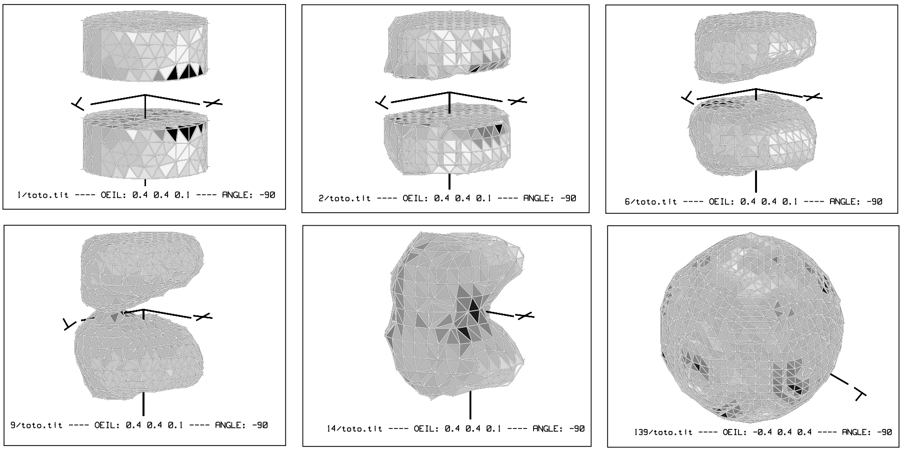

1/toto.tit ---- OEIL: 0.4 0.4 0.1 ---- ANGLE: -90
2/toto.tit ---- OEIL: 0.4 0.4 0.1 ---- ANGLE: -90
6/toto.tit ---- OEIL: 0.4 0.4 0.1 ---- ANGLE: -90
9/toto.tit ---- OEIL: 0.4 0.4 0.1 ---- ANGLE: -90
14/toto.tit ---- OEIL: 0.4 0.4 0.1 ---- ANGLE: -90
139/toto.tit ---- OEIL: -0.4 0.4 0.4 ---- ANGLE: -90

FIGURE 14.5 Two cylinders reunited in a sphere with translation.

3.11 Two cylinders reunited in a sphere with translation

Frequencies: 200 MHz, 500 MHz, 800 MHz. The ideal structure (Figure 14.5) is a metallic sphere (radius = 300 mm), translated from 200 mm with respect to y with respect to the center.

We can see again the convergence of the reconstruction process. We present here a topological reunion of two singular shapes.

The optimization method we have presented, based on deformation by isosurface, is original because for the first time the incidencies are unconstrained. We found several works on level-sets in $2D$ TM kind, which is extended to the $3D$ tubular kind. And the $3D$ reconstruction allows us to discover some different topological changes.

The success of the reconstruction process is never guaranteed (see case 1) and the only way to increase the rate of convergence is to use a good strategy of optimization. We can play on polarization or on the length between dipole incidencies and initial structure. In our validation test, the principal aspect that allows the reconstruction to succeed is the multifrequencies strategy. Work with low frequencies increases the attractive wave of the gradient (around the initial shape) because the wavelength is bigger. And work with high frequencies allows us to detect small deformations and could be considered as a design.

References

[1] M. Moubachir and J.-P. Zolésio, *Moving Shape Optimization*, Pure and Applied Mathematics, vol. 277, Chapman & Hall, Boca Raton, FL, 2006.

[2] J.-P. Zolésio and M.C. Delfour, *Shapes and Geometries: Analysis, Differential Calculus, and Optimisation*, SIAM, Philadelphia, 2001.

[3] O. Dorn, H. Bertete-Aguirre, J.G. Berryman, and G.C. Papanicolaou, *A Nonlinear Inversion Method for 3D. Electromagnetic Imaging Using Adjoint Fields N*, to appear in *Inverse Problems.*

[4] J.-P. Zolésio, *Weak Shape Formulation of Free Boundary Problem*, Scuola normale superiore pise, 1994.

[5] J. Cagnol, M. Polis, and J.-P. Zolésio. *Shape Optimisation and Optimal Design*, Marcel Dekker Inc., New York, 2001.

[6] A. Bendali, Approximation par élements finis de surface de problèmes de diffraction éléctromagnétiques, Thèse de doctorat d'etat es-sciences, Université Pierre et Marie Curie Paris VI, 1984.

[7] J.-C. Nedelec, *Acoustic and Electromagnetic Equations: Integral Representation for Harmonic Problems*, Edition Springer, Heidelberg, 2001.

[8] R. Dautray and J.L. Lions, *Analyse Mathématique et calcul numérique pour les sciences et les techniques*, Edition Masson, Paris, 1984.

[9] P.F. Combes, *Micro-ondes: Circuits passifs, Antennes et Propagations*, Edition Dunod, Paris, 1997.

[10] D. Colton and R. Kress, *Integral Equation Method in Scattering Theory*, John Wiley & Sons, Inc., New York, 1983.

[11] V. Isakov, *On Uniqueness in the Inverse Transmission Scatterring Problem*, Communications in Partial Differential Equations, 1567–1587, 1990, Marcel Dekker Inc., New York.

[12] F. Hettlich, On the Uniqueness of the Inverse Conductive Scattering Problem for the Helmholtz Equation, *Inv. Prob.* 10, 129–144, 1994.

[13] P. Dubois and J.-P. Zolésio, Shape Gradient in Inverse Scattering, Chapter 8, this volume.

[14] M. Delfour and J.-P. Zolésio, Approximation of Non Linear Problems Associated with Radiating Bodies in Space, *SIAM J. Numer. Anal.* 24, 1077–1094, 1987.

CHAPTER 15

Shape and geometric methods in image processing

Mathieu Dehaes
Department of Mathematics and Statistics,
University of Montreal,
Montreal, Quebec, Canada

Michel C. Delfour
Center for Mathematics Research
and Department of Mathematics and Statistics,
University of Montreal,
Montreal, Quebec, Canada

1 Introduction

The study of linguistic and visual perceptions has been undertaken by several pioneering authors such as H. Blum [3] in 1967, D. Marr and E. Hildreth [20] in 1980 and D. Marr [21] in 1982. It involves specialists in psychology, artificial intelligence, and experimentalists such as Hubel and Wiesel [16] in 1962 and Campbell and Robson [6] in 1968.

In the first part of this paper we revisit the pioneering work of Marr and Hildreth [20] in 1980 on the smoothing of the image by convolution with a sufficiently differentiable normalized function as a function of the *scaling parameter*. We extend the *space-frequency uncertainty principle* to N-dimensional images. It is the analogue of the *Heisenberg Uncertainty Principle* in quantum mechanics. We revisit the *Laplacian filter* and generalize the *linearity assumption* of [20] from linear to curved contours.

In the second part, we show how shape analysis methods and the shape and tangential calculi can be applied to objective functions defined on the whole contour of an image. We review shape derivatives by the velocity method and their applications to snakes, active geodesic contours, and level sets. We show that the Eulerian shape semiderivative is the basic ingredient of those representations including the case of the oriented distance function. In all cases the evolution equationfor the continuous gradient descent method is shown to have the same structure. For more related material, the reader is referred to [10].

1.1 Notation

Given an integer $N \geq 1$, m_N and H_{N-1} will denote the N-dimensional Lebesgue and $(N-1)$-dimensional Hausdorff measures. The inner product and the norm in $\mathbf{R}^{\mathrm{N}}$ will be written $x \cdot y$ and $|x|$.

2 Automatic image processing

The first level of image processing is the detection of the contours or the boundaries of the objects in the image. For an ideal image $I : D \to \mathbf{R}$ defined in an open two-dimensional frame D with values in an interval of greys continuously ranging from white to black (see Figure 15.1), the edges of an object correspond to the loci of discontinuity of the image I (see [20]). These are called “step edges” in [21]. As can be seen from Figure 15.1, the loci of discontinuity may only reveal part of an object hidden by another one and a subsequent and different level of processing is required. A more difficult case is the detection of black curves or cracks in a white frame, where the function I becomes a measure supported by the curves rather than a function.

In practice, the frame D of the image is divided into periodically spaced cells P (square, hexagon, diamond) with a quantized value or pixel from 256 grey levels. For small squares, a piecewise linear continuous interpolation or higher degree C^1-interpolation can be used to remove the discontinuities at the intercell boundaries. In addition, observations and measurements introduce noise or perturbations and the interpolated image I needs to be further *smoothed* or *filtered*. When the characteristics of the noise are known, an appropriate filter can do the job (see, for instance, the use of *low pass filters* in [23–25]).

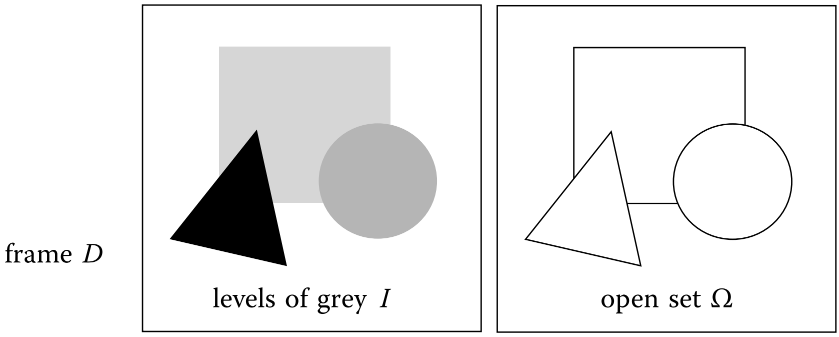


FIGURE 15.1 Image I of objects and their segmentation in the frame D.

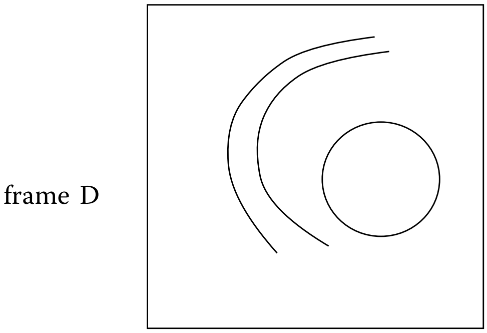


FIGURE 15.2 Image I containing black curves or cracks in the frame D.

This operation will be followed and/or combined with the use of an *edge detector* as the zero-crossings of the Laplacian. In the literature, the term filter often applies to both the filter and the detector. In this paper we shall make the distinction between the two operations.

3 Image smoothing and filtering by convolution and edge detectors

In this section we revisit the work of Marr and Hildreth [20] in 1980 on the smoothing of the image I by convolution with a sufficiently differentiable normalized function ρ as a function of the *scaling parameter* $\varepsilon > 0$. In the second part of this section we revisit the *Laplacian filter* and generalize the *linearity assumption* of [20] from linear to curved contours.

3.1 Construction of the convolution of I

Let $\rho : \mathbf{R}^N \to \mathbf{R}$, $N \geq 1$, be a sufficiently smooth function such that

$$\rho \geq 0 \quad \text{and} \quad \int_{\mathbf{R}^N} \rho(x)\, dx = 1.$$

Associate with the image I, ρ and a *scaling parameter* $\varepsilon > 0$ the normalized convolution

$$I_\varepsilon(x) \overset{\text{def}}{=} (I * \rho_\varepsilon)(x) = \frac{1}{\varepsilon^N} \int_{\mathbf{R}^N} I(y) \rho\left(\frac{x-y}{\varepsilon}\right) dy, \tag{3.1}$$

where for $x \in \mathbf{R}^N$

$$\rho_\varepsilon(x) \overset{\text{def}}{=} \frac{1}{\varepsilon^N} \rho\left(\frac{x}{\varepsilon}\right) \quad \text{and} \quad \int_{\mathbf{R}^N} \rho_\varepsilon(x)\, dx = 1. \tag{3.2}$$

The function ρ_ε plays the role of a probability density and $\rho_\varepsilon(x)\, dx$ of a probability measure. Under appropriate conditions I_ε converges to the original image I as $\varepsilon \to 0$.

For a sufficiently small $\varepsilon > 0$ the loci of discontinuity of the function I are transformed into loci of strong variation of the gradient of the convolution I_ε. When ρ has compact support, the convolution acts locally around each point in a neighborhood whose size is of order ε.

A popular choice for ρ is the *Gaussian* with integral normalized to one in $\mathbf{R}^N$ defined by

$$G^N(x) \overset{\text{def}}{=} \frac{1}{(\sqrt{2\pi})^N} e^{-\frac{|x|^2}{2}} \qquad \text{and} \qquad \int_{\mathbf{R}^N} G^N(x)\, dx = 1.$$

and the *normalized Gaussian of variance* ε

$$G^N_\varepsilon(x) = \frac{1}{\varepsilon^N} \frac{1}{(\sqrt{2\pi})^N} e^{-\frac{1}{2}\left|\frac{x}{\varepsilon}\right|^2} \qquad \text{and} \qquad \int_{\mathbf{R}^N} G^N_\varepsilon(x)\, dx = 1. \tag{3.3}$$

As noted in L. Alvarez, P.-L. Lions and J.-M. Morel [2] in dimension $N = 2$, the function $u(t) = G^2_{\sqrt{2t}} * I$ is the solution of the parabolic equation

$$\frac{\partial u}{\partial t}(t, x) = \Delta u(t, x), \quad u(0, x) = I(x).$$

3.2 Space-frequency uncertainty relationship

It is interesting to compute the Fourier transform $\mathcal{F}$ of G^N_ε for a vector x of $\mathbf{R}^N$ to explain the relationship between the mean square deviations of G^N_ε and its Fourier transform $\mathcal{F}(G^N_\varepsilon)$. For ω, a vector of $\mathbf{R}^N$, define the *Fourier transform* of a function f by

$$\mathcal{F}(f)(\omega) \overset{\text{def}}{=} \frac{1}{(\sqrt{2\pi})^N} \int_{\mathbf{R}^N} f(x) e^{-i\omega\cdot x}. \tag{3.4}$$

In applications the integral of the square of f often corresponds to an energy. We shall refer to the L^2-norm of f as the *energy norm* and the function $f^2(x)$ as the *energy density.*

The Fourier transform (3.4) of the normalized Gaussian (3.3) is given by

$$\mathcal{F}(G^N_\varepsilon)(\omega) \overset{\text{def}}{=} \frac{1}{(\sqrt{2\pi})^N} \int_{\mathbf{R}^N} G^N_\varepsilon(x) e^{-i\omega\cdot x} dx = \frac{1}{(\sqrt{2\pi})^N} e^{-\frac{1}{2}|\varepsilon\omega|^2}. \tag{3.5}$$

Marr and Hildreth [20] note that there is an uncertainty relationship between the mean square deviation with respect to the energy density $f(x)^2$ and the mean square deviation with respect to the energy density $\mathcal{F}(f)(\omega)^2$ of its Fourier transform in dimension one using a result of R. Bracewell's [5]

(160–161) (note that Bracewell [5] uses the constant $1/(2\pi)$ instead of $1/\sqrt{2\pi}$ in the definition of the Fourier transform, which yields $1/(4\pi)$ instead of $1/2$ as a lower bound to the product of the two mean-square deviations).

This *uncertainty relationship* generalizes to dimension N. First define the notions of *centroid* and *variance* with respect to the energy density $f(x)^2$. Given a function $f \in L^1(\mathbf{R}^N) \cap L^2(\mathbf{R}^N)$ and $x \in \mathbf{R}^N$, define the *centroid* as

$$\overline{x} \overset{\text{def}}{=} \frac{\int_{\mathbf{R}^N} x f(x)^2\, dx}{\int_{\mathbf{R}^N} f(x)^2\, dx} \tag{3.6}$$

and the *variance* as

$$\langle \Delta x \rangle^2 \overset{\text{def}}{=} \langle x - \overline{x} \rangle^2 = \frac{\int_{\mathbf{R}^N} |x|^2 f(x)^2\, dx}{\int_{\mathbf{R}^N} f(x)^2\, dx} - |\overline{x}|^2.$$

Theorem 3.1 (Uncertainty relationship)

Given $N \geq 1$ and a function $f \in W^{1,1}(\mathbf{R}^N) \cap W^{1,2}(\mathbf{R}^N)$ such that $-ixf \in L^1(\mathbf{R}^N) \cap L^2(\mathbf{R}^N)$,

$$\boxed{\langle \Delta x \rangle \langle \Delta \omega \rangle \geq \frac{N}{2}.} \tag{3.7}$$

For $f = G_\varepsilon^N$

$$f(x) = G_\varepsilon^N(x) = \frac{1}{\varepsilon^N} \frac{1}{(\sqrt{2\pi})^N} e^{-\frac{1}{2}\left|\frac{x}{\varepsilon}\right|^2} \quad \Rightarrow \langle \Delta x \rangle^2 = \frac{N\varepsilon^2}{2} \tag{3.8}$$

$$\hat{f}(\omega) = \mathcal{F}(G_\varepsilon^N)(\omega) = \frac{1}{(\sqrt{2\pi})^N} e^{-\frac{1}{2}|\varepsilon\omega|^2} \quad \Rightarrow \langle \Delta \omega \rangle^2 = \frac{N}{2\varepsilon^2} \tag{3.9}$$

$$\Rightarrow \boxed{\langle \Delta x \rangle \langle \Delta \omega \rangle = \frac{N}{2} \qquad \text{for } G_\varepsilon^N \text{ and } \mathcal{F}(G_\varepsilon^N).} \tag{3.10}$$

This shows that the *normalized Gaussian filter* is indeed an *optimal* filter because it achieves the lower bound in all dimensions as stated by Marr and Hildreth [20] for dimension one in 1980. In the context of quantum mechanics, this relationship is the analogue of the Heisenberg Uncertainty Principle.

3.3 Laplacian detector

One way to detect the edges of a regular object is to start from the convolution-smoothed image I_ε. Given a direction v, $|v| = 1$, and points $x \in \mathbf{R}^2$, an *edge point* will correspond to a local minimum or maximum of the directional derivative $f(t) \overset{\text{def}}{=} \nabla I_\varepsilon(x + tv) \cdot v$ with respect to t. Denote such a point by $\hat{t}$. Then a necessary condition is

$$D^2 I_\varepsilon(x + \hat{t}v)v \cdot v = 0.$$

The point $\hat{x} = x + \hat{t}v$ is a *zero-crossing* following the terminology of [20] of the second-order directional derivative in the direction v. Thus we are looking for the pairs $(\hat{x}, \hat{v})$ verifying the necessary condition

$$D^2 I_\varepsilon(\hat{x})\hat{v} \cdot \hat{v} = 0 \tag{3.11}$$

and, more precisely, lines or curves $\mathcal{C}$ such that

$$\forall x \in \mathcal{C}, \exists v(x), |v(x)| = 1, \quad D^2 I_\varepsilon(x) v(x) \cdot v(x) = 0. \tag{3.12}$$

This condition is necessary in order that, in each point $x \in \mathcal{C}$, there exists a direction $v(x)$ such that $\nabla I_\varepsilon(x) \cdot v$ be extremal.

In order to limit the search to points x rather than to pairs (x, v), the Laplacian detector was introduced in [20] under two assumptions: the linear variation of the intensity along the edges $\mathcal{C}$ of the object and the condition of zero-crossing of the second-order derivative in the direction normal to $\mathcal{C}$:

Assumption 3.1 (linear variation condition)
The intensity I_ε in a neighborhood of lines parallel to $\mathcal{C}$ is locally linear (*affine*)*.*

Assumption 3.2
The zero-crossing condition is verified in all points of $\mathcal{C}$ in the direction of the normal n at the point, that is,

$$\boxed{D^2 I_\varepsilon n \cdot n = 0 \quad \textit{on } \mathcal{C}.} \tag{3.13}$$

Under those two assumptions and for a line $\mathcal{C}$, it is easy to show that the points of $\mathcal{C}$ verify the necessary condition

$$\Delta I_\varepsilon(x) = 0 \quad \text{on } \mathcal{C}. \tag{3.14}$$

Such conditions can be investigated for edges $\mathcal{C}$, which are the boundary Γ of a smooth domain $\Omega \subset R^N$ of class C^2, by using the *intrinsic tangential calculus* developed in [12] (p. 364) and [11] for objects in R^N, $N \geq 1$, and not just in dimension $N = 2$. Indeed we want to find points $x \in \Gamma$ such that the function $\nabla I_\varepsilon(x) \cdot n(x)$ is an extreme of the function $f(t) = \nabla I_\varepsilon(x + t\nabla b_\Omega(x)) \cdot \nabla b_\Omega(x)$ in $t = 0$. This yields the *local necessary condition* $D^2 I_\varepsilon \nabla b_\Omega \cdot \nabla b_\Omega = D^2 I_\varepsilon n \cdot n = 0$ on Γ of Assumption 3.2. As the Assumption 3.1, its analogue is Assumption 3.3.

Assumption 3.3
The restriction to the curve $\mathcal{C}$ of the gradient of the intensity I_ε is a constant vector c, that is,

$$\nabla I_\varepsilon = c, \quad c \textit{ a constant vector on } \mathcal{C}. \tag{3.15}$$

Indeed the Laplacian of I_ε on $\mathcal{C}$ can be decomposed as follows (cf. Delfour and Zolésio [12], p. 364):

$$\Delta I_\varepsilon = \operatorname{div}_\Gamma \nabla I_\varepsilon + D^2 I_\varepsilon n \cdot n,$$

where $\operatorname{div}_\Gamma \nabla I_\varepsilon$ is the *tangential divergence* of ∇I_ε, which can be defined as follows

$$\operatorname{div}_\Gamma \nabla I_\varepsilon = \operatorname{div}(\nabla I_\varepsilon \circ p_{\mathcal{C}})|_{\mathcal{C}}$$

and $p_{\mathcal{C}}$ is the *projection* onto $\mathcal{C}$. Hence, under Assumption 3.3, $\operatorname{div}_\Gamma(\nabla I_\varepsilon) = 0$ on $\mathcal{C}$ because

$$\operatorname{div}_\Gamma(\nabla I_\varepsilon) = \operatorname{div}(\nabla I_\varepsilon \circ p_\Gamma)|_\Gamma = \operatorname{div}(c)|_\Gamma = 0$$

and, under Assumption (3.13), $D^2 I_\varepsilon n \cdot n = 0$ on $\mathcal{C}$. Therefore,

$$\Delta I_\varepsilon = \operatorname{div}_\Gamma(\nabla I_\varepsilon) + D^2 I_\varepsilon n \cdot n = 0 \text{ on } \mathcal{C},$$

and we obtain the necessary condition

$$\boxed{\Delta I_\varepsilon = 0 \quad \text{on } \mathcal{C}.} \tag{3.16}$$

4 Objective functions defined on the whole edge

With the pioneering work of Kass–Witkin–Terzopoulos [17] in 1988 we go from a local necessary condition at a point of the edge to a global necessary condition by introducing objective functions defined on the entire edge of an object. Here many computations and analytical studies can be simplified by adopting the point of view that a closed curve in the plane is the boundary of a set, and using the whole machinery developed for shape and geometric analysis and the tangential and shape calculi, which readily extend to higher dimensions.

4.1 Eulerian shape semiderivative

In this section we briefly summarize some of the main notions and results from the *velocity method* [12].

Definition 4.1
Given a nonempty subset D de $\mathbf{R}^N$, consider the set $\mathcal{P}(D) = \{\Omega : \Omega \subset D\}$ of subsets of D. The set D is the hold-all *or the* universe*. A* shape function

is a well-defined map $J : \mathcal{A} \to E$ from an admissible family $\mathcal{A}$ *of* $\mathcal{P}(D)$ *with values in a topological space E.*

Given a *velocity field* $V : [0, \tau] \times \mathbf{R}^N \to \mathbf{R}^N$ (the notation $V(t)(x) = V(t, x)$ will often be used), consider the transformations

$$\boxed{T(t, X) \overset{\text{def}}{=} x(t, X), \qquad t \geq 0, X \in \mathbf{R}^N,} \tag{4.17}$$

where $x(t, X) = x(t)$ is defined as the *flow* of the differential equation

$$\frac{dx}{dt}(t) = V(t, x(t)), \quad t \geq 0; \quad x(0, X) = X \tag{4.18}$$

(here the notation $x \mapsto T_t(x) = T(t, x) : \mathbf{R}^N \to \mathbf{R}^N$will be used). The *shape semiderivative* of J in Ω in the direction V is defined as

$$dJ(\Omega; V) \overset{\text{def}}{=} \lim_{t \searrow 0} \frac{J(\Omega_t) - J(\Omega)}{t} \tag{4.19}$$

(when the limit exists in E), where $\Omega_t = T_t(\Omega) = \{T_t(x) : x \in \Omega\}$. Under appropriate assumptions on the family $\{V(t)\}$, the transformations $\{T_t\}$ are homeomorphisms that transport the boundary Γ of Ω onto the boundary Γ_t of Ω_t and the interior Ω onto the interior of Ω_t.

4.2 From local to global conditions on the edge

For simplicity, we drop the subscript ε of the convolution-smoothed image I_ε and we assume that the edge of the object is the boundary Γ of an open domain Ω of class C^2. Following Caselles, Kimmel and Sapiro [8], it is important to choose objective functions that are intrinsically defined and do not depend on an arbitrary parametrization of the boundary. For instance, given a *frame* $D =]0, a[\times]0, b[$ and a *smoothed image* $I : D \to \mathbf{R}$, to find an extremum of the objective function

$$E(\Omega) \overset{\text{def}}{=} \int_\Gamma \frac{\partial I}{\partial n} d\Gamma, \tag{4.20}$$

where the integrand is the normal derivative of I. Using the velocity method the *shape directional semiderivative* is given by the expression

$$dE(\Omega; V) = \int_\Gamma \left[H \frac{\partial I}{\partial n} + \frac{\partial}{\partial n} \left(\frac{\partial I}{\partial n} \right) \right] V \cdot n \, d\Gamma, \tag{4.21}$$

where $H = \Delta b_\Omega$ is the mean curvature and $n = \nabla b_\Omega$ is the outward unit normal. Proceeding in a formal way, a necessary condition would be

$$\begin{aligned}
&H\frac{\partial I}{\partial n} + \frac{\partial}{\partial n}\left(\frac{\partial I}{\partial n}\right) = 0 \quad \text{on } \Gamma \\
&\Rightarrow \Delta b_\Omega \nabla I \cdot \nabla b_\Omega + \nabla(\nabla I \cdot \nabla b_\Omega) \cdot \nabla b_\Omega = 0 \\
&\Rightarrow \Delta b_\Omega \nabla I \cdot \nabla b_\Omega + D^2 I \nabla b_\Omega \cdot \nabla b_\Omega + D^2 b_\Omega \nabla I \cdot \nabla b_\Omega = 0 \\
&\Rightarrow \boxed{D^2 I n \cdot n + H\frac{\partial I}{\partial n} = 0 \quad \text{on } \Gamma.} \qquad (4.22)
\end{aligned}$$

This *global condition* is to be compared with the *local condition* (3.13). It can also be expressed in terms of the Laplacian and the tangential Laplacian of I as

$$\boxed{\Delta I - \Delta_\Gamma I = 0 \quad \text{on } \Gamma,} \qquad (4.23)$$

which can be compared with the local condition (3.16), $\Delta I = 0$. This arises from the decomposition of the Laplacian of I with respect to Γ using the following identity for a smooth vector function U

$$\operatorname{div} U = \operatorname{div}_\Gamma U + DUn \cdot n, \qquad (4.24)$$

where the *tangential divergence* of U can be defined as

$$\operatorname{div}_\Gamma U = \operatorname{div}(U \circ p_\Gamma)|_\Gamma$$

and p_Γ is the projection onto Γ. Applying this to $U = \nabla I$ and recalling the definition of the tangential Laplacian

$$\Delta I = \operatorname{div}_\Gamma \nabla I + D^2 I n \cdot n = \Delta_\Gamma I + H\frac{\partial I}{\partial n} + D^2 I n \cdot n,$$

where the tangential gradient $\nabla_\Gamma I$ and the *Laplace-Beltrami* operator $\Delta_\Gamma I$ can be defined as

$$\nabla_\Gamma I = \nabla(I \circ p_\Gamma)|_\Gamma \quad \text{and} \quad \Delta_\Gamma I = \operatorname{div}_\Gamma(\nabla_\Gamma I).$$

5 Snakes, geodesic active contours, and level sets

5.1 Objective functions defined on the contours

In the literature the *objective* or *energy function* is generally made up of two terms: one (image energy) that depends on the image and one (internal energy) that specifies the smoothness of Γ. A general form of the objective function is

$$E(\Omega) \overset{\text{def}}{=} \int_\Gamma g(I)\, d\Gamma, \tag{5.25}$$

where $g(I)$ is a function of I. The directional semiderivative with respect to a velocity field V is given by

$$dE(\Omega; V) = \int_\Gamma \left[H g(I) + \frac{\partial}{\partial n} g(I) \right] n \cdot V d\Gamma. \tag{5.26}$$

This is the gradient that will make the *snakes* move and that will *activate* the contours.

5.2 Snakes and geodesic active contours

If a *gradient descent method* is used to minimize (5.25) starting from an initial curve $\mathrm{C}_0 = \mathrm{C}$, the iterative process is equivalent to following the evolution C_t (boundary of the smooth domain Ω_t) of the closed curve C given by the equation

$$\boxed{\frac{\partial \mathrm{C}_t}{\partial t} = -\left[H_t g_I + (\nabla g_I \cdot n_t) \right] n_t \quad \text{on } \mathrm{C}_t,} \tag{5.27}$$

where H_t is the mean curvature, n_t the unit exterior normal, and the right-hand side of the equation is formally the derivative of (5.25) given by (5.26). Equation (5.27) is referred to as the *geodesic flow.* For $g_I = 1$ it is the *motion by mean curvature*

$$\frac{\partial \mathrm{C}_t}{\partial t} = -H_t n_t \quad \text{on } \mathrm{C}_t. \tag{5.28}$$

5.3 Level set method

The idea is to represent the contours C_t by the zero-level set of a function $\varphi_t(x) = \varphi(t,x)$ for ae function $\varphi : [0,\tau] \times \mathbf{R}^N \to \mathbf{R}$ by setting

$$\mathrm{C}_t = \left\{ x \in \mathbf{R}^N : \varphi(t,x) = 0 \right\} = \varphi_t^{-1}\{0\}$$

and replacing (5.27) by an equation for φ, which is from Osher and Sethian [22], from 1988. This approach seems to have been simultaneously introduced in image processing by Caselles, Catté, Coll and Dibos [7] in 1993 under the name "geometric partial differential equations" with, in addition to the mean curvature term, a "transport" term, and by Malladi, Sethian and Vemuri [18] in 1995 under the name "level set approach" combined with the notion of "extension velocity."

Let $(t,x) \mapsto \varphi(t,x) : [0,\tau] \times \mathbf{R}^N \to \mathbf{R}$ be a smooth function and Ω be a subset of $\mathbf{R}^N$ of boundary $\Gamma = \overline{\Omega} \cap \overline{\complement\Omega}$ such that

$$\operatorname{int}\Omega = \{x \in \mathbf{R}^N : \varphi(0,x) < 0\} \quad \text{and} \quad \Gamma = \{x \in \mathbf{R}^N : \varphi(0,x) = 0\}. \tag{5.29}$$

Let $V : [0,\tau] \times \mathbf{R}^N \to \mathbf{R}^N$ be a sufficiently smooth velocity field so that the transformations $\{T_t\}$ are *diffeomorphisms*. Moreover, assume that the images $\Omega_t = T_t(\Omega)$ verify the following properties: for all $t \in [0,\tau]$

$$\operatorname{int}\Omega_t = \left\{x \in \mathbf{R}^N : \varphi(t,x) < 0\right\} \quad \text{and} \quad \Gamma_t = \left\{x \in \mathbf{R}^N : \varphi(t,x) = 0\right\}. \tag{5.30}$$

Assuming that the function $\varphi_t(x) = \varphi(t,x)$ is at least of class C^1, and that $\nabla\varphi_t \neq 0$ on $\varphi_t^{-1}\{0\}$, the *total derivative* with respect to t of $\varphi(t,T_t(x))$ for $x \in \Gamma$ yields

$$\frac{\partial}{\partial t}\varphi(t,T_t(x)) + \nabla\varphi(t,T_t(x)) \cdot \frac{d}{dt}T_t(x) = 0. \tag{5.31}$$

By substituting the velocity field in Equation (4.18), we get

$$\begin{aligned}&\frac{\partial}{\partial t}\varphi(t,T_t(x)) + \nabla\varphi(t,T_t(x)) \cdot V(t,T_t(x)) = 0, \quad \forall T_t(x) \in \Gamma_t\\ \Rightarrow\ &\boxed{\frac{\partial}{\partial t}\varphi_t + \nabla\varphi_t \cdot V(t) = 0 \quad \text{on } \Gamma_t, t \in [0,\tau].}\end{aligned} \tag{5.32}$$

This last equation, the *level set evolution equation*, is verified *only* on the boundaries or *fronts* Γ_t, $0 \leq t \leq \tau$. Can we find a representative φ in the equivalence class

$$[\varphi]_{\Omega,V} = \left\{\varphi : \varphi_t^{-1}\{0\} = \Gamma_t \quad \text{and} \quad \varphi_t^{-1}\{<0\} = \operatorname{int}\Omega_t, \forall t \in [0,\tau]\right\},$$

of functions φ that verify conditions (5.29) and (5.30) in order to extend equation (5.32) from Γ_t to the whole $\mathbf{R}^N$ or at least almost everywhere in $\mathbf{R}^N$? Another possibility is to consider the larger equivalence class

$$[\varphi]_{\Gamma,V} = \left\{\varphi : \varphi_t^{-1}\{0\} = \Gamma_t, \forall t \in [0,\tau]\right\}.$$

5.4 Velocity carried by the normal

In Adalsteinsson and Sethian [1], Gomez and Faugeras [15], and Malladi, Sethian and Vemuri [18], the front moves under the effect of a velocity field carried by the normal with a scalar velocity that depends on the curvatures of the level set or the front. So we are led to consider velocity fields V of the form

$$V(t)|_{\Gamma_t} = v(t)n_t, \quad x \in \Gamma_t \tag{5.33}$$

for a *scalar velocity* $(t,x) \mapsto v(t)(x) : [0,\tau] \times \mathbf{R}^N \to \mathbf{R}$. By using the expression $\nabla\varphi_t/|\nabla\varphi_t|$ of the exterior normal n_t as a function of $\nabla\varphi_t$, Equation (5.32) now becomes

$$\boxed{\frac{\partial}{\partial t}\varphi_t + v(t)|\nabla\varphi_t| = 0 \quad \text{on } \Gamma_t.} \tag{5.34}$$

For instance, consider the example of the minimization of the total length of the curve C of Equation (5.27) for the metric $g(I)d\mathrm{C}$ as introduced by Caselles et al. [8]. By using the computation (5.26) of the shape semiderivative with respect to the velocity field V of the objective function (5.25), a natural direction of descent is given by

$$\boxed{v(t)n_t, v(t) = -\left[H_t g(I) + \frac{\partial}{\partial n_t} g(I)\right] \quad \text{on } \Gamma_t.} \tag{5.35}$$

In addition, the normal n_t and the mean curvature H_t can be expressed as a function of $\nabla\varphi_t$ as follows:

$$n_t = \frac{\nabla\varphi_t}{|\nabla\varphi_t|} \quad \text{and} \quad H_t = \operatorname{div}_{\Gamma_t} n_t = \operatorname{div}_{\Gamma_t}\left(\frac{\nabla\varphi_t}{|\nabla\varphi_t|}\right) = \operatorname{div}\left(\frac{\nabla\varphi_t}{|\nabla\varphi_t|}\right)\bigg|_{\Gamma_t}. \tag{5.36}$$

By substituting in (5.34), we get the following evolution equation

$$\frac{\partial\varphi_t}{\partial t} - \left(H_t g(I) + \frac{\partial}{\partial n_t} g(I)\right) \nabla\varphi_t \cdot \frac{\nabla\varphi_t}{|\nabla\varphi_t|} = 0 \quad \text{on } \Gamma_t \tag{5.37}$$

and

$$\begin{cases} \dfrac{\partial\varphi_t}{\partial t} - \left(H_t g(I) + \dfrac{\partial}{\partial n_t} g(I)\right) |\nabla\varphi_t| = 0 & \text{on } \Gamma_t \\ \varphi_0 = \varphi^0 & \text{on } \Gamma_0. \end{cases} \tag{5.38}$$

By substituting expressions (5.36) for n_t and H_t in terms of $\nabla\varphi_t$, we finally get

$$\boxed{\begin{cases} \dfrac{\partial\varphi_t}{\partial t} - \left[\operatorname{div}\left(\dfrac{\nabla\varphi_t}{|\nabla\varphi_t|}\right) g(I) + \nabla g(I) \cdot \dfrac{\nabla\varphi_t}{|\nabla\varphi_t|}\right] |\nabla\varphi_t| = 0 & \text{on } \Gamma_t \\ \varphi_0 = \varphi^0 & \text{on } \Gamma_0. \end{cases}} \tag{5.39}$$

The negative sign arises from the fact that we have chosen the outward rather than the inward normal. The main references to the existence and uniqueness theorems related to Equation (5.39) can be found in Chen, Giga and Goto [9].

5.5 Extension of the level set equations

Equation (5.39) on the fronts Γ_t is not convenient from either the theoretical or the numerical viewpoint. So it would be desirable to be able to extend Equation (5.39) in a small *tubular neighborhood* of thickness h

$$U_h(\Gamma_t) \overset{\text{def}}{=} \left\{x \in \mathbf{R}^N : d_{\Gamma_t}(x) < h\right\} \tag{5.40}$$

of the front Γ_t for a small $h > 0$. In theory, the velocity field associated with Γ_t is given by

$$V(t) = -\left[\operatorname{div}\left(\frac{\nabla\varphi_t}{|\nabla\varphi_t|}\right) g(I) + \nabla g(I) \cdot \frac{\nabla\varphi_t}{|\nabla\varphi_t|}\right] \frac{\nabla\varphi_t}{|\nabla\varphi_t|} \quad \text{on } \Gamma_t. \tag{5.41}$$

Given the identities (5.36), the expressions of n_t and H_t extend to a neighborhood of Γ_t and possibly to $\mathbf{R}^N$ if $\nabla\varphi_t$ is sufficiently smooth and $\nabla\varphi_t \neq 0$ on Γ_t.

Equation (5.39) can be extended to $\mathbf{R}^N$ at the price of violating Assumptions (5.29) through (5.30), either by the loss of smoothness of φ_t or by allowing its gradient to be zero on Γ_t

$$\boxed{\begin{cases} \dfrac{\partial\varphi_t}{\partial t} - \left[\operatorname{div}\left(\dfrac{\nabla\varphi_t}{|\nabla\varphi_t|}\right) g(I) + \nabla g(I) \cdot \dfrac{\nabla\varphi_t}{|\nabla\varphi_t|}\right] |\nabla\varphi_t| = 0 & \text{in } \mathbf{R}^N \\ \varphi_0 = \varphi^0 & \text{in } \mathbf{R}^N. \end{cases}} \tag{5.42}$$

In fact, the starting point of Caselles, Catté, Coll and Dibos [7] was the following *geometric partial differential equation*:

$$\begin{cases} \dfrac{\partial\varphi_t}{\partial t} - \left[\operatorname{div}\left(\dfrac{\nabla\varphi_t}{|\nabla\varphi_t|}\right) g(I) + \nu g(I)\right] |\nabla\varphi_t| = 0 & \text{in } \mathbf{R}^N \\ \varphi_0 = \varphi^0 & \text{in } \mathbf{R}^N, \end{cases} \tag{5.43}$$

for a constant $\nu > 0$ and the function $g(I) = 1/(1 + |\nabla I|^2)$. They prove the existence, in dimension two, of a *viscosity solution* unique for initial data $\varphi^0 \in C([0,1] \times [0,1]) \cap W^{1,\infty}([0,1] \times [0,1])$ and $g \in W^{1,\infty}(\mathbf{R}^2)$.

6 Evolution equation for the oriented distance function

Given a velocity field V and a subset Ω of $\mathbf{R}^N$ with nonempty boundary Γ, consider the *oriented distance function*

$$b_\Omega(x) \overset{\text{def}}{=} d_\Omega(x) - d_{\complement\Omega}(x), \tag{6.44}$$

where $d_A(x)$ is the usual minimum distance from a point x to a nonempty subset A of $\mathbf{R}^N$ and $\complement A$ is the complement of the set A with respect to $\mathbf{R}^N$. It is a special example of the level set function by setting $\varphi(t,x) = b_{\Omega_t}(x)$. One interesting feature of that function is that for sets Ω whose nonempty boundary Γ is *thin*, that is, Γ has zero Lebesgue measure, the gradient of the function $t \mapsto b_{\Omega_t} : [0,\tau] \to C(\overline{D})$ (for some $\tau > 0$ and all bounded open subsets D of $\mathbf{R}^N$) has its norm equal to one almost everywhere in $\mathbf{R}^N$. This applies to the evolution of sets with arbitrary components of different dimensions, such as a cloud of points, curves, surfaces, or an object of higher codimension as long as the Lebesgue measure of their boundary is zero. So we are not limited to smooth sets or even open sets. This evolution equation of b_{Ω_t} was established by Delfour and Zolésio in [13], Theorem 5.1.

Theorem 6.1

Given $\tau > 0$, let $V : [0,\tau] \times \mathbf{R}^N \to \mathbf{R}^N$ be a transformation satisfying the conditions of Theorem 5.1 in [13] and such that $V \in C([0,\tau]; C^0(\overline{\mathbf{R}^N}, \mathbf{R}^N))$. Assume that Ω is a subset of $\mathbf{R}^N$ with thin boundary $\Gamma \neq \varnothing$. Then for all bounded open subset D of $\mathbf{R}^N$ and for all p, $1 \le p < \infty$, the function $t \mapsto b_{\Omega_t}$ belongs to the space $C^1([0,\tau]; L^p(D)) \cap C^0([0,\tau]; W^{1,p}(D))$ and satisfies the evolution equation

$$\boxed{\frac{\partial}{\partial t} b_{\Omega_t} + \nabla b_{\Omega_t} \cdot (V(t) \circ p_{\Gamma_t}) = 0 \quad a.e.\ in \mathbf{R}^N, \quad b_{\Omega_0} = b_\Omega,} \tag{6.45}$$

for all t, $0 \le t \le \tau$, where p_{Γ_t} is the projection *onto Γ_t*

$$p_{\Gamma_t}(x) = x - \frac{1}{2}\nabla b^2_{\Omega_t}(x) \quad a.e.\ in\ \mathbf{R}^N. \tag{6.46}$$

Remark

Notice that the evolution equation (5.42) for level sets functions and (6.45) for the oriented distance function have similar structures. Yet only Equation (6.45) fully preserves the integrity of the oriented distance function over time. Equation (5.42) is an *incomplete* equation.

7 Acknowledgments

Mathieu Dehaes' research was supported by the National Sciences and Engineering Research Council of Canada discovery grant A–8730, by a FQRNT grant from the Ministère de l'Éducation du Québec, and by a fellowship from the John Simon Memorial Foundation.

References

[1] D. Adalsteinsson and J.A. Sethian, The fast construction of extension velocities in level set methods, *J. Computational Physics* 148 (1999), 2–22.

[2] L. Alvarez, P.-L. Lions, and J.-M. Morel, Image selective smoothing and edge detection by nonlinear diffusion. II, *SIAM J. Numer. Anal.* 29, no. 3 (1992), 845–866.

[3] H. Blum, A transformation for extracting view description of shapes, in *Models for Perception of Speech and Visual Form*, W. Wathen-Dunn, ed., 362–380, MIT Press, Cambridge, MA, 1967.

[4] R. Bracewell, *The Fourier Transform and Its Applications*, McGraw-Hill, New York, 1965.

[5] A. Brakke, *The Motion of a Surface by Its Mean Curvature*, Princeton University Press, Princeton, NJ, 1978.

[6] F.W. Campbell and J.G. Robson, Applications of Fourier analysis to the visibility of granting, *J. Physiol.* (Lond.) 197 (1968), 551–556.

[7] V. Caselles, F. Catté, T. Coll, and F. Dibos, A geometric model for active contours, *Numerische Mathematik* 66 (1993), 1–31.

[8] V. Caselles, R. Kimmel, and G. Sapiro, Geodesic active contours, *International J. Computer Vision* 22 (1997), 61–79.

[9] Y.G. Chen, Y. Giga and S. Goto, Uniqueness and existence of viscosity solutions of generalized mean curvature flow equations, *J. Differential Geometry* 33 (1991), 749–786.

[10] M. Dehaes, *Représentations analytiques des objets géométriques et contours actifs en imagerie*, Mémoire de maîtrise, Département de mathématiques et de statistique, Université de Montréal, Canada, 2004.

[11] M.C. Delfour, Tangential differential calculus and functional analysis on a $C^{1,1}$ submanifold, in *Differential-Geometric Methods in the Control of Partial Differential Equations*, R. Gulliver, W. Littman, and R. Triggiani, eds., 83–115, Contemp. Math., Vol. 268, AMS Publications, Providence, RI, 2000.

[12] M.C. Delfour and J.-P. Zolésio, *Shapes and Geometries: Analysis, Differential Calculus and Optimization*, SIAM Series on Advances in Design and Control, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2001.

[13] M.C. Delfour and J.-P. Zolésio, Oriented distance function in shape identification via metrics and its evolution equation, *SIAM J. on Control and Optim.* 42, no. 6 (2004), 2286–2304.

[14] M.C. Delfour and J.-P. Zolésio, *Shape identification via metrics constructed from the oriented distance function*, *Control and Cybernetics* 34, no. 1 (2005), 137–164.

[15] J. Gomes and O. Faugeras, Reconciling distance functions and level sets, *J. Visual Com. and Image Representation* 11 (2000), 209–223.

[16] D.H. Hubel and T.N. Wiesel, Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, *J. Physiol.* (Lond.) 160 (1962), 106–154.

[17] M. Kass, M. Witkin, and D. Terzopoulos, Snakes: Active contour models, *International J. Computer Vision* 1, no. 4 (1988), 321–331.

[18] R. Malladi, J. A. Sethian, and B.C. Vemuri, Shape modeling with front propagation: A level set approach, *IEEE Trans. on Pattern Analysis and Machine Intelligence* 17, no. 2 (1995), 158–175.

[19] D. Marr, *Vision*, W.H. Freeman and Company, New York, 1982.

[20] D. Marr and E. Hildreth, Theory of edge detection, *Proc. R. Soc. Lond. B* 207 (1980), 187–217.

[21] A. Martelli, Edge detection using heuristic search methods, *Comp. Graphics Image Processing* 1 (1972), 169–182.

[22] S. Osher and J.A. Sethian, Front propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulation, *J. Computational Physics* 79 (1988), 12–49.

[23] A. Rosenfeld and M. Thurston, Edge and curve detection using anisotropic diffusion, *IEEE Trans. on Comput.* C-20 (1971), 562–569.

[24] A.P. Witkin, Scale space filtering, *Proc. IJCAI*, Karlsruhe 1983, 1019–1021.

[25] A. Yuille and T. Poggio, Scaling theorems for zero crossings, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 8 (1986), 15–25.

CHAPTER 16

Topological derivatives for contact problems

Jan Sokołowski
Institut Élie Cartan, Laboratoire de Mathématiques, Université Henri Poincaré
Nancy I, B.P. 239, Vandoeuvre lés Nancy Cedex, France and Systems Research
Institute of the Polish Academy of Sciences, Warsaw, Poland

Antoni Zochowski
Systems Research Institute of the Polish Academy of Sciences, Warsaw, Poland

1 Introduction

The main idea we use to derive the topological derivatives for contact problems is the modification of the energy functional by an appropriate correction term and subsequent minimization of the resulting energy functional over the cone of admissible displacements. Such an approach leads to the *outer* approximations of solutions to variational inequalities.

In this paper we derive useful formulae for the correction terms of the energy functionals. We restrict ourselves to two-dimensional problems and to singular perturbations of geometrical domains in the form of small discs.

The correction terms are derived in such a form that the numerical verification of its precision is straightforward. On the other hand, the terms are directly used to establish the topological differentiability of solutions to variational inequalities. As a result, the one-term *outer* expansion of solutions is derived for a class of nonlinear problems. Outer expansion means that the expansion is precise far from the hole. The expansion precise near the hole is called *inner* expansion and usually the matching procedure is applied (see [5]) to construct the global asymptotic approximation of solutions to boundary problems in singularly perturbed geometrical domains.

1.1 Contact problem in elasticity

We establish the conical differentiability of solutions for the two-dimensional contact problem in the elasticity. We consider the bounded domain Ω with the boundary $\partial\Omega = \Gamma_0 \cup \Gamma_c$. On Γ_0 the displacement vector of the elastic body is given; on Γ_c the frictionless contact conditions are prescribed. To specify the weak formulation we need an expression for the symmetric bilinear form and for the convex set $K \subset H^1(\Omega)^2$.

The method of analysis is the same as in the case of the Signorini problem. We start with the formulation of the free boundary problem in an unperturbed domain Ω. The form of variational inequality is straightforward.

1.1.1 Contact problem in Ω

Find $\mathbf{u} = \mathbf{u}(\Omega) = (u_1, u_2)$ and $\sigma = (\sigma)_{ij}$, $i, j = 1, 2$, such that

$$-\mathbf{div}\,\sigma = \mathbf{f} \qquad \text{in } \Omega, \tag{1.1}$$

$$C\sigma - \epsilon(\mathbf{u}) = 0 \qquad \text{in } \Omega, \tag{1.2}$$

$$\mathbf{u} = 0 \qquad \text{on } \Gamma_0, \tag{1.3}$$

$$\mathbf{u}\nu \geq 0, \quad \sigma_\nu \leq 0, \quad \sigma_\nu \mathbf{u}\nu = 0 \qquad \sigma_\tau = 0 \quad \text{on } \Gamma_c. \tag{1.4}$$

Here

$$\sigma_\nu = \sigma_{ij}\nu_j\nu_i,\ \sigma_\tau = \sigma\nu - \sigma_\nu = \left\{\sigma_\tau^i\right\}_{i=1}^2,\ \sigma\nu = \left\{\sigma_{ij}\nu_j\right\}_{i=1}^2,$$
$$\epsilon_{ij}(\mathbf{u}) = \frac{1}{2}(u_{i,j} + u_{j,i}),\ i, j = 1, 2,\ \epsilon(\mathbf{u}) = (\epsilon_{ij})_{i,j=1}^2,$$
$$\{C\sigma\}_{ij} = c_{ijk\ell}\sigma_{k\ell},\ c_{ijk\ell} = c_{jik\ell} = c_{k\ell ij},\ c_{ijk\ell} \in L^\infty(\Omega).$$

The Hooke's tensor C satisfies the ellipticity condition

$$c_{ijk\ell}\xi_{ji}\xi_{k\ell} \geq c_0|\xi|^2, \quad \forall \xi_{ji} = \xi_{ij},\ c_0 > 0, \tag{1.5}$$

and we have used the summation convention over repeated indices.

When the topology of Ω is changed, we have the following contact problem in the domain Ω_ρ with the small hole $B(\rho)$.

1.1.2 Contact problem in Ω_ρ

Find $\mathbf{u} = \mathbf{u}(\Omega_\rho) = (u_1, u_2)$ and $\sigma = (\sigma)_{ij}$, $i, j = 1, 2$, such that

$$-\mathbf{div}\,\sigma = \mathbf{f} \qquad \text{in } \Omega_\rho, \tag{1.6}$$

$$C\sigma - \epsilon(\mathbf{u}) = 0 \qquad \text{in } \Omega_\rho, \tag{1.7}$$

$$\mathbf{u} = 0 \qquad \text{on } \Gamma_0, \tag{1.8}$$

$$\sigma\nu = 0 \quad \text{on } \Gamma_\rho, \tag{1.9}$$

$$\mathbf{u}\nu \geq 0, \quad \sigma_\nu \leq 0, \quad \sigma_\nu \mathbf{u}\nu = 0 \quad \sigma_\tau = 0 \quad \text{on } \Gamma_c. \tag{1.10}$$

We assume for simplicity that the case of isotropic elasticity is considered; thus the symmetric bilinear form associated with the boundary value problems (1.1) through (1.4) is given by

$$a(\mathbf{u},\mathbf{u}) = \int_\Omega [(\lambda+\mu)(\epsilon_{11}+\epsilon_{22})^2 + \mu(\epsilon_{11}-\epsilon_{22})^2 + \mu\gamma_{12}^2], \tag{1.11}$$

where the notation for isotropic elasticity is fixed in Section 2.

Problems (1.6) through (1.10) are approximated by the problem with modified bilinear form in the following way.

1.1.3 Approximation of contact problem in Ω_ρ

We determine the modified bilinear form as a sum of two terms, as it is for the energy functional; the first term defines the elastic energy in the domain Ω, and the second term is a correction term, determined in Section 2.3 by formula (2.42). The correction term is quite complicated to evaluate, and we do not provide its explicit form; such a form is actually defined by the formulae in Section 2. The values of the symmetric bilinear form $a(\rho;\cdot,\cdot)$ are given by the expression

$$a(\rho;\mathbf{v},\mathbf{v}) = a(\mathbf{u},\mathbf{u}) + \rho^2 b(\mathbf{v},\mathbf{v}). \tag{1.12}$$

The derivative $b(\mathbf{v},\mathbf{v})$ of the bilinear form $a(\rho;\mathbf{v},\mathbf{v})$ with respect to ρ^2 at $\rho = 0+$ is given by the expression

$$b(\mathbf{v},\mathbf{v}) = -2\pi e_{\mathbf{v}}(0) - \frac{2\pi\mu}{\lambda+3\mu}(\sigma_{II}\delta_1 - \sigma_{12}\delta_2), \tag{1.13}$$

where all the quantities are evaluated for the displacement field $\mathbf{v}$ according to formulae (2.26), (2.27), (2.29), (2.42), and (2.33); we provide the line integrals that define all terms in (1.13) below.

Hence, we can determine the bilinear form $a(\rho;\mathbf{v},\mathbf{w})$ for all $\mathbf{v},\mathbf{w}$, from the equality

$$2a(\rho;\mathbf{v},\mathbf{w}) = a(\rho;\mathbf{v}+\mathbf{w},\mathbf{v}+\mathbf{w}) - a(\rho;\mathbf{w},\mathbf{w}).$$

In the same way the bilinear form $b(\mathbf{v},\mathbf{w})$ is determined from the formula for $b(\mathbf{v},\mathbf{v})$.

The convex set is defined in this case by

$$\mathbf{K} = \{\mathbf{v} \in H^1(\Omega)^2 | \mathbf{v}\nu \geq 0 \quad \text{on } \Gamma_c, \qquad \mathbf{v} = \mathbf{g} \quad \text{on } \Gamma_0\}. \tag{1.14}$$

Let us consider the following variational inequality, which provides a (sufficiently precise for our purposes) approximation $\mathbf{u}_\rho$ of the solution $\mathbf{u}(\Omega_\rho)$ to the contact problems (1.6) through (1.10),

$$\mathbf{u}_\rho \in \mathbf{K} \ : \quad a(\rho;\mathbf{u},\mathbf{v}-\mathbf{u}) \geq L(\rho;\mathbf{v}-\mathbf{u}) \quad \forall v \in \mathbf{K}. \tag{1.15}$$

The result obtained is the following. For simplicity we assume that the linear form $L(\rho;\cdot)$ is independent of ρ.

Theorem 1.1

For ρ sufficiently small we have the following expansion of the solution u_ρ with respect to the parameter ρ at $0+$,

$$\mathbf{u}_\rho = \mathbf{u}(\Omega) + \rho^2 \mathbf{q} + o(\rho^2) \quad \text{in } H^1(\Omega)^2, \tag{1.16}$$

where the topological derivative $\mathbf{q}$ of the solution $\mathbf{u}(\Omega)$ to the contact problem is given by the unique solution of the following variational inequality:

$$\mathbf{q} \in \mathcal{S}_{\mathbf{K}}(\mathbf{u}) = \{\mathbf{v} \in (H^1_{\Gamma_0}(\Omega))^2 | \mathbf{v}\nu \leq 0 \quad \text{on } \Xi(\mathbf{u}), \quad a(0;\mathbf{u},\mathbf{v}) = 0\} \tag{1.17}$$

$$a(0;\mathbf{q},\mathbf{v}-\mathbf{q}) + b(\mathbf{u},\mathbf{v}-\mathbf{q}) \geq 0 \quad \forall \mathbf{v} \in \mathcal{S}_{\mathbf{K}}(\mathbf{u}). \tag{1.18}$$

The coincidence set $\Xi(\mathbf{u}) = \{x \in \Gamma_s | \mathbf{u}(x).\nu(x) = 0\}$ is well defined (see [15]) for any function $\mathbf{u} \in H^1(\Omega)^2$, and $\mathbf{u} \in \mathbf{K}$ is the solution of variational inequality (1.14) for $\rho = 0$.

Remark
In the linear case, it can be shown that $\|\mathbf{u}(\Omega_\rho) - \mathbf{u}_\rho\| = o(\rho^2)$ in the norm of appropriate weighted space. We refer the reader to [7] for the related error estimates in the Hölder weighted spaces. In general, we cannot expect that $\mathbf{u}_\rho$ is close to $\mathbf{u}(\Omega_\rho)$ in the vicinity of B_ρ; therefore the weighted spaces should be used for error estimates.

For the convenience of the reader we provide the explicit formulae for the terms in $b(\mathbf{v},\mathbf{v})$ defined by (1.13); refer to Section 2.2 for details. We have

$$\begin{aligned} 2\pi e_{\mathbf{v}}(0) = {} & \frac{\pi(\lambda+\mu)}{\pi^2 R^6} \left(\int_{\Gamma_R} (v_1 x_1 + v_2 x_2)\, ds \right)^2 + \\ & + \frac{\mu}{\pi^2 R^6} \left(\int_{\Gamma_R} \left((1-9k)(v_1 x_1 - v_2 x_2) + \frac{12k}{R^2}(v_1 x_1^3 - v_2 x_2^3) \right] ds \right)^2 + \\ & + \frac{\mu}{\pi^2 R^6} \left(\int_{\Gamma_R} \left[(1+9k)(v_1 x_2 + v_2 x_1) - \frac{12k}{R^2}(v_1 x_2^3 + v_2 x_1^3) \right] ds \right)^2, \end{aligned} \tag{1.19}$$

with

$$\sigma_{II} = \frac{\mu}{\pi R^3} \int_{\Gamma_R} \left[(1-9k)(v_1 x_1 - v_2 x_2) + \frac{12k}{R^2}(v_1 x_1^3 - v_2 x_2^3) \right] ds,$$

$$\sigma_{12} = \frac{\mu}{\pi R^3} \int_{\Gamma_R} \left[(1+9k)(v_1 x_2 + v_2 x_1) - \frac{12k}{R^2}(v_1 x_2^3 + v_2 x_1^3) \right] ds,$$

and

$$\delta_1 = \frac{9k}{\pi R^3} \int_{\Gamma_R} \left[(v_1 x_1 - v_2 x_2) - \frac{4}{3R^2}(v_1 x_1^3 - v_2 x_2^3) \right] ds,$$
$$\delta_2 = \frac{9k}{\pi R^3} \int_{\Gamma_R} \left[(v_1 x_2 + v_2 x_1) - \frac{4}{3R^2}(v_1 x_2^3 + v_2 x_1^3) \right] ds.$$

2 Transformations of the energy functional for the 2D elasticity system

2.1 Using the Poisson kernel for computing strain

As it turns out, similar reasoning may be carried out in the case of the 2D elasticity system, even if it is much more complicated. In the absence of volume forces, such a system has a form

$$\begin{aligned} &\mu \Delta u_1 + (\lambda + \mu)(u_{1/1,1} + u_{2/1,2}) = 0, \\ &\mu \Delta u_2 + (\lambda + \mu)(u_{1/1,2} + u_{2/2,2}) = 0, \end{aligned} \tag{2.20}$$

where $u = (u_1, u_2)^T$ denotes the displacement and λ, μ are Lamé constants. We shall also use the usual notation for the symmetric strain tensor $\epsilon = [\epsilon_{ij}]$, $\epsilon_{11} = u_{1/1}$, $\epsilon_{22} = u_{2/2}$, $\gamma_{12} = 2\epsilon_{12} = u_{1/2} + u_{2/1}$, as well as stress tensor $\boldsymbol{\sigma} = [\sigma_{ij}]$. Hooke's law

$$\sigma_{11} = (\lambda + 2\mu)\epsilon_{11} + \lambda\epsilon_{22}, \quad \sigma_{22} = \lambda\epsilon_{11} + (\lambda + 2\mu)\epsilon_{22}, \quad \sigma_{12} = \mu\gamma_{12} = 2\mu\epsilon_{12}$$

links both objects. In these terms, (2.20) reduces to

$$\nabla \cdot \boldsymbol{\sigma}(\boldsymbol{u}) = 0. \tag{2.21}$$

For such a system there exists an analogue to the Poisson kernel; see [1]. It is a matrix $\boldsymbol{G}(\boldsymbol{x}, \boldsymbol{y})$ allowing us to express the values of the solution inside the circle $\Gamma_R(\boldsymbol{x}_0)$ by means of its values on the circumference:

$$\boldsymbol{u}(\boldsymbol{x}) = -\frac{1}{\pi} \int_{\Gamma_R(\boldsymbol{x}_0)} \boldsymbol{G}(\boldsymbol{x} - \boldsymbol{x}_0, \boldsymbol{y} - \boldsymbol{x}_0) \cdot \boldsymbol{u}(\boldsymbol{y}) \, ds_y. \tag{2.22}$$

Let us denote $\boldsymbol{I}$ as the identity matrix and

$$k = \frac{\lambda + \mu}{\lambda + 3\mu}.$$

Then $\boldsymbol{G}(\boldsymbol{x}, \boldsymbol{y})$ has a form

$$\boldsymbol{G}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\Gamma}(\boldsymbol{x}, \boldsymbol{y}) + \boldsymbol{A}(\boldsymbol{x}, \boldsymbol{y}), \tag{2.23}$$

where

$$\boldsymbol{\Gamma}(\boldsymbol{x},\boldsymbol{y}) = \left((1-k)\boldsymbol{I} + 2k\begin{bmatrix} \left(\frac{\partial d}{\partial x_1}\right)^2, & \frac{\partial d}{\partial x_1}\frac{\partial d}{\partial x_2} \\ \frac{\partial d}{\partial x_1}\frac{\partial d}{\partial x_2}, & \left(\frac{\partial d}{\partial x_2}\right)^2 \end{bmatrix}\right)\frac{\partial}{\partial \boldsymbol{n}_y}\log\frac{1}{d}, \tag{2.24}$$

$$\boldsymbol{A}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{2R}\left((1-k)\boldsymbol{I} - k\begin{bmatrix} \frac{x_1y_1-x_2y_2}{R^2}-1, & \frac{x_1y_2+x_2y_1}{R^2} \\ \frac{x_1y_2+x_2y_1}{R^2}, & -1-\frac{x_1y_1-x_2y_2}{R^2} \end{bmatrix}\right), \tag{2.25}$$

and $d = d(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\|$.

From now on we shall assume that $\boldsymbol{x}_0 = 0$. This greatly simplifies the formulae without loss of generality.

Using the representation of displacement as given by (2.22) we may compute the values of its derivatives at 0. Before writing down the result, we must introduce some notation. Let us define $I_1(k,l)$ and $I_2(k,l)$ as

$$I_1(k,l) = \frac{1}{\alpha(k,l)}\int_{\Gamma_R} u_1x_1^kx_2^l\,ds, \qquad I_2(k,l) = \frac{1}{\beta(k,l)}\int_{\Gamma_R} u_2x_1^kx_2^l\,ds, \tag{2.26}$$

where

$$\alpha(k,l) = R^{k+l+2}\int_0^{2\pi}\cos^{k+1}\phi\sin^l\phi\,d\phi,$$
$$\beta(k,l) = R^{k+l+2}\int_0^{2\pi}\cos^{k}\phi\sin^{l+1}\phi\,d\phi,$$

whenever these expressions make sense, that is, if k is odd and l even or vice versa. Observe that $\alpha(k,0) = \beta(0,k)$ and

$$\alpha(1,0) = \pi R^3, \quad \alpha(3,0) = \frac{3}{4}\pi R^5, \quad \alpha(1,2) = \frac{1}{4}\pi R^5,$$
$$\alpha(5,0) = \frac{5}{8}\pi R^7, \quad \alpha(3,2) = \frac{1}{8}\pi R^7$$

and so on. Furthermore, let

$$\begin{aligned}\delta_1 &= 9k\left([I_1(1,0)-I_2(0,1)] - [I_1(3,0)-I_2(0,3)]\right),\\ \delta_2 &= 9k\left([I_1(0,1)+I_2(1,0)] - [I_1(0,3)+I_2(3,0)]\right).\end{aligned} \tag{2.27}$$

In terms of these symbols one may obtain, after very lengthy calculations, the formulae for the values of strain components at the point $\boldsymbol{x}_0 = 0$, which will constitute the basis of our energy transformations:

$$\begin{aligned} \epsilon_{11} \;+\; \epsilon_{22} &= I_1(1,0) \;+\; I_2(0,1), \\ \epsilon_{11} \;-\; \epsilon_{22} &= I_1(1,0) \;-\; I_2(0,1) \;-\; \delta_1, \\ \gamma_{12} &= I_1(0,1) \;+\; I_2(1,0) \;+\; \delta_2. \end{aligned} \tag{2.28}$$

Let us recall also the expression for the elastic energy density at the same point,

$$e_{\boldsymbol{u}}(0) = \frac{1}{2}\boldsymbol{\sigma} : \boldsymbol{\epsilon} = \frac{1}{2}[(\lambda+\mu)(\epsilon_{11}+\epsilon_{22})^2 + \mu(\epsilon_{11}-\epsilon_{22})^2 + \mu\gamma_{12}^2]. \tag{2.29}$$

2.2 Distortion of the stress field caused by a small circular hole

We shall recall here some formulae describing the stress field around a circular hole in the infinite 2D elastic medium. If we assume that at infinity only σ_{11} is not zero, and the hole $B(\rho)$ is centered around the origin, then the stresses for $r \geq \rho$ have the form

$$\begin{aligned} \sigma_{rr} &= \frac{1}{2}\sigma_{11}\left[\left(1-\frac{\rho^2}{r^2}\right) + \left(1 - 4\frac{\rho^2}{r^2} + 3\frac{\rho^4}{r^4}\right)\cos 2\phi\right], \\ \sigma_{\phi\phi} &= \frac{1}{2}\sigma_{11}\left[\left(1+\frac{\rho^2}{r^2}\right) - \left(1 + 3\frac{\rho^4}{r^4}\right)\cos 2\phi\right], \\ \sigma_{r\phi} &= -\frac{1}{2}\sigma_{11}\left(1 + 2\frac{\rho^2}{r^2} - 3\frac{\rho^4}{r^4}\right)\sin 2\phi. \end{aligned} \tag{2.30}$$

Here (r,ϕ) constitute the polar coordinate system around the origin and the σ–components are given in the orthogonal coordinates defined by $\{\boldsymbol{e}_r, \boldsymbol{e}_\phi\}$, with base versors at any given point directed along the radius and perpendicular counter to it, counterclockwise.

Using these expressions we may immediately construct the solution corresponding to nonzero σ_{22} at infinity by substituting $\phi := \phi + \frac{\pi}{2}$, $\sigma_{11} := \sigma_{22}$ (exchange of axis):

$$\begin{aligned} \sigma_{rr} &= \frac{1}{2}\sigma_{22}\left[\left(1-\frac{\rho^2}{r^2}\right) - \left(1 - 4\frac{\rho^2}{r^2} + 3\frac{\rho^4}{r^4}\right)\cos 2\phi\right], \\ \sigma_{\phi\phi} &= \frac{1}{2}\sigma_{22}\left[\left(1+\frac{\rho^2}{r^2}\right) + \left(1 + 3\frac{\rho^4}{r^4}\right)\cos 2\phi\right], \\ \sigma_{r\phi} &= \frac{1}{2}\sigma_{22}\left(1 + 2\frac{\rho^2}{r^2} - 3\frac{\rho^4}{r^4}\right)\sin 2\phi. \end{aligned} \tag{2.31}$$

Furthermore, we may exploit the fact that the pure shear stress σ_{12} is equivalent to simultaneous stretching and compression with the same intensity σ_{12} and $-\sigma_{12}$, but along the axis rotated by the angle $\pi/4$. Thus we make substitutions $\phi := \phi + \frac{\pi}{4}$, $\sigma_{11} := \sigma_{12}$; then $\phi := \phi - \frac{\pi}{4}$, $\sigma_{11} := -\sigma_{12}$ in (2.30) and add both solutions together, obtaining:

$$\begin{aligned}
\sigma_{rr} &= \sigma_{12}\left(1 - 4\frac{\rho^2}{r^2} + 3\frac{\rho^4}{r^4}\right)\sin 2\phi,\\
\sigma_{\phi\phi} &= \sigma_{12}\left(1 + 3\frac{\rho^4}{r^4}\right)\sin 2\phi, \qquad (2.32)\\
\sigma_{r\phi} &= \sigma_{12}\left(1 + 2\frac{\rho^2}{r^2} - 3\frac{\rho^4}{r^4}\right)\cos 2\phi.
\end{aligned}$$

Let us now denote

$$\sigma_I = \frac{1}{2}(\sigma_{11} + \sigma_{22}), \qquad \sigma_{II} = \frac{1}{2}(\sigma_{11} - \sigma_{22}). \qquad (2.33)$$

Then adding (2.30), (2.31), and (2.32) gives the solution corresponding to the general stress field at infinity:

$$\begin{aligned}
\sigma_{rr} &= \sigma_I + \sigma_{II}\cos 2\phi + \sigma_{12}\sin 2\phi\\
&\quad - \sigma_I\frac{\rho^2}{r^2} - \sigma_{II}\left(4\frac{\rho^2}{r^2} - 3\frac{\rho^4}{r^4}\right)\cos 2\phi - \sigma_{12}\left(4\frac{\rho^2}{r^2} - 3\frac{\rho^4}{r^4}\right)\sin 2\phi,\\
\sigma_{\phi\phi} &= \sigma_I - \sigma_{II}\cos 2\phi - \sigma_{12}\sin 2\phi \qquad (2.34)\\
&\quad + \sigma_I\frac{\rho^2}{r^2} - 3\sigma_{II}\frac{\rho^4}{r^4}\cos 2\phi - 3\sigma_{12}\frac{\rho^4}{r^4}\sin 2\phi,\\
\sigma_{r\phi} &= -\sigma_{II}\sin 2\phi + \sigma_{12}\cos 2\phi\\
&\quad - \sigma_{II}\left(2\frac{\rho^2}{r^2} - 3\frac{\rho^4}{r^4}\right)\sin 2\phi + \sigma_{12}\left(2\frac{\rho^2}{r^2} - 3\frac{\rho^4}{r^4}\right)\cos 2\phi.
\end{aligned}$$

Recalling the rules for the transformation of stresses under rotation of the coordinate system, we get the distortion of the stress due to the circular hole:

$$\begin{aligned}
\hat{\sigma}_{rr} &= -\sigma_I\frac{\rho^2}{r^2} - \sigma_{II}\left(4\frac{\rho^2}{r^2} - 3\frac{\rho^4}{r^4}\right)\cos 2\phi - \sigma_{12}\left(4\frac{\rho^2}{r^2} - 3\frac{\rho^4}{r^4}\right)\sin 2\phi,\\
\hat{\sigma}_{\phi\phi} &= \sigma_I\frac{\rho^2}{r^2} - 3\sigma_{II}\frac{\rho^4}{r^4}\cos 2\phi - 3\sigma_{12}\frac{\rho^4}{r^4}\sin 2\phi, \qquad (2.35)\\
\hat{\sigma}_{r\phi} &= -\sigma_{II}\left(2\frac{\rho^2}{r^2} - 3\frac{\rho^4}{r^4}\right)\sin 2\phi + \sigma_{12}\left(2\frac{\rho^2}{r^2} - 3\frac{\rho^4}{r^4}\right)\cos 2\phi.
\end{aligned}$$

2.3 Transformation of the energy functional

Now we shall consider the contribution, in the absence of volume forces, of the energy integral over the circle surrounding the origin (i.e., the potential location of the small hole)

$$e_R(\boldsymbol{u}) = \frac{1}{2}\int_{B(R)} (\boldsymbol{\sigma} : \boldsymbol{\epsilon})\, dx = \frac{1}{2}\int_{\Gamma_R} \boldsymbol{u}^T(\boldsymbol{\sigma}.\boldsymbol{n})\, ds \tag{2.36}$$

to the global elastic energy. As in the case of the Laplace equation, we shall leave the displacement as is and consider the distortion to the stress field caused by introducing the small hole. Due to (2.35) it may be expressed as

$$\delta e_R = \frac{1}{2}\int_{\Gamma_R} \boldsymbol{u}^T(\hat{\boldsymbol{\sigma}}.\boldsymbol{n})\, ds. \tag{2.37}$$

At every point on the Γ_R we shall use the same coordinate system $\{\boldsymbol{e}_r, \boldsymbol{e}_\phi\}$ as in the last section. In this system $\boldsymbol{u} = [u_r, u_\phi]^T$, $\boldsymbol{n} = [1, 0]^T$. As a result, we have to compute the integral

$$\delta e_R = \frac{1}{2}\int_{\Gamma_R} (\hat{\sigma}_{rr} u_r + \hat{\sigma}_{r\phi} u_\phi)\, ds. \tag{2.38}$$

Now we observe that $x_1^2 + x_2^2 = R^2$ on Γ_R and

$$\begin{aligned} u_r &= \tfrac{1}{R}(u_1 x_1 + u_2 x_2) \qquad & u_\phi &= \tfrac{1}{R}(-u_1 x_2 + u_2 x_1), \\ \sin\phi &= \tfrac{1}{R} x_2 & \cos\phi &= \tfrac{1}{R} x_1. \end{aligned}$$

To simplify the calculations we introduce the following notations:

$$\begin{aligned} &f = I(1,0) + I(0,1), \quad a = I(1,0) - I(0,1), \quad b = I(3,0) - I(0,3), \\ &c = I(0,1) + I(1,0), \quad d\ = I(0,3) - I(3,0). \end{aligned}$$

In these terms

$$\begin{aligned} \int_{\Gamma_R} u_r\, ds &= \pi R^2 f, \\ \int_{\Gamma_R} u_r \cos 2\phi\, ds &= \pi R^2 \left(\tfrac{3}{2}b - a\right), \\ \int_{\Gamma_R} u_r \sin 2\phi\, ds &= \pi R^2 \left(2c - \tfrac{3}{2}d\right), \\ \int_{\Gamma_R} u_\phi \cos 2\phi\, ds &= \pi R^2 \left(\tfrac{3}{2}b - 2a\right), \\ \int_{\Gamma_R} u_r \cos 2\phi\, ds &= \pi R^2 \left(\tfrac{3}{2}d - c\right). \end{aligned} \tag{2.39}$$

Now, due to (2.27) and (2.28),

$$f = \epsilon_{11} + \epsilon_{22}, \quad a = \epsilon_{11} - \epsilon_{22} + \delta_1, \quad b = \epsilon_{11} - \epsilon_{22} + \left(1 - \frac{1}{9k}\right)\delta_1,$$
$$c = \gamma_{12} - \delta_2, \quad d = \gamma_{12} - \left(1 + \frac{1}{9k}\right)\delta_2.$$

Substituting this into (2.38) gives

$$\delta e_R = -\frac{1}{2}\pi\rho^2 \left[\sigma_I(\epsilon_{11} + \epsilon_{22}) + \sigma_{II}(\epsilon_{11} - \epsilon_{22}) + \sigma_{12}\gamma_{12} + \left(1 - \frac{1}{k} + \frac{\rho^2}{R^2}\frac{1}{k}\right)(\sigma_{II}\delta_1 - \sigma_{12}\delta_2)\right]. \tag{2.40}$$

From Hooke's law follows

$$\sigma_I = (\lambda + \mu)(\epsilon_{11} + \epsilon_{22}), \quad \sigma_{II} = \mu(\epsilon_{11} - \epsilon_{22}), \quad \sigma_{12} = \mu\gamma_{12}$$

then, because of (2.29),

$$\delta e_R = -\pi\rho^2 e_u(0) - \frac{1}{2}\pi\rho^2 \left[\left(1 - \frac{1}{k} + \frac{\rho^2}{R^2}\frac{1}{k}\right)(\sigma_{II}\delta_1 - \sigma_{12}\delta_2)\right]. \tag{2.41}$$

This makes it different from the Laplace equation case, where the additional term vanishes. Observe that in order to solve the elasticity problem in the domain containing the hole with accuracy (outside Γ_R) up to $o(\rho^2)$, we do not need, due to (2.41), the solution in the intact domain. Simultaneously all the terms in (2.41) are quadratic with respect to $\boldsymbol{u}$ and introduce no difficulty into numerical procedures.

If we restrict ourselves to the terms depending strictly on ρ^2 and take into account the value of k, the energy corrections take on the form

$$\delta e_R = -\pi\rho^2 e_u(0) - \pi\rho^2 \frac{\mu}{\lambda + 3\mu}(\sigma_{II}\delta_1 - \sigma_{12}\delta_2). \tag{2.42}$$

In order to make clear that the energy correction is indeed an integral bilinear form of $\boldsymbol{u}$ defined over Γ_R, we collect below the dependences given by (2.27), (2.28), and (2.29) and write down the explicit expression for the terms appearing in (2.42):

$$\epsilon_{11} + \epsilon_{22} = \frac{1}{\pi R^3}\int_{\Gamma_R}(u_1x_1 + u_2x_2)\,ds,$$
$$\epsilon_{11} - \epsilon_{22} = \frac{1}{\pi R^3}\int_{\Gamma_R}\left[(1 - 9k)(u_1x_1 - u_2x_2) + \frac{12k}{R^2}(u_1x_1^3 - u_2x_2^3)\right] ds,$$

$$\gamma_{12} = \frac{1}{\pi R^3} \int_{\Gamma_R} \left[(1+9k)(u_1x_2 + u_2x_1) - \frac{12k}{R^2}\left(u_1x_2^3 + u_2x_1^3\right)\right] ds,$$

$$\delta_1 = \frac{9k}{\pi R^3} \int_{\Gamma_R} \left[(u_1x_1 - u_2x_2) - \frac{4}{3R^2}\left(u_1x_1^3 - u_2x_2^3\right)\right] ds,$$

$$\delta_2 = \frac{9k}{\pi R^3} \int_{\Gamma_R} \left[(u_1x_2 + u_2x_1) - \frac{4}{3R^2}\left(u_1x_2^3 + u_2x_1^3\right)\right] ds.$$

These expressions are easy to compute numerically, but unfortunately the correction formula is not as compact as in the Laplace operator case.

Acknowledgment

This research was performed with the cooperation of the Centre National de la Recherche Scientifique and the Polish Academy of Sciences, PECO-NEI, Projet 18282, and was partially supported by grant No. 4 T11A 01524 of the State Committee for the Scientific Research of the Republic of Poland.

References

[1] M.O. Bašeleĭšvili, An analog of the Poisson formula in elasticity (Georgian, with Russian abstract), *Trudy Wyčislitelnovo Centra AN Gruzinskoĭ SSR,* 1 (1960), 97–101.

[2] T. Lewinski and J. Sokolowski, *Optimal shells formed on a sphere. The topological derivative method.* RR-3495, INRIA-Lorraine, 1998.

[3] T. Lewinski and J. Sokolowski, Energy change due to appearing of cavities in elastic solids, *Int. J. Solids & Structures*, 40 (2003), 1765–1803.

[4] T. Lewinski, J. Sokolowski, A. Zochowski, *Justification of the bubble method for the compliance minimization problems of plates and spherical shells*, CD-Rom, 3rd World Congress of Structural and Multidisciplinary Optimization (WCSMO-3) Buffalo/Niagara Falls, New York, May 17–21, 1999.

[5] W. G. Mazja, S. A. Nazarov, and B. A. Plamenevskii, *Asymptotic theory of elliptic boundary value problems in singularly perturbed domains*, Vols. 1 and 2, Birkhäuser, Basel, 2000.

[6] S. A. Nazarov and B. A. Plamenevsky, *Elliptic problems in domains with piecewise smooth boundaries*, De Gruyter Exposition in Mathematics 13, Walter de Gruyter, Berlin, 1994.

[7] S. A. Nazarov and J. Sokołowski, Asymptotic analysis of shape functionals, *Journal de Mathématiques pures et appliquées*, 82, 2 (2003), 125–196.

[8] S. A. Nazarov and J. Sokołowski, *Self-adjoint extensions for the Neumann Laplacian in application to shape optimization*, Les prépublications de l'Institut Élie Cartan 9/2003; http://www.iecn.u-nancy.fr/Preprint/publis/preprints-2003.html

[9] S. A. Nazarov and J. Sokołowski, The topological derivative of the Dirichlet integral due to formation of a thin ligament, *Siberian Math. J.*, 45, 2 (March–April 2004), 341–355.

[10] S. A. Nazarov and J. Sokołowski, Self-adjoint extensions of differential operators in application to shape optimization, *Comptes Rendus Mecanique*, 331, 10 (October 2003), 667–672.

[11] S. A. Nazarov and J. Sokołowski, Techniques of asymptotic analysis in shape optimization, *French–Russian A.M. Liapunov Institute for Applied Mathematics and Computer Science Transactions*, 4 (2003), 49–57.

[12] S. A. Nazarov and J. Sokołowski, Self-adjoint extensions for elasticity system in application to shape optimization, to appear in *Bulletin of the Polish Academy of Sciences—Mathematics.*

[13] S. A. Nazarov, A.S. Slutskij, and J. Sokołowski, *Topological derivative of the energy functional due to formation of a thin ligament on the spatial body*, Les prépublications de l'Institut Élie Cartan 14/2004; http://www.iecn.u-nancy.fr/Preprint/publis/index.html

[14] M. Rao and J. Sokołowski, Non-linear balayage and applications, *Illinois J. Math.*, 44 (2000), 310–328.

[15] M. Rao and J. Sokołowski, *Tangent sets in Banach spaces and applications to variational inequalities*, Les prépublications de l'Institut Élie Cartan, 42, 2000.

[16] J. Sokołowski and J-P. Zolésio, Introduction to shape optimization. *Shape sensitivity analysis*, Springer-Verlag, Heidelberg, 1992.

[17] J. Sokołowski and A. Zochowski, *On topological derivative in shape optimisation*, INRIA-Lorraine, Rapport de Recherche No. 3170, 1997.

[18] J. Sokołowski and A. Zochowski, On topological derivative in shape optimization, *SIAM Journal on Control and Optimization*, 37, 4 (1999), 1251–1272.

[19] J. Sokołowski and A. Zochowski, Topological derivative for optimal control problems, *Control and Cybernetics*, 28, 3 (1999), 611–626.

[20] J. Sokołowski and A. Zochowski, Topological derivatives for elliptic problems, *Inverse Problems,* 15, 1 (1999), 123–134.

[21] J. Sokołowski and A. Zochowski, Topological derivatives of shape functionals for elasticity systems, *Mechanics of Structures and Machines*, 29 (2001), 333–351.

[22] J. Sokołowski and A. Zochowski, Optimality conditions for simultaneous topology and shape optimization, *SIAM Journal on Control and Optimization,* 42, 4 (2003), 1198–1221.

CHAPTER 17

The computing zoom

J. Henry
INRIA-Futurs, University of Bordeaux, Talence, France

Introduction

We consider a common situation where one is interested in the solution only in a (variable) subdomain $\omega \subset \Omega$ of an elliptic boundary value problem set in Ω.

One wants to

- have a finer description of the solution in the subdomains of interest ω;
- interact easily with the solution: recompute it only in ω when data are changed.

In order to achieve this goal we will present a method to compute transparent boundary conditions on the boundary of ω, that is, conditions that summarize exactly the behavior of the solution outside of ω. This is done by means of a spatial invariant embedding technique. The invariant embedding technique was devised by Bellman (see [1]) in the context of optimal control theory to derive optimal feedback. The original problem is embedded in a family of similar problems on a shorter horizon. We apply here the same idea spacewise for linear elliptic boundary value problems as in [2]: The original problem in a star-shaped domain is now embedded in a family of similar problems in smaller domains defined by homothety.

We first recall from [3] the factorization method of a boundary value problem resulting from this spatial invariant embedding technique in a simple geometric situation—the cylinder case. We show the link with the LU factorization for the discretized problem. We also show the link of the operators derived by this method and the Green function, and also by Hadamard's formula. Then we present the zooming technique for a star-shaped domain with a homothetic region of interest. Finally, we explain how the method is used numerically in the case of finite differences.

1 Method of factorization by space invariant embedding

In this section we recall from [3] the factorization of boundary value problems for elliptic equation by space invariant embedding. The method is most easily explained in the case of a cylinder. Let Ω be the cylinder $\Omega =]0,a[\times\mathcal{O}$, $x' = (x,z) \in \mathbf{R}^n$, where x is the coordinate along the axis of the cylinder and the $\mathcal{O}$ bounded open set in $\mathbf{R}^{n-1}$ is the section of the cylinder. Let $\Sigma =]0,a[\times\partial\mathcal{O}$ be the lateral boundary and $\Gamma_0 = \{0\}\times\mathcal{O}$, $\Gamma_a = \{a\}\times\mathcal{O}$ be the faces of the cylinder.

We consider the following Poisson equation with mixed boundary conditions:

$$(\mathcal{P}_0)\begin{cases} -\Delta y = -\frac{\partial^2 y}{\partial x^2} - \Delta_z y = f \quad \text{in } \Omega, \\ y|_\Sigma = 0, \\ \left(-\dfrac{\partial y}{\partial x} + \alpha y\right)|_{\Gamma_0} = -y_0, \quad y|_{\Gamma_a} = y_1. \end{cases}$$

We embed this problem in the family of similar problems defined over subcylinders limited by the *moving boundary* Γ_s:

$$(\mathcal{P}_{s,h})\begin{cases} -\Delta y = f \quad \text{in } \ \Omega_s =]0,s[\times\mathcal{O} \\ y|_\Sigma = 0, \\ \left(-\dfrac{\partial y}{\partial x} + \alpha y\right)\Bigg|_{\Gamma_0} = -y_0, \quad y|_{\Gamma_s} = h. \end{cases}$$

The Dirichlet-Neumann map on Γ_s: $h \longrightarrow y|_{\Gamma_s}$ is affine:

$$\left.\frac{\partial y}{\partial x}\right|_{\Gamma_s} = P(s)h + w(s). \tag{1.1}$$

Choosing $s = x$ and $h = y(x)$, solution of $(\mathcal{P}_0)$ for an arbitrary y_1 and taking the derivative with respect to x:

$$\frac{\partial^2 y}{\partial x^2} = -\Delta_z y - f = \frac{dP}{dx}y + P\frac{\partial y}{\partial x} + \frac{\partial w}{\partial x},$$

and substituting $\dfrac{\partial y}{\partial x}$ from (4.10)

$$0 = \left(\frac{dP}{dx} + P^2 + \Delta_z\right)y + \frac{\partial w}{\partial x} + Pw + f.$$

The *trajectory* of $y(x)$ being arbitrary, we can identify to zero the term depending linearly on y and the term independent. One gets the decoupled system

$$\frac{dP}{dx} + P^2 + \Delta_z = 0; \quad P(0) = \alpha I, \tag{1.2}$$

$$\frac{dw}{dx} + Pw = -f; \qquad w(0) = y_0, \tag{1.3}$$

$$-\frac{dy}{dx} + Py = -w; \qquad y(a) = y_1, \tag{1.4}$$

where P and w are to be integrated from 0 to a then y backward from a to 0. P is an operator on functions defined on $\mathcal{O}$ satisfying a Riccati equation. The initial conditions for P and w are obtained from (1.1) written at $x = 0$.

We obtain the *factorization* of the elliptic boundary value problem $(\mathcal{P}_0)$ in the product of two Cauchy problems of parabolic type in opposite directions

$$\text{“}-\Delta\text{”} = -\left(\frac{d}{dx} + P\right)\left(\frac{d}{dx} - P\right).$$

Remark

- Once the Riccati equation for P is integrated, for a new set of data y_0, y_a, f it suffices to integrate two uncoupled, first-order equations for w and y.
- For a problem set on the complementary domain $\bar{\Omega}_s =]s,\, a[\times\mathcal{O}$, the condition on Γ_s
$$-\frac{\partial y}{\partial x}|_{\Gamma_s} + P(s)y|_{\Gamma_s} = w(s), \tag{1.5}$$
sums up the solution on Ω_s: the solution of the problem set on $\bar{\Omega}_s$ with boundary condition (1.5) on Γ_s is exactly the restriction of the solution of $(\mathcal{P}_0)$ to $\bar{\Omega}_s$.

2 Link with Gauss factorization

We recall that the previous factorization can be viewed as an extension of the infinite dimensional problem of the usual Gauss LU block factorization for matrices. Let us consider a finite difference discretization of problem $(\mathcal{P}_0)$ in 2D for the sake of simplicity. Let h be the discretization step in both x and z directions. We use a classical 5-point scheme, numbering the nodes first in the z direction. The discretized Laplacian with block notations reads

$$A_h = \frac{1}{h^2}\begin{pmatrix} B_1 & -I & & & \\ -I & B_2 & -I & & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & & -I & B_{N-2} & -I \\ & & & -I & B_{N-1} \end{pmatrix}.$$

It is block tridiagonal, with $B_i = 2I - \Delta_{z,h}$, where $\Delta_{z,h}$ is the discretized Laplacian on the section. We perform a discrete invariant embedding with respect to index i in the x direction similar to the one of the previous section. Derivatives are replaced by finite differences. We set

$$\frac{y^{i+1}-y^i}{h}=P_iy^i+w^i.$$

We get

$$\begin{cases} -\frac{P_i-P_{i-1}}{h}=P_{i-1}(I+hP_{i-1})^{-1}P_{i-1}+\Delta_{z,h}, & \forall i\in\{1,...,N-1\} \\ P_0=0. \end{cases} \tag{2.6}$$

$$\begin{cases} \frac{w^i-w^{i-1}}{h}=-P_{i-1}(I+hP_{i-1})^{-1}w^{i-1}-f^i, & \forall i\in\{1,...,N-1\} \\ w^0=-y_0(a_{1/2}), \end{cases} \tag{2.7}$$

$$\begin{cases} \frac{y^{i+1}-y^i}{h}=P_iy^i+w^i, & \forall i\in\{1,...,N-1\} \\ y^N=y_1(a_N), \end{cases} \tag{2.8}$$

which is a discretized version of (1.2). Now setting $L_i=(I+hP_i)^{-1}$ and $U_i=I+hP_i$, we can rewrite (2.6), (2.7), and (2.8) in matrix form as

$$A_h=\frac{1}{h}\begin{pmatrix} I & & & \\ -L_1 & I & & 0 \\ 0 & \ddots & \ddots & \\ & & -L_{N-2} & I \end{pmatrix}\frac{1}{h}\begin{pmatrix} U_1 & -I & & \\ & \ddots & \ddots & 0 \\ 0 & & U_{N-2} & -I \\ & & & U_{N-1} \end{pmatrix},$$

which is exactly the LU Gauss block factorization of A_h.

3 Link with Green functions and Hadamard's formula

We consider now problem $(\mathcal{P}_0)$ with a Dirichlet boundary condition on Γ_0. Let $R(t,x)$ and $S(s,x)$ be the forward and backward evolution operators related to P.

$$\frac{dR}{dx}(t,x)+P(x)R(t,x)=0, \quad R(t,t)=I,$$

$$-\frac{dS}{dx}(s,x)+P(x)S(s,x)=0, \quad S(s,s)=I.$$

Then one can show from (1.2) and (2.8) that the solution of $(\mathcal{P}_0)$ is given explicitly by

$$y(x)=S(a,x)y_1+\int_x^a\int_0^t S(t,x)R(\tau,t)f(\tau)d\tau dt+\int_x^a S(t,x)R(0,t)y_0\,dt.$$

Let $R(t, x; z, \zeta)$ be the kernel of the operator $R(t, x)$ on the function of z. Then the Green function is expressed as

$$G(a; x, z; \xi, \zeta) = \int_{\mathcal{O}} \int_{x \wedge \xi}^{a} S(t, x, z, z') R(\xi, t, z', \zeta) dt\ dz'. \qquad (3.9)$$

By taking derivatives of the preceding formula (3.9) we have

$$\begin{aligned} \frac{\partial G}{\partial \xi}(a; x, z; \xi, \zeta)|_{\xi=a} &= -S(a, x, z, \zeta) \\ \frac{\partial G}{\partial x}(a; x, z; \xi, \zeta)|_{x=a} &= -R(\xi, a, z, \zeta) \\ \frac{\partial G}{\partial a}(a; x, z; \xi, \zeta) &= \int_{\mathcal{O}} S(a, x, z, z') R(\xi, a, z', \zeta) dz'. \end{aligned}$$

Then we recover Hadamard's formula

$$\frac{\partial G}{\partial a}(a; x, z; \xi, \zeta) = \int_{\mathcal{O}} \frac{\partial G}{\partial \xi}(a; x, z; \xi, z')|_{\xi=a} \frac{\partial G}{\partial x}(a; x, z'; \xi, \zeta)|_{x=a} dz'.$$

4 The zooming technique

We intend now to extend the method to more general geometries of the domain. We will apply the invariant embedding to homothetical domains. Let Ω be a bounded open set, star-shaped with respect to O of $\mathbf{R}^2$ and $\Gamma = \partial\Omega$. We consider the problem

$$(\mathcal{P}_0) \begin{cases} -\Delta y = f & \text{in } \Omega, \\ \dfrac{\partial y}{\partial n}|_\Gamma + \alpha y|_\Gamma = y_0. \end{cases}$$

Let $\mathcal{H}(O, s)$ be the homothety of center O and ratio s, $0 < s < 1$. We embed $(\mathcal{P}_0)$ in the family of problems

$$(\mathcal{P}_{s,h}) \begin{cases} -\Delta y = f & \text{in } \Omega_s, \\ \dfrac{\partial y}{\partial n}|_\Gamma + \alpha y|_\Gamma = y_0 & \text{on } \Gamma \\ y|_{\Gamma_s} = h, \end{cases}$$

where $\Omega_s = \Omega \setminus \mathcal{H}(O, s)\Omega$ is the annulus with boundaries Γ and Γ_s.
By linearity

$$\left.\frac{\partial y}{\partial n}\right|_{\Gamma_s} = P(s)h + r(s), \qquad (4.10)$$

where $P(s)$ is the Dirichlet-Neumann map on Γ_s

$$P(s) \in \mathcal{L}(H^{1/2}(\Gamma_s), H^{1/2}(\Gamma_s)').$$

Let us change the independent variables to (t, τ) where t is the curvilinear abscissa along Γ and $0 < \tau < 1$ the homothety ratio. The boundary Γ is defined by the function $\varphi(t)$

$$\Gamma = \{t, \rho = \varphi(t)\},$$

where ρ is the distance from O. At a point M with coordinates (t, τ) we denote α the angle (OM, n) where n is the outward normal to $\mathcal{H}(O, \tau)\Omega$. With these new coordinates the variational formulation of $\mathcal{P}_{s,h}$, if we use a test function z for which we allow an arbitrary value on Γ_s, is given by

$$\int_{\Omega_s} \left(\frac{\tau}{\varphi \cos\alpha} \frac{\partial y}{\partial \tau} \frac{\partial z}{\partial \tau} - \tan\alpha \left(\frac{\partial y}{\partial t} \frac{\partial z}{\partial \tau} + \frac{\partial y}{\partial \tau} \frac{\partial z}{\partial t} \right) + \frac{\varphi}{\tau \cos\alpha} \frac{\partial y}{\partial t} \frac{\partial z}{\partial t} \right) dt\, d\tau$$

$$= \int_{\Omega_s} \tau \varphi \cos\alpha f z \, dt\, d\tau + \int_{\Gamma_s} s \frac{\partial y}{\partial n} z \, dt \qquad (4.11)$$

with

$$\left.\frac{\partial y}{\partial n}\right|_{\Gamma_s} = -\frac{1}{\varphi \cos\alpha} \frac{\partial y}{\partial \tau} + \frac{\tan\alpha}{\tau} \frac{\partial y}{\partial t}.$$

In order to derive the equations for P and r we choose a function z not depending on τ, $z = z(t)$, as a test function in (4.11). Substituting $\frac{\partial y}{\partial n}|_{\Gamma_s}$ from (4.10) into (4.11) and taking the derivative of (4.11) with respect to s for $h = y|_{\Gamma_s}$ yields an equation defined on Γ_s. Then using the argument that both $y|_{\Gamma_s}$ and z are arbitrary as in the previous computation in the cylinder case, we derive a Riccati equation for P

$$-\frac{dP}{d\tau} - \frac{1}{\tau} P + \frac{\partial}{\partial t} \circ \frac{\varphi \sin\alpha}{\tau} P + \left(\frac{\partial}{\partial t} \circ \frac{\varphi \sin\alpha}{\tau} P \right)^*$$

$$+ P\varphi \cos\alpha \, P = -\frac{\partial}{\partial t} \left(\frac{\varphi \cos\alpha}{\tau^2} \frac{\partial}{\partial t} \right), \qquad (4.12)$$

$$P(1) = \alpha I,$$

and an equation for the residue r

$$-\frac{\partial r}{\partial \tau} + \frac{\partial}{\partial t} \frac{(\varphi \sin\alpha r)}{\tau} + P\varphi \cos\alpha r = \tau \varphi \cos\alpha f, \qquad (4.13)$$

with initial condition

$$r(1) = y_0.$$

Due to the nonorthogonality of the coordinates, the Riccati equation incorporates linear terms now. One can show that P is positive and self-adjoint for $\tau \le 1$ and coercive on $H^{1/2}(\Gamma_\tau)$ for $\tau < 1$. Equation (4.13) is a well-posed abstract parabolic problem. The term in $\frac{P}{\tau}$ in (4.12) is due to polar-like coordinates and can be eliminated by the classical change of coordinate $\mu = \log(\tau)$. Furthermore, we take $Q = \tau P$ as unknown, yielding

$$\begin{aligned} -\frac{dQ}{d\mu} + \frac{\partial}{\partial t} \circ \varphi \sin\alpha Q + \left(\frac{\partial}{\partial t} \circ \varphi \sin\alpha Q\right)^* \\ + Q\varphi\cos\alpha\; Q = -\frac{\partial}{\partial t}\left(\varphi\cos\alpha\frac{\partial}{\partial t}\right) \qquad (4.14) \\ -\frac{\partial r}{\partial\mu} + \frac{\partial}{\partial t}(\varphi\sin\alpha r) + Q\varphi\cos\alpha r = \tau\varphi\cos\alpha f \qquad (4.15) \end{aligned}$$

with initial conditions

$$Q(0) = \alpha I, \quad r(0) = y_0$$

to be solved backward from 0 to $-\infty$. If we would like to complete the factorization, we would have to solve forward an equation for y coming from (4.10) from $\mu = -\infty$ to $\mu = 0$. The singularity at the origin ($\mu = -\infty$) has been studied in [4].

Instead here we are interested in solving the problem in the region of interest ω, which is supposed to be homothetical to Ω, $\omega = \mathcal{H}(O,\lambda)\Omega$. Then we have to solve (4.12) and (4.13) for $\lambda \le s \le 1$, or (4.14) and (4.15) for $\mu_0 \le \mu \le 0$, $\mu_0 = \log(\lambda) < 0$.

Now, knowing $Q(\mu_0)$ and $r(\mu_0)$, due to transmission conditions on γ, problem $(\mathcal{P}_0)$ restricted to ω is equivalent to

$$(\mathcal{P}_\omega)\begin{cases} -\Delta y = f & \text{in } \omega, \\ \dfrac{\partial y}{\partial n} + \dfrac{Q(\mu_0)}{\lambda} y = -r(\mu_0) & \text{on } \gamma = \partial\omega. \end{cases} \qquad (4.16)$$

The *transparent* boundary condition on γ in (4.16) seems similar to a Robin boundary condition, although it is nonlocal due to the effect of Q. It is easy to prove that (4.16) is well posed by the Lax Milgram theorem due to the coerciveness of Q.

The *computing zoom technique* can now be summarized: to solve $(\mathcal{P}_0)$ in ω only we need to

- solve (4.14) and (4.15) from 0 to μ_0
- solve

$$(\mathcal{P}_\omega)\begin{cases} -\Delta y = f & \text{in } \omega, \\ \dfrac{\partial y}{\partial n} + \dfrac{Q(\mu_0)}{\lambda} y = -r(\mu_0) & \text{on } \gamma = \partial\omega. \end{cases}$$

As it was desired, once Q is computed, the interaction with the solution in ω is local:

- if f is changed in ω, the solution of $(\mathcal{P}_\omega)$ has to be recomputed in ω with the same boundary conditions;
- if f is changed outside ω, r has to be recomputed by (4.15) also, but Q is unchanged.

Numerical tests have been performed with a rectangular domain Ω. We used finite differences discretization for both the outside and inside ω calculations. Outside we discretize the (μ, t) coordinates with a constant step size. The equation for Q and r being of parabolic type are solved iteratively with an explicit scheme from $\mu = 0$ to $\mu = \mu_0$ taking into account the stability condition. The number of unknowns is proportional to $-\log(\frac{1}{\lambda})$, the logarithm of the magnification ratio. Inside ω we use a rectangular mesh and constant relative mesh size (i.e., with a fixed number of unknowns independent of λ). The system is solved by a conjugate gradient algorithm.

5 Conclusion

We presented a method based on homothetic spatial invariant embedding, which allows us to eliminate the unknown function y outside the domain of interest ω. Numerical experiments show that one can increase the magnification ratio of the domain of interest and interact easily with the solution in this domain. The space- and time-consuming part of the computation is the calculation of the operator Q, which is done once for all. For problems in higher dimensions, one can think of applying a domain decomposition method to compute Q. It is unclear if the method can be extended to a translation of the region of interest, as then the stability of the resolution of the Riccati equation remains to be studied.

References

[1] Bellman, R., *Introduction to the Mathematical Theory of Control Processes. Volume 1: Linear Equations and Quadratic Criteria*, Academic Press, New York, 1967.

[2] Angel, E. and Bellman, R., *Dynamic Programming and Partial Differential Equations*, Academic Press, New York, 1971.

[3] Henry, J. and Ramos, A.M., Factorization of second order elliptic boundary value problems by dynamic programming. To appear in *Nonlinear Analysis. Theory and Applications.*

[4] J. Henry, B. Louro, and M.C. Soares, *A factorization method for elliptic problems in a circular domain*, *C. R. Acad. Sci. Paris* serie 1, 339 (2004), 175–180.

[5] Lions, J.L., *Contrôle Optimal de Systèmes Gouvernés par des Équations aux Dérivées Partielles*, Dunod, Paris, 1968.

[6] Lions, J. L. and Magenes, E., *Problèmes aux Limites Non Homogènes et Applications*, Vols. 1 and 2, Dunod, Paris, 1968.

[7] Ramos, A. M., *Algunos problemas en ecuaciones en derivadas parciales relacionados con la teoría de Control*, Ph.D. Thesis, Universidad Complutense de Madrid, 1996.

[8] Reid, W. T., *Riccati Differential Equations*, Academic Press, New York, 1972.

CHAPTER 18

An optimization approach for the delamination of a composite material with nonpenetration

M. Hintermüller
Department of Mathematics and Scientific Computing,
University of Graz, Graz, Austria

V.A. Kovtunenko
Lavrent'ev Institute of Hydrodynamics,
Novosibirsk, Russia

K. Kunisch
Department of Mathematics and Scientific Computing,
University of Graz, Graz, Austria

1 Introduction

A mathematical formulation of crack problems can be given within the framework of elasticity theory [20]. To gain some insight into the 3-dimensional situation, the standard approach is to simplify the elasticity model by splitting it into two 2-dimensional in-plane and anti-plane models. This, however, leads to a loss of information concerning the 3-dimensional nature of the system. Motivated by this drawback, we introduce an intermediate 2.5-dimensional model instead of the splitting approach. Our model is a spatial one because it takes into account all 3 components of the displacement vector, and still it is formulated in a 2-dimensional domain.

For the construction of the 2.5-dimensional model we consider a homogeneous orthotropic material with a vertical plane of elastic symmetry rotated with an angle β to a reference coordinate system. We compose two pieces of such a material along the interface given by the plane $x_2 = 0$ such that the corresponding angles in the upper and lower half-spaces are β and $-\beta$, respectively. For the formulation of elasticity models in composite laminates see [21]. We further assume that a crack is situated on part of the interface. Applying the assumption of plain deformation at $x_3 = const$, then due to the rotation, this results in a spatial model.

In our numerical experiments we observe 3-dimensional effects: mixing of crack modes (mode-1 with mode-3), and contact between opposite crack surfaces. They occur under pure mode-3 loading, which is ruled out for the in-plane and anti-plane models. Due to the latter phenomenon we are required to consider (unilaterally) constrained crack problems with nonpenetration conditions as suggested in [12] and [13]. The corresponding variational formulation provides the appropriate state space for the crack problem, which has a singularity at the crack tip.

We investigate the geometric and physical features of the composite model by numerical experiments. For this purpose a semismooth Newton technique is adapted to constrained crack problems. Under suitable assumptions, semismoothness concepts will allow a locally superlinear convergence rate of the Newton iterates. For the problems under consideration, semismooth Newton methods are equivalent to primal-dual active-set algorithms (see [7] and [11]). They are an efficient tool for the numerical treatment of constrained variational problems. In numerical experiments, global and monotone convergence was observed, which is supported by the *a posteriori* analysis in [8]. For a class of variational problems subject to boundary constraints, in [9] we applied an argument based on perturbation of M-matrices guaranteeing these convergence properties. Returning to the continuous setting of the problem, a penalization technique was utilized in [10] to obtain an approximate Lagrange multiplier, which enjoys extra L^p-regularity. For the numerical treatment of curvilinear cracks we refer to [19] and [23], where extended finite element techniques are used.

One of the principal questions in fracture mechanics and structure design is to describe the stability properties of a solid with a crack and to predict its growth. By the Griffith fracture hypothesis, the propagation of a crack is determined by the energy release rate at the crack tip, which cannot exceed a given physical parameter (see [3] and [22]). A large number of papers investigated quasi-static growth of cracks in elastic media (see, e.g., [1], [6], [16], and [20]). We argue that the energy release rate is the shape derivative of the potential energy functional with respect to variations of the crack tip. In [12] and [13] methods of shape sensitivity analysis were adapted to crack problems with nonpenetration conditions to provide a formula for the shape derivative. This includes the Griffith formula as a specific case.

We observe that the Griffith fracture criterion provides a necessary optimality condition for a local minimum (if it exists) of the total potential energy, which is defined as the sum of the potential and the surface energies. On the other hand, global optimization problems require minimization over all admissible crack shapes (see [2] and [5]). For strictly convex cost functionals these two concepts coincide. In fracture mechanics this corresponds to stable crack propagation (progressive). The case of unstable (or brutal) crack growth is related to nonconvex cost functionals. It was noticed in [5] that for brutal growth the Griffith fracture law (as a local criterion) predicts a critical loading for the initiation of crack propagation larger than that needed by the global optimization approach. This fact is observed in our numerical tests, too. By the global formulation of the optimization problem, not only continuous solutions for the stable crack propagation but also solutions with jumps and discontinuous velocities of the propagation are obtained.

Well-posedness properties for time-evolution problems with cracks were analyzed in [4] and [5]. In the present work we apply the global formulation of the optimization problem to a rectilinear crack and utilize it on a set of critical points derived in a constructive way from the Griffith fracture law. Note that the delamination process suggests a predefined path (along the interface) of the crack time evolutions, which was confirmed experimentally in [14]. This problem is solved numerically to describe the delamination of composite materials with an interface crack under quasi-static linear loading.

2 Constrained crack problems for a composite

In this section we formulate a model with respect to an in-plane deformation for two identical homogeneous orthotropic materials, which are composed at a planar interface with the angle of 2β between their vertical planes of elastic symmetry, and which have a crack along a part of their interface.

2.1 Modeling of composite materials in plane deformation

Consider a homogeneous orthotropic material with planes of elastic symmetry corresponding to the (x'_1, x'_2, x'_3)-axes. First, we compose the identical materials with respect to a reference coordinate system (x_1, x_2, x_3) in the following way. In the upper half-space $\mathbb{R}^3_+ = \{x_1, x_2 \geq 0, x_3\}$ the (x'_1, x'_2, x'_3)-axes are rotated in the anti-clockwise direction to (x_1, x_2, x_3) with respect to the common $x'_2 = x_2$-axis by the angle β between x'_3 and x_3. The angle $\beta \in [-\pi/2, \pi/2]$ is arbitrarily fixed. In the lower half-space $\mathbb{R}^3_- = \{x_1, x_2 \leq 0, x_3\}$, the (x'_1, x'_2, x'_3)-axes are rotated to (x_1, x_2, x_3) with respect to $x'_2 = x_2$ in the opposite direction by the same angle. The materials are assumed to be joined along the plane $x_2 = 0$.

For a displacement vector $u = (u_1, u_2, u_3)^\top(x)$, at a point

$$x = (x_1, x_2, x_3)^\top \in \mathbb{R}^3,$$

in the composite material

$$u = u^+ \quad \text{in } \mathbb{R}^3_+, \quad u = u^- \quad \text{in } \mathbb{R}^3_-,$$

we introduce a strain tensor $\varepsilon = \{\varepsilon_{ij}\}$ according to the linear Cauchy law and a 3×3 symmetric tensor of stress $\sigma = \{\sigma_{ij}\}$ as

$$\sigma(u) = \sigma^{\beta}(u^{+}) \quad \text{in } \mathbb{R}^3_+, \quad \sigma(u) = \sigma^{-\beta}(u^{-}) \quad \text{in } \mathbb{R}^3_-. \tag{2.1}$$

Here and throughout we utilize the standard tensor notation common in linear elasticity and the summation convention for the repeated indices $i, j = 1, 2, 3$.

Second, we apply the assumption of plane deformation at every cross-section $x_3 = const$, which means that none of the three components of the displacement vector u depend on x_3. Hence $\varepsilon_{33} = 0$ and the strain tensor takes the particular form

$$\begin{gathered} \varepsilon_{11}(u) = u_{1,1}, \quad \varepsilon_{22}(u) = u_{2,2}, \quad \varepsilon_{12}(u) = 0.5(u_{1,2} + u_{2,1}), \\ \varepsilon_{13}(u) = 0.5u_{3,1}, \quad \varepsilon_{23}(u) = 0.5u_{3,2}. \end{gathered} \tag{2.2}$$

In $\mathbb{R}^3_+$, the relevant components of the stress tensor (2.1) satisfy the following constitutive relations involving a nonsymmetric matrix:

$$\begin{bmatrix} \sigma^{\beta}_{11} \\ \sigma^{\beta}_{22} \\ \sigma^{\beta}_{12} \\ \sigma^{\beta}_{23} \\ \sigma^{\beta}_{13} \end{bmatrix} = \begin{bmatrix} C^{\beta}_{11} & C^{\beta}_{12} & 0 & 0 & 2C^{\beta}_{16} \\ C^{\beta}_{12} & C_{22} & 0 & 0 & 2C^{\beta}_{26} \\ 0 & 0 & 2C^{\beta}_{44} & 2C^{\beta}_{45} & 0 \\ 0 & 0 & 2C^{\beta}_{45} & 2C^{\beta}_{55} & 0 \\ C^{\beta}_{16} & C^{\beta}_{26} & 0 & 0 & 2C^{\beta}_{66} \end{bmatrix} \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{22} \\ \varepsilon_{12} \\ \varepsilon_{23} \\ \varepsilon_{13} \end{bmatrix} \tag{2.3}$$

where 9 elasticity coefficients depending on β (except for C_{22}) have the form (see [18]):

$$\begin{aligned} C^{\beta}_{11} &= C'_{33} \sin^4 \beta + 2(C'_{13} + 2C'_{66}) \sin^2 \beta \cos^2 \beta + C'_{11} \cos^4 \beta, \\ C^{\beta}_{66} &= C'_{66} + (C'_{33} + C'_{11} - 2C'_{13} - 4C'_{66}) \sin^2 \beta \cos^2 \beta, \\ C^{\beta}_{16} &= [C'_{11} \cos^2 \beta - C'_{33} \sin^2 \beta - (C'_{13} + 2C'_{66})(\cos^2 \beta - \sin^2 \beta)] \\ &\quad \times \sin \beta \cos \beta, \\ C^{\beta}_{44} &= C'_{44} \cos^2 \beta + C'_{55} \sin^2 \beta, \\ C^{\beta}_{55} &= C'_{44} \sin^2 \beta + C'_{55} \cos^2 \beta, \\ C^{\beta}_{45} &= (C'_{44} - C'_{55}) \sin \beta \cos \beta, \\ C^{\beta}_{12} &= C'_{23} \sin^2 \beta + C'_{12} \cos^2 \beta, \\ C^{\beta}_{26} &= (C'_{12} - C'_{23}) \sin \beta \cos \beta, \\ C_{22} &= C'_{22}. \end{aligned} \tag{2.4}$$

The coefficients subscribed with *prime* are related to the rotated coordinate system (x_1', x_2', x_3'):

$$\begin{aligned}
&C_{11}' = \theta\left(\frac{1}{E_2} - \frac{\nu_{32}^2}{E_3}\right), \quad C_{12}' = \theta\left(\frac{\nu_{21}}{E_2} + \frac{\nu_{31}\nu_{32}}{E_3}\right),\\
&C_{13}' = \theta\left(\frac{\nu_{31} + \nu_{21}\nu_{32}}{E_2}\right), \quad C_{22}' = \theta\left(\frac{1}{E_1} - \frac{\nu_{31}^2}{E_3}\right),\\
&C_{23}' = \theta\left(\frac{\nu_{32}}{E_1} + \frac{\nu_{21}\nu_{31}}{E_2}\right), \quad C_{33}' = \theta\frac{E_3}{E_2}\left(\frac{1}{E_1} - \frac{\nu_{21}^2}{E_2}\right),\\
&C_{44}' = G_{21}, \quad C_{55}' = G_{32}, \quad C_{66}' = G_{31},\\
&\frac{1}{\theta} = \left(\frac{1}{E_2} - \frac{\nu_{32}^2}{E_3}\right)\left(\frac{1}{E_1} - \frac{\nu_{31}^2}{E_3}\right) - \left(\frac{\nu_{21}}{E_2} + \frac{\nu_{31}\nu_{32}}{E_3}\right)^2,
\end{aligned} \tag{2.5}$$

with the material parameters

$$E_1,\ E_2,\ E_3,\ \nu_{21},\ \nu_{32},\ \nu_{31},\ G_{21},\ G_{32},\ G_{31}.$$

The elasticity coefficients obey the following symmetry properties:

$$\begin{gathered}
C_{11}^{-\beta} = C_{11}^{\beta},\ C_{12}^{-\beta} = C_{12}^{\beta},\ C_{44}^{-\beta} = C_{44}^{\beta},\ C_{55}^{-\beta} = C_{55}^{\beta},\ C_{66}^{-\beta} = C_{66}^{\beta},\\
C_{16}^{-\beta} = -C_{16}^{\beta},\ C_{26}^{-\beta} = -C_{26}^{\beta},\ C_{45}^{-\beta} = -C_{45}^{\beta}.
\end{gathered} \tag{2.6}$$

Note that if $\beta = 0$ or $\beta = \pm\pi/2$, then we have $C_{16}^{\beta} = C_{26}^{\beta} = C_{45}^{\beta} = 0$ and (2.2) and (2.3) are split into two independent states, namely the in-plane state for $(u_1, u_2)^\top$ and the anti-plane state for u_3. If $\beta \neq 0, \pm\pi/2$ then we have a spatial model.

The substitution of (2.2) into (2.3) allows us to rewrite the constitutive law in the symmetric form:

$$\begin{aligned}
\sigma_{11}^{\beta}(u) &= C_{11}^{\beta}u_{1,1} + C_{12}^{\beta}u_{2,2} + C_{16}^{\beta}u_{3,1},\\
\sigma_{22}^{\beta}(u) &= C_{12}^{\beta}u_{1,1} + C_{22}u_{2,2} + C_{26}^{\beta}u_{3,1},\\
\sigma_{12}^{\beta}(u) &= C_{44}^{\beta}(u_{1,2} + u_{2,1}) + C_{45}^{\beta}u_{3,2},\\
\sigma_{23}^{\beta}(u) &= C_{45}^{\beta}(u_{1,2} + u_{2,1}) + C_{55}^{\beta}u_{3,2},\\
\sigma_{13}^{\beta}(u) &= C_{16}^{\beta}u_{1,1} + C_{26}^{\beta}u_{2,2} + C_{66}^{\beta}u_{3,1}.
\end{aligned} \tag{2.7}$$

In $\mathbb{R}^3_-$ the above relations hold true if we exchange β with $-\beta$ according to (2.1).

2.2 Equilibrium problem for the interface crack with nonpenetration conditions

Consider the composite of two elastic orthotropic materials joined along the plane $x_2 = 0$, which was described in Section 2.1. Assume that in each cross-section with $x_3 = const$ the solid occupies a domain $\Omega \subset \mathbb{R}^2$ consisting of two subdomains $\Omega^+ \subset \mathbb{R}^2_+$ and $\Omega^- \subset \mathbb{R}^2_-$ with the interface Σ located on the line $x_2 = 0$. Let Ω be bounded by the Lipschitz boundary $\partial\Omega = \Gamma_N \cup \Gamma_D$ with an outward normal vector $n = (n_1, n_2)^\top$, where $\Gamma_D \neq \emptyset$. We suppose that the crack Γ_C is a part of the interface Σ and define the domain with the crack as $\Omega_C = \Omega \setminus \overline{\Gamma}_C$. Its boundary $\partial\Omega_C$ is the union of Γ_N, Γ_D, and the crack surfaces $\Gamma_C^\pm$.

To prevent mutual interpenetrations between the opposite crack surfaces Γ_C^+ and Γ_C^- we impose a nonnegativity condition on the jump of the displacement normal to the crack (u_2-component); see [12]. Let $g = (g_1, g_2, g_3)^\top$ represent a surface traction given at Γ_N, and, without loss of generality, assume that the volume force is zero. Further, the solid is assumed to be fixed at Γ_D. The problem of equilibrium of the composite with a crack is finally described by the following nonlinear (at Γ_C) relations:

$$\begin{aligned}
-\sigma_{1\alpha,\alpha}(u) = -\sigma_{2\alpha,\alpha}(u) = -\sigma_{3\alpha,\alpha}(u) = 0 \quad & \text{in } \Omega_C,\\
\sigma_{12}(u) = \sigma_{23}(u) = 0 \quad & \text{on } \Gamma_C^\pm,\\
[\![\sigma_{22}(u)]\!] = 0,\ [\![u_2]\!] \geq 0,\ \sigma_{22}(u) \leq 0,\ \sigma_{22}(u)[\![u_2]\!] = 0 \quad & \text{on } \Gamma_C,\\
\left.\begin{aligned}[\![u_1]\!] = [\![u_2]\!] = [\![u_3]\!] = 0,\\ [\![\sigma_{12}(u)]\!] = [\![\sigma_{22}(u)]\!] = [\![\sigma_{23}(u)]\!] = 0\end{aligned}\right. \quad & \text{on } \Sigma\setminus\Gamma_C,\\
\sigma_{1\alpha}(u)n_\alpha = g_1,\ \sigma_{2\alpha}(u)n_\alpha = g_2,\ \sigma_{3\alpha}(u)n_\alpha = g_3 \quad & \text{on } \Gamma_N,\\
u_1 = u_2 = u_3 = 0 \quad & \text{on } \Gamma_D,
\end{aligned} \tag{2.8}$$

where the summation convention over repeated indices $\alpha = 1, 2$ is used. Here $[\![u]\!] = u^+ - u^-$ and $[\![\sigma(u)]\!] = \sigma^\beta(u^+) - \sigma^{-\beta}(u^-)$ denote the jumps across the interface.

We introduce the cone of admissible displacements, which accounts for all the boundary conditions imposed on u in (2.8) as

$$\begin{aligned}
K(\Omega_C) &= \{u \in H(\Omega_C) : \quad [\![u_2]\!] \geq 0 \quad \text{on } \Gamma_C\} \quad \text{with}\\
H(\Omega_C) &= \{u \in H^1(\Omega_C)^3 : \quad u = 0 \quad \text{on } \Gamma_D\}.
\end{aligned}$$

For given $g \in L^2(\Gamma_N)^3$ the potential energy of the composite with a crack is defined by

$$\Pi(u) = \frac{1}{2} \int\limits_{\Omega_C} \sigma_{ij}(u)\varepsilon_{ij}(u)\, dx - \int\limits_{\Gamma_N} g_i u_i \, ds. \tag{2.9}$$

The weak solution $u \in K(\Omega_C)$ to the equilibrium problem (2.8) is defined as the solution to the constrained minimization problem

$$\text{minimize } \Pi(v) \quad \text{over } v \in K(\Omega_C). \tag{2.10}$$

The optimality condition to (2.10) is expressed by the variational inequality

$$\int\limits_{\Omega_C} \sigma_{ij}(u)\varepsilon_{ij}(v-u)\,dx \geq \int\limits_{\Gamma_N} g_i(v-u)_i\,ds \quad \text{for all } v \in K(\Omega_C). \tag{2.11}$$

For unique solvability of (2.10) (or, equivalently (2.11)) uniform positivity of the quadratic term is needed, that is, the existence of an angle β and a constant $c_0(\beta) > 0$ such that

$$\int\limits_{\Omega_C} \sigma_{ij}(u)\varepsilon_{ij}(u)\,dx \geq c_0(\beta)\|u\|^2_{H(\Omega_C)} \quad \text{for every } u \in H(\Omega_C) \tag{2.12}$$

holds. If the 5×5 matrix in (2.7) has the minimal eigenvalue $\lambda_{min}(\beta) > 0$ for some β, in this case, a Korn-type argument implies (2.12).

3 Delamination of the composite via optimization

To describe the delamination between Ω^+ and Ω^- in the model introduced in Section 2, we fix the length l_0 of an initial crack at $t = 0$ and look for its time evolution $l(t) \geq l_0$ with respect to a (loading) parameter $t > 0$. With one crack tip fixed, the length parameter $l(t)$ determines the position of the second crack tip at $t \geq 0$.

In a natural way we arrive at a one-parameter optimization problem. At every time-step t, the global setting consists of the minimization of an a priori given cost functional $T(l)$ (the total potential energy) over all admissible crack lengths $l \geq l_0$. This formulation requires us to solve (2.10) with the crack Γ_C of length l to obtain $T(l)$. Employing the shape (directional) derivative of the cost function at l_0 provides the local optimality condition, which coincides with the Griffith fracture criterion. We combine these two approaches to derive a computationally constructive strategy for optimization.

3.1 Reduced potential energy function and its shape derivative

For $L > 0$ let the crack Γ_C at the interface $\Sigma = \{0 \leq x_1 \leq L, x_2 = 0\}$ be given by the set

$$\Gamma_C = \{0 < x_1 < l, \quad x_2 = 0\}, \quad 0 \leq l \leq L. \tag{3.13}$$

Specifically we assume that the left crack tip $(0,0)^\top$ is fixed on the boundary $\partial\Omega$ and the right tip $(l,0)^\top$ is located at the interface inside Ω. If $l = L$, then the right end of the crack meets $\partial\Omega$.

For the crack (3.13) a reduced potential energy function P depending on the crack-length parameter $l \in [0, L]$ is defined according to (2.9) and (2.10):

$$P(l) := \Pi(u) = \min_{v \in K(\Omega_C)} \Pi(v). \tag{3.14}$$

From (3.14) we deduce that P is a continuous, decreasing, and uniformly bounded function:

$$P \in C([0,L]), \quad 0 \geq P(\bar{l}) \geq P(l) \geq P(L) \quad \text{for } 0 \leq \bar{l} \leq l \leq L. \tag{3.15}$$

Fix $l \in (0, L)$ and let B_0, B_1 be such that $(l,0)^\top \in B_1 \subset B_0 \subset \Omega$. Let $\chi \in W^{1,\infty}(\mathbb{R}^2)$ be an arbitrary cutoff function with support in a neighborhood of the crack tip, such that $\chi = 1$ in B_1 and $\chi = 0$ outside of B_0. For the solution u of (2.10), the shape derivative $P'(l)$ of (3.14) (in direction $(\chi, 0)^\top$) is found to be (see [15]):

$$P'(l) = \int\limits_{\Omega_C} \sigma_{ij}(u) \left(\frac{1}{2} \chi_{,1} \varepsilon_{ij}(u) - E_{ij}(\nabla\chi; u) \right) dx, \tag{3.16}$$

where E denotes a 3×3-symmetric tensor (with $E_{33} = 0$) of the generalized strain

$$\begin{aligned} &E_{11}(\nabla\chi; u) = \chi_{,1} u_{1,1}, \quad E_{22}(\nabla\chi; u) = \chi_{,2} u_{2,1}, \\ &E_{12}(\nabla\chi; u) = 0.5(\chi_{,2} u_{1,1} + \chi_{,1} u_{2,1}), \\ &E_{23}(\nabla\chi; u) = 0.5\chi_{,2} u_{3,1}, \quad E_{13}(\nabla\chi; u) = 0.5\chi_{,1} u_{3,1}. \end{aligned} \tag{3.17}$$

The value of $-P'(l)$ describes the energy release rate in the vicinity of the crack, and it is independent of χ. In fact, let us integrate by parts (3.16) in $\Omega_C \backslash B_1$, using (2.8) and $\chi = 1$ in B_1. For an outward normal vector $b = (b_1, b_2)^\top$ at ∂B_1, an equivalent representation of the shape derivative by the integral over a closed contour ∂B_1 (see [15]) is obtained:

$$P'(l) = \int\limits_{\partial B_1} \sigma_{ij}(u) \left(\frac{1}{2} b_1 \varepsilon_{ij}(u) - E_{ij}(b; u) \right) ds \tag{3.18}$$

with b_1 and b_2 replacing $\chi_{,1}$ and $\chi_{,2}$ in (3.17). For $\beta = 0$ and $u_3 = const$, formula (3.18) coincides with the path-independent Cherepanov-Rice integral, which is well known in fracture mechanics.

From (3.16) and (3.15) we can conclude that

$$P' \in C(0, L), \quad P'(l) \leq 0. \tag{3.19}$$

Let us notice the general fact that for unilaterally constrained crack problems the second derivative P'' is set valued.

3.2 Evolutionary problem of optimization

We assume that the loading depends in a linear way on a parameter $t \geq 0$:

$$g(t) = tg. \tag{3.20}$$

In view of the multiplicative property of the static problem (2.11) it follows that $u(t) = tu$ is a solution of the quasi-static problem: Find $u(t) \in K(\Omega_C)$ such that

$$\int\limits_{\Omega_C} \sigma_{ij}(u(t))\varepsilon_{ij}(v - u(t))\, dx \geq t \int\limits_{\Gamma_N} g_i(v - u(t))_i\, ds \\ \text{for all } v \in K(\Omega_C). \tag{3.21}$$

We arrive at the reduced potential energy function, which is quadratic in t:

$$P(l)(t) = t^2 P(l), \quad P'(l)(t) = t^2 P'(l). \tag{3.22}$$

In addition to the potential energy, let us introduce the surface energy distributed uniformly at the crack faces $\Gamma_C^{\pm}$,

$$S(l) := \left(\int\limits_{\Gamma_C^+} + \int\limits_{\Gamma_C^-} \right) \frac{1}{2}\gamma\, ds = \gamma\, l, \tag{3.23}$$

where $\gamma > 0$ expresses the material parameter of fracture toughness. The total potential energy T is defined as the sum of P from (3.22) and S from (3.23):

$$T(l)(t) := P(l)(t) + S(l) = t^2 P(l) + \gamma l. \tag{3.24}$$

Let an initial crack with $l_0 \in (0, L)$ be fixed at $t = 0$. To determine an actual state $l(t)$ of the crack for $t > 0$ following the principle of virtual work, we have to minimize the total potential energy over all admissible cracks. The standard assumption of brittle fracture does not allow the crack to disappear. In this way, from (3.24) we arrive at an optimization problem at every *time* t subject to a constraint $l \geq l_0$:

$$\text{minimize } \gamma l + t^2 P(l) \quad \text{over } l \in [l_0, L]. \tag{3.25}$$

Due to the linearity of $S(l)$ and (3.15), the function $T(l)$ is bounded and uniformly continuous in $[0, L]$. Hence, there exists a global minimizer

$l(t) \in [l_0, L]$ for (3.25) satisfying

$$\gamma l(t) + t^2 P(l(t)) \le \gamma l + t^2 P(l) \quad \text{for all } l \in [l_0, L], \quad t \ge 0. \tag{3.26}$$

It can be verified (see [5]) that the necessary and sufficient conditions for (3.26) are given by the system:

$$l(0) = l_0, \tag{3.27a}$$

$$l(t) \ge l(s) \quad \text{for } t > s, \tag{3.27b}$$

$$\gamma l(t) + t^2 P(l(t)) \le \gamma l + t^2 P(l) \quad \text{for all } l \ge l^-(t), \tag{3.27c}$$

$$\gamma l(t) + t^2 P(l(t)) \le \gamma l(s) + t^2 P(l(s)) \quad \text{for all } s \le t. \tag{3.27d}$$

Here, we use the notation $l^-(t) = \lim_{s\to t} l(s)$ for $s < t$, and analogously $l^+(t) = \lim_{s\to t} l(s)$ for $s > t$. In fact, the initial condition at $t = 0$ implies (3.27a), the model of brittle fracture requires that $l(t)$ should be an increasing function of t as written in (3.27a), and (3.27b) through (3.27c) follow directly from (3.26). In view of (3.19), the differentiability of P and (3.27c) lead to the necessary optimality condition

$$\gamma + t^2 P'(l(t)) \ge 0. \tag{3.28}$$

It is important to note that (3.15d) holds true in the case where l is continuous as well as in the case of a jump $l^+(t) \ne l^-(t)$. The jump can be characterized by

$$\gamma[l^+(t) - l^-(t)] + t^2[P(l^+(t)) - P(l^-(t))] = 0. \tag{3.29}$$

Alternatively, if $l(t)$ were a uniformly continuous function, then $l^+(t) = l^-(t)$ in (3.29) and the Griffith law of fracture would be satisfied:

$$\begin{gathered} l(0) = l_0, \\ l'(t) \ge 0,\ \gamma + t^2 P'(l(t)) \ge 0,\ l'(t)(\gamma + t^2 P'(l(t))) = 0, \quad t \ge 0. \end{gathered} \tag{3.30}$$

Using nonpositivity of P' we define

$$G(t, l) := t - \sqrt{\gamma/(-P'(l))}, \quad P'(l) \ne 0 \tag{3.31}$$

and get the set of critical points for (3.25):

$$M_t = \begin{cases} L, \\ l_0 \quad \text{if } G(t, l_0) \le 0 \quad \text{or } P'(l_0) = 0, \\ l \quad \text{if } G(t, l) = 0 \quad \text{and } l \ge l(s) \quad \text{for } s \le t. \end{cases} \tag{3.32}$$

Further, (3.25) is equivalent to

$$\begin{aligned} &l(0) = l_0, \\ &\text{minimize } \gamma l + t^2 P(l) \quad \text{over } l \in M_t \quad \text{for } t > 0. \end{aligned} \tag{3.33}$$

The advantage of our formulation (3.33) is related to the fact that it not only uses function values $T(l)$ but also the derivatives $T'(l)$, which gives a more accurate account of the extrema. In the numerical realization we find that for (3.25) a finer discretization with respect to l is necessary to achieve the same accuracy as (3.33).

4 Numerical example

4.1 Data for numerical calculations

We choose the following symmetric geometry for the composite with a crack as presented in Figure 18.1. The domain Ω is chosen to be a square in $\mathbb{R}^2$ with its boundary decomposed as follows:

$$\begin{aligned} &\Gamma_D = \{x_1 = 1, |x_2| \leq 0.5\}, \quad \Gamma_N = \Gamma^+_{N1} \cup \Gamma^-_{N1} \cup \Gamma^+_{N2} \cup \Gamma^-_{N2}, \\ &\Gamma^\pm_{N1} = \{x_1 = 0, 0 \leq \pm x_2 \leq 0.5\}, \quad \Gamma^\pm_{N2} = \{0 < x_1 < 1, x_2 = \pm 0.5\}. \end{aligned}$$

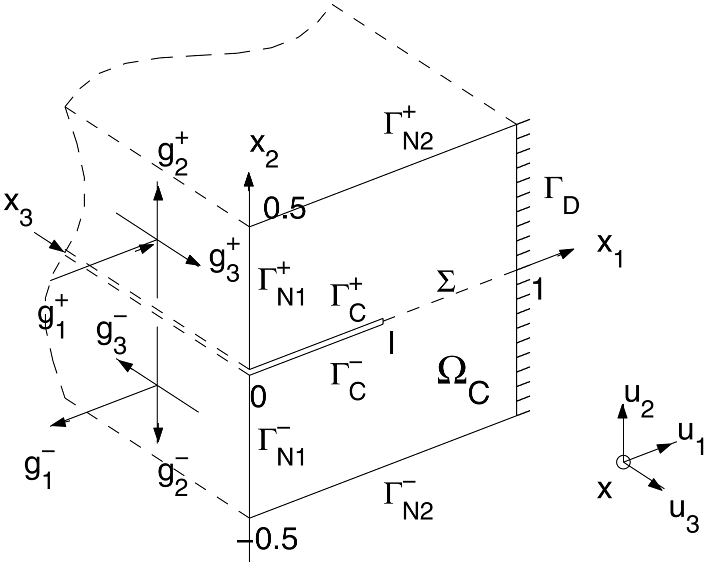


FIGURE 18.1 Geometry of domain Ω_C.

We assume that the crack Γ_C is of length $0 < l < L = 1$. The elastic problem (2.8) in Ω_C is considered with the following boundary conditions imposed on Γ_N:

$$\begin{aligned} \sigma_{12}(u) = \sigma_{22}(u) = \sigma_{23}(u) = 0 \quad \text{on } \Gamma_{N2}^{\pm}, \\ -\sigma_{11}(u) = g_1^{\pm},\ -\sigma_{12}(u) = g_2^{\pm},\ -\sigma_{13}(u) = g_3^{\pm} \quad \text{on } \Gamma_{N1}^{\pm}, \end{aligned} \tag{4.34}$$

where we assume antisymmetric loading corresponding to mode-3:

$$g_3^{\pm} = \mp g_0, \quad g_1^{\pm} = g_2^{\pm} = 0, \quad g_0 = 0.001\mu \approx 3.5376(\text{mPa}), \tag{4.35}$$

as illustrated in Figure 18.1.

We utilize the material parameters with the values from [14]:

$$\begin{aligned} &E_1 = E_2 = E = 10160(\text{mPa}), \quad E_3 = 139400(\text{mPa}), \\ &G_{31} = G_{32} = G_3 = 4600(\text{mPa}), \quad G_{21} = \tfrac{E}{2(1+\nu)} \approx 3537.6(\text{mPa}), \\ &\nu_{21} = \nu = 0.436, \quad \nu_{31} = \nu_{32} = \nu_3 = 0.3. \end{aligned}$$

The corresponding minimal eigenvalues $\lambda_{min}(\beta)$ of the matrix in (2.7) are found to be positive for $\beta \in [-\pi/2, \pi/2]$. They are approximately constant with value $\lambda_{min} \approx 3537.6$. In this case (2.12) holds, and the interface crack problem formulated in Section 2.2 is well posed.

For calculations, the angle β of fibering is taken at the six points $\beta = 0, \pi/16, \pi/8, \pi/4, 3\pi/8, \pi/2$ in $[0, \pi/2]$. This includes the limit cases of the plane isotropic model with $\beta = 0$, and the plane orthotropic model with $\beta = \pi/2$. Note that for $\beta = \pi/4$ the directions of fibering in Ω_C^+ and Ω_C^- are orthogonal to each other.

4.2 The discrete potential energy and its derivative

Following a common procedure in linear elasticity we utilize linear finite elements on a triangular mesh constructed in Ω_C, and we use a local refinement in a neighborhood of Σ. Discretization of (2.10) results in a quadratic programming problem subject to constraints associated with the nonpenetration condition. The numerical implementation of the semismooth Newton method for computing its solution is realized as a primal-dual active-set algorithm, which is described in detail in [8], [9], and [10]. Realizing this algorithm for our example one gets the following numerical results: the appearance of a mixed mode-1 ($[\![u_2]\!] \neq 0$) with mode-3 ($[\![u_3]\!] \neq 0$) crack and contact between opposite crack surfaces under pure mode-3 loading. This situation is related to the 3-dimensional elasticity state and shows the advantage of the spatial model with nonpenetration conditions, in contrast to plane isotropic ($\beta = 0$) and orthotropic ($\beta = \pi/2$) models.

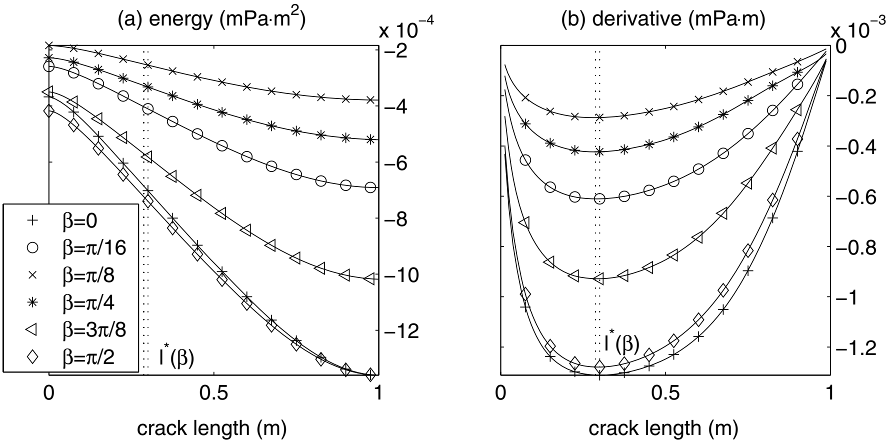


FIGURE 18.2 Potential energy and its shape derivative.

Note that there is no contact between the crack surfaces in the remaining interval $\beta \in (-\pi/2, 0)$. This case was investigated in [17] for the linear setting of the problem with the condition $\sigma_{22}(u) = 0$ describing stress-free crack faces $\Gamma_C^{\pm}$.

The reduced potential energy function P and its shape derivative are computed from (2.9) and (3.16). For numerical calculations the cutoff function χ in formula (3.16) is taken piecewise linear in Ω with $\chi = 1$ around the crack tip, $\chi = 0$ near the external boundary $\partial\Omega$, and symmetrically centered with respect to each crack tip. We approximate the functions P and P' by its discrete values in nodal points $l = 0, h, \dots, 1$, respectively $l = h, 2h, \dots, 1-h$ for $P'(l)$. The results are depicted in Figure 18.2 for various fibering angles $\beta = 0, \pi/16, \pi/8, \pi/4, 3\pi/8, \pi/2$. Here mPa stands for mega Pascal.

We find regions of convexity and concavity of P and minima of $P'(l)$:

$$P'(l^\star) \le P'(l) \quad \text{for all } l, \tag{4.36}$$

which occur for $l^\star \approx 0.3$ if $\beta \in \{0, \pi/2\}$, and for $l^\star \approx 0.2875$ if $\beta \in \{\pi/16, \pi/8, \pi/4, 3\pi/8\}$. They are marked by dotted-lines in Figure 18.2.

4.3 Delamination under mode-3 loading

For numerical tests the physical parameter γ is taken as $\gamma = 25^2/2\mu \approx 0.011$ (mPa·m).

To endow the loading parameter $t \ge 0$ with a physical scale we multiply it by g_0 and consider the linear loading $g_0 t$ (mPa) according to (3.20).

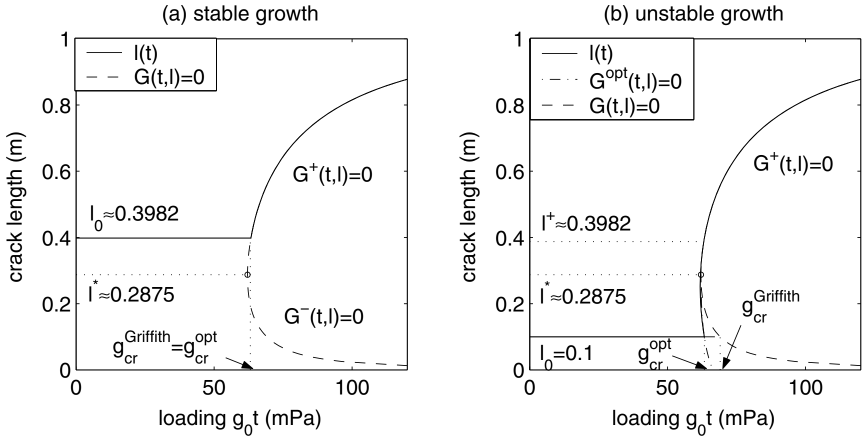


FIGURE 18.3 Quasi-static crack growth as $\beta = \pi/8$.

Because $P'(l)$ is negative $G(t,l)$ in (3.31) is well defined and $g_0 t(l)$ can be obtained from

$$0 = g_0 G(t,l) = g_0 t - g_0 \sqrt{\gamma/(-P'(l))}. \qquad (4.37)$$

For $\beta = \pi/8$ this curve is shown in Figure 18.3 (a) and (b), respectively, by a dashed line. In the remainder of this section we analyze the function T defined in (3.24) with respect to local and global minima using (3.30) and (3.33), respectively. The curve defined in (4.37) contains all critical points of T inside the optimization interval.

We start with the discussion of local minima and fix an arbitrary $l_0 \in (0,1)$. Following the Griffith fracture hypothesis, a critical loading required to start the growth of a crack of length l_0 is determined from (4.37) by

$$g_{\mathrm{cr}}^{\mathrm{Griffith}}(l_0) := g_0 t(l_0) \quad \text{where } G(t(l_0), l_0) = 0. \qquad (4.38)$$

Then the constant function $l(t) = l_0$ is the unique solution to (3.30) as long as $G(t, l_0) < 0$.

Next we seek for the solution $l(t)$ to (3.30) for t such that $G(t, l_0) \geq 0$. For this purpose, points $l^\star \in (0,1)$ of local extrema of $t(l)$ must be found. For our data we obtain one minimizer $l^\star$, which is equivalently characterized by

$$G(t, l^\star) \geq G(t,l) \quad \text{for all } l, \qquad (4.39)$$

independently of t. For $\beta = \pi/8$ we obtain $l^\star \approx 0.2875$, which is marked with a dotted line in Figure 18.3. The line $l = l^\star$ separates $G(t,l) = 0$ into two branches along which $l(t)$ is invertible. These two branches are given by $G^-(t,l) = 0$ for $l \in (0, l^\star)$, and $G^+(t,l) = 0$ for $l \in [l^\star, 1)$. The local solution

$l(t) = l_0$ of (3.30) meets either $G^-(t,l) = 0$ if $l_0 < l^\star$, or $G^+(t,l) = 0$ if $l_0 \geq l^\star$. In the latter case $l(t)$ is an increasing function. Therefore if $l_0 \in [l^\star, 1)$, then $l(t)$ satisfying $G^+(t, l(t)) = 0$ is the unique continuous solution to (3.30) for all t. Alternatively, $l(t)$ obtained from $G^-(t,l) = 0$ is a decreasing function. Hence if $l_0 \in (0, l^\star)$, then there is no solution $l(t)$ to (3.30), which is continuous at the points $t(l_0)$ satisfying $G(t, l_0) = 0$.

To explain the nonexistence of a solution to (3.30) we observe that this relation constitutes a local optimality criterion for (3.25). In our example, this results in the following: The points $l^\star$ found by (4.36) and (4.39) coincide. Thus $P(l)$ (and hence the total energy $\gamma l + t^2 P(l)$) is convex along the branch $G^+(t,l) = 0$ and concave along $G^-(t,l) = 0$. Hence, points $l(t)$ located on $G^+(t,l) = 0$ provide minima of the total potential energy, whereas points on $G^-(t,l) = 0$ give its local maxima.

Now we look for a global minimizer of the optimization problem (3.26) represented in the form (3.33). Solving it numerically we find continuous solutions for initial cracks of the length $l_0 \in [l^\star, 1)$, which coincide with those obtained by the Griffith fracture law (3.30). For $\beta = \pi/8$ and $l_0 \approx 0.3982$ the solution $l(t)$ to (3.32), (3.33) is depicted in Figure 18.3 (a) with a solid line. For initial cracks of the length $l_0 \in (0, l^\star)$, we derive discontinuous solutions with a jump $l^+ - l_0 > 0$ at the point t where the jump condition (3.29) is satisfied, that is,

$$\begin{gathered} g_{\mathrm{cr}}^{\mathrm{opt}}(l_0) := g_0 t \text{ where } t \text{ satisfies} \\ G(t, l^+) = 0, \quad \text{and } \gamma[l^+ - l_0] + t^2[P(l^+) - P(l_0)] = 0. \end{gathered} \tag{4.40}$$

For $\beta = \pi/8$ and $l_0 = 0.1$ the solution $l(t)$ to (3.32) and (3.33) is depicted in Figure 18.3 (b) with a solid line. We find numerically that $l^+ \approx 0.3982$ (this value for l_0 was chosen in the previous example of stable propagation), $g_{\mathrm{cr}}^{\mathrm{opt}} \approx 63$ (mPa) and $g_{\mathrm{cr}}^{\mathrm{Griffith}} \approx 69$ (mPa). Here the value of critical loading obtained by the optimization approach from (4.40) is smaller than the one predicted by the Griffith fracture criterion (4.38).

We obtain an improved curve of critical loading by determining $t(l)$ from the following equation

$$0 = G^{\mathrm{opt}}(t,l) := \begin{cases} G^+(t,l) & \text{for } l \in [l^\star, 1), \\ g_0 t - g_{\mathrm{cr}}^{\mathrm{opt}}(l) & \text{for } l \in (0, l^\star), \end{cases} \tag{4.41}$$

where $g_{\mathrm{cr}}^{\mathrm{opt}}(l)$ is computed according to (4.40) using (3.33) and (3.32) for all discrete length parameters $l \in (0, l^\star)$. For $\beta = \pi/8$, this curve is depicted in Figure 18.3 (b) with a dash-dotted line.

The delamination of the composite with the initial crack of length $l_0 \in (0,1)$ under linear quasi-static loading $g_0 t$ can be constructed geometrically by the following algorithm.

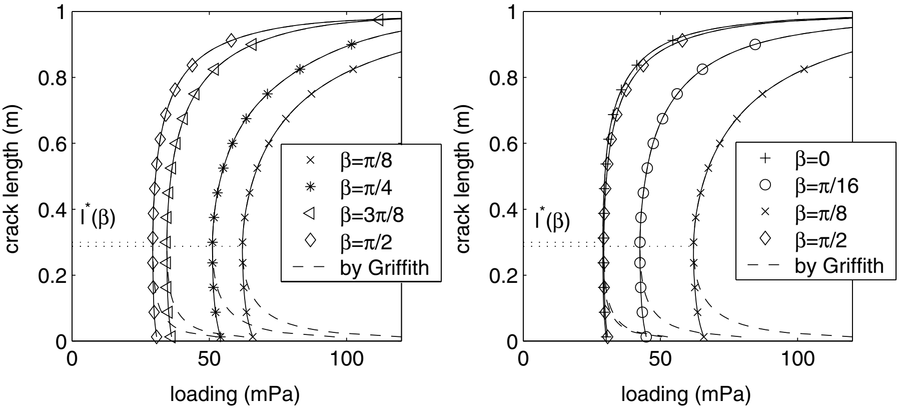


FIGURE 18.4 Curves $G^{\text{opt}}(t,l) = 0$ of critical loading.

Algorithm 1

0. *Fix the initial crack length* $l_0 \in (0,1)$, *and find* $t(l_0)$ *such that*

$$G^{\text{opt}}(t(l_0), l_0) = 0.$$

1. *For all* $t < t(l_0)$ *we have* $l(t) = l_0$ (*no growth*).

2. *At* $t = t(l_0)$ *find* $l(t) = \max\{l_0, l^+\}$, *such that* l^+ *satisfies*

$$G^{\text{opt}}(t(l_0), l^+) = 0$$

(*initiation of crack growth*).

3. *For all* $t > t(l_0)$ *find* $l(t)$ *such that* $G^+(t, l(t)) = 0$ *(crack growth).*

If $l^+ = l_0$ in Step 2 then the propagating crack is stable and it grows in a continuous way. Otherwise, if $l^+ > l_0$ then the crack propagation is unstable with the jump $l^+ - l_0$.

Next we solve $G^{\text{opt}}(t,l) = 0$ for various choices for the fibering angle $\beta = 0, \pi/16, \pi/8, \pi/4, 3\pi/8, \pi/2$. The results are depicted in Figure 18.4, and are compared to the solutions of $G(t,l) = 0$ according to Griffith's law, (4.38) indicated by dashed lines. The points $l^\star(\beta)$ separating the intervals of stable and unstable crack propagation are indicated by dotted lines. For every initial crack of length l_0, the delamination process can be constructed by Algorithm 1.

From Figure 18.4 we can report on the following features:

- The resistance to fracture with respect to the critical mode-3 loading of the composite materials is maximal at $\beta = \pi/8$.
- The curves for the limit cases $\beta = 0$ and $\beta = \pi/2$ are close to each other.

- In the interval $[l^\star, 1)$ the crack growth is stable; otherwise it is unstable.

Figure 18.4 shows clearly that $g_{\rm cr}^{\rm Griffith}(l_0) \to \infty$ as $l_0 \to 0$. To explain this behavior, note that the limit case $l_0 = 0$ corresponds to the initiation of cracking in a continuous solid, which cannot be described exactly by the above macro-crack model. Nevertheless, from Figure 18.4 we may conjecture that $g_{\rm cr}^{\rm opt}(0) < \infty$, which is more consistent physically than $g_{\rm cr}^{\rm Griffith}(0) = \infty$. The other limit behavior $g_{\rm cr}^{\rm opt}(l_0) = g_{\rm cr}^{\rm Griffith}(l_0) \to \infty$ as $l_0 \to 1$ is due to the boundary condition describing a clamped edge at $l = 1$.

Acknowledgment

The research results were obtained with the support of the Austrian Science Fund (FWF) in the framework of the SFB project F003 Optimierung und Kontrolle, and the Russian Foundation for Basic Research (03-01-00124).

References

[1] M. Bach, S.A. Nazarov, and W.L. Wendland, Stable propagation of a mode-1 planar crack in an anisotropic elastic space. Comparison of the Irwin and the Griffith approaches, *Current Problems Anal. Math. Phys.* (Taormina, 1998), 167–189, Aracne, Rome, 2000.

[2] F. Bilteryst and J.-J. Marigo, An energy based analysis of the pull-out problem, *Eur. J. Mech. A/Solids* **22** (2003), 55–69.

[3] E.G. Cherepanov, *Mechanics of Brittle Fracture*, McGraw-Hill, 1979.

[4] G. Dal Maso and R. Toader, A model for the quasistatic growth of brittle fractures: Existence and approximation results, *Arch. Rat. Mech. Anal.* **162** (2002), 101–135.

[5] G.A. Francfort and J.-J. Marigo, Revisiting brittle fracture as an energy minimization problem, *J. Mech. Phys. Solids* **46** (1998), 1319–1342.

[6] A. Friedman and Y. Liu, Propagation of cracks in elastic media, *Arch. Rat. Mech. Anal.* **136** (1996), 235–290.

[7] M. Hintermüller, K. Ito, and K. Kunisch, The primal-dual active set strategy as a semismooth Newton method, *SIAM J. Optim.* **13** (2003), 865–888.

[8] M. Hintermüller, V.A. Kovtunenko, and K. Kunisch, The primal-dual active set method for a crack problem with non-penetration, *IMA J. Appl. Math.* **69** (2004), 1–26.

[9] M. Hintermüller, V.A. Kovtunenko, and K. Kunisch, Semismooth Newton methods for a class of unilaterally constrained variational problems, *Adv. Math. Sci. Appl.* **14** (2004), 513–535.

[10] M. Hintermüller, V.A. Kovtunenko, and K. Kunisch, Generalized Newton methods for crack problems with non-penetration condition, *Numer. Methods Partial Differential Equations*, to appear.

[11] K. Ito and K. Kunisch, Semi-smooth Newton methods for the variational inequalities of the first kind, *ESAIM, Math. Modelling Numer. Anal.* **37** (2003), 41–62.

[12] A.M. Khludnev and V.A. Kovtunenko, *Analysis of Cracks in Solids*, MIT-Press, Cambridge, MA, 2000.

[13] A.M. Khludnev and J. Sokolowski, The Griffith formula and the Cherepanov-Rice integral for crack problems with unilateral conditions in nonsmooth domains, *Euro. J. Appl. Math.* **10** (1999), 379–394.

[14] M. König, R. Krüger, K. Kussmaul, M. von Alberti, and M. Gädke, Characterizing static and fatigue interlaminar fracture behavior of a first generation graphite/epoxy composite, *13th Composite Materials: Testing and Design* 13, ASTM STP 1242, J.S. Hooper (Ed.), ASTM, 1997, 60–81.

[15] V.A. Kovtunenko, Invariant energy integrals for the non-linear crack problem with possible contact of the crack surfaces, *J. Appl. Maths. Mechs.* **67** (2003), 99–110.

[16] V.A. Kovtunenko, Numerical simulation of the non-linear crack problem with non-penetration, *Math. Meth. Appl. Sci.* **27** (2003), 163–179.

[17] V.A. Kovtunenko, Interface cracks in composite orthotropic materials and their delamination via global shape optimization, Preprint.

[18] S.G. Lekhnitskii, *Theory of Elasticity of an Anisotropic Body*, Holden-Day, San Francisco, 1963.

[19] N. Moës, J. Dolbow, and T. Belytschko, A finite element method for crack growth without remeshing, *Int. J. Numer. Math. Engng.* **46** (1999), 131–150.

[20] N.F. Morozov, *Mathematical Foundation of the Crack Theory*, Nauka, Moscow, 1984, in Russian.

[21] R. Byron Pipes and N.J. Pagano, Interlaminar stresses in composite laminates under uniform axial extension, *J. Composite Materials* **4** (1970), 538–548.

[22] J.R. Rice, Elastic fracture mechanics concepts for interfacial cracks, *Trans. ASME. Ser. E. J. Appl. Mech.* **55** (1988), 98–103.

[23] N. Sukumar, N. Moës, B. Moran, and T. Belytschko, Extended finite element method for three-dimensional crack modelling, *Int. J. Numer. Math. Eng.* **48** (2000), 1549–1570.

CHAPTER 19

Adaptive refinement techniques in homogenization design method

Ronald H.W. Hoppe
Department of Mathematics, University of Houston,
Houston, Texas

Svetozara I. Petrova
Institute of Mathematics, University of Augsburg, Augsburg, Germany

Introduction

Naturally grown plants perform like wood allowing the manufacturing of cellular ceramics with unidirectional porous structures. Natural wood morphologies are characterized by an open porous system of tracheidal cells, which provide the transportation path for water and minerals in the living plant (cf., e.g., [13]). The inherent cellular highly open porous system, accessible for infiltration of various liquid or gaseous metals, is used for the design of novel porous ceramics. The transformation of carbonized wood into porous carbide ceramics can be done by infiltration–reaction processes with various carbide-forming metals (e.g., Si, Ti).

In recent years, a great deal of research and investigation has focused on the production of silicon carbide (SiC)–based biomorphic microcellular ceramics. For details of the processing scheme and mechanical properties, we refer the reader, for example, to [8] and [18]. Similar to the developed biotemplating technologies for manufacturing of SiC-ceramics, liquid and gaseous titanium (Ti)-based infiltrants yield titanium carbide (TiC)-, TiTiC-, or single phase TiC-ceramic composites (also called pure TiC-ceramics). The latter ceramic materials are obtained during the reaction in a vacuum or inert atmosphere until the carbon is totally consumed.

Ti-melt infiltration leads to almost dense TiTiC-ceramics, whereas TiC-ceramics are produced by infiltration of a gaseous Ti-based compound, which typically results in a higher porosity, but the processing step requires more time than Ti-melt infiltration. For both SiCand TiC composites it is

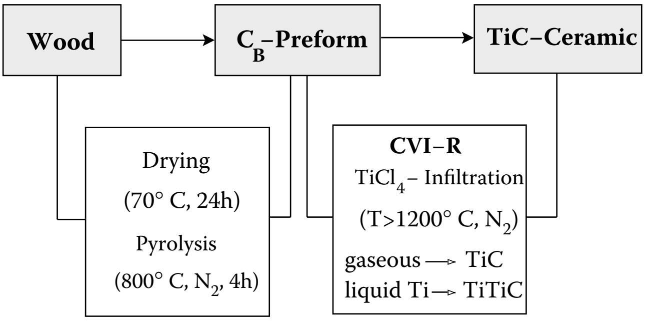


FIGURE 19.1 Processing scheme of biomorphic microcellular TiC-ceramics.

interesting to note that the morphology of the initial wood structure is reproduced on a one-to-one basis.

A novel approach proposed in [15] for synthesis of highly porous, biomorphous TiC-based ceramics by biotemplating of solid wood specimens is the CVI-R processing (chemical vapor infiltration reaction) with titanium tetrachloride ($TiCl_4$). The processing scheme is explained in Figure 19.1. The technique is similar to the technique developed for synthesis of biomorphous SiC described in [8]. The material properties of TiC-based ceramics are often inferior to those of SiC; however, a higher hardness, improved corrosion resistance in phosphoric acid, and especially a high electrical conductivity favorably characterize them.

The remaining part of this paper is organized as follows. In Section 1 we describe the homogenization design method to compute the effective (homogenized) properties of the composite materials. The shape optimization problem in Section 2 for design of the new composite TiC-ceramics is solved by the primal–dual Newton-type interior-point method (cf., e.g., [7], [9], [10]). Section 3 deals with the application of *Zienkiewicz–Zhu* (often referred to as ZZ) nodal averaging or the projection recovery technique proposed in [20] for the stress field. Based on postprocessing, this technique can be used within adaptive mesh-refinement strategies. The reliability of the ZZ a posteriori error estimator in the adaptive procedures is demonstrated by several numerical tests given in the last section.

1 Homogenization design method

Structural optimization has recently become of increasing interest in computer-aided design and optimization of composite structures in materials science (cf., e.g., [1], [2], [16] and the references therein). In particular, shape optimization is applied to structures for which the geometry is a

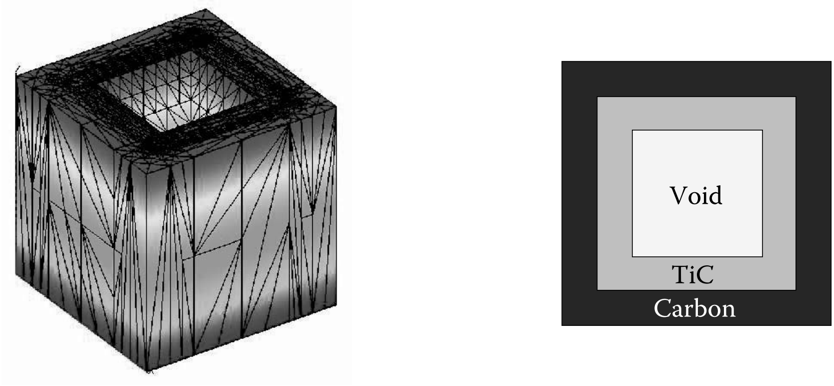


FIGURE 19.2 Left: Periodicity cell $Y = [0,1]^3$; right: Cross-section of $Y = V \cup TiC \cup C$.

design variable, that is, the discretized model associated with the structure has to be changed within the optimization algorithm.

The engineering design objectives usually depend on the mode of loading (compression, bending, shearing, etc.). Our goal is to optimize mechanical performances of the ceramic composites described above (such as the compliance or the bending strength) taking into account technological-and problem-specific constraints on the *state* and *design* parameters. Note that the anisotropic materials under consideration are microstructural, and hence optimal performances can be obtained by tuning microstructural geometric features that strongly influence the macrocharacteristics of the final material workpiece. Our macroscopic scale model is provided by the homogenization approach widely used in structural mechanics (cf., e.g. [3] and [12]).

We consider a 3-dimensional stationary microstructure with a geometrically simple tracheidal periodicity cell $Y = [0,1]^3$ (see Figure 19.2) consisting of an outer layer of carbon (C), an interior layer of TiC, and a void channel $(V$, no material). Assume that the macroscopic material is constructed by introducing an infimum of periodically distributed infinitesimal microstructures consisting of homogeneous materials in terms of carbon and TiC and a microscale void.

Among the main assumptions of homogenization are: i) the composite material is generated by periodic repetition of the unit microcell; and ii) the size of the microstructure is much smaller compared to the size of the entire macrostructure. The periodicity is represented by a small dimensionless real positive parameter $\varepsilon = x/y \ll 1$ for a macroscopic space variable x and a microscopic space variable y. This parameter plays the role of an asymptotic scale factor related to the microscale in which the properties are changing (composite microstructure scale). It allows us to define macrofunctions of the composite in terms of the microstructural behavior and vice versa. Thus, any state function can be defined as

$$f(y) = f\left(\frac{x}{\varepsilon}\right).$$

Spatial derivatives are calculated using the following differentiation rule

$$\frac{d}{dx} f\left(x, \frac{x}{\varepsilon}\right) = \frac{\partial f(x,y)}{\partial x} + \varepsilon^{-1} \frac{\partial f(x,y)}{\partial y}.$$

We consider the case of linearly elastic constituents. Denote by $\boldsymbol{\sigma}(y) = \{\sigma_{ij}(y)\}$ the symmetric stress tensor, by $\boldsymbol{u}(y) \in \mathbf{H}^1(Y)$ the corresponding displacement vector at point y of the body, and by

$$e_{ij}(\boldsymbol{u}(y)) = \frac{1}{2}\left(\frac{\partial u_i(y)}{\partial y_j} + \frac{\partial u_j(y)}{\partial y_i}\right), \qquad i, j = 1, 2, 3, \tag{1.1}$$

the components of the symmetric (small-)strain tensor.

Assume that our composite microstructure is governed by the linearized Hooke's law as the constitutive equation, which can be written in tensor notation as follows

$$\sigma_{ij}(y) = E_{ijkl}(y) e_{kl}(\boldsymbol{u}(y)). \tag{1.2}$$

Here, the Einstein summation convention is applied. The *elasticity tensor* $\boldsymbol{E}(y)$ of order 4 characterizes the material behavior at point y of the microstructure. Note that $\boldsymbol{E}(y)$ is zero if y is located in the hole channel (see Figure 19.2) and coincides with the elasticity tensor of the material (carbon or TiC) if y is located in the corresponding layer. The elasticity tensor is symmetric in the following sense

$$E_{ijkl} = E_{jikl} = E_{ijlk} = E_{klij}, \qquad \forall i, j, k, l = 1, 2, 3, \tag{1.3}$$

and satisfies the following *ellipticity conditions*

$$E_{ijkl}\, \chi_{ij} \chi_{kl} \geq \gamma\, \chi_{ij}^2, \qquad \forall \chi_{ij} = \chi_{ji},$$

for a constant $\gamma > 0$.

Denote by $\boldsymbol{u}_\varepsilon(x) := \boldsymbol{u}(x/\varepsilon)$ the macroscopic displacement vector. Following [3] for the basic concepts of the homogenization method, the unknown displacement vector is expanded asymptotically as

$$\boldsymbol{u}_\varepsilon(x) = \boldsymbol{u}^{(0)}(x,y) + \varepsilon\, \boldsymbol{u}^{(1)}(x,y) + \varepsilon^2\, \boldsymbol{u}^{(2)}(x,y) + \ldots, \quad y = x/\varepsilon, \tag{1.4}$$

where $\boldsymbol{u}^{(0)} \in \boldsymbol{H}^1(Y)$ and $\boldsymbol{u}^{(i)} \in \boldsymbol{H}_{per}(Y),\ i > 0$, is $Y-$periodic function, that is, takes equal values on opposite sides of Y. Here,

$$\boldsymbol{H}_{per}(Y) = \{\boldsymbol{v} \in \boldsymbol{H}^1(Y) | \boldsymbol{v} \text{ is } Y\text{−periodic}\}.$$

The homogenization method requires that we find the functions $\boldsymbol{\xi}^{kl} = (\xi_1^{kl}, \xi_2^{kl}, \xi_3^{kl})$, $k, l = 1, 2, 3$, satisfying the following problem in a weak formulation

$$\int_Y \left(E_{ijpq}(y) \frac{\partial \xi_p^{kl}}{\partial y_q} \right) \frac{\partial \phi_i}{\partial y_j} \, dy = \int_Y E_{ijkl}(y) \frac{\partial \phi_i}{\partial y_j} \, dy \tag{1.5}$$

for an arbitrary variational function $\boldsymbol{\phi} \in \boldsymbol{H}_{per}(Y)$. After computing the microscopic displacement field $\boldsymbol{\xi}^{kl}$, we can define the homogenized (or effective) coefficients by the homogenized formulas (we refer the reader to [3] and [12] for details)

$$E_{ijkl}^H = \frac{1}{|Y|} \int_Y \left(E_{ijkl}(y) - E_{ijpq}(y) \frac{\partial \xi_p^{kl}}{\partial y_q} \right) dy. \tag{1.6}$$

Due to the symmetry conditions—see (1.1) through (1.3)—the 4-th order homogenized elasticity tensor $\boldsymbol{E}^H = (E_{ijkl}^H)$ can be written as a symmetric 6×6 matrix

$$\mathbf{E}^H = \begin{pmatrix} E_{1111}^H & E_{1122}^H & E_{1133}^H & E_{1112}^H & E_{1123}^H & E_{1113}^H \\ & E_{2222}^H & E_{2233}^H & E_{2212}^H & E_{2223}^H & E_{2213}^H \\ & & E_{3333}^H & E_{3312}^H & E_{3323}^H & E_{3313}^H \\ & & & E_{1212}^H & E_{1223}^H & E_{1213}^H \\ & & & & E_{2323}^H & E_{2313}^H \\ \text{SYM} & & & & & E_{1313}^H \end{pmatrix}. \tag{1.7}$$

To the problem (1.5) at microscopic level in the unit cell Y we can associate a homogenized problem at the macroscopic level. Note that both problems constitute the necessary and sufficient conditions obtained by a suitable limit process where the scale parameter ε tends to zero. The latter assumption on ε combined with a double-scale asymptotic expansion of type (1.4) results in a macroscopic homogenized model.

2 Primal–dual Newton interior-point methods

The structural optimization of microcellular biomorphic TiC-ceramics is applied to the homogenized elasticity model. The state variables are the components of the displacement vector $\boldsymbol{u} = (u_1, u_2, u_3)^T$, whereas the design variables $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^T$ are given by the microstructural geometrical details of the periodicity cell (lengths and widths of the layers); see Figure 19.2b) in the case where the number of layers is $m = 2$. The dependence of the homogenized elasticity tensor $\boldsymbol{E}^H = \boldsymbol{E}^H(\boldsymbol{\alpha})$ on the design variables $\boldsymbol{\alpha} \in \mathcal{R}^m$ can be found by means of multivariate interpolation.

Denoting by Ω the domain occupied by the specimen and given an objective functional $J(\boldsymbol{u}, \boldsymbol{\alpha})$ (e.g., the mean compliance of the structure), the design optimization requires the computation of $(\boldsymbol{u}, \boldsymbol{\alpha}) \in \boldsymbol{V} \times \mathcal{R}^m$, $\boldsymbol{V} \subset \boldsymbol{H}^1(\Omega)$, as the solution of

$$J(\boldsymbol{u}, \boldsymbol{\alpha}) \;=\; \inf_{\boldsymbol{v}, \boldsymbol{\beta}} J(\boldsymbol{v}, \boldsymbol{\beta}) \tag{2.8}$$

subject to the equality constraints

$$\int_\Omega E^H_{ijk\ell}(\boldsymbol{\alpha}) \frac{\partial u_k}{\partial x_\ell} \frac{\partial v_i}{\partial x_j} \, dx := \int_\Omega \mathbf{f} \cdot \boldsymbol{v} \, dx \; + \int_{\Gamma_N} \boldsymbol{t} \cdot \boldsymbol{v} \, ds, \tag{2.9}$$

$$c(\boldsymbol{\alpha}) := C \tag{2.10}$$

and the inequality constraints

$$\alpha_i^{(min)} \;\leq\; \alpha_i \;\leq\; \alpha_i^{(max)}, \quad 1 \leq i \leq m, \tag{2.11}$$

where the volume force $\mathbf{f}$ in Ω and the surface force $\boldsymbol{t}$ on $\Gamma_N \subset \partial\Omega$ represent the load conditions. The equality constraint (2.10) as well as the bounds $\alpha_i^{(min)}, \alpha_i^{(max)}$, $1 \leq i \leq m$ in (2.11) stand for constraints motivated by both the microstructural geometry of the carbon preform and the biotemplating process.

Discretization of the objective functional (2.8), the state equation (2.9), and the equality constraint (2.10) by conforming P1 elements with respect to a simplicial tetrahedrization $\mathcal{T}_h$ of Ω leads to the finite dimensional constrained minimization problem: Find $(\boldsymbol{u}_h, \boldsymbol{\alpha}) \in \mathcal{R}^{n_h} \times \mathcal{R}^m$ such that

$$J_h(\boldsymbol{u}_h, \boldsymbol{\alpha}) \;=\; \inf_{\boldsymbol{v}_h, \boldsymbol{\beta}} J_h(\boldsymbol{v}_h, \boldsymbol{\beta}) \tag{2.12}$$

subject to the constraints

$$A_h(\boldsymbol{\alpha})\boldsymbol{u}_h \;=\; \boldsymbol{b}_h, \tag{2.13}$$

$$c_h(\boldsymbol{\alpha}) := C, \tag{2.14}$$

$$\alpha_i^{(min)} \;\leq \alpha_i \leq\; \alpha_i^{(max)}, \quad 1 \leq i \leq m, \tag{2.15}$$

where (2.13) denotes the FE discretized state equation with the stiffness matrix $A_h(\boldsymbol{\alpha}) \in \mathcal{R}^{n_h \times n_h}$ and the load vector $\boldsymbol{b}_h \in \mathcal{R}^{n_h}$.

The standard approach for the solution of (2.12) through (2.15) is to start from a given design, compute the solution of the discretized state equation for that design, update the design by means of the optimality conditions, and iterate this process until convergence to a local minimum. This is called the *alternate directions algorithm*, which has been developed, for instance, in [1] and [6]. The alternate directions algorithm is computationally cheap, if explicit formulas for the homogenized elasticity tensor are available and there are no further constraints on the state and design variables. Otherwise, as far as computational efficiency is concerned, one is better off with what is called an *all–at–once* approach, whose characteristic feature is that the numerical solution of the discretized state equation is an integral part of the optimization routine. Such an approach has been developed in [9] and [11] in terms of a primal–dual Newton interior-point method. The interior-point aspect is to take care of the inequality constraints (2.15) by parameterized logarithmic barrier functions

$$B_h^{(\rho)}(\boldsymbol{u}_h,\boldsymbol{\alpha}) := J_h(\boldsymbol{u}_h,\boldsymbol{\alpha}) - \rho\sum_{i=1}^{m}\Big[\log\ \Big(\alpha_i - \alpha_i^{(min)}\Big) + \log\ \Big(\alpha_i^{(max)} - \alpha_i\Big)\Big],$$

whereas the primal–dual aspect is to couple the remaining equality constraints (2.13) through (2.14) by Lagrangian multipliers $\boldsymbol{\lambda}_h \in \mathcal{R}^{n_h}$ and $\eta_h \in \mathcal{R}$, which gives rise to the saddle point problem

$$\inf_{\boldsymbol{u}_h,\boldsymbol{\alpha}}\ \sup_{\boldsymbol{\lambda}_h,\eta_h}\ L_h^{(\rho)}(\boldsymbol{u}_h,\boldsymbol{\alpha},\boldsymbol{\lambda}_h,\eta_h), \tag{2.16}$$

where the Lagrangian is given by

$$L_h^{(\rho)}(\boldsymbol{u}_h,\boldsymbol{\alpha},\boldsymbol{\lambda}_h,\eta_h) := B_h^{(\rho)}(\boldsymbol{u}_h,\boldsymbol{\alpha})\ +\ \boldsymbol{\lambda}_h^T\ (A_h(\boldsymbol{\alpha})\boldsymbol{u}_h - \boldsymbol{b}_h)\ +\ \eta_h\ (c_h(\boldsymbol{\alpha}) - C).$$

Finally, the Newton aspect is to apply Newton's method to the Karush–Kuhn–Tucker conditions associated with (2.16)

$$\begin{aligned}
\nabla_{\boldsymbol{u}_h} L_h^{(\rho)} &= \nabla_{\boldsymbol{u}_h} J_h + A_h(\boldsymbol{\alpha})^T\boldsymbol{\lambda}_h = 0,\\
\nabla_{\boldsymbol{\alpha}} L_h^{(\rho)} &= \nabla_{\boldsymbol{\alpha}} J_h + \partial_{\boldsymbol{\alpha}}(\boldsymbol{\lambda}_h^T A_h(\boldsymbol{\alpha})\boldsymbol{u}_h) + \eta_h\nabla c_h(\boldsymbol{\alpha}) - \rho D_1^{-1}\mathbf{e} + \rho D_2^{-1}\mathbf{e} = 0,\\
\nabla_{\boldsymbol{\lambda}_h} L_h^{(\rho)} &= A_h(\boldsymbol{\alpha})\boldsymbol{u}_h - \boldsymbol{b}_h = 0,\\
\nabla_{\eta_h} L_h^{(\rho)} &= c_h(\boldsymbol{\alpha}) - C = 0,
\end{aligned}$$

where $D_1 := \mathrm{diag}(\alpha_i - \alpha_i^{(min)}),\ D_2 := \mathrm{diag}(\alpha_i^{(max)} - \alpha_i)$, and $\mathbf{e} := (1,\ldots,1)^T$.

The resulting primal–dual Hessian system is solved by a null-space approach using right-transforming iterations (originally suggested in [19]) with respect to the special block structure of the primal–dual Hessian. An important aspect in the implementation of these iterations is the approximate solution of the discretized state equation (see e.g., [9] for details). After computation of the Newton increments, a line–search approach is performed to update the iterates and to ensure convergence to a local minimum.

3 Adaptive mesh refinement techniques

Advanced finite element applications in science and engineering provoke the extensive use of adaptive mesh refinement techniques to optimize the number of degrees of freedom and obtain accurate enough numerical solutions (cf., e.g., [6] and the references therein). The adaptive framework requires a locally refined discretization in regions where better accuracy is necessary.

The computation of the homogenized elasticity coefficients invokes the numerical solution of (1.5) with the unit cell as the computational domain. Previous works on shape and topology optimization (cf., e.g., [2], [16], [17]) strongly suggest the use of locally refined grids, particularly at material interfaces. In the context of structural optimization, such local refinements have been mostly done before the computations relying on a priori geometric informations, or in an interactive way (manual remeshing based on computational results). In the case of local singularities of the discrete solution, the a priori error estimates typically give information about the asymptotic error behavior, and thus are not the best choice to control the mesh. In those parts of the domain where the solution changes rapidly, an automatic grid refinement on the basis of reliable and robust a posteriori error estimators is highly beneficial. In practice, the main goal in adaptive mesh refinement procedures is to refine the mesh so that the discretization error is within the prescribed tolerance and as much as possible equidistributed throughout the domain.

A natural requirement for the a posteriori error estimates is to be less expensive than the cost of the numerically computed solution. Moreover, appropriate refinement techniques have to be applied to construct the adaptive mesh and implement the adaptive solver. Local reconstruction of the grid is necessary with a computational cost proportional to the number of modified elements.

The solution of our linear elasticity equation (1.5) is computed by using the adaptive finite element method based on the *Zienkiewicz–Zhu* (referred to as ZZ) error estimator. For instance, a recovery technique is analysed in [20] for determining the derivatives (stresses) of the finite element solutions at nodes. The main idea of the recovery technique is to develop smoothing procedures that recover more accurate nodal values of derivatives from the original finite element solution.

The necessity of derivative recovering arises from the fact that in the finite element approach the rate of convergence of the derivatives is usually one order less than that of the discrete solution. In particular, the accuracy of the derivatives (stresses) computed by directly differentiating the discrete solution is inferior. Therefore, in many practical problems an improved accuracy of the stresses at the nodes is needed.

Denote by $\boldsymbol{\sigma}$ the exact stress, by $\hat{\boldsymbol{\sigma}}$ the discrete finite element discontinuous stress, and by $\boldsymbol{\sigma}^*$ the smoothed continuous *recovered stress.* The computation of $\boldsymbol{\sigma}^*$ was proposed and discussed in [20] under the assumption that the same basis functions for interpolation of stresses are used as those for the displacements. The recovered stress $\boldsymbol{\sigma}^*$ is computed by smoothing the discontinuous (over the elements) numerical stress $\hat{\boldsymbol{\sigma}}$. The smoothing procedure can be accomplished by the nodal averaging method or the L_2-projection technique. Note that the components of $\boldsymbol{\sigma}^*$ are piecewise linear and continuous.

The computation of the global L_2-projection is expensive and the authors of [20] proposed to use a lumping form of the mass matrix. Thus, the value of the recovered stress $\boldsymbol{\sigma}^*$ at a node P can be computed by averaging the stresses $\hat{\boldsymbol{\sigma}}$ at the elements that share that node. Denote by $Y_P \subset Y$ the neighborhood patch as a union of all tetrahedra T having node P. Consider

$$\boldsymbol{\sigma}^*(P) = \sum_{T \in Y_P} \omega|_T \, \hat{\boldsymbol{\sigma}}|_T, \quad \omega|_T = \frac{|T|}{|Y_P|}, \ T \in Y_P, \tag{3.17}$$

that is, $\boldsymbol{\sigma}^*(P)$ is a weighted average of $\hat{\boldsymbol{\sigma}}$ with weights $\omega|_T$ defined on the tetrahedra belonging to Y_P. The least-square technique can also be applied to approximate the stress field at a given node.

It was shown in [20] that $\boldsymbol{\sigma}^*$ is a better approximation to $\boldsymbol{\sigma}$ than $\hat{\boldsymbol{\sigma}}$ and the following estimate holds

$$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}^*\|_{0,Y} \ll \|\boldsymbol{\sigma} - \hat{\boldsymbol{\sigma}}\|_{0,Y}, \tag{3.18}$$

where Y is the periodicity microcell under consideration. Furthermore, the recovered technique was used in a formulation of an a posteriori error estimator by comparing the recovered solution $\boldsymbol{\sigma}^*$ with the finite element solution $\hat{\boldsymbol{\sigma}}$. In particular, the estimate (3.18) allows us to replace the exact (unknown) stress $\boldsymbol{\sigma}$ by $\boldsymbol{\sigma}^*$ and consider

$$\|\boldsymbol{\sigma}^* - \hat{\boldsymbol{\sigma}}\|_{0,Y} \tag{3.19}$$

as an error estimator.

In many practical implementations, reliability and efficiency are highly desirable properties in a posteriori error estimation. It basically means that there exist constants independent of the discrete solution and the mesh, which limit the error (in a suitable norm) from below and above. Moreover, technically it is better to use local error estimators, which are computationally less expensive. The following local estimator is considered

$$\eta_T := \|\boldsymbol{\sigma}^* - \hat{\boldsymbol{\sigma}}\|_{0,T}. \tag{3.20}$$

In the practical computations, the nodal values of the recovered stresses are found locally. The elementwise contributions (3.20) are used further as local error indicators in the adaptive mesh refinement procedure.

We consider the global ZZ–error estimator

$$\eta_Y := \left(\sum_{T \in \mathcal{T}_n} \eta_T^2 \right)^{1/2} . \tag{3.21}$$

Based on a posteriori processing, the local estimator (3.20) is practically efficient, providing recovered values are more accurate, that is, the quality of the a posteriori error estimator strongly depends on the approximation properties and the accuracy of the recovered solution.

Arbitrary averaging techniques in low-order finite element applications for elasticity problems are the subject of investigations in [4]. In the latter study the authors considered the following global averaging estimator

$$\eta_A := \min_{\boldsymbol{\sigma}^*} \|\boldsymbol{\sigma}^* - \hat{\boldsymbol{\sigma}}\|_{0,Y} \tag{3.22}$$

and proved an equivalence to the error $\|\boldsymbol{\sigma} - \hat{\boldsymbol{\sigma}}\|_{0,Y}$ with lower and upper bounds independent of the shape-regular mesh. Note that in (3.22) $\boldsymbol{\sigma}^*$ is a smoother approximation to $\hat{\boldsymbol{\sigma}}$ obtained by any averaging procedure. In particular, the final error estimate in [5] explains the reliability and robustness of the ZZ–a posteriori error estimators in practice.

4 Numerical experiments

Some numerical experiments from the computation of the effective material properties of the composite TiC-ceramics are presented in this section. Finite element discretizations in the unit microcell are used to find the homogenized elasticity tensor (1.7). Its components are taken into account further in (2.9) when we solve the shape optimization problems (2.8) through (2.11). Note that for each distribution of materials (carbon and TiC) we have a dependence of the homogenized coefficients on the design parameters $\boldsymbol{\alpha}$, the widths of the material layers.

Denote the global density of the solid material part in the microstructure by μ, $0 < \mu < 1$. If μ is relatively small, we speak about an *early wood* (grown in spring and summer) and respectively, about *late wood* (grown in autumn and winter) for values of μ, close to 1.

We use an initial decomposition of the periodic microcell Y into hexahedra and in addition, a continuous, piecewise linear finite element discretization on tetrahedral shape-regular meshes. The elasticity equation (1.5) is solved numerically 6 times to find the elasticity periodic solutions

TABLE 19.1 Young's modulus E and Poisson's ratio ν of the solid materials.

material	E(GPa)	ν
carbon	10	0.22
titanium carbide	439	0.187

$\boldsymbol{\xi}^{11}$ (Problem 1), $\boldsymbol{\xi}^{22}$ (Problem 2), $\boldsymbol{\xi}^{33}$ (Problem 3), $\boldsymbol{\xi}^{12} = \boldsymbol{\xi}^{21}$ (Problem 4), $\boldsymbol{\xi}^{23} = \boldsymbol{\xi}^{32}$ (Problem 5), and $\boldsymbol{\xi}^{13} = \boldsymbol{\xi}^{31}$ (Problem 6). The material characteristics are given in Table 19.1.

The discrete elasticity problem is solved iteratively by using the preconditioned conjugate gradient (PCG) method with the Incomplete Cholesky (IC) or algebraic multigrid method (AMG) (see, e.g., [14]) as a preconditioner. The mesh adaptivity around the material interfaces has been realized by means of a Zienkiewicz Zhu–type a posteriori error estimator [20] which is used heuristically (as an error indicator). One computes the error (3.20) locally for each element and marks for refinement those tetrahedra $\{T\}$ for which

$$\eta_T \geq \gamma \max_{T' \in \mathcal{T}_n} \eta_{T'},$$

where $0 < \gamma < 1$ is a prescribed threshold, for instance, $\gamma = 0.5$. Note that the adaptive mesh refinement procedure is local and computationally cheap. The refinement process is visualized in Figure 19.3 on the cross-section of the microstructure Y for widths of the C- and TiC- layers $\alpha_1 = \alpha_2 = 0.15$. Additional adaptive refinement is generated in the stiffer material (TiC) and on the interface between the materials due to the different characteristic constants (see Table 19.1).

Table 19.2 contains the values of the computed homogenized coefficients with respect to the adaptive refinement level for a late wood with a density of

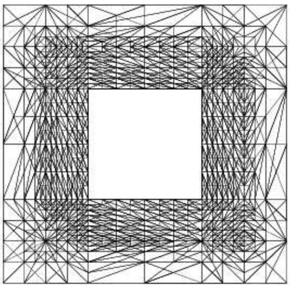


FIGURE 19.3 Cross-section of Y, density 84%, Problem 3, NT = 27527, NN = 5672.

TABLE 19.2 Homogenized coefficients for late wood, density $\mu = 75\%$.

level	E^H_{1111}	E^H_{2222}	E^H_{3333}	E^H_{1212}	E^H_{2323}	E^H_{1313}
1	171.605	178.807	189.733	64.540	68.896	62.800
2	184.221	201.439	206.704	79.012	85.663	74.757
3	173.314	181.355	198.020	64.901	73.942	68.731
4	170.736	165.584	204.236	60.853	60.465	74.665
5	156.046	156.198	203.071	40.893	65.138	69.483
6	136.914	124.878	199.132	31.535	51.781	57.224
7	117.697	116.132	199.674	38.386	46.552	52.533
8	109.720	105.920	197.962	31.669	48.928	51.209
9	100.363	97.797	197.712	27.331	46.428	46.982
10	92.994	94.291	196.055	26.018	44.807	45.497

75%. Convergence results for various values of the density and more details about the discretization parameters on successive adaptive refinement levels are presented in Table 19.3. Here, we have denoted by NT the number of tetrahedra, NN the number of nodes, NDOF the number of degrees of freedom after eliminating the Dirichlet points, ITER the number of iterations by PCG method, and the CPU time in seconds for the iterative solver using the AMG preconditioner. We get an essential efficiency of the multigrid solver compared to the IC preconditioner, both with respect to the number of iterations and CPU time.

TABLE 19.3 Convergence results with AMG preconditioners, density μ, Problem 1.

prec.	level	5	6	7	8	9	10	11
$\mu = 19\%$	NT	1028	1624	2194	3192	3511	7146	13992
	NN	307	434	552	765	837	1609	2880
	NDOF	372	606	858	1317	1515	3336	6558
	ITER	19	30	43	56	54	91	128
	CPU	0.3	0.5	0.9	1.9	2.2	7.7	23.9
$\mu = 84\%$	NT	1246	2445	4632	7125	12856	20665	32700
	NN	359	609	1103	1678	2781	4267	6824
	NDOF	474	960	1980	3345	5973	9492	15999
	ITER	20	25	33	39	69	71	78
	CPU	0.4	1.0	2.6	5.0	13.5	26.1	52.9

5 Acknowledgments

This work was partially supported by the German National Science Foundation (DFG) under Grant No. HO877/5-3 and the Bulgarian Ministry for Education and Science under Grant I1402/2004.

References

[1] G. Allaire, *Shape Optimization by the Homogenization Method*, Springer-Verlag, Berlin-Heidelberg-New York, 2002.

[2] M.P. Bendsøe and O. Sigmund, *Topology Optimization: Theory, Methods and Applications*, Springer-Verlag, Berlin-Heidelberg-New York, 2003.

[3] A. Bensoussan, J.L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, Elsevier Science Publishers, North-Holland, Amsterdam, 1978.

[4] C. Carstensen and S.A. Funken, Averaging technique for FE–a posteriori error control in elasticity. Part I: Conforming FEM, *Comput. Methods Appl. Mech. Eng.*, 190, 2001, 2483–2498.

[5] C. Carstensen and S.A. Funken, Averaging technique for FE–a posteriori error control in elasticity. Part II: $\lambda-$independent estimates, *Comput. Methods Appl. Mech. Eng.*, 190, 2001, 4663–4675.

[6] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, Introduction to adaptive methods for differential equations, *Acta Numerica*, 1995, 105–158.

[7] A. Forsgren, P.E. Gill, and M.H. Wright, *Interior methods for nonlinear optimization*, *SIAM Review*, 44, 2002, 525–597.

[8] P. Greil, T. Lifka, and A. Kaindl, Biomorphic cellular silicon carbide ceramics from wood: I. Processing and microstructure, *J. Europ. Ceramic Soc.*, 18, 1998, 1961–1973.

[9] R.H.W. Hoppe and S.I. Petrova, Applications of primal-dual interior methods in structural optimization, *Comput. Methods Appl. Math.*, 3, 2003, 159–176.

[10] R.H.W. Hoppe and S.I. Petrova, Primal-dual Newton interior point methods in shape and topology optimization, *Numer. Linear Algebra Appl.*, 11, 2004, 413–429.

[11] R.H.W. Hoppe and S.I. Petrova, Optimal shape design in biomimetics based on homogenization and adaptivity, *Math. Comput. Simul.*, 65, 2004, 257–272.

[12] V.V. Jikov, S.M. Kozlov, and O.A. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, Berlin-Heidelberg-New York, 1994.

[13] T. Ota, M. Takahashi, T. Hibi, M. Ozawa, S. Suzuki, Y. Hikichi, and H. Suzuki, Biomimetic process for producing SiC wood, *J. Amer. Ceram. Soc.*, 78, 1995, 3409–3411.

[14] J.W. Ruge and K. Stüben, Algebraic multigrid (AMG), in S.F. McCormick (Ed.), *Multigrid Methods, Frontiers in Applied Mathematics*, SIAM, Philadelphia, 1986, 5.

[15] H. Sieber, C. Zollfrank, N. Popovska, D. Almeida, and H. Gerhard, Gas phase processing of porous biomorphous TiC-ceramics, In *Proc. 8th Conference of the European Ceramic Society*, Istanbul, Turkey, 2003.

[16] J. Sokolowski and J.-P. Zolésio, *Introduction to Shape Optimization*, Springer Series in Computational Mathematics, 16, Springer-Verlag, Berlin-Heidelberg-New York, 1992.

[17] K. Suzuki and N. Kikuchi, A homogenization method for shape and topology optimization, *Comput. Meth. Appl. Mech. Eng.*, 93, 1991, 291–318.

[18] E. Vogli, H. Sieber, and P. Greil, Biomorphic SiC-ceramic prepared by Si-gas phase infiltration of wood, *J. Europ. Ceramic Soc.*, 22, 2002, 2663–2668.

[19] G. Wittum, On the convergence of multigrid methods with transforming smoothers. Theory with applications to the Navier-Stokes equations, *Numer. Math.*, 57, 1989, 15–38.

[20] O.C. Zienkiewicz and J.Z. Zhu, A simple error estimator and adaptive procedure for practical engineering analysis, *Intern. J. Numer. Methods Eng.*, 24, 1987, 337–357.

CHAPTER 20

Nonlinear stability of the flat-surface state in a Faraday experiment

Giovanna Guidoboni
Department of Mathematics, University of Houston,
Houston, Texas

1 Introduction

Faraday first observed the onset of waves on the surface of a horizontal layer of fluid subjected to a vertical oscillation [5]. This phenomenon has been widely studied both theoretically and experimentally due to the plethora of patterns that may be observed depending on the parameters of the problem, such as the frequency of oscillations, the viscosity of the fluid, or the depth of the layer (see, e.g., [4], [12], [19]).

The Faraday waves have several interesting applications. We mention here, among many others, the technology of patterned particulate films [21], and the model of amplification of earthquake waves passing through soft alluvional basins [6].

Here we are not concerned with the pattern selection of Faraday waves, but rather with the conditions under which the upper free surface of the fluid layer remains flat or becomes undulated. More precisely, even if the flat-surface state solves the equations of motion for every value of the parameters involved, it is experimentally observable only if it is stable against perturbations. Linear stability allows for perturbations of infinitesimal amplitude only, and it has been studied by Kumar and Tuckerman [11] in the case of two superimposed fluids.

The goal of this paper is to study the nonlinear stability of the flat-surface state, both in the case of a layer of a single fluid, and the case of two superimposed fluids. The motion of the upper surface of the layer (or of the interface between the two fluids in the case of superimposed layers) is not given a priori, but it is an additional unknown of the problem. This is a free-boundary problem of hyperbolic-parabolic type where the classical energy method [10] fails to give sufficient conditions for the nonlinear exponential stability of the flat-surface state. These conditions are attained by using a generalization of the energy method, which was first introduced by Padula and Solonnikov [16] to investigate the stability of a layer of heavy fluid (both incompressible and compressible) with free capillary surface. Guidoboni and Jin [7] used this method to study the nonlinear stability of the Marangoni-Bénard problem in the Boussinesq approximation with upper free boundary. Other applications of this method can be found in [1], and [13] through [17]. In this paper the generalization of the energy method is applied to the case of a large time-dependent external force.

In Section 2 we describe the mathematical model for a single layer of incompressible viscous fluid, in Section 3 we investigate the stability of the flat-surface state, and in Section 4 we treat the case of two superimposed fluids.

2 Mathematical model

We consider a horizontal layer of viscous incompressible fluid subjected to the gravity force and undergoing a periodic vertical oscillation. In a Cartesian frame of reference, which moves with the layer, the Navier-Stokes equations describing the fluid motion read as follows:

$$\nabla \cdot \mathbf{u} = 0 \tag{2.1}$$

$$\varrho[\partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u}] = -\nabla \tilde{p} + \mu \Delta \mathbf{u} - \varrho g (1 - a f(t)) \mathbf{e}_3 \tag{2.2}$$

where $\mathbf{u}$ and $\tilde{p}$ are the velocity and the pressure of the fluid, ϱ and μ are the density and the dynamic viscosity of the fluid, g is the gravitational acceleration and a is the amplitude of the vertical oscillation in units of g. The function $f(t)$ is periodic in time and takes a different form depending on the features of the oscillation. Here are the cases of one or two frequencies forcing:

$$f(t) = \begin{cases} \cos \omega t \quad (1 \text{ freq.}) \\ \cos(\chi)\cos(M_1 \omega t) + \sin(\chi)\cos(M_2 \omega t + \psi) \quad (2 \text{ freq.}), \end{cases}$$

where ω is a frequency, ψ is the phase shift, M_1 and M_2 are two integers and χ is a real number. At the bottom of the layer there is a rigid horizontal plane laying at $x_3 = -h$, on which we consider a periodicity cell Σ. The upper surface of the layer is a free capillary surface with constant surface tension σ, and it is open to an external ambient, which is exerting a given pressure $\tilde{p}_e$. We assume that the upper surface $\Gamma(t)$ admits the Cartesian representation $x_3 = \eta(x_1, x_2; t)$ and then the domain occupied by the fluid is $\Omega(t) = \Sigma \times (-h, \eta(x_1, x_2; t))$. The function η is an unknown of the problem along with $\mathbf{u}$ and $\tilde{p}$, and this is a free-boundary problem.

At the rigid bottom, $x_3 = -h$, and no-slip boundary conditions are given for the velocity field:

$$\mathbf{u} = \mathbf{0}. \tag{2.3}$$

At the upper free surface, $x_3 = \eta(x_1, x_2; t)$, and two conditions are needed. One is the kinematic condition, which says that the surface is advected by the fluid particles, and one that guarantees the continuity of stresses:

$$\mathbf{u} \cdot \mathbf{n} = \frac{\partial_t \eta}{\sqrt{1 + |\nabla_* \eta|^2}}, \quad \mathbf{T}(\mathbf{u}, \tilde{p})\mathbf{n} = (\sigma \mathcal{H}(\eta) - \tilde{p}_e)\mathbf{n}. \tag{2.4}$$

Here $\nabla_* = (\partial_{x_1}, \partial_{x_2})$ represents the horizontal gradient, $\mathbf{T}(\mathbf{u}, \tilde{p}) = -\tilde{p}\,\mathbf{I} + 2\mu \mathbf{D}(\mathbf{u})$ is the stress tensor, $\mathbf{D}(\mathbf{u}) = (\nabla \mathbf{u} + \nabla^T \mathbf{u})/2$ is the rate-of-strain tensor and $\mathcal{H}(\eta) = \nabla_* \cdot (\nabla_* \eta / \sqrt{1 + |\nabla_* \eta|^2})$ is the double mean curvature. The normal unit vector $\mathbf{n}$ is defined as follows:

$$\mathbf{n}(x_1, x_2; t) = \left(\frac{-\partial_{x_1} \eta}{\sqrt{1 + |\nabla_* \eta|^2}}, \frac{-\partial_{x_2} \eta}{\sqrt{1 + |\nabla_* \eta|^2}}, \frac{1}{\sqrt{1 + |\nabla_* \eta|^2}} \right). \tag{2.5}$$

Problems (2.1) through (2.4) can be rewritten in a simpler form. First of all we introduce a new pressure p defined as $p = \tilde{p} + \varrho g(1 - af(t))x_3$ and the dimensionless numbers:

$$C = \frac{\mu}{\varrho \sqrt{gh^3}}\,, \quad B = \frac{\sigma}{\varrho g h^2}\,. \tag{2.6}$$

C is the square of the inverse of the Galileo number and B is the inverse Bond number. Taking the unit of length, time, velocity and pressure as h, $\sqrt{h/g}$, $\sqrt{gh}$ and ϱgh, respectively, we derive the following dimensionless equations:

$$\nabla \cdot \mathbf{u} = 0 \tag{2.7}$$

$$\partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} = -\nabla p + C \Delta \mathbf{u} = \nabla \cdot \mathbf{T}(\mathbf{u}, p). \tag{2.8}$$

The boundary condition at the rigid bottom $x_3 = -1$ is $\mathbf{u} = \mathbf{0}$, while at the free boundary $x_3 = \eta(x_1, x_2; t)$ we have

$$\mathbf{u} \cdot \mathbf{n} = \frac{\partial_t \eta}{\sqrt{1 + |\nabla_* \eta|^2}}, \quad \mathbf{T}(\mathbf{u}, p)\mathbf{n} = (B\mathcal{H} - p_e)\mathbf{n} - (1 - af(t))\eta \mathbf{n}. \tag{2.9}$$

Here now $\mathbf{u}$, p and η are the dimensionless velocity, pressure and height, while p_e is the dimensionless external pressure.

3 Stability analysis

It is easy to check that the flat-surface state $\mathcal{S}_B = \{\mathbf{u}_B = 0, p_B = p_e, \eta_B = 0\}$ solves problems (2.7) through (2.9) in the domain $\Omega_B = \Sigma \times (-1, 0)$, for every value of B, C and a. The following theorem gives sufficient conditions for the nonlinear exponential stability of $\mathcal{S}_B$.

Theorem 3.1

There exist two positive constants a^* and B^* such that if

$$a \leq a* \quad \textit{and} \quad B \geq B^*$$

then the flat-surface state $\mathcal{S}_B$ is exponentially stable in the class:

$$\mathcal{V} = \left\{ \begin{array}{l} \mathbf{u} \in L^2(0,\infty; W^{1,2}(\Omega(t))) \cap L^\infty(0,\infty; L^3(\Omega(t))), \\ p \in L^\infty(0,\infty; L^2(\Omega(t))) \\ \eta \in L^\infty(0,\infty; W^{1,\infty}(\Sigma)), \\ \displaystyle\int_\Sigma \eta d\Sigma = 0, \quad \exists \, m > 0 \, : \, \sup_{\Sigma,t} |\nabla_* \eta|^2 < m \end{array} \right\}.$$

Proof

We *assume* that there exists another motion

$$\mathcal{S}(\mathbf{x}; t) = \{\mathbf{u}(\mathbf{x}; t), \ p(\mathbf{x}; t) = \pi(\mathbf{x}; t) + p_e, \ \eta(x_1, x_2; t)\} \in \mathcal{V}$$

which solves problems (2.7) through (2.9) in $\Omega(t) = \Sigma \times (-1, \eta(x_1, x_2; t))$. $\mathbf{u}$, π, η are the perturbations to velocity, pressure and height, respectively. We stress that $\mathcal{S}_B$ and $\mathcal{S}$ are defined in different domains and then the perturbations are meant as the difference between $\mathcal{S}$ and the extension of $\mathcal{S}_B$ to the whole space.

Following the classical steps of the energy method (see e.g., [10] and [18]), we write the dimensionless problem for the perturbations:

$$\nabla \cdot \mathbf{u} = 0 \tag{3.10}$$

$$\partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} = -\nabla \pi + C \Delta \mathbf{u} = \nabla \cdot \mathbf{T}(\mathbf{u}, \pi) \tag{3.11}$$

with $\mathbf{u} = \mathbf{0}$ at $x_3 = -1$, and

$$\mathbf{u} \cdot \mathbf{n} = \frac{\partial_t \eta}{\sqrt{1 + |\nabla_* \eta|^2}}, \quad \mathbf{T}(\mathbf{u}, \pi)\mathbf{n} = [B\mathcal{H}(\eta) - (1 - af(t))\eta]\mathbf{n} \tag{3.12}$$

at $x_3 = \eta(x_1, x_2; t)$. We multiply (3.11) by $\mathbf{u}$ and integrate over $\Omega(t)$ to obtain the energy identity:

$$\frac{d}{dt}E(t) = -D(t) + F(t) \tag{3.13}$$

where

$$E(t) = \frac{1}{2}\|\mathbf{u}\|^2_{L^2(\Omega(t))} + \frac{(1-af(t))}{2}\|\eta\|^2_{L^2(\Sigma)} + B\int_\Sigma [\sqrt{1+|\nabla_*\eta|^2} - 1]d\Sigma,$$

$$D(t) = 2C\|\mathbf{D}(\mathbf{u})\|^2_{L^2(\Omega(t))} \quad \text{and} \quad F(t) = -\frac{af'(t)}{2}\|\eta\|^2_{L^2(\Sigma)}.$$

Here $f' = df/dt$.

Remark

If a is small enough and if there exists $m > 0$ such that

$$\sup_{\Sigma,t} |\nabla_*\eta| < m$$

then $E(t)$ is equivalent to a norm of the perturbations. This allows the norm equivalence:

$$c_m\|\nabla_*\eta\|^2_{L^2(\Sigma)} \le \int_\Sigma [\sqrt{1+|\nabla_*\eta|^2} - 1]d\Sigma \le \|\nabla_*\eta\|^2_{L^2(\Sigma)}$$

where $c_m = 1/[\sqrt{1+m^2}+1]$.

The right-hand side in (3.13) does not have a definite sign because of the sinusoidal oscillation $f'(t)$. Due to the hyperbolic–parabolic nature of the problem, a dissipative term for η is missing, while it is present for $\mathbf{u}$, namely $-D(t)$. Therefore, to achieve an exponential stability result we need an auxiliary equation.

Let us write problems (2.7) through (2.9) in the weak form:

$$\int_{\Omega(t)} (\partial_t\mathbf{u} + \mathbf{u}\cdot\nabla\mathbf{u})\cdot\mathbf{\Phi}d\mathbf{x} = -2C\int_{\Omega(t)} \mathbf{D}(\mathbf{u}) : \mathbf{D}(\mathbf{\Phi})\, d\mathbf{x} + \int_{\Omega(t)} \pi\nabla\cdot\mathbf{\Phi}\, d\mathbf{x}$$

$$+\int_\Sigma \mathbf{\Phi}\mathbf{T}\mathbf{n}\, d\Sigma + B\int_{\Gamma(t)} \mathcal{H}(\eta)\mathbf{\Phi}\cdot\mathbf{n}\, d\Gamma - [1-af(t)]\int_{\Gamma(t)} \eta\mathbf{\Phi}\cdot\mathbf{n}\, d\Gamma \tag{3.14}$$

where $\mathbf{\Phi}$ is a vectorial test function. The energy identity (3.13) follows by choosing $\mathbf{\Phi} = \mathbf{u}$ in (3.14). If we choose as a test function the solution of the following problem

$$\nabla \cdot \mathbf{\Phi} = 0 \quad \text{in} \quad \Omega(t) \tag{3.15}$$

$$\mathbf{\Phi} = \mathbf{0} \quad \text{on} \quad \Sigma \tag{3.16}$$

$$\mathbf{\Phi} \cdot \mathbf{n} = \tfrac{\eta}{\sqrt{1+|\nabla_*\eta|^2}} \quad \text{on} \quad \Gamma(t) \tag{3.17}$$

we obtain the following identity:

$$\frac{d}{dt}\int_{\Omega(t)} \mathbf{u}\cdot\mathbf{\Phi} d\mathbf{x} = -2\int_{\Omega(t)} \mathbf{D}(\mathbf{u}) : \mathbf{D}(\mathbf{\Phi})\, d\mathbf{x} + \int_{\Omega(t)} \mathbf{u}\cdot(\partial_t\mathbf{\Phi} + \mathbf{u}\cdot\nabla\mathbf{\Phi})\, d\mathbf{x}$$

$$-B\int_\Sigma \frac{|\nabla_*\eta|^2}{\sqrt{1+|\nabla_*\eta|^2}d\Sigma - [1-af(t)]\int_\Sigma |\eta|^2} d\Sigma, \tag{3.18}$$

which contains the desired terms in the last line. This method has been introduced by Padula and Solonnikov [17] to study the nonlinear stability of the rest state for a layer of heavy fluid (both incompressible and compressible). They prove the existence of a vector field $\mathbf{\Phi} \in L^\infty(0,\infty; H^1(\Omega(t)))$, which solves problems (3.15) through (3.17) and can be taken in the form $\mathbf{\Phi}(\mathbf{x};t) = \nabla \times (\mathbf{A}(x_1,x_2;t)\chi(x_3))$, with $\nabla \times \mathbf{A}\cdot\mathbf{n} = \eta(x_1,x_2;t)$ and $\chi(x_3)$ is a cutoff function. $\mathbf{\Phi}$ satisfies the following inequalities:

$$\|\partial_t\ \mathbf{\Phi}\|_{L^2(\Omega(t))} \le c_1\|\partial_t\eta\|_{L^2(\Sigma)}, \quad \|\mathbf{\Phi}\|_{L^2(\Omega(t))} \le c_2\|\eta\|_{L^2(\Sigma)}, \tag{3.19}$$

$$\|\mathbf{\Phi}\|_{W^{1,2}(\Omega(t))} \le c_2\|\eta\|_{W^{1,2}(\Sigma)}. \tag{3.20}$$

We now multiply (3.18) by a positive constant λ, which will be determined later, and add it to (3.13) to obtain:

$$\frac{d}{dt}\mathcal{E}(t) = -\mathcal{D}(t) + \mathcal{F}(t) \tag{3.21}$$

where

$$\mathcal{E}(t) \ = E(t) + \lambda\int_{\Omega(t)} \mathbf{u}\cdot\mathbf{\Phi}\, d\mathbf{x},$$

$$\mathcal{D}(t) = D(t) + \lambda B\int_\Sigma \frac{|\nabla_*\eta|^2}{\sqrt{1+|\nabla_*\eta|^2}} d\Sigma + \lambda[1-af(t)]\|\eta\|^2_{L^2(\Sigma)},$$

$$\mathcal{F}(t) = F(t) - 2\lambda\int_{\Omega(t)} \mathbf{D}(\mathbf{u}) : \mathbf{D}(\mathbf{\Phi})\, d\mathbf{x} + \lambda\int_{\Omega(t)} \mathbf{u}\cdot\Big(\partial_t\mathbf{\Phi} + \mathbf{u}\cdot\mathbf{\Phi}\Big)\, d\mathbf{x}.$$

The coupling constant λ has to be small enough that: (1) $\mathcal{E}$ is positive definite, and (2) there exists a positive constant c such that $-\mathcal{D}+\mathcal{F} \le -c\mathcal{E}$.

This would lead to the inequality $d\mathcal{E}/dt \leq -c\mathcal{E}$, which infers exponential decay to the functional $\mathcal{E}$.

Let us consider first the positiveness of $\mathcal{E}$. Let us define

$$M_f = \sup_t |f(t)|, \quad M_{f'} = \sup_t |f'(t)|. \tag{3.22}$$

Using the estimates (3.19) and (3.20), we obtain

$$\mathcal{E}(t) \geq \frac{1}{2}\|\mathbf{u}\|^2_{L^2(\Omega_t)} + \frac{1 - aM_f}{2}\|\eta\|^2_{L^2(\Sigma)} - \lambda c_2\|\mathbf{u}\|_{L^2(\Omega_t)}\|\eta\|_{L^2(\Sigma)} \tag{3.23}$$

and then if

$$a < \frac{1 - (\lambda c_2)^2}{M_f} \quad \text{with} \quad \lambda < 1/c_2 \tag{3.24}$$

it follows that $\mathcal{E}(t) > 0 \quad \forall t$.

Now we want to find sufficient conditions for the exponential decay of $\mathcal{E}(t)$. Let us consider the right-hand side of (3.21). Using Hölder's inequality we can write

$$\begin{aligned}-\mathcal{D}(t) + \mathcal{F}(t) \leq &-2C\|\mathbf{D}(\mathbf{u})\|^2_{L^2(\Omega_t)} - \frac{\lambda B}{\sqrt{1+m^2}}\|\nabla_*\eta\|^2_{L^2(\Sigma)}\\ &- \left[\lambda(1 - aM_f) - \frac{aM_{f'}}{2}\right]\|\eta\|^2_{L^2(\Sigma)}\\ &+\lambda b\|\mathbf{D}(\mathbf{u})\|_{L^2(\Omega_t)}\|\nabla_*\eta\|_{L^2(\Sigma)} + \lambda b\|\mathbf{D}(\mathbf{u})\|^2_{L^2(\Omega_t)}\end{aligned}$$

where b is a positive constant that summarizes all the functional constants. Then, using Young's inequality, we have

$$\begin{aligned}-\mathcal{D}(t) + \mathcal{F}(t) \leq &- \left[2C - \frac{3\lambda b}{2}\right]\|\mathbf{D}(\mathbf{u})\|^2_{L^2(\Omega_t)}\\ &- \left[\lambda(1 - aM_f) - \frac{aM_{f'}}{2}\right]\|\eta\|^2_{L^2(\Sigma)}\\ &- \left[\frac{\lambda B}{\sqrt{1+m^2}} - \frac{\lambda b}{2}\right]\|\nabla_*\eta\|^2_{L^2(\Sigma)}.\end{aligned}$$

It follows that if $\lambda < 2C/3b$, $a < \lambda/(2\lambda M_f + M_{f'})$ and $B > b\sqrt{1+m^2}$, then $-\mathcal{D} + \mathcal{F} \leq -\delta_1 G$, where $G = \|D(\mathbf{u})\|^2_{L^2(\Omega(t))} + \|\eta\|^2_{L^2(\Sigma)} + \|\nabla_*\eta\|^2_{L^2(\Sigma)}$ and

$\delta_1 = \min\{C, \lambda/2, \lambda b/2\}$. On the other hand, we can estimate G in terms of $\mathcal{E}$ as follows:

$$\begin{aligned} \mathcal{E}(t) &= E(t) + \lambda \int_{\Omega(t)} \mathbf{u} \cdot \mathbf{\Phi}\, d\mathbf{x} \\ &\geq \left(\frac{1}{2} - \frac{\lambda c_2}{2}\right) \|\mathbf{u}\|^2_{L^2(\Omega(t))} + \left(\frac{1 + aM_f}{2} - \frac{\lambda c_2}{2}\right) \|\eta\|^2_{L^2(\Sigma)} \qquad (3.25) \\ &\quad + c_m \|\nabla_* \eta\|^2_{L^2(\Sigma)} \\ &\geq \delta_2 G \end{aligned}$$

where $\delta_2 = \min\{1 - \lambda c_2, 1 + aM_f - \lambda c_2, 2c_m\}/2$. Then

$$\frac{d\mathcal{E}}{dt} \leq -\mathcal{D} + \mathcal{F} \leq \delta_1 G \leq \frac{\delta_1}{\delta_2}\mathcal{E} = -\gamma \mathcal{E}, \qquad (3.26)$$

where $\gamma = \delta_1/\delta_2$. Therefore if

$$\lambda < \min\left\{\frac{1}{c_2}, \frac{2C}{3b}\right\}, \quad a < \min\left\{\frac{1 - (\lambda c_2)^2}{M_f}, \frac{\lambda}{2\lambda M_f + M_{f'}}\right\} = a^*, \qquad (3.27)$$

$$B > b\sqrt{1 + m^2} = B^*, \qquad (3.28)$$

it follows that the functional $\mathcal{E}$ is positive definite, and moreover there exists $\gamma > 0$ such that $d\mathcal{E}/dt \leq -\gamma \mathcal{E}$. This means that conditions (3.27) and (3.28) are sufficient conditions for the exponential decay of $\mathcal{E}$, and therefore for the exponential stability of the flat-surface state $\mathcal{S}_b$.

We would like to stress here that the key point of the proof is the auxiliary equation (3.18), which provides dissipative terms for the perturbation in the height of the layer.

Conditions (3.27) and (3.28) have a clear physical meaning. They say that the flat-surface state is exponentially stable provided the amplitude of the forcing oscillations is small enough and the surface tension, which is related to B, is large enough. We stress that the bounds a_* and B_* depend on the regularity class of the perturbed motion trough m, and this is a consequence of the method. Moreover, the existence of solutions in $\mathcal{V}$ other than $\mathcal{S}_B$ is assumed and not proved.

4 Case of two superposed fluids

The case of two superposed viscous incompressible fluids can be treated analogously. The fluids are confined between two rigid planes at $x_3 = -h_1$ and $x_3 = h_2$, and we assume that the separating interface admits the Cartesian representation $x_3 = \eta(x_1, x_2; t)$. The motion of each fluid is described by the dimensionless Navier-Stokes equations:

$$\nabla \cdot \mathbf{u}_i = 0 \tag{4.29}$$

$$\partial_t \mathbf{u}_i + \mathbf{u}_i \cdot \nabla \mathbf{u}_i = -\nabla p_i + C_i \Delta \mathbf{u}_i = \nabla \cdot \mathbf{T}(\mathbf{u}_i, p_i) \tag{4.30}$$

where the subscript $i = 1$ or 2 indicates the two fluids. For $i = 1$, the equations hold in $\Omega_1(t) = \Sigma \times (-1, \eta(x_1, x_2; t))$, while for $i = 2$ the domain is $\Omega_2(t) = \Sigma \times (\eta(x_1, x_2; t), \widehat{h})$, where $\widehat{h} = h_2/h_1$. At the rigid planes, no-slip boundary conditions are given for the velocities:

$$\mathbf{u}_i = \mathbf{0} \quad \text{at} \quad x_3 = -1, \widehat{h},$$

while at the interface $x_3 = \eta(x_1, x_2; t)$ we impose the continuity of the velocity and of the stresses:

$$\mathbf{u}_1 = \mathbf{u}_2, \quad [\mathbf{T}(\mathbf{u}_1, p_1) - \mathbf{T}(\mathbf{u}_2, p_2)]\mathbf{n} = (B\mathcal{H} - (1 - af(t))\eta)\mathbf{n}, \tag{4.31}$$

and the kinematic condition

$$\mathbf{u} \cdot \mathbf{n} = \frac{\partial_t \eta}{\sqrt{1 + |\nabla_* \eta|^2}}. \tag{4.32}$$

Here we have taken the difference $\varrho_1 - \varrho_2$ as the unit of density, h_1 as the unit of length, and the normal unit vector $\mathbf{n}$ is referred to as the inferior domain $\Omega_1(t)$.

A simple solution of problems (4.30) through (4.32) is the flat interface state $\mathcal{S}_B$:

$$\mathbf{u}_1 = \mathbf{u}_2 = \mathbf{0}, \quad p_1 = p_2 = const. = p_B, \ \eta_B = 0 \tag{4.33}$$

where $\mathbf{u}_1$, p_1 and $\mathbf{u}_2$, p_2 are defined in $\Omega_1 = \Sigma \times (-1, 0)$ and in $\Omega_2 = \Sigma \times (0, \widehat{h})$, respectively.

Theorem 4.1

There exist two positive constants a^* and B^* such that if

$$a \leq a* \quad \textit{and} \quad B \geq B^*$$

then the flat-interface state $\mathcal{S}_B$ is exponentially stable in the class:

$$\mathcal{V} = \left\{ \begin{array}{l} \mathbf{u}_i \in L^2(0, \infty; W^{1,2}(\Omega_i(t))) \cap L^\infty(0, \infty; L^3(\Omega_i(t))), \\ p_i \in L^\infty(0, \infty; L^2(\Omega_i(t))) \\ \eta \in L^\infty(0, \infty; W^{1,\infty}(\Sigma)), \\ \displaystyle\int_\Sigma \eta d\Sigma = 0, \ \exists \, m > 0 \ : \ \sup_{\Sigma, t} |\nabla_* \eta|^2 < m, \ i = 1, 2 \end{array} \right\}.$$

Proof

We assume that there exists another motion:

$$\mathcal{S}(\mathbf{x};t) = \{\mathbf{u}_i(\mathbf{x};t),\ p_i(\mathbf{x};t) = \pi_i(\mathbf{x};t) + p_B,\ \eta(x_1,x_2;t)\} \in \mathcal{V}$$

which solves problems (4.30) through (4.32) in $\Omega_1(t) = \Sigma \times (-1, \eta(x_1,x_2;t))$ and $\Omega_2(t) = \Sigma \times (\eta(x_1,x_2;t), \widehat{h})$ for $i = 1, 2$. The perturbations solve the following problem:

$$\nabla \cdot \mathbf{u}_i = 0 \tag{4.34}$$

$$\partial_t \mathbf{u}_i + \mathbf{u}_i \cdot \nabla \mathbf{u}_i = -\nabla \pi_i + C_i \Delta \mathbf{u}_i = \nabla \cdot \mathbf{T}(\mathbf{u}_i, \pi_i) \tag{4.35}$$

with $\mathbf{u}_i = \mathbf{0}$ at $x_3 = -1, \widehat{h}$, and

$$\mathbf{u}_1 = \mathbf{u}_2, \quad \mathbf{u}\cdot\mathbf{n} = \frac{\partial_t \eta}{\sqrt{1+|\nabla_* \eta|^2}}, \quad \mathbf{T}(\mathbf{u},\pi)\mathbf{n} = [B\mathcal{H}(\eta) - (1 - af(t))\eta]\mathbf{n} \tag{4.36}$$

at $x_3 = \eta(x_1,x_2;t)$. The energy identity is obtained by multiplying the momentum equation (4.30) for the fluid i by $\mathbf{u}_i$ and integrating over $\Omega_i(t)$ for $i = 1, 2$. Adding the resulting equations, we get:

$$\frac{d}{dt}E(t) = -D(t) + F(t), \tag{4.37}$$

where

$$E(t) = \frac{1}{2}\sum_{i=1}^{2} \|\mathbf{u}_i\|^2_{L^2(\Omega_i(t))} + \frac{(1-af(t))}{2}\|\eta\|^2_{L^2(\Sigma)} + B\int_\Sigma [\sqrt{1+|\nabla_*\eta|^2} - 1]\, d\Sigma,$$

$$D(t) = 2\sum_{i=1}^{2} C_i \|\mathbf{D}(\mathbf{u}_i)\|^2_{L^2(\Omega_i(t))},$$

$$F(t) = -\frac{af'(t)}{2}\|\eta\|^2_{L^2(\Sigma)}.$$

In order to obtain the auxiliary equation we consider now two vector fields, $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$, defined in $\Omega_1(t)$ and $\Omega_2(t)$, respectively, solutions of the following problem:

$$\nabla \cdot \mathbf{\Phi}_i = 0 \quad \text{in} \quad \Omega_i(t), \quad \text{with} \quad \mathbf{\Phi}_i = \mathbf{0} \quad \text{at} \quad x_3 = -1, \widehat{h},$$

$$\text{and with } \mathbf{\Phi}_1 = \mathbf{\Phi}_2,\ \mathbf{\Phi}\cdot\mathbf{n} = \eta/\sqrt{1+|\nabla_*\eta|^2} \quad \text{at} \quad x_3 = \eta(x_1,x_2;t),$$

where $i = 1, 2$. $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$ can be estimated in terms of η as in (3.19) through (3.20). The following auxiliary equation can now be obtained:

$$\frac{d}{dt}\sum_{i=1}^{2}\int_{\Omega_i(t)} \mathbf{u}_i \cdot \mathbf{\Phi}_i \, d\mathbf{x} = -2\sum_{i=1}^{2} C_i \int_{\Omega_i(t)} \mathbf{D}(\mathbf{u}_i) : \mathbf{D}(\mathbf{\Phi}_i)\, d\mathbf{x}$$
$$+\sum_{i=1}^{2}\int_{\Omega_i(t)} \mathbf{u}_i \cdot (\partial_t \mathbf{\Phi}_i + \mathbf{u}_i \cdot \nabla \mathbf{\Phi}_i)\, d\mathbf{x} - B\int_\Sigma \frac{|\nabla_* \eta|^2}{\sqrt{1+|\nabla_*\eta|^2}} d\Sigma$$
$$-[1 - af(t)]\int_\Sigma |\eta|^2 d\Sigma$$

Adding the energy identity and the auxiliary equation to the coupling constant λ we can write

$$\frac{d}{dt}\mathcal{E} = -\mathcal{D} + \mathcal{F}$$

where

$$\mathcal{E}(t) = E(t) + \lambda\sum_{i=1}^{2}\int_{\Omega_i(t)} \mathbf{u}_i \cdot \mathbf{\Phi}_i \, d\mathbf{x},$$
$$\mathcal{D}(t) = D(t) + \lambda B\int_\Sigma \frac{|\nabla_* \eta|^2}{\sqrt{1+|\nabla_*\eta|^2}} d\Sigma + \lambda[1 - af(t)]\int_\Sigma |\eta|^2 d\Sigma,$$
$$\mathcal{F}(t) = F(t) - 2\lambda\sum_{i=1}^{2} C_i \int_{\Omega_i(t)} \mathbf{D}(\mathbf{u}_i) : \mathbf{D}(\mathbf{\Phi}_i)\, d\mathbf{x}$$
$$+\lambda\sum_{i=1}^{2}\int_{\Omega_i(t)} \mathbf{u}_i \cdot (\partial_t \mathbf{\Phi}_i + \mathbf{u}_i \cdot \nabla \mathbf{\Phi}_i)\, d\mathbf{x}.$$

Following the line of the proof in the previous section, we see that if

$$\lambda < \min\left\{\frac{1}{2c_2}, \frac{2C_1}{3b}, \frac{2C_2}{3b}\right\}, \quad B > b\sqrt{1+m^2} = B^*, \tag{4.38}$$

$$a < \min\left\{\frac{1}{2M_f}, \frac{\lambda}{2\lambda M_f + M_{f'}}\right\} = \frac{\lambda}{2\lambda M_f + M_{f'}} = a^* \tag{4.39}$$

then the state with flat interface is exponentially stable in $\mathcal{V}$.

References

[1] R. Benabidallah and M. Padula, Stability of a heavy viscous polytropic fluid in a container bounded by perfectly heat-conducting walls, *Ann. Univ. Ferrara Sez. VII (N.S.)* **45** (2000) 127–161.

[2] B. Christiansen, P. Alstrom, and M. T. Levinsen, Ordered capillary-wave state: Quasicrystals, hexagons, and radial waves, *Phys. Rev. Lett.* **68** (1992) 2157–2160.

[3] B. Dionne, M. Silber, and A. C. Skeldon, Stability results for steady, spatially periodic planforms, *Nonlinearity* **10** (1997) 321–353.

[4] W. S. Edwards and S. Fauve, Patterns and quasi-patterns in the Faraday experiment, *J. Fluid Mech.* **278** (1994) 123–148.

[5] M. Faraday, On the forms and states of fluids on vibrating elastic surfaces, *Phil. Trans. R. Soc. Lond.* **52** (1831) 319–340.

[6] M. A. Foda, A Kelvin-Helmholtz instability mechanism of seismic faulting, *Phys. Rev. E* **54** (1991) 507–513.

[7] G. Guidoboni and B. J. Jin, On the nonlinear stability of Marangoni-Bénard problem with free surface in the Boussinesq approximation, *Math. Mod. and Meth. in Appl. Sci.*, submitted.

[8] G. Guidoboni and M. Padula, On the Bénard problem, *Proceedings of the International Conference on Trends in Partial Differential Equations of Mathematical Physics* (IC TPDE03), Obidos, Portugal (2003).

[9] B. J. Jin and M. Padula, In a horizontal layer with free upper surface, *Communications on Pure and Applied Analysis* **1**, 3 (2002) 379–415.

[10] D. D. Joseph, *Stability of Fluid Motions*, Vols. I and II, Springer Tracts in Natural Philosophy, 27 and 28, Springer-Verlag, Heidelberg, (1976).

[11] K. Kumar and L. S. Tuckerman, Parametric instability of the interface between two fluids, *J. Fluid Mech.* **279** (1994) 49–68.

[12] J. W. Miles and D. Anderson, Parametrically forced surface waves, *Ann. Rev. Fluid Mech.* **22** (1990) 143–165.

[13] M. Padula, On the exponential stability of the rest state of a viscous compressible fluid, *J. Math. Fluid Mech.* **1**, 1 (1999) 62–77.

[14] M. Padula, On direct Ljapunov method in continuum theories, *Nonlinear Problems in Mathematical Physics and Related Topics I*, In honor of Prof. Ladyzhenskaya, eds. Sh. Birman, S. Hildebrandt, V.A. Solonnikov, and O.A. Uralsteva, Kluwer Academic/Plenum, New York, Boston, Dordrecht, London, Moscow, 289–302; Novosibirsk, (2002) 271–283.

[15] M. Padula and M. Pokorny, Stability and decay to zero of the L^2 norms of perturbations to a viscous compressible heat conductive fluid motion exterior to a ball, *J. Math. Fluid Mech.* **3**, 4 (2001) 342–357.

[16] M. Padula and V. A. Solonnikov, On Rayleigh-Taylor stability, *Ann. Univ. Ferrara*—Sez. VII, Vol. XLVI (2000) 307–336.

[17] M. Padula and V. A. Solonnikov, On the global existence of nonsteady motions of a fluid drop and their exponential decay to a uniform rigid rotation, *Quaderni di matematica* **32** (2002) 187–218.

[18] J. Serrin, On the stability of viscous fluid motions, *Arch. Rat. Mech. Anal.* **3** (1959) 1–12.

[19] M. Silber, C. M. Topaz, and A. C. Skeldon, Two-frequency forced faraday waves: Weakly damped modes and pattern selection, *Physica D* **143** (2000) 205–225.

[20] V. A. Solonnikov, Probleme de frontiere libre dans l'ecoulement d'un fluide a la sortie d'un tube cylindrique, *Asymptotic Analysis* **17** (1998) 135–163.

[21] P. H. Wright and J. R. Saylor, Patterning of particulate films using Faraday waves, *Review of Scientific Instruments* **74**, 9 (2003) 4063–4070.

CHAPTER 21

A dynamic programming approach in Hilbert spaces for a family of applied delay optimal control problems

Giorgio Fabbri
Department of Mathematics, Universitá “La Sapienza,”
Rome, Italy

1 Introduction

In this paper we consider a class of one-dimensional optimal control problems of the state equation of delay type, namely, a state equation of the form (we call y the state variable and γ the control):

$$\begin{cases} y(t) = \int_{-T}^{0} \gamma(t+s)\varsigma(s)ds \\ \gamma(s) = \bar{\iota}(s) \ \forall s \in [-T, 0) \\ y(0) = \int_{-T}^{0} \bar{\iota}(s)\varsigma(s)ds \end{cases}$$

where $\varsigma : [-T, 0] \to \mathbb{R}^+$ is a positive BV-function and $\bar{\iota} \in L^2((-T, 0); \mathbb{R}^+)$ (with $y(0)$, which is determined by $\bar{\iota}$) is the initial datum that in the delay setting involve the history of the variables.

We consider a target functional (to be maximized on a set of admissible controls that will be specified below) of the form

$$J(\bar{\iota}, \gamma(\cdot)) \overset{def}{=} \int_0^\infty e^{-\rho t} \frac{(y_{\bar{\iota},\gamma}(t) - \gamma(t))^{1-\sigma}}{(1-\sigma)} ds$$

where ρ is a positive constant and $\sigma > 0$, $\sigma \neq 1$ (we have emphasized the dependency of y on the initial datum $\bar{\iota}$ and on the control γ).

The problem is quite specific but it arises directly from some economic applications: indeed it contains, as particular cases, some models that describe vintage capital, technical innovations, and obsolescence: if we take $\varsigma \equiv 1$ we have the model presented in [2] (see also [3] and [14]) where $y(t)$ is the production at time t; if we take $\varsigma(s) = \Omega e^{\frac{-\alpha}{\mu-1}s} - \eta$ we have[1] the model in [4] where $y(t)$ represents a capital-related variable that considers the costs of innovation and obsolescence. The economic models impose a state or control constraint: $\gamma(t) \in [0, y(t)]$. The set of admissible controls will be defined as in (2.9). In [14] we had already studied (in a deeper way) a more particular case and we now present a more general family of problems. We refer to [14] when the proofs are very similar, particularly in the first section.

Using the techniques developed by Delfour, Vinter and Kwong (see below for the references) we can write such DDE in a suitable infinite dimensional form. The Hilbert space in which the problem is rewritten is $M^2 \overset{def}{=} \mathbb{R} \times L^2((-T,0);\mathbb{R})$. The state $x(t)$ in M^2 (linked with the one-dimensional variables $\gamma(t)$ and $y(t)$ as described in Definition 3.1) satisfies the equation

$$\begin{cases} \frac{d}{dt}x(t) = Gx(t) + B^*\gamma(t), & t > 0 \\ x(0) = p \end{cases} \tag{1.1}$$

where G is the generator of a suitable C_0 semigroup and B a nonbounded linear functional on M^2. p is the initial datum in M^2, which is related to the one-dimensional initial datum as described in Equation (3.15).

Note that, as in the case of infinite dimensional formulation of optimal control problem modeled by PDE with boundary control (see Lasiecka and Triggiani [17]), a nonbounded term appears in the expression of the control in the Hilbert formulation that in the (1.1) is B^*.

We rewrite both the constraints and the target functional in the Hilbert space formulation. We obtain $\gamma(t) \in [0, x^0(t)]$ and

$$J_0(p, \gamma(\cdot)) \overset{def}{=} \int_0^\infty e^{-\rho s} \frac{(x^0_{p,\gamma}(t) - \gamma(t))^{1-\sigma}}{(1-\sigma)} \mathrm{d}s \tag{1.2}$$

The Hamilton-Jacobi-Bellman (HJB) equation is related to the optimal control problem with state equation (1.1) and target functional (1.2) is

$$\rho V(x^0, x^1) - \sup_{\gamma \in [0, x^0]} \Bigg\{ \langle (x^0, x^1), ADV(x^0, x^1) \rangle_{M^2} + \\ + \langle \gamma, BDV(x^0, x^1) \rangle_{\mathbb{R}} + \frac{(x^0 - \gamma)^{1-\sigma}}{(1-\sigma)} \Bigg\} = 0 \tag{1.3}$$

Note that this HJB equation cannot be treated with the existing literature; the main difficulties are the state or control constraints, the problems of domain for the term $BDV(x^0, x^1)$, and the nonanalyticity of the semigroups: we have to give a suitable definition of the solution. We require that: (i) the solution of HJB is defined on a open set Ω of M^2 and C^1 on such set, (ii) on

[1] α, μ, Ω and η are constants that can vary in suitable ranges.

a closed subset Ω_1, where the trajectories of interest for the original delay problem remain, and the solution has differential in $D(A)$ (on $D(A)$ also B makes sense), (iii) the solution satisfies the (HJB) on Ω_1. See Definition 3.4 for the formal definition of the solution.

We find an explicit solution of the HJB in Theorem 3.5. The only other example of an explicit solution of the HJB in infinite dimensions, from what we know, is with quadratic functionals (see below).

In the last section (Section 4) we come back to the original problem, and following the dynamic programming approach, we give an explicit expression for the value function of the problem in the DDE formulation and solve it in closed loop form (Proposition 4.1 and Proposition 4.2).

For the delay equation, an interesting and accurate (and quite recent) reference is the book by Diekmann, van Gils, Verduyn Lunel, and Walther [12].

The idea of a write delay system using a Hilbert setting was first described by Delfour and Mitter [10], [11]. Variants and improvements were proposed by Delfour [5–7], Vinter and Kwong [19], Delfour and Manitius [8], and Ichikawa [16] (see also the precise systematization of the argument in chapter 4 of the book by Bensoussan, Da Prato, Delfour, and Mitter [1]).

The optimal control problem in the (linear) quadratic case is studied in Vinter and Kwong [19], Ichikawa [15], and Delfour, McCalla, and Mitter [9]. In that case, the Riccati equation appears instead of that by Hamilton-Jacobi-Bellman.

2 The delay problem

We consider a one-dimensional optimal control problem in which $y(t)$ is the state variable and $\gamma(t)$ the control (both for $t \geq 0$). The state equation is of the delay type (with a finite memory T for some positive real number T) and so the initial data are represented by the history of the state $y(\cdot)$ and of the control $\gamma(\cdot)$ in the interval $[-T, 0)$. For the particular form of the state equation (see 2.6) $y(t)$ for $t \in [-T, 0)$ is not used. We call $\bar{\iota}$ the initial data given by the history of the control in the interval $[-T, 0)$. We consider only positive controls. We adopt the following notation: $\bar{\iota}\colon [-T, 0) \to \mathbb{R}^+$, which is the initial datum, $\gamma\colon [0, +\infty) \to \mathbb{R}^+$, which is the control, and $\tilde{\gamma}\colon [-T, +\infty) \to \mathbb{R}^+$, which is

$$\tilde{\gamma} = \begin{cases} \bar{\iota}(s) \quad s \in [-T, 0) \\ \gamma(s) \; s \in [0, +\infty). \end{cases} \tag{2.4}$$

Given a function $\tilde{\gamma}\colon [-T, +\infty) \to \mathbb{R}^+$ we define $\tilde{\gamma}_t$ (the history of the control) the function $\colon [-T, 0) \to \mathbb{R}^+$ given by

$$\begin{cases} \tilde{\gamma}_t \colon [-T, 0) \to \mathbb{R}^+ \\ \tilde{\gamma}_t(s) = \tilde{\gamma}(t+s). \end{cases}$$

We consider the one-dimensional optimal control problem in which the derivative of the state y at time t depends on the history $\tilde{\gamma}_t$ of the control $\tilde{\gamma}$ through an integral delay-state equation of the form

$$\begin{cases} y(t) = \int_{-T}^{0} \tilde{\gamma}_t(s)\varsigma(s)ds \\ \tilde{\gamma}(s) = \bar{\iota}(s) \ \forall s \in [-T,0) \\ y(0) = \int_{-T}^{0} \bar{\iota}(s)\varsigma(s)ds \end{cases} \tag{2.5}$$

where $\varsigma : [-T,0] \to \mathbb{R}$ is a positive BV-function. We can rewrite the equation as

$$\begin{cases} \dot{y}(t) = \tilde{\gamma}(t)\varsigma(0) - \tilde{\gamma}(t-T)\varsigma(-T) - \int_{-T}^{0} \tilde{\gamma}_t(s)d\varsigma(s) \\ \tilde{\gamma}(s) = \bar{\iota}(s) \ \forall s \in [-T,0) \\ y(0) = \int_{-T}^{0} \bar{\iota}(s)\varsigma(s)ds \end{cases} \tag{2.6}$$

where $\varsigma(-T)$ and $\varsigma(0)$ are the left and right limits. Given an initial datum $\bar{\iota} \in L^2((-T,0);\mathbb{R}^+)$ and a control $\gamma \in L^2_{loc}([0,+\infty);\mathbb{R}^+)$, we call $y_{\bar{\iota},\gamma}$ the only solution of (2.6) (sometimes we will not specify the dependency of y on $\bar{\iota}$ and γ).

We introduce the following notation, which is useful when we rewrite the problem in the Hilbert space. If we consider the continuous linear application

$$B : C[-T,0] \to \mathbb{R}$$
$$B : \kappa \mapsto \varsigma(0)\kappa(0) - \varsigma(-T)\kappa(-T) - \int_{-T}^{0} \kappa(r)d\varsigma(s)$$

and the application

$$R : L^2((-T,0);\mathbb{R}) \to \mathbb{R}$$
$$R : \bar{\iota} \mapsto \int_{-T}^{0} \bar{\iota}(s)\varsigma(s)ds$$

we can rewrite the state equation as

$$\begin{cases} \dot{y}(t) = B(\tilde{\gamma}_t) \\ \tilde{\gamma}_0(s) = \bar{\iota}(s) \ \forall s \in [-T,0) \\ y(0) = R(\bar{\iota}) \end{cases} \tag{2.7}$$

We call μ_B the measure related to the functional B. We have $y_{\bar{\iota},\gamma}(t) = R(\tilde{\gamma}_t)$.

We introduce into F the application

$$F : L^2((-T,0);\mathbb{R}) \to L^2((-T,0);\mathbb{R})$$
$$F : \gamma \mapsto F(\gamma)$$

where

$$F(\gamma)(s) \overset{def}{=} \int_{-T}^{s} \gamma(-s+r)d\mu_B(r). \tag{2.8}$$

We consider the problem of maximizing the functional

$$J(\bar{\iota}, \gamma(\cdot)) \overset{def}{=} \int_0^\infty e^{-\rho t} \frac{(y_{\bar{\iota},\gamma}(t) - \gamma(t))^{1-\sigma}}{(1-\sigma)} ds$$

in which σ and ρ are strictly positive constants with $\sigma \neq 1$, and $\gamma(\cdot)$ varies on the set of admissible controls given by

$$\mathcal{I}_{\bar{\iota}} \overset{def}{=} \{\gamma(\cdot) \in L^2_{loc}([0,+\infty);\mathbb{R}) : \gamma(t) \in [0, y_{\bar{\iota},\gamma}(t)] \text{ for almost all } t \in \mathbb{R}^+\}. \tag{2.9}$$

We call this optimal control problem *problem **P***. As emphasized in the introduction, this choice is interesting from an economic point of view. The value function of the problem is

$$\begin{cases} V(\bar{\iota}) \overset{def}{=} \sup_{\gamma(\cdot)\in\mathcal{I}_{\bar{\iota}}} J(\bar{\iota},\gamma(\cdot)) & if\ \mathcal{I}_{\bar{\iota}} \neq \emptyset \\ V(\bar{\iota}) = -\infty & if\ \mathcal{I}_{\bar{\iota}} = \emptyset\ and\ \sigma > 1 \\ V(\bar{\iota}) = 0 & if\ \mathcal{I}_{\bar{\iota}} = \emptyset\ and\ \sigma < 1. \end{cases} \tag{2.10}$$

The choice of $\gamma(\cdot)$, which maximizes the growth of the state variable $y(t)$, is $\gamma_M(t) = y(t)$ (it is quite intuitive but a precise proof can be done as in [14]). This choice gives the following DDE:

$$\begin{cases} y(t) = \int_{-T}^{0} y(t+s)\varsigma(s)ds \ \ for\ t \geq T \\ y(s) = h(s)\ \forall s \in [0,T) \end{cases} \tag{2.11}$$

for suitable initial data $h(s) \in L^2((0,T);\mathbb{R}^+)$ (for a more precise description see [14]). The *characteristic equation* of such DDE is (see [12])

$$z = \int_{-T}^{0} e^{rz} d\mu_B(r). \tag{2.12}$$

We use the following assumptions, which as already said in the introduction, are satisfied in a certain number of interesting economic cases:

(H1) Equation (2.12) has at least a positive root. We call ξ the higher positive root (which exists in view of the general theory of linear DDE, see [12]).

(H2) We have

$$\rho > \xi(1-\sigma).$$

(H3) For all $\bar{\iota} \in L^2((-T,0);\mathbb{R}^+)$ with $\bar{\iota} \not\equiv 0$ we have

$$R(\bar{\iota}) - \left(\frac{\rho - \xi(1-\sigma)}{\sigma\xi}\right)\left(\int_{-T}^{0} e^{\xi s} F(\bar{\iota})(s)ds + R(\bar{\iota})\right) > 0.$$

The meaning of hypothesis (H3) will be clearer in the next section.

Proposition 2.1

The solution of (2.11) is continuous on $\mathbb{R}^+$ and for every $\varepsilon > 0$

$$y(t) = o(e^{(\xi+\varepsilon)t}) + \sum_{j=1}^{N} p_j(t)e^{\lambda_j t} \text{ . for } t \to +\infty$$

where λ_j are the finitely many roots of the characteristic equation with a real part exceeding ξ and p_j are $\mathbb{C}$-valued polynomials in t.

Proof

See Diekmann et al. [12], page 34.

This fact justifies assumption (H2): indeed, using the last proposition it easy to see that the value function is always noninfinite when $\sigma > 0$. We now give some results, which can be argued in a way similar to [14]:

Lemma 2.2

Let $\bar{\iota}$ be in $L^2((-T,0);\mathbb{R}^+)$ and $\bar{\iota} \not\equiv 0$; then the solution of (2.11) $y(t)$ remains strictly positive for all $t \geq 0$.

Proposition 2.3

Let $\bar{\iota}$ be in $L^2((-T,0);\mathbb{R}^+)$ and $\bar{\iota} \not\equiv 0$; then $V(\bar{\iota}) < +\infty$ for all $\bar{\iota}$ in $L^2((-T,0);\mathbb{R}^+)$

Proposition 2.4

Let $\bar{\iota}$ be in $L^2((-T,0);\mathbb{R}^+)$ and $\bar{\iota} \not\equiv 0$; then there exists a control $\theta \in \mathcal{I}_{\bar{\iota}}$ such that $J(\bar{\iota},\theta) > -\infty$. So $V(\bar{\iota}) > -\infty$.

Proposition 2.5

An optimal control exists in $\mathcal{I}_{\bar{\iota}}$; that is, we can find in $\mathcal{I}_{\bar{\iota}}$ an admissible strategy $\eta(\cdot)$ such that $V(\bar{\iota}) = J(\bar{\iota},\eta(\cdot))$.

Lemma 2.6

Let $\bar{\iota}$ be in $L^2((-T,0);\mathbb{R}^+)$ and $\bar{\iota} \not\equiv 0$, and let $\eta(\cdot) \in \mathcal{I}_{\bar{\iota}}$ be an optimal strategy; then $k_{\bar{\iota},\eta}(t) > 0$ for all $t \in [0,+\infty)$.

In the next proposition we make some observations on the properties of the value function.

Proposition 2.7

Let V be the value function as defined in (2.10); then for all $\bar{\iota}_1$ and $\bar{\iota}_2$ in $L^2(-T,0;\mathbb{R}^+)$ we have:

i. $\mathcal{I}_{\bar{\iota}_1}$ *is convex*

ii. If $\lambda \in (0,1)$ *then* $\lambda \mathcal{I}_{\bar{\iota}_1} + (1-\lambda)\mathcal{I}_{\bar{\iota}_2} \subseteq \mathcal{I}_{(1-\lambda)\bar{\iota}_2+\lambda\bar{\iota}_1}$

iii. If $\bar{\iota}_1 \geq \bar{\iota}_2$ *a.e. then* $\mathcal{I}_{\bar{\iota}_2} \subseteq \mathcal{I}_{\bar{\iota}_1}$ *and* $V(\bar{\iota}_1) \geq V(\bar{\iota}_2)$

iv. V *is concave*

Proof

All properties follow easily from the linearity of the required constraints and from the concavity of the intertemporal utility considered.

3 The problem in Hilbert space

The infinite dimensional space in which we reformulate the problem is

$$M^2 \overset{def}{=} \mathbb{R} \times L^2((-T,0);\mathbb{R}).$$

The scalar product on M^2 is the natural one on a product of Hilbert spaces, that is,

$$\langle (x^0,x^1),(z^0,z^1)\rangle_{M^2} \overset{def}{=} x^0z^0 + \langle x^1,z^1\rangle_{L^2}$$

for every $(x^0,x^1),(z^0,z^1) \in M^2$.

Now we introduce the operator A on M^2

$$\begin{cases} D(A) \overset{def}{=} \{(\psi^0,\psi^1) \in M^2 : \psi^1 \in W^{1,2}(-T,0;\mathbb{R}),\ \psi^0 = \psi^1(0)\} \\ A : D(A) \to M^2 \\ A(\psi^0,\psi^1) \overset{def}{=} (0, \frac{d}{ds}\psi^1). \end{cases}$$

Abusing this notation it is also possible to confuse $D(A)$ $\psi^1(0)$ with ψ^0 and redefine

$$\begin{cases} B : D(A) \to \mathbb{R} \\ B(\psi(0),\psi) = B\psi \in \mathbb{R} \end{cases}$$

We introduce in this section the link between the initial condition for $y(t)$ and $\tilde{\gamma}_t$ (that is $y(0) = R(\bar{\iota})$), which has a clear meaning in the application but is, so to speak, quite unnatural from a mathematical point of view. We will reintroduce it in Section 4 when we find the optimal feedback for the original problem. Moreover, we also consider negative initial datum $\bar{\iota}$. So we now consider initial data given by $(y_0,\bar{\iota}) \in \mathbb{R} \times L^2((-T,0);\mathbb{R})$ where y_0 and

$\bar{\iota}$ are not related. Our problem becomes a bit more general:

$$\begin{cases} \dot{y}(t) = B(\tilde{\gamma}_t) \\ (y(0), \tilde{\gamma}_0) = (y_0, \bar{\iota}) \end{cases} \tag{3.13}$$

Its solution is

$$y_{y_0,\bar{\iota},\gamma}(t) = y_0 - \int_{-T}^{0} \bar{\iota}(s)\varsigma(s)ds + \int_{-T}^{0} \tilde{\gamma}_t(s)\varsigma(s)ds \tag{3.14}$$

Given initial data $(y_0, \bar{\iota})$, we put

$$p \overset{def}{=} (y_0, F(\bar{\iota})) \in M^2, \tag{3.15}$$

which will be the initial datum for the state equation in M^2.

Definition 3.1
Given $\bar{\iota} \in L^2((-T,0);\mathbb{R})$, $\gamma \in L^2_{loc}([0,+\infty);\mathbb{R})$, $y_0 \in \mathbb{R}$ and $y_{y_0,\bar{\iota},\gamma}(t)$ as in Equation (3.14) we define the structural state *of the system the couple $x_{p,\gamma}(t) = (x^0_{p,\gamma}(t), x^1_{p,\gamma}(t)) \overset{def}{=} (y_{y_0,\bar{\iota},\gamma}(t), F(\tilde{\gamma}_t)) \in M^2$ (where p is defined in Equation (3.15)).*

The structural state, also called the *Vinter-Kwong state*, is useful in a very general setting, for example when $y(t)$ also depends on the history of y and on a measurable $f(\cdot)$ (that is, $y(t) = Ly_t + B\gamma_t + f(t)$). The structural state is always a new couple (z^0, z^1) (obtained by original state and control variables using the so-called *structural operators*), which is a solution of a simpler equation in M^2 (see Delfour [5] or Vinter and Kwong [19] for details). Here we have used the notations of Bensoussan and others ([1], page 234).

Theorem 3.2

Assume that $\bar{\iota} \in L^2((-T,0);\mathbb{R})$, $\gamma \in L^2_{loc}([0,+\infty);\mathbb{R})$, $y_0 \in \mathbb{R}$, $p = (k_0, F(\bar{\iota}))$, then, for each $T > 0$, the structural state $x_{p,\gamma}(\cdot)$ is the unique solution in

$$\Pi \overset{def}{=} \left\{ f \in C([0,+\infty); M^2) \ : \ \frac{d}{\mathrm{d}t} j^* f \in L^2_{loc}([0,+\infty); D(A)') \right\} \tag{3.16}$$

to the equation:

$$\begin{cases} \frac{d}{dt} j^* x(t) = A^* x(t) + B^* \gamma(t), \quad t \geq 0 \\ x(0) = p = (y_0, F(\bar{\iota})) \end{cases} \tag{3.17}$$

where j^*, A^* and B^* are the dual maps of the continuous linear operators

$$j : D(A) \hookrightarrow M^2$$
$$A : D(A) \to M^2$$
$$B : D(A) \to \mathbb{R}$$

where $D(A)$ is equipped with the graph norm.

Proof

This is part of a more general theory. The proof can be found in Bensoussan, Da Prato, Delfour, and Mitter ([1], Theorem 5.1, page 258).

Now we formulate an optimal control problem in infinite dimension, which, thanks to results of the previous section, contains the original problem. First we need the following result.

Theorem 3.3

The equation

$$\begin{cases} \frac{d}{dt} j^* x(t) = A^* x(t) + B^* \gamma(t), & t \geq 0 \\ x(0) = p \end{cases}$$

for $p \in M^2$*,* $\gamma \in L^2_{\text{loc}}([0,+\infty);\mathbb{R})$ *has a unique solution in* Π (*defined in* (*3.16*))*.*

Proof

The proof can be found in Bensoussan, Da Prato, Delfour, and Mitter ([1] Theorem 5.1, page 258).

After this long preamble we can methodically formulate the optimal control problem in infinite dimensions: We consider the state equation in M^2 given by

$$\begin{cases} \frac{d}{dt} j^* x(t) = A^* x(t) + B^* \gamma(t), & t \geq 0 \\ x(0) = p \end{cases}$$

for $p \in M^2$, $\gamma \in L^2_{\text{loc}}([0,+\infty);\mathbb{R})$. Thanks to Theorem 3.3 it has a unique solution $x_{p,\gamma}(t)$ in Π, so $t \mapsto x^0_{p,\gamma}(t)$ is continuous and it makes sense to consider the set of controls

$$\mathcal{I}^0_p \overset{def}{=} \{\gamma \in L^2_{\text{loc}}([0,+\infty);\mathbb{R}) \ : \ \gamma(t) \in [0, x^0_{p,\gamma}(t)] \ \tilde{\forall}\, t \in \mathbb{R}^+\}.$$

The objective functional is

$$J_0(p, \gamma(\cdot)) \overset{def}{=} \int_0^\infty e^{-\rho s} \frac{(x^0_{p,\gamma}(t) - \gamma(t))^{1-\sigma}}{(1-\sigma)} \mathrm{d}s.$$

The value function is then:

$$\begin{cases} V_0(p) \overset{def}{=} \sup_{\gamma(\cdot) \in \mathcal{I}^0_p} J_0(p, \gamma(\cdot)) & \textit{if } \mathcal{I}^0_p \neq \emptyset \\ V_0(p) = -\infty & \textit{if } \mathcal{I}^0_p = \emptyset \textit{ and } \sigma > 1 \\ V_0(p) = 0 & \textit{if } \mathcal{I}^0_p = \emptyset \textit{ and } \sigma < 1. \end{cases}$$

Remark
Let $\bar{\iota}$ be in $L^2((-T,0);\mathbb{R}^+)$ and $\bar{\iota} \not\equiv 0$ and

$$p = (R(\bar{\iota}), F(\bar{\iota})).$$

We find $\mathcal{I}_p^0 = \mathcal{I}_{\bar{\iota}}$, $J_0(p,i(\cdot)) = J(\bar{\iota}, i(\cdot))$ and $V_0(p) = V(\bar{\iota})$ and the solution of the differential equation of Theorem 3.3 is given by Definition 3.1 as seen in Theorem 3.2.

3.1 The HJB *equation*

Thanks to Equation (3.17) we can describe the Hamiltonians of the system. First of all we introduce the *current value* Hamiltonian: it is defined on a subset of $M^2 \times M^2 \times \mathbb{R}$ given by

$$E \overset{def}{=} \{((x^0,x^1),P,\gamma) \in M^2 \times M^2 \times \mathbb{R} \ : x^0 > 0, \ \gamma \in [0,x^0], \ P \in D(A)\}.$$

The current value Hamiltonian is then defined as ($\langle \gamma, BP \rangle_{\mathbb{R}}$ is the product on $\mathbb{R}$):

$$\begin{cases} \mathcal{H}_{CV} : E \to \mathbb{R} \\ \mathcal{H}_{CV}((x^0,x^1),P,\gamma) \overset{def}{=} \langle (x^0,x^1), AP \rangle_{M^2} + \langle \gamma, BP \rangle_{\mathbb{R}} + \frac{(x^0-\gamma)^{1-\sigma}}{(1-\sigma)} \end{cases}$$

in the points in which $x^0 \neq \gamma$ or $\sigma < 1$. When $x^0 = \gamma$ and $\sigma > 1$ we define $\mathcal{H}_{CV} = -\infty$ as (we can now define the Hamiltonian of the system) we name G the subset of $M^2 \times M^2$ given by:

$$G \overset{def}{=} \{((x^0,x^1),P) \in M^2 \times M^2 \ : \ x^0 > 0, \ P \in D(A)\}.$$

The Hamiltonian then becomes:

$$\begin{cases} \mathcal{H} : G \to \overline{\mathbb{R}} \\ \mathcal{H} : ((x^0,x^1),P) \mapsto \sup_{\gamma \in [0,x^0]} \mathcal{H}_{CV}((x^0,x^1),P,\gamma). \end{cases}$$

We can finally introduce the HJB equation of the problem:

$$\rho V(x^0,x^1) - \mathcal{H}((x^0,x^1), DV(x^0,x^1)) = 0$$

$$\text{(HJB)} \quad \rho V(x^0,x^1) - \sup_{\gamma \in [0,ax^0]} \Bigg\{ \langle (x^0,x^1), ADV(x^0,x^1) \rangle_{M^2} + \\ + \langle \gamma, BDV(x^0,x^1) \rangle_{\mathbb{R}} + \frac{(x^0-\gamma)^{1-\sigma}}{(1-\sigma)} \Bigg\} = 0$$

Now we give the solution of (HJB). We want to describe a kind of regular solution; nevertheless a working definition must consider the domain problems of the definition of the Hamiltonian:

Remark

As we have already noted, this HJB equation cannot be treated with the results of the existing literature. This is due to the presence of the state or control constraint (i.e., the investments that are possible at time t depend on y at time t: $\gamma(t) \in [0, y(t)]$), to the unboundedness of the control operator (i.e., the term $(BDV(x^0, x^1))$ and the nonanalyticity of the semigroup generated by the operator A^*. To overcome these difficulties we have to give a suitable solution. We require the following facts:

(i) The solution of HJB is defined on an open set Ω of M^2 and C^1 on such set,

(ii) On a closed subset Ω_1, where the trajectories that are interesting for the original problem remain, the solution is differential in $D(A)$ (on $D(A)$ the Dirac B also makes sense),

(iii) The solution satisfies on Ω_1 the (HJB).

Definition 3.4

Let Ω be an open set of M^2 and $\Omega_1 \subseteq \Omega$ a closed subset. An application $g \in C^1(\Omega; \mathbb{R})$ is a solution of the HJB on Ω_1 if $\forall (p^0, p^1) \in \Omega_1$

$$\begin{cases} ((p^0, p^1), (Dg(p^0, p^1))) \in G \\ \rho g(p^0, p^1) - \mathcal{H}((p^0, p^1), Dg(p^0, p^1)) = 0. \end{cases}$$

Remark

If $P \in D(A)$ and $(BP)^{-1/\sigma} \in (0, x^0]$, by elementary arguments, the function

$$\mathcal{H}_{CV}(x, P, \cdot) : [0, x^0] \to \mathbb{R}$$

admits exactly a maximum at the point

$$\gamma^{MAX} = x^0 - (BP)^{-1/\sigma} \ \in [0, x^0).$$

Then we can write the Hamiltonian in a simplified form:

$$\mathcal{H}((x^0, x^1), P) = \langle (x^0, x^1), AP \rangle_{M^2} + x^0 BP + \frac{\sigma}{1-\sigma}(BP)^{\frac{\sigma-1}{\sigma}}. \tag{3.18}$$

The expression for γ^{MAX} will be used to write the solution of the original problem in closed-loop form.

We define

$$X \overset{def}{=} \left\{ (x^0, x^1) \in M^2 \ : \ x^0 > 0, \ \left(x^0 + \int_{-T}^{0} e^{\xi s} x^1(s) \mathrm{d}s \right) > 0 \right\}$$

and $\left(\text{named } \alpha = \frac{\rho - \xi(1-\sigma)}{\sigma\xi}\right)$

$$Y \overset{def}{=} \left\{ (x^0, x^1) \in X \ : \ \int_{-T}^{0} e^{\xi s} x^1(s) \mathrm{d}s \leq x^0 \frac{1-\alpha}{\alpha} \right\}. \tag{3.19}$$

It is easy to see that X is an open set of M^2 and Y a closed subset of X.

Proposition 3.5

Under the assumptions $(H1)$ *and* $(H2)$

$$v : X \to \mathbb{R}$$

$$v(x^0, x^1) \overset{def}{=} \nu \left(\int_{-T}^{0} e^{\xi s} x^1(s) \mathrm{d}s + x^0 \right)^{1-\sigma} \tag{3.20}$$

with

$$\nu = \left(\frac{\rho - \xi(1-\sigma)}{\sigma\xi} \right)^{-\sigma} \frac{1}{(1-\sigma)\xi}$$

is differentiable in all $(x^0, x^1) \in X$ *and is the solution of the HJB equation on* Y *in the sense of Definition 3.4.*

Proof

It is useful to introduce

$$\Gamma \overset{def}{=} \left(\int_{-T}^{0} e^{\xi s} x^1(s) \mathrm{d}s + x^0 \right)$$

v is of course continuous and differentiable in every point of X and its differential in (x^0, x^1) is

$$Dv(x^0, x^1) = (\nu(1-\sigma)\Gamma^{-\sigma}, \{s \mapsto (1-\sigma)\nu\Gamma^{-\sigma} e^{\xi s}\}).$$

So $Dv(x^0, x^1) \in D(A)$ everywhere in X.

We can also explicitly calculate ADv and BDv. We have (using that ξ satisfies the characteristic Equation (2.12) and then $B(\{s \mapsto e^{\xi s}\}) = \xi$):

$$ADv(x^0, x^1) = (0, \{s \mapsto (1-\sigma)\nu\Gamma^{-\sigma}\xi e^{\xi s}\}) \tag{3.21}$$

$$BDv(x^0, x^1) = (1-\sigma)\nu\Gamma^{-\sigma}\xi > 0 \tag{3.22}$$

so

$$(BDv)^{-1/\sigma} = \left(\frac{\rho - \xi(1-\sigma)}{\sigma\xi} \right) \left(\int_{-T}^{0} e^{\xi s} x^1(s) \mathrm{d}s + x^0 \right). \tag{3.23}$$

For the definition of X $(BDv)^{-1/\sigma} > 0$.

If $(x^0, x^1) \in Y$ then

$$\left(\int_{-T}^{0} e^{\xi s} x^1(s) \mathrm{d}s + x^0 \right) \leq \frac{1}{\alpha} x^0 \tag{3.24}$$

and then $(BDv)^{-1/\sigma} \leq x^0$. So we can use Remark 3.1 and use the Hamiltonian in the form of Equation (3.18).

Now it is sufficient to substitute (3.21) and (3.22) in (3.18) and verify, by easy calculations, the relation:

$$\rho v(x^0, x^1) - \langle (x^0, x^1), ADv(x^0, x^1) \rangle_{M^2} -$$
$$-x^0 BDv((x^0, x^1) - \frac{\sigma}{1-\sigma} (BDv((x^0, x^1))^{\frac{\sigma-1}{\sigma}} = 0.$$

3.2 Closed loop in infinite dimensions

In this subsection we will prove a closed-loop result for the points related to the original problem. We begin with some definitions:

Definition 3.6
Given $p \in M^2$ we call $\phi \in C(M^2)$ an admissible feedback strategy related to p *of the equation.*

$$\begin{cases} \frac{d}{\mathrm{d}t} j^* x(t) = A^* x(t) + B^*(\phi(x(t))), & t > 0 \\ x(0) = p \end{cases}$$

has a unique solution $x_\phi(t)$ in Π and $\phi(x_\phi(\cdot)) \in \mathcal{I}_p^0$. We indicate the set of admissible feedback strategies related to p with AFS_p.

Definition 3.7
Given $p \in M^2$ we call ϕ an optimal feedback strategy related to p *if it is in AFS_p and*

$$V_0(p) = \int_0^{+\infty} e^{-\rho t} \frac{\left(x_\phi^0(t) - \phi(x_\phi(t))\right)^{1-\sigma}}{(1-\sigma)} \mathrm{d}t.$$

We indicate the set of optimal feedback strategies related to p with OFS_p.

We have a solution of the HJB equation only in a part of the state space (that is Y). So we can prove a feedback result (and the optimality of the feedback) only if the admissible trajectories remain in Y. Here we will use the condition (H3) on the constants that characterize the problem.

In the following theorem we consider the point related to the points of the original problem, which are the points of the form $p = (R(\bar{\iota}), F(\bar{\iota})$ for some $\bar{\iota} \in L^2((-T, 0); \mathbb{R}^+)$

Theorem 3.8

Given $\bar\iota \in L^2((-T,0);\mathbb{R}^+)$ with $\bar\iota \not\equiv 0$, if we call $p = (R(\bar\iota), F(\bar\iota))$, the application

$$\phi\colon M^2 \to \mathbb{R}$$

$$\phi(x) \overset{def}{=} x^0 - \left(\frac{\rho - \xi(1-\sigma)}{\sigma\xi}\right)\left(\int_{-T}^{0} e^{\xi s} x^1(s)\mathrm{d}s + x^0\right) \tag{3.25}$$

is in OFS_p.

Proof

First of all, we have to observe that $\phi \in AFS_p$. We claim that

$$\begin{cases} \frac{d}{\mathrm{d}t} j^* x_\phi(t) = A^* x_\phi(t) + B^*(\phi(x_\phi(t))), \quad t > 0 \\ x_\phi(0) = p = (R(\bar\iota), F(\bar\iota)) \end{cases} \tag{3.26}$$

has a unique solution in Π:

We consider i the solution of the following DDE

$$\begin{cases} i(t) = \left(1 - \frac{\rho-\xi(1-\sigma)}{\sigma\xi}\right)\left(\int_{(-T)}^{0} i(s+t)\varsigma(s)\mathrm{d}s\right) - \\ \qquad - \frac{\rho-\xi(1-\sigma)}{\sigma\xi}\int_{-T}^{0} e^{\xi s} F(i_t)(s)\mathrm{d}s \\ i(s) = \bar\iota \quad \forall\ s \in [-T, 0), \end{cases} \tag{3.27}$$

which has an absolute continuous solution i on $[0, +\infty)$ (see for example [1], page 287, for a proof).

We now claim that $i(t) > 0$ for all $t \geq 0$: The solution is continuous and its value in 0 is strictly positive in view of (H3). Assume that at a point $\bar t$ we have $i(\bar t) = 0$. Then

$$0 = i(\bar t) = \left(1 - \frac{\rho - \xi(1-\sigma)}{\sigma\xi}\right)\left(\int_{(-T)}^{0} i(s+\bar t)\varsigma(s)\mathrm{d}s\right) - \\ - \frac{\rho - \xi(1-\sigma)}{\sigma\xi}\int_{-T}^{0} e^{\xi s} F(i_{\bar t})(s)\mathrm{d}s > 0 \tag{3.28}$$

where the last inequality follows from the fact that i is strictly positive in the interval $[0, \bar t)$ and from assumption (H3).

Then we consider the equation

$$\begin{cases} \frac{d}{\mathrm{d}t} j^* x = A^* x + B^*(i(t)), \quad t > 0 \\ x = p = (R(\bar\iota), F(\bar\iota)). \end{cases} \tag{3.29}$$

We know, thanks to Theorem 3.2, that the only solution in Π of this equation is $x(t) = (y(t), F(i_t))$ where $y(t)$ is the solution of

$$\begin{cases} \dot{y}(t) = B(i_t) \\ (y(0), i_0) = (R(\bar{\iota}),\ \bar{\iota}) \end{cases} \Big(\textit{that is} \quad y(t) = \textstyle\int_{t-T}^{t} i(s)\varsigma(s-t)\mathrm{d}s\Big). \tag{3.30}$$

We claim that $x(t)$ is the solution of (3.26); indeed,

$$\phi(x(t)) = y(t) - \left(\frac{\rho - \xi(1-\sigma)}{\sigma\xi}\right)\left(\int_{-T}^{0} e^{\xi s} F(i_t)(s)\mathrm{d}s + y(t)\right) \tag{3.31}$$

and so (by (3.27):

$$\phi(x(t)) = y(t) - \left(\int_{t-T}^{t} i(s)\varsigma(s-t)\mathrm{d}s - i(t)\right)$$

and by (3.30) we conclude that

$$\phi(x(t)) = i(t),$$

and so (3.26) is reduced to (3.29) and $x(t) = x_\phi(t)$ is a solution of (3.26) and is in Π. Moreover, thanks to the linearity of ϕ we can observe that $x_\phi(t)$ is the only solution in Π. We now observe that $i(\cdot) = \phi(x_\phi(\cdot)) \in \mathcal{I}_p^0$: we have already seen that $i(t)$ is always positive and the other inequality follows from assumption (H3).

We see now that $\phi \in OFS_p$. We consider v as defined in Proposition 3.5. From what we have just said on the admissibility of $i(t)$, it follows that $x_\phi(\cdot)$ remains in Y as defined in (3.19) and so the Hamiltonian as in the proof of Proposition 3.5 can be expressed in the simplified form of Equation (3.18) and v is a solution of HJB on the points of the trajectory.

We introduce:

$$\begin{aligned} &\tilde{v}(t,x) : \mathbb{R} \times X \to \mathbb{R} \\ &\tilde{v}(t,x) \overset{def}{=} e^{-\rho t} v(x) \quad (v \text{ is defined in (3.20)}). \end{aligned} \tag{3.32}$$

Using the fact that $(Dv(x_\phi(t))) \in D(A)$ and that the function $x \mapsto Dv(x)$ is continuous with respect to the norm of $D(A)$ (see the proof of Proposition 3.5 for the explicit form of $Dv(x)$), we find:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\tilde{v}(t, x_{\phi(t)}) &= -\rho\tilde{v}(t, x_\phi(t)) + \langle D_x\tilde{v}(t, x_\phi(t)) | A^* x_\phi(t) + B^* i(t)\rangle_{D(A)\times D(A)'} \\ &= -\rho e^{-\rho t} v(x_\phi(t)) + e^{-\rho t}(\langle ADv(x_\phi(t)), x_\phi(t)\rangle_{M^2} \\ &\quad + BDv(x_\phi(t)) i(t)). \end{aligned} \tag{3.33}$$

By definition (remember that $i(\cdot) = \phi(x_\phi)(\cdot)$):

$$v(p) - J_0(p, i(\cdot)) = v(x_\phi(0)) - \int_0^\infty e^{-\rho t} \frac{(x_\phi^0(t) - \phi(x_\phi)(t))^{1-\sigma}}{(1-\sigma)} \mathrm{d}t =$$

Then, using (3.33) (we use Proposition 2.3 to guarantee that the integral is finite and that the boundary term at ∞ vanishes), we obtain

$$= \int_0^\infty e^{-\rho t} (\rho v(x_\phi(t)) - \langle ADv(x_\phi(t)), x_\phi(t)\rangle_{M^2} - \langle BDv(x_\phi(t)), i(t)\rangle_{\mathbb{R}}) \mathrm{d}t -$$

$$- \int_0^\infty e^{-\rho t} \left(\frac{(x_\phi^0(t) - i(t))^{1-\sigma}}{(1-\sigma)} \right) \mathrm{d}t =$$

$$= \int_0^\infty e^{-\rho t} \Bigg(\rho v(x_\phi(t)) - \langle ADv(x_\phi(t)), x_\phi(t)\rangle_{M^2} -$$

$$- \langle BDv(x_\phi(t)), i(t)\rangle_{\mathbb{R}} - \frac{(x_\phi^0(t) - i(t))^{1-\sigma}}{(1-\sigma)} \Bigg) \mathrm{d}t =$$

using Theorem 3.5

$$= \int_0^\infty e^{-\rho t} (\mathcal{H}(x_\phi(t), Dv(x_\phi(t))) - \mathcal{H}_{CV}(x_\phi(t), Dv(x_\phi(t)), i(t))) \mathrm{d}t. \quad (3.34)$$

The conclusion is followed by three observations:

1. Noting that $\mathcal{H}(x_\phi(t), Dv(x_\phi(t))) \geq \mathcal{H}_{CV}(x_\phi(t), Dv(x_\phi(t)), i(t))$, Equation (3.34) implies that for every admissible control $\gamma(\cdot)$, $v(p) - J_0(p, \gamma(\cdot)) \geq 0$ and then $v(p) \geq V_0(p)$.
2. The original maximization problem is equivalent to the problem of finding a control $\gamma(\cdot)$ that minimizes $v(p) - J_0(p, \gamma(\cdot))$
3. The feedback strategy ϕ achieves $v(p) - J_0(p, i(\cdot)) = 0$, which the minimum in view of point 1. Moreover, this implies that $v(p) \geq V_0(p)$.

As a corollary of the proof (in particular from the very last observation) we have the following:

Remark
Given $p = (R(\bar{\iota}), F(\bar{\iota}))$ for some $\bar{\iota} \in L^2((-T, 0); \mathbb{R}^+)$ we have that

$$V(\bar{\iota}) = V_0(p) = v(p)$$

that is, on this point we have an explicit expression for the value function V given using v.

4 Coming back to the delay problem

We can use the result we found in the infinite dimensional setting to give some results for the original optimal control problem regulated by the delay differential equation: *problem* $\boldsymbol{P}$.

From Remark we can say the following.

Proposition 4.1

Under hypotheses (H1), (H2), and (H3), given initial data $(y(0), \tilde{\gamma}_0) = (R(\bar{\iota}), \bar{\iota})$ *in Equation (2.6) (where* $\bar{\iota}$ *is in* $L^2((-T,0);\mathbb{R}^+)$ *and* $\bar{\iota} \not\equiv 0$*), the value function* V *related to* problem $\mathbf{P}$ *is*

$$V(\bar{\iota}) = \nu\bigg(\int_{-T}^{0} e^{\xi s} F(\bar{\iota})(s)\mathrm{d}s + R(\bar{\iota})\bigg)^{1-\sigma}$$

where

$$\nu = \bigg(\frac{\rho - \xi(1-\sigma)}{\sigma\xi}\bigg)^{-\sigma} \frac{1}{(1-\sigma)\xi}.$$

Moreover, from Proposition 3.8 we can give a solution in closed form for *problem* $\boldsymbol{P}$:

Proposition 4.2

Under hypotheses (H1), (H2) and (H3), given initial data $(y(0), \tilde{\gamma}_0) = (R(\bar{\iota}), \bar{\iota})$ *in Equation (2.6) (where* $\bar{\iota}$ *is in* $L^2((-T,0);\mathbb{R}^+)$ *and* $\bar{\iota} \not\equiv 0$*), the optimal control for* problem $\mathbf{P}$ $\gamma(\cdot)$ *and the related state trajectory* $y_{\bar{\iota},\gamma}(\cdot)$ *satisfy for all* $t \geq 0$:

$$\gamma(t) = y_{\bar{\iota},\gamma}(t) - \bigg(\frac{\rho - \xi(1-\sigma)}{\sigma\xi}\bigg)\bigg(y_{\bar{\iota},\gamma}(t) + \int_{-T}^{0} e^{\xi s} F(\tilde{\gamma}_t)(s)\mathrm{d}s\bigg). \tag{4.35}$$

Then we can find a DDE whose only solution is the optimal control for *problem* $\boldsymbol{P}$:

Corollary 4.3

Under hypotheses (H1), (H2) and (H3), given initial data $(y(0), \tilde{\gamma}_0) = (R(\bar{\iota}), \bar{\iota})$ *in Equation (2.6) (where* $\bar{\iota}$ *is in* $L^2((-T,0);\mathbb{R}^+)$ *and* $\bar{\iota} \not\equiv 0$*), the optimal control for* problem $\mathbf{P}$ $\gamma(\cdot)$ *is the only absolutely continuous solution on* $[0,+\infty)$ *of the DDE*

$$\begin{cases} \tilde{\gamma}(t) = R(\tilde{\gamma}_t) - \Big(\frac{\rho-\xi(1-\sigma)}{\sigma\xi}\Big)\Big(R(\tilde{\gamma}_t) + \int_{-T}^{0} e^{\xi s} F(\tilde{\gamma}_t)(s)\mathrm{d}s\Big) \\ \tilde{\gamma}(s) = \bar{\iota}(s) \ \forall s \in [-T,0). \end{cases}$$

Eventually we can find a constant in the optimal control problem, along its optimal trajectories.

Lemma 4.4

Under hypotheses (H1), (H2) and (H3), given initial data $(y(0), \tilde{\gamma}_0) = (R(\bar{\iota}), \bar{\iota})$ *in Equation (2.6) (where* $\bar{\iota}$ *is in* $L^2((-T,0);\mathbb{R}^+)$ *and* $\bar{\iota} \not\equiv 0$*), there exists a* Λ *such that along the optimal trajectory the optimal control for* problem **P** $\gamma(\cdot)$ *and the related state trajectory* $y_{\bar{\iota},\gamma}(\cdot)$ *satisfy for all* $t \geq 0$*:*

$$y_{\bar{\iota},\gamma}(t) - \gamma(t) = \Lambda e^{gt} \tag{4.36}$$

where Λ *is a real constant and* $g = \frac{\xi-\rho}{\sigma}$*.*

Proof

In view of Proposition 4.2 along the optimal trajectory we have:

$$y_{\bar{\iota},\gamma}(t) - \gamma(t) = \left(\frac{\rho - \xi(1-\sigma)}{\sigma\xi}\right)\left(\int_{-T}^{0} e^{\xi s} F(\tilde{\gamma}_t)(s)\mathrm{d}s + y_{\bar{\iota},\gamma}(t)\right).$$

We note that

$$\int_{-T}^{0} e^{\xi s} F(\tilde{\gamma}_t)(s)\mathrm{d}s + y_{\bar{\iota},\gamma}(t) = \langle \psi, x(t) \rangle$$

where $\psi \overset{def}{=} (1, s \mapsto e^{\xi s}) \in M^2$ and $x(t)$ is as in Definition 3.1. We now calculate the derivative of this expression. It is easy to see that $\psi \in D(A)$. So we have (by Theorem 3.2)

$$\frac{d}{dt}\left(\int_{-T}^{0} e^{\xi s} F(\tilde{\gamma}_t)(s)\mathrm{d}s + y_{\bar{\iota},\gamma}(t)\right) = \frac{d}{dt}\langle \psi, x(t) \rangle_{M^2} =$$

(by Equation (3.17), the definitions of A and B and the fact that ξ is a solution of Equation (2.12))

$$= \langle A\psi, x(t) \rangle_{M^2} + \langle B(\psi), \gamma(t) \rangle_{\mathbb{R}} = \langle (0, s \mapsto \xi e^{\xi s}), x(t) \rangle_{M^2} + \langle \xi, \gamma(t) \rangle_{\mathbb{R}} =$$

using the explicit $x(t)$, the scalar products and using the Equation (4.35)

$$= \left[\xi\left(\int_{-T}^{0} e^{\xi s} F(\tilde{\gamma}_t)(s)\mathrm{d}s\right)\right] +$$

$$+\left[\xi\left(y_{\bar{\iota},\gamma}(t) - \left(\frac{\rho - \xi(1-\sigma)}{\sigma\xi}\right)\left(\int_{-T}^{0} e^{\xi s} F(\tilde{\gamma}_t)(s)\mathrm{d}s + y_{\bar{\iota},\gamma}\right)\right)\right] =$$

by simple calculations

$$= \left(\xi - \frac{\rho - \xi(1-\sigma)}{\sigma}\right)\left(\int_{-T}^{0} e^{\xi s} F(\tilde{\gamma}_t)(s)\mathrm{d}s + y_{\bar{\iota},\gamma}(t)\right)$$
$$= g\left(\int_{-T}^{0} e^{\xi s} F(\tilde{\gamma}_t)(s)\mathrm{d}s + y_{\bar{\iota},\gamma}(t)\right)$$

and so we have the thesis.

Corollary 4.5

Under hypotheses (*H1*), (*H2*) *and* (*H3*), *given initial data* $(y(0), \tilde{\gamma}_0) = (R(\bar{\iota}), \bar{\iota})$ *in Equation* (*2.6*) (*where* $\bar{\iota}$ *is in* $L^2((-T,0);\mathbb{R}^+)$ *and* $\bar{\iota} \not\equiv 0$) *if we rescale the optimal control for* problem $\mathbf{P}$ $\gamma(\cdot)$ *and the related state trajectory* $y_{\bar{\iota},\gamma}(\cdot)$ *as*

$$\bar{y}(t) \overset{def}{=} e^{-gt}y(t)$$
$$\bar{\gamma}(t) \overset{def}{=} e^{-gt}\gamma(t)$$

we have that $\bar{c}(t) \overset{def}{=} (\bar{y}(t) - \bar{\gamma}(t))$ *is constant on optimal trajectories.*

References

[1] A. Bensoussan, G. Da Prato, M.C. Delfour, and S.K. Mitter, *Representation and control of infinite dimensional systems*, Birkhäuser, Boston (1992).

[2] R. Boucekkine, L.A. Puch, O. Licandro, and F. del Rio, Vintage capital and the dynamics of the AK model, *Journal of Economic Theory*, **120**(1) (2005), 39–72.

[3] R. Boucekkine, D. la Croix, David and O. Licandro, Modelling vintage structures with *DDEs*: principles and applications, *Mathematics for Population Studies*, **11**(3) (2004) 151–179.

[4] R. Boucekkine, F. del Rio, and B. Martinez, *A vintage AK theory of obsolescence and depreciation*, working paper.

[5] M.C. Delfour, The linear quadratic optimal control problem with delays in the state and control variables: A state approach, *SIAM Journal of Control and Optimization*, **24** (1986), 835–883.

[6] M.C. Delfour, Status of the state space theory of linear hereditary differential systems with delays in state and control variables, Analysis and Optimization of Systems (Proceedings Fourth International Conference, Versailles, 1980), *Lecture Notes in Control and Information Sciences*, **28** (1980) 83–96.

[7] M.C. Delfour, Linear optimal control of systems with state and control variable delays, *Automatica Journal (of IFAC)*, **20**(1) (1984), 69–77.

[8] M.C. Delfour and A. Manitius, Control systems with delays: Areas of applications and present status of the linear theory, New Trends in Systems Analysis (Proceedings International Symposium, Versailles, 1976), *Lecture Notes in Control and Information Sciences* (1977), 420–437.

[9] M.C. Delfour, C. McCalla, and S.K. Mitter, Stability and the infinite-time quadratic cost problem for linear hereditary differential systems, *SIAM Journal of Control and Optimization*, **13** (1975), 48–88.

[10] M.C. Delfour and S.K. Mitter, Controllability and observability for infinite-dimensional systems, *SIAM Journal of Control and Optimization*, **10** (1972), 329–333.

[11] M.C. Delfour and S.K. Mitter, Hereditary differential systems with constant delays. II. A class of affine systems and the adjoint problem, *Journal of Differential Equations*, **18** (1975) 18–28.

[12] O. Diekmann, S.A. van Gils, S.M. Verduyn Lunel, and H.O. Walther, *Delay equations*, Springer-Verlag, Berlin (1995).

[13] N. Dunford and J.T. Schwartz, *Linear operators, Part I*, Wiley-Interscience, New York (1966).

[14] G. Fabbri and F. Gozzi, *An AK-type growth model with vintage capital: A dynamic programming approach*, submitted.

[15] A. Ichikawa, Quadratic control of evolution equation with delay in control, SIAM *Journal of Control and Optimization*, **20** (1982), 645–668.

[16] A. Ichikawa, *Evolution equations, quadratic control, and filtering with delay*, Analyse et contrôle de systèmes (Papers, IRIA, Rocquencourt, 1977), 117–126.

[17] I. Lasiecka and R. Triggiani, Control theory for partial differential equations: Continuous and approximation theories. I, *Encyclopedia of Mathematics and Its Applications*, **74** (2000).

[18] S. Rebelo, Long run policy analysis and long run growth, *Journal of Political Economy*, **99** (1991), 500–521.

[19] R.B. Vinter and R.H. Kwong, The infinite time quadratic control problem for linear system with state control delays: An evolution equation approach, *SIAM Journal of Control and Optimization*, **19** (1981), 139–153.

CHAPTER 22

A posteriori error estimates of recovery type for a parameter estimation problem in a linear elastic problem

Tao Feng
Department of Engineering, Physics and Mathematics, Mid Sweden University,
Sundsvall, Sweden

Mårten Gulliksson
Department of Engineering, Physics and Mathematics, Mid Sweden University,
Sundsvall, Sweden

Wenbin Liu
Institute of Mathematics and Statistics, University of Kent,
Canterbury, U.K.

1 Introduction

In the parameter estimation problem governed by the partial differential equations, the finite element approximation plays a very important role (see, for examples, [2] and [12]). It is to be expected that the efficiency of the numerical methods will be influenced by the discretization scheme. The adaptive finite element method is the most important mean to boost the reliability and efficiency of the finite element discretization. The key is to find an a posteriori error estimator that is used to guide the mesh refinement. Roughly speaking, there are residual-type error estimators and recovery-type error estimators. For the literature, readers are referred to recent books by Ainsworth-Oden [1], an earlier book by Verfürth [20], and the references cited therein.

It is well known that the so-called superconvergence patch recovery–type a posteriori error estimators are widely used in engineering applications. The a posteriori error estimators of this type were first introduced by Zienkiewicz and Zhu (see [23], and [24]). They are based on the least square fit of finite element solutions over a local patch of elements at preselected points, where the rate of convergence is higher than the global rate or at least more accurate. It has been proved that the estimated error converges to the true error if the improved solution is one or more orders more accurate than the finite element solution, a phenomenon known as superconvergence in the literature. For the mathematical explanation of the surprising robustness of this type of error estimators, see [21] and [22]. The main advantages of using the ZZ [Zienkiewicz and Zhu] error estimators are the simplicity of its implementation and its cost-effectiveness.

Recently there have been some a posteriori error estimators derived for the parameter estimate problem (see [7], [8], and [14]). However, most of the known a posteriori error estimators are of the residual type. It is well known that residual-type error estimators involve an unknown constant. Therefore, some information may be lost. In this work, an a posteriori error estimator of recovery type is introduced for our parameter estimation problem governed by a linear elastic equation with a finite number of unknown parameters.

In this paper, we consider an isotropic elastic material in two-dimensional space. We give the equations of linearized elasticity

$$\begin{aligned} -\mu\Delta\mathbf{u}-(\lambda+\mu)\nabla(\nabla\cdot\mathbf{u}) &= \mathbf{f} && \text{in } \Omega,\\ \mathbf{u} &= \mathbf{u}_D && \text{on } \Gamma_D \end{aligned} \tag{1.1}$$

in a polygonal $\Omega \subset R^2$, with Lipschitz-continuous boundary. The Dirichlet boundary Γ_D has a positive one-dimensional Lebesgue measure. As usual, $\mathbf{u}$ denotes the displacement, and $\mathbf{f}$ and $\mathbf{u}_D$ denote the body force and the boundary displacement, respectively. Let $U = \{\mathbf{u} \in (H_0^1(\Omega))^2\}$, and the space U is assumed to have the product norm

$$\mathbf{u} = (u_1, u_2) \to \|\mathbf{u}\|_{1,\Omega} = \left(\sum_{i=1}^{2} \|u_i\|_{1,\Omega}^2\right)^{\frac{1}{2}}.$$

We define the strain tensor $(\epsilon_{ij}(\mathbf{u}))$ as

$$\epsilon_{ij}(\mathbf{u}) = \epsilon_{ji}(\mathbf{u}) = \frac{1}{2}(\partial_j u_i + \partial_i u_j),\ 1 \le i,j \le 2, \quad \forall \mathbf{u} \in Y, \quad Y = U + \mathbf{u}_D$$

and by Hooke's law for isotropic bodies, the stress tensor $(\sigma_{ij}(\mathbf{u}))$ is then given by

$$\sigma_{ij}(\mathbf{u}) = \sigma_{ji}(\mathbf{u}) = \lambda\left(\sum_{k=1}^{2} \epsilon_{kk}(\mathbf{u})\right)\delta_{ij} + 2\mu\epsilon_{ij}(\mathbf{u}),\ 1 \le i,j \le 2,$$

where δ_{ij} is Kronecker's symbol. The Lamé coefficients λ and μ are given by

$$\mu = \frac{E}{2(1+\nu)}, \quad \lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}$$

with the Poisson's ratio ν and the Young's modulus E. It is well known that $\lambda \ge \lambda' > 0$ and $\mu \ge \mu' > 0$.

In our parameter estimation problem, we aim at determining the constants λ and μ by using the known measurements of displacement $\mathbf{u}$. For this purpose, the well-known output least squares formulation is used. The Gauss-Newton method with a trust region technique is employed to solve the optimization problem. For more details in the mathematical formulation, see [7] and [8].

This paper is organized as follows. The parameter estimation problem and the finite element discretization are described in Section 2. In Section 3, an a posteriori error estimator of the recovery type is derived for the parameter estimation problem. In the last section, our method is applied in determining the elastic constants of the paper.

2 Parameter estimation problem and its finite element approximation

In this section, we study the finite element approach to our parameter estimation problem. To recover the unknown parameters, we solve

$$\min_{\mathbf{m}} \frac{1}{2} ||Q\mathbf{u}(\mathbf{m}) - \mathbf{z}||_Z^2, \tag{2.2}$$

where $\mathbf{u}$ is the solution of the linear elastic equation (1.1) and the parameter $\mathbf{m} = (m_1, m_2)^T = (\lambda, \mu)^T$. The vector $\mathbf{z} \in Z$ is a given set of measurements and the observation space Z is supposed to be a Hilbert space. Further, we set $Q : Y \to Z$ as a linear bounded observation operator.

Usually, the parameter estimation problem is an ill-posed or ill-conditioned problem, see [11], and some regularization terms are added to the cost function (2.1) such that

$$\min_{\mathbf{m}} \left\{ \frac{1}{2} \left\| Q\mathbf{u}(\mathbf{m}) - \mathbf{z} \right\|_Z^2 + \frac{\beta}{2} \left\| \mathbf{m} - \mathbf{m}_{ref} \right\|^2 \right\},$$

where the penalty parameter β is assumed to be a very small positive number and $\mathbf{m}_{ref}$ is a reference model. The regularization term $\frac{\beta}{2}||\mathbf{m} - \mathbf{m}_{ref}||^2$ improves the conditioning of the inverse problem. Here $||\cdot||$ denotes the l^2 norm of the vector. A good regularization parameter β should yield a fair balance between the perturbation error and regularization error. Assume the data $\mathbf{z}$ contains noise with a known standard deviation $\mathbf{e}$; then the regularization parameter should be chosen such that

$$||Q\mathbf{u}(\mathbf{m}) - \mathbf{z}||_Z = ||\mathbf{e}||_Z,$$

see [17]. To solve the problems without known deviation, methods such as L-curve criterion, generalized cross-validation and the quasi-optimality criterion can be used for the regularization parameter selection; for more details, see [10], [17], and [19].

The weak formulation of (1.1) is given by the following. Find $\mathbf{u} \in Y$ such that

$$a(\mathbf{u}, \mathbf{v}) = (\mathbf{f}, \mathbf{v}), \quad \forall \mathbf{v} \in U,$$

where

$$\begin{aligned} a(\mathbf{u}, \mathbf{v}) &= \int_\Omega \sum_{i,j=1}^{2} \sigma_{ij}(\mathbf{u}) \epsilon_{ij}(\mathbf{v}) dx \\ &= \int_\Omega \{\lambda \; div \; \mathbf{u} \; div \; \mathbf{v} + 2\mu \sum_{i,j=1}^{2} \epsilon_{ij}(\mathbf{u}) \epsilon_{ij}(\mathbf{v})\} dx \quad \forall \mathbf{v} \in U \end{aligned}$$

and

$$(\mathbf{f}, \mathbf{v}) = \int_\Omega \sum_{i=1}^{2} f_i v_i, \quad \forall \mathbf{v} \in U,$$

where $\mathbf{f} = (f_1, f_2) \in (L^2(\Omega))^2$.

We define the composite function

$$s\,(\mathbf{m}) = \frac{1}{2} \|Q\mathbf{u}(\mathbf{m}) - \mathbf{z}\|_Z^2. \tag{2.3}$$

Assume that $s(\mathbf{m})$ is uniformly convex near the solution, at least when $\mathbf{z}$ is attainable. The optimal conditions for (2.1) are to find $(\mathbf{u}, \mathbf{p}, \mathbf{m}) \in Y \times U \times P$ such that (see, e.g., [16])

$$\begin{aligned} & a(\mathbf{u}, \mathbf{v}) = (\mathbf{f}, \mathbf{v}) && \forall \mathbf{v} \in U, \\ & a(\mathbf{q}, \mathbf{p}) = (Q^*(Q\mathbf{u} - \mathbf{z}), \mathbf{q}) && \forall \mathbf{q} \in U, \\ & (s'(\mathbf{m}), \mathbf{m} - \mathbf{w}) \le 0 && \forall \mathbf{w} \in P, \end{aligned} \tag{2.4}$$

where Q^* is the adjoint operator of Q and $(s'(\mathbf{m}), \mathbf{m} - \mathbf{w}) = (-\widetilde{\mathbf{p}}\widetilde{\mathbf{u}}, \mathbf{m} - \mathbf{w})$. We denote

$$(\widetilde{\mathbf{p}}\widetilde{\mathbf{u}}, \mathbf{m} - \mathbf{w}) = \int_\Omega \sum_{i=1}^{2} \widetilde{p}_i \widetilde{u}_i (m_i - w_i),$$

where

$$(\widetilde{p}_1 \widetilde{u}_1, \widetilde{p}_2 \widetilde{u}_2) = (\nabla \cdot \; \mathbf{u} \; \nabla \cdot \; \mathbf{p}, \, 2 \sum_{i,j=1}^{2} \epsilon_{ij}(\mathbf{u}) \epsilon_{ij}(\mathbf{p})).$$

Consider the finite element method for Equation (2.3). Here we are only interested in n-simplex elements and conforming finite elements. Let Ω^h be a polygonal approximation of problem (2.3). Let T^h be a partitioning of Ω^h into disjoint regular n-simplex τ, so that $\overline{\Omega}^h = \cup_{\tau \in T^h} \overline{\tau}$. Each element has at most one face on $\partial\Omega^h$; $\overline{\tau}$ and $\overline{\tau}'$ have at most either one common vertex or a whole edge l if $\overline{\tau}$ and $\overline{\tau}' \in T^h$. The maximum diameter of τ is denoted by h_τ and the mesh parameter h is defined as a cellwise constant function by setting $h\mid_\tau = h_\tau$. We further require that $E_i \in \partial\Omega^h \Rightarrow E_i \in \partial\Omega$ where $\{E_i\}$ is the vertex set associated with the triangulation T^h. For simplicity, we assume that $\Omega^h = \Omega$, that is, Ω is a convex polygon.

Associated with T^h is a finite element subspace W^h of $C(\overline{\Omega}^h)$, such that $\varkappa\mid_\tau$ are polynomials of k-th order ($k \geq 1$) for $\varkappa \in W^h$ and $\tau \in T^h$. Let $V^h = W^h \cap U$; then it is easy to see that $V^h \subset U$.

In order to implement the local mesh refinement, we keep the information about the whole hierarchy of grids starting with the macrotriangulation up to the actual one (see [15]). Every element in this triangulation can be infinitely uniformly refined so that we can obtain an infinite hierarchy tree. Thus, every adaptive mesh is only a section of this hierarchy tree.

The corresponding Galerkin solution $(\mathbf{u}_h, \mathbf{p}_h, \mathbf{m}_h) \in Y^h \times V^h \times P$ is then given by

$$\begin{aligned}
&a_h(\mathbf{u}_h, \mathbf{v}_h) = (\mathbf{f}, \mathbf{v}_h) && \forall \mathbf{v}_h \in V^h,\\
&a_h(\mathbf{q}_h, \mathbf{p}_h) = (Q^*(Q\mathbf{u}_h - \mathbf{z}), \mathbf{q}_h) && \forall \mathbf{q}_h \in V^h, \qquad (2.5)\\
&(s'_h(\mathbf{m}_h), \mathbf{m}_h - \mathbf{w}_h) \leq 0 && \forall \mathbf{w}_h \in P,
\end{aligned}$$

where $(s'_h(\mathbf{m}_h), \mathbf{m}_h - \mathbf{w}_h) = (-\widetilde{\mathbf{p}}_h \widetilde{\mathbf{u}}_h, \mathbf{m}_h - \mathbf{w}_h)$, and

$$a_h(\mathbf{u}_h, \mathbf{v}_h) = \int_\Omega \left\{ \lambda_h \; div\; \mathbf{u}_h \; div\; \mathbf{v}_h + 2\mu_h \sum_{i,j=1}^{2} \epsilon_{ij}(\mathbf{u}_h)\epsilon_{ij}(\mathbf{v}_h) \right\} dx \quad \forall \mathbf{v}_h \in V^h.$$

3 Recovery-type a posteriori error estimates

Before we construct the recovery operator, let us define

$$|\mathbf{t}| = \left(\sum_{i,j=1}^{2} |\epsilon_{ij}(\mathbf{t})|^2_{0,\Omega} \right)^{\frac{1}{2}},$$

which is a norm, equivalent to the product norm (see [5]). We already know that

$$a(\mathbf{v}, \mathbf{v}) \geq 2\mu\, |\mathbf{v}|^2 .$$

Further, we let

$$\frac{\partial \mathbf{u}}{\partial \mathbf{n}} = \left(\sum_{j=1}^{2} \sigma_{1j}(\mathbf{u}) n_j, \sum_{j=1}^{2} \sigma_{2j}(\mathbf{u}) n_j \right)^T,$$

$$\frac{\partial \mathbf{p}}{\partial \mathbf{n}} = \left(\sum_{j=1}^{2} \sigma_{1j}(\mathbf{p}) n_j, \sum_{j=1}^{2} \sigma_{2j}(\mathbf{p}) n_j \right)^T.$$

Suppose that $\mathbf{u}(\mathbf{m}_h)$ and $\mathbf{p}(\mathbf{m}_h)$ are solutions of the equation

$$\begin{aligned} &a_h(\mathbf{u}(\mathbf{m}_h),\mathbf{v}) = (\mathbf{f},\mathbf{v}) && \forall \mathbf{v}\in U,\\ &a_h(\mathbf{q},\mathbf{p}(\mathbf{m}_h)) = (Q^*(Q\mathbf{u}(\mathbf{m}_h)-\mathbf{z}),\mathbf{q}) && \forall \mathbf{q}\in U. \end{aligned} \tag{3.6}$$

Let $(\mathbf{u},\mathbf{p},\mathbf{m})$ and $(\mathbf{u}_h,\mathbf{p}_h,\mathbf{m}_h)$ be the solutions of (2.3) and (2.4), respectively. Then

$$\|\mathbf{m}-\mathbf{m}_h\|^2 \le C(|\mathbf{u}(\mathbf{m}_h)-\mathbf{u}_h|^2 + |\mathbf{p}(\mathbf{m}_h)-\mathbf{p}_h|^2). \tag{3.7}$$

For more details, please refer to [7].

In order to derive recovery-type a posteriori upper error estimates, we need a recovery operator $G_{ij}, 1\le i,j\le 2$. In our work, the recovery operator G_{ij} is similar to the one introduced by Zienkiewicz and Zhu ([23], and [24]). It should be noted that G_{ij} is the same as the ZZ gradient recovery in our piecewise linear case. From V^h to $(V^h)^2$, we have

$$G_{ij}\mathbf{v}_h = \sum_{z\in\partial^2 T^h} S_{ij}\mathbf{v}_h(z)\phi_z,$$

where $\partial^2 T^h$ is the set of all node points, ϕ_z is the base function on the node z, and

$$S_{ij}\mathbf{v}_h(z) = \int_{\omega_z}\frac{\epsilon_{ij}(\mathbf{v}_h)}{|\omega_z|},$$

where ω_z is the support of ϕ_z. For more details, please refer to [1], [23], and [24].

Based on the operator G_{ij} defined above, we can construct the recovery-type a posteriori error estimator as follows.

$$\eta^2 = \sum_\tau \sum_{i,j=1}^{2} (|G_{ij}\mathbf{u}_h - \epsilon_{ij}(\mathbf{u}_h)|^2_{0,\tau} + |G_{ij}\mathbf{p}_h - \epsilon_{ij}(\mathbf{p}_h)|^2_{0,\tau}). \tag{3.8}$$

For the above gradient recovery-type a posteriori error estimator, we can derive the upper error bounds for the parameters on general meshes. To derive the main results of this paper, we need the following lemma.

Lemma 3.1

Assume that πv is the weighted Clément-type interpolation of v defined in [4]. For all $v\in H^1_0(\Omega), \pi v\in V^h\subset H^1_0(\Omega)$, we have

$$\sum_{\tau\in T^h} ||h_\tau^{-1}(v-\pi v)||^2_{0,\tau} \le C|v|^2_{1,\Omega},$$

$$|\pi v|^2_{1,\Omega} \le C|v|^2_{1,\Omega}.$$

Furthermore, if $f \in L^2(\Omega)$*,*

$$\int_\Omega f(v - \pi v) \le C|v|_{1,\Omega} \left(\sum_{z \in \Lambda} \int_{\omega_z} h_z^2 |f - \overline{f}_z|^2 \right)^{1/2},$$

with

$$\overline{f}_z = \frac{\int_{\omega_z} f}{\int_{\omega_z} 1},$$

where Λ *is the set of all inner nodes and* h_z *is the size of* ω_z*. For more details on the proof, see [4].*

Then we derive the recovery-type a posteriori error estimator in the following theorem. In order to simplify the presentation, only the linear element case is considered in our analysis. Further, we denote $[\mathbf{r}]_l$ the jump of $\mathbf{r}$ on the edge l, and $[\mathbf{r}]_l = 0$ when $l \subset \Gamma_D$.

Theorem 3.2

Suppose that $(\mathbf{u}, \mathbf{p}, \mathbf{m})$ *and* $(\mathbf{u}_h, \mathbf{p}_h, \mathbf{m}_h)$ *are the solutions of (2.3) and (2.4), respectively. Then*

$$\|\mathbf{m}_h - \mathbf{m}\|^2 + |\mathbf{u}_h - \mathbf{u}|^2 + |\mathbf{p}_h - \mathbf{p}|^2 \le C\eta^2 + C(\varepsilon_1^2 + \varepsilon_2^2), \tag{3.9}$$

where η *is defined in (3.3) and*

$$\varepsilon_1^2 = \sum_{z \in \Lambda} \int_{\omega_z} h_z^2 |\mathbf{f} - \overline{\mathbf{f}}_z|^2,$$

$$\varepsilon_2^2 = \sum_{z \in \Lambda} \int_{\omega_z} h_z^2 |Q^*(Q\mathbf{u}_h - \mathbf{z}) - \overline{Q^*(Q\mathbf{u}_h - \mathbf{z})}_z|^2.$$

Proof

Let $\mathbf{e}^p = \mathbf{p}(\mathbf{m}_h) - \mathbf{p}_h$*,* $\mathbf{e}^p \in U$*. Then by using the positive definite property of the elasticity matrix and Korn's inequality, also by (2.4), (3.1), and Lemma 3.1, we can obtain*

$$\begin{aligned}
\sum_{i,j=1}^{2} |\epsilon_{ij}(\mathbf{e}^p)|_{0,\Omega}^2 &\le Ca(\mathbf{e}^p, \mathbf{e}^p) \\
&= C\{(Q^*(Q\mathbf{u}(\mathbf{m}_h) - Q\mathbf{u}_h), \mathbf{e}^p) - a((\mathbf{e}^p - \pi\mathbf{e}^p), \mathbf{p}_h) \\
&\quad -(Q^*(Q\mathbf{u}_h - \mathbf{z}), \pi\mathbf{e}^p - \mathbf{e}^p)\} \\
&\le C\left\{ \|\mathbf{u}(\mathbf{m}_h) - \mathbf{u}_h\|_{0,\Omega} \, \|\mathbf{e}^p\|_{0,\Omega} + \sum_{l \bigcap \partial\Omega = \emptyset} \int_l \left[\frac{\partial \mathbf{p}_h}{\partial \mathbf{n}} \right] (\pi\mathbf{e}^p - \mathbf{e}^p) \right.
\end{aligned}$$

$$
\left. + (Q^*(Q\mathbf{u}_h - \mathbf{z}), \mathbf{e}^p - \pi\mathbf{e}^p) \right\}
$$

$$
\leq C \left\{ \|\mathbf{u}(\mathbf{m}_h) - \mathbf{u}_h\|_{0,\Omega} + \delta \|\mathbf{e}^p\|_{1,\Omega} + \sum_{l \bigcap \partial\Omega = \emptyset} h_l \int_l \left[\frac{\partial \mathbf{p}_h}{\partial \mathbf{n}}\right]^2 \right.
$$

$$
\left. + \sum_{z\in\Lambda} \int_{\omega_z} h_z^2 |Q^*(Q\mathbf{u}_h - \mathbf{z}) - \overline{Q^*(Q\mathbf{u}_h - \mathbf{z})}_z|^2 \right\}
$$

$$
\leq C \left\{ \sum_{l \bigcap \partial\Omega = \emptyset} h_l \int_l \left[\frac{\partial \mathbf{p}_h}{\partial \mathbf{n}}\right]^2 + \|\mathbf{u}(\mathbf{m}_h) - \mathbf{u}_h\|_{0,\Omega} + \delta \|\mathbf{e}^p\|_{1,\Omega} + \varepsilon_2^2 \right\}.
$$

We define

$$
\overline{G}\mathbf{p}_h = \begin{pmatrix} (\lambda + 2\mu)G_{11}\mathbf{p}_h + \lambda G_{22}\mathbf{p}_h & 2\mu G_{12}\mathbf{p}_h \\ 2\mu G_{21}\mathbf{p}_h & (\lambda + 2\mu)G_{22}\mathbf{p}_h + \lambda G_{11}\mathbf{p}_h \end{pmatrix}
$$

and

$$
\widetilde{\nabla}\mathbf{p}_h = \begin{pmatrix} (\lambda + 2\mu)\epsilon_{11}(\mathbf{p}_h) + \lambda\epsilon_{22}(\mathbf{p}_h) & 2\mu\epsilon_{12}(\mathbf{p}_h) \\ 2\mu\epsilon_{21}(\mathbf{p}_h) & (\lambda + 2\mu)\epsilon_{22}(\mathbf{p}_h) + \lambda\epsilon_{11}(\mathbf{p}_h) \end{pmatrix}.
$$

Note that $G_{ij}\mathbf{p}_h$ is continuous in Ω, so is $\overline{G}\mathbf{p}_h$. And the components in $\widetilde{\nabla}\mathbf{p}_h$ and $\overline{G}\mathbf{p}_h$ are all polynomials on any element. Then it follows from an inverse inequality (see [5]) that

$$
\begin{aligned}
\sum_{l \bigcap \partial\Omega = \emptyset} h_l \int_l \left[\frac{\partial \mathbf{p}_h}{\partial \mathbf{n}}\right]^2 &\leq \sum_{l \bigcap \partial\Omega = \emptyset} h_l \int_l |[\widetilde{\nabla}\mathbf{p}_h]|^2 \\
&= \sum_{l \bigcap \partial\Omega = \emptyset} h_l \int_l |[\widetilde{\nabla}\mathbf{p}_h - \overline{G}\mathbf{p}_h]|^2 \\
&\leq C \sum_{\tau} h_\tau \int_\tau |\widetilde{\nabla}\mathbf{p}_h - \overline{G}\mathbf{p}_h|^2 \\
&\leq C \sum_{\tau} \sum_{i,j=1}^{2} |G_{ij}\mathbf{p}_h - \epsilon_{ij}(\mathbf{p}_h)|_{0,\tau}^2.
\end{aligned}
$$

Hence,

$$
|\mathbf{p}(\mathbf{m}_h) - \mathbf{p}|^2 \leq C\eta^2 + C\varepsilon_2^2 + C\|\mathbf{u}(\mathbf{m}_h) - \mathbf{u}_h\|_{0,\Omega}. \tag{3.10}
$$

Similarly, let $\mathbf{e}^u = \mathbf{u}(\mathbf{m}_h) - \mathbf{u}_h$*,* $\mathbf{e}^u \in U$*, we have*

$$\begin{aligned}
&\sum_{i,j=1}^{2} |\epsilon_{ij}(\mathbf{e}^u)|^2_{0,\Omega} \le Ca(\mathbf{e}^u, \mathbf{e}^u) \\
&= a(\mathbf{u}(\mathbf{m}_h) - \mathbf{u}_h, \mathbf{e}^u - \pi\mathbf{e}^u) \\
&= C\left\{(\mathbf{f}, \mathbf{e}^u - \pi\mathbf{e}^u) + \sum_{l\bigcap\partial\Omega=\emptyset} \int_l \left[\frac{\partial \mathbf{u}_h}{\partial \mathbf{n}}\right](\pi\mathbf{e}^u - \mathbf{e}^u)\right\} \\
&\le C\left\{\sum_{z\in\Lambda}\int_{\omega_z} h_z^2|\mathbf{f} - \bar{\mathbf{f}}_z|^2 + \sum_{l\bigcap\partial\Omega=\emptyset} \int_l \left[\frac{\partial \mathbf{u}_h}{\partial \mathbf{n}}\right](\mathbf{e}^u - \pi\mathbf{e}^u)\right\} \\
&\le C\left\{\sum_{l\bigcap\partial\Omega=\emptyset} h_l \int_l \left[\frac{\partial \mathbf{u}_h}{\partial \mathbf{n}}\right]^2 + \delta\left\|\mathbf{e}^u\right\|_{1,\Omega} + \varepsilon_1^2\right\}.
\end{aligned}$$

Again,

$$\sum_{l\bigcap\partial\Omega=\emptyset} h_l \int_l \left[\frac{\partial \mathbf{u}_h}{\partial \mathbf{n}}\right]^2 \le C\sum_{\tau}\sum_{i,j=1}^{2} |G_{ij}\mathbf{u}_h - \epsilon_{ij}(\mathbf{u}_h)|^2_{0,\tau},$$

and

$$|\mathbf{u}(\mathbf{m}_h) - \mathbf{u}|^2 \le C\eta^2 + C\varepsilon_1^2. \tag{3.11}$$

Therefore, it follows from (3.2), (3.5), (3.6) that

$$\|\mathbf{m}_h - \mathbf{m}\|^2 + |\mathbf{u}_h - \mathbf{u}(\mathbf{m}_h)|^2 + |\mathbf{p}_h - \mathbf{p}(\mathbf{m}_h)|^2 \le C\eta^2 + C(\varepsilon_1^2 + \varepsilon_2^2), \tag{3.12}$$

According to (2.3) and (3.1), it can be shown that

$$\begin{aligned}
&|\mathbf{p}(\mathbf{m}_h) - \mathbf{p}| \le C(\|\mathbf{m}_h - \mathbf{m}\| + \|\mathbf{u}(\mathbf{m}_h) - \mathbf{u}\|_{0,\Omega}), \\
&|\mathbf{u}(\mathbf{m}_h) - \mathbf{u}| \le C\left\|\mathbf{m}_h - \mathbf{m}\right\|.
\end{aligned} \tag{3.13}$$

It is easy to see that

$$\begin{aligned}
|\mathbf{p}_h - \mathbf{p}| &\le |\mathbf{p}(\mathbf{m}_h) - \mathbf{p}| + |\mathbf{p}(\mathbf{m}_h) - \mathbf{p}_h|, \\
|\mathbf{u}_h - \mathbf{u}| &\le |\mathbf{u}(\mathbf{m}_h) - \mathbf{u}| + |\mathbf{u}(\mathbf{m}_h) - \mathbf{u}_h|,
\end{aligned} \tag{3.14}$$

and (3.4) follows from (3.7) through (3.9).

It is easy to see if $\mathbf{f}, Q^*(Q\mathbf{u}_h - \mathbf{z})$ are smooth, ε_1^2 and ε_2^2 are all high-order terms. Then, the extra terms $\varepsilon_1^2, \varepsilon_2^2$ can be ignored in Theorem 3.2 and η^2 provides an upper-error estimator.

Remark
In a practical situation, observation data may not match the exact data due to measurement errors. The statistical quality of the estimated parameter is proportional to the least square residual $Q\mathbf{u}(\mathbf{m}) - \mathbf{z}$; for more details, see [3]. This means that it is in general not efficient to reduce the error in the parameter by mesh refinement with the discretization error smaller than the statistical error.

Remark
Because we assume that $Q : Y \to Z$ is a linear bounded observation operator, we are able to obtain the adjoint operator $Q^* : Z \to Y$, which is a linear bounded operator too [13]. However, in some applications, one might need the measurement data in some particular points $\mathbf{x}_i, i = 1, 2, \cdots, n_m$, where n_m is the number of measurement points. In the case where $\mathbf{u} \in (H^1(\Omega))^2$, the solution $\mathbf{u}$ may not be continuous when the geometrical dimension d is equal to or higher than two. This implies that the solution in the measurement points may not be defined and Q is not a linear bounded observation operator anymore. In order to circumvent this issue, one may work with the average value of the quantity in a small neighborhood of the measurement points. We use mollification; see [1]. It is customary to choose the mollifiers k_ϵ of the form

$$k_\epsilon(\mathbf{x} - \mathbf{x}_i) = \begin{cases} Cexp[-\epsilon^2/(\epsilon^2 - (\mathbf{x} - \mathbf{x}_i)^2)] & if \quad |\mathbf{x} - \mathbf{x}_i| < \epsilon, \\ 0 & if \quad |\mathbf{x} - \mathbf{x}_i| \geq \epsilon, \end{cases}$$

where the constant C, which depends on d, ϵ and $\mathbf{x}_i$, is selected to satisfy

$$\int_\Omega k_\epsilon(\mathbf{x} - \mathbf{x}_i) d\mathbf{x} = 1.$$

We reformulate our optimization problems as

$$\min_{\mathbf{m}} \frac{1}{2} \sum_i \int_\Omega k_\epsilon(\mathbf{x} - \mathbf{x}_i)(\mathbf{u}(\mathbf{m}) - \mathbf{z}_i)^2 d\mathbf{x},$$

where $\mathbf{z}_i$ denotes the observation data on the point $\mathbf{x}_i$. Then our approach, developed in section 3, is well suited to the pointwise case; see [8] for more details.

4 Numerical example

In this section, we carry out one numerical experiment to demonstrate the error estimators obtained in Section 3. We also compare our error estimators with the residual-type error estimators [8]. The finite element method is defined on a triangular mesh and we consider both piecewise linear and quadratic polynomials for the finite element space. Furthermore, the same mesh will be used for the state and adjoint variables. This means that $\eta_1^2 + \eta_2^2$ will be the indicator of the mesh refinement. All computations are done with AFEPack, a generic C++ adaptive finite element library [15].

In practice, there are four major types of adaptive finite element methods, namely, the h-method (mesh refinement), the p-method (order enrichment), the r-method (mesh motion), and the hp-method. In this paper, we use the h-method. The general idea is to refine the mesh such that the error indicators are equally distributed over the computational mesh. To this end, an equidistribution strategy for mesh refinement is used in our algorithm [6]. We assume that an a posteriori error estimator η has the form $\eta^2 = \Sigma_{e_i}\eta_{e_i}^2$, where e_i is a finite element. In the process of mesh refinement, at each iteration, an average value η_{avg}^2 is calculated. A parameter $0 < \theta < 1$ is defined. Then the element e_i will be refined if $\eta_{e_i}^2 > \theta\eta_{avg}^2$. Generally, the algorithms for refinement and coarseness should be completely different and they are very complex. However, in AFEPack, these two algorithms are the same; moreover there is, in fact, only one algorithm for both refinement and coarseness [15].

In order to solve the optimization problem, we employ the standard Gauss-Newton algorithm. The literature in this area is huge—see, for example, [9] and [18]—and the references cited therein. As we know, there are two main possibilities to ensure the globalization of convergence, line search and trust region methods. In our work, we apply a trust region technique for improving the global convergence. In this method, each iteration step is restricted by the region of validity of the Taylor series. The proof of the global convergence can be found in [9]. It is well known that there are many types of trust-region methods. In our paper, we apply Levenberg-Marquardt methods. For more detail on the techniques directly relevant to our work, see [9] and [18].

We consider determining the elastic constants of the paper from the measured displacements where the model is the equilibrium equations of the linear elastic problem for orthotropic case. In our application, the thickness of the paper is very small compared to its length and width. We fix the lower side of the paper and pull the paper from the upside of the paper. We can measure the displacement $\mathbf{u}$ at some random distinct points, and our aim is to find the Young's modulus and the Poisson's ratio. Here we denote the unknown parameters $\mathbf{m} = (m_1, m_2, m_3)$. For more details on the mathematic model and further information about the application, please refer to [8].

Throughout, we assume the exact parameters $\mathbf{m} = (2, 0.3, 0.8)$. In our example, the computation domain $\Omega = \{(x, y) \mid 0 \le x \le 1,\ 0 \le y \le 1\}$. The boundary conditions on the top edge of the computational domain

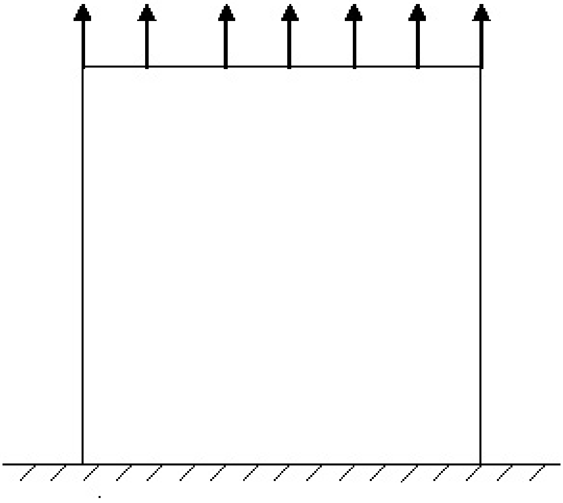


FIGURE 22.1 The domain.

are set to $\mathbf{u} = (0, 0.01)$ and the boundary conditions on the bottom edge are set to $\mathbf{u} = \mathbf{0}$; the boundary conditions are set to homogeneous natural boundary conditions on the right and left edges (see Figure 22.1). We will estimate the parameters by the displacement measurements at 9 different points (0.25,0.75), (0.25,0.5), (0.25,0.25), (0.5,0.75), (0.5,0.5), (0.5,0.25), (0.75,0.75), (0.75,0.5), (0.75,0.25). The exact solution for this problem is not known analytically, so we use a very fine mesh to solve the problem to get the *exact* solutions in the measurement points.

Because point measurements are used, very fine meshes are normally needed around these measurement points; see Figure 22.2. On the one hand, if uniform meshes are used over the computational domain, computational work will be expensive. On the other hand, if we construct nonuniform meshes in an a priori way, generally it is difficult to know how dense the meshes should be around the measurement points and other parts of domain. Therefore, an adaptive mesh refinement strategy is needed to generate the meshes with a good balance between the refined and unrefined regions.

We will now compare the efficiency of our error estimator of recovery type with the error estimator of residual type using a linear element and a quadratic element. The results are shown in Figure 22.3. It demonstrates that our a posteriori error estimator of recovery type is more efficient in our application.

Finally, we add random noise to our pointwise displacement measurements. Figure 22.4 shows the errors in m_1, m_2 and m_3 with different noise levels. Furthermore, in our mesh refinement strategy, $\theta\eta_{avg}^2$ is replaced by a fixed tolerance to avoid mesh overrefinement.

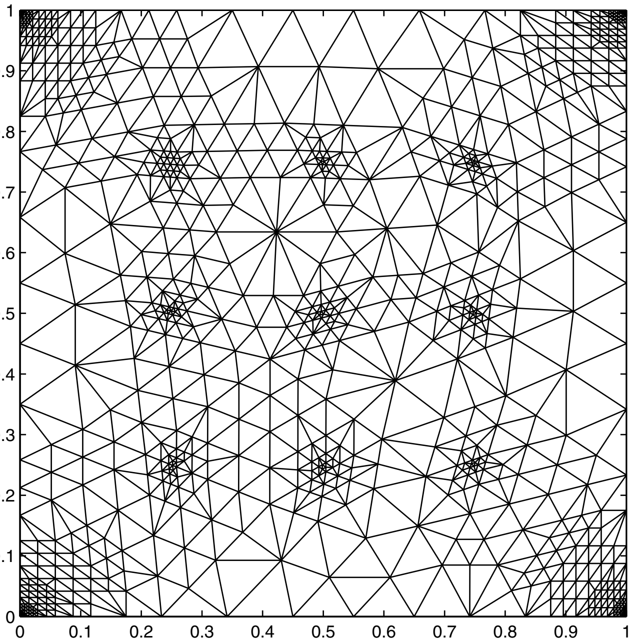


FIGURE 22.2 Typical meshes produced by our error estimator.

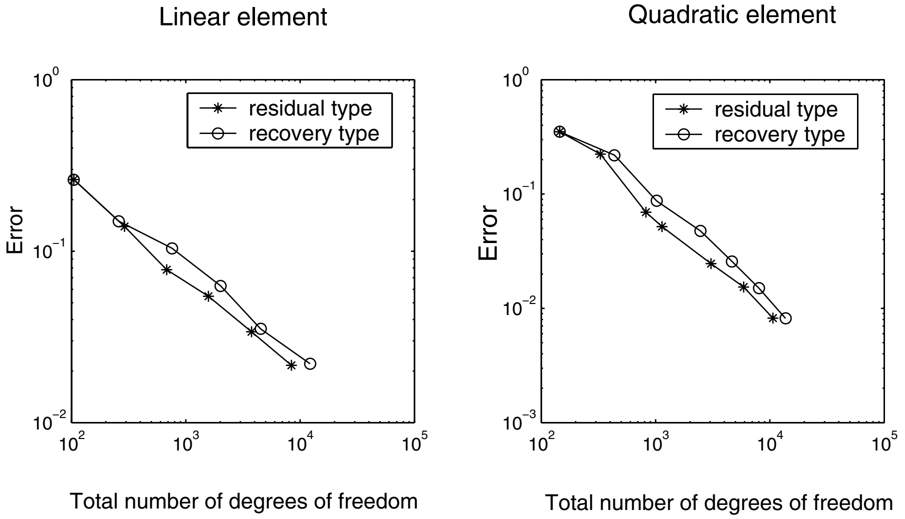


FIGURE 22.3 Error reduction in $\|m - m_h\|$ using recovery-type error estimators and residual-type error estimators.

δ	DOF	$\mid m_1 - m_{1,h} \mid$	$\mid m_2 - m_{2,h} \mid$	$\mid m_3 - m_{3,h} \mid$
5%	3344	0.45898	0.01044	0.01888
1%	3202	0.18419	3.1230e-03	0.06621
0.5%	3365	0.10508	1.7916e-03	0.03707
0.1%	3419	0.01447	1.3124e-04	9.5502e-03
0.0	3033	0.00379	3.1573e-05	1.9598e-03

FIGURE 22.4 Computation with noisy data.

As shown in Figure 22.4, when the noise level is large, the error in the parameters is large as well. If the noise level is further increased, then the computational results are no longer reliable. However, our adaptive meshes are quite efficient for the low-level noise case.

Acknowledgment

The first author would like to thank Professor Ningning Yan for her careful reading of the manuscript and some helpful comments.

References

[1] M. Ainsworth and J. T. Oden, *A Posterior Error Estimation in Finite Element Analysis*, John Wiley & Sons, New York, 2000.

[2] H. T. Banks and K. Kunisch, *Estimation Techniques for Distributed Parameter Systems*, Birkhäuser, Boston, 1989.

[3] D. Bates and D. Watts, *Nonlinear Regression Analysis and Its Applications*, John Wiley & Sons, New York, 1988.

[4] C. Carstensen, Quasi-interpolation and a posteriori analysis in finite element methods, *RAIRO Modél. Math. Anal. Numér.* 33(1999), 1187–1202.

[5] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[6] K. Eriksson and C. Johnson, Adaptive finite element methods for parabolic problems I: A linear model problem, *SIAM J. Numer. Anal.*, 28(1991), 43–77.

[7] T. Feng, M. Gulliksson, and W. B. Liu, Adaptive finite element methods for parameter estimation problem involving linear elastic problem, submitted.

[8] T. Feng, M. Gulliksson, and W. B. Liu, Adaptive finite element methods for identification of elastic constants, *J. Sci. Comput.*, to appear.

[9] R. Fletcher, *Practical Methods of Optimization*, John Wiley & Sons, New York, 2000.

[10] P. C. Hansen and D. P. O'Leary, The use of L-curve in the regularization of discrete ill-posed problems, *SIAM J. Sci. Comput.*, 14(1993), 1487–1503.

[11] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems*, SIAM, Philadelphia, 1998.

[12] T. Kärkkäinen, Error Estimates for Distributed Parameter Identification Problems, Ph.D. Thesis, University of Jyväskylä, 1995.

[13] A. Kirsch, *An Introduction to the Mathematical Theory of Inverse Problems*, Springer-Verlag, New York, 1991.

[14] K. Kunisch, W. B. Liu, and N. N. Yan, *A posteriori error estimates for a model parameter estimation problem*, EUNMA'01 Proceedings, (2002), 723–730.

[15] R. Li, On multi-mesh h-adaptive algorithm, *J. Sci. Comput.*, to appear.

[16] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, 1971.

[17] V. A. Morozov, *Methods for Solving Incorrectly Posed Problems*, Springer-Verlag, New York, 1984.

[18] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Series in Operations Research, Springer-Verlag, New York, 1999.

[19] G. Wahba, *Spline Models for Observation Data*, CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia, 1990.

[20] R. Verfürth, *A Review of a Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, John Wiley & Sons, New York, 1995.

[21] Z. Zhang and J. Z. Zhu, Analysis of the superconvergent patch recovery technique and a posteriori error estimator in the finite element method. Part 1, *Comput. Methods Appl. Mech. Eng.*, 123(1995), 173–187.

[22] Z. Zhang and J. Z. Zhu, Analysis of the superconvergent patch recovery technique and a posteriori error estimator in the finite element method. Part 2, *Comput. Methods Appl. Mech. Eng.*, 163(1998), 159–170.

[23] O. C. Zienkiewicz and J. Z. Zhu, The superconvergent patch recovery and a posteriori error estimates. Part 1: The recovery technique, *Internat. J. Numer. Methods Eng.*, 33(1992), 1331–1364.

[24] O. C. Zienkiewicz and J. Z. Zhu, The superconvergent patch recovery and a posteriori error estimates. Part 2: Error estimates and adaptivity, *Internat. J. Numer. Methods Eng.*, 33(1992), 1365–1382.

CHAPTER 23

Tube derivative of noncylindrical shape functionals and variational formulations

Raja Dziri
Department of Mathematics, University of Tunis,
Tunis, Tunisia

Jean-Paul Zolésio
CNRS and INRIA, Sophia Antipolis, France

1 Introduction

Shape optimization problems for systems modeled by noncylindrical evolution partial differential equations (PDE) are encountered, for example, in fluid-structure and free boundary issues. Cost functionals involved in such problems are expressed in terms of integrals over the noncylindrical evolution domain and/or its lateral boundary. Following [32], this noncylindrical evolution domain will be called a tube and is of the following form:

$$Q = \bigcup_{0<t<\tau} (\{t\} \times \Omega_t); \quad \text{At } \ t = 0, \ \Omega_0 = \Omega.$$

The initial geometry Ω is called the base of the tube.

Nonsmooth tubes are described in [32] with help of time convection of the base by the non-Lipschitzian vector field. However, if the lateral boundary Σ of the tube Q is smooth enough to ensure the existence of the outward normal field $\vec{\nu}$ to Σ, there exists a smooth nonautonomous vector field V such that

$$T_t(V)\Omega = \Omega_t \ \subset \mathbb{R}^N, \ \forall t \in [0, \tau), \tag{1.1}$$

where $T(V)$ is the flow associated with V (cf. paragraph 2). Conversely, to any sufficiently smooth nonautonomous vector field V, one can associate, in the time interval $[0, \tau]$, a tube $Q(V)$ (also denoted Q_V) by setting $\Omega_t = T_t(V)(\Omega)$. The functionalsconsidered may depend not only on the tube containing the evolution, but also on the field that builds this tube. For that reason we consider now as an independent variable in the problem, not the tube itself but the vector field V whose flow mapping builds the tube of base Ω. Of course, extra physical conditions could be imposed on the vector field (cf. [18]).

We designate by Q_V the tube built by an admissible vector field V and we consider functionals in the form

$$\mathbf{j}(V) = J(V, Q_V)$$

where J is a functional depending on V and on the tube $Q(V)$ of base Ω. The exclusive dependence on the shape of Q_V is characterized by:

$$\mathbf{j}(V+W) = \mathbf{j}(V), \ \forall\, W \ \text{ s.t. } \ W \cdot n_{\Omega_t(V)} = \ 0 \quad \text{on } \Sigma_V.$$

The functional $\mathbf{j}$ is thus called a *tube functional.* The dependence of J on a field V comes, generally, from boundary conditions associated with the state equation, for example, sticking boundary conditions in fluid mechanics.

For a bounded tube Q, there exists a bounded open set D (called a hold-all) such that Q and its perturbations remain in $(0,\tau) \times D$. This cylindrical tube will be invariant under all the transformations we shall consider. So the vector field V, as well as the perturbation directions represented by W, will satisfy the condition:

$$W(t).\vec{n}_D = 0 \quad \text{on} \ \ (0,\tau) \times \partial D \qquad (1.2)$$

where $\vec{n}_D$ is the outward normal field to D, cf. for example [22] or [25].

The aim of this work is to compute the derivative with respect to the field of the functional $\mathbf{j}$ and to deduce the expression of the shape derivative. In the first step we obtain a general expression in terms of a *transverse* field denoted $\mathbf{Z}$. After studying the problem whose $\mathbf{Z}$ is the solution, we determine the gradient with the use of the transposed equation and its solution Λ.

We give sufficient conditions under which $\mathbf{j}$ is Gâteaux differentiable at a field V. We prove that if the state depends on the field only by the boundary conditions, the gradient field $G(V)$ is supported by Σ_V. And if it depends only on the shape of the tube, $G(V)$ has the following form

$$G(V)(t) = \gamma^*_{\Gamma_t(V)}[g(V)(t) n_{\Omega_t(V)}]$$

where $\Gamma_t(V) = \partial(\Omega_t(V))$ and $\gamma^*_{\Gamma_t(V)}$ is the transposition of the trace operator on $\Gamma_t(V)$, $g(V)(t) \in [\mathcal{D}^{k-1}(\Gamma_t(V))]'$. For these results one has to assume Ω_t to be $\mathcal{C}^k$ $(k \geq 1)$, $\forall t \in [0,\tau]$.

We conclude the paper by applying the previous work to the cylindrical shape minimization problem. This consists of considering, for given Ω and

$t_0 > 0$, the minimization problem

$$\min_V J(\Omega_{t_0}).$$

Using the gradient method, we get a strategy for building a domain Ω_{k+1} starting from Ω_k such that $J(\Omega_{k+1}) \leq J(\Omega_k)$. As an illustration we consider, as a cost functional, the compliance in structural mechanics. An existence result and an optimal condition are presented.

2 Field characterization for smooth tubes

In $\mathbb{R}^N$, consider an open and bounded set Ω of class C^k $(k \geq 1)$.

Let $Q = \cup_{0<t<\tau}(\{t\} \times \Omega_t)$ be a bounded tube of base Ω. There exists D a smooth and bounded open set in $\mathbb{R}^N$ such that Q remains in $(0, \tau) \times D$. We assume Ω_t of class C^k $(\forall t \in [0, \tau))$ and Q to be of class C^1. The intrinsic outward normal field $\vec{\nu} \in \mathbb{R}^{N+1}$ defined on the lateral boundary Σ of Q, can be written as (cf. [25])

$$\vec{\nu}(t) = \frac{1}{\sqrt{1+v_\nu^2}}(-v_\nu(t), \vec{n}_{\Omega_t})$$

where $\vec{n}_{\Omega_t} \in \mathbb{R}^N$ is the *horizontal* outward normal field to Ω_t and $v_\nu(t)$ is the normal component of the boundary velocity on $\partial\Omega_t$. Introduce the Banach space

$$\mathcal{V}_o^k(D) = \{v \in C^k(\overline{D}, \mathbb{R}^N) \mid v \cdot n_D = 0 \quad \text{on} \quad \partial D\}$$

and consider, in $C([0, \tau], \mathcal{V}_o^k(D))$, the vector field

$$\begin{aligned} V : [0, \tau] \times \overline{D} \longrightarrow \mathbb{R}^N, (t, x) \longmapsto V(t)(x) \overset{def}{=} V(t, x) \ \text{ such that} \\ V(t) \cdot n_{\Omega_t} = v_\nu(t) \quad \text{on} \quad \partial\Omega_t, \forall t \in [0, \tau]. \end{aligned} \tag{2.3}$$

Moreover, we assume

$$(V) \begin{cases} \forall x \in \overline{D}, V(., x) \in C([0, \tau]; \mathbb{R}^N) \\ \exists c > 0, \forall x, y \in \overline{D}, \| V(., x) - V(., y) \|_{C([0,\tau];\mathbb{R}^N)} \leq c|x - y| \end{cases}$$

where $V(., x)$ is the function $t \longmapsto V(t, x)$. For any $X \in \overline{D}$, associate the solution $x_V(., X)$ of the ordinary differential equation

$$\begin{cases} \dfrac{dx}{dt}(t) = V(t, x(t)), \quad t \in [0, \tau] \\ x(0) = X. \end{cases}$$

For any $t \in [0, \tau]$, we have a transformation (cf. [25])

$$T_t(V) : \overline{D} \longrightarrow \overline{D};\; X \longmapsto T_t(V)(X) \overset{def}{=} x_V(t, X)$$

such that $\Omega_t = T_t(V)(\Omega)$. The mapping $(t, X) \longmapsto T_t(V)(X)$ is denoted $T(V)$ or T if no confusion is possible. For any $v \in \mathcal{V}_o^k(D) = \{v \in C^k(\overline{D}, \mathbb{R}^N) \mid v \cdot n_D = 0 \quad \text{on } \partial D\}$, set:

$$c(v) \overset{def}{=} \sup_{y \neq x} \frac{|v(y) - v(x)|}{|y - x|};\; c_k(v) \overset{def}{=} \sum_{|\alpha| = l} c(\partial^\alpha v)\, \forall v \in \mathcal{V}_o^k(D).$$

The regularity result stated below for the flow associated to a given vector field is developed in [29].

Proposition 2.1

For all $W \in C([0, \tau], \mathcal{V}_o^k(D))$ such that

$$c_k(W(t)) \leq c \quad (\textit{for some constant } c > 0 \textit{ independent of } t), \qquad (2.4)$$

it is associated a unique flow $T(W)$ such that

$$T(W) \in C^1([0, \tau], C^k(\overline{D}, \mathbb{R}^N)) \cap C([0, \tau], W^{k+1,\infty}(D, \mathbb{R}^N)).$$

Moreover $T(W)^{-1}$ is in $C([0, \tau], C^k(\overline{D}, \mathbb{R}^N))$ and the mapping

$$t \longrightarrow W(t) \circ T_t \textit{ is in } L^\infty(0, \tau; W^{k+1,\infty}(D, \mathbb{R}^N)).$$

The transformation $T(W)_t^{-1}$ is the flow, at $s = t$, of the vector field $\tilde{W}_t$ defined by $\tilde{W}_t(s) = -W(t - s)$ (cf. [25], [29]). We assume the field V satisfies condition (2.4). Using the implicit function theorem, we obtain the following regularity result.

Proposition 2.2

There exists $\tau' \in]0, \tau]$ such that the mapping $[0, \tau'] \longrightarrow \mathcal{C}^k(\overline{D}, \mathbb{R}^N), t \longmapsto T_t(V)^{-1}$ is continuously differentiable.

Proof

Let $t \in [0, \tau]$. The map $T_t \in \mathcal{C}^k(\overline{D}, \mathbb{R}^N)$. Consider the mapping

$$\Phi : [0, \tau] \times \mathcal{C}^{k-1}(\overline{D}, \mathbb{R}^N) \longrightarrow \mathcal{C}^{k-1}(\overline{D}, \mathbb{R}^N)$$

defined by

$$\Phi(t, S) = T(V)(t, S) - Id.$$

It is clear that $\Phi(t, T_t(V)^{-1}) = 0$. Moreover, Φ is continuously differentiable. The partial derivatives are

$$\partial_t \Phi(t, S) = V(t, T_t(V) \circ S); \;\; \partial_S \Phi(t, S) = (DT_t) \circ S.$$

There exists $\tau' \in]0, \tau]$ such that $\det DT_t \neq 0$, $\forall t \in [0, \tau']$. And so the linear operator $\partial_S \Phi(t, T_t^{-1})$ is in $\mathrm{aut}(\mathcal{C}^{k-1}(\overline{D}, \mathbb{R}^N))$, $\forall t \in [0, \tau']$. We deduce from the implicit function theorem that $T^{-1} \in \mathcal{C}^1([0, \tau']; \mathcal{C}^{k-1}(\overline{D}, \mathbb{R}^N))$.

An immediate consequence is that the mapping $t \longrightarrow DT_t^{-1} = (DT_t)^{-1} \circ T_t^{-1}$ is also in $\mathcal{C}^1([0, \tau']; \mathcal{C}^{k-1}(\overline{D}, \mathbb{R}^N))$.

3 Transverse perturbations

Perturbations of the tube Q_V (based on Ω) in a given direction can be obtained by considering *transverse* transformations (which operate only on domains built at the same time).

For any small positive parameter s, we define the moving domain $Q_{(V+sW)}$ as the perturbation of the tube Q_V in the direction of the field W. It is composed of the sets

$$\Omega_t(V + sW) = T_t(V + sW)(\Omega), \, \forall t \in [0, \tau].$$

A transverse transformation associated with this pertubation is a function which maps $\Omega_t(V)$ onto $\Omega_t(V + sW)$, for any $t \in [0, \tau]$ (and $\overline{D}$ onto $\overline{D}$). A quite natural one is

$$\mathcal{T}_s^t = T_t(V + sW) \circ T_t(V)^{-1}.$$

If the regularity assumptions (T_1) through (T_3) are satisfied by the mapping $(s, x) \longmapsto \mathcal{T}_s^t(x)$ for any $t \in [0, \tau]$, this transformation can be considered as the flow of the vector field (see, for instance, [9])

$$\mathcal{Z}^t(s, .) = \left(\frac{\partial}{\partial s} \mathcal{T}_s^t\right) \circ \mathcal{T}_s^t(\, . \,)^{-1} = [\partial_s T_t(V + sW)] \circ T_t(V + sW)^{-1}. \quad (3.5)$$

Lemma 3.1

Let I_0 be a neighborhood of zero and W in $\mathcal{C}([0, \tau]; \mathcal{V}_o^k(D))$ satisfying condition (2.4). The mapping

$$I_0 \longrightarrow \mathcal{C}([0, \tau]; C^{k-1}(\overline{D}, \mathbb{R}^N))$$

$$s \longrightarrow T(V + sW)$$

is continuously differentiable and $\partial_s(T_t(V+sW))$ *satisfies for any* $t \in [0,\tau]$,

$$\partial_s[T_t(V+sW)] = \int_0^t D(V+sW)(\mu, T_\mu(V+sW))\partial_s[T_\mu(V+sW)]\,d\mu$$
$$+\int_0^t W(\mu, T_\mu(V+sW))\,d\mu. \tag{3.6}$$

Proof

$$T_t(V+sW) - T_t(V+s_oW)$$
$$= \int_0^t (V+sW)(\mu, T_\mu(V+sW)) - (V+s_oW)(\mu, T_\mu(V+s_oW))\,d\mu.$$

$$||T_t(V+sW) - T_t(V+s_oW)||_{C^{k-1}(\overline{D})} \leq |s-s_o| \int_0^t ||W(\mu)||_{C^{k-1}(\overline{D})}\,d\mu$$
$$+ \max_{t\in[0,\tau]} ||DV(t) + s_oDW(t)||_{C^{k-1}(\overline{D})} \int_0^t ||T_\mu(V+sW)$$
$$-T_\mu(V+s_oW)||_{C^{k-1}(\overline{D})}\,d\mu$$

Applying the Gronwall inequality, we derive

$$||T_t(V+sW) - T_t(V+s_oW)||_{C^{k-1}(\overline{D})} \leq \tau|s-s_o|\,||W||_{C([0,\tau];C^{k-1}(\overline{D}))}$$
$$+\tau|s-s_o|\,||W||_{C([0,\tau];C^{k-1}(\overline{D}))} \int_0^t \exp{(t-\mu)}\,d\mu.$$

Thus for any $t \in [0,\tau]$:

$$||T_t(V+sW) - T_t(V+s_oW)||_{C^{k-1}(\overline{D})} \leq \tau|s-s_o|\,||W||_{C([0,\tau];C^{k-1}(\overline{D}))}\,e^t.$$

This proves that the considered map is in $W^{1,\infty}(I_0; C([0,\tau]; C^{k-1}(\overline{D};\mathbb{R}^N)))$ and also gives the following uniform boundedness (with respect to s):

$$\frac{1}{|s-s_o|}\,||(V+sW)(\mu, T_\mu(V+sW))$$
$$-(V+s_oW)(\mu, T_\mu(V+s_oW))||_{C^{k-1}(\overline{D})}$$
$$\leq ||W(\mu)||_{C^{k-1}(\overline{D})} + \tau(\exp\tau) \max_{t\in[0,\tau]} ||DV(t)$$
$$+s_oDW(t)||_{C^{k-1}(\overline{D})}\,||W(\mu)||_{C^{k-1}(\overline{D})}.$$

Then according to the Lesbegue theorem, the derivative exists everywhere in I_0 and satisfies (3.6). It has the following expression:

$$\begin{aligned}\partial_s[T_t(V+sW)] = & \int_0^t \exp\left\{\int_\xi^t D(V+sW)(\mu, T_\mu(V+sW))\,d\mu\right\} \\ & W(\xi, T_\xi(V+sW))\,d\xi. \\ = & \int_0^t DT_t(V+sW).[DT_\xi(V+sW)]^{-1}W(\xi, T_\xi(V+sW))\,d\xi.\end{aligned}$$

It is clear that this expression is continuous in I_0.

Let $\mathcal{S}^t(s) = \partial_s[T_t(V+sW)]$. Then $\mathcal{Z}^t(s,x) = \mathcal{S}^t(s) \circ T_t(V+sW)^{-1}$. In the next section, it will be shown that the field derivatives of noncylindrical functionals are expressed in terms of the transverse vector field $\mathbf{Z}(t,x) = \mathcal{Z}^t(0,x)$. Therefore, one has to know more about this vector field. In particular, it will be shown that $\mathbf{Z}$ can be characterized as the unique solution of

$$\partial_t \mathbf{Z} + [\mathbf{Z}, V] = W \quad \text{in} \quad (0,\tau)\times D \tag{3.7}$$

$$\mathbf{Z}(0,.) = 0 \quad \text{in} \quad D \tag{3.8}$$

where [,] denotes the Lie Brackets.

For that purpose let us consider the vector field S, $S(t,.) \overset{def}{=} \mathcal{S}^t(0,.) = \mathbf{Z}(t)\circ T_t(V)$.

Lemma 3.2

The function S is the unique vector field, in $\mathcal{C}^1([0,\tau]; C^{k-1}(\overline{D}, \mathbb{R}^N))$, satisfying

$$S(t) = \int_0^t W(\mu, T_\mu(V))d\mu + \int_0^t DV(\mu, T_\mu(V))S(\mu)\,d\mu. \tag{3.9}$$

Proof

Let $\mathcal{F}$ be the mapping defined by

$$\begin{aligned}\mathcal{F} : [0,\tau]\times C^{k-1}(\overline{D}, \mathbb{R}^N) &\to C^{k-1}(\overline{D}, \mathbb{R}^N) \\ (t,\varphi) &\to DV(t, T_t(V))\varphi + W(t, T_t(V)).\end{aligned}$$

For any $t \in [0,\tau]$ and any $\varphi \in C^{k-1}(\overline{D})$, $\mathcal{F}(t)$ is affine and $\mathcal{F}(.,\varphi)$ is continuous. So the existence and uniqueness of (3.9) are given by the

Cauchy-Lipschitz theorem. Moreover the solution has the following expression

$$S(t) = \int_0^t \exp\left\{\int_s^t DV(\mu, T_\mu(V))\,d\mu\right\} W(s, T_s(V))\,ds, \;\; \forall t \in [0,\tau].$$
$$= \int_0^t DT_t(V).[DT_s(V)]^{-1} W(s, T_s(V))\,ds.$$

Lemma 3.3

If $\psi \in C^1([0,\tau], C^{k-1}(\overline{D}, \mathbb{R}^N))$ is such that

$$\partial_t \psi + D\psi.V \;\in C([0,\tau], C^{k-1}(\overline{D}, \mathbb{R}^N))$$

and satisfies (3.7) through (3.8), then $\psi \circ T(V)$ belongs to $C^1([0,\tau], C^{k-1}$ $(\overline{D}, \mathbb{R}^N))$ and satisfies (3.9). Conversely, if $\varphi \in C^1([0,\tau], C^{k-1}(\overline{D}, \mathbb{R}^N))$ is the solution of (3.9), then $\varphi \circ T(V)^{-1} \in C^1([0,\tau], C^{k-1}(\overline{D}, \mathbb{R}^N))$ such that

$$\partial_t[\varphi \circ T(V)^{-1}] + D[\varphi \circ T(V)^{-1}] \cdot V \;\; \textit{is in} \;\; C([0,\tau], C^{k-1}(\overline{D}, \mathbb{R}^N))$$

and satisfies (3.7) through (3.8).

Proof

If ψ belongs to $C^1([0,\tau], C^{k-1}(\overline{D}, \mathbb{R}^N))$, then it is the same for the mapping $t \longmapsto \psi(t) \circ T_t$ and we have:

$$\partial_t(\psi(t, T_t)) = [DV(t)] \circ T_t.\, \psi(t, T_t) + W(t) \circ T_t.$$

Thus $\psi \circ T(V)$ satisfies (3.9). Conversely, because $\varphi \in C^1([0,\tau], C^{k-1}(\overline{D}, \mathbb{R}^N))$ is the solution of (3.9) and T^{-1} is in $C^1([0,\tau], C^{k-1}(\overline{D}, \mathbb{R}^N))$, it follows that $\varphi \circ T(V)^{-1}$ is in $C^1([0,\tau], C^{k-1}(\overline{D}, \mathbb{R}^N))$ and we have:

$$\begin{aligned}\partial_t(\varphi(t, T_t^{-1})) &= [\partial_t \varphi] \circ T_t^{-1} + [D\varphi] \circ T_t^{-1}.\partial_t(T_t^{-1})\\ &= \{[DV] \circ T_t.\varphi + W \circ T_t\} \circ T_t^{-1} - (D\varphi) \circ T_t^{-1} D(T_t^{-1}).V(t)\\ &= DV.(\varphi \circ T_t^{-1}) + W - D(\varphi \circ T_t^{-1}).V(t)\end{aligned}$$

which concludes the proof.

Theorem 3.4

The field $\mathbf{Z}$ is the unique vector field, in $C^1([0,\tau], C^{k-1}(\overline{D}, \mathbb{R}^N))$ such that $\partial_t \mathbf{Z} + D\mathbf{Z}.V \in C([0,\tau], C^{k-1}(\overline{D}, \mathbb{R}^N))$; it is the solution of problem (3.7) through (3.8).

Proof

Consider the solution S of (3.9), compose by $T(V)^{-1}$ and make use of the lemma 3.3.

Remark

Taking into account the characterization of $\mathbf{Z}$, we deduce that it has the following expression:

$$\mathbf{Z}(t) = \left\{ \int_0^t DT_t(V).[DT_s(V)]^{-1} W(s, T_s(V))\, ds \right\} \circ T_t(V)^{-1}. \tag{3.10}$$

4 Properties of the transverse field Z

4.1 Transverse field normal component

As we will see in the applications, the expression of the Eulerian derivative depends on $\mathbf{Z}$. Hence, it seems to be necessary to introduce two adjoint states. One associated with the state equation and the other with the field $\mathbf{Z}$. For a functional defined on a tube Q_V (of base Ω), this unusual situation might be avoided by considering the function $\mathbf{z}$ defined by

$$\mathbf{z}(t) = (\mathbf{Z}(t) \cdot n_t) \circ T_t(V) \quad \text{on} \quad (0, \tau) \times \Gamma(\Gamma = \partial\Omega).$$

Lemma 4.1

The mapping $t \longmapsto T_t(V)$ being in $C^1([0, \tau]; C^k(\overline{D}, \mathbb{R}^N))$ therefore

$$t \longmapsto n_t \circ T_t = \frac{{}^*(DT_t)^{-1}\, n}{||\, {}^*(DT_t)^{-1}\, n||} \quad \textit{is in} \;\; C^1([0, \tau]; C^{k-1}(\Gamma))$$

n and n_t are the outward normal fields respectively to Ω and $\Omega_t(V)$, on Γ and Γ_t. Its derivative is given by:

$$\partial_t(n_t \circ T_t) = \langle DV \cdot n_t, n_t \rangle \, \circ T_t \; n_t \circ T_t - \, {}^*DV \circ T_t \; n_t \circ T_t.$$

It follows from this lemma that

$$n_t \circ T_t = \left\{ \exp \int_0^t \langle DV \cdot n_s, n_s \rangle \, \circ T_s \, id - \, {}^*DV \circ T_s \, ds \right\} n.$$

Proposition 4.2

The function $\mathbf{z} \in C^1([0, \tau]; C^{k-1}(\Gamma))$ is the unique solution of

$$\partial_t \mathbf{z}(t) \; - \; \alpha(t) \circ T_t(V)\, \mathbf{z}(t) = \beta(t) \circ T_t(V) \quad \textit{on} \;\; (0, \tau) \times \Gamma \tag{4.11}$$

$$\mathbf{z}(0) = 0 \quad \textit{on} \;\; \Gamma \tag{4.12}$$

where $\alpha(t) = \langle DV \cdot n_t, n_t \rangle$, $\beta(t) = W(t) \cdot n_t$.

Proof

The mapping $t \longmapsto \mathbf{Z}(t, T_t) = S(t)$ is differentiable. Thus

$$\begin{aligned}\partial_t\left[\mathbf{Z}(t,T_t).\, n_t \circ T_t\right] &= \partial_t(\mathbf{Z} \circ T_t)\,.\, n_t \circ T_t + \mathbf{Z}(t,T_t).\partial_t(n_t \circ T_t)\\ &= \langle (W(t) + DV(t).\mathbf{Z}(t)) \circ T_t, n_t \circ T_t\rangle\\ &\quad + \langle DV.\, n_t, n_t\rangle \circ T_t\,(\mathbf{Z}(t)\cdot n_t)\circ T_t - \langle (DV.\mathbf{Z})\circ T_t, n_t \circ T_t\rangle\\ &= (W(t).\, n_t)\circ T_t + \langle DV\cdot n_t, n_t\rangle \circ T_t\;(\mathbf{Z}(t).\, n_t)\circ T_t.\end{aligned}$$

Eventually the desired result is obtained.

Remark

Expression of $\mathbf{z}$ in terms of the data:

From (4.11) we deduce that

$$\partial_t\left[\mathbf{z}\,\exp\left(-\int_0^t \alpha(s)\circ T_s\,ds\right)\right] = \left[-\alpha(t)\circ T_t\,\mathbf{z} + \partial_t\mathbf{z}\,\right]\exp\left(-\int_0^t \alpha(s)\circ T_s\,ds\right)$$

$$= \beta(t)\circ T_t \exp\left(-\int_0^t \alpha(s)\circ T_s\,ds\right).$$

Hence $\mathbf{z}$ can be expressed as follows

$$\mathbf{z}(t) \;=\; \int_0^t \beta(s)\circ T_s(V)\;\exp\left(\int_s^t \alpha(r)\circ T_r(V)\,dr\right)\,ds. \qquad (4.13)$$

4.2 Volume Preservation

A useful property of $\mathbf{Z}$ is that if V and W are of free divergence then $\mathbf{Z}$ is too.

Proposition 4.3

Let W in $C([0,\tau], \mathcal{V}_o^k(D))$ such that (2.4) holds. Assume

$$div V = div W = 0 \;\; in\ D.$$

Then the field $\mathbf{Z}$ is divergence free:

$$div\,\mathbf{Z} = 0 \;\; in\ D\,.$$

Proof

Let $f \in \mathcal{D}(D)$. The transformations $T_t(V+sW)$ and $T_t(V)^{-1}$ map D onto D. Then,

$$\int_D f \circ \mathcal{T}_s^t\;dx = \int_D f\;dx.$$

Indeed, since V and W are of free divergence, we have

$$\int_D f\,dx = \int_{T_t(V+sW)(T_t(V)^{-1}(D))} f\,dx = \int_{T_t(V)^{-1}(D)} f \circ T_t(V+sW)\,dx$$

$$= \int_D f \circ \mathcal{T}_s^t\,dx.$$

From this we deduce that

$$\frac{d}{ds}\left(\int_D f \circ \mathcal{T}_s^t\,dx\right) = 0.$$

The mapping $s \longmapsto \mathcal{T}_s^t$ is in $C^1(I_0; C^{k-1}(\overline{D}, \mathbb{R}^N))$, thus

$$\int_D \nabla f \,.\, \mathbf{Z}\,dx = 0, \quad \forall f \in \mathcal{D}(D).$$

or equivalently, $\operatorname{div}\mathbf{Z} = 0$ in $\mathcal{D}'(D)$.

5 Adjoint problem associated with the transverse field

As shown in the proof of Lemma 3.3, the solution of (3.7) through (3.8) is obtained via a change of variable. Then if $H(D)$ is a Banach space of functions defined on D, stable by multiplication by functions in $C^{k-1}(\overline{D})$, the same process generates the solution of the adjoint problem associated with $\mathbf{Z}$.

5.1 General case

Theorem 5.1

Let $F \in L^2((0,\tau); H(D))$. There exists a unique $\Lambda \in C([0,\tau]; H(D))$ such that $\partial_t\Lambda + D\Lambda.V \in L^2((0,\tau); H(D))$ solution of

$$-\partial_t\Lambda - D\Lambda \cdot V - {}^*DV \cdot \Lambda - (divV)\,\Lambda = F \tag{5.14}$$

$$\Lambda(\tau) = 0. \tag{5.15}$$

Proof

Consider $\theta \in C^1([0,\tau]; H(D))$, the unique solution of the backward problem

$$-\partial_t\theta - [\,{}^*(DV)\circ T_t + (\operatorname{div}V)\circ T_t\,]\;\theta = F \circ T_t$$

$$\theta(\tau) = 0.$$

Applying $\exp \displaystyle\int_0^t [({}^*DV(s)) \circ T_s + (\mathrm{div} V(s)) \circ T_s \, \mathbf{I}] \; ds$, we get

$$-\partial_t \left[\exp \left\{ \int_0^t [{}^*(DV(s)) \circ T_s + (\mathrm{div} V(s)) \circ T_s \, \mathbf{I}] \; ds \right\} \theta(t) \right]$$
$$= \exp \left\{ \int_0^t [{}^*(DV(s)) \circ T_s + (\mathrm{div} V(s)) \circ T_s \, \mathbf{I}] \; ds \right\} F \circ T_t.$$

By integration we deduce an explicit expression of θ:

$$\theta(t) = \int_t^\tau \exp \left\{ - \int_s^t [{}^*DV(\xi) \circ T_\xi + (\mathrm{div} V(\xi)) \circ T_\xi \, \mathbf{I}] \, d\xi \right\} F(s) \circ T_s \; ds.$$
$$= \int_t^\tau {}^*(DT_t)^{-1} \, {}^*(DT_s) F(s) \circ T_s \det DT_s (\det DT_t)^{-1} \, ds$$

Then taking $\Lambda = \theta \circ T_t^{-1}$, it is easy to see that Λ is the unique solution of (5.14) through (5.15) which, for a suitable right-hand term depending on the given cost functional, will represent the adjoint problem associated with $\mathbf{Z}$.

5.2 A right-hand term supported by the lateral boundary

Let $f \in L^2(\Sigma(V))$ and assume that the mapping $t \longmapsto \gamma^*_{\Gamma_t}(f(t) n_t)$ belongs to $L^2((0,\tau), H(D))$. We proved in Theorem 5.1 the existence of a unique $\Lambda \in C([0,\tau], H(D))$ such that

$$-\partial_t \Lambda - D\Lambda . V - {}^*DV \cdot \Lambda - \mathrm{div} V \; \Lambda = \gamma^*_{\Gamma_t}(f(t) n_t) \qquad (5.16)$$
$$\Lambda(\tau) = 0.$$

We shall prove that the solution Λ is, in fact, supported by the lateral boundary $\Sigma(V)$ because the right-hand term in this problem is itself supported by $\Sigma(V)$ and there is no diffusion term.

Lemma 5.2

Let $f \in L^2(0,\tau;\, L^2(\Gamma_t))$. There exists a unique solution in $C([0,\tau];\, L^2(\Gamma_t))$ such that $\partial_t \lambda + \nabla_{\Gamma_t} \lambda . V \in L^2(0,\tau; L^2(\Gamma_t))$ of the following problem

$$\begin{cases} \partial_t \lambda(t) + \nabla_{\Gamma_t} \lambda . V + \lambda \, \mathit{div} V = f(t) \quad \mathit{on} \;\; \cup_t \, (\{t\} \times \Gamma_t). \\ \lambda(\tau) \qquad\qquad\qquad\qquad\quad = 0 \quad \mathit{on} \;\; \Gamma_\tau \end{cases}$$

Proof

Notice that

$$[\partial_t \lambda(t) + \nabla_{\Gamma_t} \lambda . V]_{|\Gamma_t} \circ T_t = \partial_t (\lambda \circ T_t)_{|\Gamma}$$

and consider $\mu \in C([0,\tau]; L^2(\Gamma))$ the unique solution of

$$\begin{aligned}&\partial_t\mu + (\mathrm{div}V)\circ T_t \;\; \mu = f(t)\circ T_t \quad \text{on} \quad (0,\tau)\times\Gamma\\ &\mu(\tau) = 0 \quad \text{on} \quad \Gamma.\end{aligned}$$

which can be expressed as:

$$\begin{aligned}\mu(t) &= -\int_t^\tau \exp\left\{\int_t^s (\mathrm{div}V)\circ T_r\, dr\right\}\, f(s)\circ T_s\, ds\\ &= -\int_t^\tau \mathrm{det}DT_s.(\det DT_t)^{-1}\, f(s)\circ T_s\, ds.\end{aligned}$$

Then λ, defined by $\lambda(t) = \mu(t)\circ T_t^{-1}$, is the solution of the problem. Uniqueness is obvious.

Theorem 5.3

Let $f \in L^2(0,\tau; L^2(\Gamma_t))$. The solution Λ of (5.16) is supported by $\Sigma(V)$. Precisely

$$\Lambda(t) = -\gamma^*_{\Gamma_t}(\lambda(t)n_t), \qquad t\in(0,\tau), \tag{5.17}$$

where λ is defined in Lemma 5.2.

Proof

Set $\mathbf{X}(t) = -\gamma^*_{\Gamma_t}(\lambda(t)n_t)\,(\in H^{-1}(D,\mathbb{R}^N))$. We should identify the distribution

$$-\partial_t\mathbf{X} - D\mathbf{X}.V - {}^*DV.\mathbf{X} - (\mathrm{div}V)\,\mathbf{X}.$$

For that let $\varphi\in\mathcal{D}((0,\tau)\times D)$, thus

$$\langle -\partial_t\mathbf{X} - D\mathbf{X}.V - {}^*DV.\mathbf{X} - (\mathrm{div}V)\,\mathbf{X}, \varphi\rangle_{\mathcal{D}'((0,\tau)\times D),\mathcal{D}((0,\tau)\times D)}$$

$= -\int_0^\tau\int_{\Gamma_t} \lambda(t)\,\langle\partial_t\varphi, n_t\rangle\, d\Gamma_t dt + \int_0^\tau\int_{\Gamma_t}\lambda(t)\langle -D\varphi.V + DV.\varphi, n_t\rangle\, d\Gamma_t dt$ The first term $E_1 = -\int_0^\tau\int_{\Gamma_t}\lambda\,(\partial_t\varphi)\cdot n_t\, d\Gamma_t dt$ is treated as follows: Using the transformation $T_t(V)$

$$\begin{aligned}E_1 &= -\int_0^\tau\int_\Gamma \lambda\circ T_t\,[(\partial_t\varphi)\circ T_t]\,.\,n_t\circ T_t\,\omega(t)\, d\Gamma dt,\ \omega(t)\\ &= det(DT_t)\,\|\,{}^*DT_t^{-1}\cdot n\,\|_{\mathbb{R}^N}\\ &= -\int_0^\tau\int_\Gamma\langle\partial_t(\varphi\circ T_t) - (D\varphi.V)\circ T_t\,,\ n_t\circ T_t\rangle\,\lambda\circ T_t\,\omega(t)\, d\Gamma dt\\ &= \int_0^\tau\int_\Gamma\langle\varphi\circ T_t\,,\, n_t\circ T_t\rangle\,\partial_t(\lambda\circ T_t)\,\omega(t)\end{aligned}$$

$$
\begin{aligned}
&+\langle \varphi\circ T_t\,,\, \partial_t(\omega(t)\, n_t\circ T_t)\rangle\; \lambda\circ T_t\, d\Gamma dt\\
&+\int_0^\tau\int_\Gamma \langle (D\varphi.V)\circ T_t\,,\, n_t\circ T_t\rangle\;\; \lambda\circ T_t\;\omega(t)\; d\Gamma dt.
\end{aligned}
$$

But $\omega(t)n_t\circ T_t = \gamma(t)^* DT_t^{-1} n$; where $\gamma(t) = \det DT_t$ so

$$
\begin{aligned}
E_1 = &\int_0^\tau\int_\Gamma \partial_t(\lambda\circ T_t)\langle \varphi\circ T_t\,,\, n_t\circ T_t\rangle\,\omega(t)\; d\Gamma dt\\
&+\int_0^\tau\int_\Gamma \langle \varphi\circ T_t\,,\, \partial_t(\gamma(t)\,{}^*DT_t^{-1}n)\rangle\; \lambda\circ T_t\; d\Gamma dt,\\
&+\int_0^\tau\int_{\Gamma_t} \lambda\langle D\varphi.V\,,\, n_t\rangle\; d\Gamma_t dt,\\
&=\int_0^\tau\int_\Gamma [\partial_t\lambda+\nabla_{\Gamma_t}\lambda.V]\circ T_t\langle \omega(t)\; n_t\circ T_t\,, \varphi\circ T_t\rangle\, d\Gamma dt\\
&+\int_0^\tau\int_\Gamma \lambda\circ T_t\langle \varphi\circ T_t\,,\, \partial_t({}^*DT_t^{-1})n\rangle\,\gamma(t)\, d\Gamma dt\\
&+\int_0^\tau\int_\Gamma \lambda\circ T_t\langle \varphi\circ T_t\,,\, \partial_t(\gamma(t))\;{}^*DT_t^{-1}n\rangle\, d\Gamma dt\\
&+\int_0^\tau\int_{\Gamma_t} \lambda\langle D\varphi.V\,,\, n_t\rangle\; d\Gamma_t dt\\
&=\int_0^\tau\int_{\Gamma_t} [\partial_t\lambda+\nabla_{\Gamma_t}\lambda.V]\;\varphi\cdot n_t\, d\Gamma_t dt+\int_0^\tau\int_{\Gamma_t}\langle D\varphi.V\,,\, n_t\rangle\,\lambda\; d\Gamma_t dt\\
&+\int_0^\tau\int_\Gamma \lambda\circ T_t\langle \varphi\circ T_t\,,\, n_t\circ T_t\rangle\,\omega(t)\;(\mathrm{div} V)\circ T_t\, d\Gamma dt\\
&-\int_0^\tau\int_\Gamma \lambda\circ T_t\langle \varphi\circ T_t\,,\, {}^*(DV)\circ T_t\; n_t\circ T_t\rangle\,\omega(t)\; d\Gamma dt\\
&=\int_0^\tau\int_{\Gamma_t} (\partial_t\lambda+\nabla_{\Gamma_t}\lambda\cdot V)\,\varphi\cdot n_t\, d\Gamma_t dt+\int_0^\tau\int_{\Gamma_t}\lambda\langle D\varphi.V\,,\, n_t\rangle\, d\Gamma_t dt\\
&+\int_0^\tau\int_{\Gamma_t} \lambda\,\varphi\cdot n_t\,\mathrm{div} V\; d\Gamma_t dt-\int_0^\tau\int_{\Gamma_t}\lambda\langle DV.\varphi\,,\, n_t\rangle\, d\Gamma_t dt.
\end{aligned}
$$

Finally we obtain

$$
\begin{aligned}
&\langle -\partial_t\mathbf{X} - D\mathbf{X}.V - {}^*DV.\mathbf{X} - \mathrm{div} V\,\mathbf{X}\,,\, \varphi_{\mathcal{D}',\mathcal{D}}\rangle\\
&\quad=\int_0^\tau\!\!\int_{\Gamma_t}(\partial_t\lambda+\nabla_{\Gamma_t}\lambda.V+\lambda\mathrm{div} V)\,\varphi\cdot n_t\; d\Gamma_t dt=\int_0^\tau\!\!\int_{\Gamma_t} f(t)\,\varphi(t)\cdot n_t\; d\Gamma_t dt
\end{aligned}
$$

which is equivalent, in a distribution sense, to

$$
-\partial_t\mathbf{X} - D\mathbf{X}.V - {}^*DV.\mathbf{X} - \mathrm{div} V\,\mathbf{X} = \gamma_{\Gamma_t}^*(f(t)n_t).
$$

Moreover, it is clear that $\mathbf{X}(\tau) = 0$.

From the uniqueness Theorem 5.1, we deduce that $\Lambda(t) = -\gamma^*_{\Gamma_t}(\lambda(t)n_t)$.

6 Derivability with respect to the field

As mentioned in the beginning, we are interested in the structure of the Eulerian derivative for a class of noncylindrical functionals of the following type

$$\mathbf{j}(V) = \int_{Q(V)} F(t,x,u(V)(t,x))\,dxdt$$

where:
$Q(V) = \bigcup_{0<t<\tau}(\{t\} \times T_t(V)(\Omega))$,
τ is a non-negative scalar,
Ω is a bounded open subset of $\mathbb{R}^N$ of class $\mathcal{C}^k$, $k \geq 2$,
$F : I(=[0,\tau]) \times D \times \mathbb{R}^{N'} \to \mathbb{R}\,(N' \in \mathbb{N}^*)$ is of class $\mathcal{C}^1$.

The function u is the solution of a well-posed non-cylindrical parabolic PDE of order $2m$, $m \leq 2k$ $(m \in \mathbb{N}^*)$ defined in $Q(V)$.

In the sequel, we adopt the following hypotheses.

6.1 Assumption-Definition

For all admissible V and W, we assume

- *the existence of the derivative, at $s = 0$, of the mapping $s \to u^s(.,.) = u(V + sW)(., T_s(\mathcal{Z}(s)(.))$ in $L^2(I, H^{2m}(\Omega_t(V))$ (it will be denoted $\dot{u}(V;W)$)*
- *the mapping*

$$W \longrightarrow \dot{u}(V;W) \text{ to be linear and continuous}$$

and we introduce
$$u'(V;W) \overset{def}{=} \dot{u}(V;W) - \partial_x u.\mathbf{Z}.$$

Lemma 6.1

Under the previous assumptions, the functional $\mathbf{j}(.)$ is Gâteaux differentiable at V and there exists a time-dependent distribution $G(V)$, $G(V)(t) \in \mathcal{D}'(D,\mathbb{R}^N)$ for a.e. t, such that

- *The support of $G(V)(t)$, $spt[G(V)(t)]$, is contained in $\overline{\Omega_t}(V)$*
- *For all $W \in \mathcal{C}^1([0,\tau], \mathcal{D}(D,\mathbb{R}^N))$, the mapping $t \longrightarrow \langle G(V)(t), W(t)\rangle_{\mathcal{D}'(D,\mathbb{R}^N),\mathcal{D}(D,\mathbb{R}^N)}$ is in $L^1(0,\tau)$*
and
$$\mathbf{j}'(V;W) = \int_0^\tau \langle G(V)(t), W(t)\rangle_{\mathcal{D}'(D,\mathbb{R}^N),\mathcal{D}(D,\mathbb{R}^N)}\,dt.$$

Proof

In the perturbed tube $Q(V+sW)$, the cost functional has the following expression:

$$\begin{aligned}\mathbf{j}(V+sW) &= \int_0^\tau \int_{\Omega_t(V+sW)} F(t,x,u(V+sW)(t,x))\,dxdt\\ &= \int_0^\tau \int_{T_s(\mathcal{Z}^t)(\Omega_t(V))} F(t,x,u(V+sW)(t,x))\,dxdt\\ &= \int_0^\tau \int_{\Omega_t(V)} F(t,T_s(\mathcal{Z}^t)(x),u^s(t,x))\,\det D\mathcal{Z}^t(s)\,dxdt\end{aligned}$$

The assumptions ensure the existence of

$$\mathbf{j}'(V;W) = \frac{d}{ds}\mathbf{j}(V+sW)_{|s=0}.$$

Precisely we have

$$\begin{aligned}\mathbf{j}'(V;W) = \int_0^\tau \int_{\Omega_t(V)} &\partial_x F(t,x,u(t,x)).\mathbf{Z}(t,x) + \partial_y F(t,x,u(t,x)).\dot{u}(t,x)\\ &+ F(t,x,u(t,x))\,\mathrm{div}_x\mathbf{Z}(t,x)\,dxdt\end{aligned}$$

which is equivalent to

$$\begin{aligned}\mathbf{j}'(V;W) = &\int_0^\tau \partial_y F(t,x,u(t,x)).[\dot{u}(t,x) - \partial_x u(t,x).\mathbf{Z}(t,x)]\,dt\\ &+\int_0^\tau \int_{\Omega_t(V)} \mathrm{div}_x[F(t,x,u(t,x))\mathbf{Z}(t)]\,dxdt.\end{aligned}$$

Finally we can rewrite

$$\begin{aligned}\mathbf{j}'(V;W) = &\int_0^\tau \int_{\Omega_t(V)} \partial_y F(t,x,u(t,x)).u'(t,x)\,dxdt \qquad (6.18)\\ &+\int_0^\tau \int_{\Gamma_t(V)} F(t,x,u(t,x))\mathbf{Z}(t)\cdot n_t\,d\Gamma_t dt.\end{aligned}$$

Because u' and $\mathbf{Z}$ depend linearly on W, we obtain the linear dependence of $\mathbf{j}'(V;W)$ on W. So a.e. there exists $G(V)(t) \in \mathcal{D}'(D,\mathbb{R}^N)$ such that for any $W \in \mathcal{C}^1([0,\tau];\mathcal{D}(D,\mathbb{R}^N))$

$$\mathbf{j}'(V;W) = \int_0^\tau \langle G(V)(t), W(t)\rangle_{\mathcal{D}'(D),\mathcal{D}(D)}\,dt.$$

To localize the support of $G(V)(t)$ it suffices to consider W such that, for any $t \in [0,\tau]$, $\text{spt}\{W(t)\} \subset \overline{\Omega_t(V)}^c$ and compute $\mathbf{j}(V+sW)$. The assumption on W implies that $W(t)\cdot n_{\Omega_t(V)} = 0$ on $\Gamma_t(V)$, so $\Omega_t(V+sW) = \Omega_t(V), \forall t \in [0,\tau]$. Also, because $W(t)_{|\Omega_t(V)} = 0$, we get $u(V+sW) = u(V)$. So $\mathbf{j}(V+sW) = \mathbf{j}(V)$ and $\mathbf{j}'(V;W) = 0$.

Considering Λ the solution of problems (5.14) through (5.15), we can express the boundary integral on $\mathbf{Z}(t)\cdot n_t$ explicitly in terms of $W(t)\cdot n_t$. This is the object of the following lemma.

Lemma 6.2

Let F be a sufficiently smooth function defined on $\Sigma(V)$. Then

$$\int_0^\tau \int_{\Gamma_t(V)} F(t)\,\mathbf{Z}(t)\cdot n(t)\,d\Gamma_t dt = -\int_0^\tau \int_{\Gamma_t(V)} \lambda(t)\,W(t)\cdot n(t)\,d\Gamma_t dt =$$

$$\int_0^\tau \int_{\Gamma_t(V)} \left\{\int_t^\tau F(s)\circ T_s(V)\gamma(s)\gamma(t)^{-1}\,ds\right\}\circ T_t(V)^{-1}\,W(t)\cdot n(t)\,d\Gamma_t dt$$

where λ is defined in Lemma 5.2.

Proof

Using the adjoint state Λ associated to $\mathbf{Z}$, it comes

$$\int_0^\tau \int_{\Gamma_t} F(t)\,\mathbf{Z}\cdot n_t\,d\Gamma_t dt = \int_0^\tau \langle \gamma^*_{\Gamma_t}(F(t)n_t), \mathbf{Z}(t)\rangle\,dt$$

$$= \int_0^\tau \langle -\partial_t\Lambda - D\Lambda\cdot V - {}^*DV\cdot\Lambda - (\text{div}V)\Lambda, \mathbf{Z}\rangle\,dt$$

$$= \int_0^\tau \langle \partial_t\mathbf{Z} + D\mathbf{Z}\cdot V - DV.\mathbf{Z}, \Lambda\rangle\,dt = -\int_0^\tau \int_{\Gamma_t(V)} \lambda(t)\,W\cdot n_t\,d\Gamma_t dt$$

$$= \int_0^\tau \int_{\Gamma_t} \int_t^\tau F(s)\circ T_s\circ T_t(V)^{-1}\,W(t)\cdot n(t)[\gamma(s)\gamma(t)^{-1}]\circ T_t(V)^{-1}\,ds\,d\Gamma_t dt.$$

Remark

If we use the explicit expression of $\mathbf{Z}$ given by (3.10) we obtain the following expression of the integral in terms of W:

$$\int_0^\tau \int_{\Gamma_t} F(t)\,\mathbf{Z}\cdot n_t\,d\Gamma_t dt =$$

$$\int_0^\tau \int_{\Gamma_t} \int_0^t F(t)[DT_t(V)(DT_s(V))^{-1}W(s)\circ T_s(V)]\circ T_t(V)^{-1}\cdot n(t)\;\,ds\,d\Gamma_t dt.$$

In the sequel, it will be shown that the Eulerian derivative coincides with the shape derivative when the functional depends only on the shape of the tube. First let us define a *tube function* (resp. *tube functional*).

Definition 6.3 *If $u(V+W)=u(V)$ (resp. $\mathbf{j}(\mathbf{V}+\mathbf{W})=\mathbf{j}(\mathbf{V})$) for any sufficiently smooth V and W s.t. $W(t)\cdot n_{\Omega_t(V)}=0$ on $\Sigma(V)$, then u (resp. $\mathbf{j}$) depends only on the shape of the considered tube. It is called a tube function (resp. tube functional).*

Proposition 6.4

Assume the hypotheses of Lemma 6.1.

1. *If $u(V)$ depends only on the trace of the field V on the lateral boundary $\Sigma(V)$, then the gradient $G(V)$ is supported on $\Sigma(V)$ and there exists a time-dependent boundary distribution $R(V)$ such that $R(V)(t)\in\mathcal{D}'(\Gamma_t(V))$ for a.e. t and*

$$G(V)(t)=\gamma^*_{\Gamma_t(V)}(R(V)(t))$$

where $\gamma^*_{\Gamma_t(V)}$ is the transposition of the trace operator on $\Gamma_t(V)$.

2. *If $u(V)$ is a tube function, then*

$$\mathbf{j}'(V;W)=0 \quad \textit{for any } W \ \textit{ s.t. } \ W(t)\cdot n_{\Omega_t(V)}=0 \ \textit{ in } \ \Gamma_t(V) \\ \textit{for a.e. } \ t\in[0,\tau].$$

3. *The initial domain Ω being of class $\mathcal{C}^k$, if the linear mapping $W\longmapsto \mathbf{j}'(V;W)$ is continuous in $\mathcal{C}([0,\tau];C^k(\overline{D},\mathbb{R}^N))$ and if $u(V)$ is a tube function, then there exists a time-dependent distribution g, $g(t)\in[\mathcal{D}^{k-1}(\Gamma_t(V))]'$, such that*

$$\mathbf{j}'(V;W)=\int_0^\tau\langle\gamma^*_{\Gamma_t(V)}(g(V)(t)n_t),W(t)\rangle_{\mathcal{D}^{k-1}(D)',\mathcal{D}^{k-1}(D)}\,dt$$

Proof

1. We already know, from Lemma 6.1, that $\mathrm{spt}G(V)(t)\subset\overline{\Omega_t(V)}$ a.e. in $[0,\tau]$. By similar arguments and considering vector fields W such that $W(t)\in\mathcal{D}(\Omega_t(V),\mathbb{R}^N)$ for any $t\in[0,\tau)$, we prove that $\mathrm{spt}G(V)(t)\subset\Omega_t(V)^c$ for any $t\in[0,\tau)$. Hence, we can conclude that $\mathrm{spt}G(V)(.)\subset\Sigma(V)$ and there exists a time-dependent boundary distribution $R(V)$, $R(V)(t)\in\mathcal{D}'(\Gamma_t(V);\mathbb{R}^N))$ such that

$$G(V)(t)=\gamma^*_{\Gamma_t}(R(V)(t)).$$

2. Expressing the functional $\mathbf{j}$ at the point $(V + sW)$, we obtain
$$\begin{aligned}\mathbf{j}(V+sW) &= J(V+sW, Q(V+sW))\\ &= J(V+sW, \cup_{0<t<\tau}[\{t\} \times \mathcal{T}_s^t\Omega_t(V)])\end{aligned}$$
where $\mathcal{T}_s^t = T_t(V+sW) \circ T_t(V)^{-1}$.
The condition $W(t) \cdot n_{\Omega_t(V)} = 0$ for any $t \in [0, \tau)$ implies that $T_t(V + sW)\Omega = T_t(V)\Omega$ $(= \Omega_t(V))$. It follows that $Q(V + sW) = Q(V)$. Moreover, $u(V + sW) = u(V)$, then $\mathbf{j}(V + sW) = \mathbf{j}(V)$. Therefore $\mathbf{j}'(V; W) = 0$.
3. The continuity of the mapping $W \longmapsto \mathbf{j}'(V; W)$ in $\mathcal{C}([0, \tau]; C^k(\overline{D}, \mathbb{R}^N))$ and the fact that the gradient $G(V)(t) = \gamma^*_{\Gamma_t(V)}(R(V)(t))$ give that $R(V)(t) \in \mathcal{D}^{-k}(\Gamma_t(V)))$ for a.e. $t \in [0, \tau]$. Moreover, the outward normal field $n_t(V)$ being in $\mathcal{D}^{k-1}(\Gamma_t(V))$, let $W \in \mathcal{C}([0, \tau]; \mathcal{D}^{k-1}(D, \mathbb{R}^N))$ and C an admissible vector field such that
$$C(t) = W(t) - (W(t) \cdot n_t(V))\, n_t(V) \quad \text{on} \quad \Gamma_t(V).$$
Therefore
$$\begin{aligned}\mathbf{j}'(V;W) &= \int_0^\tau \langle G(V)(t), W(t)\rangle_{\mathcal{D}^{-k+1},\mathcal{D}^{k-1}}\, dt\\ &= \int_0^\tau \langle G(V)(t), C(t)\rangle\, dt + \int_0^\tau \langle G(V)(t), W(t) - C(t)\rangle\, dt\\ &= \int_0^\tau \langle R(V)(t), (W(t) \cdot n_t(V))\, n_t(V)\rangle\, dt.\end{aligned}$$
Indeed, $C(t)$ is a tangential vector field, so from the first part of the proof we deduce that $\mathbf{j}'(V; C) = 0$. So under the specified hypotheses there exists a boundary density $g(V)$, $g(V)(t) = R(V)(t) \cdot n_t(V)$, such that
$$G(V)(t) = \gamma^*_{n_t}(g(V)(t))$$
where γ_{n_t} is the normal trace.

Remark

1. More generally if u depends only on the trace, on the lateral boundary $\Sigma(V)$, of the field V and if $R(V)(t) \in \mathcal{D}^{-k+1}(\Gamma_t(V))$ a.e. then $\mathbf{j}'(V; W) =$
$$\int_0^\tau (R(V)(t) \cdot n_t(V))\, (W(t) \cdot n_t(V))\, dt + \int_0^\tau \langle R(V)(t), C(t)\rangle\, dt.$$

The first integral of the right-hand term is the shape derivative (in the direction W) and the second one is exclusively dynamic, that is, due to the variations of the tangential component of W (which does not affect the shape).

2. If the vector function $R(V)$ is such that $R(V)(t) \in L^p(\Gamma_t(V), \mathbb{R}^N)$ $(p \geq 1)$, then under the assumptions of Lemma 6.1 and if the density $g(t) = R(t) \cdot n_t$ is in $L^p(\Gamma_t(V))$, we have an integral representation for the derivative
$$\mathbf{j}'(V; W) = \int_0^\tau \int_{\Gamma_t(V)} g(V)(t)\, W(t) \cdot n_t \, d\Gamma_t \, dt.$$

7 Application to a cylindrical shape optimization

Examples that motivate our work can be found in [18] and [17]. In this section we consider a shape optimization problem associated with a stationary PDE and show that the previous tube analysis could be applied to select, while using the minimization process, the optimal velocity field (regarding some criteria) to go from one domain to another. The original problem is to solve the following minimization problem: $\min_{\Omega \in \mathcal{A}} J(\Omega)$. Using a gradient method, the purpose is to build a domain Ω_{k+1} starting from a given domain Ω_k such that

$$J(\Omega_{k+1}) \leq J(\Omega_k).$$

If $t_k > 0$ is the time step, the vector field V^k that makes the job (i.e., such that $\Omega_{k+1} = T_{t_k}(V^k)\Omega_k$) will be chosen as the solution of

$$\min_V J(T_{t_k}(V)\Omega_k).$$

More generally, the problem can be stated as follows: The domain Ω and $t_0 > 0$ being given, consider the minimization problem

$$\min_V J(\Omega_{t_0}(V)).$$

Under some assumptions, we get the classical expansion of the mapping $t \longrightarrow J(\Omega_t(V))$,

$$J(\Omega_{t_0}(V)) = J(\Omega) + \int_0^{t_0} dJ(\Omega_s(V); V)\, ds, \quad t_0 \in (0, \tau].$$

That expansion can be treated following the previous tube functional approach. Indeed, with the initial domain Ω being given as well as the time $t_0 \in (0, \tau]$, the functional $J(\Omega_{t_0}(V))$ becomes a functional $\mathbf{j}(V)$ of the speed vector field V. We apply our previous field derivative calculus with the use of the transverse field $\mathbf{Z}$ associated with a perturbation field W. Concerning the shape analysis, we adopt the classical terminology introduced in [3] and [22]. There the notions of shape derivative (resp. boundary shape derivative) of distributions defined on a domain (resp. on a boundary or surface) are accurate. For example, the boundary shape derivative of the normal vector field, denoted by $n'_\Gamma(V)$, is given by $n'_\Gamma(V) = -\nabla_\Gamma(\langle V(0), n\rangle)$; see [31].

Assume Ω to be smooth enough—say of class C^k with the integer $k \geq 1$. We know that for any

$$V \in \mathcal{E}_{0,k} \overset{def}{=} C^0([0,\tau], \mathcal{V}_o^k(D))$$

the associated flow, $T(V)$, is in $\mathcal{C}^1([0,\tau], \mathcal{C}^k(\overline{D}, \mathbb{R}^N))$.

Consider a shape functional $J(.)$ defined on a family of domains $\mathcal{A}$ containing the set

$$\{\Omega_s(Y) = T_s(Y)\Omega; \quad \forall s \in [0,\tau], \quad \forall Y \in \mathcal{E}_{0,k}\}.$$

Suppose J is shape differentiable on any $B \in \mathcal{A}$. We denote by $dJ(B;Y)$ the Eulerian semiderivative of J at $B \in \mathcal{A}$ in the direction Y. If B is C^k, there exists a distribution $G(B)$ of finite order, $G(B) \in \mathcal{D}^{k-1}(D, \mathbb{R}^n)'$, such that

$$dJ(B;Y) = \langle G(B), Y\rangle_{\mathcal{D}^{k-1}(D,\mathbb{R}^n)', \mathcal{D}^{k-1}(D,\mathbb{R}^n)}.$$

This distribution is called the shape gradient of J at B.

When $k = 1$, that gradient is a (vector) measure supported by the boundary (cf. [25]). Assume the gradient is smooth enough in time; say $s \to \langle G(\Omega_s(V)),\ Z\rangle$ in $L^1(0,\tau)$, for any $Z \in \mathcal{D}(D, \mathbb{R}^n)$. Then, for $t_0 \in]0,\tau]$, we obtain the following expansion:

$$J(\Omega_{t_0}(V)) = J(\Omega) + \int_0^{t_0} \langle G(\Omega_s(V))\,,\ V(s)\rangle_{\mathcal{D}',\mathcal{D}}\, ds \tag{7.19}$$

Therefore we can define a field functional $\mathbf{j}$ by: $\mathbf{j}(V) = J(\Omega_{t_0}(V))$.

We assume J to be twice shape differentiable at any $B \in \mathcal{A}$. Thus the shape derivative of $G(.)$ exists $\forall B \in \mathcal{A}$ and its derivative at B, in the direction Z, is denoted $G'(B;Z)$. Hence, for $B = \Omega_s(V)$, the mapping $Z \to G'(\Omega_s(V), Z)$ is linear and continuous. Moreover, because $\partial\Omega_s(V)$ is smooth, from the generic argument developed in [2], we derive that $G'(\Omega_s(V), Z)$ will depend only on the normal trace on the boundary of the autonomous field Z. So $G'(\Omega_s(V), Z) = \tilde{G}'(\Omega_s(V)).z$ where $z = \langle Z, n_s(V)\rangle$.

The continuous linear operator $\tilde{G}'(\Omega_s(V))$ $(\in \mathcal{L}(\mathcal{D}(\Gamma_s(V), \mathbb{R}), \mathcal{D}(D, \mathbb{R}^N)'))$ defines the *Shape Hessian* of J at $\Omega_s(V)$ (cf. [11]). Under the previous assumptions, it becomes possible to compute the derivative of $\mathbf{j}$ on V in the direction W:

$$\mathbf{j}'(V;W)=\int_0^{t_0}\langle G'(\Omega_s(V),\mathbf{Z}(s))\,,\,V(s)\rangle_{\mathcal{D}',\mathcal{D}}+\langle G(\Omega_s(V)),W(s)\rangle_{\mathcal{D}',\mathcal{D}}\,ds$$

where $\mathbf{Z}$ is the transverse field introduced in Section 3, Theorem 3.1.

7.1 Derivative with density gradient formulation

We rewrite $\mathbf{j}$ in terms of the density gradient $g(\Gamma_s(V))$ associated with $G(\Omega_s(V))$, using the fact that

$$G(\Omega_s(V))=\gamma^*_{\Gamma_s(V)}(g(\Gamma_s(V))\,n_s(V)), \tag{7.20}$$

so that

$$\mathbf{j}(V)=J(\Omega)+\int_0^{t_0}\int_{\Gamma_s(V)}g(\Gamma_s(V))\,\langle V(s),n_s(V)\rangle\,d\Gamma_s ds.$$

The derivative of $\mathbf{j}$ at V, in the direction W, has a more explicit expression.

Proposition 7.1

Assume

1. *the data are sufficiently smooth*
2. *the shape derivative $g'(\Gamma_s(V);.)$ exists for any $s\in[0,t_0]$.*

Then

$$\begin{aligned}\mathbf{j}'(V;W)=\int_0^{t_0}\int_{\Gamma_s(V)} & g'(\Gamma_s(V);\mathbf{Z}(s))\,\langle V(s),n_s(V)\rangle \\ +\,g(\Gamma_s(V))\,\langle W(s),n_s(V)\rangle + & \ \mathrm{div}_{\Gamma_s(V)}[\,g(\Gamma_s(V))\,V(s)]\,\langle\mathbf{Z}(s),\,n_s(V)\rangle \\ +\,g(\Gamma_s(V))\,\langle DV(s)\cdot n_s(V),\,n_s(V)\ \rangle & \ \langle\mathbf{Z}(s),\ n_s(V)\ \rangle\ d\Gamma_s ds\end{aligned} \tag{7.21}$$

and it depends on W through $\mathbf{Z}$.

Proof

By direct calculation, following derivatives rules and notations ([29], [31]), κ being the mean curvature of the manifold, we get:

$$\begin{aligned}\mathbf{j}'(V;W)=&\int_0^{t_0}\int_{\Gamma_s(V)}g'(\Gamma_s(V);\mathbf{Z}(s))\langle V(s),n_s(V)\rangle+g(\Gamma_s(V))\,\langle W(s),n_s(V)\rangle\\ &+g(\Gamma_s(V))\,\langle V(s),n'_s(V;\mathbf{Z}(s))\rangle+\kappa\,g(\Gamma_s(V))\,\langle V(s),n_s(V)\rangle\,\langle\mathbf{Z}(s),n_s(V)\rangle\\ &+g(\Gamma_s(V))\langle DV(s)\cdot n_s(V),\,n_s(V)\rangle\langle\mathbf{Z}(s),\ n_s(V)\rangle\ d\Gamma_s ds\end{aligned}$$

Now as $n'_s(V;\mathbf{Z}(s))=-\nabla_{\Gamma_s(V)}(\langle\mathbf{Z}(s),n_s(V)\rangle)$, using tangential by part integration we obtain (7.21).

By hypothesis, the following linearity holds:

$$g'(\Gamma_s(V);\mathbf{Z}(s)) = \tilde{g}'(\Gamma_s(V)).\langle \mathbf{Z}(s),\, n_s(V)\rangle$$

because the structure theorem ([22], [25]) is applicable as soon as $Z \longmapsto g'(\Gamma_s(V);Z)$ is linear and continuous with respect to Z, together with the regularity of the boundary. Then the first term of $\mathbf{j}'(V;W)$ can be rewritten as $\displaystyle\int_0^{t_0}\int_{\Gamma_s(V)} g'(\Gamma_s(V);\mathbf{Z}(s))\ \langle V(s), n_s(V)\rangle\, d\Gamma_s ds =$

$$\int_0^{t_0}\int_{\Gamma_s(V)} \tilde{g}'(\Gamma_s(V))^*.\langle V(s),\quad n(\Gamma_s(V))\rangle\,\langle \mathbf{Z}(s),\quad n(\Gamma_s(V))\rangle\, d\Gamma_s ds.$$

Set

$$\begin{aligned} F(s) = {} & \tilde{g}'(\Gamma_s(V))^*.\langle V(s), n(\Gamma_s(V))\rangle + g(\Gamma_s(V))\mathrm{div}V(s) \\ & +\langle \nabla_{\Gamma_s(V)} g(\Gamma_s(V)),\, V(s)\rangle. \end{aligned}$$

Then

$$\mathbf{j}'(V;W) = \int_0^{t_0}\int_{\Gamma_s(V)} F(s)\mathbf{Z}(s)\cdot n(s) + g(\Gamma_s(V))W(s)\cdot n(s)\ d\Gamma ds.$$

Finally we derive the result concerning the necessary optimality condition for $\mathbf{j}(V) = J(\Omega_{t_0}(V))$.

Proposition 7.2

The necessary optimality condition associated with the considered minimization problem is: for a.e. $s \in (0,t_0)$

$$\lambda(s) + g(\Gamma_s(V))\langle W(s), n(s)\rangle = 0 \quad \textit{a.e. in} \quad \Gamma_s(V)$$

where λ solves the backward problem

$$\partial_t\lambda(t) + \nabla_{\Gamma_t}\lambda.V + \lambda\, \mathit{div}V = F(t) \quad \textit{on}\ \cup_{0<t<t_0}(\{t\}\times\Gamma_t), \quad \lambda(t_0) = 0 \quad \textit{on}\ \Gamma_{t_0}.$$

The proof is based on Lemma 6.2. Particularly on the fact that

$$\int_0^{t_0}\int_{\Gamma_t(V)} F(t)\,\mathbf{Z}(t)\cdot n_t\, d\Gamma_t dt = -\int_0^{t_0}\int_{\Gamma_t(V)} \lambda(t)\, W(t)\cdot n_t\ d\Gamma_t dt$$

so that

$$\mathbf{j}'(V;W) = \int_0^{t_0}\int_{\Gamma_s(V)} (\lambda(s) + g(\Gamma_s(V)))\langle W(s), n(s)\rangle\ d\Gamma ds.$$

8 Example

Let Ω be a bounded open domain in D with smooth boundary Γ, and let $y = y(\Omega)$ be the solution of the homogeneous Dirichlet problem: $f \in L^2(D)$,

$$-\Delta y = f \quad \text{in } \Omega, \qquad y = 0 \quad \text{on } \Gamma.$$

Consider the classical compliance encountered in structural mechanics

$$J(\Omega) = \int_\Omega f y \, dx.$$

which is the work of the loading f as minimized with respect to the design variable in order to render the structure more resistant under the given loading f. It turns out that (cf. [23]),

$$\int_\Omega f y \, dx = \int_\Omega |\nabla y|^2 \, dx = -2 \min_{\phi \in H_0^1(\Omega)} \int_\Omega \left(\frac{1}{2} |\nabla \phi|^2 - f\phi \right) dx$$

so that

$$\min_\Omega J(\Omega) = -2 \max_\Omega \min_{\phi \in H_0^1(\Omega)} \int_\Omega \left(\frac{1}{2} |\nabla \phi|^2 - f\phi \right) dx.$$

Classically, we get the expression of the derivative

$$dJ(\Omega; V) = -\frac{1}{2} \int_\Gamma \left| \frac{\partial y}{\partial n} \right|^2 V \cdot n d\Gamma.$$

For a fixed $t_0 > 0$, we define the functional $\mathbf{j}(V) = J(\Omega_{t_0}(V))$, which can be rewritten as follows:

$$\mathbf{j}(V) = J(\Omega) + \int_0^{t_0} dJ(\Omega_t(V), V(t)) \, dt.$$

Thus the expression of $\mathbf{j}$ in terms of the density gradient associated with the shape derivative of J is

$$\mathbf{j}(V) = J(\Omega) - \frac{1}{2} \int_0^{t_0} \int_{\Gamma_t(V)} \left(\frac{\partial y_t}{\partial n_t} \right)^2 \langle V(t), \ n_t \rangle \, d\Gamma dt$$

where y_t is the unique solution of

$$-\Delta y_t = f \quad \text{in } \Omega_t(V), \qquad y_t = 0 \quad \text{on } \Gamma_t(V). \tag{8.22}$$

According to the previous results, the Eulerian derivative of $\mathbf{j}$ in the direction W is given by:

$$\mathbf{j}'(V;W) = -\frac{1}{2}\int_0^{t_0}\int_{\Gamma_t(V)} 2\,\frac{\partial y_t}{\partial n_t}\frac{\partial y_t'}{\partial n_t}\,\langle V(t), n_t\rangle \,+\, \text{div}\,(|\nabla y_t|^2 V)\,\langle \mathbf{Z}(t), n_t\rangle \\ + \left(\frac{\partial y_t}{\partial n_t}\right)^2 \langle W, n_t\rangle\, d\Gamma dt$$

where y_t' is the solution of

$$\Delta y_t' = 0 \quad \text{in } \Omega_t(V), \qquad y_t' = -\frac{\partial y_t}{\partial n_t}\mathbf{Z}(t)\cdot n_t \quad \text{on } \Gamma_t(V).$$

The calculus of $\mathbf{j}'$ is obtained easily using the following expression of $\mathbf{j}$:

$$\mathbf{j}(V) = J(\Omega) - \frac{1}{2}\int_0^{t_0}\int_{\Omega_t(V)} \text{div}(|\nabla y_t|^2 V(t))\,dxdt.$$

Using the R_t solution of

$$\Delta R_t = 0 \quad \text{in } \Omega_t(V) \qquad R_t = \frac{\partial y_t}{\partial n_t}V(t)\cdot n_t \quad \text{on } \Gamma_t(V),$$

we obtain the following expression:

$$\mathbf{j}'(V;W) = -\frac{1}{2}\int_0^{t_0}\int_{\Gamma_t(V)} -2\,\frac{\partial y_t}{\partial n_t}\frac{\partial R_t}{\partial n_t}\langle \mathbf{Z}(t), n(t)\rangle \,+\, \text{div}\,(|\nabla y_t|^2 V)\,\langle \mathbf{Z}(t), n_t\rangle \\ + \left(\frac{\partial y_t}{\partial n_t}\right)^2 \langle W(t), n_t\rangle\, d\Gamma dt.$$

In terms of W:

$$\mathbf{j}'(V,W) = -\frac{1}{2}\int_0^{t_0}\int_{\Gamma_t(V)} \left[\lambda(t) + \left(\frac{\partial y_t}{\partial n_t}\right)^2\right] \langle W(t), n_t\rangle\, d\Gamma_t dt$$

where λ solves the backward problem:

$$\partial_t\lambda(t) + \nabla_{\Gamma_t}\lambda.V + \lambda\,\text{div}V = -2\,\frac{\partial y_t}{\partial n_t}\frac{\partial R_t}{\partial n_t} + \text{div}\,(|\nabla y_t|^2 V) \;\text{ on } \cup_{0<t<t_0}(\{t\}\times\Gamma_t)$$

and $\lambda(t_0) = 0 \quad \text{on } \Gamma_{t_0}$.

Note that only the normal components, on $\Sigma(V)$, of the fields V and W are needed. Also according to Lemma 5.2 and its proof, we have the explicit expression of the solution λ.

The distributed expression of the field derivative is, for any admissible direction W,

$$\mathbf{j}'(V;W) = \int_0^{t_0} \frac{1}{2} \int_D -\Lambda(t).W(t) + \operatorname{div}(|\nabla y_t|^2 W(t))\, dx\, dt$$

where Λ is defined in (5.16) with a right-hand term equal to

$$\gamma^*_{\Gamma_t(V)} \left(\left[-2 \frac{\partial y_t}{\partial n_t} \frac{\partial R_t}{\partial n_t} + \operatorname{div}(|\nabla y_t|^2 V) \right] n_t \right)$$

and y_t is the solution of (8.22) extended to D by $y_t = 0$ in $D \setminus \Omega_t(V)$.

8.1 Existence result and necessary optimality condition

For simplicity we study the existence result in $\mathbb{R}^3$. Assume Ω is of class C^2. Denote by $\mathbf{H}_0^m(D)$ the Hilbert space $\{v \in H^m(D)^3, v \cdot n = 0 \quad \text{on } \partial D\}$ $(m \geq 6)$.

For any $\theta \in L^2(0,t_0;\mathbf{H}_0^m(D))$, consider the velocity field $V_\theta \in H^1(0,t_0;\mathbf{H}_0^m(D))$ defined as follows

$$V_\theta(\mu) = \int_0^\mu \theta(s)ds, \quad \mu \in (0,t_0). \tag{8.23}$$

When there is no confusion, V_θ is denoted V. Denote by $T(V)$ or simply T, the associated transformation.

This section is devoted to the study of the next minimization problem: *A domain* Ω, *and an instant* $t_0 \in [0,t_0]$ *being given, find a solution* θ *of*

$$\min_{\theta \in L^2(0,t_0;\mathbf{H}_0^m(D))} \left\{ \mathbf{j}(V_\theta) + \frac{\alpha}{2} \parallel \theta \parallel^2_{L^2(0,t_0;\mathbf{H}_0^m(D))} \right\}, \tag{8.24}$$

where $\mathbf{j}(V_\theta) = J(T_{t_0}(V_\theta)(\Omega))$ and $\alpha > 0$ arbitrarily small.

Let θ_n $(n \in \mathbb{N})$ and θ in $L^2(0,t_0;\mathbf{H}_0^m(D))$.

Lemma 8.1

Assume that $\theta_n \rightharpoonup \theta$, $n \to \infty$, *weakly in* $L^2(0,t_0;\mathbf{H}_0^m(D))$. *Then* $V_n \to V$ *strongly in* $\mathcal{C}([0,t_0];\mathbf{H}_0^{m'}(D))$, $m' = m - \mu$, $\mu > 0$ *arbitrarily small.*

Proof

Because, for almost every $t \in (0,t_0)$, $V_n(t) = \int_0^t \theta_n(s)ds$, we have

$$\parallel V_n(t) \parallel_{\mathbf{H}_0^m} \leq \int_0^t \parallel \theta_n(s) \parallel_{\mathbf{H}_0^m} ds.$$

By hypothesis, the sequence $\{\theta_n\}$ is bounded in $L^2(0, t_0; \mathbf{H}_0^m(D))$ and a fortiori in $L^1(0, t_0; \mathbf{H}_0^m(D))$. Hence,

$$\| V_n(t) \|_{\mathbf{H}_0^m} \leq C, \;\; C \text{ is a constant.}$$

This means that

$$V_n \text{ is bounded in } L^\infty(0, t_0; \mathbf{H}_0^m(D)). \qquad (8.25)$$

On the other hand, we deduce from equation

$$\frac{d}{dt} V_n(t) = \theta_n(t)$$

the boundedness of $\dfrac{d}{dt} V_n$ in $L^2(0, t_0; \mathbf{H}_0^m(D))$.

Under those boundedness results, one can use a compactness result (see [23]), and get the convergence of

$$V_n \to V \text{ strongly in } \mathcal{C}([0, t_0]; \mathbf{H}_0^{m'}(D)), \qquad (8.26)$$

(the injection $H^m(D) \hookrightarrow H^{m'}(D)$ is compact).

In particular, because $m > 5/2$, we obtain Corollary 8.2.

Corollary 8.2

Assume that $\theta_n \rightharpoonup \theta, n \to \infty$, weakly in $L^2(0, t_0; \mathbf{H}_0^m(D))$. Then, for k such that

$$\frac{1}{2} < \frac{m - (k+1)}{3}, \;\; V_n \to V \; \textit{strongly in } \mathcal{C}([0, t_0]; W^{k,\infty}(D)).$$

Proof

As $1/2 < (m - (k+1))/3$ implies $1/2 < (m' - k))/3$ then in addition to the injection $H^m(D) \hookrightarrow W^{k+1,\infty}(D)$, we have $H^{m'}(D) \hookrightarrow W^{k,\infty}(D)$ and both of them are compact.

Lemma 8.3

Assume that, $n \to \infty$,

$$V_n \to V \; \textit{strongly in } \mathcal{C}([0, t_0]; W^{k,\infty}(D)), k \in \mathbb{N}^*.$$

Then,

$$T(V_n) \; (= T^n) \longrightarrow T(V) \; (= T) \; \textit{strongly in } \mathcal{C}^1([0, t_0]; W^{k,\infty}(D)).$$

Before proving this convergence result, we recall Lemma 8.4.

Lemma 8.4

Let $F \in W^{m,\infty}(D)$, $m \geq 1$, be a homeomorphic transformation such that F^{-1} is Lipschitz-continuous in $\overline{D}$. Then,

$$V \circ F \in W^{m,\infty}(D) \ \ \textit{and} \quad \| V \circ F \|_{W^{m,\infty}} \leq c \| V \|_{W^{m,\infty}} .$$

Proof

(See [13].)

Proof **of Lemma 8.4**

$$\begin{aligned} T_t^n(x) - T_t(x) &= \int_0^t V_n(s, T_s^n(x)) - V(s, T_s(x))ds \\ &= \int_0^t V_n(s, T_s^n(x)) - V(s, T_s^n(x))ds + \int_0^t V(s, T_s^n(x)) \\ &\quad -V(s, T_s(x))ds \end{aligned}$$

Set $r_n(t) = \| T_t^n - T_t \|_{W^{k,\infty}}$ and $R_n(t) = \int_0^t r_n(s)ds$.

Because $V(s)$ belongs to $W^{k+1,\infty}(D) \hookrightarrow \mathcal{C}^1(D)$, $V(s, T_s^n(x)) - V(s, T_s(x)) = DV(s, T_s(x) + \eta(T_s^n(x) - T_s(x)))(T_s^n(x) - T_s(x))$, $\eta = \eta(n, s, x)$, we have

$$r_n(t) \leq c_V \int_0^t \| V_n(s) - V(s) \|_{W^{k,\infty}} ds + K_V R_n(t),$$

$$K_V = \| DV \|_{\mathcal{C}([0,t_0];W^{k,\infty}(D))}$$

$$R_n^{'}(t) - K_V R_n(t) \leq c_V t_0 \| V_n - V \|_{\mathcal{C}([0,t_0];W^{k,\infty}(D))}$$

$$(\exp(-K_V t) R_n(t))' \leq c_V t_0 \exp(-K_V t) \| V_n - V \|_{\mathcal{C}([0,t_0];W^{k,\infty}(D))} .$$

By integration, we get

$$\exp(-K_V t) R_n(t) \leq c_V t_0 \| V_n - V \|_{\mathcal{C}([0,t_0];W^{k,\infty}(D))} \int_0^t \exp(-K_V s)ds$$

$$R_n(t) \leq c_V t_0 \| V_n - V \|_{\mathcal{C}([0,t_0];W^{k,\infty}(D))} \frac{\exp(K_V t) - 1}{K_V}.$$

Then,

$$\begin{aligned} r_n(t) &\leq t_0 \| V_n - V \|_{\mathcal{C}([0,t_0];W^{k,\infty}(D))} + t_0 \| V_n - V \|_{\mathcal{C}([0,t_0];W^{k,\infty}(D))} \\ &\qquad (\exp(K_V t) - 1) \\ &\leq t_0 \| V_n - V \|_{\mathcal{C}([0,t_0];W^{k,\infty}(D))} \exp(K_V t_0). \end{aligned}$$

This last estimation gives the convergence of T^n, strongly in $\mathcal{C}([0,t_0];W^{k,\infty}(D))$, to T. Moreover, because T^n and T are in $\mathcal{C}^1([0,t_0];W^{k,\infty}(D))$ and

$$\frac{d}{dt}T^n = V_n \circ T^n$$

we deduce

$$T^n \longrightarrow T,\ n \to \infty,\ \text{ strongly in }\ \mathcal{C}^1([0,t_0];W^{k,\infty}(D)). \tag{8.27}$$

In the sequel m will be sufficiently large to allow $k \geq 3$.

For an existence result we shall study the continuity of the solution y with respect to the field. Let $y_n(t)$, $n \in \mathbb{N}^*$, the unique solution of

$$\begin{aligned} -\Delta y_n(t) = f_n(t) \quad &\text{in} \quad L^2(\Omega_n(t)) \\ y_n(t) = 0 \quad &\text{in} \quad \Gamma_n(t). \end{aligned}$$

where $f_n(t) = f_{|\Omega_n(t)}$ and

$$\left| \begin{array}{l} T_t^n(\Omega) = \Omega_n(t), \quad T_t^n(\Gamma) = \Gamma_n(t);\ t \in (0,t_0) \\ T_0^n(\Omega) = \Omega. \end{array} \right.$$

Lemma 8.6

Assume that, $n \to \infty$,

$$V_n \longrightarrow V \quad \textit{strongly in}\ \ \mathcal{C}([0,t_0];\mathbf{H}_0^{m'}(D)).$$

Then,

$$y_n(t) \rightharpoonup y(t), \quad \textit{weakly in}\ H_0^1(\Omega_t(V)), \tag{8.28}$$

where $y(t)$ is the unique solution of (8.22).

Proof

By composition we obtain, $\forall \varphi \in H_0^1(\Omega)$,

$$\int_\Omega \langle A(t)\nabla(y_n \circ T_t^n - y \circ T_t), \nabla\varphi\rangle\, dx$$

$$= \int_\Omega (j_n(t) f \circ T_t^n - j(t) f \circ T_t)\varphi\, dx - \int_\Omega \langle (A_n(t) - A(t))\nabla y_n \circ T_t^n, \nabla\varphi\rangle\, dx$$

where

$$j(t) = |\text{det} DT_t(V)|, \quad A(t) = j(t) DT_t(V)^{-1}.{}^*DT_t(V)^{-1}$$

and

$$j_n(t) = |\det DT_t(V_n)|, \quad A_n(t) = j_n(t) DT_t(V_n)^{-1}. {}^*DT_t(V_n)^{-1}.$$

The assumption implies the convergence $T^n \longrightarrow T$ strongly in $C^1([0,t_0]; W^{k,\infty}(D))$. Thus we deduce the convergence

$$\begin{array}{ll} DT^n \longrightarrow DT & \text{in } C^1([0,t_0]; W^{k-1,\infty}(D)) \\ j_n \longrightarrow j & \text{in } C^1([0,t_0]; W^{k-1,\infty}(D)) \\ A_n(t) \longrightarrow A(t) & \text{in } C^1([0,t_0]; W^{k-1,\infty}(D)) \\ f \circ T^n \longrightarrow f \circ T & \qquad \text{in } L^2(D). \end{array}$$

The convergence of $(y_n \circ T(V_n))$ to $y \circ T(V)$ weakly in $H^1_0(\Omega)$ follows immediately involving the desired convergence for (y_n).

Eventually we obtain Proposition 8.6.

Proposition 8.7

There exists a vector field $\theta_{op} \in L^2((0,t_0), \mathbf{H}^m_0(D))$, the solution of the minimization problem (8.24).

We turn our attention to the first-order necessary optimality condition associated with the minimization problem (8.24). For that purpose we need to compute the Gâteaux derivative of the mapping

$$E : \theta \longrightarrow \mathbf{j}(V_\theta) + \frac{\alpha}{2} \parallel \theta \parallel^2_{L^2(0,t_0;\mathbf{H}^m_0(D))}$$

$dE(\theta;\xi) = \mathbf{j}'(V_\theta; V_\xi) + \alpha \ \langle \theta, \xi \rangle_{L^2((0,t_0),\mathbf{H}^m_0(D))}$.

Making use of the expression of $\mathbf{j}'$ computed in the previous section, the first-order optimality condition associated with problem (8.24) can be stated as follows.

Proposition 8.8

The optimal field θ_{op}, of (8.24), satisfies: $\forall \xi \in L^2((0,t_0), \mathbf{H}^m_0(D))$

$$\frac{1}{2}\int_0^{t_0}\int_D -\Lambda(t)\cdot V_\xi(t) + \operatorname{div}\left(|\nabla y_t|^2\, V_\xi(t)\right) dxdt + \alpha \int_0^{t_0} \langle \theta_{op}, \xi \rangle_{\mathbf{H}^m_0(D)}\, dxdt = 0$$

where $V_{op} = V_{\theta_op}$, $y_t = y(\Omega_t(V_{op}))$ is solution of (8.22) considered in the open set $\Omega_t(V_{op})$ and Λ is the solution of

$$-\partial_t \Lambda - D\Lambda \cdot V_{op} - {}^*DV_{op} \cdot \Lambda - \operatorname{div} V_{op}\, \Lambda = \gamma^*_{\Gamma_t(V_{op})}\,(h(t) n_t) \quad \text{in } (0,t_0) \times D$$
$$\Lambda(t_0) = 0 \text{ in } D,$$

with $h(t) = -2\,\dfrac{\partial y_t}{\partial n_t}\dfrac{\partial R_t}{\partial n_t} + \operatorname{div}\ (|\nabla y_t|^2 V_{op})$.

Thanks to this optimality condition it is possible, under regularity assumptions, to select the best velocity to build a domain starting from a given one, while using the shape minimization process.

References

[1] Zolésio, J.-P, *Variational formulation for Euler équations*, eds. Hoffman and Leugring, Proc. conf. IFIP Chemnitz, 1998.

[2] Zolésio, J.-P., *Introduction to shape optimization problems and free boundary problems*, Nato Adv. Sci. Inst. Kluwer Acad. Publ., Montreal, 1992, 397–457.

[3] Zolésio, J.-P., *The material derivative (or speed) method for shape optimization*, Nato Adv. Study Ser. E, E. J. Haug and J. Céa, eds., Nijhoff, The Hague, 1981, 1089–1151.

[4] Rudin, W., *Real and complex analysis*, McGraw-Hill, New York, 1966.

[5] Weinberger, H. F., *Variational methods for eigenvalue approximation*, SIAM, Philadelphia, 1974.

[6] Chen, G. and Zhou, J., *Vibration and damping in distributed systems*, vol. II, CRC Press, Boca Raton, FL, 1993.

[7] Lemaire, B., *Thèse de doctorat d'état*, Faculté des sciences de Paris, 1970.

[8] Banks, H. T., Smith, R. C., and Wang, Y., *Smart material structures: modeling, estimation, and control*, Research in applied mathematics, Wiley, New York, 1996.

[9] Delfour, M.-C. and Zolésio, J.-P., Structure of shape derivatives for non smooth domains, *Journal of Functional Analysis*, 1992, 104.

[10] Cannarsa, C., Da Prato, G., and Zolésio, J.-P., The damped wave equation in a moving domain, *Journal of Differential Equations*, 85, 1990, 1–16.

[11] Bucur, D and Zolésio, J.-P., Anatomy of the shape Hessian via Lie brackets, *Ann. Mat. Pura Appl.*, 1997, 173, 4, 127–143.

[12] Delfour, M. C. and Zolésio, J.-P., Shape analysis via oriented distance functions, *J. Funct. Anal.*, 1994, 123, 1, 1–16.

[13] Necas, J., *Les méthodes directes en théorie des équations elliptiques*, Masson, Paris, 1967.

[14] Aubin, J.-P., *Mutational and morphological analysis: tools for shape evolution and morphogenesis*, Birkhäuser, Boston, 1998.

[15] Delfour, M.-C. and Zolésio, J.-P., Shape optimization, INRIA-Sophia Antipolis, Comett Matari Programme, Mathematical Toolkit for artificial intelligence and regulation of Macro-systems, 1993.

[16] Dziri, R. and Zolésio, J.-P., Shape existence in Navier-Stokes flow with heat convection, *Ann. della Scuola Normale di Pisa*, 1997, XXIV, IV, 165–192.

[17] Dziri, R. and Zolésio, J.-P., Dynamical shape control in non-cylindrical hydrodynamics, *J. Inverse Probl.*, 1999, 15, 1, 113–122.

[18] Dziri, R. and Zolésio, J.-P., Dynamical shape control in non-cylindrical Navier-Stokes equations, *J. Convexe Anal.*, 1999, 123–153.

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

[20] Lions, J.-L., *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Springer-Verlag, Berlin, 1971.

[21] Correa, R. and Seeger, A., Directional derivative of a minmax function, *J. Nonlinear Anal.*, 1985, 9, 13–22.

[22] Sokolowski, J. and Zolésio, J.-P., *Introduction to shape optimization*, Springer-Verlag, Berlin, SCM 16, 1992.

[23] Temam, R. *Theory and numerical analysis of the Navier-Stokes equations*, North-Holland, Amsterdam, 1977.

[24] Lions, J.-L. and Magenes, E., *Problèmes aux limites non homogènes*, Dunod, Paris, 1968.

[25] Zolésio, J.-P., *Identification de domaines par déformations*, Université de Nice, France, 1979, Thèse de Doctorat d'état.

[26] Zolésio, J.-P., Semi-derivatives of repeated eigenvalues, *Distributed parameter structures*, E. J. Haug and J. Céa, Eds., Nijhoff, The Hague, 1981, 1457–1473.

[27] Myslinski, A., Bimodal optimal design of vibrating plates using theory and methods of nondifferentiable optimization, *J. Optim. Theory Appl.*, 1985, 46, 2, 187–203.

[28] Delfour, M. C. and Zolésio, J.-P., Shape sensitivity analysis via min max differentiability, *SIAM J. Control Optim.*, 1988, 26, 4, 834–862.

[29] Delfour, M. C. and Zolésio, J.-P., *Shapes and geometries*, Advances in Design and Control, 4, SIAM, Philadelphia, 2001.

[30] Ekeland, I. and Temam, R., *Convex analysis and variational problems*, North-Holland, Amsterdam, 1976.

[31] Desaint, F. R. and Zolésio, J.-P., J. Manifold derivative in the Laplace-Beltrami equation, *Funct. Anal.*, 1997, 151, 1, 234–269.

[32] Zolésio, J.-P., *Weak form of the shape differential equation*, Proc. of the AFORS Conf., September 1997, Washington, J. Burns, ed., Birkhäuser, Basel, 1997.

[33] Dziri, R. and Zolésio, J.-P., Interface minimisant l'énergie dans un écoulement stationnaire de Navier-Stokes, *C.R. Acad. Sci. Paris*, t. 324, 12, 1997.

[34] Da Prato, G. and Zolésio, J.-P., Dynamical programming for non cylindrical parabolic equation, *Sys. Control Lett.*, 1988, 11.

[35] Da Prato, G. and Zolésio, J.-P., Existence and optimal control for wave equation in moving domain, *Stabilization of flexible structures*, Lecture Notes in Control and Information Sciences, 147, Springer-Verlag, Berlin, 1990, 167–190.

CHAPTER 24

A stochastic Riccati equation for a hyperbolic-like system with point and boundary control

Cavit Hafizoglu
Department of Mathematics, University of Virginia, Charlottesville, Virginia

1 Introduction

The goal of this paper is to present a full optimal feedback synthesis along with well-posedness of the stochastic Riccati equation for an infinite dimensional stochastic linear quadratic control problem with unbounded control operators. The finite dimensional counterpart of this problem is well known in the literature; see, for instance, [7] and [15]. The infinite dimensional problem has been studied in the literature within the context of analytic semigroups e^{At} and the corresponding stochastic processes driven by relatively (with respect to A) bounded control operators B [4,6]. Our aim is to extend the theory presented in [4] and [6] to a larger class of dynamics that are not analytic.

More specifically, we shall consider a class of control problems whose dynamics is governed by C_0-semigroups e^{At} defined on a Hilbert space $\mathcal{H}$ along with unbounded control operators $B : U \to [D(A^*)]'$, for which the pair (A, B) satisfies the so-called *singular estimate*. By this we mean that the following estimate is valid $|e^{At}B|_{\mathcal{L}(U,\mathcal{H})} \leq Ct^{-\gamma}$ for some $\gamma \in [0, 1)$. Such systems are referred to as systems with singular estimates.

It is clear that a singular estimate is always satisfied when (i) the operator B is bounded or when (ii) the semigroup e^{At} is analytic and the operator B is relatively bounded, that is, $R(\lambda, A)^{-\theta}B \in \mathcal{L}(U, H)$ for some $\theta < 1$ and $\lambda \in \rho(A)$. Thus, our framework is a strict generalization and extension of results obtained so far in the literature [4,6].

On the other hand, systems with singular estimates arise frequently in the context of coupled structures that contain parabolic and hyperbolic components in their dynamics (fluid structure interactions, structural acoustic interactions, composite materials, etc.). In such mixed configurations, it often happens that thesmoothing effect of parabolicity is propagated onto the entire system leading to a singular estimate, though the entire system may still have a predominantly hyperbolic character. Strong physical motivations and an abundance of control applications have attracted considerable attention to singular estimate control systems [2]. In fact, a complete Riccati theory has been developed for this class of problems within a *deterministic* framework [9,10]. The main technical difficulty in the analysis of these systems is caused by the unboundedness of the control operator along with the lack of smoothing generated by the analyticity of the semigroup e^{At}. In the deterministic case, this difficulty was overcome by using a variational approach whose starting point is an explicit representation of optimal quantities, including the Riccati operator.

Unfortunately, such explicit formulas are not available in the stochastic framework. Thus, the analysis of stochastic singular estimate control systems and of the corresponding Riccati feedback synthesis—a main goal of this work—meets with a new set of technical difficulties to be overcome. These will be discussed later, after we introduce our main results.

2 Definitions and main results

Let $\mathcal{U}$, $\mathcal{H}$ and $\mathcal{Z}$ be real separable Hilbert spaces, $(\Omega, \mathcal{F}, P)$ a complete probability space, and W_t a real-valued Brownian motion defined on $(\Omega, \mathcal{F}, P)$. Let $\mathcal{F}_t$ be the right continuous σ-algebra generated by the Wiener process W_t. Let $R \in \mathcal{L}(\mathcal{H}, \mathcal{Z})$ be the linear bounded observation operator, $A : D(A) \subset \mathcal{H} \to \mathcal{H}$ an unbounded linear operator that generates a strongly continuous semigroup on $\mathcal{H}$, and let the linear operator $B : \mathcal{U} \to [D(A^*)]'$ satisfy the following condition:

$$|R(\lambda, A)Bu| \leq C_{Re\lambda}|u|_{\mathcal{U}}, \quad \lambda \in \rho(A) \tag{2.1}$$

Here the duality is considered with respect to the pivot space $\mathcal{H}$, $R(\lambda, A)$ is the resolvent of A, and $\rho(A)$ is the resolvent set of A. We assume $0 \in \rho(A)$ for clarity. The noise operator $C \in \mathcal{L}(\mathcal{H})$, where the symbol $\mathcal{L}$ denotes the linear and bounded operators.

We shall study the problem under the following setup:

2.1 The singular estimate assumption

B is subject to the singular estimate

$$|e^{At}Bu|_H \leq c\frac{e^{\alpha t}}{t^\gamma}|u|_{\mathcal{U}}, \quad 0 < \gamma < 1 \tag{2.2}$$

This estimate is trivially satisfied if $B \in \mathcal{L}(\mathcal{U}, \mathcal{H})$. It is also satisfied for analytic semigroups e^{At} and relatively bounded control operators

$B:\mathcal{U} \to [D(A^\gamma)]'$, that is, $A^{-\gamma}B \in \mathcal{L}(\mathcal{U},\mathcal{H})$; this follows from the estimate $|A^\gamma e^{At}|_{\mathcal{L}(\mathcal{H})} \le \frac{C}{t^\gamma}$ for the analytic semigroups, [12].

Remark

It should be noted that a first prototype of a singular estimate control system was a Dirichlet boundary control system associated with a heat equation [14]. In fact, in [14] it was shown that the singularity for that system is equal to $\gamma = 3/4 + \varepsilon$, a bound that is referred to in the literature as the Balakrishnan-Washburn bound. The above model is a special case of a control system governed by an analytic semigroup and a relatively bounded control operator B.

Denote by $L_2([0,T];L_2(\Omega,\mathcal{U}))$ the set of all square integrable $\mathcal{F}_t$-predictable $\mathcal{U}$-valued stochastic processes u with the norm

$$|u|_{L_2([0,T];L_2(\Omega,\mathcal{U}))} = \left(\int_0^T E(|u(t)|^2_{\mathcal{U}})dt\right)^{1/2}$$

and let $C([0,T];L_2(\Omega,\mathcal{H}))$ be the set of all $\mathcal{F}_t$-predictable continuous $\mathcal{H}$-valued stochastic processes y endowed with the norm

$$|y|_{C([0,T];L_2(\Omega,\mathcal{H}))} = \sup_{t\in[0,T]} \left(E(|y(t)|^2_{\mathcal{H}})\right)^{1/2}.$$

Denote the linear continuous symmetric positive definite operators from $\mathcal{H}$ to $\mathcal{H}$ *by* $\Sigma^+(\mathcal{H})$.

A one-dimensional Wiener process will be used in order to simplify the exposition. Extension to the case of the infinite dimensional Wiener processes having trace-class covariance operators is fairly transparent [5].

We consider a control problem given by the state equation

$$dy = (Ay + Bu)dt + CydW_t \tag{2.3}$$

$$y(0) = y_0 \in \mathcal{H} \tag{2.4}$$

and the performance index

$$\text{minimize } J(u, y(y_0,u)) \equiv E\left(\int_0^T |Ry(t)|^2_{\mathcal{Z}} + |u(t)|^2_{\mathcal{U}}\, dt\right)$$

over all $u \in L_2(0,T;L_2(\Omega,\mathcal{H}))$.

The optimal trajectory and optimal control will be denoted by y° and u°, respectively.

The main result for the finite horizon problem under assumptions (2.1) and (2.2) is the following:

Theorem 2.1

There exists an operator

$$P \in C([0,T]; \Sigma^+(\mathcal{H}))$$

such that

$$B^* P \in C([0,T]; \mathcal{L}(\mathcal{H}))$$

and P is the unique solution to the following differential Riccati equation:

$$\frac{d}{dt}\langle Px, y\rangle + \langle A^* Px, y\rangle + \langle PAx, y\rangle + \langle C^* PCx, y\rangle \\ + \langle R^* Rx, y\rangle = \langle B^* Px, B^* Py\rangle \tag{2.5}$$

for all $x, y \in D(A)$ with the initial condition $P(T) = 0$. Moreover the value function is

$$\min_{u \in L_2([0,T]; L_2(\Omega, \mathcal{U}))} J(u, y(y_0, u)) = J(u^\circ, y^\circ(y_0, u^\circ)) = \langle P(0) y_0, y_0\rangle_{\mathcal{H}}$$

for all $y_0 \in \mathcal{H}$ and the optimal synthesis takes the form

$$u^\circ(t) = -B^* P(t) y^\circ(t) \ , \ T \geq 0$$

Remark

This result generalizes the results of studies [4] and [6], which are confined to analytic semigroups. In addition to these references, the stochastic boundary control problem for hyperbolic dynamics with a terminal cost condition is investigated in [13] by solving the dual stochastic Riccati equation under the following trace regularity assumption, $\int_0^T |B^* e^{A^* t} x|^2 dt \leq c_T |x|^2$. However, the dual stochastic Riccati equation is a version of the Riccati equation where the nonlinear term does not involve the operator B but involves the bounded observation operator R and the bounded noise operator C. Because in our case the stochastic Riccati equation is solved directly, the main difficulty is the unboundedness of B.

As already stated in the introduction, the deterministic version of Theorem 2.1 can be found in [9] and [10]. For the stochastic dynamics driven by *analytic semigroups*, the conclusion of Theorem 2.1 is given in [4] and [6].

The main difficulty in both cases is how to deal with a priori unboundedness of the feedback operator $B^* P(t)$. In the analytic case, this is dealt with by exploiting strong smoothing of the analytic dynamics; thus, analyticity of the underlying semigroup is an essential tool of analysis. In the nonanalytic but deterministic case, the key and the starting point of the approach is the variational formulation of the optimal solution. Unfortunately, this approach is not available in the stochastic setting. The presence of the noise term causes rather complicated structure of the adjoint equation, which cannot be used effectively to carry out the analysis of singularities caused by the unboundedness of control operator B.

By necessity, our proof of Theorem 2.1 relies on a new approach that involves specially constructed fixed-point techniques applied to a *system of three equations* defining the feedback operator, the Riccati operator, and the expected value of the stochastic evolution. This step provides local solutions to the deterministic quantities. The passage from local to global solutions is obtained via a careful limit process on Ito's formula, which involves approximation of the noise rather than approximations of operators A and B as is done in the literature [4–6].

In fact, the classical [4–6] Yosida approximations of control processes run into technical difficulties due to the combined unboundedness (in fact, uncloseability) of the operator B and lack of analyticity of the semigroup, which would provide for the smoothing effect transferred on $P(t)$. This latter property is not the case for the singular estimate class of control problems and therefore the limit process on the quantity $B^*P(t)$ cannot be carried out. We overcome the problem by introducing a suitable approximation to the multiplicative noise operator C.

3 An example: The stochastic structural acoustic equation with point/boundary control

This example illustrates the theory presented. We consider the interaction between a wave equation and a dynamic plate equation through an interface consisting of a portion of the boundary. Let $G \subset R^3$ be a bounded, sufficiently smooth domain with boundary Γ consisting of two pieces $\overline{\Gamma_0}$ and $\overline{\Gamma_1}$, for $G \subset R^2$. We assume Γ_0 is flat and $\Gamma_0 \cap \Gamma_1 = \emptyset$. In applications, Γ_0 is a flexible wall and Γ_1 is a hard wall of the acoustic chamber Γ.

$$\begin{aligned}
&d(z_t) = c^2\Delta z dt + (a\nabla z + C_{22}z_t + C_{23}v + C_{24}v_t)dW_t, \quad \text{in } G \times (0,T)\\
&\frac{\partial}{\partial\nu}z + z = 0, \quad \text{on } \Gamma_1 \times (0,T)\\
&\frac{\partial}{\partial\nu}z = v_t, \quad \text{on } \Gamma_0 \times (0,T)\\
&d(v_t) + \Delta^2 v dt + \Delta^2 v_t dt + z_t|_{\Gamma_0} dt + (C_{41}z + b\nabla v + C_{44}v_t)\\
&\quad dW_t = \mathcal{B}u dt, \quad \text{in } \Gamma_0 \times (0,T)\\
&v = \frac{\partial}{\partial\nu}v = 0, \quad \text{on } \partial\Gamma_0 \times (0,T)
\end{aligned} \tag{3.6}$$

and the initial data

$$z(0,\cdot) = z_0, \quad z_t(0,\cdot) = z_1, \quad v(0,\cdot) = v_0, \quad v_t(0,\cdot) = v_1$$

The acoustic medium inside the chamber is described by z. In the equation of z, c^2 is the speed of sound. Here, the equation of v is the plate equation with the static damping through the interface Γ_0. The multiplicative noise is described by the bounded linear operators C_{ij}, for $i = 1, \cdots, 4$ and $j = 1, \cdots, 4$. The functions $a : \bar{G} \to \mathbb{R}$ and $b : \bar{\Gamma_0} \to \mathbb{R}$ are differentiable. Define the bounded linear operator $C = (C_{ij})$, for $i = 1, \cdots, 4$ and $j = 1, \cdots, 4$, by $C_{21} = a\nabla \in \mathcal{L}(H^1(G) \to L_2(G))$, $C_{22} \in \mathcal{L}(L_2(G))$, $C_{23} \in \mathcal{L}(L_2(\Gamma_0) \to L_2(G))$, $C_{24} \in \mathcal{L}(H^1(\Gamma_0) \to L_2(G))$, $C_{41} \in \mathcal{L}(H^1(G) \to L_2(\Gamma_0))$, $C_{43} = b\nabla \in \mathcal{L}(H^1(\Gamma_0) \to L_2(\Gamma_0))$, $C_{44} \in \mathcal{L}(L_2(\Gamma_0))$ and $C_{ij} = 0$ for all other indices. It can easily be seen that the noise operator C, defined as above, is a bounded linear operator.

We also assume that

$$\mathcal{B} : \mathcal{U} \to \mathcal{H}^{-2+\varepsilon}(G) \text{ is a bounded linear operator.} \tag{3.7}$$

Remark
There are several applications of piezoelectric actuators that satisfy the conditions above (see [2]).

The stochastic performance criteria is

$$J(u, z, w) = E \int_0^T \{|\nabla z|^2_{0,G} + |z_t|^2_{0,G} + |\Delta v|^2_{0,\Gamma_0} + |v_t|^2_{0,\Gamma_0} + |u(t)|^2_U\} dt$$

subject to the dynamics of (3.6).

To describe the problem in the semigroup formulation, we set $y(t) = [z(t), z_t(t), v(t), v_t(t)]$. And we define a set of operators below:

Let $A_R : L_2(G) \supset \mathcal{D}(A_R) \to L_2(G)$ be the nonnegative, self-adjoint operator defined by

$$A_R h = -c^2 \Delta h, \ \mathcal{D}(A_R) = \left\{ h \in H^2(G) \, : \, \left(\frac{\partial}{\partial \nu} h + h|_\Gamma = 0 \right\} .$$

Let N be the Nuemann-Robin map from $L_2(\Gamma_0)$ to $L_2(G)$, defined by

$$\psi = Ng \iff \left\{ \Delta \psi = 0 \text{ in } G \, : \, \frac{\partial}{\partial \nu} \psi|_{\Gamma_0} = g \left(\frac{\partial}{\partial \nu} \psi + d_1 \psi)|_{\Gamma_1} = 0 \right\} .$$

N is continuous from $L_2(\Gamma_0)$ to $H^{3/2}(G) \subset \mathcal{D}(A_R^{3/4-\varepsilon})$ for all $\varepsilon > 0$; see [11]. By Green's second theorem, the following trace result holds true with any $h \in \mathcal{D}(A_N)$:

$$N^* A_R h = \begin{cases} h|_{\Gamma_0} & \text{on } \Gamma_0 \\ 0 & \text{on } \Gamma_1 \end{cases} .$$

Because $N^* A_R$ is bounded on $\mathcal{D}(A_R^{1/2})$ and $\mathcal{D}(A_R^{1/2})$ is dense in $\mathcal{D}(A_R)$, $N^* A_R$ extends to $H^1(G) \equiv \mathcal{D}(A_R^{1/2})$.

Define $\mathcal{A} : \mathcal{D}(\mathcal{A}) \subset L_2(\Gamma_0) \to L_2(\Gamma_0)$ by $\mathcal{A}g = \Delta^2 g$ with the clamped boundary condition $\mathcal{D}(\mathcal{A}) = \{g \in H^2(\Gamma_0) : g|_{\partial\Gamma_0} = \frac{\partial}{\partial\nu} g|_{\partial\Gamma_0} = 0\}$.

By using the above definitions, the second-order abstract model can be written as

$$dz_t + A_R z dt + d_0 z_t dt - A_R N v_t dt = (a\nabla z + C_{22} z_t + C_{23} v + C_{24} v_t) dW_t$$

on $[\mathcal{D}(A_R)]'$ and

$$dv_t + \mathcal{A} v dt + \mathcal{A} v_t dt + N^* A_R z_t dt = \mathcal{B} u dt + (C_{41} z + b\nabla v + C_{44} v_t) dW_t$$

on $[\mathcal{D}(\mathcal{A})]'$.

Define $H_z \equiv \mathcal{D}(A_R^{1/2}) \times L_2(G)$, $H_v \equiv \mathcal{D}(\mathcal{A}^{1/2}) \times L_2(\Gamma_0)$. On H_z, define $A_z : H_z \to H_z$ by

$$A_z = \begin{pmatrix} 0 & I \\ -A_R & -d_0 \end{pmatrix}.$$

Similarly, on H_v define $A_v : H_v \to H_v$ by

$$A_v = \begin{pmatrix} 0 & I \\ -\mathcal{A} & -\mathcal{A} \end{pmatrix},$$

and define the trace operator $\mathcal{T} : H_z \supset \mathcal{D}(\mathcal{T}) \to H_v$ by

$$\mathcal{T} = \begin{pmatrix} 0 & 0 \\ 0 & N^* A_R \end{pmatrix}$$

$$\mathcal{D}(\mathcal{T}) = \{(z_1, z_2) \in H_z \,:\, N^* A_R z_2 = z_2|_{\Gamma_0} \in L_2(\Gamma_0)\} \supset \mathcal{D}(A_R^{1/2}) \times \mathcal{D}(A_R^{1/4+\varepsilon}).$$

Finally, define the dynamics operator A acting on $H_z \times H_v$ by

$$A = \begin{pmatrix} A_z & \mathcal{T}^* \\ -\mathcal{T} & A_v \end{pmatrix}$$

with the domain

$$\mathcal{D}(A) = \{[z_1, z_2, v_1, v_2] \in H : \, z_2 \in \mathcal{D}(A_R^{1/2}), \quad v_2 \in \mathcal{D}(\mathcal{A}), \\ \mathcal{A}^{1/2}(v_1 + v_2) \in \mathcal{D}(\mathcal{A}^{1/2}), \quad z_1 - N v_2 \in \mathcal{D}(A_R)\}.$$

By the Lummer Phillips theorem it is known that A generates a strongly continuous semi-group of contractions on H. Define the control operator $B : U \to [\mathcal{D}(A^*)]'$; here the duality is with respect to H, by setting $B = [0, 0, 0, \mathcal{B}]^T$. The assumption (3.7) implies that the singular estimate assumption for (A, B) is satisfied with $\gamma = 1/2 - \varepsilon/4$. The proof of the singular estimate for the above system can be found in [1] and [3].

Finally, the deterministic part of the equation becomes

$$\frac{d}{dt}y = Ay + Bu \ \text{ in } [\mathcal{D}(A^*)]', \ \ y(0) = y_0.$$

Then, the stochastic structural acoustic equation can be written as

$$dy = (Ay + Bu)dt + Cydw_t.$$

Therefore, the abstract framework of this study applies to the stochastic structural acoustic equation with point and boundary control.

4 Sketch of the proof

A fixed-point problem, consisting of three coupled equations, is set up to prove the local existence, uniqueness and continuity of the Riccati operator, the expectation of the evolution operator and the gain operator, (P, Φ, B^*P), respectively (see [8]).

4.1 Expected evolution operator Φ

Differentiability of Φ in the first variable is easily obtained by using the singular estimate assumption and assumption (2.1). As a consequence, the expected evolution operator Φ is the solution of the equation

$$\frac{d}{dt}\left(\Phi(t,s)x, y\right) = \left(\left(A - BB^*P(t)\right)\Phi(t,s)x, y\right) \tag{4.8}$$

satisfying the initial condition $\Phi(s,s) = I$, for $x \in \mathcal{H}$, $y \in D(A^*)$. It should be noted that $\Phi(t,0)x$ represents the expected value of the optimal evolution initiating from point x. Although the structure of the evolution operator is the same as the deterministic problem, in fact it is different. This is because P solves the stochastic Riccati equation, not the deterministic Riccati equation.

The evolution operator Φ is the key ingredient in obtaining results about the stochastic Riccati equation. Therefore, we discuss some of its properties below.

4.2 Singular estimate for the expected evolution operator Φ

Lemma 4.1

For $T_{max} \leq s \leq t \leq T$

$$|\Phi(t,s)B|_{L(\mathcal{U}\to\mathcal{H})} \leq C_T \frac{1}{(t-s)^{\gamma}}$$

The proof mainly involves the equality

$$\Phi(t,s) = \left(I + LL^* \left[R^*R + C^*P(\cdot)C\right]\right)^{-1} e^{A(t-s)}$$

where L is the control-to-state operator. On a Banach space X define

$$C_r([s,T];X) = \{f \in C((s,T];X) \, : \, \sup_{s \le t \le T} |t-s|^r |f(t)|_X < \infty\}.$$

Similar to the deterministic case, the invertibility of

$$[I + L^*(R^*R + C^*P(\cdot)C)L]$$

in $C_r([s,T];\mathcal{U})$ provides the claimed estimate.

4.3 Differentiability of the evolution Φ

Differentiability of Φ in the first variable is established as described above. Differentiability in the second variable is not straightforward. Let $x \in D(A)$

$$\begin{aligned}
\Phi(t,s)x &= e^{A(t-s)}x - LB^*P(\cdot)\Phi(\cdot,s)x \\
\Phi(t,s)x &= [I + LB^*P(\cdot)]^{-1} e^{A(\cdot - s)}x \\
\frac{d}{ds}\Phi(t,s)x &= -Ae^{A(t-s)}x + e^{A(t-s)}BB^*P(s)x + \\
&\quad - \int_s^t e^{A(t-\tau)}BB^*P(\tau)\frac{d}{ds}\Phi(\tau,s)x d\tau
\end{aligned}$$

For $t - s > 0$ small, $I + LB^*P$ is invertible in C_γ. Therefore,

$$\frac{d}{ds}\Phi(t,s)x = -[I + LB^*P(\cdot)]^{-1} e^{A(t-s)}[A - BB^*P(s)]x,$$

by using the bound obtained in the previous lemma for $\Phi(t,s)B$,

$$\begin{aligned}
&= -\Phi(t,s)Ax + \Phi(t,s)BB^*Px \\
&= -\Phi(t,s)[A - BB^*P(s)]x
\end{aligned}$$

Finally, the derivative of the expected optimal evolution operator Φ with respect to its second variable is

$$\begin{aligned}
\frac{d}{ds}(\Phi(t,s)x, y) &= -\left(\Phi(t,s)\left(A - BB^*P(s)\right)x, y\right), \quad x \in D(A), \\
\Phi(s,s) &= I, \quad \text{for } s > T_{max}.
\end{aligned}$$

This local result can be extended to the maximal interval on which the triple (P, Φ, B^*P) uniquely exists. As a consequence of the above results, P satisfies the Riccati equation locally in time.

4.4 Global solvability

The optimality is needed to obtain the global bound on P. Optimality follows from the dynamic programming argument. Because the state equation doesn't have strong solutions, the dynamic programming needs a rather special regularity lemma in order to justify the application of Ito's lemma in this unbounded framework.

4.5 The regularity lemma

Lemma 4.2

Assume that the following holds together with the singular estimate assumption
$B^*P(\cdot) \in L(\mathcal{H} \to C([0,T];\mathcal{U}))$
$u(\cdot) \in L_2(\Omega, H_0^2(0,T;\mathcal{U}))$
$AC \in L(\mathcal{H})$
$P(\cdot) \in L(\mathcal{H} \to C([0,T];\mathcal{H}))$. *Then* $\forall t \in [0,T]$

$$E\,|\langle AX(t), P(t)X(t)\rangle| \leq c_T$$

for any solution $X(t)$ *of*

$$dX(t) = (AX(t) + Bu(t))dt + CX(t)dw_t$$

with $X(0) = x_0 \in D(A)$. *Here,* c_T *is a function of* $|B^*P|$, $|u|_{L_2(\Omega;H^2(0,T;\mathcal{U}))}$, $|AC|$, $|P|_{L(\mathcal{H}\to C([0,T];\mathcal{H})}$, $|A^{-1}B|_{L(\mathcal{U},H)}$.

The main aspect of the proof is using the regularity of the solution with smooth data and also using the regularity provided by the smooth control. The boundedness of B^*P is a key ingredient for obtaining a priori estimates.

4.6 Dynamic programming argument

Applying the Ito Formula formally for $\psi(t,x) = \langle P(t)x, x\rangle$ and

$$dy_n = (Ay_n + Bu)dt + C_n y_n dw_t$$

$y_n(t) = x$ where $AC_n = nAR(n,A)C$. Thus $C_n \in \mathcal{L}(\mathcal{H})$ and $AC_n \in \mathcal{L}(\mathcal{H})$.

$$J_n(t,x,u) = E\left(\int_t^T |u + B^*P(s)y_n(s)|_U^2 ds\right) +$$

$$+(P(t)x,x)_H + \int_t^T \langle P(s)C_n y(s)_n, C_n y_n(s)\rangle\, ds +$$

$$-\int_t^T \langle P(s)Cy_n(s), Cy_n(s)\rangle ds$$

However, in the infinite dimensional setting, Ito's lemma is applicable only to strong solutions. Therefore, the formal representation has to be justified. By a change of variables and using the regularity lemma, it is shown that a transformed version of the formal calculation is valid (see [8]).

4.7 The optimal control and the value function

By passing to the limit, we obtain

$$J(t,x,u) = E\left(\int_t^T |u + B^*P(s)y(s)|_U^2 ds\right) + (P(t)x,x)_H$$

Hence, the optimal feedback control is given by $u_{opt} = -B^*P(s)y(s)$.

4.8 The global existence of B^*P, Φ^P and P

The priori bound for P follows from the optimality. However, the main issue is to show the global bounds for Φ and B^*P. This is straightforward when B is bounded. In this unbounded B case, the following representation is the starting point for proving the global bounds.

$$\Phi(t,s) = \left(I + LL^*\left[R^*R + C^*P(\cdot)C\right]\right)^{-1} e^{A(t-s)}$$

The invertibility of $(I + LL^*\left[R^*R + C^*P(\cdot)C\right])$ on C_r shows that $\Phi \in C_r$ for all r. By using the smoothing properties of LL^*, it follows that Φ is continuous and bounded.

$$|B^*P(t)x| = \left|B^*\int_t^T e^{A^*(\tau-t)}(R^*R + C^*P(\tau)C)\Phi(\tau,t)x d\tau\right| \leq$$

$$\leq \int_t^T \frac{K_0|x|_{\mathcal{H}}}{(\tau-t)^\gamma} d\tau = K_0(T-t)^{1-\gamma}|x|_{\mathcal{H}}$$

References

[1] Avalos, G. and Lasiecka, I., Differential Riccati equation for the active control of a problem in structural acoustics, *J. Optim. Theory Appl.*, 91, 3, 1996, 695–728.

[2] Banks, H. T. and Smith, R. C., Numerical techniques for simulation, parameter estimation, and noise control in structural acoustic systems, *Dynamics and Control of Distributed Systems*, Cambridge University Press, Cambridge, 1998, 202–263.

[3] Bucci, F., Lasiecka, I., and Triggiani, R., Singular estimates and uniform stability of coupled systems of hyperbolic/parabolic PDEs, *Abstr. Appl. Anal.*, 7, 4, 2002, 169–237.

[4] Da Prato, G., Direct solution of a Riccati equation arising in stochastic control theory, *Appl. Math. Optim.*, 11, 3, 1984, 191–208.

[5] Da Prato, G. and Zabczyk, J., Stochastic equations in infinite dimensions, *Encyclopedia of Mathematics and Its Applications*, 44, Cambridge University Press, Cambridge, 1992.

[6] Flandoli, Franco, Direct solution of a Riccati equation arising in a stochastic control problem with control and observation on the boundary, *Appl. Math. Optim.*, 14, 2, 1986, 107–129.

[7] Fleming, W. H. and Soner, H. M., *Controlled Markov processes and viscosity solutions*, Applications of Mathematics, 25, Springer-Verlag, New York, 1993.

[8] Hafizoglu, C., *An infinite dimensional stochastic LQ control problem for hyperbolic-like systems with boundary/point controls*, preprint, 2005.

[9] Lasiecka, I. and Triggiani, R., Control theory for partial differential equations: continuous and approximation theories. I, *Encyclopedia of Mathematics and Its Applications*, 74, Abstract parabolic systems, Cambridge University Press, Cambridge, 2000.

[10] Lasiecka, I. and Triggiani, R., Control theory for partial differential equations: Continuous and approximation Theories. II, *Encyclopedia of Mathematics and Its Applications*, 75, Abstract hyperbolic-like systems over a finite time horizon, Cambridge University Press, Cambridge, 2000.

[11] Lions, J.-L. and Magenes, E., *Non-Homogeneous Boundary Value Problems and Applications.* Vols. I and II, Springer-Verlag, New York, 1972.

[12] Pazy, A., *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Applied Mathematical Sciences, 44, Springer-Verlag, New York, 1983.

[13] Tessitore, G., Linear quadratic optimal control for a stochastic system with control on the boundary and hyperbolic dynamics, *J. Math. Systems Estim. Control*, 2, 4, 1992, 453–482.

[14] Washburn, D., A bound on the boundary input map for parabolic equations with application to time optimal control, *SIAM J. Control Optim.*, 17. 5, 1979, 652–671.

[15] Yong, J. and Zhou, X. Y, Stochastic Controls, Applications of Mathematics (New York), 43, Hamiltonian Systems and HJB Equations, Springer-Verlag, New York, 1999.

Mathematics

Addressing algebraic problems found in biomathematics and energy, **Free and Moving Boundaries: Analysis, Simulation and Control** discusses moving boundary and boundary control in systems described by partial differential equations. With contributions from international experts, the book emphasizes numerical and theoretical control of moving boundaries in fluid structure couple systems, arteries, shape stabilization level methods, family of moving geometries, and boundary control.

Using numerical analysis, the contributors examine the problems of optimal control theory applied to partial differential equations that arise from continuum mechanics. The book presents several applications to electromagnetic devices, flow, control, computing, images analysis, topological changes, and free boundaries. It specifically focuses on the topics of boundary variation and control, dynamical control of geometry, optimization, free boundary problems, stabilization of structures, controlling fluid-structure devices, electromagnetism 3D, and inverse problems that occur in areas such as biomathematics.

Free and Moving Boundaries: Analysis, Simulation and Control explains why the boundary control of physical systems can be viewed as a moving boundary control, empowering the future research of select algebraic areas.

Features

- Emphasizes numerical and theoretical control of moving boundaries
- Explores the problems of optimal control theory applied to partial differential equations arising from continuum mechanics
- Addresses boundary variation and control, dynamical control of geometry, optimization, and inverse problems
- Presents numerical simulation of suspensions, liquids, and shape gradients
- Discusses boundary conditions, including Neumann, Dirichlet, and Robin

Roland Glowinski is professor of mathematics and mechanical engineering at the University of Houston, Texas, USA.
Jean-Paul Zolésio is director of research at CNRS and INRIA, Sophia Antipolis, France.

C6064

ISBN 1-58488-606-4

90000

9 781584 886068

Chapman & Hall/CRC
Taylor & Francis Group
an **informa** business
www.taylorandfrancisgroup.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
270 Madison Avenue
New York, NY 10016
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

www.crcpress.com