IFIP AICT 391

Dietmar Hömberg
Fredi Tröltzsch
(Eds.)

System Modeling and Optimization

25th IFIP TC 7 Conference, CSMO 2011
Berlin, Germany, September 2011
Revised Selected Papers

Springer

IFIP Advances in Information and Communication Technology 391

IFIP – The International Federation for Information Processing

IFIP was founded in 1960 under the auspices of UNESCO, following the First World Computer Congress held in Paris the previous year. An umbrella organization for societies working in information processing, IFIP's aim is two-fold: to support information processing within its member countries and to encourage technology transfer to developing nations. As its mission statement clearly states,

> *IFIP's mission is to be the leading, truly international, apolitical organization which encourages and assists in the development, exploitation and application of information technology for the benefit of all people.*

IFIP is a non-profitmaking organization, run almost solely by 2500 volunteers. It operates through a number of technical committees, which organize events and publications. IFIP's events range from an international congress to local seminars, but the most important are:

- The IFIP World Computer Congress, held every second year;
- Open conferences;
- Working conferences.

The flagship event is the IFIP World Computer Congress, at which both invited and contributed papers are presented. Contributed papers are rigorously refereed and the rejection rate is high.

As with the Congress, participation in the open conferences is open to all and papers may be invited or submitted. Again, submitted papers are stringently refereed.

The working conferences are structured differently. They are usually run by a working group and attendance is small and by invitation only. Their purpose is to create an atmosphere conducive to innovation and development. Refereeing is also rigorous and papers are subjected to extensive group discussion.

Publications arising from IFIP events vary. The papers presented at the IFIP World Computer Congress and at open conferences are published as conference proceedings, while the results of the working conferences are often published as collections of selected and edited papers.

Any national society whose primary activity is about information processing may apply to become a full member of IFIP, although full membership is restricted to one society per country. Full members are entitled to vote at the annual General Assembly, National societies preferring a less committed involvement may apply for associate or corresponding membership. Associate members enjoy the same benefits as full members, but without voting rights. Corresponding members are not represented in IFIP bodies. Affiliated membership is open to non-national societies, and individual and honorary membership schemes are also offered.

Dietmar Hömberg Fredi Tröltzsch (Eds.)

System Modeling and Optimization

25th IFIP TC 7 Conference, CSMO 2011
Berlin, Germany, September 12-16, 2011
Revised Selected Papers

Volume Editors

Dietmar Hömberg
Weierstrass Institute for Applied Analysis and Stochastics (WIAS)
Mohrenstraße 39, 10117 Berlin, Germany
E-mail: dietmar.hoemberg@wias-berlin.de

Fredi Tröltzsch
Technische Universität Berlin
Institut für Mathematik
Straße des 17. Juni 136, 10623 Berlin, Germany
E-mail: troeltzsch@math.tu-berlin.de

ISSN 1868-4238 e-ISSN 1868-422X
ISBN 978-3-642-36061-9 e-ISBN 978-3-642-36062-6
DOI 10.1007/978-3-642-36062-6
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012955348

CR Subject Classification (1998): G.1.6-8, G.1.4, I.2.8, C.4, G.3, F.4.1, I.6.3-5

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

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

Preface

This volume contains selected papers presented at the 25th IFIP TC 7 Conference held in Berlin, September 12–16, 2011, as part of a biennial conference series. The preceding ones were held in Buenos Aires (2009), Cracow (2007), and Turin (2005). The IFIP TC 7 conference series stands for a fairly unique scientific orientation focusing on the links between research in abstract mathematical optimization and control theory and on building a bridge to numerical methods and applications in various fields. It also includes contributions to mathematical modeling of applied problems. This unique flavor attracts a community of scientists that can hardly be found together at other conferences on optimization and control.

The active interplay between the different fields of optimization, mathematical modeling, and control theory was a characteristic feature of the scientific program. It showed that modern key technologies influence the research in applied optimization and control theory and, in turn, these applications often give rise to new and challenging questions of basic mathematical research. It became obvious that modern optimization and control is a vivid area that has reached a new scientific level.

All in all, the conference provided an excellent survey on the latest trends in optimization and control theory and on the tremendous progress in widespread applications in this field, and we hope you will agree after reading that the selected papers reflect this.

We would like to thank our sponsors, the German Science Foundation (DFG) and the European Science Foundation (ESF), the National Institute for Research in Computer Science and Control in France (INRIA), the DFG Research Center MATHEON, the Weierstrass Institute for Applied Analysis and Stochastics (WIAS), and the European Patent Office for their financial support. We are grateful to Technische Universität Berlin and the Institute of Mathematics for their hospitality. Last but not least, we would like to acknowledge Anke Giese (WIAS) and Frank Holzwarth of Springer for their support during the preparation of this volume.

November 2012

Dietmar Hömberg
Fredi Tröltzsch

Scientific Committee

Jacques Henry	INRIA/Université Bordeaux 1, France
Irena Lasiecka	University of Virginia, Charlottesville, USA
Alampallam V. Balakrishnan	University of California, Los Angeles, USA
Héctor Cancela	Universidad de la República, Montevideo, Uruguay
Gianni di Pillo	Università di Roma "La Sapienza", Italy
Yury G. Evtushenko	Russian Academy of Sciences, Moscow
Michael Havbro Faber	ETH Zurich, Switzerland
Raimo P. Hämäläinen	Aalto University, Finland
Peter Kall	University of Zurich, Switzerland
Alfred Kalliauer	VERBUND-Austrian Power Trading AG, Wien
Hisao Kameda	University of Tsukuba, Japan
Spiros D. Likothanassis	University of Patras, Greece
Kurt Marti	Universität der Bundeswehr München, Germany
Jiri Outrata	Academy of Sciences of the Czech Republic, Praha
Hugo D. Scolnik	Universidad de Buenos Aires, Argentinia
Lukasz Stettner	Polish Academy of Sciences, Warsaw
Philippe L. Toint	University of Namur, Belgium
Jean-Paul Zolesio	CNRS and INRIA, Sophia Antipolis, France

Table of Contents

Plenary Talks

Control of Distributed Parameter Systems

Stochastic Optimization and Control

Stabilization, Feedback, and Model Predictive Control

Flow Control

Shape and Structural Optimization

Applications and Control of Lumped Parameter Systems

Second Order Conditions for L^2 Local Optimality in PDE Control*

Eduardo Casas

Departmento de Matemática Aplicada y Ciencias de la Computación
E.T.S.I. Industriales y de Telecomunicación
Universidad de Cantabria, Av. Los Castros s/n, 39005 Santander, Spain
eduardo.casas@unican.es

Abstract. In the second order analysis of infinite dimension optimization problems, we have to deal with the so-called two-norm discrepancy. As a consequence of this fact, the second order optimality conditions usually imply local optimality in the L^∞ sense. However, we have observed that the L^2 local optimality can be proved for many control problems of partial differential equations. This can be deduced from the standard second order conditions. To this end, we make some quite realistic assumptions on the second derivative of the cost functional. These assumptions do not hold if the control does not appear explicitly in the cost functional. In this case, the optimal control is usually of bang-bang type. For this type of problems we also formulate some new second order optimality conditions that lead to the strict L^2 local optimality of the bang-bang controls.

Keywords: optimal control of partial differential equations, semilinear partial differential equations, second order optimality conditions, bang-bang controls.

1 Introduction

This paper is split into three parts. In the first part, we consider the following infinite dimensional abstract optimization problem. Let U_∞ and U_2 be Banach and Hilbert spaces, respectively, endowed with the norms $\|\cdot\|_\infty$ and $\|\cdot\|_2$. We assume that $U_\infty \subset U_2$ with continuous embedding; in particular, the choice $U_\infty = U_2$ is possible. A nonempty convex subset $\mathcal{K} \subset U_\infty$ is given, and $\mathcal{A} \subset U_\infty$ is an open set covering $\mathcal{K}$. Moreover, an objective function $J : \mathcal{A} \longrightarrow \mathbb{R}$ is given. We consider the abstract optimization problem

$$\text{(P)} \qquad \min_{u \in \mathcal{K}} J(u),$$

where we assume that J is of class C^2 with respect to the norm $\|\cdot\|_\infty$. In the next section, we will impose some other assumptions on J so that the first order

* This work was partially supported by the Spanish Ministerio de Economía y Competitividad under project MTM2011-22711.

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 1–12, 2013.

optimality conditions and the inequality $J''(\bar{u})v^2 > 0$ for every $v \in C_{\bar{u}} \setminus \{0\}$ imply that $\bar{u}$ is a strict local minimum of (P) in the U_2 sense. Here $C_{\bar{u}}$ denotes the usual cone of critical directions that we will define later. This result is new in the sense that the classical theory claims the local optimality only in the U_∞ sense due to the non-differentiability of J in with respect to $\|\cdot\|_2$. Moreover, a stronger inequality $J''(\bar{u})v^2 \geq \delta\|v\|_2^2$ is usually required.

In the second part of the paper, contained in §3, we prove that the abstract assumptions are fulfilled by a typical Neumann control problem. The method used for this control problem can be extended in an easy form to many other control problems associated with elliptic or parabolic equations; see [8]. Finally, the last part of the paper is considered in §4. There, we analyze the case of bang-bang control problems, which do not satisfy the assumptions of §2. For these problems we also give some second order conditions leading to the strict L^2 local optimality of the controls.

2 An Abstract Optimization Problem in Banach Spaces

The results presented in this section were obtained in collaboration with Fredi Tröltzsch. The reader is referred to [8] for the proofs and details.

In this section, we study the abstract optimization problem (P) formulated in the introduction. Besides the hypotheses established in §1 on U_2 and U_∞, we require the following assumptions on (P).

(A1) The functional $J : \mathcal{A} \longrightarrow \mathbb{R}$ is of class C^2. Furthermore, for every $u \in \mathcal{K}$ there exist continuous extensions

$$J'(u) \in \mathcal{L}(U_2, \mathbb{R}) \quad \text{and} \quad J''(u) \in \mathcal{B}(U_2, \mathbb{R}), \tag{2.1}$$

where $\mathcal{L}(U_2, \mathbb{R})$ and $\mathcal{B}(U_2, \mathbb{R})$ denote the spaces of continuous linear and bilinear forms on U_2, respectively.

(A2) For any sequence $\{(u_k, v_k)\}_{k=1}^\infty \subset \mathcal{K} \times U_2$ with $\|u_k - \bar{u}\|_2 \to 0$ and $v_k \rightharpoonup v$ weakly in U_2, the conditions

$$J'(\bar{u})v = \lim_{k\to\infty} J'(u_k)v_k, \tag{2.2}$$

$$J''(\bar{u})v^2 \leq \liminf_{k\to\infty} J''(u_k)v_k^2, \tag{2.3}$$

$$\text{if } v = 0, \text{ then } \quad \Lambda \liminf_{k\to\infty} \|v_k\|_2^2 \leq \liminf_{k\to\infty} J''(u_k)v_k^2, \tag{2.4}$$

hold for some $\Lambda > 0$.

The reader might have the impression that Assumptions *(A1)* and *(A2)*, mainly *(A2)*, are too strong. However, we will see in the next sections that they are fulfilled by many optimal control problems.

Associated with $\bar{u}$, we define the sets

$$\begin{aligned} S_{\bar{u}} &= \{v \in U_\infty : v = \lambda(u - \bar{u}) \text{ for some } \lambda > 0 \text{ and } u \in \mathcal{K}\}, \\ C_{\bar{u}} &= cl_2(S_{\bar{u}}) \cap \{v \in U_2 : J'(\bar{u})v = 0\} \\ D_{\bar{u}} &= \{v \in S_{\bar{u}} : J'(\bar{u})v = 0\}, \end{aligned} \tag{2.5}$$

where $cl_2(S_{\bar u})$ denotes the closure of $S_{\bar u}$ in U_2. The set $S_{\bar u}$ is called the cone of feasible directions and $C_{\bar u}$ is said to be the critical cone. It is obvious that $cl_2(D_{\bar u}) \subset C_{\bar u}$. However, the equality can fail. In fact, this equality is a regularity condition equivalent to the notion of polyhedricity of $\mathcal{K}$; see [2] or [1, §3.2]. This property is enjoyed by control problems with pointwise control constraints.

Now, we formulate the necessary first and second order optimality conditions. The second order conditions hold under the mentioned regularity assumption; we refer to [1, §3.2] or [7] for the proof.

Theorem 2.1. *Assume that (A1) holds and let $\bar u$ be a local solution of* (P) *in U_∞, then $J'(\bar u)(u-\bar u) \geq 0$ $\forall u \in \mathcal{K}$. Moreover, if the regularity condition $C_{\bar u} = cl_2(D_{\bar u})$ is satisfied, then $J''(\bar u)v^2 \geq 0$ holds for all $v \in C_{\bar u}$.*

Now, we state our result about sufficient sufficient second order optimality conditions. As the reader may check, the gap between the necessary and sufficient second order conditions is minimal, the same as in finite dimension.

Theorem 2.2. *Suppose that assumptions (A1) and (A2) hold. Let $\bar u \in \mathcal{K}$ satisfy the first order optimality condition as formulated in Theorem 2.1, and*

$$J''(\bar u)v^2 > 0 \quad \forall v \in C_{\bar u} \setminus \{0\}. \tag{2.6}$$

Then, there exist $\varepsilon > 0$ and $\delta > 0$ such that

$$J(\bar u) + \frac{\delta}{2}\|u-\bar u\|_2^2 \leq J(u) \quad \forall u \in \mathcal{K} \cap B_2(\bar u;\varepsilon). \tag{2.7}$$

Above $B_2(\bar u;\varepsilon)$ denotes the ball of U_2 with center at $\bar u$ and radius ε.

This theorem can be proved arguing by contradiction. To this end, we assume that for any positive integer k there exists $u_k \in \mathcal{K}$ such that

$$\|u_k - \bar u\|_2 < \frac{1}{k} \quad \text{and} \quad J(\bar u) + \frac{1}{2k}\|u_k-\bar u\|_2^2 > J(u_k). \tag{2.8}$$

Setting $\rho_k = \|u_k - \bar u\|_2$ and $v_k = (u_k - \bar u)/\rho_k$, we can assume that $v_k \rightharpoonup v$ in U_2; if necessary, we select a subsequence. Then we prove that $v \in C_{\bar u}$ and $J''(\bar u)v^2 = 0$. Because of (2.6), this is only possible if $v = 0$. With the help of (2.4) the contradiction is obtained from the identity $\|v_k\|_2 = 1$; see [8] for the details.

As a consequence of Theorem 2.2, we can not only prove that $\bar u$ is the unique local minimum in a certain U_2 neighborhood. We are even able to show the non-existence of other stationary points in such a neighborhood. Recall that $\tilde u \in \mathcal{K}$ is said to be a stationary point if

$$J'(\tilde u)(u-\tilde u) \geq 0 \quad \text{for all} \quad u \in \mathcal{K}. \tag{2.9}$$

Corollary 2.3. *Under the assumptions of Theorem 2.2, there exists $\varepsilon > 0$ such that there is no stationary point $\tilde{u} \in B_2(\bar{u};\varepsilon) \cap \mathcal{K}$ different from $\bar{u}$.*

Assumption (2.6) has another consequence that was known up to now only in an U_∞-neighborhood of $\bar{u}$. The result expresses some alternative formulation of second-order sufficient conditions that is useful for applications in the numerical analysis.

Theorem 2.4. *Under the assumptions of Theorem 2.2, there exist numbers $\varepsilon > 0$, $\nu > 0$ and $\tau > 0$ such that*

$$J''(u)v^2 \geq \frac{\nu}{2}\|v\|_2^2 \quad \forall\, v \in E_{\bar{u}}^\tau \quad \text{and} \quad \forall\, u \in \mathcal{K} \cap B_2(\bar{u};\varepsilon), \tag{2.10}$$

where $$E_{\bar{u}}^\tau = \{v \in cl_2(S_{\bar{u}}) : |J'(\bar{u})v| \leq \tau\|v\|_2\}.$$

3 Application. An Elliptic Neumann Control Problem

In this section we study the optimal control problem

$$(\mathrm{P}_1) \qquad \min_{u\in\mathcal{K}} J(u),$$

where
$$J(u) = \int_\Omega L(x, y_u(x))\, dx + \int_\Gamma l(x, y_u(x), u(x))\, d\sigma(x), \tag{3.1}$$
$$\mathcal{K} = \{u \in L^\infty(\Gamma) : \alpha \leq u(x) \leq \beta \ \text{ for a.a. } x \in \Gamma\},$$

σ denotes the Lebesgue surface measure, $-\infty < \alpha < \beta < +\infty$, and y_u is the solution of the following Neumann problem

$$\begin{cases} -\Delta y + f(y) = 0 & \text{in } \Omega, \\ \partial_\nu y = u & \text{on } \Gamma. \end{cases} \tag{3.2}$$

We impose the following assumptions on the functions and parameters appearing in the control problem (P_1).

Assumption (N1): Ω is an open, bounded and connected subset of $\mathbb{R}^n$, $n \geq 2$, with Lipschitz boundary Γ and $f : \mathbb{R} \longrightarrow \mathbb{R}$ is a function of class C^2 such that $f'(t) \geq c_o > 0$ for all $t \in \mathbb{R}$. The reader is referred to [5] for more general non-linear terms in the state equation.

Assumption (N2): We assume that $L : \Omega \times \mathbb{R} \longrightarrow \mathbb{R}$ and $l : \Gamma \times \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$ are Carathéodory functions of class C^2 with respect to the second variable for L and with respect to the second and third variables for l, with $L(\cdot, 0) \in L^1(\Omega)$, $l(\cdot, 0, 0) \in L^1(\Gamma)$. For every $M > 0$ there exist functions $\psi_M \in L^{\bar{p}}(\Omega)$, $\bar{p} > n/2$, and $\phi_M \in L^{\bar{q}}(\Gamma)$, $\bar{q} > n - 1$, and a constant $C_M > 0$ such that

$$\begin{cases} \left|\dfrac{\partial^j L}{\partial y^j}(x,y)\right| \le \psi_M(x), \quad \text{with } j=1,2, \\ \left|\dfrac{\partial^j l}{\partial y^j}(x,y,u)\right| \le \phi_M(x), \quad \text{with } j=1,2, \\ \left|\dfrac{\partial^{i+j} l}{\partial u^i \partial y^j}(x,y,u)\right| \le C_M, \ 1 \le i+j \le 2 \text{ and } i \ge 1 \end{cases}$$

are satisfied for a.a. $x \in \Omega$ and every $u, y \in \mathbb{R}$, with $|y| \le M$ and $|u| \le M$.

Moreover, for every $\varepsilon > 0$ there exists $\eta > 0$ such that for a.a. $x \in \Omega$ and all $u_i, y_i \in \mathbb{R}$, with $i = 1, 2$,

$$\begin{cases} |y_2 - y_1| \le \eta \Rightarrow \left|\dfrac{\partial^2 L}{\partial y^2}(x,y_2) - \dfrac{\partial^2 L}{\partial y^2}(x,y_1)\right| \le \varepsilon, \\ |u_2 - u_1| + |y_2 - y_1| \le \eta \Rightarrow \left|D^2_{(y,u)} l(x,y_2,u_2) - D^2_{(y,u)} l(x,y_1,u_1)\right| \le \varepsilon. \end{cases}$$

Here $D^2_{(y,u)} l(x,y,u)$ denotes the Hessian matrix of l with respect to the variables (y,u). We also assume the Legendre-Clebsch type condition

$$\exists \Lambda > 0 \text{ such that } \frac{\partial^2 l}{\partial u^2}(x,y,u) \ge \Lambda \ \text{ for a.a. } x \in \Gamma \text{ and } \forall y, u \in \mathbb{R}. \tag{3.3}$$

It is obvious that the usual quadratic integrands $L(x,y) = \frac{1}{2}(y - y_{Ld}(x))^2$ and $l(x,y,u) = \frac{1}{2}(y - y_{ld}(x))^2 + \frac{\Lambda}{2}u^2$ satisfy Assumption *(N2)* if $y_{Ld} \in L^{\bar{p}}(\Omega)$ and $y_{ld} \in L^{\bar{q}}(\Gamma)$.

The hypothesis (3.3) is crucial for satisfying the assumptions (2.3) and (2.4). In §4 we will consider the case where (3.3) doe not hold.

On the state equation (2.1), the following result is known.

Theorem 3.1. *Under the Assumption (N1), for every $u \in L^{\bar{q}}(\Gamma)$ the equation (3.2) has a unique solution $y_u \in H^1(\Omega) \cap C(\bar{\Omega})$. Furthermore, the mapping $G : L^{\bar{q}}(\Gamma) \longrightarrow H^1(\Omega) \cap C(\bar{\Omega})$, defined by $G(u) = y_u$, is of class C^2. For elements u, v, v_1 and v_2 of $L^{\bar{q}}(\Gamma)$, the functions $z_v = G'(u)v$ and $z_{v_1 v_2} = G''(u)(v_1, v_2)$ are the solutions of the problems*

$$\begin{cases} Az + f'(y_u)z = 0 & \text{in } \Omega, \\ \partial_{\nu_A} z = v & \text{on } \Gamma, \end{cases} \tag{3.4}$$

and

$$\begin{cases} Az + f'(y_u)z + f''(y_u)z_{v_1} z_{v_2} = 0 & \text{in } \Omega, \\ \partial_{\nu_A} z = 0 & \text{on } \Gamma, \end{cases} \tag{3.5}$$

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

The proof of existence and uniqueness of a solution y_u in $H^1(\Omega) \cap L^\infty(\Omega)$ is standard; see, for instance, [3]. For the continuity of y_u, the reader is referred to [11] or [12]. As usual, the differentiability of G can be obtained from the implicit function theorem.

As a consequence of this theorem and the chain rule the next result follows.

Theorem 3.2. *Assuming (N1) and (N2), then the mapping $J : L^\infty(\Gamma) \longrightarrow \mathbb{R}$, defined by (3.1), is of class C^2. For all u, v, v_1 and v_2 of $L^\infty(\Gamma)$ we have*

$$J'(u)v = \int_\Gamma \left(\varphi_u + \frac{\partial l}{\partial u}(x, y_u, u) \right) v \, d\sigma \tag{3.6}$$

$$\begin{aligned} J''(u)(v_1, v_2) = &\int_\Omega \left(\frac{\partial^2 L}{\partial y^2}(x, y_u) - \varphi_u f''(y_u) \right) z_{v_1} z_{v_2} \, dx \\ &+ \int_\Gamma \left(\frac{\partial^2 l}{\partial y^2}(x, y_u, u) z_{v_1} z_{v_2} + \frac{\partial^2 l}{\partial y \partial u}(x, y_u, u)(v_1 z_{v_2} + v_2 z_{v_1}) \right) d\sigma \\ &+ \int_\Gamma \frac{\partial^2 l}{\partial u^2}(x, y_u, u) v_1 v_2 \, d\sigma, \end{aligned} \tag{3.7}$$

where $z_{v_i} = G'(u)v_i$, $i = 1, 2$, and $\varphi_u \in H^1(\Omega) \cap C(\bar{\Omega})$ is the solution of

$$\begin{cases} -\Delta \varphi + f'(y_u)\varphi = \dfrac{\partial L}{\partial y}(x, y_u) & \text{in } \Omega, \\ \partial_\nu \varphi = \dfrac{\partial l}{\partial y}(x, y_u, u) & \text{on } \Gamma. \end{cases} \tag{3.8}$$

From the above expressions for $J'(u)$ and $J''(u)$ and Assumption *(N2)* we deduce that $J'(u)$ and $J''(u)$ can be extended to linear and bilinear forms, respectively, on $L^2(\Gamma)$. Even more, there exist two constants $M_1 > 0$ and $M_2 > 0$ such that for every $v, v_1, v_2 \in L^2(\Gamma)$ and $u \in \mathcal{K}$

$$|J'(u)v| \le M_1 \|v\|_{L^2(\Gamma)} \quad \text{and} \quad |J''(u)(v_1, v_2)| \le M_2 \|v_1\|_{L^2(\Gamma)} \|v_2\|_{L^2(\Gamma)}. \tag{3.9}$$

This shows that (2.1) holds with $U_2 = L^2(\Gamma)$ and $U_\infty = L^\infty(\Gamma)$. The most delicate issue in the proof of (2.2)-(2.4) is the verification of (2.3), which can be done with the help of the following lemma.

Lemma 3.1 *Let (X, Σ, μ) be a measure space with $\mu(X) < +\infty$. Suppose that $\{g_k\}_{k=1}^\infty \subset L^\infty(X)$ and $\{v_k\}_{k=1}^\infty \subset L^2(X)$ satisfy the assumptions*

- *$g_k \ge 0$ a.e. in X, $\{g_k\}_{k=1}^\infty$ is bounded in $L^\infty(X)$ and $g_k \to g$ in $L^1(X)$.*
- *$v_k \rightharpoonup v$ in $L^2(X)$.*

Then there holds the inequality

$$\int_X g(x) v^2(x) \, d\mu(x) \le \liminf_{k \to \infty} \int_X g_k(x) v_k^2(x) \, d\mu(x). \tag{3.10}$$

The proof of this lemma can be obtained by an application of Egorov's theorem; see [8]. To confirm (2.3) we apply Lemma 3.1 with $X = \Gamma$, $\mu = \sigma$ and

$$0 < \Lambda \le g_k(x) = \frac{\partial^2 l}{\partial u^2}(x, y_{u_k}(x), u_k(x)) \to g(x) = \frac{\partial^2 l}{\partial u^2}(x, y_u(x), u(x)) \quad \text{in } L^1(\Gamma).$$

Finally, we apply Theorems 2.1 and 2.2 to the problem (P_1). Given $\bar{u} \in \mathcal{K}$, we see that the cone of critical directions $C_{\bar{u}}$ defined in §2 can be expressed for the problem (P_1) in the form

$$C_{\bar{u}} = \{v \in L^2(\Gamma) : v(x) = \begin{cases} \ge 0 \text{ if } \bar{u}(x) = \alpha \\ \le 0 \text{ if } \bar{u}(x) = \beta \\ 0 \quad \text{if } \bar{d}(x) \ne 0 \end{cases} \text{ a.e. in } \Gamma\},$$

where

$$\bar{d}(x) = \bar{\varphi}(x) + \frac{\partial l}{\partial u}(x, \bar{y}(x), \bar{u}(x))$$

and $\bar{y} = y_{\bar{u}}$ and $\bar{\varphi} = \varphi_{\bar{u}}$ denote the state and adjoint state associated to $\bar{u}$, respectively. It is not difficult to check that the regularity assumption stated in Theorem 2.1 is fulfilled by $C_{\bar{u}}$. Then we have the following corollaries.

Corollary 3.1 *Let the Assumption (N1) be satisfied and suppose that $\bar{u}$ is a local minimum of (P_1) in the $L^\infty(\Gamma)$ sense. Then there holds $J'(\bar{u})(u - \bar{u}) \ge 0$ for all $u \in \mathcal{K}$ and $J''(\bar{u})v^2 \ge 0$ $\forall v \in C_{\bar{u}}$. Conversely, if $\bar{u} \in \mathcal{K}$ obeys*

$$J'(\bar{u})(u - \bar{u}) \ge 0 \quad \forall u \in \mathcal{K}, \tag{3.11}$$

$$J''(\bar{u})v^2 > 0 \quad \forall v \in C_{\bar{u}} \setminus \{0\}, \tag{3.12}$$

then there exist $\varepsilon > 0$ and $\delta > 0$ such that

$$J(\bar{u}) + \frac{\delta}{2}\|u - \bar{u}\|^2_{L^2(\Gamma)} \le J(u) \quad \forall u \in \mathcal{K} \cap B_2(\bar{u}; \varepsilon). \tag{3.13}$$

Let us underline that the mapping G is only differentiable in $L^q(\Gamma)$ for $q > n-1$. Consequently, for all $n \ge 3$, G is not differentiable in $L^2(\Gamma)$. Moreover, the general nonlinear cost functional J is only differentiable in $L^\infty(\Gamma)$. Hence, for any dimension n, the classical theory of second order conditions would only assure the local optimality of $\bar{u}$ in the $L^\infty(\Gamma)$ sense. In contrast to this, our result guarantees local optimality in the sense of $L^2(\Gamma)$.

Corollary 3.2 *Under the assumption (N1) and (N2), there exists a ball $B_2(\bar{u}; \varepsilon)$ in $L^2(\Gamma)$ such that there is no other stationary point in $B_2(\bar{u}; \varepsilon) \cap \mathcal{K}$ than $\bar{u}$. Moreover, there exist numbers $\nu > 0$ and $\tau > 0$ such that*

$$J''(u)v^2 \ge \frac{\nu}{2}\|v\|^2_{L^2(\Gamma)} \quad \forall\, v \in C^\tau_{\bar{u}} \quad \text{and} \quad \forall\, u \in \mathcal{A} \cap B_2(\bar{u}; \varepsilon), \tag{3.14}$$

where $\mathcal{A}$ is a bounded open subset of $L^\infty(\Gamma)$ containing $\mathcal{K}$ and

$$C_{\bar u}^\tau = \{v \in L^2(\Gamma) : v(x) = \begin{cases} \geq 0 \text{ if } \bar u(x) = \alpha \\ \leq 0 \text{ if } \bar u(x) = \beta \\ 0 \quad \text{if } |\bar d(x)| > \tau \end{cases} \text{a.e. in } \Gamma\}.$$

In the above corollaries, $B_2(\bar u;\varepsilon)$ denotes the $L^2(\Gamma)$-ball of radius ε centered at $\bar u$.

Observe that the above cone $C_{\bar u}^\tau$ is not equal to the cone $E_{\bar u}^\tau$ defined in Theorem 2.4. However, if $v \in C_{\bar u}^\tau$, then

$$|J'(\bar u)v| = \int_\Gamma |\bar d(x)v(x)|\,dx \leq \tau \int_{\{x:|\bar d(x)|\leq\tau\}} |v(x)|\,dx \leq \tau\sqrt{|\Gamma|}\|v\|_{L^2(\Gamma)}.$$

Thus, we have that $C_{\bar u}^\tau \subset E_{\bar u}^{\tau_\Gamma}$, with $\tau_\Gamma = \tau\sqrt{|\Gamma|}$. Hence, Theorem 2.4 can be applied.

4 A Bang-Bang Control Problem

The reader is referred to [4] for proofs and extensions of the results stated below. Let Ω be an open and bounded domain in $\mathbb{R}^n$, $n \leq 3$, with a Lipschitz boundary Γ. In this domain, we consider the following control problem

$$(\mathrm{P}_2) \quad \begin{cases} \min J(u) = \displaystyle\int_\Omega L(x, y_u(x))\,dx \\ \alpha \leq u(x) \leq \beta \end{cases}$$

where y_u is the solution of the Dirichlet problem

$$\begin{cases} -\Delta y + f(y) = u \text{ in } \Omega, \\ \qquad\qquad y = 0 \text{ on } \Gamma, \end{cases} \tag{4.1}$$

$-\infty < \alpha < \beta < +\infty$ and L and f satisfy the following assumptions.

Assumption (D1) The function $f : \mathbb{R} \longrightarrow \mathbb{R}$ is of class C^2 and $f'(t) \geq 0$ for every $t \in \mathbb{R}$.

Assumption (D2) The function $L : \Omega \times \mathbb{R} \longrightarrow \mathbb{R}$ is measurable with respect to the first variable and of class C^2 with respect to the second. Moreover, $L(\cdot, 0) \in L^1(\Omega)$, and for all $M > 0$ there is a constant $C_{L,M} > 0$ and a function $\psi_M \in L^{\bar p}(\Omega)$ such that

$$\left|\frac{\partial L}{\partial y}(x,y)\right| \leq \psi_M(x), \quad \left|\frac{\partial^2 L}{\partial y^2}(x,y)\right| \leq C_{L,M}.$$

For every $M > 0$ and $\varepsilon > 0$ there exists $\delta > 0$, depending on M and ε such that

$$\left|\frac{\partial^2 L}{\partial y^2}(x,y_2) - \frac{\partial^2 L}{\partial y^2}(x,y_1)\right| < \varepsilon \text{ if } |y_1|, |y_2| \leq M,\ |y_2 - y_1| \leq \delta, \text{ for a.a. } x \in \Omega.$$

Hereafter, we will denote

$$\mathcal{K} = \{u \in L^\infty(\Omega) : \alpha \le u(x) \le \beta \ \text{ for a.e. } \ x \in \Omega\}.$$

For every $u \in L^p(\Omega)$, with $p > n/2$, the state equation (4.1) has a unique solution $y_u \in H_0^1(\Omega) \cap C(\bar{\Omega})$. The proof of this result is a quite standard combination of Schauder's fixed point theorem and the $L^\infty(\Omega)$ estimates [12]. For the continuity of the solution in $\bar{\Omega}$ see, for instance, [10, Theorem 8.30]. Moreover, the mapping $G : L^p(\Omega) \longrightarrow H_0^1(\Omega) \cap C(\bar{\Omega})$, with $G(u) = y_u$, is of class C^2. In the sequel, we will take $p = 2$ and we will denote by $z_v = G'(u)v$ the solution of

$$\begin{cases} -\Delta z + f'(y_u)z = v \text{ in } \Omega, \\ z = 0 \text{ on } \Gamma, \end{cases} \tag{4.2}$$

where $y_u = G(u)$ is the state corresponding to u. As usual, we consider the adjoint state equation associated with a control u

$$\begin{cases} -\Delta\varphi + f'(y_u)\varphi = \dfrac{\partial L}{\partial y}(x, y_u) \text{ in } \Omega, \\ \varphi = \quad 0 \qquad \text{on } \Gamma, \end{cases} \tag{4.3}$$

denoted by φ_u. Because of the assumptions on L, we have that $\varphi \in H_0^1(\Omega) \cap C(\bar{\Omega})$. Moreover, there exists $M > 0$ such that

$$\|y_u\|_\infty + \|\varphi_u\|_\infty \le M \quad \forall u \in \mathcal{K}. \tag{4.4}$$

Under the above assumptions, the problem (P_2) has at least one solution $\bar{u}$ with an associated state $\bar{y} \in H_0^1(\Omega) \cap C(\bar{\Omega})$. The cost functional $J : L^2(\Omega) \longrightarrow \mathbb{R}$ is of class C^2 and the first and second derivatives are given by

$$J'(u)v = \int_\Omega \varphi_u(x)v(x)\,dx, \tag{4.5}$$

and

$$J''(u)(v_1, v_2) = \int_\Omega \left(\frac{\partial^2 L}{\partial y^2}(x, y_u(x)) - \varphi_u(x) f''(y_u(x)) \right) z_{v_1}(x) z_{v_2}(x)\,dx, \tag{4.6}$$

where $z_{v_i} = G'(v_i)$ are the solution of (4.2) for $v = v_i$, $i = 1, 2$.

Any local solution $\bar{u}$ satisfies the optimality system

$$\begin{cases} -\Delta\bar{y} + f(\bar{y}) = \bar{u} \text{ in } \Omega, \\ \bar{y} = 0 \text{ on } \Gamma, \end{cases} \tag{4.7}$$

$$\begin{cases} -\Delta\bar{\varphi} + f'(\bar{y})\bar{\varphi} = \dfrac{\partial L}{\partial y}(x, \bar{y}) \text{ in } \Omega, \\ \bar{\varphi} = \quad 0 \qquad \text{on } \Gamma, \end{cases} \tag{4.8}$$

$$\int_\Omega \bar{\varphi}(x)(u(x) - \bar{u}(x))\,dx \ge 0 \quad \forall u \in \mathcal{K}. \tag{4.9}$$

From the last condition, we deduce as usual for a.a. $x \in \Omega$

$$\bar{u}(x) \begin{cases} = \alpha \text{ if } \bar{\varphi}(x) > 0, \\ = \beta \text{ if } \bar{\varphi}(x) < 0, \end{cases} \quad \text{and} \quad \bar{\varphi}(x) \begin{cases} > 0 \text{ if } \bar{u}(x) = \alpha, \\ < 0 \text{ if } \bar{u}(x) = \beta, \\ = 0 \text{ if } \alpha < \bar{u}(x) < \beta. \end{cases} \tag{4.10}$$

The cone of critical directions associated with $\bar{u}$ is defined by

$$C_{\bar{u}} = \{v \in L^2(\Omega) : v(x) \begin{cases} \geq 0 \text{ if } \bar{u}(x) = \alpha \\ \leq 0 \text{ if } \bar{u}(x) = \beta \\ = 0 \text{ if } \bar{\varphi}(x) \neq 0 \end{cases} \}$$

Then, the necessary second order condition satisfied is written in the form

$$J''(\bar{u})v^2 \geq 0 \quad \forall v \in C_{\bar{u}}. \tag{4.11}$$

For the above results the reader is referred to [5] or [6], where similar cases were studied. Let us remark that in the case where the set of zeros of $\bar{\varphi}$ has a zero Lebesgue measure, then $\bar{u}(x)$ is either α or β for almost all points $x \in \Omega$, i.e. $\bar{u}$ is a bang-bang control. Moreover, in this case, $C_{\bar{u}} = \{0\}$, therefore (4.11) does not provide any information. Consequently, it is unlikely that the sufficient second order conditions could be based on the set $C_{\bar{u}}$. To overcome this drawback we are going to increase the set $C_{\bar{u}}$. For every $\tau \geq 0$ we define

$$C_{\bar{u}}^{\tau} = \{v \in L^2(\Omega) : v(x) \begin{cases} \geq 0 \text{ if } \bar{u}(x) = \alpha \\ \leq 0 \text{ if } \bar{u}(x) = \beta \\ = 0 \text{ if } |\bar{\varphi}(x)| > \tau \end{cases} \}$$

It is obvious that $C_{\bar{u}}^0 = C_{\bar{u}}$. An example due to Dunn [9] proves that, in general, the second order condition based on the cone $C_{\bar{u}}$ is not sufficient for the local optimality. Before analyzing (P_2), let us take a look on its Tikhonov regularization. For any $\Lambda > 0$, let us consider the problem

$$(\mathrm{P}_{2,\Lambda}) \qquad \min_{u \in \mathcal{K}} J_\Lambda(u) = \int_\Omega L(x, y_u(x))\, dx + \frac{\Lambda}{2} \int_\Omega u^2(x)\, dx.$$

Then, we have

$$J'_\Lambda(u)v = \int_\Omega (\varphi_u + \Lambda u) v\, dx$$

and

$$J''_\Lambda(u)(v_1, v_2) = \int_\Omega \left(\frac{\partial^2 L}{\partial y^2}(x, y_u) - \varphi_u \frac{\partial^2 f}{\partial y^2}(x, y_u) \right) z_{v_1} z_{v_2} dx + \Lambda \int_\Omega v_1 v_2 dx.$$

Now, we apply Theorem 2.2 to $(\mathrm{P}_{2,\Lambda})$ and we obtain the following result.

Theorem 4.1. *Let $\bar{u} \in \mathcal{K}$ satisfy that*

$$J'_\Lambda(\bar{u})(u - \bar{u}) \geq 0 \quad \forall u \in \mathcal{K} \quad \textit{and}$$
$$J''_\Lambda(\bar{u})v^2 > 0 \quad \forall v \in C_{\bar{u}} \setminus \{0\}.$$

Then, there exists $\delta > 0$ and $\varepsilon > 0$ such that

$$J_\Lambda(\bar{u}) + \frac{\delta}{2}\|u - \bar{u}\|^2_{L^2(\Omega)} \leq J_\Lambda(u) \quad \forall u \in B_2(\bar{u};\varepsilon) \cap \mathcal{K}.$$

In the above theorem and hereafter, $B_2(\bar{u};\varepsilon)$ denotes the $L^2(\Omega)$-ball of center at $\bar{u}$ and radius ε. Now, invoking Theorem 2.4 and observing that $C^\tau_{\bar{u}} \subset E^{\tau_\Omega}_{\bar{u}}$ for $\tau_\Omega = \sqrt{|\Omega|}\tau$, we get the following theorem.

Theorem 4.2. *Let $\bar{u} \in \mathcal{K}$ satisfy $J'_\Lambda(\bar{u})(u - \bar{u}) \geq 0$ for every $u \in \mathcal{K}$. Then, the following assumptions are equivalent*

1. $J''_\Lambda(\bar{u})v^2 > 0 \quad \forall v \in C_{\bar{u}} \setminus \{0\}$.
2. $\exists \nu > 0$ *and* $\tau > 0$ *s.t.* $J''_\Lambda(\bar{u})v^2 \geq \nu\|v\|^2_{L^2(\Omega)} \quad \forall v \in C^\tau_{\bar{u}}$.
3. $\exists \nu > 0$ *and* $\tau > 0$ *s.t.* $J''_\Lambda(\bar{u})v^2 \geq \nu\|z_v\|^2_{L^2(\Omega)} \quad \forall v \in C^\tau_{\bar{u}}$,

where $z_v = G'(\bar{u})v$.

In the case $\Lambda = 0$, Dunn's example shows that *1* is not enough, in general, to assure the local optimality of $\bar{u}$. We will see below that *2* does not hold for $\Lambda = 0$. Then, it remains to analyze if the assumption *3* is enough for the local optimality of $\bar{u}$ when $\Lambda = 0$. The next theorem proves that it is sufficient.

Theorem 4.3. *Let us assume that $\bar{u}$ is a feasible control for problem* (P_2) *satisfying the first order optimality conditions* (4.7)-(4.9) *and suppose that there exist $\delta > 0$ and $\tau > 0$ such that*

$$J''(\bar{u})v^2 \geq \delta\|z_v\|^2_{L^2(\Omega)} \quad \forall v \in C^\tau_{\bar{u}}, \tag{4.12}$$

where $z_v = G'(\bar{u})v$ is the solution of (4.2) *for $y = \bar{y}$. Then, there exists $\varepsilon > 0$ such that*

$$J(\bar{u}) + \frac{\delta}{8}\|z_{u-\bar{u}}\|^2_{L^2(\Omega)} \leq J(u) \quad \forall u \in B_2(\bar{u};\varepsilon) \cap \mathcal{K}, \tag{4.13}$$

with $z_{u-\bar{u}} = G'(\bar{u})(u - \bar{u})$.

Corollary 4.1 *Under the hypotheses of Theorem 4.3, there exists $\varepsilon > 0$ such that*

$$J(\bar{u}) + \frac{\delta}{9}\|y_u - \bar{y}\|^2_{L^2(\Omega)} \leq J(u) \quad \forall u \in B_2(\bar{u};\varepsilon) \cap \mathcal{K}. \tag{4.14}$$

We finish by showing that the statement *2* of Theorem 4.2 does not hold for $\Lambda = 0$. Indeed, let us assume that it holds. Then, a simple modification of the proof of Theorem 4.3, see [4], leads to the inequality

$$J(\bar{u}) + \frac{\nu}{2}\|u - \bar{u}\|^2_{L^2(\Omega)} \leq J(u) \quad \forall u \in B_2(\bar{u};\varepsilon) \cap \mathcal{K}, \tag{4.15}$$

for some $\nu > 0$ and $\varepsilon > 0$. Then, $\bar{u}$ is a solution of the problem

$$(P_\nu) \qquad \min_{u \in B_2(\bar{u};\varepsilon)\cap\mathcal{K}} J(u) - \frac{\nu}{2}\int_\Omega (u - \bar{u})^2\, dx.$$

The Hamiltonian of this control problem is given by

$$H(x, y, u, \varphi) = L(x, y) + \varphi(u - f(x, y)) - \frac{\nu}{2}(u - \bar{u}(x))^2.$$

From the Pontryagin's principle we deduce

$$H(x, \bar{y}(x), \bar{u}(x), \bar{\varphi}(x)) = \min_{t \in [\alpha, \beta]} H(x, \bar{y}(x), t, \bar{\varphi}(x)) \text{ for almost all } x \in \Omega.$$

However, invoking (4.10) we obtain that this is a contradiction to the following facts that can be easily checked

$$\begin{cases} \text{If } 0 < \bar{\varphi}(x) < \dfrac{\nu}{2}(\beta - \alpha) \text{ then } H(x, \bar{y}(x), \beta, \bar{\varphi}(x)) < H(x, \bar{y}(x), \alpha, \bar{\varphi}(x)), \\ \text{If } 0 > \bar{\varphi}(x) > \dfrac{\nu}{2}(\alpha - \beta) \text{ then } H(x, \bar{y}(x), \alpha, \bar{\varphi}(x)) < H(x, \bar{y}(x), \beta, \bar{\varphi}(x)). \end{cases}$$

References

1. Bonnans, F., Shapiro, A.: Perturbation Analysis of Optimization Problems. Springer (2000)
2. Bonnans, J.: Second-order analysis for control constrained optimal control problems of semilinear elliptic systems. Appl. Math. Optim. 38, 303–325 (1998)
3. Casas, E.: Boundary control of semilinear elliptic equations with pointwise state constraints. SIAM J. Control Optim. 31, 993–1006 (1993)
4. Casas, E.: Second order analysis for bang-bang control problems of pde (submitted, 2012)
5. Casas, E., Mateos, M.: Second order optimality conditions for semilinear elliptic control problems with finitely many state constraints. SIAM J. Control Optim. 40, 1431–1454 (2002)
6. Casas, E., Tröltzsch, F.: Optimality conditions for a class of optimal control problems with quasilinear elliptic equations. SIAM J. Control Optim. 48, 688–718 (2009)
7. Casas, E., Tröltzsch, F.: A general theorem on error estimates with application to elliptic optimal control problems. Comp. Optim. Appls. (to appear)
8. Casas, E., Tröltzsch, F.: Second order analysis for optimal control problems: Improving results expected from abstract theory. SIAM J. Optim. (to appear)
9. Dunn, J.: Second-order optimality conditions in sets of L^∞ functions with range in a polyhedron. SIAM J. Control Optim. 33, 1603–1635 (1995)
10. Gilbarg, D., Trudinger, N.: Elliptic Partial Differential Equations of Second Order. Springer, Heidelberg (1977)
11. Haller-Dintelmann, R., Meyer, C., Rehberg, J., Schiela, A.: Hölder continuity and optimal control for nonsmooth elliptic domains. Appl. Math. Optim. 60, 397–428 (2009)
12. Stampacchia, G.: Problemi al contorno ellittici con dati discontinui dotati di soluzioni Hölderiane. Ann. Mat. Pura Appl. 51, 1–38 (1960)

Quadratic ODE and PDE Models of Drug Release Kinetics from Biodegradable Polymers

Michel C. Delfour and André Garon

[1] Centre de Recherches Mathématiques et Département de Mathématiques et de Statistique, Université de Montréal, CP 6128, Succ Centre-ville, Montréal (Qc), Canada H3C 3J7
delfour@crm.umontreal.ca
[2] Département de Génie Mécanique, École Polytechnique de Montréal, C.P. 6079, Succ. Centre-ville, Montréal (Qc), Canada H3C 3A7
Andre.Garon@polymtl.ca

Abstract. In order to achieve prescribed drug release kinetics over long therapeutic periods, bi-phasic and possibly multi-phasic releases from blends of biodegradable polymers are currently envisioned. The modelling of drug release in the presence of degradation of the polymer matrix and surface erosion is quite complex. Yet, simple reliable mathematical models validated against experimental data are now available to help in classifying neat polymers and in predicting the release dynamics from polymer blends. In this paper, we survey a two-parameter quadratic ODE model that has been validated against experimental data for the release of paclitaxel from a broad range of biodegradable polymers and a quadratic semi-permeable membrane PDE model that mimics the ODE model and could readily be extended to drug eluding stents.

Keywords: Drug release models, biodegradable polymers, paclitaxel.

1 Introduction

Stents are used in interventional cardiology to keep a diseased vessel open after angioplasty. This procedure is known to damage the endothelium at the insertion site and thus to favour the occurrence of in-stent restenosis through the proliferation of smooth muscle cells (SMC) within the vessel lumen. To control the abnormal behaviour of SMC, stents are coated with polymers that slowly release drug through diffusion into the vessel wall (drug-eluting stents or DES). These drugs are designed to control the rate of mitosis of SMC until the regeneration of the endothelium. The reader is referred to T. Kataoka et als [15] in 2002 and Joner et als [14] in 2006 for a fairly well-documented account of DES for the prevention of neointimal growth (see, for instance, [15, Figure 1, p. 1791]).

If endothelial cells do not recover to effectively control the proliferation of SMC's, a sustained dose will be required over the therapeutic period and even forever. In order to achieve prescribed drug release kinetics the current design strategies focus on bi-phasic and possibly multi-phasic releases from blends of

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 13–24, 2013.

biodegrable polymers (see, for instance, Batycky et al [1] in 1997) to achieve specific drug release kinetics profiles over long therapeutic windows.

Recently, Lao and Venkatraman [16] and Lao, Venkatraman, and Peppas [18] have proposed a semi-empirical model to predict the release profile of paclitaxel from three neat polymer matrices: PCL (Polycaprolactone), PLGA (dl-lactide-co-glycolide) and PLGAPEG (PLGA with polyethylene glycol). They are representative of a broad family of biodegradable polymers ranging from hydrophobic to hydrophilic. In hydrophilic polymers the internal bounds between the chains are weakened and this adds to the surface erosion phenomenon. The drug release mechanism within a polymer matrix depends on many factors such as the affinity of the drug with the surrounding medium (water). Specifically, paclitaxel is hydrophobic and this might explain the fact that some of the drug blended into the polymer matrix is not released and cannot participate to the treatment of the disease wall. This is a difficult subject. The main criticism expressed in [18] of available models for drug release from eroding surfaces is that they fail to faithfully reproduce experimental data for highly degradable polymers (the S-curve behaviour). The reader is referred to the introduction of the paper of Lao et als [18] for a comprehensive review of the literature.

A quick look at the paclitaxel release profiles suggests two types of release: S-curve type and exponential type. S-curve behaviours are similar to the ones encountered in the study of the logistic equation of populations. In [2] we introduced a simple two-parameter Ordinary Differential Equation (ODE) model that completely describes the paclitaxel release profiles from neat PCL, PLGA PEG, and PLGA polymer matrices. This model describes with greater accuracy the drug-release than the semi-empirical model of Lao et als [18] using 5 to 8 parameters.

The simplicity of our model for such a broad range of polymers indicates that somehow the quadratic structure captures the complex microphysics and chemistry of the release and degradation processes. Using a purely mathematical intuition to modelling, we have introduced in [6] a time-space three dimensional partial differential equation (PDE) model of the paclitaxel release that mimics the ODE model. The film of neat polymer is modelled as a thin flat domain whose polymer/medium interface is a quadratic semi-permeable membrane with a concentration jump at the interface.

In this approach, the diffusion process through a semi-permeable membrane is modelled as a diffusion through an interface with cracks (not to be confused with holes) where the rate of transfer of the product is proportional to the size of the concentration jump across the interface. Since the cracks have zero surface, their *size* is measured in terms of the mathematical notion of *capacity*. What is very nice about this approach is that it is based on a mathematically well documented linear model coming from the study of the Neumann sieve by Damlamian [5] in 1985. It provides a variational formulation and a mathematically tractable approach to the asymptotic analysis of a punctured membrane as the size of the holes goes to zero while preserving a strictly positive capacity that accounts for

the diffusion of the drug through the cracks.[1] Adding the non-linearity captures the effect of the internal degradation of the polymer by making the rate of mass transfer proportional to the *size of the concentration jump* across the interface.

Our approach is different from others in the literature since it deals with the nonlinearity through a quadratic condition at the interface between the polymer and the surrounding medium instead of using a time-dependent or a nonlinear diffusion. This model can be seen as a first step towards a three dimensional modelling of the release of paclitaxel from drug eluding stents coated with biodegradable polymers. It is capable of covering a wide range of biodegradable polymers potentially including the ones for which an incomplete release is experimentally observed (recall that the paclitaxel is hydrophobic).

To complete the experimental approach to this modelling, the next step would be to set up an experimental benchmark to check if the model and the mathematical assumptions on the coefficients of the model are realistic. The validation of such a model would improve the modelling of the drug release part of the global three-dimensional model of a blood vessel incorporating the lumen, the blood, the aggregated wall, and the coated stent (cf. for instance, [8]) and the subsequent studies of the effect of the pattern of the stent in [3] and the effect of the pulsative nature of the blood in [7]. Such global studies are important to determine the set of features in the modelling of the blood vessel and of the stent that should be retained in the design of the stent and the drug release dynamics.

2 ODE Model and Gradient Flow Interpretation

In the previous paper [2] we have shown an excellent fit between experimental release data [16] of paclitaxel from biodegradable neat polymers and a two-parameter quadratic ODE model of the Riccati type. We briefly recall this model.

Given an initial mass $M_0 > 0$ of drug uniformly impregnated into a *polymeric matrix*, denote by $M(t) > 0$ the *mass* of drug *released* outside the *polymer* as a function of the time $t > 0$. Denote by M_∞, $0 \le M_\infty \le M_0$, the asymptotic mass of the drug released. The ODE model was chosen of the form

$$\frac{dM}{dt}(t) = h(M(t)), \quad t > 0, \quad M(0) = 0, \tag{2.1}$$

for some quadratic right-hand side

$$h(M) \stackrel{\text{def}}{=} A_1 \left(M_\infty - M(t)\right) + A_2 \left(M_\infty - M(t)\right)^2 \tag{2.2}$$

such that $M'(0) = (A_1 + A_2 M_\infty) M_\infty > 0$. By introducing the *normalized released mass*

$$m(t) \stackrel{\text{def}}{=} M(t)/M_0, \tag{2.3}$$

[1] See also the more recent comprehensive paper [4, Theorem 5.5] using the very nice theory of periodic unfolding.

we get the following quadratic ODE model

$$\frac{dm}{dt}(t) = \left[A_1 + A_2\, M_0 \left(\frac{M_\infty}{M_0} - m(t)\right)\right] \left(\frac{M_\infty}{M_0} - m(t)\right), \quad m(0) = 0. \tag{2.4}$$

Assuming that the ratio $0 < M_\infty/M_0 \leq 1$ is known, the model is completely specified by the two parameters A_1 and $A_2 M_\infty$. When $A_2 = 0$, the model is linear; when $A_2 \neq 0$, the right-hand side is of the form

$$h(m) \stackrel{\text{def}}{=} A_2\, M_0\, (m_2 - m)(m_1 - m), \quad m_1 \stackrel{\text{def}}{=} \frac{M_\infty}{M_0}, \quad m_2 \stackrel{\text{def}}{=} \frac{A_1 + A_2\, M_\infty}{A_2\, M_0}.$$

It was shown in [2] that the following four cases can occur under the conditions $m(0) = 0$ and $m'(0) = A_1 + A_2\, M_\infty > 0$:

Case 1) (True S type)

$$A_1 > 0,\ A_2 < 0,\ \text{and}\ -m_1 < \frac{1}{2}\frac{A_1}{A_2\, M_0} \quad (\text{that is},\ -m_1 < m_2 < 0), \tag{2.5}$$

with solution

$$m(t) = m_1\, m_2\, \frac{1 - e^{-A_1\, t}}{m_2 - m_1\, e^{-A_1\, t}}$$

for which the *point of inflexion* occurs at time $t_c = -(\log(-m_2/m_1))/A_1 > 0$;

Case 2) (S type)

$$A_1 > 0,\ A_2 < 0,\ \text{and}\ \frac{1}{2}\frac{A_1}{A_2\, M_0} \leq -m_1 \quad (\text{that is},\ m_2 \leq -m_1), \tag{2.6}$$

with the solution and the *point of inflexion*

$$m(t) = m_1\, m_2\, \frac{1 - e^{-A_1\, t}}{m_2 - m_1\, e^{-A_1\, t}}, \quad t_c = -(\log(-m_2/m_1))/A_1 \leq 0;$$

Case 3) (Exponential type)

$$A_1 \geq 0 \text{ and } A_2 > 0 \quad (\text{that is},\ m_2 \geq 1), \tag{2.7}$$

with the solution and the *blow up time*

$$\begin{cases} m(t) = m_1\, m_2\, \dfrac{1 - e^{-A_1\, t}}{m_2 - m_1\, e^{-A_1\, t}}, \\ t_c = -\dfrac{\log(m_2/m_1)}{A_1} < 0, \end{cases} \quad \text{for } A_1 > 0 \text{ since } m_2 > m_1,$$

$$\begin{cases} m(t) = m_1 \dfrac{A_2\, M_\infty\, t}{1 + A_2\, M_\infty\, t}, \\ t_c = -\dfrac{1}{A_2\, M_\infty} < 0, \end{cases} \quad \text{for } A_1 = 0 \text{ since } m_2 = m_1; \tag{2.8}$$

Case 4) (True exponential) $A_1 > 0$ and $A_2 = 0$ with the solution

$$m(t) = \frac{M_\infty}{M_0}\left(1 - e^{-A_1\, t}\right), \quad t_c = -\infty.$$

The generic behaviours of the solution $m(t)$ in the above cases are illustrated in Figures 1 and 2 and the parameters tabulated in Table 1 of [6].

The ODE model has an interesting gradient flow interpretation by introducing the function

$$E(m) \stackrel{\text{def}}{=} \frac{1}{2} A_1 \left(\frac{M_\infty}{M_0} - m\right)^2 + \frac{1}{3} A_2\, M_0 \left(\frac{M_\infty}{M_0} - m\right)^3 \tag{2.9}$$

with gradient (derivative)

$$E'(m) = -A_1 \left(\frac{M_\infty}{M_0} - m\right) - A_2\, M_0 \left(\frac{M_\infty}{M_0} - m\right)^2 \tag{2.10}$$

and Hessian (second order derivative)

$$E''(m) = A_1 + 2\, A_2\, M_0 \left(\frac{M_\infty}{M_0} - m\right). \tag{2.11}$$

The ODE can now be rewritten in the form of a *gradient flow* equation

$$\frac{dm}{dt}(t) + E'(m(t)) = 0, \quad m(0) = 0. \tag{2.12}$$

This is the continuous version of a steepest descent method to minimize the functional E. So, it is expected that starting from $m(0) = 0$ with $m'(0) > 0$ the asymptotic value m_1 of the solution of the ODE (2.12) would achieve a local minimum of $E(m)$. To do that, we compute the second derivative of E under the assumption that $A_1 \geq 0$ and $m'(0) > 0$ which is equivalent to $E'(0) = -(M_\infty/M_0)[A_1 + A_2 M_\infty] < 0$. It turns out that in all cases except the second part of case 3), m_1 is a local minimum of $E(m)$. The exception corresponds to a point of inflection that can be changed into a global minimum by modifying the function E to $E(m) = (A_2\, M_0/3)\,|M_\infty/M_0 - m|^3$.

3 PDE Model of Quadratic Semi-permeable Membranes

3.1 Equations in the Polymer and the Surrounding Medium

The experimental benchmark of [16] is contained in a vial. The polymer film is deposited flat at the bottom of the vial and the vial is filled with a fluid that we shall call the *surrounding medium* (see Figure 1). The vial is closed without circulation of the fluid. Denote by Ω_p the open domain occupied by the polymer and by Ω_m the open domain occupied by the surrounding medium. Let Γ_p and Γ_m be the respective boundaries of Ω_p and Ω_m. The polymer occupies a thin square parallelepipedic region at the bottom of the vial. Its boundary is made up of the *interface* $\Gamma_{int} = \Gamma_p \cap \Gamma_m$ between the polymer and the medium

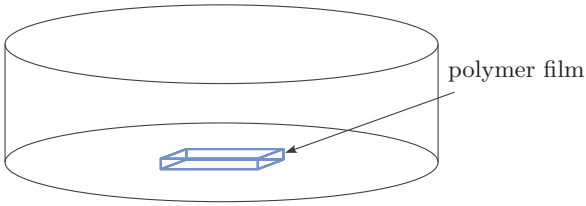


Fig. 1. The polymer film and the surrounding fluid in the vial

(top boundary and lateral boundary of Ω_p) and the bottom square boundary of Ω_p that we shall denote Γ_0.

The vial is closed without circulation of the fluid filling the vial (the medium). Within the (surrounding) medium only linear diffusion is expected with zero Neumann boundary conditions at the boundary of the vial $\Gamma_{ext} = (\Gamma_p \cup \Gamma_m) \backslash \Gamma_{int}$.

At time t, denote by $c_p(x,t)$ the concentration of the drug at the point $x \in \Omega_p$ and by $c_m(x,t)$ the concentration of the drug at point $x \in \Omega_m$. *Assume linear diffusion equations in the polymer and the surrounding medium*

$$\frac{\partial c_p}{\partial t} = \text{div}\,(D_p \nabla c_p) \text{ in } \Omega_p \tag{3.1}$$

$$\frac{\partial c_m}{\partial t} = \text{div}\,(D_m \nabla c_m) \text{ in } \Omega_m \tag{3.2}$$

with constant diffusion constants D_p and D_m and initial conditions

$$c_p(x,0) = c_0(x) = M_0/|\Omega_p| \text{ in } \Omega_p, \quad c_m(x,0) = 0 \text{ in } \Omega_m, \tag{3.3}$$

where $|\Omega_p|$ is the volume of Ω_p. *Assume that the experimental set up is closed:*

$$D_p \frac{\partial c_p}{\partial n_p} = 0 \text{eq.constriant} \Gamma_p \backslash \Gamma_{int} \quad D_m \frac{\partial c_m}{\partial n_m} = 0 \text{ on } \Gamma_m \backslash \Gamma_{int}, \tag{3.4}$$

where the unit normals n_p and n_m are exterior to the respective domains Ω_p and Ω_m. *Assume that there is no loss of product*: this yields the (affine) *constraint*

$$\forall t \geq 0, \quad M_0 \stackrel{\text{def}}{=} \int_{\Omega_p} c_0(x)\,dx = \int_{\Omega_p} c_p(x,t)\,dx + \int_{\Omega_m} c_m(x,t)\,dx, \tag{3.5}$$

where M_0 is the total mass of product. By integrating (3.1) over Ω_p and (3.2) over Ω_m and by using the constraint (3.5), we get

$$\begin{aligned} \Rightarrow 0 &= \int_{\Omega_p} \frac{\partial c_p}{\partial t}(x,t)\,dx + \int_{\Omega_m} \frac{\partial c_m}{\partial t}(x,t)\,dx \\ &= \int_{\Omega_p} \text{div}\,(D_p \nabla c_p)(x,t)\,dx + \int_{\Omega_m} \text{div}\,(D_m \nabla c_m)(x,t)\,dx \\ &= \int_{\Gamma_p} D_p \frac{\partial c_p}{\partial n_p}(x,t)\,d\Gamma + \int_{\Gamma_m} D_m \frac{\partial c_m}{\partial n_m}(x,t)\,d\Gamma. \end{aligned} \tag{3.6}$$

Finally, by using the boundary conditions (3.4) we get

$$\int_{\Gamma_{int}} \left[D_p \frac{\partial c_p}{\partial n_p}(x,t) + D_m \frac{\partial c_m}{\partial n_m}(x,t) \right] d\Gamma = 0. \tag{3.7}$$

It remains to specify the conditions at the interface Γ_{int}.

3.2 Conditions at the Interface

In order to incorporate the microphysics taking place in the thin film of polymer, it is assumed that the interface behaves as a semi-permeable membrane with micro fissures through which the drug diffuses into the surrounding medium. Many empirical and theoretical models of such membranes have been studied in the literature and in different contexts. One mathematically interesting model of a semi-permeable membrane is to assume that the interface is a membrane punctured with small holes whose size goes to zero while preserving a strictly positive *capacity*[2] in the limiting process. In other words the membrane is *fissured* or *cracked* and the drug diffuses through the cracks. This problem has been studied from the mathematical point of view under the name of the *Neumann sieve* by A. Damlamian [5] in 1985. From the physical point of view, it can be assimilated with a *semi-permeable membrane*.

In this section we consider an evolution equation of the form

$$\frac{\partial c}{\partial t}(t) + A(c(t)) = 0, \quad c(0) = M_0/|\Omega_p|\, \chi_{\Omega_p}, \tag{3.8}$$

where the operator A is now quadratic in $c(t)$. Since the domain Ω_p is thin, it is reasonable to put the nonlinearity at the interface Γ_{int} rather than on Ω_p via a diffusion coefficient $D_p(c)$ that depends on c:

$$\begin{aligned} -\frac{d}{dt}\int_{\Omega_p} c_p(t)\, dx &= \frac{d}{dt}\int_{\Omega_m} c_m(t)\, dx \\ &= \int_{\Gamma_{int}} \left[k_1 + k_2 \frac{|\Omega_p|}{M_0} |c_p(t) - c_m(t)| \right] (c_p(t) - c_m(t))\, d\Gamma \end{aligned} \tag{3.9}$$

for some constant k_2. Note that we have introduced a scaling by the initial concentration of product $M_0/|\Omega_p|$ of the drug so that k_1 and k_2 are parameters of the same physical dimension.

Now consider the (cubic) functional

$$\begin{aligned} E(v) \overset{\text{def}}{=} &\frac{1}{2}\int_{\Omega_p} D_p |\nabla v_p|^2\, dx + \frac{1}{2}\int_{\Omega_m} D_m |\nabla v_m|^2\, dx \\ &+ \int_{\Gamma_{int}} \frac{1}{2} k_1 |v_p - v_m|^2 + \frac{1}{3} k_2 \frac{|\Omega_p|}{M_0} |v_p - v_m|^3\, d\Gamma \end{aligned} \tag{3.10}$$

$$v_p \overset{\text{def}}{=} v|_{\Omega_p}, \quad v_m \overset{\text{def}}{=} v|_{\Omega_m} \tag{3.11}$$

[2] The *capacity* of a set is a mathematical notion. For instance a finite segment in the plane has zero area but finite capacity. Roughly speaking, the capacity is a "measure" of the cracks.

defined on the space $H^1(\Omega_p \cup \Omega_m)$ with a *crack* $\Gamma_{int} = \Gamma_p \cap \Gamma_m$ in $\Omega_p \cup \Omega_m$ along which the function v can have a *jump discontinuity* $[v] = v_m - v_p$. This convex non quadratic variational formulation is similar to the T^4 radiation law for the temperature T of a radiating body in free space (cf., for instance, [9]).

We do not impose the continuity of the concentrations at the interface. Taking into account the constraint on the total mass of product, we look for a solution $c(t)$ at time $t > 0$ in the affine subspace

$$V_{M_0}^{pm} \stackrel{\text{def}}{=} \begin{cases} \left\{ c \in H^1(\Omega_p \cup \Omega_m) : \displaystyle\int_{\Omega_p \cup \Omega_m} c(x)\, dx = M_0 \right\}, & \text{if } k_2 = 0,\ k_1 > 0, \\ \left\{ \begin{array}{c} c \in H^1(\Omega_p \cup \Omega_m) : \displaystyle\int_{\Omega_p \cup \Omega_m} c(x)\, dx = M_0 \\ v_p - v_m \in L^3(\Gamma_{int}) \end{array} \right\}, & \text{if } k_2 > 0, \end{cases}$$

of $H^1(\Omega_p \cup \Omega_m)$. In the first case

$$\Big[\int_{\Omega_p} |\nabla v_p|^2 \, dx + \int_{\Omega_m} |\nabla v_m|^2 \, dx + \int_{\Gamma_{int}} |v_p - v_m|^2 \, d\Gamma \Big]^{1/2} \tag{3.12}$$

is an equivalent norm on $V_{M_0}^{pm}$; in the second case

$$\left[\int_{\Omega_p} |\nabla v_p|^2 \, dx + \int_{\Omega_m} |\nabla v_m|^2 \, dx \right]^{1/2} + \left[\int_{\Gamma_{int}} |v_p - v_m|^3 \, d\Gamma \right]^{1/3} \tag{3.13}$$

is an equivalent norm on $V_{M_0}^{pm}$ (cf., for instance, [9] with the T^4 radiation law for the temperature T in free space).

The directional derivative of E is

$$\begin{aligned} dE(u; v) = & \int_{\Omega_p} D_p \nabla u \cdot \nabla v \, dx + \int_{\Omega_m} D_m \nabla u \cdot \nabla v \, dx \\ & + \int_{\Gamma_{int}} k_2 \frac{|\Omega_p|}{M_0} \, |u_p - u_m| \, (u_p - u_m)\, (v_p - v_m) \\ & \qquad + k_1 \, (u_p - u_m)\, (v_p - v_m) \, d\Gamma \end{aligned} \tag{3.14}$$

$$u_p \stackrel{\text{def}}{=} u|_{\Omega_p}, \quad u_m \stackrel{\text{def}}{=} u|_{\Omega_m}, \quad v_p \stackrel{\text{def}}{=} v|_{\Omega_p}, \quad v_m \stackrel{\text{def}}{=} v|_{\Omega_m}. \tag{3.15}$$

We are interested in the *stationary points* $c = (c_p, c_m) \in V_{M_0}$ of E that are the solutions of the variational equation

$$\exists c \in V_{M_0}^{pm}, \quad dE(c; v) = 0, \quad \forall v \in V_0^{pm}, \tag{3.16}$$

$$V_0^{pm} \stackrel{\text{def}}{=} \begin{cases} \left\{ c \in H^1(\Omega_p \cup \Omega_m) : \displaystyle\int_{\Omega_p \cup \Omega_m} c(x)\, dx = 0 \right\}, & \text{if } k_2 = 0,\ k_1 > 0, \\ \left\{ \begin{array}{c} c \in H^1(\Omega_p \cup \Omega_m) : \displaystyle\int_{\Omega_p \cup \Omega_m} c(x)\, dx = 0 \\ v_p - v_m \in L^3(\Gamma_{int}) \end{array} \right\}, & \text{if } k_2 > 0. \end{cases} \tag{3.17}$$

Again $dE(c;v) = 0$ for all constant functions v and V_0^{pm} can be replaced by $H^1(\Omega_p \cup \Omega_m)$:

$$\exists c \in V_{M_0}^{pm}, \quad dE(c;v) = 0, \quad \forall v \in H^1(\Omega_p \cup \Omega_m).$$

It yields a complete set of conditions at the interface and the following system of equations

$$\operatorname{div}(D_p \nabla c_p) = 0 \text{ in } \Omega_p, \quad \operatorname{div}(D_p \nabla c_m) = 0 \text{ in } \Omega_m \tag{3.18}$$

$$D_p \frac{\partial c_p}{\partial n_p} + k_2 \frac{|\Omega_p|}{M_0} |c_p - c_m| (c_p - c_m) + k_1 (c_p - c_m) = 0 \text{ on } \Gamma_{int} \tag{3.19}$$

$$D_m \frac{\partial c_m}{\partial n_m} - \left[k_2 \frac{|\Omega_p|}{M_0} |c_p - c_m| (c_p - c_m) + k_1 (c_p - c_m) \right] = 0 \text{ on } \Gamma_{int} \tag{3.20}$$

$$D_p \frac{\partial c_p}{\partial n_p} = 0 \text{ on } \Gamma_p \backslash \Gamma_{int}, \quad D_m \frac{\partial c_m}{\partial n_m} = 0 \text{ on } \Gamma_m \backslash \Gamma_{int} \tag{3.21}$$

$$\int_{\Omega_p} c_p \, dx + \int_{\Omega_m} c_m \, dx = M_0. \tag{3.22}$$

From the mathematical viewpoint, the condition involving $|c_p - c_m| (c_p - c_m)$ is the analogue of the condition $|T - T_m|^3 (T - T_m)$ (usually written $(T - T_m)^4$) on the temperature of a radiating body (cf., for instance, [9]). The thin layer of polymer behaves as a *nonlinear* semi-permeable membrane. The second order directional derivative of E is

$$\begin{aligned} d^2E(u;v;w) = &\int_{\Omega_p} D_p \nabla w \cdot \nabla v \, dx + \int_{\Omega_m} D_m \nabla w \cdot \nabla v \, dx \\ &+ \int_{\Gamma_{int}} \left[2\, k_2 \frac{|\Omega_p|}{M_0} |u_p - u_m| + k_1 \right] (w_p - w_m)(v_p - v_m) \, d\Gamma \end{aligned} \tag{3.23}$$

$$\begin{aligned} \Rightarrow \; d^2E(u;v;v) = &\int_{\Omega_p} D_p |\nabla v|^2 \, dx + \int_{\Omega_m} D_m |\nabla v|^2 \, dx \\ &+ \int_{\Gamma_{int}} \left[2\, k_2 \frac{|\Omega_p|}{M_0} |u_p - u_m| + k_1 \right] |v_p - v_m|^2 \, d\Gamma. \end{aligned} \tag{3.24}$$

Since E is a cubic functional, local minima and local maxima can both occur depending on the signs and magnitudes of the constants k_1 and k_2. A local minimum $u \in V_{M_0}^{pm}$ is characterized by

$$\forall v \in V_0^{pm} \quad dE(u;v) = 0 \quad \text{and} \quad \forall 0 \neq v \in V_0^{pm} \quad d^2E(u;v;v) > 0$$

and a local maximum $u \in V_{M_0}^{pm}$ by

$$\forall v \in V_0^{pm} \quad dE(u;v) = 0 \quad \text{and} \quad \forall 0 \neq v \in V_0^{pm} \quad d^2E(u;v;v) < 0.$$

Going back to the evolution equation (3.8) using the above conditions at the interface, we get the following system of equations

$$
\begin{aligned}
\frac{\partial c_p}{\partial t} = \operatorname{div}(D_p \nabla c_p) \text{ in } \Omega_p, \quad \frac{\partial c_m}{\partial t} = \operatorname{div}(D_m \nabla c_m) \text{ in } \Omega_m \\
c_p(x,0) = M_0/|\Omega_p|\, \chi_{\Omega_p}(x) \text{ in } \Omega_p, \quad c_m(x,0) = 0 \text{ in } \Omega_m \\
D_p \frac{\partial c_p}{\partial n_p} + k_2 \frac{|\Omega_p|}{M_0} |c_p - c_m| (c_p - c_m) + k_1 (c_p - c_m) = 0 \text{ on } \Gamma_{int} \\
D_m \frac{\partial c_m}{\partial n_m} - \left[k_2 \frac{|\Omega_p|}{M_0} |c_p - c_m| (c_p - c_m) + k_1 (c_p - c_m) \right] = 0 \text{ on } \Gamma_{int} \\
D_p \frac{\partial c_p}{\partial n_p} = 0 \text{ on } \Gamma_p \backslash \Gamma_{int}, \quad D_m \frac{\partial c_m}{\partial n_m} = 0 \text{ on } \Gamma_m \backslash \Gamma_{int} \\
\int_{\Omega_p} c_p \, dx + \int_{\Omega_m} c_m \, dx = M_0.
\end{aligned}
\tag{3.25}
$$

The nonlinear condition on Γ_{int}

$$
\begin{aligned}
D_m \frac{\partial c_m}{\partial n_m} &= k_2 \frac{|\Omega_p|}{M_0} |c_p - c_m| (c_p - c_m) + k_1 (c_p - c_m) \\
&= \underbrace{\left(k_2 \frac{|\Omega_p|}{M_0} |c_p - c_m| + k_1 \right)}_{k(c)} (c_p - c_m)
\end{aligned}
$$

says that $k(c)$ is an affine function of the *size* of the jump. This means that the rate of transfer of the product across the interface is large when the absolute value of the concentration jump is large. Assuming that $k_1 \geq 0$, when $k_2 > 0$ it decreases to k_1 when the size of the jump goes to zero; when $k_2 < 0$ it increases to k_1 when the size of the jump goes to zero.

Remark 1. *When $k_2 > 0$, it would not be appropriate to remove the absolute value on $c_p - c_m$ in the term k' of the previous identity. This would give the expression*

$$
D_m \frac{\partial c_m}{\partial n_m} = \underbrace{\left(k_2 \frac{|\Omega_p|}{M_0} (c_p - c_m) + k_1 \right)}_{k'(c)} (c_p - c_m),
$$

where, if the size of the jump is large, $k'(c) > 0$ is large, $\partial c_m / \partial n_m > 0$ is large, and the diffusion of product would be from the medium to the polymer even when $c_p > c_m$, that is, when the concentration in the polymer is larger than the one in the medium. However, it is interesting to note that various behaviours can be modelled by replacing $|c_p - c_m|$ by the plus $[c_p - c_m]^+ = \max\{0, c_p - c_m\}$ or the minus $[c_p - c_m]^- = \max\{0, -(c_p - c_m)\}$ functions or introducing a threshold $\theta > 0$ $\max\{|c_p - c_m| - \tau, \theta\}$.

3.3 Relation between the PDE and the ODE Models

Since $|\Omega_p|$ is much smaller than $|\Omega_m|$, this last equation is related to the quadratic ODE model by making the same assumptions on the concentrations on Γ_{int} as in the previous section:

$$c_p(x,t) \simeq \frac{1}{|\Omega_p|}\int_{\Omega_p} c_p(x,t)\,dx \quad \text{and} \quad c_m(x,t) \simeq \frac{1}{|\Omega_m|}\int_{\Omega_m} c_m(x,t)\,dx \tag{3.26}$$

$$\Rightarrow\ c_p(x,t) - c_m(x,t) \simeq \frac{1}{|\Omega_p|}\left[M_0 - M_m(t)\right],$$

where

$$M_m(t) \stackrel{\text{def}}{=} \int_{\Omega_m} c_m(x,t)\,dx \tag{3.27}$$

is the mass released at time t in the medium and

$$\frac{dM_m}{dt}(t) = \frac{|\Gamma_{int}|}{|\Omega_p|}\left[k_1 + \frac{k_2}{M_0}\left|M_0 - M_m(t)\right|\right]\left(M_0 - M_m(t)\right) \tag{3.28}$$

$$\Rightarrow\ \frac{dm_m}{dt}(t) = \frac{1}{h}\left[k_1 + k_2\left|1 - m_m(t)\right|\right]\left(1 - m_m(t)\right), \quad m_m(t) \stackrel{\text{def}}{=} \frac{M_m(t)}{M_0}, \tag{3.29}$$

where $h = |\Omega_p|/|\Gamma_{int}|$ is the thickness of the polymer. This would correspond to $A_1 = k_1/h$ and $A_2 = k_2/h$ in the ODE model. The thickness h is an important *parameter*: the thinner the polymer the faster the release. If k_1 and k_2 are constants, m_m can be normalized through the change of variable $t \mapsto \tau = t/h$.

References

1. Batycky, R.P., Hanes, J., Langer, R., Edwards, D.A.: A theoretical model of erosion and macromolecular drug release from biodegrading microspheres. J. Pharm. Sci. 86, 1464–1477 (1997)
2. Blanchet, G., Delfour, M.C., Garon, A.: Quadratic models to fit experimental data of Paclitaxel release kinetics from biodegradable polymers. SIAM J. on Applied Mathematics (Special Issue on Mathematical Modeling of Controlled Drug Delivery) 71(6), 2269–2286 (2011)
3. Bourgeois, É., Delfour, M.C.: General patterns and asymptotic dose in the design of coated stents. Computer Methods in Biomechanics and Biomedical Engineering 11(4), 323–334 (2008)
4. Cioranescu, D., Damlamian, A., Griso, G., Onofrei, D.: The periodic unfolding method for perforated domains and Neumann sieve models. J. Math. Pures Appl. 89, 248–277 (2008)
5. Damlamian, A.: Le problème de la passoire de Neumann (French) [The Neumann sieve problem]. Rend. Sem. Mat. Univ. Politec. Torino 43, 427–450 (1985)
6. Delfour, M.C.: Drug release kinetics from biodegradable polymers via partial differential equations models. Acta Appl. Math. 118, 161–183 (2012)
7. Delfour, M.C., Garon, A.: New equations for the dose under pulsative/periodic conditions in the design of coated stents. Computer Methods in Biomechanics and Biomedical Engineering 13(1), 19–34 (2010)

8. Delfour, M.C., Garon, A., Longo, V.: Modeling and design of stents to optimize the effect of the dose. SIAM J. on Applied Mathematics 65(3), 858–881 (2005)
9. Delfour, M.C., Payre, G., Zolésio, J.-P.: Approximation of nonlinear problems associated with radiating bodies in space. SIAM J. on Numerical Analysis 24, 1077–1094 (1987)
10. Faisant, N., Akiki, J., Siepmann, J., Benoit, J.P., Siepmann, J.: Effects of the type of release medium on drug release from PLGA-based microparticles: experiment and theory. Int. J. Pharm. 314, 189–197 (2006)
11. Farb, A., Heller, P.F., Shroff, S., Cheng, L., Kolodgie, F.D., Carter, A.J., Scott, D.S., Froehlich, J., Virmani, R.: Pathological analysis of local delivery of paclitaxel via a polymer-coated stent. Circulation 104(4), 473–479 (2001)
12. Gopferich, A.: Polymer bulk erosion. Macromolecules 30, 2598–2604 (1997)
13. Higuchi, T.: Mechanism of sustained action mediation: theoretical analysis of rate of release of solid drugs dispersed in solid matrices. J. Pharm. Sci. 52, 1145–1149 (1963)
14. Joner, M., Finn, A.V., Farb, A., Mont, E.K., Kolodgie, F.D., Ladich, E., et al.: Pathology of drug-eluting stents in humans – delayed healing and late thrombotic risk. J. Am. Coll. Cardiol. 48(1), 193–202 (2006)
15. Kataoka, T., Grube, E., Honda, Y., Morino, Y., Hur, S.-H., Bonneau, H.N., Colombo, A., Di Mario, C., Guagliumi, G., Hauptmann, K.E., Pitney, M.R., Lansky, A.J., Stertzer, S.H., Yock, P.G., Fitzgerald, P.J.: 7-Hexanoyltaxol-Eluting Stent for Prevention of Neointimal Growth: An Intravascular Ultrasound Analysis From the Study to Compare REstenosis rate between QueST and QuaDS-QP2 (SCORE). Circulation 106, 1788–1793 (2002)
16. Lao, L.L., Venkatraman, S.S.: Adjustable paclitaxel release kinetics and its efficacy to inhibit smooth muscle cells proliferation. J. Control. Release 130, 9–14 (2008)
17. Lao, L.L., Venkatraman, S.S.: Paclitaxel release from single and double layered poly (DL-lactide-co-glycolide)/poly (L-lactide) film for biodegradable coronary stent application. J. Biomed. Mater. Res. A 87A(1), 1–7 (2008)
18. Lao, L.L., Venkatraman, S.S., Peppas, N.A.: Modeling of drug release from biodegradable polymer blends. Eur. J. Pharm. Biopharm. 70, 796–803 (2008)
19. Lao, L.L., Venkatraman, S.S., Peppas, N.A.: A novel model and experimental analysis of hydrophilic and hydrophobic agent release from biodegradable polymers. J. Biomed. Mater. Res. A 90(4), 1054–1065 (2009)
20. Lemaire, V., Bélair, J., Hildgen, P.: Structural modeling of a drug release from biodegradable porous matrices based on a combined diffusion/erosion process. Int. J. Pharm. 258, 95–107 (2003)
21. Regar, E., Sianos, G., Serruys, P.W.: Stent development and local drug delivery. British Medical Bulletin 59(1), 227–248 (2001)
22. Siepmann, J., Gopferich, A.: Mathematical modeling of bioerodible, polymeric drug delivery systems. Adv. Drug Deliv. Rev. 48, 229–247 (2001)

A Critical Note on Empirical (Sample Average, Monte Carlo) Approximation of Solutions to Chance Constrained Programs

René Henrion*

Weierstrass Institute Berlin, Mohrenstr. 39, 10117 Berlin, Germany

Abstract. The solution of chance constrained optimization problems by means of empirical approximation of the underlying multivariate distribution has recently become a popular alternative to conventional methods due to the efficient application of appropriate mixed integer programming techniques. As the complexity of required computations depends on the sample size used for approximation, exponential estimates for the precision of optimal solutions or optimal values have become a key argument for controlling the sample size. However, these exponential estimates may involve unknown constants such that the required sample size to approximate the solution of a problem may become arbitrarily large. We will illustrate this effect for Gaussian distributions.

Keywords: chance constrained programming, probabilistic constraints, empirical approximation, sample average approximation, convergence of solution sets, sample size.

1 Introduction

A chance constrained optimization problem has the general form

$$\min\{g(x)|\mathbb{P}(h(x,\xi)\geq 0)\geq p, x\in C\}, \tag{1}$$

where $x\in\mathbb{R}^n$ is a decision vector, $g:\mathbb{R}^n\to\mathbb{R}$ is an objective function, ξ is an s-dimensional random vector defined on some probability space $(\Omega,\mathcal{A},\mathbb{P})$, $h:\mathbb{R}^n\times\mathbb{R}^s\to\mathbb{R}^m$ is a Borel measurable with respect to the second argument mapping, $C\subseteq\mathbb{R}^n$ represents some abstract deterministic constraint and $p\in[0,1]$ is a fixed probability level. The random inequality system $h(x,\xi)\geq 0$ may reflect some technological constraints in engineering problems which are affected by uncertainty. Since usually a decision x has to be taken before the uncertain parameter ξ is observed, it has become a standard approach of robust modeling to define x as feasible, whenever the probability of satisfying the random inequality is at least p. This is expressed in the so-called chance constraint $\mathbb{P}(h(x,\xi)\geq 0)\geq p$. For a standard introduction to chance constrained programming we refer to the classical monograph [6] and to the more recent treatise in [7].

* This work was supported by the DFG Research Center Matheon "Mathematics for key technologies" in Berlin

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 25–37, 2013.

Recent progress in mixed integer programming techniques tailored to chance constraints has led to the idea of solving (1) by empirical approximation (or sample average approximation) of the original random vector ξ (e.g., [4,5]). This means that an i.i.d. sample $\xi^1, \ldots, \xi^N$ of size N is drawn from the distribution of ξ and that the law $\mathbb{P} \circ \xi^{-1}$ of ξ in (1) is replaced by the empirical measure

$$N^{-1} \sum_{i=1}^{N} \delta(\xi^i),$$

where $\delta(z)$ is the Dirac measure centered on z. Doing so, the original chance constraint $\mathbb{P}(h(x,\xi) \geq 0) \geq p$ turns into its empirical counterpart

$$\# \left\{ i | h(x,\xi^i) \geq 0 \right\} \geq pN.$$

Of course this change of constraint leads to another optimization problem whose solution deviates from the solution of (1) and one has to answer two questions: do the approximating solutions converge with $N \to \infty$ and if so, how large should N be chosen in order to guarantee a given precision of the solution obtained? The first question is answered by the stability theory of chance constrained programming mainly developed in [1,2] which applies to arbitrary approximations of the original distribution and, in particular, to empirical ones. Under some explicitly verifiable conditions, not only qualitative convergence of approximating solutions can be guaranteed but also rates for this convergence can be derived. The latter allow us to obtain exponential bounds (in terms of sample size) for the precision of solutions in case of empirical approximations. However, one has to take into account that the exponential term involves apart from the sample size N also some other constants which may depend on the conditioning of the problem and may be hard to estimate. Thus, exponential estimates of solutions do not exclude the need for a large sample size even in small dimension in order to arrive at a reasonable precision of the solution. This situation occurs in particuar if the law of the original random vector ξ has unbounded support. We will illustrate and explain this effect for a multivariate Gaussian distribution (but similar observations could be made for other classes of multivariate distributions such as log-normal or t-). In order to keep the presentation as simple as possible we restrict ourselves to the simplest yet meaningful instance of problem (1):

$$\min\{c^T x \mid \mathbb{P}(\xi \leq x) \geq p\}.$$

This means that we consider just linear objective functions, we forget about additional abstract deterministic constraints and we assume the chance constraint being in elementary separated form. Recalling the definition of the distribution function $F_\xi(x) := \mathbb{P}(\xi \leq x)$ of a random vector ξ, we may rewrite this problem as

$$\min\{c^T x \mid F_\xi(x) \geq p\} \qquad (\mathrm{P}_{\mathrm{c},\xi,\mathrm{p}}). \tag{2}$$

In order to emphasize the dependence on the problem data c, ξ and p, we label problem (2) as $(\mathrm{P}_{\mathrm{c},\xi,\mathrm{p}})$. Before coming back to the issue of empirical distributions

discussed above, we derive in the next section our main result demonstrating the difficulty of approximating a chance constrained program with unbounded support of the underlying distribution by means of distributions with bounded support. Note that, in particular, empirical measures have bounded support.

2 Main Result

We start by recalling the following well-known representation for partial derivatives of Gaussian distribution functions. We make use of the familiar notation $\xi \sim \mathcal{N}(\mu, \Sigma)$ to designate a Gaussian random vector with expectation μ and covariance matrix Σ.

Theorem 1 ([6], p. 204). *Let $\xi \sim \mathcal{N}(\mu, \Sigma)$ with some positive definite covariance matrix $\Sigma = (\sigma_{ij})$ of order (s, s). Then, the distribution function F_ξ is continuously differentiable at any $z \in \mathbb{R}^s$ and*

$$\frac{\partial F_\xi}{\partial z_j}(z) = f_{\xi_j}(z_j) \cdot F_{\tilde{\xi}(z_j)}(z_1, \ldots, z_{j-1}, z_{j+1} \ldots, z_s) \quad (j = 1, \ldots, s).$$

Here, f_{ξ_j} denotes the one-dimensional Gaussian density of the component ξ_j, $\tilde{\xi}(z_j)$ is an (s-1)-dimensional Gaussian random vector distributed according to $\tilde{\xi}(z_j) \sim \mathcal{N}(\hat{\mu}, \hat{\Sigma})$, $\hat{\mu}$ results from the vector $\mu + \sigma_{jj}^{-1}(z_j - \mu_j)\sigma_j$ by deleting component j and $\hat{\Sigma}$ results from the matrix $\Sigma - \sigma_{jj}^{-1}\sigma_j\sigma_j^T$ by deleting row j and column j, where σ_j refers to column j of Σ.

Corollary 1. *In the context of the previous Theorem, one has that*

$$\frac{\partial F_\xi}{\partial z_j}(z) > 0 \quad \forall z \in \mathbb{R}^s, \ \forall j \in \{1, \ldots, s\}.$$

Proof. This follows immediately from the formula in Theorem 1 and the fact that both the density and the distribution function of a regular Gaussian distribution are strictly positive.

With each problem $(\mathrm{P}_{c,\xi,\mathrm{p}})$ in (2) we associate its (possibly empty) solution set

$$\Psi_{c,\xi,p} := \arg\min\{c^T x \mid \mathbb{P}(\xi \leq x) \geq p\}.$$

Lemma 1. *For problem $(\mathrm{P}_{c,\xi,\mathrm{p}})$ in (2) assume that $c_i > 0$ for $i = 1, \ldots, s$. Then,*

$$\Psi_{c,\xi,p} \subseteq [a, b] := \{x \in \mathbb{R}^s \mid a_i \leq x_i \leq b_i \quad (i = 1, \ldots, s)\},$$

where, with 'supp' denoting the support of a random vector,

$$a_i := \inf\{z_i | z \in \operatorname{supp} \xi\} \quad b_i := \sup\{z_i | z \in \operatorname{supp} \xi\} \quad (i = 1, \ldots, s).$$

Proof. Assume that there is some $x^* \in \Psi_{c,\xi,p}$ and some i such that $x_i^* > b_i$. Then, $b_i < \infty$ and we may define $\bar{x}$ by

$$\bar{x}_i := (x_i^* + b_i)/2, \quad \bar{x}_j := x_j^* \quad (j \neq i).$$

From x^* being feasible for problem $(\mathrm{P}_{\mathrm{c},\xi,\mathrm{p}})$, we conclude that

$$p \leq \mathbb{P}(\xi \leq x^*) = \mathbb{P}(\xi \leq \bar{x}) + \mathbb{P}(\bar{x}_i \leq \xi_i \leq x_i^*,\ \xi_j \leq \bar{x}_j \quad (j \neq i)).$$

Now, since $\bar{x}_i > b_i$, it follows that $\{x \in \mathbb{R}^n \mid \bar{x}_i \leq x_i \leq x_i^*\} \cap \operatorname{supp} \xi = \emptyset$, whence

$$\mathbb{P}(\bar{x}_i \leq \xi_i \leq x_i^*,\ \xi_j \leq \bar{x}_j \quad (j \neq i)) = 0$$

and $\mathbb{P}(\xi \leq \bar{x}) \geq p$. Therefore, $\bar{x}$ too is feasible for problem $(\mathrm{P}_{\mathrm{c},\xi,\mathrm{p}})$. On the other hand, $c^T \bar{x} < c^T x^*$ due to $c_i > 0$, $\bar{x}_i < x_i^*$ and $\bar{x}_j = x_j^*$ for $j \neq i$. This contradicts the assumption $x^* \in \Psi_{c,\xi,p}$. Consequently, $x^* \leq b$ for any $x^* \in \Psi_{c,\xi,p}$. Similarly one shows that $x^* \geq a$ for any $x^* \in \Psi_{c,\xi,p}$. It follows that $\Psi_{c,\xi,p} \subseteq [a,b]$.

Now, we are in a position to state our main result:

Theorem 2. *Let $s > 1$. Assume that ξ has a regular normal distribution according to $\xi \sim \mathcal{N}(\mu, \Sigma)$ and that η is a random vector with compact support. Then, for any $p \in (0,1)$ there exists a sequence $c^{(n)} \in \mathbb{R}^n$ with $c_i^{(n)} > 0$ for $i = 1, \ldots, s$ such that $\Psi_{c^{(n)},\xi,p} \neq \emptyset$ and*

$$\inf\{\|x - y\| \mid x \in \Psi_{c^{(n)},\xi,p}, y \in \Psi_{c^{(n)},\eta,p}\} > n \quad \forall n \in \mathbb{N}. \tag{3}$$

Proof. Fix an arbitrary $p \in (0,1)$ and an arbitrary $n \in \mathbb{N}$. Since $\operatorname{supp} \eta$ is compact, we may apply Lemma 1 to η in order to derive the existence of some compact(!) rectangle $[a,b]$ such that

$$\Psi_{c,\eta,p} \subseteq [a,b] \quad \forall c \in \mathbb{R}^s : c_i > 0 \quad (i = 1, \ldots, s). \tag{4}$$

With $[a,b]$ being compact, we may choose $L_n > 0$ such that

$$\|y - z\| \geq n \quad \forall y \in [a,b],\ \forall z : \|z\| \geq L_n. \tag{5}$$

Since $s > 1$ by assumption, we may define the ratio

$$\varkappa(z) := \frac{\partial F_\xi}{\partial z_1}(z) \Big/ \frac{\partial F_\xi}{\partial z_2}(z) \quad (z \in \mathbb{R}^s). \tag{6}$$

Note that the partial derivatives of F_ξ are continuous (see Theorem 1) and strictly positive (see Corollary 1), hence $\varkappa$ is correctly defined and continuous. Consequently the quantity

$$\bar{\varkappa} := \sup\{\varkappa(z) | z \in \mathbb{B}(0, L_n)\} \tag{7}$$

is finite. Next, let q_p be the p-quantile of the first marginal distribution of ξ, i.e., of the distribution of the first component ξ_1. Since ξ_1 has a one-dimensional normal distribution and $p \in (0,1)$, $q_p \in \mathbb{R}$ is uniquely defined by $\mathbb{P}(\xi_1 \leq q_p) = p$. We claim that for each $k \in \mathbb{N}$ there exists some $t_k \in \mathbb{R}$ such that

$$F_\xi\left(q_p + k^{-1}, t_k, \ldots, t_k\right) = p. \tag{8}$$

Indeed, for arbitrarily fixed k one has that

$$\lim_{\tau \to -\infty} F_\xi\left(q_p + k^{-1}, \tau, \ldots, \tau\right) = 0$$

as a general property of distribution functions and that

$$p = \mathbb{P}(\xi_1 \leq q_p) < \mathbb{P}\left(\xi_1 \leq q_p + k^{-1}\right) = \lim_{\tau \to \infty} F_\xi\left(q_p + k^{-1}, \tau, \ldots, \tau\right).$$

Now, the existence of t_k with the desired property (8) follows from continuity of F_ξ and from $p > 0$. Next we claim that $t_k \to_k \infty$. If there existed some subsequence t_{k_l} and some $r \in \mathbb{R}$ such that $t_{k_l} \leq r$ for all $l \in \mathbb{N}$, then

$$F_\xi\left(q_p + k_l^{-1}, r, \ldots, r\right) \geq F_\xi\left(q_p + k_l^{-1}, t_{k_l}, \ldots, t_{k_l}\right) = p \quad \forall l \in \mathbb{N}$$

which again by continuity of F_ξ as a function of each of its components yields the contradiction

$$p \leq F_\xi(q_p, r, \ldots, r) = \mathbb{P}(\xi_1 \leq q_p, \xi_2 \leq r, \ldots, \xi_s \leq r) < \mathbb{P}(\xi_1 \leq q_p) = p.$$

Here, the strict inequality relies on the fact that a regular Gaussian distribution has a density which is strictly positive everywhere. Hence, we have shown that $t_k \to_k \infty$. Consider the sequence

$$z^{(k)} := \left(q_p + k^{-1}, t_k, \ldots, t_k\right) \quad (k \in \mathbb{N}).$$

Then, by (8),

$$F_\xi\left(z^{(k)}\right) = p \quad (k \in \mathbb{N}). \tag{9}$$

Moreover, Theorem 1 yields that

$$\frac{\partial F_\xi}{\partial z_1}(z^{(k)}) = f_{\xi_1}(q_p + k^{-1}) \cdot F_{\tilde{\xi}\left(z_1^{(k)}\right)}(t_k, \ldots, t_k), \tag{10}$$

where f_{ξ_1} denotes the one-dimensional Gaussian density of the component ξ_1 and $\tilde{\xi}\left(z_1^{(k)}\right)$ is an (s-1)-dimensional Gaussian random vector distributed according to $\tilde{\xi}\left(z_1^{(k)}\right) \sim \mathcal{N}(\hat{\mu}^{(k)}, \hat{\Sigma})$ where $\hat{\mu}^{(k)}$ and $\hat{\Sigma}$ result from the original parameters μ and Σ, respectively, of ξ as detailed in Theorem 1. In particular,

$$\hat{\mu}^{(k)} = (\mu_2, \ldots, \mu_s) + \sigma_{11}^{-1}\left(q_p + k^{-1} - \mu_1\right)(\sigma_{21}, \ldots, \sigma_{s1})$$

and we observe that

$$\hat{\mu}^{(k)} \to_k \hat{\mu} := (\mu_2, \ldots, \mu_s) + \sigma_{11}^{-1} (q_p - \mu_1) (\sigma_{21}, \ldots, \sigma_{s1}) . \tag{11}$$

Note also that in contrast to $\hat{\mu}^{(k)}$, the covariance matrix $\hat{\Sigma}$ does not depend on the index k. Now we define the centered random vector

$$\hat{\xi} := \tilde{\xi} \left(z_1^{(k)} \right) - \hat{\mu}^{(k)} \sim \mathcal{N}(0, \hat{\Sigma}),$$

whose distribution does no longer depend on the index k. Exploiting the relation

$$F_{\tilde{\xi}\left(z_1^{(k)}\right)} (t_k, \ldots, t_k) = F_{\hat{\xi}} \left(t_k - \hat{\mu}_1^{(k)}, \ldots, t_k - \hat{\mu}_{s-1}^{(k)} \right)$$

and noting that all components of the argument $\left(t_k - \hat{\mu}_1^{(k)}, \ldots, t_k - \hat{\mu}_{s-1}^{(k)} \right)$ tend to infinity due to (11) and $t_k \to_k \infty$, we conclude that

$$F_{\hat{\xi}} \left(t_k - \hat{\mu}_1^{(k)}, \ldots, t_k - \hat{\mu}_{s-1}^{(k)} \right) \to_k 1$$

because the values of the (fixed) distribution function $F_{\hat{\xi}}$ tend to one if all its components tend to infinity. This implies

$$F_{\tilde{\xi}\left(z_1^{(k)}\right)} (t_k, \ldots, t_k) \to_k 1,$$

whence (10) leads to

$$\frac{\partial F_\xi}{\partial z_1}(z^{(k)}) \to_k f_{\xi_1}(q_p) > 0 \tag{12}$$

by continuity and positivity of the density f_{ξ_1}. Similarly, the second partial derivative of F_ξ calculates from Theorem 1 as

$$\frac{\partial F_\xi}{\partial z_2}(z^{(k)}) = f_{\xi_2}(t_k) \cdot F_{\tilde{\xi}\left(z_2^{(k)}\right)} \left(q_p + k^{-1}, t_k, \ldots, t_k \right), \tag{13}$$

where f_{ξ_2} denotes the one-dimensional Gaussian density of the component ξ_2 and $\tilde{\xi}\left(z_2^{(k)}\right)$ is a certain (s-1)-dimensional Gaussian random vector. From $f_{\xi_2}(t_k) \to_k 0$ (due to $t_k \to_k \infty$) and from the fact that distribution functions are bounded between zero and one, we infer that

$$\frac{\partial F_\xi}{\partial z_2}(z^{(k)}) \to_k 0,$$

which along with (12) and (6) provides that $\varkappa(z^{(k)}) \to_k \infty$. Therefore, with our arbitrarily fixed number $n \in \mathbb{N}$ we may associate an index $k_n \in \mathbb{N}$ such that $\varkappa(z^{(k_n)}) > \bar{\varkappa}$ where $\bar{\varkappa}$ is defined in (7). Now, we assign to n the cost vector $c^{(n)} := \nabla F_\xi \left(z^{(k_n)} \right)$ for the linear objective function in problem $(\mathrm{P}_{c^{(n)},\xi,\mathrm{p}})$ in (2). Then, by Corollary 1, we have that $c_i^{(n)} > 0$ for $i = 1, \ldots, s$ as required in the

statement of our theorem. Knowing that $\log F_\xi$ is a concave function (see [6]), the problem $(\mathrm{P}_{c^{(n)},\xi,p})$ in (2) may be written equivalently as a convex optimization problem

$$\min\{c^{(n)T}x \mid -\log F_\xi(x) \leq -\log p\}. \qquad (\mathrm{P}_{c^{(n)},\xi,p}) \tag{14}$$

With $c^{(n)} \neq 0$, a solution x^* to this problem is equivalently characterized by the conditions

$$-\log F_\xi(x^*) = -\log p \quad \text{and} \quad c^{(n)} + \lambda \nabla\left(-\log F_\xi\right)(x^*) = 0 \quad \text{for some} \quad \lambda > 0.$$

Simplifying these leads to the equivalent conditions

$$F_\xi(x^*) = p \quad \text{and} \quad c^{(n)} = \lambda \nabla\left(F_\xi\right)(x^*) \quad \text{for some} \quad \lambda > 0. \tag{15}$$

Now, since $c^{(n)} = \nabla F_\xi\left(z^{(k_n)}\right)$ and $F_\xi(z^{(k_n)}) = p$ by (9), we conclude that $z^{(k_n)}$ is a solution to $(\mathrm{P}_{c^{(n)},\xi,p})$. This shows that $\Psi_{c^{(n)},\xi,p} \neq \emptyset$ as asserted in our theorem. Finally, we show that $\Psi_{c^{(n)},\xi,p} \cap \mathbb{B}\,(0, L_n) = \emptyset$ with L_n defined in (5). Assume the contrary and choose some $x^* \in \Psi_{c^{(n)},\xi,p}$ with $\|x^*\| \leq L_n$. From (6) and (15) we derive

$$\begin{aligned}\varkappa\,(x^*) &= \frac{\partial F_\xi}{\partial z_1}(x^*) \Big/ \frac{\partial F_\xi}{\partial z_2}(x^*) = c_1^{(n)} \Big/ c_2^{(n)} = \frac{\partial F_\xi}{\partial z_1}(z^{(k_n)}) \Big/ \frac{\partial F_\xi}{\partial z_2}(z^{(k_n)}) \\ &= \varkappa(z^{(k_n)}) > \bar{\varkappa}.\end{aligned}$$

which is a contradiction with (7). Consequently, $\Psi_{c^{(n)},\xi,p} \cap \mathbb{B}\,(0, L_n) = \emptyset$. Now, select arbitrary $x \in \Psi_{c^{(n)},\xi,p}$ and $y \in \Psi_{c^{(n)},\eta,p}$. Then, $\|x\| > L_n$. Since also $\Psi_{c^{(n)},\eta,p} \subseteq [a, b]$ by (4), it follows from (5) that $\|x - y\| \geq n$. Since x and y were arbitrarily chosen, we end up at the final assertion (3) of our theorem.

Theorem 2 can be interpreted as follows in the context of empirical approximation upon observing that the support of empirical measures is finite, hence compact: no matter how large the sample size N for the empirical approximation of the original random vector ξ is chosen, there is always an instance of problem (2) (by choosing an appropriate cost vector c) such that the solutions between the original problem and its empirical approximation are arbitrarily far from each other. Note that relation (3) implies (and actually is much stronger than) the Hausdorff distance between both solution sets being larger than any prescribed n. Moreover, this effect of ill-conditioning is not caused by letting the probability level tend to one, because the result of the theorem holds true for any fixed p. In the following section we look at the same phenomenon from a slightly different viewpoint.

3 Exponential Estimates with Ill-Conditioned Constants

The recent literature on empirical or sample average approximation on chance constraints [4,5] compiles several convergence results for feasible sets, optimal

values and solutions, most of them of qualitative nature (continuity, upper semicontinuity), some of them providing exponential estimates (convergence of feasible sets, lower bounds for optimal values). For the general stability theory of chance constrained programming with arbitrary approximations (not just empirical ones) and quantitative convergence results even for solution sets, we refer to [1,2]. In this section we will appy these results to the special case of empirical approximations in order to obtain an exponential bound for the convergence (in the sense of Hausdorff distance!) of solution sets. Despite this positive result we will show then, that the existence of exponential estimates does not exclude the need for a possibly excessive sample size in the empirical approximation. We start by citing the following stability result for chance constraints whch we present here in a simplified form sufficient for our purposes

Theorem 3 ([2], Corollary 3). *In problem* $(\mathrm{P}_{c,\xi,p})$ *in (2) let* $p \in (0,1)$ *and the following assumptions be satisfied:*

1. $\log F_\xi$ *is a strongly concave function.*
2. $\Psi_{c,\xi,p}$ *is nonempty and compact.*

Then, there exist $L, \delta > 0$ *such that*

$$d_H\left(\Psi_{c,\xi,p}, \Psi_{c,\eta,p}\right) \leq L\sqrt{\sup_{z\in\mathbb{R}^s}|F_\xi(z) - F_\eta(z)|} \quad \forall \eta : \sup_{z\in\mathbb{R}^s}|F_\xi(z) - F_\eta(z)| < \delta. \tag{16}$$

Here, d_H *refers to the Hausdorff distance.*

A prototype example for a problem (2) which automatically satisfies all assumptions of Theorem 3 is given by a random vector ξ having a standard Gaussian distribution. As a preparation we show the following property which is of independent interest:

Proposition 1. *In problem* $(\mathrm{P}_{c,\xi,p})$ *in (2) let* $p \in (0,1)$, $c_i > 0$ *for* $i = 1,\ldots,s$ *and* $\xi \sim \mathcal{N}(0, I_s)$. *Then, the problem has a solution.*

Proof. Referring back to the proof of Theorem 2, a solution of problem $(\mathrm{P}_{c,\xi,p})$ is equivalently characterized by the conditions (15) applied to c rather than $c^{(n)}$. The distribution assumption implies that for all $z \in \mathbb{R}^s$ and all $i = 1,\ldots,s$,

$$F_\xi(z) = \Phi(z_1)\cdots\Phi(z_s); \quad \frac{\partial F_\xi}{\partial z_i}(z) = f(z_i)\,\Phi(z_1)\cdots\Phi(z_{i-1})\,\Phi(z_{i+1})\cdots\Phi(z_s),$$

where f and Φ denote the one-dimensional standard normal density and distribution function, respectively. From (15) we derive that x^* is a solution to $(\mathrm{P}_{c,\xi,p})$ if there exists some $\lambda > 0$ such that

$$\Phi(x_1^*)\cdots\Phi(x_s^*) = p; \quad \lambda\frac{f(x_i^*)}{\Phi(x_i^*)} = c_i \quad (i = 1,\ldots,s). \tag{17}$$

(recall that Φ is strictly positive). Defining $\alpha := f/\Phi$, we observe that due to $c_i > 0$ the second relation of (17) amounts to the fact that $\alpha(x_i^*)/c_i$ is constant for all i. We may assume, without loss of generality, that c_1 is the largest of the coefficients c_i. Then, x^* is a solution to $(\mathrm{P}_{\mathrm{c},\xi,\mathrm{p}})$ if there exist coefficients $\rho_1, \ldots, \rho_{s-1} \geq 1$ such that

$$\Phi(x_1^*) \cdots \Phi(x_s^*) = p; \quad \alpha(x_1^*) = \rho_1 \alpha(x_2^*) = \cdots = \rho_{s-1}\alpha(x_s^*). \tag{18}$$

We recall from the properties of the one-dimensional standard normal density and distribution function that $\alpha(t) \to 0$ for $t \to \infty$. Consequently, given any $t \in \mathbb{R}$, there exist values $\beta_i(t)$ for $i = 1, \ldots, s-1$ such that $\beta_i(t) \geq t$ and

$$\alpha(t) = \rho_1 \alpha(\beta_1(t)) = \cdots = \rho_{s-1}\alpha(\beta_{s-1}(t)). \tag{19}$$

Taking into account that $\lim_{t\to-\infty} \Phi(t) = 0$ and $\Phi(\beta_i(t)) \leq 1$ on the one hand and

$$\lim_{t\to\infty} \Phi(t) = \lim_{t\to\infty} \Phi(\beta_i(t)) = 1$$

due to $\beta_i(t) \geq t$, on the other hand, we conclude that

$$\lim_{t\to-\infty(+\infty)} \Phi(t)\,\Phi(\beta_1(t)) \cdots \Phi(\beta_{s-1}(t)) = 0(1).$$

For continuity reasons, there exists some t^* such that

$$\Phi(t^*)\,\Phi(\beta_1(t^*)) \cdots \Phi(\beta_{s-1}(t^*)) = p.$$

Setting $x_1^* := t^*$and $x_i^* := \beta_{i-1}(t^*)$ for $i = 2, \ldots, s$, one verifies via (19) that (18) is satisfied and, hence, x^* is a solution to $(\mathrm{P}_{\mathrm{c},\xi,\mathrm{p}})$.

Corollary 2. *Under the assumptions of Proposition 1 the estimate (16) holds true.*

Proof. We have to check that the assumptions of Theorem 3 are satisfied. The strong concavity of the log of Gaussian distribution functions with independent components is easy to verify (see [2, Prop. 14]). As shown in Proposition 1, the solution set $\Psi_{c,\xi,p}$ is nonempty. On the other hand, there may not exist more than one solution to problem $(\mathrm{P}_{\mathrm{c},\xi,\mathrm{p}})$ because in its equivalent description

$$\min\{c^T x \mid -\log F_\xi(x) \leq -\log p\} \qquad (\mathrm{P}_{\mathrm{c},\xi,\mathrm{p}})$$

the inequality constraint is strongly convex according to what we have mentioned in the begining of this proof.

We emphasize that in the result of Theorem 3 the approximating random vector η can be arbitrary. In the special case that η is an empirical approximation, one may exploit exponential bounds from empirical process theory (e.g., [8]) to further interpret the obtained stability result. In order to keep the presentation simple, we refer here to a classical inequality by Kiefer [3] stating in any dimension s the existence of constants k_1 and $k_2 < 2$ (where k_2 may be chosen

arbitrarily close to 2) such that for all $\tilde{\varepsilon} > 0$ and all η^N having an empirical distribution of an i.i.d. sample of ξ with size N the following estimate applies:

$$\mathbb{P}\left(\sup_{z\in\mathbb{R}^s} |F_\xi(z) - F_{\eta^N}(z)| \geq \tilde{\varepsilon}\right) \leq k_1 \exp\left(-k_2\tilde{\varepsilon}^2 N\right). \tag{20}$$

Since by (16) one has for all $\varepsilon > 0$ the implication

$$\sup_{z\in\mathbb{R}^s} |F_\xi(z) - F_\eta(z)| < \min\{\delta, (\varepsilon/L)^2\} \Longrightarrow d_H\left(\Psi_{c,\xi,p}, \Psi_{c,\eta,p}\right) < \varepsilon,$$

it follows from (20) with $\tilde{\varepsilon} := \min\{\delta, (\varepsilon/L)^2\}$ that

$$\mathbb{P}\left(d_H\left(\Psi_{c,\xi,p}, \Psi_{c,\eta^N,p}\right) \geq \varepsilon\right) \leq k_1 \exp\left(-k_2\left(\min\{\delta, (\varepsilon/L)^2\}\right)^2 N\right). \tag{21}$$

This last relation establishes an exponential bound for the convergence of Hausdorff distance between the solution sets of the original problem and the problem approximated by a sample of size N. Obviously, the quantity

$$k_2\left(\min\{\delta, (\varepsilon/L)^2\}\right)^2$$

determines the exponential decay of the required sample size. However, for practical use, one would have to know the values or at least estimates for δ and L which is difficult or impossible in general. Then, the availability of an exponential convergence result does not exclude excessive sample sizes even in order to give a sense to (21), i.e., to ensure that the right-hand side is smaller than one as an upper probability estimate. Revisiting Theorem 3, one observes that the couple (δ, L) in (16) is not uniquely determined. Therefore, let us define the best possible coefficient of exponential decay by

$$\vartheta(c, \xi, p, \varepsilon) := \sup\left\{k_2\left(\min\{\delta, (\varepsilon/L)^2\}\right)^2 \middle| (\delta, L) \text{ satisfy (16) for } (\mathrm{P}_{c,\xi,p})\right\}.$$

Then, (21) can be formally improved to

$$\mathbb{P}\left(d_H\left(\Psi_{c,\xi,p}, \Psi_{c,\eta^N,p}\right) \geq \varepsilon\right) \leq k_1 \exp\left(-\vartheta(c, \xi, p, \varepsilon) \cdot N\right).$$

The following result demonstrates that for a given significant problem class, this coefficient of exponential decay may be arbitrarily close to zero, thus driving the required sample size to infinity.

Theorem 4. *In (2), let $s > 1$, $\xi \sim \mathcal{N}(0, I_s)$ and $p \in (0, 1)$ be arbitrarily given. Then, for any $\varepsilon > 0$ one has that*

$$\inf\{\vartheta(c, \xi, p, \varepsilon) \,|\, c \succ 0\} = 0,$$

where '$c \succ 0$' means $c_i > 0$ for $i = 1, \ldots, s$.

Proof. Denote by τ the infimum above and assume that $\tau > 0$. Then, $\vartheta(c, \xi, p, \varepsilon) \geq \tau$ for all $c \succ 0$. By definition of $\vartheta(c, \xi, p, \varepsilon)$, we infer that

$$\forall c \succ 0 \; \exists (\delta, L) : (\delta, L) \text{ satisfy (16) for } (\mathrm{P}_{c,\xi,\mathrm{p}}) \text{ and } k_2 \left(\min\{\delta, (\varepsilon/L)^2\}\right)^2 \geq \tau/2.$$

The last relation entails that

$$\delta \geq \sqrt{\frac{\tau}{2k_2}} =: \bar{\delta}, \quad L \leq \varepsilon \sqrt[4]{\frac{2k_2}{\tau}} =: \bar{L}.$$

Note that $\bar{\delta}$ and $\bar{L}$ do not depend on c. Consequently, we have shown that there exist $\bar{\delta} > 0$ and $\bar{L}$ such that

$$\forall c \succ 0 \; \exists \delta \geq \bar{\delta}, L \leq \bar{L} : \text{ (16) holds true for } (\mathrm{P}_{c,\xi,\mathrm{p}}).$$

This statement can be evidently reduced to:

$$\forall c \succ 0 : d_H(\Psi_{c,\xi,p}, \Psi_{c,\eta,p}) \leq \bar{L} \sqrt{\sup_{z \in \mathbb{R}^s} |F_\xi(z) - F_\eta(z)|} \tag{22}$$

$$\forall \eta : \sup_{z \in \mathbb{R}^s} |F_\xi(z) - F_\eta(z)| < \bar{\delta}.$$

From (20) we infer that, for any N,

$$\mathbb{P}\left(\sup_{z \in \mathbb{R}^s} |F_\xi(z) - F_{\eta^N}(z)| < \bar{\delta}/2\right) \geq 1 - k_1 \exp\left(-k_2 \bar{\delta}^2 N/4\right),$$

where η^N has an empirical distribution of an i.i.d. sample of ξ with size N. For $N \to \infty$, the right-hand side tends to one such that the probability on the left-hand side is at least strictly positive. As a consequence, for some fixed N large enough, there exists a discrete random vector η^N with N atoms such that

$$\sup_{z \in \mathbb{R}^s} |F_\xi(z) - F_{\eta^N}(z)| < \bar{\delta}/2.$$

Then, (22) implies that

$$\forall c \succ 0 : \quad d_H\left(\Psi_{c,\xi,p}, \Psi_{c,\eta^N,p}\right) \leq \bar{L} \sqrt{\bar{\delta}/2}. \tag{23}$$

Now, since $\operatorname{supp} \eta^N$ is compact, Theorem 2 yields the existence of some $\tilde{c} \succ 0$ such that

$$\inf\{\|x - y\| \mid x \in \Psi_{\tilde{c},\xi,p}, y \in \Psi_{\tilde{c},\eta^N,p}\} > \bar{L} \sqrt{\bar{\delta}/2}.$$

This, however, is a contradiction with (23) because

$$d_H(A, B) \geq \inf\{\|x - y\| \mid x \in A, y \in B\}$$

for any closed sets A, B. Hence, $\tau = 0$, as was to be shown.

The effect of the previous theorem can already be illustrated in two dimensions:

Example 1. Consider the following 2-dimensional problem:

$$\min\{x_1 + 10^{-4}x_2 \mid \mathbb{P}(\xi_1 \leq x_1, \xi_2 \leq x_2) \geq 0.99\}, \quad \xi \sim \mathcal{N}\left((0,0), \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

By independence of components, we may rewrite the chance constraint as

$$\Phi(x_1)\Phi(x_2) \geq 0.99,$$

where Φ denotes the one-dimensional standard normal distribution function. Referring to the optimality conditions as in (15), the solution of this problem is equivalently characterized by the following three nonlinear equations in the three variables x_1, x_2, λ:

$$\Phi(x_1)\Phi(x_2) = 0.99, \quad \varphi(x)\Phi(y) = \lambda, \quad \varphi(y)\Phi(x) = 10^{-4}\lambda.$$

Here, φ is the density of the one-dimensional standard normal distribution. This system is easily solved numerically, providing a unique optimal solution $x_1^* = 2.33$, $x_2^* = 4.88$.

Now, suppose that we want to approximate this solution empirically up to a precision of $\varepsilon = 0.1$. Then, in particular, the second component of the solution to the problem with empirical approximation has to exceed the value 4.78. Now, the second components $\xi_2^1, \ldots, \xi_2^N$ of an i.i.d. sample of ξ are i.i.d. standard normal. Hence,

$$\mathbb{P}(\max_{i=1,\ldots,N} \xi_2^i \leq t) = \Phi^N(t).$$

For instance, for $N = 10^6$ and $t = 4.78$, one has $\Phi^N(t) \approx 0.42$. This means that the probability of obtaining a one-digit precise solution by empirical approximation with a huge sample size like one million is less than $1 - 0.42 = 0.58$.

Certainly, the effect of the example relies on the highly unbalanced cost vector $c = \left(1, 10^{-4}\right)$. Making it more reasonably balanced like $c = (1, 0.1)$, one would still need a sample size of $N \approx 6.300$ for estimating the solution of the problem with a precision of 0.1 at a reasonably high probability of 0.99. Taking into account that N corresponds to the number of binary variables required in the discrete optimization problem, this is already a considerable quantity given the trivial dimension $s = 2$ of the problem. Of course, things may be expected to become much worse in larger dimension.

References

1. Henrion, R., Römisch, W.: Metric regularity and quantitative stability in stochastic programs with probabilistic constraints. Math. Program. 84, 55–88 (1999)
2. Henrion, R., Römisch, W.: Hölder and Lipschitz Stability of Solution Sets in Programs with Probabilistic Constraints. Math. Program. 100, 589–611 (2004)

3. Kiefer, J.: On large deviations of the empiric d.f. of vector chance variables and a law of the iterated logarithm. Pacific J. Math. 11, 649–660 (1961)
4. Luedtke, J., Ahmed, S.: A sample approximation approach for optimization with probabilistic constraints. SIAM J. Optim. 19, 674–699 (2008)
5. Pagnoncelli, B., Ahmed, S., Shapiro, A.: Sample Average Approximation Method for Chance Constrained Programming: Theory and Applications. J. Optim. Theory Appl. 142, 399–416 (2009)
6. Prékopa, A.: Stochastic Programming. Kluwer, Dordrecht (1995)
7. Shapiro, A., Dentcheva, D., Ruszczyński, A.: Lectures on Stochastic Programming. MPS-SIAM series on optimization, vol. 9 (2009)
8. Talagrand, M.: Sharper bounds for Gaussian and empirical processes. Ann. Probab. 22, 28–76 (1994)

Convergence Rates for the Iteratively Regularized Landweber Iteration in Banach Space

Barbara Kaltenbacher*

University of Klagenfurt,
Universitätsstraße 65–67, 9020 Klagenfurt, Austria
barbara.kaltenbacher@aau.at

Abstract. In this paper we provide a convergence rates result for a modified version of Landweber iteration with a priori regularization parameter choice in a Banach space setting.

Keywords: regularization, nonlinear inverse problems, Banach space, Landweber iteration.

An increasing number of inverse problems is nowadays posed in a Banach space rather than a Hilbert space setting, cf., e.g., [2,6,13] and the references therein.

An Example of a model problem, where the use of non-Hilbert Banach spaces is useful, is the identification of the space-dependent coefficient function c in the elliptic boundary value problem

$$-\Delta u + cu = f \quad \text{in } \Omega \tag{1}$$

$$u = 0 \quad \text{on } \partial\Omega \tag{2}$$

from measurements of u in $\Omega \subseteq \mathbb{R}^d$, $d \in \{1,2,3\}$, where f is assumed to be known. Here e.g., the choices $p = 1$ for recovering sparse solutions, $q = \infty$ for modelling uniformly bounded noise, or $q = 1$ for dealing with impulsive noise are particulary promising, see, e.g., [3] and the numerical experiments in Section 7.3.3 of [13].

Motivated by this fact we consider nonlinear ill-posed operator equations

$$F(x) = y \tag{3}$$

where F maps between Banach spaces X and Y.

In the example above, the forward operator F maps the coefficient function c to the solution of the boundary value problem (1), (2), and is well-defined as an operator

$$F : \mathcal{D}(F) \subseteq L^p(\Omega) \to L^q(\Omega)\,,$$

* Support by the German Science Foundation DFG under grant KA 1778/5-1 and within the Cluster of Excellence in Simulation Technology (EXC 310/1) at the University of Stuttgart is gratefully acknowledged.

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 38–48, 2013.

where $\mathcal{D}(F) = \{c \in X \mid \exists\, \hat{c} \in L^\infty(\Omega), \hat{c} \geq 0 \text{ a.e. } : \ \|c - \hat{c}\|_X \leq r\}$, r sufficiently small, for any

$$\begin{array}{ll} p, q \in [1, \infty], \quad f \in L^1(\Omega) & \text{if } d \in \{1, 2\} \\ p \in [1, \infty],\ q \in (\frac{d}{2}, \infty],\ f \in L^s(\Omega),\ s > \frac{d}{2} & \text{if } d \geq 3\,, \end{array}$$

see Section 1.3 in [13].

Since the given data y^δ are typically contaminated by noise, regularization has to be applied. We are going to assume that the noise level δ in

$$\|y - y^\delta\| \leq \delta \tag{4}$$

is known and provide convergence results in the sense of regularization methods, i.e., as δ tends to zero. In the following, x_0 is some initial guess and we will assume that a solution $x^\dagger$ to (3) exists.

Variational methods in Banach space have been extensively studied in the literature, see, e.g., [2,10,6] and the references therein.

Since these generalizations of Tikhonov regularization require computation of a global minimizer, iterative methods are an attractive alternative especially for large scale problems. After convergence results on iterative methods for nonlinear ill-posed operator equations in Banach spaces had already been obtained in the 1990's (cf. the references in [1]) in the special case $X = Y$, the general case $X \neq Y$ has only been treaten quite recently, see e.g. [5], [7], and [9] for an analysis of gradient and Newton type iterations. While convergence rates have already been established for the iteratively regularized Gauss-Newton iteration in [7], the question of convergence rates remains challenging and will be tackled in this paper; we refer to [14] for a different approach.

In order to formulate and later on analyze the method, we have to introduce some basic notations and concepts.

Consider, for some $q \in (1, \infty)$, the duality mapping $J_q^X(x) := \partial \left\{ \frac{1}{q} \|x\|^q \right\}$, which maps from X to its dual X^*. To analyze convergence rates we employ the Bregman distance

$$D_{j_q}(\tilde{x}, x) = \frac{1}{q}\|\tilde{x}\|^q - \frac{1}{q}\|x\|^q - \langle j_q^X(x), \tilde{x} - x \rangle_{X^*,X}$$

(where $j_q^X(x)$ denotes a single valued selection of $J_q^X(x)$) or its shifted version

$$D_q^{x_0}(\tilde{x}, x) := D_{j_q}(\tilde{x} - x_0, x - x_0)\,.$$

Throughout this paper we will assume that X is smooth, which means that the duality mapping is single-valued, and moreover, that X is q-convex, i.e.,

$$D_{j_q}(x, y) \geq c_q \|x - y\|^q \tag{5}$$

for some constant $c_q > 0$. As a consequence, X is reflexive and we also have

$$D_{j_{q^*}}(x^*, y^*) \leq C_{q^*} \|x^* - y^*\|^{q^*}\,, \tag{6}$$

for some C_{q^*}. Here q^* denotes the dual index $q^* = \frac{q}{q-1}$. Moreover, the duality mapping is bijective and $J_q^{-1} = J_{q^*}^{X^*}$, the latter denoting the (by q-convexity also single-valued) duality mapping on X^*. We will also make use of the identities

$$D_{j_q}(x,y) = D_{j_q}(x,z) + D_{j_q}(z,y) + \langle J_q^X(z) - J_q^X(y), x - z\rangle_{X^*,X} \tag{7}$$

and

$$D_{j_q}(y,x) = D_{j_{q^*}}(J_q^X(x), J_q^X(y))\,. \tag{8}$$

For more details on the geometry of Banach spaces we refer, e.g., to [12] and the references therein.

We here consider the iteratively regularized Landweber iteration

$$\begin{aligned} J_q^X(x_{n+1}^\delta - x_0) &= (1-\alpha_n) J_q^X(x_n^\delta - x_0) - \mu_n A_n^* j_p^Y(F(x_n^\delta) - y^\delta)\,, \\ x_{n+1}^\delta &= x_0 + J_{q^*}^{X^*}(J_q^X(x_{n+1}^\delta - x_0))\,, \qquad n = 0,1,\ldots \end{aligned} \tag{9}$$

where we abbreviate

$$A_n = F'(x_n^\delta)\,,$$

which, for an appropriate choice of the sequence $\{\alpha_n\}_{n\in\mathbb{N}} \in (0,1]$, has been shown to be convergent with rates under a source condition

$$x^\dagger - x_0 \in \mathcal{R}(F'(x^\dagger)^* F'(x^\dagger))^{\nu/2}), \tag{10}$$

with $\nu = 1$ in a Hilbert space setting in [11]. Since the linearized forward operator $F'(x)$ typically has some smoothing property (reflecting the ill-posedness of the inverse problems) condition (10) can often be interpreted as a regularity assumption on the initial error $x^\dagger - x_0$, which is stronger for larger ν.

In the Hilbert space case the proof of convergence rates for the plain Landweber iteration (i.e., (9) with $\alpha_n = 0$) under source conditions (10) relies on the fact that the iteration errors $x_n^\delta - x^\dagger$ remain in the range of $(F'(x^\dagger)^* F'(x^\dagger))^{\nu/2}$ and their preimages under $(F'(x^\dagger)^* F'(x^\dagger))^{\nu/2}$ form a bounded sequence (cf., Proposition 2.11 in [8]). Since carrying over this approach to the Banach space setting would require more restrictive assumptions on the structure of the spaces even in the special case $\nu = 1$, we here consider the modified version with an appropriate choice of $\{\alpha_n\}_{n\in\mathbb{N}} \in (0,1]$.

In place of the Hilbert space source condition (10), we consider variational inequalities

$$\begin{aligned} &\exists\, \beta > 0\ \forall x \in \mathcal{B}_\rho^D(x^\dagger) : \\ &|\langle J_q^X(x^\dagger - x_0), x - x^\dagger\rangle_{X^*\times X}| \le \beta D_q^{x_0}(x^\dagger, x)^{\frac{1-\nu}{2}} \|F'(x^\dagger)(x - x^\dagger)\|^\nu\,, \end{aligned} \tag{11}$$

cf., e.g., [4], where

$$\mathcal{B}_\rho^D(x^\dagger) = \{x \in X \mid D_q^{x_0}(x^\dagger, x) \le \rho^2\}$$

with $\rho > 0$ such that $x_0 \in \mathcal{B}_\rho^D(x^\dagger)$. Using the interpolation and the Cauchy-Schwarz inequality, it is readily checked that in the Hilbert space case (10)

implies (11). For more details on such variational inequalities we refer to Section 3.2.3 in [13] and the references therein.

The assumptions on the forward operator besides a condition on the domain

$$\mathcal{B}_\rho^D(x^\dagger) \subseteq \mathcal{D}(F) \tag{12}$$

include a structural condition on its degree of nonlinearity (cf. [4]

$$\begin{aligned} &\left\|(F'(x^\dagger + v) - F'(x^\dagger))v\right\| \le K \left\|F'(x^\dagger)v\right\|^{c_1} D_q^{x_0}(x^\dagger, v + x^\dagger)^{c_2}\,, \\ &v \in X,\ x^\dagger + v \in \mathcal{B}_\rho^D(x^\dagger)\,, \end{aligned} \tag{13}$$

whose strength depends on the smoothness index in (11). Namely, we assume that

$$c_1 = 1 \text{ or } c_1 + c_2 p > 1 \text{ or } (c_1 + c_2 p \ge 1 \text{ and } K \text{ is sufficiently small}) \tag{14}$$

$$c_1 + c_2 \tfrac{2\nu}{\nu+1} \ge 1\,, \tag{15}$$

so that in case $\nu = 1$, a Lipschitz condition on F', corresponding to $(c_1, c_2) = (0, 1)$ is sufficient.

Here F' denotes the Gateaux derivative of F, hence a Taylor remainder estimate

$$\begin{aligned} &\left\|F(x_n^\delta) - F(x^\dagger) - F'(x^\dagger)(x_n^\delta - x^\dagger)\right\| \qquad (16)\\ &= \left\|g(1) - g(0) - F'(x^\dagger)(x_n^\delta - x^\dagger)\right\| \\ &= \left\|\int_0^1 g'(t)\,dt - F'(x^\dagger)(x_n^\delta - x^\dagger)\right\| \\ &= \left\|\int_0^1 F'(x^\dagger + t(x_n^\delta - x^\dagger))(x_n^\delta - x^\dagger)\,dt - F'(x^\dagger)(x_n^\delta - x^\dagger)\right\| \\ &\le K \left\|F'(x^\dagger)(x_n^\delta - x^\dagger)\right\|^{c_1} D_q^{x_0}(x^\dagger, x_n^\delta)^{c_2} \qquad (17) \end{aligned}$$

where $g : t \mapsto F(x^\dagger + t(x_n^\delta - x^\dagger))$, follows from (13).

We will assume that in each step the step size $\mu_n > 0$ in (9) is chosen such that

$$\mu_n \frac{1 - 3C(c_1)K}{3(1 - C(c_1)K)} \|F(x_n^\delta) - y^\delta\|^p - 2^{q^*+q-2} C_{q^*} \mu_n^{q^*} \|A_n^* j_p^Y(F(x_n^\delta) - y^\delta)\|^{q^*} \ge 0 \tag{18}$$

where $C(c_1) = c_1^{c_1}(1 - c_1)^{1-c_1}$, and c_1, K are as in (13), which is possible, e.g., by a choice $0 < \mu_n \le C_\mu \frac{\|F(x_n^\delta) - y^\delta\|^{\frac{q-p}{q-1}}}{\|A_n\|^{q^*}} =: \overline{\mu}_n$ with $C_\mu := \frac{2^{2-q^*-q}}{3} \frac{1-3C(c_1)K}{(1-C(c_1)K)C_{q^*}}$ If

$$p \ge q \tag{19}$$

and F, F' are bounded on $\mathcal{B}_\rho^D(x^\dagger)$, it is possible to bound $\overline{\mu}_n$ away from zero

$$\overline{\mu}_n \ge C_\mu \left(\sup_{x \in \mathcal{B}_\rho^D(x^\dagger)} (\|F(x) - y\| + \overline{\delta})^{p-q} \|F'(x)\|^q \right)^{-1/(q-1)} =: \underline{\mu} \tag{20}$$

for $\delta \in [0, \overline{\delta}]$, provided the iterates remain in $\mathcal{B}_\rho^D(x^\dagger)$ (which we will show by induction in the proof of Theorem 1). Hence, there exist $\underline{\mu}, \overline{\mu} > 0$ independent of n and δ such that we can choose

$$0 < \underline{\mu} \le \mu_n \le \overline{\mu}\,, \tag{21}$$

(e.g., by simply setting $\mu_n \equiv \underline{\mu}$).

Moreover, we will use an a priori choice of the stopping index n_* according to

$$n_*(\delta) = \min\{n \in \mathbb{N} \ : \ \alpha_n^{\frac{\nu+1}{p(\nu+1)-2\nu}} \le \tau\delta\}\,, \tag{22}$$

and of $\{\alpha_n\}_{n\in\mathbb{N}}$ such that

$$\left(\frac{\alpha_{n+1}}{\alpha_n}\right)^{\frac{2\nu}{p(\nu+1)-2\nu}} + \tfrac{1}{3}\alpha_n - 1 \ge c\alpha_n \tag{23}$$

for some $c \in (0, \frac{1}{3})$ independent of n, where $\nu \in [0,1]$ is the exponent in the variational inequality (11).

Remark 1. A possible choice of $\{\alpha_n\}_{n\in\mathbb{N}}$ satisfying (23) and smallness of $\alpha_{\max}$ is given by

$$\alpha_n = \frac{\alpha_0}{(n+1)^x}$$

with $x \in (0,1]$ such that $3x\theta < \alpha_0$ sufficiently small, since then with $c := \frac{1}{3} - \frac{x\theta}{\alpha_0} > 0$, using the abbreviation $\theta = \frac{2\nu}{p(\nu+1)-2\nu} \in [0, \frac{1}{p-1}]$ we get by the Mean Value Theorem

$$\begin{aligned}
&\left(\frac{\alpha_{n+1}}{\alpha_n}\right)^\theta + (\tfrac{1}{3} - c)\alpha_n - 1 \\
&= \frac{\alpha_n}{\alpha_0}\left\{\alpha_0(\tfrac{1}{3} - c) - \frac{(n+2)^{x\theta} - (n+1)^{x\theta}}{(n+2)^{x\theta}}(n+1)^x\right\} \\
&= \frac{\alpha_n}{\alpha_0}\left\{\alpha_0(\tfrac{1}{3} - c) - \frac{x\theta(n+1+t)^{x\theta-1}}{(n+2)^{x\theta}}(n+1)^x\right\} \\
&\ge \frac{\alpha_n}{\alpha_0}\left\{\alpha_0(\tfrac{1}{3} - c) - x\theta\frac{(n+1)^x}{n+1+t}\right\} \ \ge 0\,,
\end{aligned}$$

for some $t \in [0,1]$.

Theorem 1. *Assume that X is smooth and q-convex, that x_0 is sufficiently close to $x^\dagger$, i.e., $x_0 \in \mathcal{B}_\rho^D(x^\dagger)$, (which by (5) implies that $\|x^\dagger - x_0\|$ is also small), that a variational inequality (11) with $\nu \in (0,1]$ and β sufficiently small is satisfied, that F satisfies (13) with (14), (15), that F and F' are continuous and uniformly bounded in $\mathcal{B}_\rho^D(x^\dagger)$, that (12) holds and that*

$$q^* \ge \frac{2\nu}{p(\nu+1) - 2\nu} + 1\,. \tag{24}$$

Let $n_*(\delta)$ be chosen according to (22) with τ sufficiently large. Moreover assume that (19) holds and the sequence $\{\mu_n\}_{n\in\mathbb{N}}$ is chosen such that (21) holds for $0 < \underline{\mu} < \overline{\mu}$ according to (20), and let the sequence $\{\alpha_n\}_{n\in\mathbb{N}} \subseteq [0,1]$ be chosen such that (23) holds, and $\alpha_{\max} = \max_{n\in\mathbb{N}} \alpha_n$ is sufficiently small.

Then, the iterates x_{n+1}^δ remain in $\mathcal{B}_\rho^D(x^\dagger)$ for all $n \le n_*(\delta) - 1$ with n_* according to (22). Moreover, we obtain optimal convergence rates

$$D_q^{x_0}(x^\dagger, x_{n_*}) = O(\delta^{\frac{2\nu}{\nu+1}}), \quad \text{as } \delta \to 0 \tag{25}$$

as well as in the noise free case $\delta = 0$

$$D_q^{x_0}(x^\dagger, x_n) = O\left(\alpha_n^{\frac{2\nu}{p(\nu+1)-2\nu}}\right) \tag{26}$$

for all $n \in \mathbb{N}$.

Remark 2. Note that the rate exponent in (26) $\frac{2\nu}{p(\nu+1)-2\nu} = \frac{2\nu}{\nu+1}(p - \frac{2\nu}{\nu+1})^{-1}$, always lies in the interval $[0, \frac{1}{p-1}]$, since $\frac{2\nu}{\nu+1} \in [0,1]$.

Moreover, note that Theorem 1 provides a results on rates only, but no convergence result without variational inequality. This corresponds to the situation from [11] in a Hilbert space setting.

Proof. First of all, for $x_n^\delta \in \mathcal{B}_\rho^D(x^\dagger)$, (13) allows us to estimate as follows (see also (16)) in case $c_1 \in [0,1)$:

$$\begin{aligned}
&\left\|F(x_n^\delta) - F(x^\dagger) - A(x_n^\delta - x^\dagger)\right\| \\
&\le K \left\|A(x_n^\delta - x^\dagger)\right\|^{c_1} D_q^{x_0}(x^\dagger, x_n^\delta)^{c_2} \\
&\le C(c_1)K \left(\left\|A(x_n^\delta - x^\dagger)\right\| + D_q^{x_0}(x^\dagger, x_n^\delta)^{\frac{c_2}{1-c_1}}\right),
\end{aligned} \tag{27}$$

where we have used the abbreviation $A = F'(x^\dagger)$ and the elementary estimate

$$a^{1-\lambda}b^\lambda \le C(\lambda)(a+b) \text{ with } C(\lambda) = \lambda^\lambda(1-\lambda)^{1-\lambda} \text{ for } a,b \ge 0\,,\ \lambda \in (0,1)\,, \tag{28}$$

and therewith, by the second triangle inequality,

$$\left\|A(x_n^\delta - x^\dagger)\right\| \le \frac{1}{1 - C(c_1)K}\left(\left\|F(x_n^\delta) - F(x^\dagger)\right\| + C(c_1)K D_q^{x_0}(x^\dagger, x_n^\delta)^{\frac{c_2}{1-c_1}}\right) \tag{29}$$

as well as analogously

$$\begin{aligned}
&\left\|F(x_n^\delta) - F(x^\dagger) - A_n(x_n^\delta - x^\dagger)\right\| \\
&\le 2C(c_1)K \left(\left\|A(x_n^\delta - x^\dagger)\right\| + D_q^{x_0}(x^\dagger, x_n^\delta)^{\frac{c_2}{1-c_1}}\right) \\
&\le \frac{2C(c_1)K}{1 - C(c_1)K}\left(\left\|F(x_n^\delta) - F(x^\dagger)\right\| + D_q^{x_0}(x^\dagger, x_n^\delta)^{\frac{c_2}{1-c_1}}\right).
\end{aligned} \tag{30}$$

For any $n \le n_*$ according to (22), by (7) we have

$$\begin{aligned}
&D_q^{x_0}(x^\dagger, x_{n+1}^\delta) - D_q^{x_0}(x^\dagger, x_n^\delta) \\
&= D_q^{x_0}(x_n^\delta, x_{n+1}^\delta) + \langle J_q^X(x_n^\delta - x_0) - J_q^X(x_{n+1}^\delta - x_0), x^\dagger - x_n^\delta \rangle_{X^* \times X} \\
&= D_q^{x_0}(x_n^\delta, x_{n+1}^\delta) - \mu_n \langle j_p^Y(F(x_n^\delta) - y^\delta), A_n(x_n^\delta - x^\dagger) \rangle_{Y^* \times Y} \\
&\quad + \alpha_n \langle J_q^X(x^\dagger - x_0), x^\dagger - x_n^\delta \rangle_{X^* \times X} \\
&\quad - \alpha_n \langle J_q^X(x^\dagger - x_0) - J_q^X(x_n^\delta - x_0), x^\dagger - x_n^\delta \rangle_{X^* \times X}
\end{aligned} \tag{31}$$

where the terms on the right hand side can be estimated as follows.

By (6) and (8) we have

$$\begin{aligned}
&D_q^{x_0}(x_n^\delta, x_{n+1}^\delta) \qquad (32) \\
&\le C_{q^*} \| J_q^X(x_{n+1}^\delta - x_0) - J_q^X(x_n^\delta - x_0) \|^{q^*} \\
&= C_{q^*} \| \alpha_n J_q^X(x_n^\delta - x_0) + \mu_n A_n^* j_p^Y(F(x_n^\delta) - y^\delta) \|^{q^*} \\
&\le 2^{q^*-1} C_{q^*} \left(\alpha_n^{q^*} \| x_n^\delta - x_0 \|^q + \mu_n^{q^*} \| A_n^* j_p^Y(F(x_n^\delta) - y^\delta) \|^{q^*} \right) \\
&\le 2^{q^*-1} C_{q^*} \left(\alpha_n^{q^*} (2^{q-1}(\|x^\dagger - x_0\|^q + \frac{1}{c_q} D_q^{x_0}(x^\dagger, x_n^\delta)) + \mu_n^{q^*} \| A_n^* j_p^Y(F(x_n^\delta) - y^\delta) \|^{q^*} \right) \qquad (33)
\end{aligned}$$

where we have used the triangle inequality in X^* and X, the inequality

$$(a+b)^\lambda \le 2^{\lambda-1}(a^\lambda + b^\lambda) \quad \text{for } a, b \ge 0\,,\ \lambda \ge 1\,, \tag{34}$$

and (5).

For the second term on the right hand side of (31) we get, using (30), (28), (34),

$$\begin{aligned}
&\langle j_p^Y(F(x_n^\delta) - y^\delta), A_n(x_n^\delta - x^\dagger) \rangle_{Y^* \times Y} \\
&= \langle j_p^Y(F(x_n^\delta) - y^\delta), F(x_n^\delta) - y^\delta \rangle_{Y^* \times Y} \\
&\quad - \langle j_p^Y(F(x_n^\delta) - y^\delta), F(x_n^\delta) - y^\delta - A_n(x_n^\delta - x^\dagger) \rangle_{Y^* \times Y} \\
&\ge \frac{1 - 3C(c_1)K}{1 - C(c_1)K} \|F(x_n^\delta) - y^\delta\|^p \\
&\quad - \|F(x_n^\delta) - y^\delta\|^{p-1} \left(\frac{2C(c_1)K}{1 - C(c_1)K} D_q^{x_0}(x^\dagger, x_n^\delta)^{\frac{c_2}{1-c_1}} + \frac{1 + C(c_1)K}{1 - C(c_1)K} \delta \right) \\
&= \frac{1 - 3C(c_1)K}{1 - C(c_1)K} \|F(x_n^\delta) - y^\delta\|^p \\
&\quad - \left(\frac{1 - 3C(c_1)K}{3C(\frac{p-1}{p})(1 - C(c_1)K)} \|F(x_n^\delta) - y^\delta\|^p \right)^{\frac{p-1}{p}} \left(\frac{(3C(\frac{p-1}{p}))^{p-1}}{(1 - C(c_1)K)} \right)^{\frac{1}{p}} \\
&\qquad \left(2C(c_1)K D_q^{x_0}(x^\dagger, x_n^\delta)^{\frac{c_2}{1-c_1}} + (1 + C(c_1)K)\delta \right)
\end{aligned}$$

$$\geq \frac{1-3C(c_1)K}{1-C(c_1)K}\|F(x_n^\delta)-y^\delta\|^p - C(\tfrac{p-1}{p})\left\{\frac{1-3C(c_1)K}{3C(\frac{p-1}{p})(1-C(c_1)K)}\|F(x_n^\delta)-y^\delta\|^p \right.$$
$$\left. +\frac{(3C(\frac{p-1}{p})^{p-1}}{(1-C(c_1)K)}2^{p-1}\left((2C(c_1)K)^p D_q^{x_0}(x^\dagger,x_n^\delta)^{\frac{c_2 p}{1-c_1}}+(1+C(c_1)K)^p\delta^p\right)\right\}. \tag{35}$$

Using the variational inequality (11), (29), and

$$(a+b)^\lambda \leq (a^\lambda + b^\lambda) \quad \text{for } a,b\geq 0\,,\ \lambda\in[0,1]\,, \tag{36}$$

we get

$$|\alpha_n\langle J_q^X(x^\dagger-x_0),x^\dagger-x_n^\delta\rangle_{X^*\times X}|$$
$$\leq \beta\alpha_n D_q^{x_0}(x^\dagger,x_n^\delta)^{\frac{1-\nu}{2}}\|F'(x^\dagger)(x_n^\delta-x^\dagger)\|^\nu$$
$$\leq \beta\alpha_n D_q^{x_0}(x^\dagger,x_n^\delta)^{\frac{1-\nu}{2}}\frac{1}{(1-C(c_1)K)^\nu}\left(\left\|F(x_n^\delta)-y^\delta\right\|+\delta+C(c_1)KD_q^{x_0}(x^\dagger,x_n^\delta)^{\frac{c_2}{1-c_1}}\right)^\nu$$

$$\leq \beta\alpha_n D_q^{x_0}(x^\dagger,x_n^\delta)^{\frac{1-\nu}{2}}\epsilon^{-\nu}\left(\epsilon\frac{1}{(1-C(c_1)K)^\nu}(\|F(x_n^\delta)-y^\delta\|+\delta)\right)^\nu$$
$$+\beta\alpha_n\left(\frac{C(c_1)K}{(1-C(c_1)K)}\right)^\nu D_q^{x_0}(x^\dagger,x_n^\delta)^{\frac{1-\nu}{2}+\frac{\nu c_2}{1-c_1}}$$
$$\leq C(\tfrac{\nu}{p})\left\{\left(\beta\alpha_n D_q^{x_0}(x^\dagger,x_n^\delta)^{\frac{1-\nu}{2}}\epsilon^{-\nu}\right)^{\frac{p}{p-\nu}}+\left(\epsilon\frac{1}{(1-C(c_1)K)^\nu}(\|F(x_n^\delta)-y^\delta\|+\delta)\right)^p\right\}$$
$$+\beta\alpha_n\left(\frac{C(c_1)K}{(1-C(c_1)K)}\right)^\nu D_q^{x_0}(x^\dagger,x_n^\delta)^{\frac{1-\nu}{2}+\frac{\nu c_2}{1-c_1}}$$

$$= C(\tfrac{\nu}{p})\left\{(\beta\epsilon^{-\nu})^{\frac{p}{p-\nu}}(3C(\tfrac{\nu}{p})C(\tfrac{p(1-\nu)}{2(p-\nu)}))^{\frac{p(1-\nu)}{2(p-\nu)}}\alpha_n^{\frac{p(1+\nu)}{2(p-\nu)}}\left(\frac{\alpha_n D_q^{x_0}(x^\dagger,x_n^\delta)}{3C(\frac{\nu}{p})C(\frac{p(1-\nu)}{2(p-\nu)})}\right)^{\frac{p(1-\nu)}{2(p-\nu)}}\right.$$
$$\left.+\left(\epsilon\frac{1}{(1-C(c_1)K)^\nu}(\|F(x_n^\delta)-y^\delta\|+\delta)\right)^p\right\}$$
$$+\beta\alpha_n\left(\frac{C(c_1)K}{(1-C(c_1)K)}\right)^\nu D_q^{x_0}(x^\dagger,x_n^\delta)^{\frac{1-\nu-c_1+\nu c_1+2\nu c_2}{2(1-c_1)}}$$

$$\leq C(\tfrac{\nu}{p})\left\{C(\tfrac{p(1-\nu)}{2(p-\nu)})\left[\left(\beta\epsilon^{-\nu}(3C(\tfrac{\nu}{p})C(\tfrac{p(1-\nu)}{2(p-\nu)}))^{\frac{1-\nu}{2}}\right)^{\frac{2p}{p(\nu+1)-2\nu}}\alpha_n^{\frac{p(1+\nu)}{p(\nu+1)-2\nu}}\right.\right.$$
$$\left.\left.+\left(\frac{\alpha_n D_q^{x_0}(x^\dagger,x_n^\delta)}{3C(\frac{\nu}{p})C(\frac{p(1-\nu)}{2(p-\nu)})}\right)\right]\right.$$
$$\left.+\left(\epsilon\frac{1}{(1-C(c_1)K)^\nu}(\|F(x_n^\delta)-y^\delta\|+\delta)\right)^p\right\}$$
$$+\frac{1}{3}\alpha_n D_q^{x_0}(x^\dagger,x_n^\delta) \tag{37}$$

where we have used (28) two times and $\epsilon > 0$ will be chosen as a sufficiently small number below. Moreover, by (15), the exponent $\frac{1-\nu-c_1+\nu c_1+2\nu c_2}{2(1-c_1)} = 1 + \frac{1+\nu}{2(1-c_1)}(c_1 + \frac{2\nu}{\nu+1}c_2 - 1)$ is larger or equal to one and β is sufficiently small so that $\beta \left(\frac{C(c_1)K}{(1-C(c_1)K)}\right)^{\nu} \rho^{\frac{1-\nu-c_1+\nu c_1+2\nu c_2}{2(1-c_1)}-1} < \frac{1}{3}$.

Finally, we have that

$$\begin{aligned}\langle J_q^X(x^\dagger - x_0) - J_q^X(x_n^\delta - x_0), x^\dagger - x_n^\delta\rangle_{X^*\times X} &= D_q^{x_0}(x^\dagger, x_n^\delta) + D_q^{x_0}(x_n^\delta, x^\dagger) \\ &\geq D_q^{x_0}(x^\dagger, x_n^\delta) \end{aligned} \tag{38}$$

Inserting estimates (32)-(38) with $\epsilon = 2^{p-1}\mu_n^{1/p}\left(\frac{1-3C(c_1)K}{3(1-C(c_1)K)}\right)^{1/p}\frac{(1-C(c_1)K)^\nu}{C(\frac{\nu}{p})}$ into (31) and using boundedness away from zero of μ_n and the abbreviations

$$\begin{aligned}
\mathrm{d}_n &= D_q^{x_0}(x^\dagger, x_n^\delta)^{1/2} \\
C_0 &= 6^{p-1}C(\tfrac{p-1}{p})^p \frac{(2C(c_1)K)^p}{(1-C(c_1)K)} \\
C_1 &= 2^{q^*+q-2}\frac{C_{q^*}}{c_q} \\
C_2 &= C(\tfrac{\nu}{p})C(\tfrac{p(1-\nu)}{2(p-\nu)})\left(\beta\epsilon^{-\nu}(3C(\tfrac{\nu}{p})C(\tfrac{p(1-\nu)}{2(p-\nu)})^{\frac{1-\nu}{2}}\right)^{\frac{2p}{p(\nu+1)-2\nu}} \\
C_3 &= 2^{q^*+q-2}C_{q^*}\|x^\dagger - x_0\|^q \\
C_4 &= 2^{p-1}C(\frac{\nu}{p})\overline{\epsilon}\frac{1}{(1-C(c_1)K)^\nu} + 6^{p-1}C(\tfrac{p-1}{p})^p\frac{(1+C(c_1)K)^p}{1-C(c_1)K}
\end{aligned}$$

$$\begin{aligned}
\underline{\epsilon} &= 2^{p-1}\underline{\mu}^{1/p}\left(\frac{1-3C(c_1)K}{3(1-C(c_1)K)}\right)^{1/p}\frac{(1-C(c_1)K)^\nu}{C(\frac{\nu}{p})} \\
\overline{\epsilon} &= 2^{p-1}\overline{\mu}^{1/p}\left(\frac{1-3C(c_1)K}{3(1-C(c_1)K)}\right)^{1/p}\frac{(1-C(c_1)K)^\nu}{C(\frac{\nu}{p})}
\end{aligned}$$

we obtain

$$\begin{aligned}\mathrm{d}_{n+1}^2 &\leq C_0\mathrm{d}_n^{\frac{2c_2p}{1-c_1}} + (1-\tfrac{1}{3}\alpha_n + C_1\alpha_n^{q^*})\mathrm{d}_n^2 + C_2\alpha_n^{\frac{p(1+\nu)}{p(\nu+1)-2\nu}} + C_3\alpha_n^{q^*} + C_4\delta^p \\ &- \left(\mu_n\frac{1-3C(c_1)K}{3(1-C(c_1)K)}\|F(x_n^\delta) - y^\delta\|^p - 2^{q^*+q-2}C_{q^*}\mu_n^{q^*}\|A_n^* j_p^Y(F(x_n^\delta) - y^\delta)\|^{q^*}\right).\end{aligned}$$

Here the last term is nonpositive due to the choice (18) of μ_n, so that we arrive at

$$\mathrm{d}_{n+1}^2 \leq C_0\mathrm{d}_n^{\frac{2c_2p}{1-c_1}} + (1-\tfrac{1}{3}\alpha_n + C_1\alpha_n^{q^*})\mathrm{d}_n^2 + \underbrace{(C_2 + C_3 + C_4\tau^{-p})}_{=:C_5}\alpha_n^{\frac{p(1+\nu)}{p(\nu+1)-2\nu}} \tag{39}$$

where we have used (24) and the stopping rule (22). Denoting

$$\gamma_n := \frac{\mathrm{d}_n^2}{\alpha_n^{\frac{2\nu}{p(\nu+1)-2\nu}}}$$

we get the following recursion

$$\gamma_{n+1} \le C_0 \left(\frac{\alpha_n}{\alpha_{n+1}}\right)^{\theta} \alpha_n^{\theta\omega}\gamma_n^{\omega} + \left(\frac{\alpha_n}{\alpha_{n+1}}\right)^{\theta} (1 - \tfrac{1}{3}\alpha_n + C_1\alpha_n^{q^*})\gamma_n + C_5 \left(\frac{\alpha_n}{\alpha_{n+1}}\right)^{\theta} \alpha_n \tag{40}$$

with

$$\theta = \frac{2\nu}{p(\nu+1) - 2\nu} \qquad \omega = \frac{c_2 p}{1 - c_1}\,,$$

where

$$\omega \ge 1$$

by (14) and

$$\theta\omega = \frac{p}{p - \frac{2\nu}{\nu+1}} \frac{c_2 \frac{2\nu}{\nu+1}}{1 - c_1} \ge 1$$

due to assumption (15). Hence as sufficient conditions for uniform boundedness of $\{\gamma_n\}_{n \le n_*}$ by $\overline{\gamma}$ and for $x_{n+1}^{\delta} \in \mathcal{B}_{\rho}^{D}(x^{\dagger})$ we get

$$\overline{\gamma} \le \rho^2 \tag{41}$$

$$C_0\alpha_n^{\theta\omega-1}\overline{\gamma}^{\omega} - \left\{ \left(\frac{\alpha_{n+1}}{\alpha_n}\right)^{\theta} + \tfrac{1}{3}\alpha_n - 1 - C_1\alpha_n^{q^*} \right\} \alpha_n^{-1}\overline{\gamma} + C_5 \le 0\,, \tag{42}$$

where by $q^* > 1$, (15) the factors $C_0\alpha_n^{\theta\omega-1}$, $C_1\alpha_n^{q^*-1}$ and C_5 can be made small for small $\alpha_{\max}$, β, $\|x^{\dagger} - x_0\|$ and large τ. We use this fact to achieve

$$C_0\alpha_n^{\theta\omega-1}\rho^{\omega-1} + C_1\alpha_n^{q^*-1} \le \tilde{c} < c$$

with $\tilde{c}$ independent of n, which together with (23) yields sufficiency of

$$\frac{C_5}{c - \tilde{c}} \le \overline{\gamma} \le \rho^2$$

for (41), (42), which for any (even small) prescribed ρ is indeed enabled by possibly decreasing β, $\|x^{\dagger} - x_0\|$, τ^{-1}, and therewith C_5.

In case $c_1 = 1$, estimates (29), (30) simplify to

$$\left\|A(x_n^{\delta} - x^{\dagger})\right\| \le \frac{1}{1 - \rho^{2c_2}K} \left\|F(x_n^{\delta}) - F(x^{\dagger})\right\| \tag{43}$$

and

$$\left\|F(x_n^{\delta}) - F(x^{\dagger}) - A_n(x_n^{\delta} - x^{\dagger})\right\| \le \frac{2\rho^{2c_2}K}{1 - \rho^{2c_2}K} \left\|F(x_n^{\delta}) - F(x^{\dagger})\right\|\,. \tag{44}$$

Therewith, the terms containing $D_q^{x_0}(x^{\dagger}, x_n^{\delta})^{\frac{c_2}{1-c_1}}$ are removed and $C(c_1)$ is replaced by ρ^{2c_2} in (32)-(38), so that we end up with a recursion of the form (40) (with C_0 replace by zero) as before. Hence the remainder of the proof of uniform boundedness of γ_n can be done in the same way as in case $c_1 < 1$.

In case $\delta = 0$, i.e., $n_* = \infty$, uniform boundedness of $\{\gamma_n\}_{n\in\mathbb{N}}$ implies (26). For $\delta > 0$ we get (25) by using (22) in

$$D_q^{x_0}(x^\dagger, x_{n_*}) = \gamma_{n_*}\alpha_{n_*}^{\frac{2\nu}{p(\nu+1)-2\nu}} \le \overline{\gamma}\alpha_{n_*}^{\frac{2\nu}{p(\nu+1)-2\nu}} \le \overline{\gamma}(\tau\delta)^{\frac{2\nu}{\nu+1}}$$

Remark 3. In view of estimate (39), an optimal choice of α_n would be one that minimizes the right hand side. At least in the special case that the same power of α_n appears in the last two terms, i.e., $\frac{p(1+\nu)}{p(\nu+1)-2\nu} = q^*$, elementary calculus yields

$$(\alpha_n^{opt})^{\frac{2\nu}{p(\nu+1)-2\nu}} = \frac{D_q^{x_0}(x^\dagger, x_n^\delta)}{3q^*(C_1 D_q^{x_0}(x^\dagger, x_n^\delta) + C_5)},$$

which shows that the obtained relation $D_q^{x_0}(x^\dagger, x_n^\delta) \sim \alpha_n^{\frac{2\nu}{p(\nu+1)-2\nu}}$ is indeed reasonable and probably even optimal.

References

1. Bakushinsky, A.B., Kokurin, M.Y.: Iterative methods for approximate solution of inverse problems. Springer, Dordrecht (2004)
2. Burger, M., Osher, S.: Convergence rates of convex variational regularization. Inverse Problems 20(5), 1411–1421 (2004)
3. Clason, C., Jin, B.: A semi-smooth Newton method for nonlinear parameter identification problems with impulsive noise. SIAM J. Imaging Sci. 5, 505–538 (2012)
4. Hein, T., Hofmann, B.: Approximate source conditions for nonlinear ill-posed problems – chances and limitations. Inverse Problems 25, 035003 (16pp) (2009)
5. Hein, T., Kazimierski, K.: Accelerated Landweber iteration in Banach spaces. Inverse Problems 26, 055002 (17pp) (2010)
6. Hofmann, B., Kaltenbacher, B., Pöschl, C., Scherzer, O.: A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators. Inverse Problems 23(3), 987–1010 (2007)
7. Kaltenbacher, B., Hofmann, B.: Convergence rates for the iteratively regularized Gauss-Newton method in Banach spaces. Inverse Problems 26, 035007 (21pp) (2010)
8. Kaltenbacher, B., Neubauer, A., Scherzer, O.: Iterative Regularization Methods for Nonlinear Ill-posed Problems. de Gruyter (2007)
9. Kaltenbacher, B., Schöpfer, F., Schuster, T.: Convergence of some iterative methods for the regularization of nonlinear ill-posed problems in Banach spaces. Inverse Problems 25, 065003 (2009), doi: 10.1088/0266-5611/25/6/065003
10. Neubauer, A., Hein, T., Hofmann, B., Kindermann, S., Tautenhahn, U.: Improved and extended results for enhanced convergence rates of Tikhonov regularization in Banach spaces. Appl. Anal. 89(11), 1729–1743 (2010)
11. Scherzer, O.: A modified Landweber iteration for solving parameter estimation problems. Appl. Math. Optim. 38, 45–68 (1998)
12. Schöpfer, F., Louis, A.K., Schuster, T.: Nonlinear iterative methods for linear ill-posed problems in Banach spaces. Inverse Problems 22(1), 311–329 (2006)
13. Schuster, T., Kaltenbacher, B., Hofmann, B., Kazimierski, K.: Regularization Methods in Banach Spaces. de Gruyter (2007)
14. Hein, T., Kazimierski, K.S.: Modified Landweber iteration in Banach spaces - convergence and convergence rates. Numerical Functional Analysis and Optimization 31(10), 1158–1189 (2010)

Weak Compactness in the Space of Operator Valued Measures and Optimal Control

Nasiruddin Ahmed

EECS, University of Ottawa, Ottawa, Canada

Abstract. In this paper we present a brief review of some important results on weak compactness in the space of vector valued measures. We also review some recent results of the author on weak compactness of any set of operator valued measures. These results are then applied to optimal structural feedback control for deterministic systems on infinite dimensional spaces.

Keywords: Space of Operator valued measures, Countably additive operator valued measures, Weak compactness, Semigroups of bounded linear operators, Optimal Structural control.

AMS(MOS) Subject Classification: 46A50,46B50,46E27,46E99,47A55.

1 Introduction

Necessary and sufficient conditions for weak compactness in the space of vector measures has been a subject of great interest over half a century. One of the seminal results in this topic is the well known Bartle-Dunford-Schwartz theorem [1, Theorem 5, p.105] for countably additive bounded vector measures with values in Banach spaces satisfying, along with their duals, the Radon-Nikodym property. This result was extended to finitely additive vector measures by Brooks [3] and Brooks and Dinculeanu [1, Corollary 6, p.106]. We present in this section some of the celebrated results on this topic. For vector measures see [1] and [2]. First we present the Bartle-Dunford-Schwartz theorem (BDS) [1, Theorem 5, p.105].

Theorem 1.1(BDS). Let D be any set and $\Sigma \equiv \sigma(D)$ denote the sigma algebra of subsets of the set D, and X, X^* a dual pair of B-spaces satisfying (Radon-Nikodym Property) RNP. A set $\Gamma \subset M_{ca}(\Sigma, X)$ is relatively weakly compact if, and only if,

(i) Γ is bounded

(ii) $\{|\mu|, \mu \in \Gamma\}$ is uniformly c.a

(iii) for each $\sigma \in \Sigma$ the set $\{\mu(\sigma), \mu \in \Gamma\} \subset X$ is weakly relatively compact.

This result was extended to finitely additive (f.a) vector measures by Brooks [3] for reflexive Banach spaces X, and then by Brooks and Dinculeanu [1, Corollary 6, 106] for nonreflexive spaces.

We state here the later result.

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 49–58, 2013.

Theorem 1.2 (Brooks & Dinculeanu). Let Σ be an algebra of subsets of a set D, $M_{ba}(\Sigma, X)$ the space of finitely additive vector measures with values in X and $\{X, X^*\}$ satisfy RNP. A set $\Gamma \subset M_{ba}(\Sigma, X)$ is weakly relatively compact if, and only if, the following three conditions are satisfied

(i) Γ is bounded

(ii) there exists a **f.a** nonnegative measure ν such that $\lim_{\nu(\sigma)\to 0} |\mu|(\sigma) = 0$ uniformly w.r.t $\mu \in \Gamma$.

(iii) for each $\sigma \in \Sigma$, the set $\{\mu(\sigma), \mu \in \Gamma\} \subset X$ is relatively (conditionally) weakly compact.

These results have been used by the author in the study of optimal control of impulsive systems on Banach spaces [13].

Weak sequential compactness for regular vector measures have been studied by Kuo [4, Theorem 1.6, Theorem 3.3] where he gives several results on weak sequential compactness based on set-wise weak convergence.

In physical sciences and engineering, involving control theory and optimization, one has the freedom to choose from a given class of vector or operator valued measures the best one that minimizes or maximizes certain functionals representing a measure of performance of the system. This is where compactness is useful. These problems arise naturally in the area of optimization, optimal control, system identification, Kalman filtering, structural control etc [8,9,10,13,14].

2 Basic Properties of Operator Valued Measures

Let D be a compact Hausdorff space and Σ an algebra of subsets of D, $\{X, Y\}$ a pair of B-spaces and $\mathcal{L}(X, Y)$ is the space of bounded linear operators from X to Y. The function

$$B : \Sigma \longrightarrow \mathcal{L}(X, Y)$$

is generally a finitely additive (f.a) set function with values in $\mathcal{L}(X, Y)$. This class, denoted by $M_{ba}(\Sigma, \mathcal{L}(X, Y))$, is called the space of operator valued measures. Clearly, this is a B-space with respect to the topology induced by the supremum of the operator norm on Σ.

Now we introduce the notion of countable additivity of operator valued measures. Unlike vector measures (Banach space valued), the notion of countable additivity of operator valued measures depends on the topology used for the space of bounded linear operators. Thus if we limit ourselves to the most popular topologies such as uniform, strong, weak operator topologies we have at least three kinds of countable additivity. This is described below.

Definition 2.1 $(ca - \tau_{uo})$ An element $B \in M_{ba}(\Sigma, \mathcal{L}(X, Y))$ is countably additive in the uniform operator topology $(ca - \tau_{uo})$ if for any family of pairwise disjoint sets $\{\sigma_i\} \in \Sigma, \sigma_i \subset D$ and $\cup\sigma_i \in \Sigma$,

$$\lim_{n\to\infty} \| B(\bigcup \sigma_i) - \sum_{i=1}^{n} B(\sigma_i) \|_{\mathcal{L}(X,Y)} = 0.$$

Similarly we define countable additivity in the strong operator topology as follows.

Definition 2.2 $(ca-\tau_{so})$ An element $B \in M_{ba}(\Sigma, \mathcal{L}(X,Y))$ is said to be countably additive in the strong operator topology $(ca-\tau_{so})$ if for any family of pair wise disjoint sets $\{\sigma_i\} \in \Sigma, \sigma_i \subset D$, $\cup\sigma_i \in \Sigma$, and for every $x \in X$,

$$\lim_{n\to\infty} |B(\bigcup \sigma_i)x - \sum_{i=1}^{n} B(\sigma_i)x|_Y = 0.$$

Note that if $X = R$ then B reduces to an Y valued vector measure and the countable additivity in the strong operator topology reduces to the usual (norm) countable additivity. Similarly, if $Y = R$ then B reduces to an X^*-valued vector measure and the countable additivity in the strong operator topology reduces to countable additivity in the weak star topology. One can also define countable additivity in the weak operator topology. Since we do not use it, it is not necessary to include it here.

By Orlicz-Pettis Theorem [1,2], a vector measure is countably additive if and only if it is weakly countably additive. Thus it follows from this result, that $ca - \tau_{so} \cong ca - \tau_{wo}$. That is, countable additivity in the strong operator topology is equivalent to countable additivity in the weak operator topology. Next we consider the question of variation. It is known that for Banach space valued vector measures there are two notions of variation, strong variation usually known as variation, and weak variation called semivariation. Again in the case of operator valued measures there are as many possibilities of variations as there are topologies.

For any set $J \in \Sigma$, let us denote the class of finite disjoint Σ measurable partitions of J by $\Pi_\Sigma(J)$.

Definition 2.3-τ_{uo} :(Variation in the τ_{uo}) For any $B \in M_{ba}(\Sigma, \mathcal{L}(X,Y))$ its variation in the uniform operator topology on the set J is given by

$$|B|_u(J) = \sup_{\pi\in\Pi_\Sigma(J)} \sum_{\sigma\in\pi} \parallel B(\sigma) \parallel_{\mathcal{L}(X,Y)},$$

where the supremum is taken over $\Pi_\Sigma(J)$.

Clearly, if $X = R$ then the operator valued measure B reduces to an Y -valued vector measure and the above expression gives the standard variation of vector measures.

Definition 2.4-τ_{so}(Variation in the τ_{so}). The variation of B on J in the strong operator topology is given by:

$$|B|_s(J) = \sup\left\{ |\sum_{i=1}^{n} B(\sigma_i)x_i|_Y, x_i \in B_1(X), \{\sigma_i, 1 \le i \le n,\} \in \Pi_\Sigma(J), n \in N \right\}.$$

Again if $X = R$ then B reduces to an Y valued vector measure and the above expression gives the standard semivariation of vector measures. Similarly one can define variation in the weak operator topology.

The uniform, strong, and weak variations of B over D are given respectively by

$$|B|_u(D) \equiv \sup\{|B|_u(\sigma), \sigma \in \Sigma\}, \quad |B|_s(D) \equiv \sup\{|B|_s(\sigma), \sigma \in \Sigma\}, \quad \text{and}$$
$$|B|_w(D) \equiv \sup\{|B|_w(\sigma), \sigma \in \Sigma\}.$$

It is easy to verify that $|B|_w \leq |B|_s \leq |B|_u$. Clearly, this result means that an element $B \in M_{ba}(\Sigma, \mathcal{L}(X,Y))$ may have finite strong variation while it has infinite uniform variation. That is , $|B|_s < \infty$ but $|B|_u = \infty$.

Let $B_\infty(D,X)$ denote the vector space of bounded Σ-measurable X valued functions which are uniform limits of Σ-measurable simple functions $\mathcal{S}(D,X)$. Endowed with sup norm topology, $B_\infty(D,X)$ is a B-Space. Let $\mathcal{L}_1(X,Y)$ denote the B-space of nuclear operators. It is well known that the Grothendieck characterization of $L \in \mathcal{L}_1(X,Y)$ is given by

$$Lx \equiv \sum \lambda_i x_i^*(x) y_i, x_i^* \in \partial B_1(X^*), y_i \in \partial B_1(Y),$$

with $\sum |\lambda_i| < \infty$, where $\{x_i, x_i^*\} \in X \times X^*, \{y_i, y_i^*\} \in Y \times Y^*$ are the normalized bi-orthogonal basis of the spaces X and Y, respectively.

3 Weak Compactness

Now we can present some recent results on the characterization of conditionally weakly compact sets in the space of operator valued measures. The first result presented here involves Hilbert spaces and nuclear operator valued measures.

Theorem 3.1. Let $\{X,Y\}$ be a pair of separable Hilbert spaces with complete ortho-normal basis $\{x_i, y_i\}$. A set $\Gamma \subset M_{ba}(\Sigma, \mathcal{L}_1(X,Y))$ is conditionally weakly compact if, and only if, the following conditions hold:

(c1): Γ is bounded,

(c2): for each $\sigma \in \Sigma$, $\sum_{i=1}^{\infty} |(M(\sigma)x_i, y_i)_Y|$ is convergent uniformly with respect to $M \in \Gamma$,

(c3): for each $i \in N$, the set of scalar valued measures $\{\mu_M(\cdot) = (M(\cdot)x_i, y_i), M \in \Gamma\}$ is a conditionally weakly compact subset of $M_{ba}(\Sigma)$.

Proof. [10, Theorem 3.2, PMD,(2010),p1-15]

This result was recently extended to more general spaces of operator valued measures. Here we consider $\{X,Y\}$ to be a pair of Banach spaces and replace the space of nuclear operators by $\mathcal{L}(X,Y)$, the space of bounded linear operators. Let

$$M_{casbsv}(\Sigma, \mathcal{L}(X,Y)) \subset M_{ba}(\Sigma, \mathcal{L}(X,Y))$$

denote the space of operator valued measures countably additive in the strong operator topology having bounded semivariations (variation in the strong operator topology). To proceed further, we need to consider the question of integration

of vector valued functions with respect to operator valued measures. The most general theory of integration was introduced by Dobrakov [5,6]. This generalizes the theory of Lebesgue integral, Bochner integral, Bartle bilinear integral and Dinculeanu integral etc.

For convenience of the reader we recall that a formal series $\sum x_n$, $x_n \in X$, is said to be unconditionally convergent if $\sum x_{\pi(n)}$ is convergent for every permutation $\pi : N \to N$. Orlicz-Pettis Theorem [1, Corollary 4, p.22] states that if every subseries of the series is weakly convergent then the series is unconditionally strongly convergent. This is the foundation of Dobrakov integral.

Dobrakov Integral: For any $f \in B_\infty(D, X)$ and $T \in M_{casbsv}(\Sigma, \mathcal{L}(X, Y))$ the integral,

$$I_T(f) \equiv \int_D T(ds)f(s) \quad \in Y,$$

is well defined in the sense of Dobrakov [5]. As usual the integral is first defined for simple functions $\mathcal{S}(D, X)$ and then extended to $B_\infty(D, X)$ by density argument. The most important point is that the limit is taken in the sense of unconditional convergence of the sum arising from the simple functions. This limit is the Dobrakov integral. This is unlike the Lebesgue and Bochner integrals which are based on absolute convergence. This is where the main difference is.

Now we can introduce the notion of Dobrakov semivariation as follows.

Definition 3.2. (Dobrakov semivariation) For any $T \in M_{casbsv}(\Sigma, \mathcal{L}(X, Y))$ and $\sigma \in \Sigma$ define the set function given by

$$\hat{T}(\sigma) \equiv \sup\left\{ | \int_\sigma T(ds)f(s)|_Y, f \in \mathcal{S}(D, X), \| f \|_\infty \leq 1 \right\}.$$

Then the Dobrakov semivariation of T over D is given by $\hat{T}(D) \equiv \sup\{\hat{T}(\sigma), \sigma \in \Sigma\}$.

The reader can easily verify that $\hat{T}(D) = |T|_s$. In other words, Dobrakov semi varition is the same as the variation in the strong operator topology.

We need few more concepts before we can return to the compactness issue.

Definition 3.3 (F-Space): A compact Hausdorff space D is said to be an $\mathcal{F}$-space if every pair of disjoint open F_σ set has disjoint closure, [4, Kuo].

Definition 3.4(Grothendieck Space): A Banach space X is said to be a Grothendieck space if weak star convergence in its dual X^* is equivalent to weak convergence. In other words the X topology of X^* is equivalent to the X^{**} topology of X^*.

Well known examples of Grothendieck spaces are reflexive Banach space and separable dual spaces. For more examples see Diestel & Uhl [1]. Another characterization of Grothedieck space X is that, for every separable Banach space

Y, every bounded linear operator from X to Y is weakly compact. If K is a compact metric space then $C(K)$, the space of continuous functions on K, is a Grothendieck space. The space $L_\infty(\mu)$ is a Grothendieck space if μ is a positive measure.

A General Result on Weak Compactness: Now we are prepared to present a general result characterizing weakly compact sets in $M_{casbsv}(\Sigma, \mathcal{L}(X,Y))$. Let $\Gamma \subset M_{casbsv}(\Sigma, \mathcal{L}(X,Y))$ and $f \in B_\infty(D,X)$. Define the set

$$\Gamma(f) \equiv \big\{\mu \in M_{ba}(\Sigma,Y) : \mu(\sigma) = \int_\sigma T(ds)f(s), \sigma \in \Sigma, T \in \Gamma\big\}.$$

It is easy to verify that $\Gamma(f) \subset M_{ca}(\Sigma,Y) \subset M_{ba}(\Sigma,Y)$.

Theorem 3.5. Suppose D is a compact Hausdorff $\mathcal{F}$-space, and $\{X,Y\}$ is a pair of B-spaces with Y being reflexive. Then a set $\Gamma \subset M_{casbsv}(\Sigma, \mathcal{L}(X,Y))$ is conditionally weakly compact if, and only if, the following conditions hold:

(i): Γ is bounded in the sense that $\sup\{\hat{T}(D) \equiv |T|_s, T \in \Gamma\} < \infty$.

(ii): For each $f \in B_\infty(D,X)$, the set $\{|\mu|(\cdot), \mu \in \Gamma(f)\}$ is uniformly c.a.

Proof. Detailed proof appears in Ahmed [[11], Theorem 1]. Here we present only a brief outline. Since Γ is bounded, for each $f \in B_\infty(D,X)$, the set $\Gamma(f)$ is a bounded subset of $M_{ca}(\Sigma,Y)$. By hypothesis (ii), $\Gamma(f)$ is uniformly countably additive. Thus, since Y is reflexive, it follows from Brooks theorem [3, Main Theorem, Cor.1, p284] that for each $f \in B_\infty(D,X)$, the set $\Gamma(f)$ is a conditionally weakly compact subset of $M_{ca}(\Sigma,Y)$. Rest of the proof presents arguments to demonstrate that this implies conditional weak compactness of Γ itself and conversely. The tools used are: a result of Kuo [4] that asserts that if D is an F-space and Y is reflexive then $C(D,Y^*)$ is a Grothedieck space. Thus w^* and weak convergence in $M_{ca}(\Sigma,Y)$ are equivalent. Next we use Nikodym uniform boundedness principle [1, Theorem 1, p14] and the classical uniform boundedness principle for linear operators. Then we use the fact that any closed bounded convex subset of $\mathcal{L}(X,Y)$ is compact in the weak operator topology if and only if Y is reflexive. We also use Hahn-Banach theorem to prove boundedness of the semivariation of the limit of any convergent sequence from Γ. Finally to prove countable additivity in the strong operator topology, we use Pettis theorem which states that a weakly countably additive vector measure defined on a sigma algebra is (strongly) countably additive. This completes the outline of our proof.

Some Remarks and Open Problems

(R1): Note that we have characterized conditionally weakly compact sets in the space $M_{casbsv}(\Sigma, \mathcal{L}(X,Y))$. This is a subspace of the space of finitely additive operator valued measures having finite semivariations denoted by $M_{fabsv}(\Sigma,(X,Y))$. The author believes that following our approach one can obtain characterization of conditional weak compactness in this space also. In any case it would be interesting to characterize weakly compact sets in the two larger spaces:

$$M_{fabsv}(\Sigma, \mathcal{L}(X,Y)) \subset M_{ba}(\Sigma, \mathcal{L}(X,Y)).$$

(R2): Another fact that we have used in the proof of the above theorem is that the closed unit ball $B_1(\mathcal{L}(X,Y))$ is compact in the weak operator topology. This means that Y is reflexive. In fact this is a necessary and sufficient condition for weak compactness of $B_1(\mathcal{L}(X,Y))$. It would be interesting to extend our result to cases where Y is not necessarily reflexive. In other words to improve our result one must avoid using the compactness of $B_1(\mathcal{L}(X,Y))$ in the weak operator topology.

(R3): One of the very important topic in vector measure theory is the representation theory like the Riesz representation theorem. By virtue of Dobrakov theory, every

$$T \in M_{casbsv}(\Sigma, \mathcal{L}(X,Y))$$

determines a linear operator $L_T \in \mathcal{L}(B_\infty(D,X),Y)$, the space of bounded linear operators from $B_\infty(D,X)$ to Y, satisfying $\parallel L_T \parallel = |T|_s$, and so we have the embedding

$$M_{casbsv}(\Sigma, \mathcal{L}(X,Y)) \hookrightarrow \mathcal{L}(B_\infty(D,X),Y).$$

The question is: does every $L \in \mathcal{L}(B_\infty(D,X),Y)$ have the integral representation with some $T \in M_{casbsv}(\Sigma, \mathcal{L}(X,Y))$. The answer seems to be no. Note that every $T \in M_{fabsv}(\Sigma, \mathcal{L}(X,Y))$ determines a continuous linear operator L_T on $B_\infty(D,X)$ to Y through the Dobrakov integral

$$L_T(f) = \int_D T(ds) f(s).$$

This follows from the fact that

$$|L_T(f)|_Y = |\int_D T(ds) f(s)|_Y \leq |T|_s \parallel f \parallel_{B_\infty(D,X)} .$$

This also shows that $M_{fabsv}(\Sigma, \mathcal{L}(X,Y)) \hookrightarrow \mathcal{L}(B_\infty(D,X),Y)$. In fact we can prove that

$$M_{fabsv}(\Sigma, \mathcal{L}(X,Y)) \cong \mathcal{L}(B_\infty(D,X),Y).$$

(**R4**): Since Y is a reflexive Banach space, every operator $L \in \mathcal{L}(B_\infty(D,X),Y)$ is weakly compact in the sense that it maps any bounded subset of $B_\infty(D,X)$ into a relatively weakly compact subset of Y. Also for the same reason the closed unit ball $\mathcal{B}_1(\mathcal{L}(B_\infty(D,X),Y))$ is compact in the weak operator topology. Thus every net $\{L_\alpha, \alpha \in \Lambda\} \in \mathcal{B}_1$ has a subnet that converges in the weak operator topology to some $L_o \in \mathcal{B}_1(\mathcal{L}(B_\infty(D,X),Y))$. If the net $\{L_\alpha, \alpha \in \Lambda\}$ were generated by a net $B_\alpha \in M_{casbsv}(\Sigma, \mathcal{L}(X,Y))$, the limit operator L_o may not be represented by a measure B_o from $M_{casbsv}(\Sigma, \mathcal{L}(X,Y))$. In view of the isomorphism stated above, the representing measure B_o corresponding to L_o may very well be an element of $M_{fabsv}(\Sigma, \mathcal{L}(X,Y))$.

4 Applications to Optimal Feedback Control

In this section we present two applications of weak compactness to structural control problems in infinite dimension involving deterministic systems. The result also applies to stochastic systems on Hilbert spaces.

Consider the structural control system on a real Banach space X

$$dx = Axdt + B(dt)y + f(x)dt, x(0) = \xi \tag{1}$$

$$y = Lx + \eta \quad \text{(output)} \tag{2}$$

over the time interval $t \in [0, T]$. The state space X is a reflexive B-space and the output space Y is any real Banach space. The operator $L \in \mathcal{L}(X, Y)$ represents the sensor and η is a deterministic bounded Y valued perturbation. The objective functional is given by

$$J(B) \equiv \int_0^T \ell(t, x(t))dt + |B|_s, \tag{3}$$

where $|B|_s$ denotes the semivariation of B over the set I. The admissible set of structural controls is given by a set $\Gamma \subset M_{casbsv}(\Sigma_I, \mathcal{L}(Y, X))$. The objective is to find a control that minimizes this functional.

Let $\mathcal{G}_0(M, \omega)$ denote the class of infinitesimal generators $\{A\}$ of C_0-semigroups of linear operators on X with stability parameters (M, ω) for $M \geq 1$ and $\omega \in R$.

Theorem 4.1. Suppose $A \in \mathcal{G}_0(M, \omega)$ generating the semigroup $S(t), t \geq 0$, compact for $t > 0$, Γ a weakly compact subset of $M_{casbsv}(\Sigma_I, \mathcal{L}(Y, X))$, f locally Lipschitz with at most linear growth, $L \in \mathcal{L}(X, Y)$, $\eta \in B_\infty(I, Y)$. There exists $\nu \in M^+_{cabv}(\Sigma_I)$ such that $|B|_s(\sigma) \leq \nu(\sigma)$ for $\sigma \in \Sigma_I$ uniformly w.r.t $B \in \Gamma$. The cost integrand ℓ is measurable in t and lower semicontinuous in x on X and there exists $\alpha \in L^+_1(I)$ and $\beta \geq 0$ satisfying

$$|\ell(t, x)| \leq \alpha(t) + \beta|x|^p_X, \text{ for any } p \in (0, \infty).$$

Then, there exists a $B_o \in \Gamma$ at which J attains its minimum.

Proof. For detailed proof see [11, Theorem 1]. We present a brief outline. For existence and uniqueness of (mild) solutions of the system the reader may see [8, Theorem 3.5, p106]. According to this theorem, the mild solutions are elements of $B_\infty(I, X)$. We concentrate on the question of existence of optimal controls. For $\xi \in X$, and $B \in \Gamma$, let $x(B)(\cdot) \in B_\infty(I, X)$ denote the mild solution of the system (1) (2). Under the given assumptions, it is easy to verify that there exists a ball $B_r \subset X, r \in (0, \infty)$, such that $x(B)(t) \in B_r$ for all $t \in I$ and all $B \in \Gamma$. Since the set $\Gamma \subset M_{casbsv}(\Sigma_I, \mathcal{L}(Y, X))$ is weakly compact, it suffices to prove that $B \longrightarrow J(B)$ is weakly lower semicontinuous. Let $B_n \xrightarrow{w} B_o$ in $M_{casbsv}(\Sigma_I, \mathcal{L}(Y, X))$ and let $\{x_n, x_o\}$ denote the corresponding mild solutions of the output feedback system

$$dx = Axdt + B(dt)Lx + f(x)dt + B(dt)\eta(t),$$
$$x(0) = \xi, \tag{4}$$

corresponding to $B = B_n$ and $B = B_o$, respectively. Using compactness property of the semigroup $S(t), t > 0$, and a generalized Gronwall type inequality [12] relative to a nonnegative countably additive measure, we show that $x_n(t) \xrightarrow{s} x_o(t)$ in X for each $t \in I$. Then it follows from lower semicontinuity of ℓ in the second argument that

$$\ell(t, x_o(t)) \leq \liminf \ell(t, x_n(t)) \ a.a \quad t \in I.$$

By use of Hahn-Banach theorem and weak convergence of B_n to B_o in the space $M_{casbsv}(\Sigma_I, \mathcal{L}(Y, X))$, one can verify without any difficulty that $|B_o|_s \leq \liminf |B_n|_s$. By assumption on f and ℓ, and $\{\alpha, \beta, p\}$, we have $\ell(\cdot, x_o(\cdot)) \in L_1(I)$ implying $J(B) > -\infty$ for all $B \in \Gamma$. Using these results and Fatou's lemma, we obtain

$$J(B_o) \leq \liminf J(B_n).$$

This shows that $B \longrightarrow J(B)$ is w.ℓ.s.c on $M_{casbsv}(\Sigma_I, \mathcal{L}(Y, X))$ and bounded away from $-\infty$. Since by assumption Γ is compact with respect to the weak topology, J attains its minimum on Γ. •

Remark. Assumption on compactness of the semigroup $S(t)$ can be relaxed by imposing a stronger assumption on the admissible set Γ. For example, one may assume that Γ is compact in the sense that any sequence $B_n \in \Gamma$ has a subsequence that converges in the strong operator topology set-wise on Σ to an element of Γ. That is, there exists a $B_o \in \Gamma$ such for each $\sigma \in \Sigma$, $B_{n_k}(\sigma) \xrightarrow{\tau_{so}} B_o(\sigma)$ in $\mathcal{L}(X, Y)$.

Time Optimal Control. Given the initial state $\xi \in X$ and a nonempty closed target set $C \subset X$ not containing ξ, the problem is to find a control $B \in \Gamma$ that drives the system to C in minimum time. Since the solutions $\{x^B, B \in \Gamma\}$ are elements of $B_\infty(I, X)$, and so not necessarily continuous, we must formulate the objective functional as follows:

$$J(B) \equiv inf\{t \geq 0 : \int_0^t I_C(x^B(s)) \ ds > \varepsilon\}, \tag{5}$$

where $I_C(x)$ is the characteristic function of the set C taking value 1 for $x \in C$ and 0 outside C and $\varepsilon \in (0, 1)$. Note that $\varepsilon > 0$ can chosen as small as necessary. We use the convention that $\inf(\emptyset) = +\infty$. The problem is to find $B_o \in \Gamma$ so that $J(B_o) \leq J(B)$ for all $B \in \Gamma$.

Theorem 4.2. Suppose the assumptions of Theorem 4.1 related to $\{A, f, L, \eta, \Gamma\}$ hold. Consider the output feedback system (4) and the time optimal control problem as stated above with the target set C, a nonempty closed subset of X, not containing ξ. Suppose there exists at least one $B \in \Gamma$ for which the set $\{t \geq 0 : \int_0^t I_C(x^B(s))ds > \varepsilon\} \neq \emptyset$. Then there exists a time optimal control.

Proof. For lack of space we can only present a brief outline. Since C is a closed set the characteristic function I_C is upper semicontinuous on X. Using this fact

we prove that the map $B \longrightarrow J(B)$ given by the expression (5) is weakly lower semicontinuous on Γ and since this set is weakly compact J attains its minimum on Γ. Thus time optimal control exists. •

References

1. Diestel, J., Uhl Jr., J.J.: Vector Measures. American Mathematical Society, Providence (1977)
2. Dunford, N., Schwartz, J.T.: Linear Operators, Part 1, General Theory, Second Printing (1964)
3. Brooks, J.K.: Weak Compactness in the Space of Vector Measures. Bulletin of the American Mathematical Society 78(2), 284–287 (1972)
4. Kuo, T.: Weak Convergence of Vector Measures on F-Spaces. Math. Z. 143, 175–180 (1975)
5. Dobrakov, I.: On integration in Banach spaces I. Czechoslovak Mathematical Journal 20(95), 511–536 (1970)
6. Dobrakov, I.: On Integration in Banach Spaces IV. Czechoslovak Mathematical Journal 30(105), 259–279 (1980)
7. Brooks, J.K., Lewis, P.W.: Linear Operators and Vector Measures. Trans. American Math. Soc. 192, 139–162 (1974)
8. Ahmed, N.U.: Vector and Operator Valued Measures as Controls for Infinite Dimensional Systems: Optimal Control. Differential Inclusions, Control and Optimization 28, 95–131 (2008)
9. Ahmed, N.U.: Impulsive Perturbation of C_0-Semigroups by Operator Valued Measures. Nonlinear Funct. Anal. & Appl. 9(1), 127–147 (2004)
10. Ahmed, N.U.: Weak Compactness in the Space of Operator Valued Measures. Publicationes Mathematicae, Debrechen (PMD) 77(3-4), 399–413 (2010)
11. Ahmed, N.U.: Weak Compactness in the Space of Operator Valued Measures $M_{ba}(\Sigma, \mathcal{L}(X,Y))$ and its Applications. Differential Inclusions, Control and Optimization 31, 231–247 (2011)
12. Ahmed, N.U.: Some Remarks on the Dynamics of Impulsive Systems in Banach Spaces. Dynamics of Continuous, Discrete and Impulsive Systems 8, 261–274 (2001)
13. Ahmed, N.U.: Existence of Optimal Controls for a General Class of Impulsive Systems on Banach Spaces. SIAM J. Control. Optim. 42(2), 669–685 (2003)
14. Ahmed, N.U.: Dynamics of Hybrid systems Induced by Operator Valued Measures. Nonlinear Analysis: Hybrid Systems 2, 359–367 (2008)

Adaptive Methods for Control Problems with Finite-Dimensional Control Space*

Saheed Akindeinde and Daniel Wachsmuth

Johann Radon Institute for Computational and Applied Mathematics (RICAM)
Austrian Academy of Sciences
Altenbergerstraße 69
A-4040 Linz, Austria

Abstract. We investigate adaptive methods for optimal control problems with finitely many control parameters. We analyze a-posteriori error estimates based on verification of second-order sufficient optimality conditions. Reliability and efficiency of the error estimator is shown. The estimator is used in numerical tests to guide adaptive mesh refinement.

Keywords: optimal control, numerical approximation, a-posteriori error estimates, adaptive refinement.

1 Introduction

We study optimal control problems of the following type: Minimize the functional J given by

$$J(y,u) = g(y) + j(u) \tag{P}$$

over all $(y,u) \in Y \times U$ that satisfy the non-linear elliptic partial differential equation

$$E(y,u) = 0$$

and the control constraints

$$u \in U_{ad}.$$

Here, Y is a real Banach space, $U = \mathbb{R}^n$. The set $U_{ad} \subset U$ is a non-empty, convex and closed set given by $U_{ad} = \{u \in U : \ u_a \leq u \leq u_b\}$, where the inequalities are to be understood component-wise. Here, the cases $u_a = -\infty$ and $u_b = +\infty$ are allowed, such that problems with one-sided constraints or without control constraints are included in the analysis as well. Examples that are covered by this framework include parameter identification and optimization problems with finitely many parameters, see for instance our previous work [1].

Adaptive mesh refinement remains a valuable tool in scientific computation. The main objective of an adaptive procedure is to find a discrete solution to a problem while maintaining as few as possible numbers of unknowns with respect to a desired error estimate. As the solution and hence the error distributions on the mesh are unknown a-priori, one has to rely on a-posteriori error estimates.

* This work was funded by Austrian Science Fund (FWF) grant P21564-N18.

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 59–69, 2013.

A-posteriori error estimates for non-linear control and identification problems can be found for instance in [2,5,6,9]. However, they depend on two crucial *a-priori* assumptions: the first is that a second-order sufficient condition (SSC) has to hold at the solution of the continuous problem. With this assumption, error estimates of the type $\|\bar{u} - u_h\|_U \leq c\,\eta + \mathcal{R}$ can be derived, where η is a computable error indicator and $\mathcal{R}$ is a second-order remainder term. Here, the second a-priori assumption comes into play: one has to assume that $\mathcal{R}$ is small enough, in order to guarantee that mesh refinement solely based on η is meaningful. A different approach with respect to mesh refinement was followed in [13]. There the residuals in the first-order necessary optimality condition were used to derive an adaptive procedure. However, smallness of residuals does not imply smallness of errors without any further assumption. Here again, SSC as well as smallness of remainder terms is essential to draw this conclusion.

In our previous work [1], we applied a different strategy: There the sufficient optimality condition as well as smallness of remainders is checked *a-posteriori*. If both conditions are fulfilled, an error-estimator of the form

$$\|u - u_h\|_U \leq \frac{2}{\alpha}(\omega_y r_y + \omega_p r_p)$$

is available, see [1, Thm 3.22]. This error estimator is localizable if r_y and r_p are localizable error estimates for the norm of the residual in the state and adjoint equations, respectively. For earlier and related work on a special problem calls with infinite-dimensional control space, we refer to [7,8].

In this article, we will prove a lower bound of the error estimator. For the setting $Y = H_0^1(\Omega)$, we obtain

$$\begin{aligned} r_y + r_p \leq c(\|u - u_h\|_U + \|y - y_h\|_Y + \|\nabla y - \sigma_h\|_{L^2(\Omega)} \\ + \|p - p_h\|_Y + \|\nabla p - \tau_h\|_{L^2(\Omega)} + \tilde{o}), \end{aligned}$$

where y, p and y_h, p_h are solutions of continuous and discrete state and adjoint equations, respectively, and σ_h and τ_h are approximations of ∇y_h and ∇p_h in $H(\mathrm{div})$. The term $\tilde{o}$ is a higher-order oscillation term. In addition, we have localized lower bounds for the residuals in the state and adjoint equations, respectively. This justifies the use of the a-posteriori estimator above in an adaptive mesh-refinement procedure.

1.1 The Abstract Framework

Let Ω be a polygonal domain in $\mathbb{R}^m$, $m = 2, 3$. The function space for the states of the optimal control problem is chosen as $Y := H_0^1(\Omega)$. Let us now specify assumptions on the abstract problem (P).

Assumption 1. *The mapping $E : Y \times U \to Y^*$, $g : Y \to \mathbb{R}$, and $j : U \to \mathbb{R}$ are twice continuously Fréchet-differentiable with locally Lipschitz continuous second-derivatives. Furthermore, we assume that the mapping E is strongly monotone with respect to the first variable at all points $u \in U_{ad}$.*

The assumptions on E are met for instance for semilinear elliptic equations with monotone nonlinearities. Under Assumption (1), the state equation $E(y,u)=0$ is uniquely solvable for each admissible control $u \in U_{ad}$, [12, Theorem 26.A, p. 557]. We remark that the differentiability assumption on E can be relaxed to accomodate a more general class of problems, see [1, Remark 1.2], e.g. differentiability of E from $(Y \cap L^\infty(\Omega)) \times U$ to Y^* is sufficient.

Let us define the Lagrange functional for the abstract problem:

$$\mathcal{L}(u,y,p) := g(y) + j(u) - \langle E(y,u),\, p\rangle_{Y^*,Y}.$$

Let $(\bar{y},\bar{u})$ be locally optimal for (P). Then the first-order necessary optimality conditions can be expressed as $\mathcal{L}'_y(\bar{y},\bar{u},\bar{p})=0$ and $\mathcal{L}_u(\bar{y},\bar{u},\bar{p})(u-\bar{u}) \geq 0$ for all $u \in U_{ad}$, which is equivalent to

$$\begin{aligned} E_y(\bar{u},\bar{y})^*\bar{p} &= g'(\bar{y}) \\ \langle j'(\bar{u}) - E_u(\bar{u},\bar{y})^*\bar{p},\, u-\bar{u}\rangle_{U^*,U} &\geq 0 \quad \forall u \in U_{ad}. \end{aligned}$$

Since the problem (P) is in general non-convex, the fulfillment of these necessary conditions does not imply optimality. In order to guarantee this, one needs additional sufficient optimality conditions of the type: There exists $\alpha > 0$ such that

$$\mathcal{L}''(\bar{u},\bar{y},\bar{p})[(z,v)^2] \geq \alpha \|v\|_U^2 \tag{1}$$

holds for all $v = u-\bar{u}$, $u \in U_{ad}$, and z solves the linearized equation $E_y(\bar{u},\bar{y})z + E_u(\bar{u},\bar{y})v = 0$. This condition can be weakened taking strongly active inequality constraints into account, see e.g. [1,3]. For simplicity, we chose to work with this stronger condition. The results of this article hold also under the weakened sufficient condition.

Although, the sufficient condition is of high interest, it is difficult to check numerically even when $(\bar{u},\bar{y},\bar{p})$ are given, see e.g. [1,7,8]. The main difficulty here is that the function z appearing in (1) is given as solution of a partial differential equation, which cannot be solved explicitly. Any discretization of this equation introduces another error that has to be analyzed.

1.2 Discretization

In order to solve (P) numerically, we discretize the problem. Let Y_h be a finite-dimensional subspace of Y. Here and in the following, the index h denotes a discrete quantity. Then a discretization of the state equation can be obtained in the following way: A function $y_h \in Y_h$ is a solution of the discretized equation for given $u \in U_{ad}$ if and only if

$$\langle E(y_h,u),\, \phi_h\rangle_{Y^*,Y} = 0 \qquad \forall \phi_h \in Y_h. \tag{2}$$

The discrete optimization problem is then given by: Minimize the functional $J(y_h,u_h)$ over all $(y_h,u_h) \in Y_h \times U_{ad}$, where y_h solves the discrete equation.

Let $(\bar{y}_h, \bar{u}_h)$ be a local solution of the discrete problem. Then it fulfills the discrete first-order necessary optimality condition, which is given as: there exists a uniquely determined discrete adjoint state $\bar{p}_h \in Y_h$ such that it holds

$$\begin{aligned} \langle E_y(\bar{y}_h, \bar{u}_h)^* \bar{p}_h,\, \phi_h \rangle_{Y^*,Y} &= \langle g'(\bar{y}_h),\, \phi_h \rangle_{Y^*,Y} \quad \forall \phi_h \in Y_h \\ \langle j'(\bar{u}_h) - E_u(\bar{y}_h, \bar{u}_h)^* \bar{p}_h, u - \bar{u}_h \rangle_{U^*,U} &\geq 0 \ \forall u \in U_{ad}. \end{aligned} \tag{3}$$

Throughout this work, we will assume that errors in discretizing the optimality system are controllable in the following sense. We will not make any further assumptions on the discretization, in particular, we do not assume a sufficient fine discretization.

Assumption 2. *For a fixed finite-dimensional subspace Y_h, let (u_h, y_h, p_h) be approximations of the discrete optimal control and the corresponding state and adjoint. There are positive constants r_y, r_p such that the following holds*

$$\|E(y_h, u_h)\|_{Y^*} \leq r_y, \tag{4}$$

$$\|g'(y_h) - E_y(y_h, u_h)^* p_h\|_{Y^*} \leq r_p, \tag{5}$$

$$\langle j'(u_h) - E_u(y_h, u_h)^* p_h, u - u_h \rangle_{U^*,U} \geq 0 \ \forall u \in U_{ad}. \tag{6}$$

Here, r_y and r_p are dual norms of residuals in the state and adjoint equation, respectively, and hence reflect the discretization error. We report on the computation of these residuals in Section 2.

As already mentioned, without any further assumption, smallness of the residuals in (4)–(6) does not imply smallness of the error $\|u - u_h\|_U$ in the control. In order to establish such a bound, it is essential to check that a second-order sufficient optimality condition is satisfied.

Here it is important to recognize that sufficient optimality conditions for the *discrete* problem alone are still not enough. The sufficient optimality condition for the discrete problem is given by: There exists $\alpha_h > 0$ such that

$$\mathcal{L}''(\bar{u}_h, \bar{y}_h, \bar{p}_h)[(z_h, v)^2] \geq \alpha_h \|v\|_U^2 \tag{7}$$

holds for all $v = u - \bar{u}$, $u \in U_{ad}$, and z_h solves the linearized discrete equation

$$\langle E_y(\bar{u}_h, \bar{y}_h) z_h + E_u(\bar{u}_h, \bar{y}_h) v,\ \phi_h \rangle_{Y^*,Y} = 0 \quad \forall \phi_h \in Y_h. \tag{8}$$

This condition is equivalent to positive definiteness of a certain computable matrix, see [1, Section 3.5]. Moreover we have the following estimate relating the coercivity constants α and α_h appearing in (11) and (7):

$$\alpha \geq \alpha_h - \|\mathcal{E}\|_2,$$

where $\|\mathcal{E}\|_2$ is the norm of an error matrix taking the discretization error in the linearized equation $E_y(\bar{u}, \bar{y})z + E_u(\bar{u}, \bar{y})v = 0$ into account, see [1, Section 3.5] for the details. If the computable lower bound $\alpha_h - \|\mathcal{E}\|_2$ of α is positive, then it follows that (11) is satisfied. Moreover, we have the following result:

Theorem 1 (Upper bound of the error). *Let Assumptions 1 and 2 be satisfied. Let (y_h, u_h, p_h) be a solution of the discrete optimal control problem. If $\alpha_h - \|\mathcal{E}\|_2 > 0$ holds and the residuals r_y and r_p are small enough, then there exists a local solution $\bar{u}$ of (P) that satisfies the error bound*

$$\|\bar{u} - u_h\|_U \leq \frac{2}{\alpha_h - \|\mathcal{E}\|_2} \left(\omega_y \, r_y + \omega_p \, r_p\right), \tag{9}$$

where the weights ω_y, ω_p depend on the discrete solution (y_h, u_h, p_h). If for different discretizations the discrete solutions $\{(y_h, u_h, p_h)\}_{h>0}$ are uniformly bounded in $Y \times U \times Y$ then the weights ω_y, ω_p are bounded as well.

For the proof, we refer to [1, Thm. 3.22]. There, precise estimates of the weights ω_y, ω_p are given. Moreover, a quantification of the smallness assumption on r_y and r_p is given, which makes this assumption verifiable *a-posteriori*.

Corollary 1. *Let the assumptions of Theorem 1 be satisfied. Let $\bar{y}$, $\bar{p}$ denote the solutions of the state and adjoint equations to $\bar{u}$, respectively. Then it holds*

$$\begin{aligned}
\|\bar{y} - y_h\|_Y &\leq \upsilon_{yu} \|\bar{u} - u_h\|_U + \delta^{-1} r_y, \\
\|\bar{p} - p_h\|_Y &\leq \upsilon_{pu} \|\bar{u} - u_h\|_U + \delta^{-1} r_p + \upsilon_{py} r_y,
\end{aligned}$$

with δ^{-1} being the global bound of $\|E_y^{-1}(y,u)\|_{\mathcal{L}(Y^*,Y)}$, and weights υ_{yu}, υ_{pu}, and υ_{py} depending on (y_h, u_h, p_h) in the same way as the weights ω_y, ω_p in Theorem 1.

Proof. The result is a consequence of Theorem 1 and [1, Lemma 3.1, 3.3].

2 Lower Error Bounds

In this section, we assume that the general operator E can be written as $E(y,u) = -\Delta y + d(y,u)$ with d being a superposition operator induced by a smooth function $d : \mathbb{R}^2 \to \mathbb{R}$. We remark that the subsequent analysis can be easily extended to operators in divergence form with bounded coefficients possibly depending on u [1]. We will work with a classical finite-element discretization: The discrete space Y_h is the classical space of piecewise quadratic and continuous elements (P2) on a given conforming triangulation $\mathcal{T}_h$ of Ω. The diameter of an element $T \in \mathcal{T}_h$ is denoted by h_T. We denote by $\Sigma_h \subset H(\mathrm{div})$, a conforming Raviart-Thomas (RT_1) discretization of the space $H(\mathrm{div})$.

Let us endow $Y = H_0^1(\Omega)$ with the norm $\|y\|_Y^2 := \|\nabla y\|_{L^2(\Omega)}^2 + \|y\|_{L^2(\Omega)}^2$. In the sequel, let us denote the norm of the embedding $H_0^1(\Omega) \hookrightarrow L^2(\Omega)$ by I_2.

Now let us report on the computation of the residual r_y in the state equation. As required by Assumption 2, we are interested in constant-free error estimates, i.e. all constants appearing in the a-posteriori error estimate must be computable. Here, we apply the results of Vohralík [11].

Theorem 2. *Let $y_h \in Y_h \subset H_0^1(\Omega)$, $u_h \in U_{ad}$ satisfy the discrete equation (2). Let $\sigma_h \in \Sigma_h \subset H(\mathrm{div})$ be given such that*

$$(\mathrm{div}\,\sigma_h,\,1)_{L^2(T)} = (d(y_h,u),\,1)_{L^2(T)} \quad \text{for all cells } T \in \mathcal{T}_h. \tag{10}$$

Let us define the cell-wise indicator $\eta_{y,T}$, $T \in \mathcal{T}_h$,

$$\eta_{y,T} := 2\|\nabla y_h - \sigma_h\|_{L^2(T)} + \pi^{-1} h_T \|d(y_h,u_h) - \mathrm{div}\,\sigma_h\|_{L^2(T)} \tag{11}$$

Then it holds

$$\|-\Delta y_h + d(y_h,u_h)\|^2_{H^{-1}(\Omega)} \le (1-I_2^2)^{-1} \sum_{K\in\mathcal{T}_h} \eta_{y,T}^2 =: r_y^2. \tag{12}$$

If moreover, $\mathcal{T}_h$ is shape-regular, then it holds

$$\eta_{y,T} \le C\|\nabla(\tilde{y}-y_h)\|_{L^2(T)} + c\|\nabla\tilde{y}-\sigma_h\|_{L^2(T)} + ch_T\|d(y_h,u_h) - \Pi_h d(y_h,u_h)\|_{L^2(T)}, \tag{13}$$

where $\tilde{y} := \Delta^{-1} d(y_h,u_h)$ and Π_h denotes the orthogonal L^2-projection onto Y_h. The constants C, c depend only on the spatial dimension m and the shape regularity of the triangulation.

Proof. The upper bound (12) is a consequence of [11, Thm. 6.8, 6.12] taking [11, Remark 6.3] into account for σ_h satisfying $(\mathrm{div}\,\sigma_h,\,1)_{L^2(T)} = (d(y_h,u),\,1)_{L^2(T)}$, $T \in \mathcal{T}_h$. The lower bound (13) follows from [11, Thm. 6.16], see also [10, Lemma 7.6]. Since $d(y_h,u)$ is in general not in the discrete space Y_h, we obtain the additional oscillation term $h_T\|d(y_h,u) - \pi_h d(y_h,u)\|_{L^2(T)}$ by a standard argument.

Estimates of the residual in the adjoint equation can be obtained after obvious modifications: for $\tau_h \in \Sigma_h$ satisfying

$$(\mathrm{div}\,\tau_h,\,1)_{L^2(T)} = (d'(y_h,u_h)p_h - g'(y_h),\,1)_{L^2(T)} \quad \text{for all cells } T \in \mathcal{T}_h \tag{14}$$

and with the local error indicators defined by

$$\eta_{p,T} := 2\|\nabla p_h - \tau_h\|_{L^2(T)} + \pi^{-1} h_T \|d'(y_h,u_h)p_h - g'(y_h) - \mathrm{div}\,\tau_h\|_{L^2(T)} \tag{15}$$

we obtain the upper bound

$$\|-\Delta p_h + d'(y_h,u_h)p_h - g'(y_h)\|^2_{H^{-1}(\Omega)} \le (1-I_2^2)^{-1} \sum_{K\in\mathcal{T}_h} \eta_{p,T}^2 =: r_p^2. \tag{16}$$

as well as the lower bound

$$\begin{aligned}\eta_{p,T} \le C\|\nabla(\tilde{p}-p_h)\|_{L^2(T)} + c\|\nabla\tilde{p}-\tau_h\|_{L^2(T)} \\ + ch_T\|(I-\Pi_h)(d'(y_h,u_h)p_h - g'(y_h))\|_{L^2(T)},\end{aligned} \tag{17}$$

where $\tilde{p} := \Delta^{-1}(d'(y_h,u_h)p_h - g'(y_h))$.

We remark that the upper bounds (12) and (16) are constant-free, making them explicitly computable. In our computations, we computed the functions

σ_h and τ_h as a minimizer of the right-hand side in (12) and (16), respectively, using Raviart-Thomas elements for discretization of $H(\mathrm{div})$. This shows that the requirements of Assumption 2 on the computability of upper bounds on the residuals can be fulfilled.

Now let us argue that under the assumptions of Theorem 1 we also obtain lower bounds for the error, which proves efficiency of the error bound.

Theorem 3. *Let the assumptions of Theorem 1 be fulfilled. Let r_y and r_p be computed according to (12) and (16). Let $(\bar{y}, \bar{u}, \bar{p})$ be the local solution of (P) provided by Theorem 1. Then it holds*

$$\sum_{T\in\mathcal{T}_h} r_{y,T} \le C\Big(\|\bar{u}-u_h\|_U + \|\bar{y}-y_h\|_Y + \|\nabla\bar{y}-\sigma_h\|_{L^2(\Omega)} + \|h_T(I-\Pi_h)d(y_h,u_h)\|_{L^2(\Omega)}\Big), \quad (18)$$

$$\sum_{T\in\mathcal{T}_h} r_{p,T} \le C\Big(\|\bar{u}-u_h\|_U + \|\bar{y}-y_h\|_Y + \|\bar{p}-p_h\|_Y + \|\nabla\bar{p}-\tau_h\|_{L^2(\Omega)} + \|h_T(I-\Pi_h)(d'(y_h,u_h)p_h - g'(y_h))\|_{L^2(\Omega)}\Big), \quad (19)$$

where $C > 0$ depends only on the spatial dimension m, the shape regularity of the triangulation, and global bounds of derivatives d_y, d_u, d_{yy}, and d_{yu} of $d : Y \times U \to Y^*$ near (y_h, u_h).

Proof. Let $\tilde{y}$ be given by $\tilde{y} := \Delta^{-1}d(y_h, u_h)$. Then we can estimate

$$\|\nabla(\tilde{y}-y_h)\|_{L^2(\Omega)} + \|\nabla\tilde{y}-\sigma_h\|_{L^2(\Omega)} \le 2\|\nabla(\tilde{y}-\bar{y})\|_{L^2(\Omega)} + \|\nabla(\bar{y}-y_h)\|_{L^2(\Omega)} + \|\nabla\bar{y}-\sigma_h\|_{L^2(\Omega)}.$$

Using Lipschitz continuity of d, we find

$$\|\nabla(\tilde{y}-\bar{y})\|_{L^2(\Omega)} \le C(\|\bar{u}-u_h\| + \|\bar{y}-y_h\|_Y),$$

with C depending on bounds of $\|d'\|_{\mathcal{L}(Y\times U, Y^*)}$ near (y_h, u_h). This together with (13) proves (18). The estimate (19) can be obtained analogously.

These lower bounds together with (9) and the local lower bounds (13) and (17) justify the use of the error indicators in an adaptive mesh-refinement procedure.

Remark 1. Another possibility of constant-free a-posteriori error estimators based on $H(\mathrm{div})$-functions is described in [4]. There, fluxes across edges in a dual mesh are prescribed instead of the integrals on elements as in (10) and (14). In [4] it is proven that the resulting error estimate is reliable and efficient. Moreover, the terms $\|\nabla y_h - \sigma_h\|_{L^2(\Omega)}$ and $\|\nabla p_h - \tau_h\|_{L^2(\Omega)}$ do not appear in the lower error bound when compared to (18) and (19), respectively.

3 Adaptivity

We will compare the performance of adaptive mesh refinement using different strategies to mark elements for refinement. The first one, referred to as *'verified adaptive'*, is implemented as follows: in each step the verification procedure of [1] is carried out. If it confirms that the assumptions of Theorem 1 are satisfied, then the error indicator $\omega_y r_y + \omega_p r_p$ given by (9) is used to guide the mesh-refinement. If the requirements of Theorem 1 cannot be verified, then a uniform refinement step is carried out. Here, we expect that after a small number of uniform refinement steps the requirements of Theorem 1 are confirmed a-posteriori, which coincides with the numerical experiments done in earlier work [1]. After these initial uniform refinements steps, we expect that the method proceeds with adaptive steps.

A second strategy, called *'fully adaptive'*, omits the verification step, and simply uses $\omega_y r_y + \omega_p r_p$ from (9) without checking the validity of this bound.

4 Numerical Results

Let us report about the outcome of the above described adaptive methods for a selected example, taken from [1]. The functional J was chosen as

$$J(y,u) := \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} + \frac{\kappa}{2}\|u\|^2_{\mathbb{R}^n}.$$

The nonlinear mapping E represents a semi-linear elliptic equation given by

$$E(y,u) := -\Delta y + \sum_{k=1}^{n} u_k d_k(y) - g, \tag{20}$$

where the functions d_k are chosen as $d_1(y) = 1$, $d_j(y) = y|y|^{j-2}$ for $j = 2 \dots n$. This example is motivated by parameter identification: given a state y_d and source term g, find the set of coefficients u such that the resulting solution y of $E(y,u) = 0$ is as close as possible to y_d.

In order to make the operator E strongly monotone, we require positivity of the coefficients u_k, i.e. we set $U_{ad} = \{u \in \mathbb{R}^n : \ u_k \geq 0 \quad \forall k = 1 \dots n\}$. For the computations we used the following data: the source term $g = 10.0001$ and

$$\Omega = (0,1)^2, \ u_a = 0, \ u_b = 0.5, \ \kappa = 10^{-2}, \ y_d(x_1,x_2) = 0.5\sin(2\pi x_1 x_2), \ n = 4.$$

Let us remark, that the function d_3 is not of class C^2 globally. Since g is non-negative, every solution y of (20) to $u \in U_{ad}$ will be non-negative. For non-negative functions y it holds $d_3(y) = y^2$, which is C^2, so the assumptions on E are satisfied. See also the discussion in [1, Section 4.3].

We employed a discretization scheme as described in Section 2. After the resulting non-linear optimization problem is solved, the error indicators according to the chosen strategy are computed. For an adaptive refinement, a subset

$\tilde{\mathcal{T}} \subset \mathcal{T}$ of elements T with large local error contributions η_T were selected for refinement that satisfies $\sum_{T\in\tilde{\mathcal{T}}} \eta_T^2 \geq \theta^2 \sum_{T\in\mathcal{T}} \eta_T^2$ with $\theta = 0.8$.

Let us report on the outcome of the different adaptive strategies as described in Section 3. For all the methods, we compare the residual norms as given in (11), i.e. with the notation of that section

$$\epsilon_{\text{residual}} := \omega_y r_y + \omega_p r_p.$$

Moreover, we employed the verification procedure of Theorem 1, see e.g. [1], and report about the upper error bound

$$\epsilon_{\text{bound}} := \frac{2}{\alpha_h - \|\mathcal{E}\|_2} \left(\omega_y\, r_y + \omega_p\, r_p\right).$$

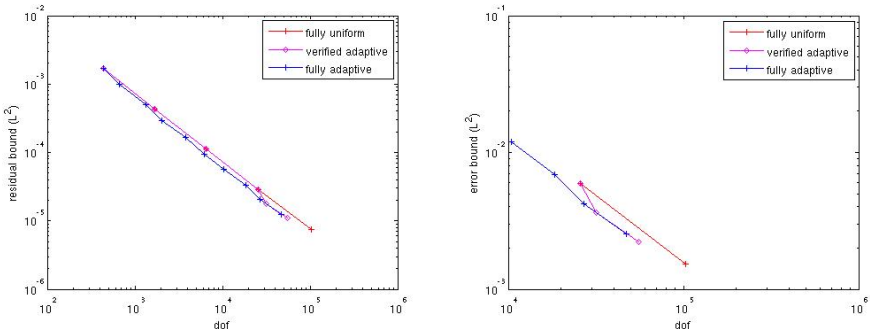


Fig. 1. (a) upper bound of residuals versus number of unknowns, (b) verified error bound versus number of unknowns

As can be expected, the assumptions of Theorem 1 are only fulfilled on a sufficiently fine discretization. This is reflected by our numerical results.

Plots of $\epsilon_{\text{residual}}$ and ϵ_{bound} versus the number of degrees of freedom can be seen in Figure 4. For reference, we provided the numerical values in Tables 1 and 2. In the tables, L refers to refinement level, where $L = 0$ is the initial mesh, which is the same for all the different adaptive methods. Moreover, dof denotes the number of degrees of freedom.

Let us comment on the observed behavior of the verified adaptive methods. The conditions of Theorem 1 are fulfilled for the first time after three uniform refinement steps. The fourth and all further refinement levels were reached by using adaptive refinement according to the error indicator based on (11). The fully adaptive scheme, which refines according to the residuals in the optimality system, obtains verified error bounds as of level 7. After the verified adaptive methods actually start adaptive refinement, they quickly reach the same ratio of error bound versus number of degrees of freedom as the full adaptive method. That means, the early (unverified) adaptive refinements of the full adaptive methods

does not seem to give this method an advantage over the verified method. The same observation also applies to the residual error versus number of degrees of freedom ratio, as can be seen in Figure 4.

Table 1. Error bound estimates

fully uniform			*verified adaptive*			*fully adaptive*		
L	$\#\,dof$	ϵ_{error}	L	$\#\,dof$	ϵ_{error}	L	$\#\,dof$	ϵ_{error}
0	441	–	1	1681	–	5	6177	–
1	1681	–	2	6561	–	6	10341	–
2	6561	–	3	25921	$7.6226 \cdot 10^{-3}$	7	18427	$9.0724 \cdot 10^{-3}$
3	25921	$7.6226 \cdot 10^{-3}$	4	31491	$4.8745 \cdot 10^{-3}$	8	27155	$5.6376 \cdot 10^{-3}$
4	103041	$1.9183 \cdot 10^{-3}$	5	55061	$2.8728 \cdot 10^{-3}$	9	46979	$3.3167 \cdot 10^{-3}$

Table 2. Residual error bound estimates

fully uniform			*verified adaptive*			*fully adaptive*		
L	$\#dof$	$\epsilon_{\text{residual}}$	L	$\#dof$	$\epsilon_{\text{residual}}$	L	$\#dof$	$\epsilon_{\text{residual}}$
1	1681	$5.4976 \cdot 10^{-4}$	1	1681	$5.4976 \cdot 10^{-4}$	2	1311	$6.6336 \cdot 10^{-4}$
2	6561	$1.4181 \cdot 10^{-4}$	2	6561	$1.4181 \cdot 10^{-4}$	5	6193	$1.2537 \cdot 10^{-4}$
3	25921	$3.6666 \cdot 10^{-5}$	3	25921	$3.6669 \cdot 10^{-5}$	7	18427	$4.3361 \cdot 10^{-5}$
			4	31491	$2.3746 \cdot 10^{-5}$	8	27155	$2.7370 \cdot 10^{-5}$
4	103041	$9.4714 \cdot 10^{-6}$	5	55061	$1.4121 \cdot 10^{-5}$	9	46979	$1.6271 \cdot 10^{-5}$

References

1. Akindeinde, S., Wachsmuth, D.: A-posteriori verification of optimality conditions for control problems with finite-dimensional control space. Num. Funct. Anal. Opt. 33(5), 473–523 (2012)
2. Becker, R.: Estimating the control error in discretized pde-constrained optimization. Journal of Numerical Mathematics 14, 163–185 (2006)
3. Casas, E., Tröltzsch, F., Unger, A.: Second-order sufficient optimality conditions for a nonlinear elliptic control problem. J. for Analysis and its Applications 15, 687–707 (1996)
4. El Alaoui, L., Ern, A., Vohralík, M.: Guaranteed and robust a posteriori error estimates and balancing discretization and linearization errors for monotone nonlinear problems. Comput. Methods Appl. Mech. Engrg. 200(37-40), 2782–2795 (2011)
5. Griesbaum, A., Kaltenbacher, B., Vexler, B.: Efficient computation of the Tikhonov regularization parameter by goal-oriented adaptive discretization. Inverse Problems 24(2), 025025, 20 (2008)
6. Kunisch, K., Liu, W., Chang, Y., Yan, N., Li, R.: Adaptive finite element approximation for a class of parameter estimation problems. J. Comput. Math. 28(5), 645–675 (2010)

7. Rösch, A., Wachsmuth, D.: Numerical verification of optimality conditions. SIAM J. Control Optim. 47(5), 2557–2581 (2008)
8. Rösch, A., Wachsmuth, D.: How to check numerically the sufficient optimality conditions for infinite-dimensional optimization problems. In: Control of Coupled Partial Differential Equations. Internat. Ser. Numer. Math., vol. 158, pp. 297–317. Birkhäuser, Basel (2009)
9. Vexler, B., Wollner, W.: Adaptive finite elements for elliptic optimization problems with control constraints. SIAM J. Control Optim. 47(1), 509–534 (2008)
10. Vohralík, M.: A posteriori error estimates for lowest-order mixed finite element discretizations of convection-diffusion-reaction equations. SIAM J. Numer. Anal. 45(4), 1570–1599 (2007)
11. Vohralík, M.: Unified primal formulation-based a priori and a posteriori error analysis of mixed finite element methods. Math. Comp. 79(272), 2001–2032 (2010)
12. Zeidler, E.: Nonlinear functional analysis and its applications. II/B. Springer, New York (1990), Nonlinear monotone operators
13. Ziems, J.C., Ulbrich, S.: Adaptive multilevel inexact SQP methods for PDE-constrained optimization. SIAM J. Optim. 21(1), 1–40 (2011)

Dynamic Contact Problem for Viscoelastic von Kármán-Donnell Shells

Igor Bock[1] and Jiří Jarušek[2]

[1] Institute of Computer Science and Mathematics FEI,
Slovak University of Technology, 812 19 Bratislava 1, Slovakia
igor.bock@stuba.sk
[2] Institute of Mathematics, Academy of Sciences of the Czech Republic,
Žitná 25, 115 67 Praha 1, Czech Republic
jarusek@math.cas.cz

Abstract. We deal with initial-boundary value problems describing vertical vibrations of viscoelastic von Kármán-Donnell shells with a rigid inner obstacle. The short memory (Kelvin-Voigt) material is considered. A weak formulation of the problem is in the form of the hyperbolic variational inequality. We solve the problem using the penalization method.

Keywords: Von Kármán-Donnell shell, unilateral dynamic contact, viscoelasticity, solvability, penalty approximation.

1 Introduction

Contact problems represent an important but complex topic of applied mathematics. Its complexity profounds if the dynamic character of the problem is respected. For elastic problems there is only a very limited amount of results available (cf. [3] and there cited literature). Viscosity makes possible to prove the existence of solutions for a broader set of problems for membranes, bodies as well as for linear models of plates. The presented results extend the research made in [2], where the problem for a viscoelastic short memory von Kármán plate in a dynamic contact with a rigid obstacle was considered. Our results also extend the research made for the quasistatic contact problems for viscoelastic shells (cf. [1]). A thin isotropic shallow shell occupies the domain

$$G = \{(x, z) \in R^3 : \ x = (x_1, x_2) \in \Omega, \ |z - \mathcal{Z}| < h/2\},$$

where $h > 0$ is the thickness of the shell, $\Omega \subset \mathbb{R}^2$ is a bounded simply connected domain in $\mathbb{R}$ with a sufficiently smooth boundary Γ. We set $I \equiv (0, T)$ a bounded time interval, $Q = I \times \Omega$, $S = I \times \Gamma$. The unit outer normal vector is denoted by $\mathbf{n} = (n_1, n_2)$, $\tau = (-n_2, n_1)$ is the unit tangent vector. The displacement is denoted by $\mathbf{u} \equiv (u_i)$. The strain tensor is defined as

$$\varepsilon_{ij}(\mathbf{u}) = \frac{1}{2}(\partial_i u_j + \partial_j u_i + \partial_i u_3 \partial_j u_3) - k_{ij} u_3 - x_3 \partial_{ij} u_3, \ i, j = 1, 2$$

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 70–75, 2013.

with $k_{12} = k_{21} = 0$ and the curvatures $k_{ii} > 0,\ i = 1, 2$.

Further, we set

$$[u, v] \equiv \partial_{11} u \partial_{22} v + \partial_{22} u \partial_{11} v - 2\partial_{12} u \partial_{12} v.$$

In the sequel, we denote by $W_p^k(M)$, $k \geq 0$, $p \in [1, \infty]$ the Sobolev spaces defined on a domain or an appropriate manifold M. By $\mathring{W}_p^k(M)$ the spaces with zero traces are denoted. If $p = 2$ we use the notation $H^k(M)$, $\mathring{H}^k(M)$. The duals to $\mathring{H}^k(M)$ are denoted by $H^{-k}(M)$. For the anisotropic spaces $W_p^k(M)$, $k = (k_1, k_2) \in \mathbb{R}_+^2$, k_1 is related with the time variable while k_2 with the space variables. We shall use also the Bochner-type spaces $W_p^k(I; X)$ for a time interval I and a Banach space X. Let us remark that for $k \in (0, 1)$ their norm is defined by the relation

$$\|w\|^p_{W_p^k(I;X)} \equiv \int_I \|w(t)\|^p_X dt + \int_I \int_I \frac{\|w(t) - w(s)\|^p_X}{|s - t|^{1+kp}} ds\, dt.$$

By $C(M)$ we denote the spaces of continuous functions on a (possibly relatively) compact manifold M. They are equipped with the max-norm. Analogously the spaces $C(M; X)$, are introduced for a Banach space X. The following generalization of the Aubin's compactness lemma verified in [4] Theorem 3.1 will be essentially used:

Lemma 1. *Let $B_0 \hookrightarrow\hookrightarrow B \hookrightarrow B_1$ be Banach spaces, the first reflexive and separable. Let $1 < p < \infty,\ 1 \leq r < \infty$. Then*

$$W \equiv \{v;\ v \in L_p(I; B_0),\ \dot{v} \in L_r(I, B_1)\} \hookrightarrow\hookrightarrow L_p(I; B).$$

2 Short Memory Material

2.1 Problem Formulation

Employing the Einstein summation, the constitutional law has the form

$$\sigma_{ij}(\mathbf{u}) = \frac{E_1}{1-\mu^2}\partial_t\big((1-\mu)\varepsilon_{ij}(\mathbf{u}) + \mu\delta_{ij}\varepsilon_{kk}(\mathbf{u})\big) + \frac{E_0}{1-\mu^2}\big((1-\mu)\varepsilon_{ij}(\mathbf{u}) + \mu\delta_{ij}\varepsilon_{kk}(\mathbf{u})\big).$$

The constants $E_0,\ E_1 > 0$ are the Young modulus of elasticity and the modulus of viscosity, respectively. We shall use the abbreviation $b = h^2/(12\varrho(1-\mu^2))$, where $h > 0$ is the shell thickness and ϱ is the density of the material. We involve the rotation inertia expressed by the term $a\Delta\ddot{u}$ in the first equation of the considered system with $a = \frac{h^2}{12}$. It will play the crucial role in the deriving a strong convergence of the sequence of velocities $\{\dot{u}_m\}$ in the appropriate space. We assume the shell clamped on the boundary. We generalize the dynamic elastic model due to the von Kármán-Donnell theory mentioned in [6]. The classical

formulation for the deflection $u_3 \equiv u$ and the Airy stress function v is then the initial-value problem

$$\left.\begin{aligned} &\ddot{u} + a\Delta\ddot{u} + b(E_1\Delta^2\dot{u} + E_0\Delta^2 u) - [u,v] - \Delta_k * v = f + g,\\ &u - \Psi \geq 0,\ g \geq 0,\ (u-\Psi)g = 0,\\ &\Delta^2 v + E_1\partial_t(\tfrac{1}{2}[u,u] + k_{11}\partial_{22}u + k_{22}\partial_{11}u)\\ &+E_0(\tfrac{1}{2}[u,u] + \Delta_k u) = 0 \end{aligned}\right\} \text{ on } Q, \tag{1}$$

$$u = \partial_n u = v = \partial_n v = 0 \text{ on } S, \tag{2}$$

$$u(0,\cdot) = u_0,\ \dot{u}(0,\cdot) = u_1 \text{ on } \Omega. \tag{3}$$

The obstacle function $\Psi \in L_\infty(\Omega)$ is fulfilling $0 < U_0 \leq u_0 - \Psi$ in Ω and

$$\Delta_k u \equiv \partial_{11}(k_{22}u) + \partial_{22}(k_{11}u), \tag{4}$$

$$\Delta_k^* v \equiv k_{22}\partial_{11}v + k_{11}\partial_{22}v. \tag{5}$$

We define the operators $L : H^2(\Omega) \to \mathring{H}^2(\Omega)$, $\Phi : H^2(\Omega) \times H^2(\Omega) \to \mathring{H}^2(\Omega)$ by uniquely solved equations

$$(\Delta Lu, \Delta w) \equiv (\Delta_k u, w)\, \forall w \in \mathring{H}^2(\Omega), \tag{6}$$

$$(\Delta\Phi(u,v), \Delta w) \equiv ([u,v], w)\, \forall w \in \mathring{H}^2(\Omega). \tag{7}$$

with the inner product $(\cdot,\cdot)$ in the space $L_2(\Omega)$. The operator L is linear and compact. The bilinear operator Φ is symmetric and compact. Moreover due to Lemma 1 from [5] $\Phi : H^2(\Omega)^2 \to W_p^2(\Omega)$, $2 < p < \infty$ and

$$\|\Phi(u,v)\|_{W_p^2(\Omega)} \leq c\|u\|_{H^2(\Omega)}\|v\|_{W_p^1(\Omega)}\ \forall u \in H^2(\Omega),\ v \in W_p^1(\Omega). \tag{8}$$

We have also $L : H^2(\Omega) \mapsto W_p^2(\Omega)$, $2 < p < \infty$ and

$$\|Lu\|_{W_p^2(\Omega)} \leq c\|u\|_{H^2(\Omega)}\ \forall u \in H^2(\Omega). \tag{9}$$

For $u, y \in L_2(I; H^2(\Omega))$ we define the bilinear form A by

$$A(u,y) := b\big(\partial_{kk}u\partial_{kk}y + \mu(\partial_{11}u\partial_{22}y + \partial_{22}u\partial_{11}y) + 2(1-\mu)\partial_{12}u\partial_{12}y\big).$$

We introduce shifted cone $\mathcal{K}$ by

$$\mathcal{K} := \{y \in H^{1,2}(Q);\ \dot{y} \in L_2(I, \mathring{H}^1(\Omega);\ y \geq \Psi\}. \tag{10}$$

Then the variational formulation of the problem (1-3) has the form of

Problem $\mathcal{P}$. *Find $u \in \mathcal{K}$ such that $\dot{u} \in L_2(I; \mathring{H}^2(\Omega))$ and*

$$\begin{aligned} &\textstyle\int_Q (E_1 A(\dot{u}, y-u) + E_0 A(u, y-u))\, dx\, dt\\ &+ \textstyle\int_Q [u, E_1\partial_t(\frac{1}{2}\Phi(u,u) + Lu) + E_0(\frac{1}{2}\Phi(u,u) + Lu)](y-u)\, dx\, dt\\ &+ \textstyle\int_Q \Delta_k\left(E_1\partial_t(\frac{1}{2}\Phi(u,u) + Lu) + E_0(\frac{1}{2}\Phi(u,u) + Lu)\right)(y-u)\, dx\, dt\\ &- \textstyle\int_Q (a\nabla\dot{u}\cdot\nabla(\dot{y}-\dot{u}) + \dot{u}(\dot{y}-\dot{u}))\, dx\, dt\\ &+ \textstyle\int_\Omega (a\nabla\dot{u}\cdot\nabla(y-u) + \dot{u}(y-u))(T,\cdot)\, dx\\ &\geq \textstyle\int_\Omega (a\nabla u_1\cdot\nabla(y(0,\cdot)-u_0) + u_1(y(0,\cdot)-u_0))\, dx\\ &+ \textstyle\int_Q f(y_1-u)\, dx\, dt\ \forall y \in \mathcal{K}. \end{aligned} \tag{11}$$

2.2 The Penalization

For any $\eta > 0$ we define the *penalized problem*
Problem $\mathcal{P}_\eta$. *Find* $u \in H^{1,2}(Q)$ *such that* $\dot{u} \in L_2(I; \mathring{H}^2(\Omega))$, $\ddot{u} \in L_2(I; \mathring{H}^1(\Omega))$,

$$\begin{aligned} &\int_Q \left(\ddot{u}z + a\nabla\ddot{u}\cdot\nabla z + E_1A(\dot{u},z) + E_0A(u,z)\right)dx\,dt \\ &+\int_Q [u, E_1\partial_t(\tfrac{1}{2}\Phi(u,u)+Lu) + E_0(\tfrac{1}{2}\Phi(u,u)+Lu)]z\,dx\,dt \\ &+\int_Q \Delta_k\left(E_1\partial_t(\tfrac{1}{2}\Phi(u,u)+Lu) + E_0(\tfrac{1}{2}\Phi(u,u)+Lu)\right)z\,dx\,dt \\ &=\int_Q (f+\eta^{-1}(u-\Psi)^-)z\,dx\,dt\ \forall z\in L_2(I;H^2(\Omega)) \end{aligned} \tag{12}$$

and the conditions (3) remain valid.

Lemma 2. *Let* $f \in L_2(Q)$, $u_0 \in \mathring{H}^2(\Omega)$, *and* $u_1 \in \mathring{H}^1(\Omega)$. *Then there exists a solution* u *of the problem* $\mathcal{P}_\eta$.

Proof. Let us denote by $\{w_i \in \mathring{H}^2(\Omega);\ i = 1, 2, ...\}$ a basis of $\mathring{H}^2(\Omega)$ orthonormal in $H^1(\Omega)$ with respect to the inner product

$$(u,v)_a = \int_\Omega (uv + a\nabla u\cdot\nabla v)\,dx,\ u,v\in H^1(\Omega).$$

We construct the Galerkin approximation u_m of a solution in a form

$$u_m(t) = \sum_{i=1}^m \alpha_i(t)w_i,\ \alpha_i(t)\in\mathbb{R},\ i=1,...,m,\ m\in N, \tag{13}$$

$$\begin{aligned} &(\ddot{u}_m(t), w_i)_a + \int_\Omega \left(E_1A(\dot{u}_m(t),w_i) + E_0A(u_m(t),w_i)\right)dx + \\ &\int_\Omega \Delta\big(E_1\partial_t(\tfrac{1}{2}\Phi(u_m,u_m)+Lu_m) + E_0(\tfrac{1}{2}\Phi(u_m,u_m)+Lu_m)\big) \\ &\times\Delta(\Phi(u_m,w_i)+Lw_i)\,dx \\ &=\int_\Omega \left(f(t)+\eta^{-1}(u_m(t)-\Psi)^-\right)w_i\,dx,\ i=1,...,m, \end{aligned} \tag{14}$$

$$u_m(0) = u_{0m},\ \dot{u}_m(0) = u_{1m},\ u_{0m}\to u_0 \text{ in } \mathring{H}^2(\Omega),\ u_{1m}\to u_1 \text{ in } \mathring{H}^1(\Omega). \tag{15}$$

After multiplying the equation (14) by $\dot{\alpha}_i(t)$, summing up with respect to i , taking in mind the definitions of the operators Φ, L and integrating we obtain the *a priori* estimates not depending on m:

$$\begin{aligned} &\|\dot{u}_m\|^2_{L_2(I;\mathring{H}^2(\Omega))} + \|\dot{u}_m\|^2_{L_\infty(I;\mathring{H}^1(\Omega))} + \|u_m\|^2_{L_\infty(I;\mathring{H}^2(\Omega))} \\ &+\|\partial_t\Phi(u_m,u_m)\|^2_{L_2(I;\mathring{H}^2(\Omega))} + \|\partial_t Lu_m\|^2_{L_2(I;\mathring{H}^2(\Omega))} \\ &+\eta^{-1}\|(u_m-\Psi)^-\|_{L_\infty(I;L_2(\Omega))} \le c \equiv c(f,u_0,u_1). \end{aligned} \tag{16}$$

Moreover the estimates (8), (9) imply

$$\|\partial_t\Phi(u_m,u_m)\|_{L_2(I;W_p^2(\Omega))} + \|\partial_t Lu_m\|_{L_2(I;W_p^2(\Omega))} \le c_p\ \forall\, p>2. \tag{17}$$

After multiplying the equation (14) by $\ddot{\alpha}_i(t)$, summing up and integrating we obtain the estimate of $\ddot{u}_m$

$$\|\ddot{u}_m\|_{L_2(I;H^1(\Omega))} \le c_\eta,\ m\in\mathbb{N}. \tag{18}$$

Applying the estimates (16)-(18), the compact imbedding theorem and the interpolation, we obtain for any $p \in [1, \infty)$, a subsequence of $\{u_m\}$ (denoted again by $\{u_m\}$), a function u and the convergences

$$\begin{aligned}
&\ddot{u}_m \rightharpoonup \ddot{u} \text{ in } L_2(I; H^1(\Omega)),\\
&\dot{u}_m \rightharpoonup^* \dot{u} \text{ in } L_\infty(I; \mathring{H}^1(\Omega)),\\
&\dot{u}_m \rightharpoonup \dot{u} \text{ in } L_2(I; \mathring{H}^2(\Omega)),\\
&\dot{u}_m \to \dot{u} \text{ in } L_p(I; \mathring{H}^(\Omega)) \cap L_\infty(I; H^{2-\varepsilon}(\Omega)) \ \forall \varepsilon > 0,\\
&u_m \to u \text{ in } C(\bar{I}; W^1_p(\Omega),\\
&\partial_t\big(\tfrac{1}{2}\Phi(u_m, u_m) + Lu_m\big) \to \partial_t\big(\tfrac{1}{2}\Phi(u, u) + Lu\big) \text{ in } L_2(I; W^2_p(\Omega))
\end{aligned} \tag{19}$$

implying that a function u fulfils the identity (12). The initial conditions (3) follow due to (15) and the proof of the existence of a solution is complete.

2.3 Solving the Original Problem

We verify the existence theorem

Theorem 1. *Let $f \in L_2(Q)$, $u_i \in \mathring{H}^2(\Omega)$, $i = 0,\ 1$, $0 < U_0 \le u_0 - \Psi$. Then there exists a solution of the Problem $\mathcal{P}$.*

Proof. We perform the limit process for $\eta \to 0$. We write u_η for the solution of the problem $\mathcal{P}_{1,\eta}$. The *a priori* estimates (16) imply the estimates

$$\begin{aligned}
&\|\dot{u}_\eta\|^2_{L_2(I;\mathring{H}^2(\Omega))} + \|\dot{u}_\eta\|^2_{L_\infty(I;\mathring{H}^1(\Omega))} + \|u_\eta\|^2_{L_\infty(I;\mathring{H}^2(\Omega))}\\
&+\|\partial_t\Phi(u_\eta, u_\eta)\|^2_{L_2(I;W^2_p(\Omega))} + \|\partial_t L u_\eta\|^2_{L_2(I;W^2_p(\Omega))}\\
&+\eta^{-1}\|(u_\eta - \Psi)^-\|_{L_\infty(I;L_2(\Omega))} \le c_p,\ p > 2.
\end{aligned} \tag{20}$$

To get the crucial estimate for the penalty, we put $z = u_0 - u_\eta(t, \cdot)$ in (12) and obtain the estimate

$$\begin{aligned}
0 \le U_0 \textstyle\int_Q \eta^{-1}(u_\eta - \Psi)^- dx\, dt &\le \textstyle\int_Q \|\eta^{-1}(u_\eta - \Psi)^-(u_0 - \Psi) dx\, dt\\
&\le \textstyle\int_Q \|\eta^{-1}(u_\eta - \Psi)^-(u_0 - u_\eta) dx\, dt\\
&= \textstyle\int_Q \big(\dot{u}_\eta^2 + a|\nabla \dot{u}_\eta|^2 + A(\,(E_1\partial_t u_\eta + E_0 u_\eta), u_0 - u_\eta)\\
&+E_1\partial_t(\Delta(Lu_\eta + \tfrac{1}{2}\Phi(u_\eta, u_\eta)))\Delta(L(u_0 - u_\eta) + \Phi(u_\eta, u_0 - u_\eta))\\
&+E_0\Delta(Lu_\eta + \tfrac{1}{2}\Phi(u_\eta, u_\eta))\Delta(L(u_0 - u_\eta) + \Phi(u_\eta, u_0 - u_\eta))\big)\, dx\, dt\\
&- \textstyle\int_Q f(u_0 - u_\eta)\, dx\, dt + \int_\Omega \big(\big(\dot{u}_\eta(u_0 - u_\eta) + a\nabla\dot{u}_\eta \cdot \nabla(u_0 - u_\eta)\big)(T, \cdot)\big)\, dx.
\end{aligned}$$

Applying the *a priori* estimates (20) we obtain

$$\|\eta^{-1}u_\eta^-\|_{L_1(Q)} \le c(f, u_0, u_1, \Psi). \tag{21}$$

With respect to Dirichlet conditions we obtain from (12) and (21) the dual estimate

$$\| - a\Delta\ddot{u}_\eta + \ddot{u}_\eta\|_{L_1(I;H^{-2}(\Omega))} \le c. \tag{22}$$

We take the sequence $\{u_k\} \equiv \{u_{\eta_k}\},\ \eta_k \to 0+$.

After applying the Lemma 1 with the spaces

$$B_0 = L_2(\Omega),\ B = H^{-1}(\Omega),\ B_1 = H^{-2}(\Omega)$$

we obtain the relative compactness of the sequence $\{-a\Delta\dot{u}_k + \dot{u}_k\}$ in $L_2(I; H^{-1}(\Omega)$ and with the help of the test function $\dot{u}_k - \dot{u}$ the crucial strong convergence

$$\dot{u}_k \to \dot{u} \text{ in } L_2(I; \mathring{H}^1(\Omega)). \tag{23}$$

Simultaneously we have the convergences

$$\begin{aligned} &\dot{u}_k \rightharpoonup \dot{u} \text{ in } L_2(I; \mathring{H}^2(\Omega)),\\ &\dot{u}_k \to \dot{u} \text{ in } L_2(I; W_p^1(\Omega)),\\ &\tfrac{1}{2}\partial_t \Phi(u_k, u_k) + \partial_t L u_k \rightharpoonup \tfrac{1}{2}\partial_t \Phi(u,u) + \partial_t L u \text{ in } L_2(I; W_p^2(\Omega)). \end{aligned} \tag{24}$$

It can be verified after inserting the test function $z = y - u_k$ in (12) for $y \in \mathcal{K}$, performing the integration by parts in the terms containing $\ddot{u}$, applying the convergences (23), (24), using the definitions of the operators L, Φ in (6), (7) and the weak lower semicontinuity that the limit function u is a solution of the original problem $\mathcal{P}$.

Remark 1. The existence Theorem 1 can be after some modification verified also for another types of boundary conditions.

Acknowledgments. The work presented here was supported by the Czech Academy of Sciences under grant P201/12/0671 and under the Institutional research plan RVD 67985840, by the Czech Ministery of Education under grant MEB 0810045, by the APVV Agency of Slovak Republic under grant 0011-09 and by Ministry of Education of the Slovak Republic under VEGA grant 1/0426/12.

References

1. Bock, I.: On a pseudoparabolic system for a viscoelastic shallow shell. PAMM Proc. Appl. Math. Mech. 6, 621–622 (2006)
2. Bock, I., Jarušek, J.: Unilateral dynamic contact of viscoelastic von Kármán plates. Advances in Math. Sci. and Appl. 16, 175–187 (2006)
3. Eck, C., Jarušek, J., Krbec, M.: Unilateral Contact Problems in Mechanics. Variational Methods and Existence Theorems. Monographs & Textbooks in Pure & Appl. Math., vol. 270. Chapman & Hall/CRC (Taylor & Francis Group), Boca Raton, London, New York, Singapore (2005)
4. Jarušek, J., Málek, J., Nečas, J., Šverák, V.: Variational inequality for a viscous drum vibrating in the presence of an obstacle. Rend. Mat., Ser. VII 12, 943–958 (1992)
5. Koch, H., Stachel, A.: Global existence of classical solutions to the dynamical von Kármán equations. Math. Methods in Applied Sciences 16, 581–586 (1993)
6. Vol̆mir, A.G.: Gibkije plastinky i oboločky. Gosizdat, Moskva (1956) (in Russian)

On Existence, Uniqueness, and Convergence, of Optimal Control Problems Governed by Parabolic Variational Inequalities

Mahdi Boukrouche[1,⋆] and Domingo A. Tarzia[2,⋆⋆]

[1] Lyon University UJM F-42023, CNRS UMR 5208, ICJ, France
[2] CONICET and Austral University, Rosario, Argentina

Abstract. I) We consider a system governed by a free boundary problem with Tresca condition on a part of the boundary of a material domain with a source term g through a parabolic variational inequality of the second kind. We prove the existence and uniqueness results to a family of distributed optimal control problems over g for each parameter $h > 0$, associated to the Newton law (Robin boundary condition), and of another distributed optimal control problem associated to a Dirichlet boundary condition. We generalize for parabolic variational inequalities of the second kind the Mignot's inequality obtained for elliptic variational inequalities (Mignot, J. Funct. Anal., 22 (1976), 130-185), and we obtain the strictly convexity of a quadratic cost functional through the regularization method for the non-differentiable term in the parabolic variational inequality for each parameter h. We also prove, when $h \to +\infty$, the strong convergence of the optimal controls and states associated to this family of optimal control problems with the Newton law to that of the optimal control problem associated to a Dirichlet boundary condition.

II) Moreover, if we consider a parabolic obstacle problem as a system governed by a parabolic variational inequalities of the first kind then we can also obtain the same results of Part I for the existence, uniqueness and convergence for the corresponding distributed optimal control problems.

III) If we consider, in the problem given in Part I, a flux on a part of the boundary of a material domain as a control variable (Neumann boundary optimal control problem) for a system governed by a parabolic variational inequality of second kind then we can also obtain the existence and uniqueness results for Neumann boundary optimal control problems for each parameter $h > 0$, but in this case the convergence when $h \to +\infty$ is still an open problem.

Keywords: Parabolic variational inequalities, convex combination of solutions, regularization method, optimal control problems, strict convexity of cost functional.

⋆ Lyon University, UJM F-42023, CNRS UMR 5208, Institut Camille Jordan, 23 Paul Michelon, 42023, Saint-Etienne, France.

⋆⋆ CONICET and Austral University, Mathematics Department, Paraguay 1950, S2000FZF Rosario, Argentina.

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 76–84, 2013.

1 Introduction

The goal of this paper is to show the existence and uniqueness results to a family of distributed (see Sections 2 and 3) or Neumann boundary (see Section 4) optimal control problems for each parameter $h > 0$, associated to the Newton law (Robin boundary condition on a part of the boundary of the material domain), and of another distributed optimal control problem associated to a Dirichlet boundary condition. The system of these optimal control problems are governed by free boundary problems (with Tresca boundary condition (see Sections 2 and 4) or of an obstacle type problem (see Section 3) through a parabolic variational inequalities of the first (see Section 3) or second (see Sections 2 and 4) kind [2], [6]. An optimal control problem for elliptic variational inequality of the second kind is given in [9].

In order to prove the existence and uniqueness results we generalize for parabolic variational inequalities of the second kind the Mignot's inequality obtained for elliptic variational inequalities [18], and then we obtain the strictly convexity of a quadratic cost functional through the regularization method for the non-differentiable term for each parameter $h > 0$.

We also prove, when $h \to +\infty$, the strong convergence of the optimal controls and states associated to this family of optimal control problems with the Newton law to that of the optimal control problem associated to a Dirichlet boundary condition.

We obtain these convergence without using the adjoint state which is a great advantage with respect to the proof given previously for optimal control problems governed by elliptic and parabolic variational equalities [3], [11], [12], [17].

These convergence when $h \to +\infty$ are valid for the optimal control problems given in Sections 2 and 3, and it is still an open problem for the Neumann boundary optimal control problem given in Section 4.

2 Distributed Optimal Control Problems Governed by Parabolic Variational Inequality of Second Kind

Let Ω a bounded open set in $\mathbb{R}^N$ with smooth boundary $\partial\Omega = \Gamma_1 \cup \Gamma_2$ such that $\Gamma_1 \cap \Gamma_2 = \emptyset$, and $meas(\Gamma_1) > 0$. We set $V = H^1(\Omega)$, $V_0 = \{v \in V : v_{|_{\Gamma_1}} = 0\}$, $H = L^2(\Omega)$, $\mathcal{H} = L^2(0,T;H)$, $\mathcal{V} = L^2(0,T;V)$, and the closed convex set $K_b = \{v \in V : \quad v_{|_{\Gamma_1}} = b\}$. Let given

$$b \in L^2(0,T;H^{1/2}(\Gamma_1)), \quad b > 0, \quad g \in \mathcal{H}, \quad g \geq 0,$$
$$q \in L^2((0,T) \times \Gamma_2), \quad q > 0, \quad u_b \in K_b. \tag{1}$$

We consider the following variational problems [6]

Problem 1. Let given g, q, b and u_b as in (1). Find $u = u_g \in \mathcal{C}(0,T,H) \cap L^2(0,T;K_b)$ with $\dot{u} \in \mathcal{H}$, such that $u(0) = u_b$, and solution of the parabolic variational inequality of second kind:

$$< \dot{u}, v - u > + a(u, v - u) + \Phi(v) - \Phi(u) \geq (g, v - u), \quad \forall v \in K_b, \quad t \in (0,T).$$

Problem 2. Let given g, q, b and u_b as in (1). For all $h > 0$, find $u = u_{hg}$ in $\mathcal{C}(0,T,H) \cap \mathcal{V}$ with $\dot{u} \in \mathcal{H}$, such that $u(0) = u_b$, and solution of the parabolic variational inequality of second kind

$$< \dot{u}, v-u > + a_h(u, v-u) + \Phi(v) - \Phi(u) \geq (g, v-u) \\ + h \int_{\Gamma_1} b(v-u) ds, \; \forall v \in V, \quad t \in (0,T).$$

Where $\dot{u} = u_t$, $<,>$ denotes the duality brackets between V' and V, a is a symmetric, continuous and coercive bilinear form over V_0, and Φ is given by

$$\Phi(v) = \int_{\Gamma_2} q|v| ds, \tag{2}$$

and

$$a(u,v) = \int_\Omega \nabla u \nabla v dx, \quad a_h(u,v) = a(u,v) + h \int_{\Gamma_1} uv ds, \quad (g,v) = \int_\Omega g v dx.$$

Moreover from [15], [20], [21] we have that:

$$\exists \lambda_1 > 0 \quad \text{such that } \lambda_h \|v\|_V^2 \leq a_h(v,v) \quad \forall v \in V, \text{ with } \lambda_h = \lambda_1 \min\{1\,,\, h\}$$

that is, a_h is also a bilinear continuous, symmetric and coercive form on V.

We remark that on $\Gamma_1 \times (0,T)$, *Problem* 1 is with the Dirichlet condition $u_{|_{\Gamma_1}} = b$, while *Problem* 2 is with the Robin's condition $-\nabla u \cdot n = h(u-b)$, where n is the exterior unit vector normal to Γ. The functional Φ comes from the Tresca condition on Γ_2 [1], [4].

The existence and uniqueness of the solution to each of the above *Problem* 1 and *Problem* 2, is well known see for example [7], [8], [10]. Therefore, it allows us to consider $g \mapsto u_g$ as a function from $\mathcal{H}$ to $\mathcal{C}(0,T,H) \cap \mathcal{V}$.

Let $M > 0$ be a constant and $\mathcal{H}_+ = \{g \in \mathcal{H} : g \geq 0\}$. We consider the following distributed optimal control problems defined by:

$$\text{Find } g_{op} \in \mathcal{H}_+ \quad \text{such that} \quad J(g_{op}) = \min_{g \in \mathcal{H}_+} J(g), \tag{3}$$

$$\text{Find } g_{op_h} \in \mathcal{H}_+ \quad \text{such that} \quad J(g_{op_h}) = \min_{g \in \mathcal{H}_+} J_h(g), \tag{4}$$

where the cost functional $J : \mathcal{H} \to \mathbb{R}$ and $J_h : \mathcal{H} \to \mathbb{R}$ such that [16] (see also [13], [14], [22])

$$J(g) = \frac{1}{2}\|u_g\|_{\mathcal{H}}^2 + \frac{M}{2}\|g\|_{\mathcal{H}}^2, \quad \text{and} \quad J_h(g) = \frac{1}{2}\|u_{hg}\|_{\mathcal{H}}^2 + \frac{M}{2}\|g\|_{\mathcal{H}}^2, \tag{5}$$

being here u_g, u_{hg} the unique solutions of the parabolic variational *Problem* 1, and *Problem* 2 respectively, and corresponding to the control g in $\mathcal{H}$. In order to prove the strict convexity of the cost functional J and J_h, we generalize for

parabolic variational inequalities a main property [18] that : For any two control $g_i \in \mathcal{H}$, $i = 1$ or $i = 2$, we have

$$u_{\mu g_1+(1-\mu)g_2} \leq \mu u_{g_1} + (1-\mu)u_{g_2}, \quad \forall \mu \in [0,1],$$

$$u_{h(\mu g_1+(1-\mu)g_2)} \leq \mu u_{hg_1} + (1-\mu)u_{hg_2}, \quad \forall \mu \in [0,1],$$

by using a regularization method for the non-differentiable functional Φ (see [6]). Then we prove the following

Theorem 1. *[6] Let $u_{hg_{op_h}}$, g_{op_h} and $u_{g_{op}}$, g_{op} be the states and the optimal controls defined in Problem 1 and Problem 2 respectively. Then, we obtain the following asymptotic behavior:*

$$\lim_{h\to+\infty} \|u_{hg_{op_h}} - u_{g_{op}}\|_{\mathcal{V}} = 0, \tag{6}$$

$$\lim_{h\to+\infty} \|g_{op_h} - g_{op}\|_{\mathcal{H}} = 0. \tag{7}$$

3 Distributed Optimal Control Problems Governed by Parabolic Variational Inequality of First Kind

We will examine in this section, some distributed optimal control problems, for which the strong formulation can be linked to a free boundary problems of complementarity type (Obstacle problems [19]), given for example by the following conditions:

$$u \geq 0, \quad u(\dot{u} - \Delta u - g) = 0, \quad \dot{u} - \Delta u - g \geq 0 \quad \text{in } \Omega, \tag{8}$$

$$u = b \geq 0 \text{ on } \Gamma_1, \quad -\frac{\partial u}{\partial n} = f \text{ on } \Gamma_2, \quad \text{and} \quad u(0) = u_b \tag{9}$$

and

$$u \geq 0, \quad u(\dot{u} - \Delta u - g) = 0, \quad \dot{u} - \Delta u - g \geq 0 \quad \text{in } \Omega, \tag{10}$$

$$-\frac{\partial u}{\partial n} = h(u-b) \text{ on } \Gamma_1, \quad -\frac{\partial u}{\partial n} = f \text{ on } \Gamma_2, \quad \text{and} \quad u(0) = u_b \tag{11}$$

where Ω is a multidimentional regular domain whose boundary is $\partial\Omega = \Gamma_1 \cup \Gamma_2$ with $\Gamma_1 \cap \Gamma_2 = \emptyset$. Let consider the convex set K_b as in Section 2. It is classical that, for a given positive $b \in L^2(0,T;H^{\frac{1}{2}}(\Gamma_1))$, $f \in L^2(0,T;L^2(\Gamma_2))$, and $g \in \mathcal{H}$, the variational formulations of *Problems* (8)-(9) and (10)-(11) are respectively given by the following parabolic variational problems:

Problem 3. Let given g, b and u_b as in (1) and $f \in L^2(0,T;L^2(\Gamma_2))$. Find $u = u_g \in \mathcal{C}(0,T,H) \cap L^2(0,T;K_b)$ with $\dot{u} \in \mathcal{H}$, such that $u(0) = u_b$, and

$$< \dot{u}, v-u > + a(u, u-v) \geq (g, v-u) - \int_{\Gamma_2} f(v-u)ds, \quad \forall v \in K_b, \forall t \in (0,T).$$

Problem 4. Find $u = u_{hg} \in \mathcal{C}(0,T,H) \cap \mathcal{V}$ with $\dot{u} \in \mathcal{H}$, such that $u(0) = u_b$, and

$$< \dot{u}, v-u > + a_h(u, u-v) \geq (g, v-u) + h \int_{\Gamma_1} b(v-u)ds$$
$$- \int_{\Gamma_2} f(v-u)ds, \quad \forall v \in V, \quad \forall t \in (0,T).$$

where a and a_h are as in Section 2. Then the existence and uniqueness of the solution to *Problem* 3 and *Problem* 4, is also well known see for example [7], [8], [10]. Then it allows us to consider $g \mapsto u_g$ as a function from $\mathcal{H}$ to $\mathcal{C}(0,T,H) \cap \mathcal{V}$. Let $M > 0$ be a constant. We consider the same family of distributed optimal control problems (3)-(4) and we obtain the same results of the previous Theorem 1.

Theorem 2. *Let g, b, u_b as in (7) and $f \leq 0$ in $\Gamma_2 \times (0,T)$, we can obtain the same results as in Section 2, for the corresponding distributed optimal control problems (3)-(4) when $g \geq 0$ is the control variable.*

4 Neumann Boundary Optimal Control Problem Governed by Parabolic Variational Inequalities of Second Kind

We assume in this section that the boundary of a multidimensional regular domain Ω is decomposed in three parts $\partial\Omega = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3$ with $meas(\Gamma_1) > 0$ and $meas(\Gamma_3) > 0$.

We consider a Neumann boundary optimal control problem whose system is governed by a free boundary problem with Tresca conditions on a portion Γ_2 of the boundary, with the flux f on Γ_3 *as the control variable*, given by:

Problem 5.

$$\dot{u} - \Delta u = g \quad \text{in} \quad \Omega \times (0,T),$$

$$\left|\frac{\partial u}{\partial n}\right| < q \Rightarrow u = 0, \text{ on } \Gamma_2 \times (0,T),$$
$$\left|\frac{\partial u}{\partial n}\right| = q \Rightarrow \exists k > 0 : \quad u = -k\frac{\partial u}{\partial n}, \text{ on } \Gamma_2 \times (0,T),$$

$$u = b \quad \text{on} \quad \Gamma_1 \times (0,T),$$

$$-\frac{\partial u}{\partial n} = f \quad \text{on} \quad \Gamma_3 \times (0,T),$$

with the initial condition

$$u(0) = u_b \quad \text{on} \quad \Omega,$$

and the compatibility condition on $\Gamma_1 \times (0,T)$

$$u_b = b \quad \text{on} \quad \Gamma_1 \times (0,T),$$

where $q > 0$ is the Tresca friction coefficient on Γ_2 ([1], [4], [10]). We define the space $\mathcal{F} = L^2(0,T;L^2(\Gamma_3))$.

The variational formulation of *Problem* 5 leads to the following parabolic variational problem:

Problem 6. Let given g, q, b and u_b as in (1) and $f \in \mathcal{F}$, $f \leq 0$. Find $u = u_f$ in $\mathcal{C}(0,T,H) \cap L^2(0,T;K_b)$ with $\dot{u} \in \mathcal{H}$, such that $u(0) = u_b$, and for $t \in (0,T)$

$$< \dot{u}, v-u > + a(u,u-v) + \Phi(v) - \Phi(u) \geq (g, v-u) - \int_{\Gamma_3} f(v-u)ds, \ \forall v \in K_b.$$

where a and Φ are defined as in Section 2.

We consider also the following problem where we change, in *Problem* 5, only the Dirichlet condition on $\Gamma_1 \times (0,T)$ by the Newton law or a Robin boundary condition.

Problem 7.

$$\dot{u} - \Delta u = g \quad \text{in} \quad \Omega \times (0,T),$$

$$\left|\frac{\partial u}{\partial n}\right| < q \Rightarrow u = 0, \ \text{on} \ \Gamma_2 \times (0,T),$$

$$\left|\frac{\partial u}{\partial n}\right| = q \Rightarrow \exists k > 0 : \quad u = -k\frac{\partial u}{\partial n}, \ \text{on} \ \Gamma_2 \times (0,T),$$

$$-\frac{\partial u}{\partial n} = h(u-b) \quad \text{on} \quad \Gamma_1 \times (0,T),$$

$$-\frac{\partial u}{\partial n} = f \quad \text{on} \quad \Gamma_3 \times (0,T),$$

with the initial condition

$$u(0) = u_b \quad \text{on} \quad \Omega,$$

and the condition of compatibility on $\Gamma_1 \times (0,T)$

$$u_b = b \quad \text{on} \quad \Gamma_1 \times (0,T).$$

The variational formulation of the problem (7) leads to the the following parabolic variational problem

Problem 8. Let given g, q, b, u_b and f as in *Problem* 6. For all $h > 0$, find $u = u_{hf} \in \mathcal{C}(0,T,H) \cap \mathcal{V}$ with $\dot{u} \in \mathcal{H}$, such that $u(0) = u_b$, and for $t \in (0,T)$

$$< \dot{u}, v-u > + a_h(u,u-v) + \Phi(v) - \Phi(u) \geq (g, v-u) - \int_{\Gamma_3} f(v-u)ds + h\int_{\Gamma_1} b(v-u)ds, \quad \forall v \in V,$$

where a_h and Φ are defined as in Section 2.

4.1 Neumann Boundary Optimal Control Problems

Let $M > 0$ be a constant and we define the space $\mathcal{F}_- = \{f \in \mathcal{F} : f \leq 0\}$. We consider the new following Neumannn boundary optimal control problems defined by:

Problem 9. Find the optimal control $f_{op} \in \mathcal{F}_-$ such that

$$J(f_{op}) = \min_{f \in \mathcal{F}_-} J(f) \tag{12}$$

where the cost functional $J : \mathcal{F} \to \mathbb{R}_0^+$ is given by

$$J(f) = \frac{1}{2}\|u_f\|_{\mathcal{H}}^2 + \frac{M}{2}\|f\|_{\mathcal{F}}^2 \quad (M > 0) \tag{13}$$

and u_f is the unique solution of the *Problem* 6.

Problem 10. Find the optimal control $f_{op_h} \in \mathcal{F}_-$ such that

$$J(f_{op_h}) = \min_{f \in \mathcal{F}_-} J_h(f) \tag{14}$$

where the cost functional $J_h : \mathcal{F} \to \mathbb{R}_0^+$ is given by

$$J_h(f) = \frac{1}{2}\|u_{hf}\|_{\mathcal{H}}^2 + \frac{M}{2}\|f\|_{\mathcal{F}}^2 \quad (M > 0, \quad h > 0) \tag{15}$$

and u_{hf} is the unique solution of *Problem* 8.

Theorem 3. *Under the assumptions given in Problem 6, we have the following properties:*

a) *The cost functional J is strictly convex on $\mathcal{F}_-$,*
b) *There exists a unique optimal $f_{op} \in \mathcal{F}_-$ solution of the new Neumannn boundary optimal control Problem 9.*

Proof. We give some sketch of the proof.

i) We generalize for parabolic variational inequalities of the second kind the estimates obtained for convex combination of solutions for elliptic variational inequalities [5] that is, the estimate between

$$u_4(\mu) = u_{\mu f_1 + (1-\mu) f_2}, \text{ and } u_3(\mu) = \mu u_{f_1} + (1-\mu) u_{f_2},$$

for any two element f_1 and f_2 in $\mathcal{F}$.

ii) The main difficulty, to prove this result comes from the fact that the functional Φ is not differentiable. To overcome this difficulty, we use the regularization method and consider for $\varepsilon > 0$ the following approach of Φ defined by:

$$\Phi_\varepsilon(v) = \int_{\Gamma_2} q\sqrt{\varepsilon^2 + |v|^2} ds, \qquad \forall v \in V, \tag{16}$$

which is Gateaux differentiable, with

$$\langle \Phi'_{\varepsilon}(w)\,,\, v\rangle = \int_{\Gamma_2} \frac{qwv}{\sqrt{\varepsilon^2+|w|^2}} ds \qquad \forall (w,v)\in V^2.$$

We define u^{ε} as the unique solution of the corresponding parabolic variational inequality for all $\varepsilon > 0$. We obtain that for all $\mu \in [0,1]$ we have $u_4^{\varepsilon}(\mu) \leq u_3^{\varepsilon}(\mu)$ for all $\varepsilon > 0$.

iii) When $\varepsilon \to 0$ we have that:

$$u_i^{\varepsilon} \to u_i \text{ strongly in } \mathcal{V}\cap L^{\infty}(0,T;H) \text{ for } i=1,2,3,4, \tag{17}$$

for all $\mu \in [0,1]$ and therefore we get:

$$0 \leq u_4(\mu) \leq u_3(\mu) \quad in \quad \Omega\times[0,T], \quad \forall \mu\in[0,1]. \tag{18}$$

iv) For all $\mu \in]0,1[$, and for all f_1, f_2 in $\mathcal{F}$, and by using $f_3(\mu) = \mu f_1 + (1-\mu) f_2$ we obtain that:

$$\begin{aligned}\mu J(f_1) + (1-\mu)J(f_2) - J(f_3(\mu)) = \frac{1}{2}\left(\|u_3(\mu)\|^2_{\mathcal{H}} - \|u_4(\mu)\|^2_{\mathcal{H}}\right)\\ +\frac{1}{2}\mu(1-\mu)\|u_{f_1}-u_{f_2}\|^2_{\mathcal{H}} + \frac{M}{2}\mu(1-\mu)\|f_1-f_2\|^2_{\mathcal{F}}.\end{aligned} \tag{19}$$

Then J is strictly convex functional on $\mathcal{F}_-$ and therefore there exists a unique optimal $f_{op} \in \mathcal{F}_-$ solution of the new Neumannn boundary optimal control Problem 9.

Theorem 4. *Under the assumptions given in Problem 6, we have the following properties:*

a) *The cost functional J_h are strictly convex on $\mathcal{F}_-$, for all $h > 0$,*
b) *There exists a unique optimal $f_{op_h} \in \mathcal{F}_-$ solution of the new Neumannn boundary optimal control Problem 10, for all $h > 0$.*

Proof. We follow a similar method to the one developed in Theorem 3 for all $h > 0$.

4.2 Open Problem

The convergence of the new Neumann boundary optimal control *Problem* 10 to the new Neumann boundary optimal control *Problem* 9 when $h \to \infty$ is an open problem.

Acknowledgements. The autors would like to thank very much the unknown referee for helpful comments which allowed to improve the paper. This paper was partially sponsored by the Institut Camille Jordan ST-Etienne University for first author and the project PICTO Austral # 73 from ANPCyT and Grant AFOSR FA9550-10-1-0023 for the second author.

References

1. Amassad, A., Chenais, D., Fabre, C.: Optimal control of an elastic contact problem involving Tresca friction law. Nonlinear Analysis 48, 1107–1135 (2002)
2. Barbu, V.: Optimal control of variational inequalities. Research Notes in Mathematics, vol. 100. Pitman (Advanced Publishing Program), Boston (1984)
3. Ben Belgacem, F., El Fekih, H., Metoui, H.: Singular perturbation for the Dirichlet boundary control of elliptic problems. ESAIM: M2AN 37, 833–850 (2003)
4. Boukrouche, M., El Mir, R.: On a non-isothermal, non-Newtonian lubrication problem with Tresca law: Existence and the behavior of weak solutions. Nonlinear Analysis: Real World Applications 9(2), 674–692 (2008)
5. Boukrouche, M., Tarzia, D.A.: On a convex combination of solutions to elliptic variational inequalities. Electro. J. Diff. Equations (31), 1–10 (2007)
6. Boukrouche, M., Tarzia, D.A.: Convergence of distributed optimal controls for second kind parabolic variational inequalities. Nonlinear Analysis: Real World Applications 12(4), 2211–2224 (2011)
7. Brézis, H.: Problèmes unilatéraux. J. Math. Pures Appl. 51, 1–162 (1972)
8. Chipot, M.: Elements of nonlinear Analysis. Birkhäuser Advanced Texts (2000)
9. De Los Reyes, J.C.: Optimal control of a class of variational inequalities of the second kind. SIAM J. Control Optim. 49, 1629–1658 (2011)
10. Duvaut, G., Lions, J.L.: Les inéquations en Mécanique et en Physique. Dunod, Paris (1972)
11. Gariboldi, C.M., Tarzia, D.A.: Convergence of distributed optimal controls on the internal energy in mixed elliptic problems when the heat transfer coefficient goes to infinity. Appl. Math. Optim. 47(3), 213–230 (2003)
12. Gariboldi, C.M., Tarzia, D.A.: Convergence of boundary optimal controls problems with restrictions in mixed elliptic Stefan-like problems. Adv. Diff. Eq. and Control Processes 1, 113–132 (2008)
13. Kesavan, S., Muthukumar, T.: Low-cost control problems on perforated and non-perforated domains. Proc. Indian Acad. Sci. (Math. Sci.) 118(1), 133–157 (2008)
14. Kesavan, S., Saint Jean Paulin, J.: Optimal control on perforated domains. J. Math. Anal. Appl. 229, 563–586 (1997)
15. Kinderlehrer, D., Stampacchia, G.: An introduction to variational inequalities and their applications. Academic Press, New York (1980)
16. Lions, J.L.: Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles. Dunod, Paris (1968)
17. Menaldi, J.L., Tarzia, D.A.: A distributed parabolic control with mixed boundary conditions. Asymptotic Anal. 52, 227–241 (2007)
18. Mignot, F.: Contrôle dans les inéquations variationelles elliptiques. J. Functional Anal. 22(2), 130–185 (1976)
19. Rodrigues, J.F.: Obstacle problems in mathematical physics. North-Holland, Amsterdam (1987)
20. Tabacman, E.D., Tarzia, D.A.: Sufficient and/or necessary condition for the heat transfer coefficient on Γ_1 and the heat flux on Γ_2 to obtain a steady-state two-phase Stefan problem. J. Diff. Equations 77(1), 16–37 (1989)
21. Tarzia, D.A.: Una familia de problemas que converge hacia el caso estacionario del problema de Stefan a dos fases. Math. Notae 27, 157–165 (1979)
22. Tröltzsch, F.: Optimal control of partial differential equations: Theory, methods and applications. American Math. Soc., Providence (2010)

A Note on Linear Differential Variational Inequalities in Hilbert Space

Joachim Gwinner

Institut für Mathematik und Rechneranwendung, Fakultät für Luft- und Raumfahrttechnik, Universität der Bundeswehr München, 85577 Neubiberg/München, Germany

Abstract. Recently a new class of differential variational inequalities has been introduced and investigated in finite dimensions as a new modeling paradigm of variational analysis to treat many applied problems in engineering, operations research, and physical sciences. This new subclass of general differential inclusions unifies ordinary differential equations with possibly discontinuous right-hand sides, differential algebraic systems with constraints, dynamic complementarity systems, and evolutionary variational systems. In this short note we lift this class of nonsmooth dynamical systems to the level of a Hilbert space, but focus to linear input/output systems. This covers in particular linear complementarity systems where the underlying convex constraint set in the variational inequality is specialized to an ordering cone.

The purpose of this note is two-fold. Firstly, we provide an existence result based on maximal monotone operator theory. Secondly we are concerned with stability of the solution set of linear differential variational inequalities. Here we present a novel upper set convergence result with respect to perturbations in the data, including perturbations of the associated linear maps and the constraint set.

1 Introduction

Recently Pang and Stewart [18] introduced and investigated a new class of differential variational inequalities in finite dimensions as a new modeling paradigm of variational analysis to treat many applied problems in engineering, operations research, and physical sciences. This new subclass of general differential inclusions unifies ordinary differential equations with possibly discontinuous right-hand sides, differential algebraic systems with constraints, dynamic complementarity systems, and evolutionary variational systems.

Here we lift differential variational inequalities to the more general level of a Hilbert space, but focus to the case of a linear input/output regime, where the operators in the differential equation and in the additional constraint equation are linear. This covers in particular linear complementarity systems, where the underlying convex constraint set in the variational inequality is specialized to an ordering cone. Linear complementarity systems are of much use in mechanical end electrical engineering as well as in optimization [13, 20].

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 85–91, 2013.

In this note we provide an existence result that relies on maximal monotone operator theory. Furthermore we are concerned with stability of the solution set to differential variational inequalities. In this connection let us refer to [19], where at first several sensitivity results are established for initial value problems of ordinary differential equations with nonsmooth right hand sides and then applied to treat differential variational inequalities. This has to be distinguished from asymptotic Lyapunov stability that has been investigated in [1, 8, 9] for solutions of evolution variational inequalities and nonsmooth dynamical systems. Here we present a novel upper set convergence result with respect to perturbations in the data, including perturbations of the associated linear maps and of the constraint set.

2 Setting of Linear Differential Variational Inequalities

Let X, V be two real, separable Hilbert spaces that are endowed with norms $\|\cdot\|_X$, $\|\cdot\|_V$ respectively and with scalar products denoted by $\langle\cdot,\cdot\rangle, (\cdot,\cdot)$ respectively. Further let there be given $T > 0$, a convex closed subset $K \subset V$, some functions f, g on $[0,T]$ with values in X, respectively in V, and some fixed $x_0 \in X$. Then we consider the following problem: Find an X - valued function x and an V - valued function u both defined on $[0,T]$ that satisfy for a.a. (almost all) $t \in [0,T]$

$$\text{(LDVI)}(\mathcal{A}, f, g, K; x_0) \quad \begin{cases} \begin{pmatrix} \dot{x}(t) \\ q(t) \end{pmatrix} = \mathcal{A} \begin{pmatrix} x(t) \\ u(t) \end{pmatrix} + \begin{pmatrix} f(t) \\ g(t) \end{pmatrix} \\ u(t) \in K, \ \big(q(t), v - u(t)\big) \geq 0, \ \forall v \in K\,, \end{cases} \tag{1}$$

complemented by the initial condition $x(0) = x_0$. Here $\dot{x}(t)$ denotes the time derivative of $x(t)$ and $\mathcal{A} : X \times V \to X \times V$ is a given linear continuous operator that is defined by

$$\mathcal{A} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

with appropriate linear operators A, B, C, D.

For the closed convex subset K of V and for any $w \in V$, the *tangent cone* (also *support cone* or *contingent cone*, see e.g. [3]) to K at w, denoted by $T_K(w)$, is the closure of the convex cone $\bigcup\{\lambda(K - w) : \lambda > 0\}$. Then $T_K(w)$ is clearly a closed convex cone with vertex 0 and is the smallest cone S whose translate $w + S$ has vertex w and contains K. Taking polars with respect to the scalar product in V gives $(T_K(w))^0 = (T_K(w))^- =: N_K(w)$, the normal cone to K at w, which is the subdifferential of the convex indicator function on K; for notions of convex analysis see e.g. [14]. Thus the variational inequality in (1) writes as the generalized equation $-q(t) \in N_K(u(t))$.

The fixed finite time interval $[0,T]$ gives rise to the Hilbert space $L^2(0,T;V)$ endowed with the scalar product

$$[u_1,u_2] := \int_0^T \big(u_1(t),u_2(t)\big)\ dt, \quad u_1,u_2 \in L^2(0,T;V)\,.$$

Also we introduce the closed convex subset

$$\mathcal{K} := L^2(0,T;K) := \{w \in L^2(0,T;V) \mid w(t) \in K, \forall a.a.\ t \in (0,T)\} \tag{2}$$

As in [18] we consider weak solutions of a LDVI in the sense of Caratheodory. In particular, the $X-$ valued function x has to be absolutely continuous with derivative $\dot{x}(t)$ defined almost everywhere. Moreover to define the initial condition, the trace $x(0)$ is needed. Therefore (see [7], Theorem 1, p. 473) we are led to the function space

$$W(0,T;X) := \{x \mid x \in L^2(0,T;X),\ \dot{x} \in L^2(0,T;X)\}\,,$$

a Hilbert space endowed with the scalar product

$$[x_1,x_2] + [\dot{x}_1,\dot{x}_2], \quad x_1,x_2 \in W(0,T;X)\,.$$

Note that $W(0,T;X)$ is continuously and densely embedded in the space $C[0,T;X]$ of X-valued continuous functions on $[0,T]$, where the latter space is equipped with the norm of uniform convergence.

3 Solvability of Linear Differential Variational Inequalities

In this section we provide an existence result for linear differential variational inequalities based on maximal monotonicity theory [6, 16]. Here we assume that the given function g is constant, so that shortly $g \in V$.

First we rewrite the variational inequality $(LDVI)_3$ as $-q \in N_K(u)$. By $(LDVI)_2$, $-Cx \in g + Du + N_K(u)$ follows. Hence with the affine map $D_g, D_g v = g + Dv$, we can insert $u \in (D_g + N_K)^{-1}(-Cx)$ in $(LDVI)_1$ and obtain

$$\dot{x} \in Ax + B(D_g + N_K)^{-1}(-Cx) + f. \tag{3}$$

Now we adopt an argument due to Brogliato and Goeleven [5] from finite dimension to Hilbert space and assume there exists a coercive selfadjoint operator $P \in \mathcal{L}(X,X)$ such that $B = P\ C^*$. Then P admits a square root $Q \in \mathcal{L}(X,X)$, i.e. $P = Q\ Q^*$ with $Q > 0$ (coercive), hence invertible and therefore $Q^*\ C^* = Q^{-1}\ B$. With $x = -Qz$ (3) transforms to

$$\dot{z} \in Q^{-1}AQ\ z - Q^{-1}B(D_g + N_K)^{-1}(CQ\ z) - Q^{-1}f. \tag{4}$$

Let us assume that $D \geq 0$, i.e. $(D\,v, v) \geq 0$. Then $D_g + N_K$ is maximal monotone by [6, Proposition 2.4, Corollaire 2.7]. Clearly, also the inverse $(D_g + N_K)^{-1}$ is maximal monotone.

Furthermore we use the notion of the relative interior denoted by rint and assume the regularity condition $0 \in \text{rint}\left[\text{im}\,(C\ Q) - \text{dom}\left((D_g + N_K)^{-1}\right)\right]$. Then by [17, Cor. 4.4], [21, Theorem 4], also $Q^{-1}B(D_g + N_K)^{-1}CQ$ is maximal monotone in virtue of $(C\ Q)^* = Q^{-1}B$. Since $Q^{-1}AQ$ is a Lipschitz perturbation, [6, Theorem 3.17; Corollaire 3.2], [16, Theorem 2.1, Remark 2.1] applies to conclude the existence of a unique strong solution $z \in W^{1,\infty}(0,T;X)$ to (4) with $z(0) = -Q^{-1}x_0$, provided $f \in W^{1,1}(0,T;X)$ and $z_0 := -Q^{-}1\ x_0$ satisfies $CQz_0 \in \text{dom}\,(D_g + N_K)^{-1} = \text{im}\,(D_g + N_K)$.

If moreover $D > 0$ with a coercivity constant $\delta > 0$, then from the variational inequality $(LDVI)_3$ we get uniqueness of u and the estimate

$$\|u(s) - u(t)\|_V \leq \frac{\|C\|}{\delta}\|x(s) - x(t)\|_X \quad s,t \in [0,T]\,,$$

that shows that u is $W^{1,\infty}$ on $(0,T)$, too.

Thus we have proven the following existence result.

Theorem 1. *Suppose $D \geq 0$ and there exists $P = P^* > 0$ such that $B = PC^*$. Moreover assume the regularity condition $0 \in$ rint $\left[im\ (C\ Q) - dom\left((D + N_K)^{-1}\right)\right]$, where $P = Q\,Q^*$. Then for any $f \in W^{1,1}(0,T;X)$, $g \in V$, and for any x_0 such that $-Cx_0 - g \in im\ (D + N_K)$, (LDVI) is uniquely solvable with $x \in W^{1,\infty}(0,T;X)$ and $x(0) = x_0$. - If moreover $D > 0$, then u is unique, too, and $u \in W^{1,\infty}(0,T;V)$.*

Remark. Let $\mathcal{M}$ be a general maximal monotone map that replaces the above normal cone map N_K. Then by a similar reasoning as above we obtain an existence result for the multivalued Luré dynamical system

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) + f(t);\ x(0) = x_0 \\ u(t) \in \mathcal{M}\big[Cx(t) + Du(t)\big]. \end{cases}$$

Luré dynamical systems with $\mathcal{M} = \partial\varphi$, φ a convex closed and proper function have been recently studied by Brogliato and Goeleven [5] in finite dimensions with applications in nonsmooth electronics.

4 Stability of Linear Differential Variational Inequalities

In this section we study stability of linear differential variational inequalities formulated as LDVI and admit perturbations $x_{0,n}$ of x_0 in the initial condition $x(0) = x_0$, $\mathcal{A}_n = (A_n, B_n, C_n, D_n)$ of the linear map $\mathcal{A} = (A, B, C, D)$, f_n, g_n of the functions f, g, and K_n of the convex closed subset $K \subset V$. Suppose that (x^n, u^n) solves (LDVI)$(\mathcal{A}_n, f_n, g_n, K_n; x_{0,n})$ and assume that $(x^n, u^n) \to$

(x, u) with respect to an appropriate convergence for X-valued, respectively V-valued functions on $[0, T]$. Then we seek conditions on $\mathcal{A}_n \to \mathcal{A}, f_n \to f, g_n \to g, K_n \to K, x_{0,n} \to x_0$ that guarantee that (x, u) solves the limit problem (LDVI)$(\mathcal{A}, f, g, K; x_0)$. Such a stability result can be understood as a result of upper set convergence for the solution set of the LDVI.

4.1 Preliminaries; Mosco Convergence of Sets

As the convergence of choice in variational analysis we employ Mosco set convergence for a sequence $\{K_n\}$ of closed convex subsets which is defined as follows. A sequence $\{K_n\}$ of closed convex subsets of the Hilbert space V is called Mosco convergent to a closed convex subset K of V, written $K_n \xrightarrow{M} K$, if and only if

$$\sigma - \limsup_{n\to\infty} K_n \subset K \subset s - \liminf_{n\to\infty} K_n.$$

Here the prefix σ means sequentially weak convergence in contrast to strong convergence denoted by the prefix s; lim sup, respectively lim inf are in the sense of Kuratowski upper, resp. lower limits of sequences of sets (see [2, 4] for more information on Mosco convergence).

As a preliminary result we need that Mosco convergence of convex closed sets K_n inherits to Mosco convergence of the associated sets $\mathcal{K}_n = L^2(0, T; K_n)$, derived from K_n similar to (2).

Lemma 1. *Let $K_n \xrightarrow{M} K$. Then $\mathcal{K}_n \xrightarrow{M} \mathcal{K}$ in $L^2(0, T; V)$.*

For the proof we refer to [10, 12].

As a further tool in our stability analysis we recall from [11] the following technical result.

Lemma 2. *Let H be a separable Hilbert space and let $T > 0$ be fixed. Then for any sequence $\{z_n\}_{n\in\mathbf{N}}$ converging to some z in $L^1(0, T; H)$ there exists a subsequence $\{z_{n_k}\}_{k\in\mathbf{N}}$ such that for some set N of zero measure, $z_{n_k}(t) \xrightarrow{s} z(t)$ for all $t \in [0, T]\backslash N$.*

4.2 The Stability Result

We need the following hypotheses on the convergence of the perturbations:

(H1) Convergence $\mathcal{A}_n \to \mathcal{A}$ holds in the operator norm topology. - All operators D_n are monotone, i.e. for any $v \in V$, $(D_n v, v) \geq 0$ holds.

(H2) Convergence of the functions $f_n \to f, g_n \to g$ holds in $L^2(0, T; X)$, respectively in $L^2(0, T; V)$.

Now we can state the following stability result.

Theorem 2. *Let (x^n, u^n) solve (LDVI)$(\mathcal{A}_n, f_n, g_n, K_n; x_{0,n})$. Suppose, $\mathcal{A}_n$ and $\mathcal{A}$ satisfy (H1), and that f_n, g_n and f, g satisfy (H2). Let the convex closed sets K_n Mosco-converge to K and let $x_{0,n} \xrightarrow{s} x_0$. Assume that $x^n \xrightarrow{s} x$ in $W(0, T; X)$*

and that $u^n \in L^2(0,T;V)$ converges weakly to u pointwise in V for a.a. $t \in (0,T)$ with $\|u^n(t)\|_V \leq m(t)$, $\forall$ a.a. $t \in (0,T)$ for some $m \in L^2(0,T)$. Then (x,u) is a solution to (LDVI)$(\mathcal{A}, f, g, K; x_0)$.

Proof.

The proof consists of three parts.

1. Feasibility: $u \in \mathcal{K}, x(0) = x_0$.

First we observe that for any $w \in L^2(0,T;V)$, in virtue of Lebesgue's theorem of dominated convergence,

$$[u^n, w] = \int_0^T \big(u^n(t), w(t)\big)\ dt \ \to [u,w]\,.$$

Thus $u^n \overset{\sigma}{\to} u$ and $u \in L^2(0,T;V)$. Moreover directly by Mosco convergence of $\{K_n\}$ or invoking lemma 1, $u \in \mathcal{K}$ follows. - Since by continuous embedding $x^n \overset{s}{\to} x$ in $C[0,T;X]$, we conclude $x^n(0) = x_{0,n} \overset{s}{\to} x(0) = x_0$.

2. u solves the variational inequality in (LDVI)$(\mathcal{A}, f, g, K; x_0)$:

Fix an arbitrary $w \in \mathcal{K}$. Then by lemma 1, there exist $w^n \in \mathcal{K}_n$ such that $w^n \overset{s}{\to} w$ in $L^2(0,T;V)$. Moreover, by extracting eventually a subsequence, we have by lemma 2 that $w^n(t), g_n(t)$ strongly converges to $w(t), g(t)$,respectively, for a.a. $t \in (0,T)$. For any measureable set $A \subset (0,T)$ we can define $w^n_A \in L^2(0,T;V)$ by $w^n_A = w^n$ on A, $w^n_A = u^n$ on $(0,T) \setminus A$. Hence $w^n_A \in \mathcal{K}_n$ and by construction,

$$\int_A \big(q^n(t), w^n(t) - u^n(t)\big)\ dt \ \geq\ 0\,,$$

where $q^n(t) = C_n\ x^n(t) + D_n\ u^n(t) + g_n(t)$. Hence a contradiction argument shows that we have pointwise for a.a. $t \in (0,T)$, $\big(q^n(t), w^n(t) - u^n(t)\big) \geq 0$. By (H1), monotonicity entails $\big(C_n\ x^n(t) + D_n\ w^n(t) + g_n(t), u^n(t) - w^n(t)\big) \leq 0$. By (H1) and (H2), in the limit $\big(C\ x(t) + D\ w(t) + g(t), u(t) - w(t)\big) \leq 0$. In virtue of the linear growth of the linear operators we arrive at

$$[G(x,w), u - w] := \int_0^T \big(C\ x(t) + D\ w(t) + g(t), u(t) - w(t)\big)\ dt \ \leq\ 0,\ \forall w \in \mathcal{K}\,.$$

Hence by a well-known argument in monotone operator theory (see e.g. [22]) we obtain that $u \in \mathcal{K}$ satisfies the variational inequality

$$[G(x,u), w - u] \geq 0,\ \forall w \in \mathcal{K}\,.$$

3. (x,u) solves the limit problem (LDVI)$(\mathcal{A}, f, g, K; x_0)$:

By Lemma 2 applied to $\{f_n\}$, $\{x^n\}$, and $\{\dot{x}^n\}$, we can extract a subsequence such that $f_n(t) \to f(t)$, $x^n(t) \to x(t)$, and $\dot{x}^n(t) \to \dot{x}(t)$ strongly in X pointwise for all $t \in (0,T)\backslash N_0$, where N_0 is a null set. Fix $t \in (0,T)\backslash N_0$. Then by assumption, for all $n \in \mathbf{N}$ we have $\dot{x}^n(t) = A_n\ x^n(t) + B_n\ u^n(t) + f_n(t)$. Then in virtue of (H1) and (H2), $\dot{x}(t) = A\ x(t) + B\ u(t) + f(t)$ follows and (x,u) solves (LDVI)$(\mathcal{A}, f, g, K; x_0)$. ∎

References

[1] Adly, S., Goeleven, D.: A stability theory for second-order nonsmooth dynamical systems with application to friction problems. J. Math. Pures Appl. 83(9), 17–51 (2004)

[2] Attouch, H.: Variational convergence for functions and operators. Pitman, Boston (1984)

[3] Aubin, J.-P., Cellina, A.: Differential Inclusions. Set-valued Maps and Viability Theory. Springer, Berlin (1984)

[4] Aubin, J.-P., Frankowska, H.: Set-Valued Analysis. Birkhäuser, Boston (1990)

[5] Brogliato, B., Goeleven, D.: Well-posedness, stability and invariance results for a class of multivalued Luré dynamical systems. Nonlinear Analysis 74, 195–212 (2011)

[6] Brézis, H.: Operateurs maximaux monotones. North-Holland, Amsterdam (1973)

[7] Dautray, R., Lions, J.L.: Mathematical Analysis and Numerical Methods for Science and Technology. Evolution Problems I, vol. 5. Springer, Berlin (1992)

[8] Goeleven, D., Brogliato, B.: Necessary conditions of asymptotic stability for unilateral dynamical systems. Nonlinear Anal., Theory Methods Appl. A 61, 961–1004 (2005)

[9] Goeleven, D., Motreanu, D., Motreanu, V.V.: On the stability of stationary solutions of first order evolution variational inequalities. Adv. Nonlinear Var. Inequal. 6, 1–30 (2003)

[10] Gwinner, J.: A class of random variational inequalities and simple random unilateral boundary value problems – existence, discretization, finite element approximation. Stochastic Anal. Appl. 18, 967–993 (2000)

[11] Gwinner, J.: On differential variational inequalities and projected dynamical systems – equivalence and a stability result. Discrete Contin. Dyn. Syst. 2007, 467–476 (2007)

[12] Gwinner, J.: On a new class of differential variational inequalities and a stability result. Math. Programming (to appear)

[13] Heemels, W.P.M.H., Schumacher, J.M., Weiland, S.: Linear complementarity systems. SIAM J. Appl. Math. 60, 1234–1269 (2000)

[14] Ioffe, A.D., Tihomirov, V.M.: Theory of extremal problems. Translated from the Russian by K. Makowski. North-Holland, Amsterdam (1979)

[15] Kinderlehrer, D., Stampacchia, G.: An Introduction to Variational Inequalities and Their Applications. Academic Press, New York (1984)

[16] Morosanu, G.: Nonlinear Evolution Equations and Applications. D. Reidel, Dordrecht (1988)

[17] Pennanen, T.: Dualization of generalized equations of maximal monotone type. SIAM J. Opt. 10, 809–835 (2000)

[18] Pang, J.-S., Stewart, D.E.: Differential variational inequalities. Math. Program. 113, 345–424 (2008)

[19] Pang, J.-S., Stewart, D.E.: Solution dependence on initial conditions in differential variational inequalities. Math. Program. 116, 429–460 (2009)

[20] Schumacher, J.M.: Complementarity systems in optimization. Math. Progr., Ser. B 101, 263–295 (2004)

[21] Robinson, S.M.: Composition duality and maximal monotonicity. Math. Progr. 85, 1–13 (1999)

[22] Zeidler, E.: Nonlinear Functional Analysis and its Applications. Nonlinear Monotone Operators, vol. II/B. Springer, New York (1990)

Model Order Reduction for Networks of ODE and PDE Systems

Michael Hinze and Ulrich Matthes

Department of Mathematics, University of Hamburg, Bundesstr. 55, 20146 Hamburg, Germany
michael.hinze@uni-hamburg.de
ulrich.matthes@math.uni-hamburg.de

Abstract. We propose a model order reduction (MOR) approach for networks containing simple and complex components. Simple components are modeled by linear ODE (and/or DAE) systems, while complex components are modeled by nonlinear PDE (and/or PDAE) systems. These systems are coupled through the network topology using the Kirchhoff laws. As application we consider MOR for electrical networks, where semiconductors form the complex components which are modeled by the transient drift-diffusion equations (DDEs). We sketch how proper orthogonal decomposition (POD) combined with discrete empirical interpolation (DEIM) and passivity-preserving balanced truncation methods for electrical circuits (PABTEC) can be used to reduce the dimension of the model. Furthermore we investigate residual-based sampling to construct reduced order models which are valid over a certain parameter range.

Keywords: Model Order Reduction, Parametrized Dynamical Systems, Drift-Diffusion Equations, Integrated Circuits.

AMS subject classifications: 93A30, 65B99, 65M60, 65M20.

1 Introduction

In this paper we propose a simulation-based MOR approach for the reduction of networks consisting of (many) simple and (only few) complex components. We assume that the simple and complex components are modeled by systems of linear ODEs (DAEs) and nonlinear PDEs (PDAEs), respectively, which are coupled through the network topology using the Kirchhoff laws.

As application we consider electrical networks where the simple components consist of resistors, capacitors, voltage sources, current sources, and inductors, and the complex components are formed by e.g. semi-conductors, see Figure 1. The overall system is then represented by a nonlinear partial differential algebraic equation (PDAE) system, see e.g. [3,8]. In this paper we address the following issues:

1. construction of reduced order models for the complex components;
2. reduction of the complete network while retaining the structure of a network;
3. parametric MOR for complex components.

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 92–101, 2013.

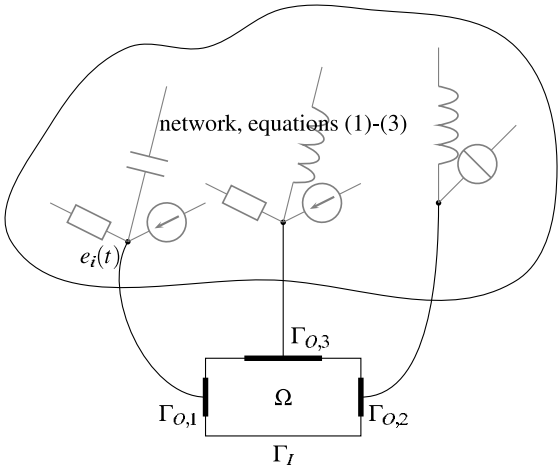


Fig. 1. Sketch of a coupled system with one semiconductor forming the complex component

2 Example: Modeling of an Electrical Network

In electrical networks resistors, capacitors, and inductors form the simple components which in general are modeled by linear ODEs. Complex components are given by e.g. semiconductors which are modeled by PDAE systems, see below. Considering additional voltage and current sources the overall network can be modeled by a PDAE which is obtained as follows. First the network containing only the simple components is modeled by a differential algebraic equation (DAE) which is obtained by a modified nodal analysis (MNA), including the Ohmic contacts $\Gamma_{O,k}$ of the semiconductors as network nodes, see Figure 1. Denoting by e the node potentials and by j_L, j_V, and j_S the currents of inductive, voltage source, and semiconductor branches, the DAE reads (see [8,12,19])

$$A_C \frac{d}{dt} q_C(A_C^\top e, t) + A_R g(A_R^\top e, t) + A_L j_L + A_V j_V + A_S j_S = -A_I i_s(t), \tag{1}$$

$$\frac{d}{dt} \phi_L(j_L, t) - A_L^\top e = 0, \tag{2}$$

$$A_V^\top e = v_s(t). \tag{3}$$

Here, the incidence matrix $A = [A_R, A_C, A_L, A_V, A_S, A_I] = (a_{ij})$ represents the network topology, e.g. at each non mass node i, $a_{ij} = 1$ if the branch j leaves node i and $a_{ij} = -1$ if the branch j enters node i and $a_{ij} = 0$ elsewhere. The indices R, C, L, V, S, I denote the capacitive, resistive, inductive, voltage source, semiconductor, and current source branches, respectively. In particular the matrix A_S denotes the semiconductor incidence matrix. The vector valued functions q_C, g and ϕ_L are continuously differentiable defining the voltage-current

relations of the network components. The continuous vector valued functions v_s and i_s are the voltage and current sources. For details we refer to [10].

In a second step the semiconductors are modeled by PDAE systems, which are then coupled to the DAE of the network. Here we use the transient drift-diffusion equations as a continuous model for semiconductors, see e.g. [1,3] and the references cited there. Using the notation and scaling introduced there, we obtain the following scaled system of PDEs for the electrostatic potential $\psi(t,x)$, the electron and hole concentrations $n(t,x)$ and $p(t,x)$ and the current densities $J_n(t,x)$ and $J_p(t,x)$:

$$\lambda \Delta \psi = n - p - C, \tag{4}$$

$$-\partial_t n + \nu_n \operatorname{div} J_n = \quad R(n,p), \tag{5}$$

$$\partial_t p + \nu_p \operatorname{div} J_p = -R(n,p), \tag{6}$$

$$J_n = \quad \nabla n - n \nabla \psi, \tag{7}$$

$$J_p = -\nabla p - p \nabla \psi. \tag{8}$$

Here $(t,x) \in [0,T] \times \Omega$ and $\Omega \subset \mathbb{R}^d$. The nonlinear function R describes the rate of electron/hole recombination, $\lambda > 0$ is the scaled Debye length, ν_n and ν_p are the scaled mobilities of electrons and holes. The temperature is assumed to be constant which leads to a constant thermal voltage U_T. The function C is the time independent doping profile.

This system is supplemented with the boundary conditions

$$\psi(t,x) = \psi_{bi}(x) + (A_S^\top e(t))_k = U_T \log\left(\frac{\sqrt{C(x)^2 + 4n_i^2} + C(x)}{2n_i}\right) + (A_S^\top e(t))_k, \tag{9}$$

$$n(t,x) = \frac{1}{2}\left(\sqrt{C(x)^2 + 4n_i^2} + C(x)\right), \quad p(t,x) = \frac{1}{2}\left(\sqrt{C(x)^2 + 4n_i^2} - C(x)\right), \tag{10}$$

for $(t,x) \in [0,T] \times \Gamma_{O,k}$, where the potential of the nodes which are connected to a semiconductor interface enter in the boundary conditions for ψ. Here, $\psi_{bi}(x)$ denotes the build-in potential and n_i the constant intrinsic concentration. All other parts of the boundary are isolation boundaries $\Gamma_I := \Gamma \setminus \Gamma_O$, where $\nabla \psi \cdot \nu = 0$, $J_n \cdot \nu = 0$ and $J_p \cdot \nu = 0$ holds. The semiconductor model (4)-(8) is coupled to the network through the semiconductor current vector j_S with the components

$$j_{S,k} = \int_{\Gamma_{O,k}} (J_n + J_p - \varepsilon \partial_t \nabla \psi) \cdot \nu \, d\sigma, \tag{11}$$

where ν denotes the unit outward normal to the interface $\Gamma_{O,k}$. More details, including a precise description of the coupling, are given in [10]. The analytical and numerical analysis of PDAE systems of the presented form is subject to current research, see [3,7,16,19].

3 Reduced Order Models for Complex Components

We assume that every complex component is modeled by a time-dependent PDE or PDAE system which is amenable to a numerical treatment with Galerkin

methods. After appropriate spatial discretization the method of lines then yields a large, nonlinear ODE system representing the spatially discrete complex component. This nonlinear ODE or DAE system now represents the complex component in the network. The reduction of the complex components is based on simulation-based MOR with POD. In this approach time snapshots of the complex components are extracted from snapshots of the simulation of the complete network. POD for the complex component then is performed using the extracted parts of the snapshots. In combination with DEIM [5] this now delivers low dimensional, nonlinear surrogate models for the complex components, see [9] for details.

Among other things it is an important feature of this reduction technique that it delivers distinct reduced order models for the same complex component at different locations in the network.

As example let us consider the rectifier network in Figure 2 (left). The POD basis functions of two identical semiconductors may be different due to different operating states of the semiconductors. Simulation results for this network are plotted in Figure 2 (right). Details of the implementation are sketched in Section 4. The distance between the linear spaces U^1 and U^2 which are spanned, e.g., by the POD-basis-functions U^1_ψ associated to ψ for the diode S_1 and U^2_ψ associated to ψ for the diode S_2 respectively, is measured by

$$d(U^1, U^2) := \max_{\substack{u \in U^1 \\ \|u\|_2 = 1}} \min_{\substack{v \in U^2 \\ \|v\|_2 = 1}} \|u - v\|_2.$$

Exploiting the orthonormality of the bases U^1_ψ and U^2_ψ and using a Lagrange framework, we find

$$d(U^1, U^2) = \sqrt{2 - 2\sqrt{\lambda}},$$

where λ is the smallest eigenvalue of the positive definite matrix $SS^\top$ with $S_{ij} = \langle u^1_{\psi,i}, u^2_{\psi,j} \rangle_2$. Here, $u^1_{\psi,i}$ denotes the i-th node in U^2_ψ, $u^2_{\psi,j}$ the j-th node in U^2_ψ. The distances for the rectifier network are given in Table 1. While the reduced model for the diodes S_1 and S_3 are almost equal, the reduced models for the diodes S_1 and S_2 are significantly different. Similar results are obtained for the reduction of the variables n, p, etc.

Table 1. Distances between reduced models in the rectifier network

Δ	$d(U^1, U^2)$	$d(U^1, U^3)$
10^{-4}	0.61288	$5.373 \cdot 10^{-8}$
10^{-5}	0.50766	$4.712 \cdot 10^{-8}$
10^{-6}	0.45492	$2.767 \cdot 10^{-7}$
10^{-7}	0.54834	$1.211 \cdot 10^{-6}$

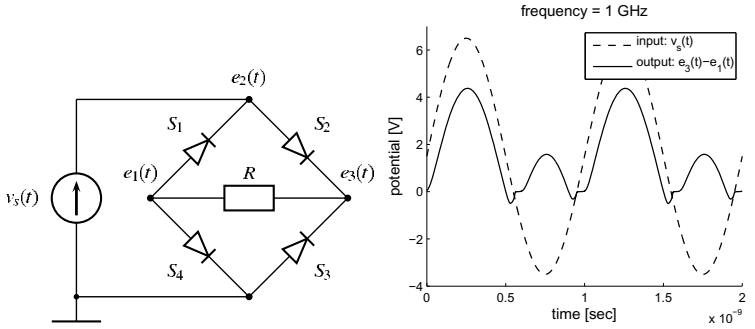


Fig. 2. Left: Rectifier network with 4 identical semiconductors. Right: Simulation results for the rectifier network. The input v_s is sinusoidal with frequency 1 GHz and offset $+1.5$ V. The time integration of the underlying nonlinear DAE system is performed with DASPK [4,14].

4 Reduction of the Whole Network

Let us assume that the overall network with simple and complex components now is represented by a nonlinear DAE system, where the linear part stems from the simple components, and the nonlinear part from the spatially-discrete complex components. The reduction for the complex components now is performed as in the previous section, whereas the linear part is approximated by a reduced order linear model of lower dimension. In the case of an electrical network the passivity preserving reduction method PABTEC [18] can be used to perform the reduction of the linear part of the network. Finally, the reduced order models, for the linear an the nonlinear part have to be recoupled appropriately, for details we refer to e.g. [17]. To illustrate the performance of this approach we report on the numerical results obtained in [11] for an electrical network formed by an RC chain with one diode, see Figure 3.

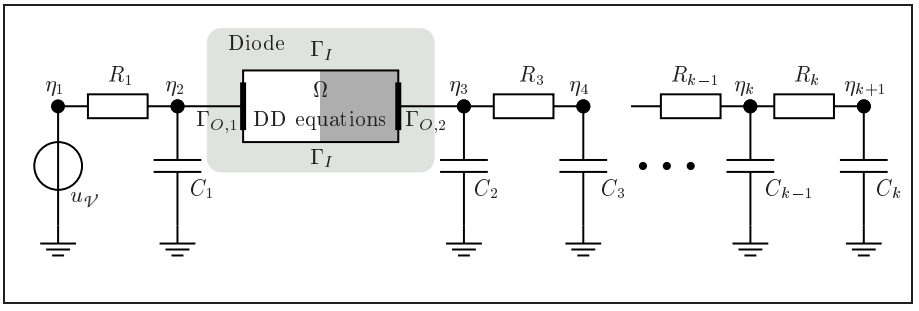


Fig. 3. RC chain with a diode

For model reduction of the linear circuit equations we use the MATLAB Toolbox PABTEC [15]. The POD method is implemented in C++ based on the FEM library deal.II [2] for discretizing the drift-diffusion equations. The obtained large and sparse nonlinear DAE system as well as the small and dense reduced-order model are integrated using the DASPK software package [4] based on a BDF method, where the nonlinear equations are solved using Newton's method. Furthermore, the direct sparse solver SuperLU [6] is employed for solving linear systems.

For the RC circuit with one diode in Figure 3 we use the input

$$u(t) = u_{\mathcal{V}}(t) = 10\sin(2\pi f_0 t)^4$$

with the frequency $f_0 = 10^4$ Hz. The output of the system is $y(t) = -\imath_{\mathcal{V}}(t)$. We simulate the models over the fixed time horizon $[0, \frac{2.5}{f_0}]$. The linear resistors have the same resistance $R = 2\,\mathrm{k}\Omega$ and the linear capacitors have the same capacitance $C = 0.02\,\mu\mathrm{F}$.

We use the transient drift-diffusion equations to model the diode. For the parameters of the diode and the related scaling we refer to [11]. In Table 2 we collect the numerical results for our reduction strategy. The outputs of the systems with the reduced network and POD-reduced diode are compared to the fully, spatially semidiscretized model with 7510 variables.

Here we construct a POD-reduced model for the diode based on a FE simulation with 500 nodes, where we apply DEIM for the reduction of the nonlinearity. The resulting reduced-order model for the diode is a dense nonlinear DAE of dimension 105 while the original spatially discrete model of the diode has dimension 6006. In Table 2 we summarize the results of the numerical simulations for the full nonlinear DAE system and the recoupled reduced system. The results demonstrate that the recoupling of the PABTEC reduced order model with the POD-MOR model for the semiconductor delivers an overall reduced-order model for the circuit-device system which allows significantly faster simulations (speedup-factor is about 20) while keeping the relative errors below 10 %.

In Figure 4 the evolution of the output currents is depicted for the full and the reduced systems. In addition, the evolutions of the output currents for the partially reduced systems (only reduction of the linear network, and only reduction of the diode) are shown.

Table 2. Statistics for model reduction of the coupled circuit-device system

network (MNA equations)	diode (DD equations)	dim.	simul. time	Jacobian evaluations	absolute error	relative error $\|y-\hat{y}\|_{\mathbb{L}_2}\|y-\hat{y}\|_{\mathbb{L}_2}/\|y\|_{\mathbb{L}_2}$
unreduced	unreduced	7510	23.37s	20		
reduced	reduced	130	1.19s	11	$2.954 \cdot 10^{-6}$	$1.000 \cdot 10^{-1}$

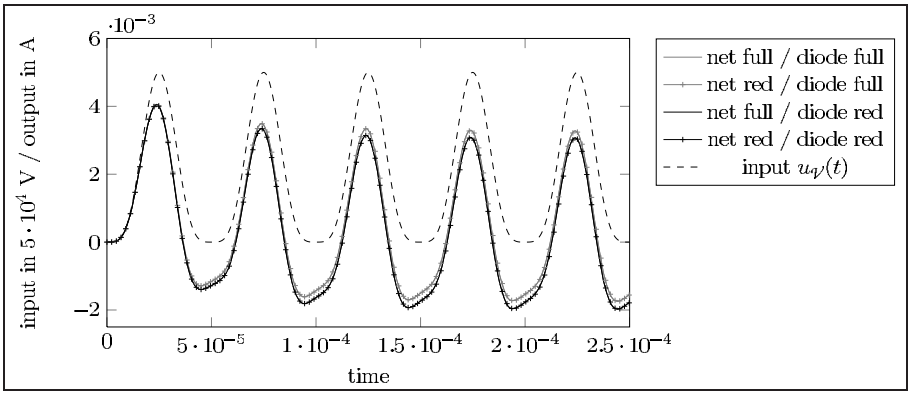


Fig. 4. Input voltage and output currents for different model reduction setups

5 Parametric Model Order Reduction with Residual Based Sampling

One major difficulty in simulation based MOR for complex components modeled by e.g. nonlinear PDE systems consists in the construction of reduced order models which are valid over a certain input parameter range, where the latter for electrical networks may be given by the input frequency. To obtain reduced order models for the complex components we propose residual based sampling which detects extreme parameters by evaluating the residual $\mathcal{R}$ of the reduced models over the parameter span. The greedy approach proposed in [13] then is used to enrich the simulation basis for the construction of a new reduced order model of the complex component, see Algorithm 1.

We summarize our ideas in the following sampling algorithm, for details see [10]. Let $\mathcal{P}$ denote the parameter space and $\omega \in \mathcal{P}$ a parameter. Furthermore, let $\mathcal{R}(z^{POD}(\omega, P))$ denote the residual obtained by evaluation of the unreduced model at the solution of the reduced order model $z^{POD}(\omega, P)$ based on snapshots taken on the parameter set $P \subset \mathcal{P}$.

Algorithm 1 (Sampling)

1. *Select* $\omega_1 \in \mathcal{P}$, $P_{test} \subset \mathcal{P}$, $tol > 0$, *and set* $k := 1$, $P_1 := \{\omega_1\}$. *Simulate the unreduced model at* ω_1 *and calculate the reduced model with POD basis functions* U^1.
2. *Calculate the residual* $\|\mathcal{R}(z^{POD}(\omega, P_k))\|$ *for all* $\omega \in P_{test}$.
3. *Check termination conditions, e.g.*
   - $\max_{\omega \in P_{test}} \|\mathcal{R}(z^{POD}(\omega, P_k))\| < tol$, *or*
   - *no further reduction of residual, then STOP.*
4. *Calculate* $\omega_{k+1} := \arg\max_{\omega \in P_{test}} \|\mathcal{R}(z^{POD}(\omega, P_k))\|$.

5. *Simulate the unreduced model at ω_{k+1} and create a new reduced model with POD basis U^{k+1} using also the already available information at $\omega_1, \ldots, \omega_k$.*
6. *Set $P_{k+1} := P_k \cup \{\omega_{k+1}\}$, $k := k+1$ and goto 3.*

The step 5 in Algorithm 1 can be executed in different ways. If offline time and offline memory requirements are not critical one may combine snapshots from all simulations of the full model and redo the model order reduction on the large snapshot ensemble. Otherwise a new reduced model at reference frequency ω_{k+1} may be constructed using the current POD-basis $\bar{U}$ and then perform an additional POD step on $(U_k, \bar{U})$.

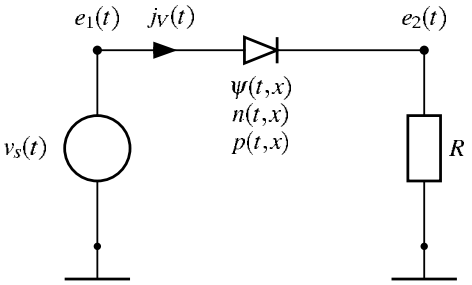


Fig. 5. Basic test circuit with one diode

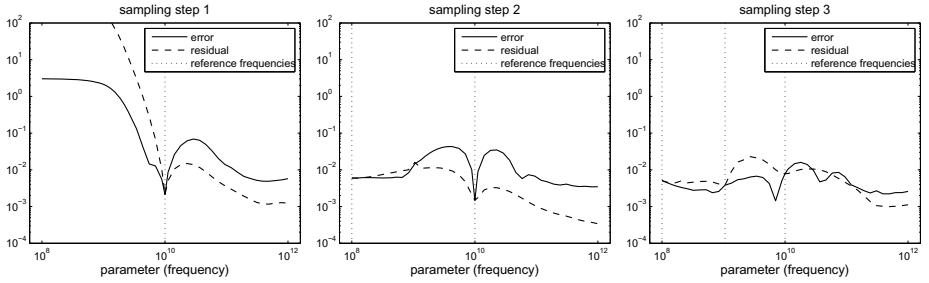


Fig. 6. Relative reduction error (solid line) and residual (dashed line) plotted over the frequency parameter space. The reduced model is based on simulations at the frequency parameter space. The reduced model is based on simulations at the reference frequencies $\omega_1 := 10^{10}\ Hz$ (left), ω_1 and $\omega_2 := 10^8\ Hz$ (middle), and ω_1, ω_2, and $\omega_3 := 1.0608 \cdot 10^9\ Hz$ (right). The reference frequencies are marked by vertical dotted lines.

To illustrate the performance of the sampling procedure we now apply Algorithm 1 to provide a reduced order model of the basic circuit shown in Figure 5. We choose the frequency of the input voltage v_s as model parameter with parameter space $\mathcal{P} := [10^8,\ 10^{12}]\ Hz$. We initialize with a reduced model which

Table 3. Performance of Algorithm 1

step k	reference parameters P_k	max. residual (at frequency)	max. relative error (at frequency)
1	$\{1.0000 \cdot 10^{10}\}$	$9.9864 \cdot 10^{2}$ $(1.0000 \cdot 10^{8})$	$3.2189 \cdot 10^{0}$ $(1.0000 \cdot 10^{8})$
2	$\{1.0000 \cdot 10^{8}, 1.0000 \cdot 10^{10}\}$	$1.5982 \cdot 10^{-2}$ $(1.0608 \cdot 10^{9})$	$4.3567 \cdot 10^{-2}$ $(3.4551 \cdot 10^{9})$
3	$\{1.0000 \cdot 10^{8}, 1.0608 \cdot 10^{9}, 1.0000 \cdot 10^{10}\}$	$2.2829 \cdot 10^{-2}$ $(2.7283 \cdot 10^{9})$	$1.6225 \cdot 10^{-2}$ $(1.8047 \cdot 10^{10})$

is constructed from the simulation of the full model at the reference frequency $\omega_1 := 10^{10}\ Hz$. The number of POD basis functions s is chosen such that the lack of information content $\Delta(s)$ is approximately 10^{-7}. The relative error and the residual are plotted in Figure 6 (left). We observe that the residual admits a structure similar to that of the approximation error. Using Algorithm 1 the next additional reference frequency is $\omega_2 := 10^{8}\ Hz$ since it maximizes the residual.

The next two iterations of the sampling algorithm are also depicted in Figure 6. Based on the residual in step 2, one selects $\omega_3 := 1.0608 \cdot 10^{9}\ Hz$ as the next reference frequency. Since no further reduction of the residual is achieved in step 3, the algorithm terminates. The maximal errors and residuals are given in Table 3. We note that in practical applications the error is not amenable over the whole parameter span. However the residual at least in the presented example seams to deliver a reliable indicator for the expected model error.

Acknowledgements. The work reported in this paper was supported by the German Federal Ministry of Education and Research (BMBF), grant no. 03HIPAE5. Responsibility for the contents of this publication rests with the authors.

References

1. Anile, A., Mascali, G., Romano, V.: Mathematical problems in semiconductor physics. Lectures given at the C. I. M. E. summer school, Cetraro, Italy, July 15-22, 1998. Lecture Notes in Mathematics. Springer, Berlin (2003)
2. Bangerth, W., Hartmann, R., Kanschat, G.: deal.II — a general-purpose object-oriented finite element library. ACM Trans. Math. Softw. 33(4) (2007)
3. Bodestedt, M., Tischendorf, C.: PDAE models of integrated circuits and index analysis. Math. Comput. Model. Dyn. Syst. 13(1), 1–17 (2007)
4. Brown, P., Hindmarsh, A., Petzold, A.: A description of DASPK: A solver for large-scale differential-algebraic systems. Tech. rep., Lawrence Livermore National Report UCRL (1992)
5. Chaturantabut, S., Sorensen, D.: Nonlinear model reduction via discrete empirical interpolation. SIAM J. Sci. Comput. 32(5), 2737–2764 (2010)

6. Demmel, J.W., Eisenstat, S.C., Gilbert, J.R., Li, X.S., Liu, J.W.H.: A supernodal approach to sparse partial pivoting. SIAM J. Matrix Analysis and Applications 20(3), 720–755 (1999)
7. Günther, M.: Partielle differential-algebraische Systeme in der numerischen Zeitbereichsanalyse elektrischer Schaltungen. VDI Fortschritts-Berichte, Reihe 20, Rechnerunterstützte Verfahren, vol. 343 (2001)
8. Günther, M., Feldmann, U., ter Maten, J.: Modelling and discretization of circuit problems. In: Schilders, W.H.A., et al. (eds.) Handbook of Numerical Analysis. Special volume: Numerical methods in electromagnetics, vol. XIII, pp. 523–629. Elsevier/North Holland, Amsterdam (2005)
9. Hinze, M., Kunkel, M.: Discrete empirical interpolation in pod model order reduction of drift-diffusion equations in electrical networks. In: SCEE Proceedings 2010, Toulouse (2010)
10. Hinze, M., Kunkel, M.: Residual based sampling in POD model order reduction of drift-diffusion equations in parametrized electrical networks. Z. Angew. Math. Mech. 92, 91–104 (2012)
11. Hinze, M., Kunkel, M., Steinbrecher, A., Stykel, T.: Model order reduction of coupled circuit-device systems. Int. J. Numer. Model. (2012), doi:10.1002/jnm.840
12. Ho, C., Ruehli, A., Brennan, P.: The modified nodal approach to network analysis. IEEE Trans. Circuits Syst. 22, 504–509 (1975)
13. Patera, A., Rozza, G.: Reduced Basis Approximation and A Posteriori Error Estimation for Parametrized Partial Differential Equations. Version 1.0. Copyright MIT 2006–2007, to appear in (tentative rubric) MIT Pappalardo Graduate Monographs in Mechanical Engineering (2007)
14. Petzold, L.R.: A description of DASSL: A differential/algebraic system solver. IMACS Trans. Scientific Computing 1, 65–68 (1993)
15. Salih, H., Steinbrecher, A., Stykel, T.: MATLAB Toolbox PABTEC - A users guide. Technical Report, Institut für Mathematik, Technische Universität Berlin, Germany (2011)
16. Selva Soto, M., Tischendorf, C.: Numerical analysis of DAEs from coupled circuit and semiconductor simulation. Appl. Numer. Math. 53(2-4), 471–488 (2005)
17. Steinbrecher, A., Stykel, T.: Model order reduction of nonlinear circuit equations. Technical Report 2011/02, Institut für Mathematik, Technische Universität Berlin, Germany (2011)
18. Stykel, T., Reis, T.: The PABTEC algorithm for passivity-preserving model reduction of circuit equations. In: Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems (MTNS 2010), July 5-9. ELTE, Budapest (2010) (paper 363)
19. Tischendorf, C.: Coupled Systems of Differential Algebraic and Partial Differential Equations in Circuit and Device Simulation. Habilitation thesis, Humboldt-University of Berlin (2003)

Path-Planning with Collision Avoidance in Automotive Industry

Chantal Landry[1], Matthias Gerdts[2], René Henrion[1], and Dietmar Hömberg[1]

[1] Weierstrass Institute, Mohrenstr. 39, 10117 Berlin, Germany
{chantal.landry,rene.henrion,dietmar.hoemberg}@wias-berlin.de
[2] Institute of Mathematics and Applied Computing, University of the Federal Armed Forces at Munich, Werner-Heisenberg-Weg 39, 85577 Neubiberg, Germany
matthias.gerdts@unibw.de

Abstract. An optimal control problem to find the fastest collision-free trajectory of a robot is presented. The dynamics of the robot is governed by ordinary differential equations. The collision avoidance criterion is a consequence of Farkas's lemma and is included in the model as state constraints. Finally an active set strategy based on backface culling is added to the sequential quadratic programming which solves the optimal control problem.

Keywords: Optimal control, collision avoidance, backface culling, active set strategy.

1 Collision Avoidance

In automotive industry robots have to work simultaneously on the same workpiece. The challenge is to find for each robot the fastest trajectory that avoids any collision with the surrounding obstacles and the other robots. We start with the establishment of the collision avoidance criterion.

For simplicity, we suppose that only one obstacle is present in the workspace. As in [7,8] a collision detection can be obtained when the robot is approximated by a union of convex polyhedra. This union is called P and it given by

$$P = \bigcup_{i=1}^{n_P} P^{(i)}, \text{ with } P^{(i)} = \{y \in \mathbb{R}^3 | A^{(i)} y \leq b^{(i)}\}$$

where n_P is the number of polyhedra in P. If p_i denotes the number of faces in $P^{(i)}$, then $A^{(i)} \in \mathbb{R}^{p_i \times 3}$ and $b^{(i)} \in \mathbb{R}^{p_i}$ for $i = 1, \ldots, n_P$.

Similarly, the obstacle is approximated by the following union of convex polyhedra, called Q

$$Q = \bigcup_{j=1}^{n_Q} Q^{(j)}, \text{ with } Q^{(j)} = \{y \in \mathbb{R}^3 | C^{(j)} y \leq d^{(j)}\}$$

where n_Q is the number of polyhedra in Q. If q_j is the number of faces in $Q^{(j)}$, then $C^{(j)} \in \mathbb{R}^{q_j \times 3}$ and $d^{(j)} \in \mathbb{R}^{q_j}$ for $j = 1, \ldots, n_Q$. In the following, n_P, A, b

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 102–111, 2013.

and i are always associated with the robot, and n_Q, C, d, and j are related to the obstacle. Furthermore, the robot will be identified with its approximation P and the obstacle with Q.

A first characterization of the collision-freeness between P and Q is given by

$$P^{(i)} \cap Q^{(j)} = \emptyset, \quad \forall i = 1, \ldots, n_P \text{ and } \forall j = 1, \ldots, n_Q.$$

The definition of the polyhedra $P^{(i)}$ and $Q^{(j)}$ implies that $P^{(i)}$ does not collide with $Q^{(j)}$ if and only if there does not exist any point $y^{(i,j)} \in \mathbb{R}^3$ satisfying

$$\begin{pmatrix} A^{(i)} \\ C^{(j)} \end{pmatrix} y^{(i,j)} \leq \begin{pmatrix} b^{(i)} \\ d^{(j)} \end{pmatrix}.$$

According to Farkas's lemma [1], this linear system does not have any solution if and only if there exists a vector $w^{(i,j)} \in \mathbb{R}^{p_i+q_j}$ such that

$$w^{(i,j)} \geq 0, \begin{pmatrix} A^{(i)} \\ C^{(j)} \end{pmatrix}^\top w^{(i,j)} = 0 \text{ and } \begin{pmatrix} b^{(i)} \\ d^{(j)} \end{pmatrix}^\top w^{(i,j)} < 0.$$

In conclusion, the pair of polyhedra $(P^{(i)}, Q^{(j)})$ is collision-free if and only if such a vector $w^{(i,j)}$ exists. This forms the collision avoidance characterization between a pair of polyhedra. Between the robot and the obstacle, the characterization reads as follows:

Proposition 1. *Two unions of convex polyhedra $P = \bigcup_{i=1}^{n_P} P^{(i)}$ and $Q = \bigcup_{j=1}^{n_Q} Q^{(j)}$ do not collide if and only if for each pair of polyhedra $(P^{(i)}, Q^{(j)})$, $i = 1, \ldots, n_P$, $j = 1, \ldots, n_Q$, there exists a vector $w^{(i,j)} \in \mathbb{R}^{p_i+q_j}$ such that*

$$w^{(i,j)} \geq 0, \begin{pmatrix} A^{(i)} \\ C^{(j)} \end{pmatrix}^\top w^{(i,j)} = 0 \text{ and } \begin{pmatrix} b^{(i)} \\ d^{(j)} \end{pmatrix}^\top w^{(i,j)} < 0.$$

2 Optimal Control Problem

To express the dynamics of the robot, we need to describe P differently from the previous section. As an industrial robot, P is composed by m links and is asked to move from its current position to a desired point [12]. Let $q = (q_1, \ldots, q_m)$ denote the vector of joint angles at the joints of the robot. The vector $v = (v_1, \ldots, v_m)$ contains the joint angle velocities and $u = (u_1, \ldots, u_m)$ describes the torques applied at the center of gravity of each link. The Lagrangian form of the dynamics of the robot depends on these three vectors as follows

$$q'(t) = v(t) \text{ and } \mathcal{M}(q(t))\, v'(t) = \mathcal{G}(q(t), v(t)) + \mathcal{F}(q(t), u(t)), \tag{1}$$

where $\mathcal{M}(q)$ is the symmetric and positive definite mass matrix, $\mathcal{G}(q, v)$ contains the generalized Coriolis forces and $\mathcal{F}(q, u)$ is the vector of applied joint torques and gravity forces [10,12].

For the remainder of the paper, let us define the vector $x = (q, v) \in \mathbb{R}^{n_x}$ with $n_x := 2m$. With the definition of x and the non-singularity of the matrix $\mathcal{M}$, we can define the function $f : \mathbb{R}^{n_x} \times \mathbb{R}^m \to \mathbb{R}^{n_x}$ as follows

$$f(x,u) = \begin{pmatrix} v \\ \mathcal{M}^{-1}(q)\left(\mathcal{G}(q,v) + \mathcal{F}(q,u)\right) \end{pmatrix}.$$

The fastest trajectory of a robot is the solution of an optimal control problem where the system of ordinary differential equations (ODE) are given by (1), see [3,6]. If an obstacle is present in the workspace, the collision-freeness is assured as soon as the vector $w^{(i,j)}$ of Proposition 1 is found at each time t and for all pairs of polyhedra. However, to be written as state constraints, the strict inequality in Proposition 1 has to be relaxed. Furthermore, since the robot moves, the matrices $A^{(i)}$ and the vectors $b^{(i)}$ evolves in time. Their evolution depends explicitly on $x(t)$. A complete formulation of $A^{(i)}(x(t))$ and $b^{(i)}(x(t))$ is given in [6].

Before writing down the model, let us define the index transformation $I = (i-1)n_Q + j$. Hence, to each pair $(i,j) \in \{1,\dots,n_P\} \times \{1,\dots,n_Q\}$ there corresponds an index I in $\{1,\dots,n_P n_Q\}$, and reciprocally. In the sequel, the index I is used instead of the pair (i,j). The variable $w^{(i,j)}$ is then numbered as w_I. Let us also define the functions $G_I : \mathbb{R}^{n_x} \to \mathbb{R}^{(p_i+q_j)\times 3}$ and $g_I : \mathbb{R}^{n_x} \to \mathbb{R}^{p_i+q_j}$ for $I = 1,\dots,n_P n_Q$ as follows

$$G_I(x) = \begin{pmatrix} A^{(i)}(x) \\ C^{(j)} \end{pmatrix} \text{ and } g_I(x) = \begin{pmatrix} b^{(i)}(x) \\ d^{(j)} \end{pmatrix}.$$

G_I and g_I allow us to write Proposition 1 as a function of time. Finally let set $M := n_P n_Q$ the number of indices I and $n_I := p_i + q_j$ the size of w_I for $I = 1,\dots,M$, and let t_f denote the travel time. Then, after transformation onto the fixed time interval $T := [0,1]$ the optimal control problem reads as follows:

(OCP): *minimize* $\varphi(x(0), x(1), t_f)$

with respect to the state variable $x \in W^{n_x}_{1,\infty}(T)$*, the control variables* $u \in L^m_\infty(T)$ *and* $w_I \in L^{n_I}_\infty(T)$*,* $I = 1,\dots,M$*, and* $t_f \geq 0$*, subject to:*

- *ODE:* $x'(t) - t_f f(x(t), u(t)) = 0$, *a.e. in* T,
- *state constraints:* $G_I(x(t))^\top w_I(t) = 0$, $I = 1,\dots,M$, *a.e. in* T, (2)
$g_I(x(t))^\top w_I(t) \leq -\varepsilon$, $I = 1,\dots,M$, *a.e. in* T, (3)
- *boundary conditions:* $\psi(x(0), x(1)) = 0$,
- *box constraints:* $w_I(t) \geq 0$, $I = 1,\dots,M$, *a.e. in* T,
$u(t) \in \mathcal{U} := \{u \in \mathbb{R}^m \,|\, u_{min} \leq u \leq u_{\max}\}$.

where the function $\psi : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}^{2n_x}$ is given as follows $\psi(x(t_0), x(t_f)) = (R(q(t_0)) - R_0, 0, R(q(t_f)) - R_f, 0)$ where $R(q)$ denotes the position of the barycenter of the last link of the robot and $R_0, R_f \in \mathbb{R}^m$ are given. The vectors u_{min} and u_{max} are also given. The relaxation parameter ε is positive and small. As usual $L^m_\infty(T)$ denotes the Banach space of essentially bounded functions mapping

from T into $\mathbb{R}^m$ and $W_{1,\infty}^{n_x}(T)$ denotes the Banach space of absolutely continuous functions with essentially bounded derivative that map from T into $\mathbb{R}^{n_x}$.

If multiple obstacles are present in the workspace, the *anti-collision constraints* (2)-(3) and the associated control variables w_I are defined for each obstacle.

Depending on the number M of anti-collision constraints, the problem is inherently sparse since the artificial control variables w_I, $I = 1, \ldots, M$, do not enter the dynamics, the boundary conditions, and the objective function of the problem, but only appear linearly in the anti-collision constraints with one-sided coupling through the state.

We attempt to solve the problem (OCP) numerically with a reduced discretization approach [4]. Let us consider the control grid $\mathbb{G}_N := \{t_k = kh \mid k = 0, 1, \ldots, N\}$, which, for simplicity, is chosen equidistantly with the fixed step-size $h = 1/N$. We use B-spline of second order to approximate the control variables:

$$u_h(t; u_0, \ldots, u_N) := \sum_{i=0}^{N} u_i \, B_i(t),$$

$$w_{I,h}(t; w_{I,0}, \ldots, w_{I,N}) := \sum_{i=0}^{N} w_{I,i} \, B_i(t), \quad I = 1, \ldots, M$$

where $(u_0, \ldots, u_N)^\top \in \mathbb{R}^{m(N+1)}$ and $(w_{I,0}, \ldots, w_{I,N})^\top \in \mathbb{R}^{n_I(N+1)}$ are the vector of de Boor points, and B_i, $i = 0, \ldots, N$, denote elementary B-splines. As the elementary B-splines sum up to one for all $t \in T$, the box constraints $u_h(t) \in \mathcal{U}$ are satisfied, if $u_i \in \mathcal{U}$, $i = 0, \ldots, N$. The choice of B-splines is convenient as it is easy to create approximations with prescribed smoothness properties and, even more important, the elementary B-splines B_i have a local support only.

We solve the differential equations for the initial value x_0 and a given t_f by the classical explicit Runge-Kutta method of order 4. The state approximations at the grid points t_k, $k = 0, \ldots, N$ depend on the vector $z := (x_0, u_0, \ldots, u_N, t_f)^\top \in \mathbb{R}^{n_z}$ with $n_z = n_x + (N+1)m + 1$.

Let us define $J(z) := \varphi(x_0, x_N(z), t_f)$, $h(z) := \psi(x_0, x_N(z))$, as well as $\bar{G}_{I,k}(z) := G_I(x_k(z))$ and $\bar{g}_{I,k}(z) := g_I(x_k(z))$ for $I = 1, \ldots, M$, $k = 0, \ldots, N$. With these new notations the discretized form of (OCP) can be formulated as follows

(DOCP): *Minimize* $J(z)$ *with respect to* $z \in \mathbb{R}^{n_z}$ *and*
$w = (w_{1,0}, \ldots, w_{1,N}, \ldots, w_{M,0}, \ldots, w_{M,N})^\top \in \mathbb{R}^{(N+1)\sum_{I=1}^{M} n_I}$

$$\begin{aligned}
\textit{subject to: } & h(z) = 0, \\
& w_{I,k} \geq 0, && I = 1, \ldots, M, \; k = 0, \ldots, N, \\
& \bar{G}_{I,k}(z)^\top w_{I,k} = 0, && I = 1, \ldots, M, \; k = 0, \ldots, N, \\
& \bar{g}_{I,k}(z)^\top w_{I,k} \leq -\varepsilon, && I = 1, \ldots, M, \; k = 0, \ldots, N, \\
& z \in \mathcal{Z} := \{z \in \mathbb{R}^{n_z} \mid z_\ell \leq z \leq z_u\}.
\end{aligned}$$

Herein, $z_\ell \leq z_u$ define box constraints for z, where the settings $\pm\infty$ are permitted if a component of z is not restricted from above or below.

The above nonlinear optimization problem is solved by a sequential quadratic programming (SQP) method [5,9]. As in [13] we use an Armijo type line-search procedure for the augmented Lagrangian function in our implementation. However, (DOCP) contains a lot of constraints: at each time step t_k, $k = 0, \ldots, N$, and for every pair of polyhedra $(P^{(i)}, Q^{(j)})$, four anti-collision constraints are defined (compare (2)-(3)). To overcome this difficulty, we add an active set strategy based on the following observation: the anti-collision constraints are superfluous when the robot is far from the obstacle or moves in the opposite direction. The establishment of the active set strategy is the purpose of the next section.

3 Backface Culling Active Set Strategy

Backface culling comes from computer graphics and consists of working only with visible objects, see [14]. We apply here the same concept to define our active set strategy and develop four criteria to determine which objects are visible.

In this section P, resp. Q, is a polyhedron belonging to the approximation of the robot, resp. obstacle. The first criterion is similar to the *broad phase* in [2,11] and consists of checking if P is far from Q. If this is the case, no collision can occur and the anti-collision constraints are superfluous.

The distance between P and Q is roughly computed by defining the axis-aligned bounding box of each polyhedron. Let (x_i^P, y_i^P, z_i^P), $i = 1, \ldots, s^P$, denote the vertices of P where s^P is the number of vertices, and define

$$
\begin{aligned}
x_m^P &= \min_{i=1,\ldots,s^P} x_i^P, \quad & y_m^P &= \min_{i=1,\ldots,s^P} y_i^P, \quad & z_m^P &= \min_{i=1,\ldots,s^P} z_i^P, \\
x_M^P &= \max_{i=1,\ldots,s^P} x_i^P, \quad & y_M^P &= \max_{i=1,\ldots,s^P} y_i^P, \quad & z_M^P &= \max_{i=1,\ldots,s^P} z_i^P.
\end{aligned}
$$

Then, the tuple $(x_m^P, y_m^P, z_m^P, x_M^P, y_M^P, z_M^P)$ represents the smallest axis-aligned bounding box around P. Similarly the tuple $(x_m^Q, y_m^Q, z_m^Q, x_M^Q, y_M^Q, z_M^Q)$ denotes the smallest bounding box of Q. Let $\delta > 0$ and define slightly bigger boxes:

$$
\begin{aligned}
\mathcal{B}^P &= [x_m^P - \delta,\, x_M^P + \delta] \times [y_m^P - \delta,\, y_M^P + \delta] \times [z_m^P - \delta,\, z_M^P + \delta] \\
\mathcal{B}^Q &= [x_m^Q - \delta,\, x_M^Q + \delta] \times [y_m^Q - \delta,\, y_M^Q + \delta] \times [z_m^Q - \delta,\, z_M^Q + \delta].
\end{aligned}
$$

With these definitions the first criterion reads

Criterion 1. *If $\mathcal{B}^P$ and $\mathcal{B}^Q$ are separated, then P is far from Q and the associated anti-collision constraints are superfluous.*

The vertices of P evolve in time since they belong to the robot. Hence the box $\mathcal{B}^P$ has to be determined at each grid point t_k, $k = 0, \ldots, N$. Because the obstacle does not move, the box $\mathcal{B}^Q$ is computed only once.

Let us assume now that Q is close enough to P and consider the situation depicted in Figure 2: P is moving downwards, v_c indicates the velocity of the

center of gravity and $\tilde{Q}$ is generated by the faces e_1 and e_2 of Q. According to Proposition 1, $\tilde{Q}$ does not collide with P if and only if

$$\exists \tilde{w} > 0, \text{ such that } \begin{pmatrix} A \\ C_{1,2} \end{pmatrix}^{\top} \tilde{w} = 0 \text{ and } \begin{pmatrix} b \\ d_{1,2} \end{pmatrix}^{\top} \tilde{w} < 0, \tag{4}$$

where $C_{1,2}$ is the matrix composed of the first two rows of C and $d_{1,2}$ is the vector composed of the first two components of d.

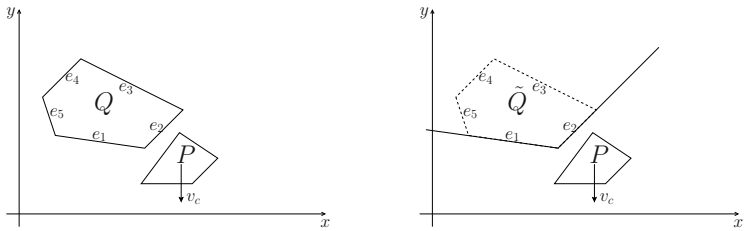


Fig. 1. (a) The polyhedron P is moving downwards. The faces of Q are denoted by $e_1, \ldots, e_5$. (b) The set $\tilde{Q}$ is generated by the faces of Q visible to P.

Suppose now that $\tilde{w}$ exists. By setting $w = (\tilde{w}, 0, 0, 0)$, we obtain:

$$\begin{pmatrix} A \\ C \end{pmatrix}^{\top} w = 0 \text{ and } \begin{pmatrix} b \\ d \end{pmatrix}^{\top} w < 0.$$

Then, Proposition 1 implies that P and Q do not collide. In summary, if no collision occurs between $\tilde{Q}$ and P, then Q and P do not collide. The dimension of $\tilde{w}$ is always smaller than that of w, because the polyhedra are supposed to be compact. Consequently, the problem of finding $\tilde{w}$ is always smaller and there is an advantage in replacing the anti-collision constraints by (4). In (4) only the faces visible to P are taken into consideration. The next criteria concern the determination of the visible faces of Q relative to P.

The faces of Q which are located behind P, are invisible to P. P is always looking towards its velocity, v_c. Hence, all objects located in the lower halfspace $\mathcal{H}$ generated by v_c and S_R, the vertex of P located furthest in the opposite direction of v_c, are behind P. Then, the second criterion reads

Criterion 2. *A face e of Q is invisible to P if $e \subset \mathcal{H} = \{y \in \mathbb{R}^n | v_c^{\top}(y - S_R) < 0\}$.*

An example is given in Figure 2-(a) where the faces e_3 and e_4 satisfy Criterion 2. The case where all faces are located behind P means that P is moving in the opposite direction to Q. In this situation no collision can occur and we have:

Criterion 3. *If all faces of Q are invisible to P according to Criterion 2, then the anti-collision constraints are superfluous.*

Remark 1. If the velocity v_c is zero, Criteria 2 and 3 are not applied.

Not all remaining faces of Q are visible to P. Some of them can be hidden by other faces of Q. This is the case of the face e_5 in Figure 2-(b). The vertex S, which is the closest point of P to e_5, cannot see e_5 because e_5 is hidden by e_1. The face e_5 will be visible to S as soon as S is no more located in the halfspace $\mathcal{H}_5 = \{y \in \mathbb{R}^n \,|\, C_5\, y < d_5\}$. This yields the last criterion

Criterion 4. *The face e_i of Q is invisible to P if $P \subset \mathcal{H}_i = \{y \in \mathbb{R}^n | C_i y < d_i\}$ with C_i, the i^{th} line of C, and d_i, the i^{th} component of d, in the definition of Q.*

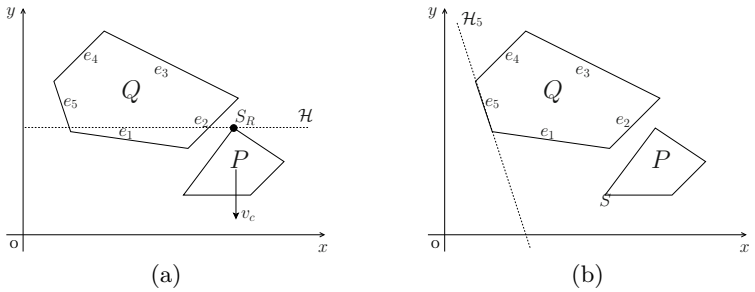


Fig. 2. (a) The faces e_3 and e_4 of Q are located behind P. (b) The face e_5 of Q is invisible to P.

A limit case exists with Criterion 4 when P is included in Q. In that case all faces of Q are invisible to P according to Criterion 4. But in fact all these faces must be considered in the anti-collision constraints. Hence, Criterion 4 must not be applied in this particular case.

Criterion 4 can also be applied to detect which faces of P are visible to Q. Then the anti-collision constraints defined for the pair (P, Q) can be reduced as it was done in (4).

In this section criteria to determine the visible faces of Q were developed, provided Q is visible. All criteria were depending on the position of P which is given by the state variable q. In the next section we show how the backface culling is included in the SQP algorithm to solve (DOCP).

4 Algorithm and Numerical Examples

Let us recall the index transformation that associates to each pair (i, j) the new index I via the formula: $I = (i-1)n_Q + j$ and define the set of indices

$$\mathcal{K} := \{(I, k) \,|\, \text{the polyhedron } Q^{(j)} \text{ is visible to } P^{(i)} \text{ at } t_k\}.$$

$\mathcal{K}$ is determined by applying Criteria 1, 2 and 3. Let us also recall that w_I belongs to $\mathbb{R}^{p_i+q_j}$. The first p_i components of w_I are associated to the faces of

$P^{(i)}$ and the next q_j components are related to the faces of $Q^{(j)}$. Then, let us define the following set of indices for each $(I,k) \in \mathcal{K}$

$$\mathcal{J}_{I,k} := \{c \in \{1,\dots,p_i\} \mid \text{the face } c \text{ of } P^{(i)} \text{ is invisible to } Q^{(j)} \text{ at } t_k\} \cup$$
$$\{c \in \{p_i+1,\dots,p_i+q_j\} \mid \text{the face } c-p_i \text{ of } Q^{(j)} \text{ is invisible to } P^{(i)} \text{ at } t_k\}.$$

This set contains the index of the faces of the pair $(P^{(i)}, Q^{(j)})$ which are invisible at t_k. The invisibility of a face is determined using Criteria 2 and 4.

Backface culling involves considering the anti-collision constraints whose pair of indices (I,k) belongs to $\mathcal{K}$ and write these constraints according to (4). The algorithm to solve $(DOCP)$ is the SQP method presented below in which the backface culling is added as an active set strategy. This means that at each iteration we update the set $\mathcal{K}$ and then build the quadratic problem by considering only the constraints whose pair of indices belongs to $\mathcal{K}$.

Backface Culling Active Set Strategy

(0) Choose $\varepsilon > 0$, $z^{(0)} \in \mathcal{Z}$ and $w^{(0)} \geq 0$.
Determine the sets of indices $\mathcal{K}^{(0)}$ and $\mathcal{J}_{I,k}^{(0)}$ for all $(I,k) \in \mathcal{K}^{(0)}$.
Set $B_0 := I$, the identity matrix and $\ell := 0$.

(1) If $(z^{(\ell)}, w^{(\ell)})$ is a KKT point of the optimization problem, STOP.

(2) Compute a KKT point of the following linear-quadratic optimization problem: *Minimize* $\frac{1}{2} d^\top B_\ell d + J'(z^{(\ell)}) d_z$

with respect to $d = (d_z, d_{w_{I,k}})$, $(I,k) \in \mathcal{K}^{(\ell)}$, *subject to the constraints*

$$\begin{aligned}
h(z^{(\ell)}) + h'(z^{(\ell)}) d_z &= 0, \\
w_{I,k}^{(\ell)} + d_{w_{I,k}} &\geq 0, \qquad (I,k) \in \mathcal{K}^{(\ell)}, \\
\bar{G}_{I,k}(z^{(\ell)})^\top w_{I,k}^{(\ell)} + \bar{G}_{I,k}(z^{(\ell)})^\top d_{w_{I,k}} & \\
+ \bar{G}'_{I,k}(z^{(\ell)})^\top (w_{I,k}^{(\ell)}, d_z) &= 0, \qquad (I,k) \in \mathcal{K}^{(\ell)}, \\
\bar{g}_{I,k}(z^{(\ell)})^\top w_{I,k}^{(\ell)} + \bar{g}_{I,k}(z^{(\ell)})^\top d_{w_{I,k}} & \\
+ \bar{g}'_{I,k}(z^{(\ell)})^\top (w_{I,k}^{(\ell)}, d_z) &\leq -\varepsilon, \qquad (I,k) \in \mathcal{K}^{(\ell)}, \\
z^{(\ell)} + d_z &\in \mathcal{Z}, \\
d_{w_{I,k},c} &= 0, \qquad c \in \mathcal{J}_{I,k}^{(\ell)}, (I,k) \in \mathcal{K}^{(\ell)}.
\end{aligned}$$

Note: The constraints $d_{w_{I,k},c} = 0$ are only included for notational simplicity. In practice these variables are actually eliminated.

(3) Set
$$z^{(\ell+1)} := z^{(\ell)} + d_z^{(\ell)}, \quad w_{I,k}^{(\ell+1)} := w_{I,k}^{(\ell)} + d_{w_{I,k}}^{(\ell)}, \ (I,k) \in \mathcal{K}^{(\ell)}.$$

(4) Update the sets of indices $\mathcal{K}^{(\ell+1)}$ and $\mathcal{J}_{I,k}^{(\ell+1)}$ for $(I,k) \in \mathcal{K}^{(\ell+1)}$ according to Criteria 1 to 4 which depend on $z^{(\ell+1)}$. Update $B_{\ell+1}$ according to BFGS update formulas, set $\ell := \ell + 1$ and go to (1).

At the step (4), if a pair of indices (I,k) newly appears in $\mathcal{K}^{(\ell+1)}$ (i.e. if $(I,k) \in \mathcal{K}^{(\ell+1)} \setminus \mathcal{K}^{(\ell)}$), then the variable $w_{I,k}^{(\ell+1)}$ must be initialized. We choose to take $w_{I,k}^{(\ell+1)}$ as the solution of

$$\min_{w} \bar{g}_{I,k}(z^{(\ell+1)})^{\top} w \text{ such that } \bar{G}_{I,k}(z^{(\ell+1)})^{\top} w = 0,$$
$$w_c > 0, \text{ if } c \notin \mathcal{J}_{I,k}^{(\ell+1)}, \text{ and } w_c = 0, \text{ if } c \in \mathcal{J}_{I,k}^{(\ell+1)}.$$

Thereby, we are assured to satisfy the state constraints as close as possible.

In our first example we consider an obstacle and a robot composed by a socket and 3 links. A load is fixed at the end of the last link. A complete description of the example is given in [6]. In this example, the collision avoidance needs to be applied only between the load and the obstacle. The obstacle is always close to the load. Consequently, the number of state contraints is not reduced with the backface culling. For this example, we take 21 control grid points and $\varepsilon = 10^{-5}$. In Figure 3 the visible faces of the obstacle are in white. We can observe that only 3 faces of the obstacle are visible. The computational time is $52\,s$. If we do not use the backface culling, the computational time is equal to $3\,min\,44$. So, with about half of the unknowns, the code runs about four times faster.

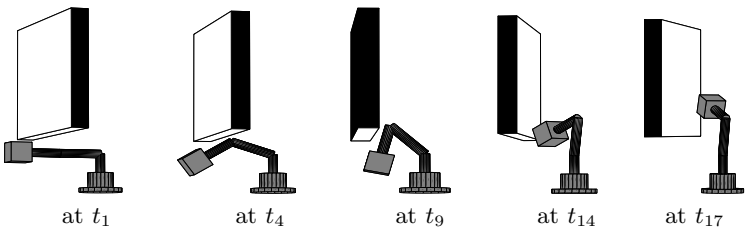


Fig. 3. Snapshots of the motion of the robot avoiding an obstacle. The visible faces of the obstacle are in white.

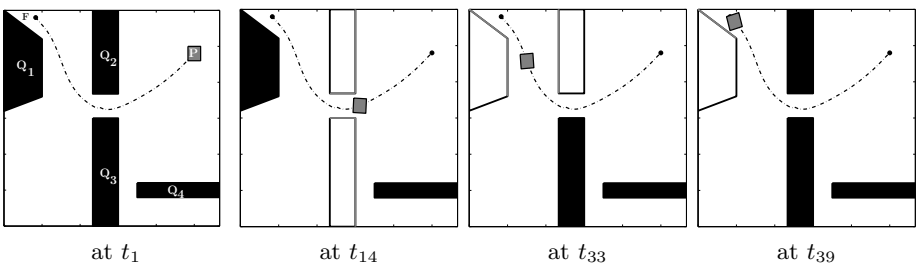


Fig. 4. Snapshots of the motion of the robot P moving to F and avoiding four obstacles. The visible obstacles are in white and their visible faces in gray.

The second example is in 2 dimensions and uses all criteria of the backface culling. The robot P is a square and 4 obstacles, Q_1 to Q_4, are present in the workspace. We take for this example 42 control grid points and $\varepsilon = 10^{-2}$. Snapshots of the motion of the robot, which must reach the point F, are given in Figure 4. The visible obstacles are in white and their visible faces in gray. The computational time is equal to $36\,min\,50$ when no backface culling is used. With the backface culling strategy the CPU time is $27\,s$. Hence, Criteria 1 to 4 induce a large decrease in the computational time.

Acknowledgments. This work has been supported by the DFG Research Center MATHEON Mathematics for key technologies.

References

1. Berkovitz, L.D.: Convexity and Optimization in $\mathbb{R}^n$. John Wiley & Sons, New York (2001)
2. Cohen, J.D., Lin, M.C., Manocha, D., Ponamgi, M.K.: I-collide: An Interactive and Exact Collision Detection System for Large-Scaled Environments. In: Symposium on Interactive 3D Graphics, pp. 189–196. ACM Siggraph (1995)
3. Diehl, M., Bock, H.G., Diedam, H., Wieber, P.B.: Fast Direct Multiple Shooting Algorithms for Optimal Robot Control. In: Fast Motions in Biomechanics and Robotics, pp. 65–94. Springer (2005)
4. Gerdts, M.: Direct Shooting Method for the Numerical Solution of Higher-Index DAE Optimal Control Problems. Journal of Optimization Theory and Applications 117, 267–294 (2003)
5. Gerdts, M.: Optimal Control of Ordinary Differential Equations and Differential-Algebraic Equations. Fakultät für Mathematik und Physik, Universität Bayreuth, Bayreuth (2006)
6. Gerdts, M., Henrion, R., Hömberg, D., Landry, C.: Path Planning and Collision Avoidance for Robots. Numer. Algebra Control Optim. 3, 437–463 (2012)
7. Gilbert, E.G., Hong, S.M.: A New Algorithm for Detecting the Collision of Moving Objects. In: IEEE Proc. Int. Conf. on Robotics and Automation, vol. 1, pp. 8–14 (1989)
8. Gilbert, E.G., Johnson, D.W.: Distance Functions and their Application to Robot Path Planning in the Presence of Obstacles. IEEE Journal of Robotics and Automation 1, 21–30 (1985)
9. Gill, P.E., Murray, W., Saunders, M.A.: SNOPT: An SQP Algorithm for Large-Scale Constrained Optimization. SIAM Rev. 47, 99–131 (2005)
10. LaValle, S.M.: Planning Algorithms. Cambridge University Press, Cambridge (2006)
11. Mirtich, B.: V-Clip: Fast and Robust Polyhedral Collision Detection. Mitsubishi Electronics Research Laboratory (1997)
12. Murray, R.M., Sastry, S.S., Li, Z.: A Mathematical Introduction to Robotic Manipulation. CRC Press, Inc., Boca Raton (1994)
13. Schittkowski, K.: On the Convergence of a Sequential Quadratic Programming Method with an Augmented Lagrangean Line Search Function. In: Mathematische Operationsforschung und Statistik. Series Optimization, vol. 14, pp. 197–216 (1983)
14. Vaněček Jr., G.: Back-Face Culling Applied to Collision Detection of Polyhedra. Journal of Visualization and Computer Animation 5, 55–63 (1994)

Regularized Extremal Shift in Problems of Stable Control

Vyacheslav Maksimov*

Ural Federal University and
Institute of Mathematics and Mechanics,
Ural Branch of the Russian Academy of Sciences,
Ekaterinburg, 620990 Russia
maksimov@imm.uran.ru

Abstract. We discuss a technical approach, based on the method of regularized extremal shift (RES), intended to help solve problems of stable control of uncertain dynamical systems. Our goal is to demonstrate the essence and abilities of the RES technique; for this purpose we construct feedback controller for approximate tracking a prescribed trajectory of an inaccurately observed system described by a parabolic equation. The controller is "resource-saving" in a sense that control resource spent for approximate tracking do not exceed those needed for tracking in an "ideal" situation where the current values of the input disturbance are fully observable.

Keywords: parabolic equation, stable control, extremal shift.

1 Introduction

In the present work, the problem of tracking a solution of a system with distributed parameters is discussed. The essence of this problem can be formulated in the following way. A parabolic equation is considered on a given time interval $T = [t_0, \vartheta]$, $\vartheta < +\infty$. The solution of this equation $w(\cdot) = w(\cdot; v(\cdot))$ depends on a time-varying control $v = v(\cdot)$. This solution is inaccurately measured at frequent enough time moments. It is required to organize a control process for the equation by the feedback principle in such a way that it is possible to preserve given properties of the solution. The quality of the solution constructed is estimated by the distance from a given (prescribed, standard) solution $x(\cdot)$. The latter is a solution of the parabolic equation generated by some input $u = u(\cdot)$. The problem in question is treated as the problem of constructing a control $v = v(\cdot)$ providing the retention of the trajectory $w(\cdot) = w(\cdot; v(\cdot))$ nearby $x(\cdot)$. This is the conceptual statement of the control problem under consideration.

* This work was supported in part by the Russian Foundation for Basic Research (No. 11-01-12112-ofi-m), by the Program on Basic Research of the Presidium of the Russian Acad. Sci. (No. 12-P-1-1012) and by the Program for support of leading scientific schools of Russia (No. 6512.2012.1).

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 112–121, 2013.

2 Problem Statement

Let H and V be real Hilbert spaces. The space V is a dense subspace of H and $V \subset H \subset V^*$ algebraically and topologically, $(\cdot,\cdot)$ stands for the inner product in H, $\langle\cdot,\cdot\rangle$ stands for the duality relation between V and V^*. We consider a system Σ which is described by the parabolic equation

$$\dot{w}(t) + Aw(t) = Bv(t) + f(t), \quad \text{for a. a.} \quad t \in T, \quad w(t_0) = w_0. \tag{1}$$

Here $A : V \to V^*$ is a linear continuous ($A \in \mathcal{L}(V; V^*)$) and symmetrical operator satisfying (for some $c_* > 0$ and real ω) the coercitivity condition

$$\langle Aw, w\rangle + \omega|w|_H^2 \geq c_*|w|_V^2 \quad \forall y \in V, \tag{2}$$

U is a Hilbert space, $f \in L_2(T; H)$ is a given function, $|\cdot|_H$, $|\cdot|_U$ and $|\cdot|_V$ stand for the norms in H, U and V, respectively, $B : U \to H$ is a linear continuous operator ($B \in \mathcal{L}(U; H)$). Let the following condition be fulfilled.

Condition 1. Operator B is invertible.

Let $w(t_0) = w_0 \in D(A_H)$, where $D(A_H) = \{w \in V : A_H \in H\}$. It is known that under such conditions, for any $v(\cdot) \in L_2(T; U)$, there exists a unique solution $w(\cdot) = w(\cdot; t_0, w_0, v(\cdot))$ of equation (1) with the following properties [1]: $w(\cdot) \in W(T) = W^{1,2}(T; H) \cap L_2(T; V)$. Here, $W^{1,2}(T; H) = \{w(\cdot) \in L_2(T; H) : \dot{w}(\cdot) \in L_2(T; H)\}$, the derivative $\dot{w}(\cdot)$ is understood in the sense of distributions.

Assume that along with equation (1) we have another equation of the same form:

$$\dot{x}(t) + Ax(t) = Bu(t) + f(t) \quad \text{for a.a.} \quad t \in T \tag{3}$$

with an initial state $x(t_0) = x_0 \in D(A_H)$. This equation (in what follows, we call it reference) is subject to the action of some reference control $u(\cdot) \in L_2(T; U)$. The reference control as well as the corresponding solution $x(\cdot) = x(\cdot; t_0, x_0, u(\cdot))$ of equation (3) are a priori unknown. At discrete, frequent enough, time moments $\tau_i \in \Delta = \{\tau_i\}_{i=0}^m$ ($\tau_0 = t_0$, $\tau_m = \vartheta$, $\tau_{i+1} = \tau_i + \delta$), the states $w(\tau_i) = w(\tau_i; t_0, w_0, v(\cdot))$ of equation (1) as well as the states $x(\tau_i) = x(\tau_i; t_0, x_0, u(\cdot))$ of reference equation (3) are measured. The states $w(\tau_i)$ are measured with an error. The results of measurements are elements $\xi_i^h \in H$ satisfying the inequalities

$$|w(\tau_i) - \xi_i^h|_H \leq h, \quad i \in [1 : m-1]. \tag{4}$$

Here, the value $h \in (0, 1)$ is the measurement accuracy. It is required to design an algorithm for forming the control $v = v^h(\cdot)$ in equation (1) allowing us to track the solution $x(\cdot)$ of equation (3) by the solution $w(\cdot)$ of equation (1). Thus, we consider the problem consisting in constructing an algorithm, which (on the basis of current measurements of the values $w(\tau_i)$ and $x(\tau_i)$) forms in real time mode (by the feedback principle) the control $v = v^h(\cdot)$ in the right-hand part of inequality (1) such that the deviation of $w(\cdot) = w(\cdot; t_0, w_0, v^h(\cdot))$ from $x(\cdot) = x(\cdot; t_0, x_0, u(\cdot))$ in metric of the space $C(T; H) \cap L_2(T; V)$ is small if the measurement accuracy h is small enough. We also want the constructed algorithm to be resource-saving. This means that the resources of the synthetic

control $v = v^h(\cdot)$ (i.e., the value $\int_{t_0}^{\vartheta} |v^h(\tau)|_U^2 \, d\tau$ should exceed the resources of the reference control by a small value depending on the measurement accuracy h. This value tends to zero as h tends to zero. Thus, we require the validity of the inequality

$$\int_{t_0}^{\vartheta} |v^h(\tau)|_U^2 \, d\tau \le \int_{t_0}^{\vartheta} |u(\tau)|_U^2 \, d\tau + \varphi(h), \tag{5}$$

where $\varphi(h) \to 0$ as $h \to 0$.

In the case when the reference control u as well as the control v in inequality (1) are subject to instantaneous constraints ($u \in P$, $v \in P$, where $P \subset U$ is a given bounded and closed set), the problem above can be solved by means of the method of extremal shift [2]. Namely, if the control $v = v^h(\cdot)$ in the right-hand part of (1) is calculated by the formula

$$v^h(t) = v(\tau_i, \xi_i^h, x(\tau_i)) = \arg\min\{(x(\tau_i) - \xi_i^h, Bv) : v \in P\} \text{ for } t \in [\tau_i, \tau_{i+1}), \tag{6}$$

then, as it follows from [3], for any $\varepsilon > 0$ one can find numbers $h_1 > 0$ and $\delta_1 > 0$ such that the inequality

$$\sup_{t \in T} |w(t; t_0, w_0, v^h(\cdot)) - x(t; t_0, x_0, u(\cdot))|_H \le \varepsilon$$

is fulfilled if $h \in (0, h_1)$ and $\delta \in (0, \delta_1)$. The last inequality is valid for any reference control, i.e., for any Lebesgue measurable function $u(t) \in P$ for almost all $t \in T$. Here and below, we assume that $\omega > 0$ and $w_0 \in D \subset V$, where D is a bounded set,

$$|w_0 - x_0|_H \le h. \tag{7}$$

Thus, the method of extremal shift allows us to solve the problem of tracking the solution of the reference equation under instantaneous constraints on the controls ($v, u \in P$). In the present paper, we assume that any function from the space $L_2(T; U)$ can be the admissible control (both reference, $u(\cdot)$, and real, $v(\cdot)$). No additional information on the functions $v(\cdot)$ and $u(\cdot)$ is required. We construct a corresponding modification of the method of extremal shift, using, according to [4–9], the idea of its local regularization. Along with measuring the phase states at discrete time moments (see (4)), we also consider the case of "continuous" measuring of the states $x(t)$ and $w(t)$. Namely, it is assumed that, at every time $t \in T$, the phase states of equations (1) and (3) are measured; as a result, we have functions $\xi^h(t) \in H$ with the properties

$$|\xi^h(t) - w(t)|_H \le h, \quad t \in T. \tag{8}$$

The functions $\xi^h(t)$, $t \in T$, are Lebesgue measurable.

In control theory for distributed systems, a linear quadratic control problem (LQP) is widely known. Its solution methods have been studied rather well (see, for example, [11,12]). This problem consists in the minimization of some

quadratic functional depending on a phase trajectory and control (for example, in the minimization of the deviations in L_2-norm from a reference control and state trajectory). The problem in question, which is in essence close to LQP, has, at the same time, several distinctive features. Among them, at the first turn, it is worth while noticing the following two features. Firstly, an LQP solution, as a rule, does not guarantee that we find a control generating a trajectory that is close to the reference trajectory in uniform metric. In addition, some apriori information on the reference control is rather often required. The second distinction is in the essence of solving methods. Namely, the method suggested in the present paper is based on constructions of the well-known in the theory of guaranteed control principle of extremal shift.

3 Control Algorithm. Case of Continuous Measuring of Solutions

First, we consider the case of "continuous" measuring of solutions of equations (1) and (3). In this case, inequalities (8) are valid (for simplicity, we set $\xi^h(t_0) = w_0$). The problem consists in designing a rule forming (by the feedback principle) the control $v = v(t, \xi^h(t), w(t))$. Fix a function $\alpha = \alpha(h) : (0,1) \to (0,1)$. Let the control $v^{\alpha,h}(t)$ in equation (1) be defined by the formula

$$v = \tilde{v}^{\alpha,h}(t) = v^{\alpha,h}(t) + \tilde{v}^h(t), \tag{9}$$

where

$$\tilde{v}^h(t) = cB^{-1}(x(t) - \xi^h(t)), \quad v^{\alpha,h}(t) = \alpha^{-1}B^*(x(t) - \xi^h(t)). \tag{10}$$

Here, B^* denotes the adjoint operator, $c = \text{const} > 2\omega$. Thus, we obtain system (1), (3); i.e., we have the pair of equations

$$\dot{x}(t) + Ax(t) = Bu(t) + f(t),$$

$$\dot{w}^{\alpha,h}(t) + Aw^{\alpha,h}(t) = \alpha^{-1}BB^*(x(t) - \xi^h(t)) + c(x(t) - \xi^h(t)) + f(t)$$

with the initial condition $x(t_0) = x_0$, $w^{\alpha,h}(t_0) = w_0$. Here, we denote by $w^{\alpha,h}(\cdot)$ the solution of equation (1) corresponding to the function $v = v^{\alpha,h}(\cdot)$ of form (9).

The second formula in (10) is an analog of relation (6). If the constraint in the form of the set P is absent then the application of formula (6) for calculating the control v is impossible, since in this case it is required to solve the problem of minimization of the linear functional $l_i(u) = (\xi_i^h - y(\tau_i), Bu)$ over the whole space U. It is natural to replace this problem by a new regularized problem with a smoothing functional of the form $\alpha(h)|v|_U^2$, i.e., to replace problem (6) by the problem of finding the function $v^{\alpha,h}(t)$ by the rule

$$v^{\alpha,h}(t) = \arg\min\{\alpha|v|_U^2 - 2(B^*(x(t) - \xi^h(t)), v)_U : v \in U\}.$$

Formula (10) provides the solution of the new problem. Thus, to calculate $v^{\alpha,h}(t)$, we realize the regularization of the method of extremal shift by means of the method of smoothing functional, which is known in the theory of ill-posed problems.

Theorem 1. *Let $\alpha = \alpha(h) \to 0$. Then the following inequalities*

$$|x(t) - w^{\alpha,h}(t)|_H^2 + 2c \int_{t_0}^{t} |x(\tau) - w^{\alpha,h}(\tau)|_V^2 \, d\tau \le d_0(h + \alpha(h)), \quad t \in T, \tag{11}$$

$$\int_{t_0}^{\vartheta} |\tilde{v}^{\alpha,h}(\tau)|_U^2 \, d\tau \le \int_{t_0}^{\vartheta} |u(\tau)|_U^2 \, d\tau + d_*(h\alpha^{-2}(h) + h^{1/2} + \alpha^{1/2}(h)) \tag{12}$$

are fulfilled. Here, $d_0, d_* = \text{const} > 0$ are constants, which do not depend on $h \in (0, 1)$.

Proof. Due to (10), it holds that $|v^{\alpha,h}(t)|_U^2 \le 2b^2\alpha^{-2}(h^2 + |\mu_{\alpha,h}(t)|_H^2)$, $t \in T$, where $\mu_{\alpha,h}(t) = x(t) - w^{\alpha,h}(t)$, $b = |B^*|_{L(H;U)}$ is the norm of the linear operator $B^* \in L(H; U)$. In this case, we have

$$\int_{t_0}^{t} |v^{\alpha,h}(\tau)|_U^2 d\tau \le 2b^2\alpha^{-2}\varrho_h(t) + c_1h^2\alpha^{-2}, \quad \varrho_h(t) = \int_{t_0}^{t} |\mu_{\alpha,h}(\tau)|_H^2 d\tau. \tag{13}$$

Due to coercivity condition (2), we obtain,

$$\dot{\varepsilon}_h(t) \le -2(v^{\alpha,h}(t), B^*(x(t) - \xi^h(t)))_U + \alpha|v^{\alpha,h}(t)|_U^2 + 2(u(t), B^*(x(t) - \xi^h(t)))_U -$$
$$- \alpha|u(t)|_U^2 + 2bh\{|u(t)|_U + |v^{\alpha,h}(t)|_U\} + (2\omega - c)|\mu_{\alpha,h}(t)|_H^2 + 2ch^2, \tag{14}$$

where $\varepsilon_h(t) = |\mu_{\alpha,h}(t)|_H^2 + 2c_* \int_{t_0}^{t} |\mu_{\alpha,h}(\tau)|_V^2 \, d\tau + \alpha \int_{t_0}^{t} \{|v^{\alpha,h}(\tau)|_U^2 - |u(\tau)|_U^2\} \, d\tau$.
From (14) and (10), it follows that

$$\varepsilon_h(t) \le \varepsilon_h(t_0) + c_2 + \int_{t_0}^{t} 2bh\{|u(\tau)|_U + |v^{\alpha,h}(\tau)|_U\} d\tau + (2\omega - c)\varrho_h(t). \tag{15}$$

From (13), (15), and the inequality $2\omega - c < 0$, we derive

$$\varepsilon_h(t) \le c_3(h + h^3\alpha^{-2}) + c_4(h\alpha^{-2} + h)\varrho_h(t). \tag{16}$$

Therefore, from (16) we get the bound

$$|\mu_{\alpha,h}(t)|_H^2 \le c_5(h + \alpha + h^3\alpha^{-2}) + c_4(h\alpha^{-2} + h)\varrho_h(t). \tag{17}$$

By the Gronwall inequality and (17), we obtain

$$|\mu_{\alpha,h}(t)|_H^2 \le c_5(h + \alpha + h^3\alpha^{-2}) \exp\{c_4(t - t_0)(h\alpha^{-2} + h)\}. \tag{18}$$

Note that $h\alpha^{-2} \le \text{const}$. Then

$$|\mu_{\alpha,h}(t)|_H^2 \le c_6(h+\alpha). \tag{19}$$

From (16) and (19) we derive

$$\varepsilon_h(t) \le c_3(h+h^3\alpha^{-2}) + c_7(h\alpha^{-2}+1)(h+\alpha) \le c_9(h+\alpha). \tag{20}$$

Relation (11) follows from (10). Let us verify (12). By virtue of inequality (19), from (16) we obtain

$$\varepsilon_h(t) \le c_8\{h+h^3\alpha^{-2}+(h\alpha^{-2}+h)(h+\alpha)\} \le c_9\{h+h^2\alpha^{-2}+h\alpha^{-1}\}. \tag{21}$$

Using (21), we get for $t \in T$

$$\int_{t_0}^{t} |v^{\alpha,h}(\tau)|_U^2 \, d\tau \le \int_{t_0}^{t} |u(\tau)|_U^2 \, d\tau + c_{10}h\alpha^{-2}, \tag{22}$$

$$|\tilde{v}^h(t)|_U \le c_{11}(h+\alpha)^{1/2}. \tag{23}$$

Relation (12) follows from (22) and (23). The theorem is proved.

4 Control Algorithm. Case of Discrete Measuring of Solutions

Let us describe the algorithm for solving the problem in the case of discrete measuring of phase states. In this case, we assume that relations (4) are fulfilled. Let $l(\cdot) : W^{1,2}(T;H) \cap L_2(T;V) \to \mathbb{R}^+$, $l(y(\cdot)) = |y(\cdot)|_{C(T;H)} + |\dot{y}(\cdot)|_{L_2(T;H)} + |y(\cdot)|_{L_2(T;V)}$. In a standard way, we establish the validity of the following lemma.

Lemma 1. *There exists a number* $K = K(\omega, D, c_*, |B|_{L(U;H)})$ *such that the inequality* $l(x(\cdot;t_0,x,u(\cdot))) \le K(1+|u(\cdot)|_{L_2(T;U)})$ *is fulfilled uniformly for any* $x \in D$ *and* $u(\cdot) \in L_2(T;U)$.

Let a family of partitions $\Delta_h = \{\tau_{h,i}\}_{i=0}^{m_h}$, $\tau_{h,0} = t_0$, $\tau_{h,m_h} = \vartheta$, $\tau_{h,i+1} = \tau_{h,i} + \delta(h)$ and a function $\alpha(h) : (0,1) \to (0,1)$ be fixed. First, before the moment t_0, the value h and the partition Δ_h of the interval T are chosen and fixed. The work of the algorithm is decomposed into $m-1$ ($m = m_h$) identical steps. At the ith step, which is carried out on the time interval $\delta_i = [\tau_i, \tau_{i+1})$, $\tau_i = \tau_{h,i}$, the following sequence of actions is fulfilled. First, at the moment τ_i, the element

$$v_i^h = \alpha^{-1}B^*(x(\tau_i) - \xi_i^h) \tag{24}$$

is calculated. Then, the control defined by the formula

$$v = \tilde{v}^h(t) = v_i^h + cB^{-1}(x(\tau_i) - \xi_i^h), \quad t \in \delta_i, \tag{25}$$

is fed onto the input of equation (1), where $c = \text{const} > 2\omega$. Under the action of this control, instead of the state $w^h(\tau_i) = w^h(\tau_i; \tau_{i-1}, w^h(\tau_{i-1}), v^h_{i-1})$, the state $w^h(\tau_{i+1}) = w^h(\tau_{i+1}; \tau_i, w^h(\tau_i), v^h_i)$ is realized. The work of the algorithm stops at the time moment ϑ.

Let the family of partitions Δ_h of the time interval T and the function $\alpha(h)$ have the following properties:

$$h\delta^{-1}(h) \le C_1, \quad \delta(h)\alpha^{-2}(h) \to 0, \quad h\alpha^{-1}(h) \to 0, \tag{26}$$

$$\alpha(h) \to 0, \ \delta(h) \to 0 \text{ as } h \to 0+.$$

Here $C_1 = \text{const} > 0$ is a constant, which does not depend on h.

Theorem 2. *Uniformly with respect to $h \in (0,1)$, the inequalities*

$$\lambda_h(t) \equiv |x(t) - w^h(t)|^2_H + 2c \int_{t_0}^{t} |x(\tau) - w^h(\tau)|^2_V \, d\tau \le d_1(h + \alpha + \delta) \quad \forall t \in T, \tag{27}$$

$$\int_{t_0}^{\vartheta} |\tilde{v}^h(\tau)|^2_U \, d\tau \le \int_{t_0}^{\vartheta} |u(\tau)|^2_U \, d\tau + d_2(h\alpha^{-1} + \delta\alpha^{-2}) + d_3(h + \alpha + \delta)^{1/2} \tag{28}$$

are true. Here, d_1, d_3 ($d_1 - d_3 = \text{const} > 0$) are constants, which do not depend on h, $\alpha = \alpha(h)$, and $\delta = \delta(h)$.

Proof. First, we verify inequality (27). Using the invertibility of the operator B as well as coercitivity condition (2), we obtain for a.a. $t \in \delta_i$ the inequality

$$0.5\frac{d|\mu^h(t)|^2_H}{dt} + c_*|\mu^h(t)|^2_V - \omega|\mu^h(t)|^2_H \le (B(u(t) - v^h(t)) - c(x(\tau_i) - \xi^h_i), \mu^h(t))_U,$$

where $\mu^h(t) = x(t) - w^h(t)$ for $t \in T$, $v^h(t) = v^h_i$ for $t \in \delta_i$. From the inequality

$$c(\xi^h_i - x(\tau_i), \mu^h(t)) \le -0.5c|\mu^h(t)|^2_H + 4ch^2 + 8c(t - \tau_i) \int_{\tau_i}^{t} \{|\dot{x}(\tau)|^2_H + |\dot{w}^h(\tau)|^2_H\} \, d\tau,$$

we have for a.a. $t \in \delta_i$

$$(B(u(t) - v^h(t)) - c(x(\tau_i) - \xi^h_i), \mu^h(t))_U \le$$

$$(B(u(t) - v^h(t)), x(\tau_i) - \xi^h_i)_U + \varrho_i(t,h) + \chi_i(t,h) - 0.5c|\mu^h(t)|^2_H.$$

Here,

$$\chi_i(t,h) = 4ch^2 + 8c(t - \tau_i) \int_{\tau_i}^{t} \{|\dot{x}(\tau)|^2_H + |\dot{w}^h(\tau)|^2_H\} \, d\tau,$$

$$\varrho_i(t,h) = b(|u(t)|_U + |v^h(t)|_U)(h + \int_{\tau_i}^{t} \{|\dot{w}^h(\tau)|_H + |\dot{x}(\tau)|_H\}\, d\tau).$$

For a.a. $t \in \delta_i$, we deduce that

$$\dot{\varepsilon}^h(t) \le -2(v^h(t), B^*(x(\tau_i) - \xi_i^h))_U + \alpha|v^h(t)|_U^2 +$$

$$+ 2(u(t), B^*(x(\tau_i) - \xi_i^h))_U - \alpha|u(t)|_U^2 + 2\varrho_i(t,h) + 2\chi_i(t,h) + (2\omega - c)|\mu^h(t)|_H^2,$$

where $\varepsilon^h(t) = |\mu^h(t)|_H^2 + 2c_* \int_{t_0}^{t} |\mu^h(\tau)|_V^2\, d\tau + \alpha \int_{t_0}^{t} \{|v^h(\tau)|_U^2 - |u(\tau)|_U^2\}\, d\tau$. Therefore, by virtue of the rule of forming the control $\tilde{v}^h(\cdot)$ (see (24) and (25)), we conclude that, for a.a. $t \in \delta_i$,

$$\varepsilon^h(t) \le \varepsilon^h(\tau_i) + c_1 h^2 + c_2\delta \int_{\tau_i}^{t} \{|u(\tau)|_U^2 + |v^h(\tau)|_U^2\}\, d\tau + \tag{29}$$

$$+ c_3\delta \int_{\tau_i}^{t} \{|\dot{w}^h(\tau)|_H^2 + |\dot{x}(\tau)|_H^2\} d\tau + (2\omega - c) \int_{\tau_i}^{t} |\mu^h(\tau)|_H^2\, d\tau.$$

Summing the right-hand and left-hand parts of (29) over i and taking into account Lemma 1, we obtain for $t \in T$

$$\varepsilon^h(t) \le \varepsilon^h(t_0) + c_4 h^2 \delta^{-1} + c_6\delta + c_7\gamma_{h,\delta}(t). \tag{30}$$

Here, $\gamma_{h,\delta}(t) = \delta^2 \sum_{j=0}^{i(t)} |v_j^h|_U^2$. Using (4) and the rule of forming v_i^h (see (24)), we get

$$|v_i^h|_U^2 \le 2b^2(\varrho_i^h + h^2)\alpha^{-2} \le c_8(\varrho_i^h + h^2)\alpha^{-2}, \tag{31}$$

where $\varrho_i^h = |x(\tau_i) - w^h(\tau_i)|_H^2$. Due to (7), we have $\varepsilon^h(t_0) \le h^2$. Therefore, we derive from (30) the estimate

$$\lambda_h(t) \le c_9(\delta + h^2\delta^{-1} + \alpha + \gamma_{h,\delta}(t)). \tag{32}$$

Note that $\varrho_i^h \le \lambda_i^h$, where $\lambda_j^h = \lambda_h(\tau_j)$. Therefore, for $t \in [\tau_i, \tau_{i+1}]$, due to (31), the inequality

$$\gamma_{h,\delta}(t) \le c_8\delta^2 \sum_{j=0}^{i(t)} (\lambda_j^h + h^2)\alpha^{-2} \tag{33}$$

is valid. Consequently, (32) implies the inequality

$$\lambda_i^h \le c_{10}(\delta + h^2\delta^{-1} + \alpha) + c_{11}\delta h^2\alpha^{-2} + c_{12}\delta^2\alpha^{-2} \sum_{j=0}^{i} \lambda_j^h. \tag{34}$$

By the discrete Gronwall inequality [10], (34), and the inequalities $h\delta^{-1}(h) \le C_1$, $\delta\alpha^{-2}(h) \le C_2$ as $h \to 0$ (see (26)), we have

$$\lambda_i^h \le c_{14}(h + \delta + \alpha), \quad i \in [0:m].$$

This and (33) imply $\gamma_{h,\delta}(t) \le c_{15}(h+\delta+\alpha)$ $t \in T$. Moreover, from the last inequality and (32), we have $\lambda_h(t) \le c_{16}(\delta + h^2\delta^{-1} + \alpha + \gamma_{h,\delta}(t)) \le c_{17}(h+\delta+\alpha)$. Relation (27) follows from the last inequality. The proof of (28) is similar to the proof of (12). The theorem is proved.

It follows from Theorems 1 and 2 that the algorithms presented above are resource-saving.

5 Example

The second algorithm was tested. The parabolic equation

$$w_t(t,\eta) - \partial^2 w(t,\eta)/\partial\eta^2 = v(t,\eta), \quad \eta \in [0,1] \tag{35}$$

with the boundary $w(t,0) = w(t,1)$, $t \in T = [0,2]$ and initial $w(0,\eta) = 0$, $\eta \in [0,1]$ conditions was considered. The reference equation (see (3)) was of the form

$$x_t(t,\eta) - \partial^2 x(t,\eta)/\partial\eta^2 = u(t,\eta), \quad \eta \in [0,1] \tag{36}$$

$$x(t,0) = x(t,1) = 0, \quad t \in T, \qquad x(0,\eta) = 0, \quad \eta \in [0,1].$$

Equations (35) and (36) were solved by the grid method [10]. The grid $\{\eta_j\}_{j=0}^n$, $\eta_0 = 0$, $\eta_n = 1$ with the step $\gamma_N = 1/n$ was taken on the interval $[0,1]$. The control $v = v^h(t,\eta)$ in the right-hand part of (35) was calculated by formula (25) taking the form

$$v^h(t,\eta_j) = (\alpha^{-1} + c)(x(\tau_i,\eta_j) - \xi_i^h(\eta_j)), \quad t \in [\tau_i, \tau_{i+1}), \quad j \in [0:n].$$

During the experiment, we assumed that $\xi_i^h(\eta_j) = w(\tau_i,\eta_j) + h$. In figs. 1–4, the cross-sections of the trajectories x (dashed line) and w (solid line) by the hyperplane $\eta = 0,4$ are presented, as well as the variations of the values $p(t) = \int_0^t |v^h(\tau)|^2_{L_2([0,1])}\, d\tau$ (solid line) and $q(t) = \int_0^t |u^h(\tau)|^2_{L_2([0,1])}\, d\tau$ (dashed line). Figs. 1 and 3 correspond to the case $\delta = 2/m_h$, $m_h = 800$, $n = 10$, $h = 0.05$;

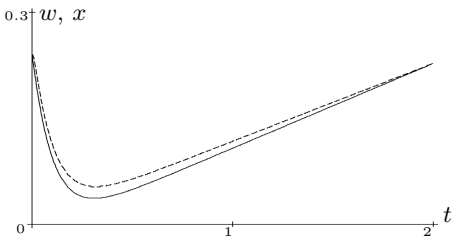


Fig. 1.

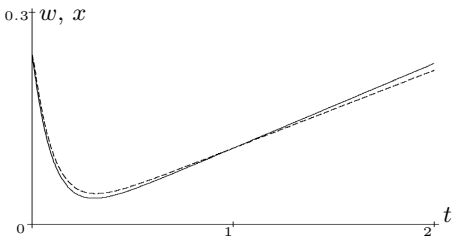


Fig. 2.

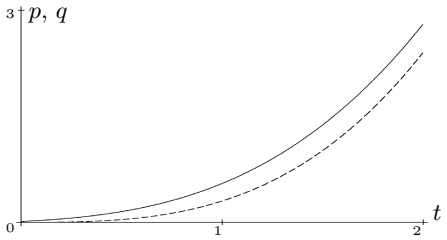


Fig. 3.

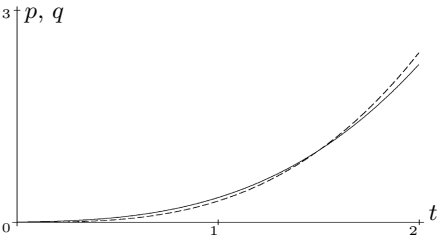


Fig. 4.

figs. 2 and 4, to the case $\delta = 2/m_h$, $m_h = 800$, $n = 10$, $h = 0.01$. As the numerical experiment showed,

$$\max_{\substack{i\in[0:m_h]\\ j\in[0:n]}} |x(\tau_i,\eta_j) - w^h(\tau_i,\eta_j)| = \begin{cases} 0,01926, & \text{in the first case} \\ 0,00972, & \text{in the second case.} \end{cases}$$

References

1. Bensoussan, A., Da Prato, G., Delfour, M., Mitter, S.: Representation and Control of Infinite Dimensional Systems. Birkhauser, Boston (1992)
2. Krasovskii, N.N., Subbotin, A.I.: Game-theoretical Control Problems. Springer, New-York (1988)
3. Vaysburd, I.F., Osipov, Y.S.: Differential pursuit game for systems with distributed parameters. Prikl. Mat. Mech. 39(5), 772–779 (1975) (in Russian)
4. Osipov, Y.S., Kryazhimskii, A.V.: Inverse Problems for Ordinary Differential Equations: Dynamical Solutions. Gordon and Breach, London (1995)
5. Maksimov, V.I.: Dynamical Inverse Problems of Distributed Systems. VSP, Utrecht (2002)
6. Osipov, Y.S., Kryazhimskii, A.V., Maksimov, V.I.: Methods of Dinamical Reconstruction of Inputs of Controlled Systems. Ekaterinburg (2011) (in Russian)
7. Maksimov, V., Pandolfi, L.: Dynamical reconstruction of inputs for constraction semigroup systems: the boudary input case. J. Optim. Theor. Appl. 103, 401–420 (1999)
8. Maksimov, V., Pandolfi, L.: On reconstruction of unbounded controls in nonlinear dynamical systems. Prikl. Mat. Mech. 65(3), 35–42 (2001) (in Russian)
9. Maksimov, V., Tröltzsch, F.: Dynamical state and control reconstruction for a phase field model. Dynamics of Continuous, Discrete and Impulsive Systems. A: Mathematical Analysis 13(3-4), 419–444 (2006)
10. Samarskii, A.A.: Introduction to the Theory of Difference Schemes. Nauka, Moscow (1971) (in Russian)
11. Lasiecka, I., Triggiani, R.: Control theory for partial differential equations: continuous and approximation theories. I. Abstract parabolic systems, Cambridge (2000)
12. Curtain, R.F., Pritchard, A.J.: Infinite dimensional linear systems theory. LNCIS, vol. 8. Springer (1978)

New Necessary Conditions for Optimal Control Problems in Discontinuous Dynamic Systems

Ekaterina Kostina[1], Olga Kostyukova[2], and Werner Schmidt[3]

[1] Department of Mathematics and Computer Science, University of Marburg, Hans-Meerwein-Str., 35032 Marburg, Germany
[2] Institute of Mathematics, Belarus Academy of Sciences, Surganov Str. 11, 220072 Minsk, Belarus
[3] Institute of Mathematics and Computer Science, University of Greifswald, Rathenaustr. 47, 17487 Greifswald, Germany

Abstract. In the paper we derive new necessary optimality conditions for optimal control of differential equations systems with discontinuous right hand side. The main attention is paid to a situation when an optimal trajectory slides on the discontinuity surface. The new conditions, derived in the paper, are essential and do not follow from any known necessary conditions for such systems.

Keywords: Optimal control problem, maximum principle, discontinuous dynamic system.

1 Introduction

Optimal control of systems of differential equations with discontinuous right hand side are widely used to describe numerous applications in natural sciences and engineering, where, e.g., there is a necessity to model dynamics with different scales or even with jumps. Such models appear e.g. in economics, mechanics (e.g. optimal control of mechanical systems with Coulomb friction), chemical processes, electrical and radio engineering, aerodynamics, automatic control theory, and theory of hybrid systems [3,5,8,10,19].

There is a row of papers devoted to the numerical solution of such optimal control problems, see [18] and references therein. However, due to the complexity of these optimal control problems there are few papers devoted to their theoretical studies.

The aim of this paper is to present new necessary optimality conditions for optimal control of differential equations systems with discontinuous right hand side. It is assumed that the discontinuity of the function, which describes the dynamic system, appears at some surface defined in the state space. This surface is called the discontinuity or switching surface. Most papers on the topic consider the situation when an optimal trajectory crosses the surface [2,4,17]. In this case the optimality conditions are a slight modification of the maximum principle of Pontryagin. However, the more interesting and practically non-studied case is the situation when an optimal trajectory slides on the surface. There

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 122–135, 2013.

are only very few papers [2,12,15,16], which deal with this case, however, the maximum conditions presented there are weaker than the conditions presented in this paper.

Dynamic systems with discontinuous right hand side may be presented as differential inclusions. Necessary optimality conditions in optimal control problems for differential inclusions have been intensively studied, see e.g. [6,14,20] and references therein. However, the results obtained there are based on the assumption that the differential inclusion is Lipschitz continuous or possesses a modified one-sided Lipschitz property, which is not the case for the differential inclusion describing the dynamic system of this paper.

2 Problem Statement and Assumptions

Let us consider the following optimal control problem with discontinuous right hand side:

$$\begin{aligned} &\min_{u(\cdot)} \ f_0(x(t^*)), \\ &\dot{x} = \begin{cases} \bar{f}^-(x,u) \text{ if } S(x(t)) < 0, \\ \bar{f}^+(x,u) \text{ if } S(x(t)) > 0, \end{cases} \\ &x(t_*) = x_0, \ h(x(t^*)) = 0, \ |u(t)| \le 1, \ t \in T = [t_*, t^*]. \end{aligned} \tag{1}$$

Here $\bar{f}^{\pm}(x,u) = f^{\pm}(x) + b(x)u$, $x = x(t) \in \mathbb{R}^n$ denotes the state n-vector, $u = u(t) \in \mathbb{R}$ is a scalar control, $S(x) := d^T x - \gamma = 0$ is a surface of discontinuity (switching surface), the functions $f_0(x) \in \mathbb{R}, f^{\pm}(x), b(x) \in \mathbb{R}^n$, $h(x) \in \mathbb{R}^{s_*}$ are given sufficiently smooth scalar and vector functions, d, x_0 are given vectors, γ, t_*, t^* are given numbers, $s_* = \dim(h)$ is the dimension of the vector-function $h = h(x)$.

In the general case, even for a fixed control $u(\cdot) = (u(t), t \in T)$ and a given initial value $x(t_*) = x_0$, the system (1) may not have a classical solution, since the system is not defined at the switching surface. Therefore, we redefine a solution of the system (1) at the switching surface following Filippov [8]. Then the problem may be reformulated as

$$\begin{aligned} &\min_{u(\cdot),\alpha(\cdot)} \ f_0(x(t^*)), \\ &\dot{x} = \begin{cases} \bar{f}^-(x,u), \text{ if } d^T x(t) < \gamma, \\ \bar{f}^+(x,u), \text{ if } d^T x(t) > \gamma, \\ F(x,u,\alpha), \text{ if } d^T x(t) = \gamma, \end{cases} \\ &x(t_*) = x_0, \ h(x(t^*)) = 0, \\ &|u(t)| \le 1, \ 0 \le \alpha(t) \le 1, \ t \in T = [t_*, t^*], \end{aligned} \tag{2}$$

where $F(x,u,\alpha) := \alpha \ \bar{f}^-(x,u) + (1-\alpha) \ \bar{f}^+(x,u) = f^+(x) + a(x)\alpha + b(x)u$, $a(x) := f^-(x) - f^+(x)$. In this problem the control is $u(t)$, $\alpha(t)$, $t \in T$.

Using classical optimal control theory [13], we may conclude that if the problem (2) is feasible, it has an optimal solution in the class of measurable functions $u(t)$, $\alpha(t)$, $t \in T$.

Let $u^*(\cdot) = (u^*(t), t \in T)$, $\alpha^*(\cdot) = (\alpha^*(t), t \in T)$, $x^*(\cdot) = (x^*(t), t \in T)$ be an optimal control and the corresponding trajectory of the system (2). Here and in what follows, we formally suppose that

$$\alpha^*(t) = 1 \text{ if } d^T x^*(t) < \gamma,\ \alpha^*(t) = 0 \text{ if } d^T x^*(t) > \gamma,$$
$$0 \le \alpha^*(t) \le 1 \text{ if } d^T x^*(t) = \gamma,\ t \in T.$$

For the formulation and the proof of the maximum principle we need several assumptions.

Assumption 1. *The functions $u^*(\cdot) = (u^*(t), t \in T)$, $\alpha^*(\cdot) = (\alpha^*(t),\ t \in T)$ are piecewise continuous and piecewise smooth.*

We denote $T_s = \{t \in T : d^T x^*(t) = \gamma\}$. In the general case this set contains points, which correspond to crossing the discontinuity surface by a trajectory, and segments, which correspond to the case when the trajectory lies on the surface. As it was mentioned before, the case when the trajectory crosses the discontinuity surface is well-studied. The aim of this paper is to study the case when the trajectory may lie on the discontinuity surface. For this purpose we need the following assumption.

Assumption 2. *The set T_s consists of a finite number of segments $[\tau_k, \tau^k]$, $k = 1, ..., p$,*

$$t_* < \tau_1 < \tau^1 < \tau_2 < \tau^2 < ... < \tau_p < \tau^p < t^*,$$

and the following inequalities hold,

$$d^T \dot{x}^*(\tau_k - 0) \ne 0,\ d^T \dot{x}^*(\tau^k + 0) \ne 0,\ k = 1, ..., p. \tag{3}$$

Here and in what follows, $z(\bar{t} - 0)$ and $z(\bar{t} + 0)$ for a given function $z(t)$, $t \in T$, are defined by $z(\bar{t} + 0) := \lim\limits_{t \to \bar{t}, t \ge \bar{t}} z(t)$, $z(\bar{t} - 0) := \lim\limits_{t \to \bar{t}, t \le \bar{t}} z(t)$.

We denote

$$T_s^0 = \{t \in T_s : \alpha^*(t) = 0\},\ \ T_s^1 = \{t \in T_s : \alpha^*(t) = 1\},\ \ T_s^* = T_s \setminus (T_s^0 \cup T_s^1).$$

Assumption 3. *The "active" set $T_s^0 \cup T_s^1$ does not contain isolated points and the following relations hold true for an optimal control:*

$$\exists\ \epsilon_0 > 0,\ \ |u^*(t)| \le 1 - \epsilon_0,\ t \in T_s^0 \cup T_s^1,$$
$$|d^T b(x^*(t))| \ge \epsilon_0,\ t \in T_s^0 \cup T_s^1;\ \ \ |d^T a(x^*(t))| \ge \epsilon_0,\ t \in T_s^*,$$
$$rank\ \frac{\partial h(x^*(t^*))}{\partial x} = s_*.$$

3 Necessary Optimality Conditions

The main goal of this section is to formulate and prove new necessary optimality conditions in the form of the maximum principle. For this purpose we will need some auxiliary results.

Without loss of generality, we will suppose for simplicity that the following relations are satisfied

$$d^T x^*(t) > \gamma, t \in [t_*, \tau_1), d^T x^*(t) = \gamma, t \in [\tau_1, \tau^1], d^T x^*(t) < \gamma, t \in (\tau^1, t^*], \quad (4)$$
$$\alpha^*(t) = 1,\ |u^*(t)| < 1,\ t \in (\tau_1, \tau_0),\ 0 < \alpha^*(t) < 1,\ t \in (\tau_0, \tau^1).$$

Therefore, for the case under consideration we have

$$p = 1,\ T_s = [\tau_1, \tau^1],\ T_\alpha^1 = (\tau_1, \tau_0),\ T_\alpha^0 = \emptyset,\ T_\alpha^* = (\tau_0, \tau^1),$$
$$T^+ = [t_*, \tau_1),\ T^- = (\tau^1, t^*].$$

We introduce the set of parameters

$$\mu = (y_0, y, \lambda_1, \lambda^1), \quad (5)$$

where $y_0 \geq 0$, $\lambda_1,\ \lambda^1$ are scalars, $y \in \mathbb{R}^{s_*}$, and denote by $\psi(t|\mu), t \in T$, a solution of the system

$$\dot{\psi}^T(t) = -\psi^T(t)\frac{\partial \bar{f}^+(x^*(t), u^*(t))}{\partial x},\ t \in [t_*, \tau_1);$$
$$\dot{\psi}^T(t) = -\psi^T(t)\frac{\partial \bar{f}^-(x^*(t), u^*(t))}{\partial x},\ t \in [\tau^1, t^*];$$
$$\dot{\psi}^T(t) = -\psi^T(t)\left(\frac{\partial F(x^*(t), u^*(t), \alpha^*(t))}{\partial x} - q_1^*(t)d^T\right),\ t \in [\tau_1, \tau_0),$$
$$\dot{\psi}^T(t) = -\psi^T(t)\left(\frac{\partial F(x^*(t), u^*(t), \alpha^*(t))}{\partial x} - q_2^*(t)d^T\right),\ t \in [\tau_0, \tau^1), \quad (6)$$
$$\psi(t^*) = -y_0\frac{\partial f_0(x^*(t^*))}{\partial x} - \frac{\partial h^T(x^*(t^*))}{\partial x}y,$$
$$\psi(\tau_1 - 0) = \psi(\tau_1 + 0) + d\lambda_1,\ \psi(\tau^1 - 0) = \psi(\tau^1 + 0) + d\lambda^1.$$

Further we pick up any $m \in \mathbb{N}$, $m \geq 2$, and consider a set of points $\bar{t}_1,\ \bar{t}_2, ...,\ \bar{t}_{2m}$, such that

$$\tau_1 = \bar{t}_1 < \bar{t}_2 < ... < \bar{t}_{2m-1} = \tau_0 < \bar{t}_{2m} = \tau^1. \quad (7)$$

This set of points satisfies the following lemma:

Lemma 1. *Let $u^*(\cdot)$, $\alpha^*(\cdot)$, $x^*(\cdot)$ be an optimal control and the trajectory in the problem (2), for which Assumptions 1-3 are fulfilled. For any choice of the points (7) there exists a vector of parameters μ, $\|\mu\| = 1$, (5), such that along*

the corresponding solution $\psi(t) = \psi(t|\mu), t \in T$, *of the system* (6) *the following relations hold true,*

$$\psi^T(t)b(x^*(t))u^*(t) = \max_{|u|\le 1} \psi^T(t)b(x^*(t))u, \ \ a.e.\ t \in T; \tag{8}$$

$$\psi^T(t)a(x^*(t))\alpha^*(t) = \max_{0\le\alpha\le 1} \psi^T(t)a(x^*(t))\alpha, \tag{9}$$

$$a.e.\ t \in [\bar{t}_{2i-1}, \bar{t}_{2i}], \ \ i = 1, 2, ..., m,$$

$$\psi^T(t)q_1^*(t) \ge 0, \quad a.e.\ t \in [\bar{t}_{2i}, \bar{t}_{2i+1}], \quad i = 1, 2, ..., m-1. \tag{10}$$

The proof of the lemma can be found in [11] and its main idea is sketched here. For a fixed $m \ge 2$ we consider a set of parameters $\theta = (t_1, t_2, ..., t_{2m})$ and formulate the optimal control problem of a hybrid system:

$$\begin{aligned}
\min_{u(\cdot),\alpha(\cdot),\theta} \ & f_0(x(t^*)), \\
& \dot{x} = \bar{f}^+(x,u), \ d^T x(t) \ge \gamma, \ t \in [t_*, t_1[\\
& \dot{x} = F(x,u,\alpha), \ d^T x(t) = \gamma, \ t \in [t_{2i-1}, t_{2i}[, \\
& \dot{x} = \bar{f}^-(x,u), \ d^T x(t) \le \gamma, \ t \in [t_{2i}, t_{2i+1}[, \ i = 1, ..., m, \\
& x(t_*) = x_0, \ h(x(t_*)) = 0, \\
& t_* = t_0 \le t_1 \le ... \le t_{2m} \le t_{2m+1} = t^*, \\
& |u(t)| \le 1, \ t \in T; \ 0 \le \alpha(t) \le 1, \ t \in \bigcup_{i=1}^{m} [t_{2i-1}, t_{2i}].
\end{aligned} \tag{11}$$

Let us note that in the problem (11), the decision variables are the control $u(\cdot), \alpha(\cdot)$ and a vector θ.

With the notations

$$\begin{aligned}
z_i(\tau) &= x(t_{i-1} + \tau(t_i - t_{i-1})), \ \ i = 1, 2, ..., 2m+1, \\
v_i(\tau) &= u(t_{i-1} + \tau(t_i - t_{i-1})), \ \ i = 1, 2, ..., 2m+1, \\
\beta_i(\tau) &= \alpha(t_{i-1} + \tau(t_i - t_{i-1})), \ \ i = 2, 4, ..., 2m,
\end{aligned} \tag{12}$$

we form the extended state vector

$$Z(\tau) = (z_i(\tau), i = 1, ..., 2m+1; \ t_i(\tau), i = 1, ..., 2m) \in \mathbb{R}^{n\times(2m+1)+2m}, \quad \tau \in [0,1], \tag{13}$$

and the extended control vector

$$V(\tau) = (v_i(\tau), i = 1, ..., 2m+1, \beta_i(\tau), i = 2, 4, ..., 2m) \in \mathbb{R}^{3m+1}, \quad \tau \in [0,1]. \tag{14}$$

Using the introduced notations we may rewrite the problem (11) as follows,

$$\begin{aligned}
&\min_{V(\cdot)} \ f_0(z_{2m+1}(1)),\\
&\quad \dot{Z}(\tau) = \mathcal{F}(Z(\tau), V(\tau)),\ \Phi(Z(0), Z(1)) = 0, \qquad (15)\\
&\quad G_1^T Z(\tau) \geq \gamma,\ G_{2i}^T Z(\tau) = \gamma,\ G_{2i+1}^T Z(\tau) \leq \gamma,\ i = 1, 2, ..., m;\\
&\quad |v_i(\tau)| \leq 1,\ i = 1, ..., 2m+1,\\
&\quad 0 \leq \beta_i(\tau) \leq 1,\ i = 2, 4, ..., 2m,\\
&\quad t_i(\tau) \leq t_{i+1}(\tau), i = 0, ..., 2m,\ \tau \in [0, 1]. \qquad (16)
\end{aligned}$$

Here $V(\tau)$ is the control vector (14), $Z(\tau)$ is the state vector (13),

$$\begin{aligned}
\mathcal{F}^T(Z, V) = \Big(&(t_1 - t_0)\bar{f}^+(z_1, v_1),\\
&(t_{2i} - t_{2i-1})F(z_{2i}, v_{2i}, \beta_{2i}),\ (t_{2i+1} - t_{2i})\bar{f}^-(z_{2i+1}, v_{2i+1}),\\
&i = 1, ..., m, \underbrace{0, ..., 0}_{2m}\Big),\\
\Phi(Z(0), Z(1)) = &\begin{pmatrix} z_1(0) - x_0 \\ z_i(1) - z_{i+1}(0),\ i = 1, ..., 2m \\ h(z_{2m+1}(1)) \\ d^T z_i(1) - \gamma,\ i = 1, ..., 2m \end{pmatrix},\\
G_i^T = (&\underbrace{\mathbb{O}^T, ..., \mathbb{O}^T}_{i-1}, d^T, \underbrace{\mathbb{O}^T, ..., \mathbb{O}^T}_{2m+1-i}, \underbrace{0, ..., 0}_{2m}),\ i = 1, ..., 2m+1,\\
&\mathbb{O} \in \mathbb{R}^n \text{ is a vector of zeros},\\
&t_0(\tau) \equiv \bar{t}_0 = t_*,\ t_{2m+1}(\tau) \equiv \bar{t}_{2m+1} = t^*.
\end{aligned}$$

Now consider the set of points $\bar{t}_1,\ \bar{t}_2, ...,\ \bar{t}_{2m}$ satisfying (7) and denote by $Z^*(\tau), V^*(\tau),\ \tau \in [0, 1]$, the functions (12) - (14), constructed using this set, the optimal control $u^*(\cdot)$, $\alpha^*(\cdot)$ and the trajectory $x^*(\cdot)$ of the problem (2).

Since the control $u^*(\cdot)$, $\alpha^*(\cdot)$ and the trajectory $x^*(\cdot)$ are optimal in the problem (2), it is obvious that $V^*(\tau), Z^*(\tau), \tau \in [0, 1]$, are an optimal control and the corresponding trajectory of the problem (15), (16). By the assumptions (see (3) and (4)) we have

$$\begin{aligned}
&G_1^T Z^*(\tau) > \gamma,\ \tau \in [0, 1),\ G_1^T Z^*(1) = \gamma,\ G_1^T \dot{Z}^*(1 - 0) \neq 0,\\
&G_{2m+1}^T Z^*(\tau) < \gamma,\ \tau \in (0, 1],\ G_{2m+1}^T Z^*(0) = \gamma,\ G_{2m+1}^T \dot{Z}^*(+0) \neq 0.
\end{aligned}$$

Hence, $V^*(\tau), Z^*(\tau), \tau \in [0,1]$, is also a strong local extremal in the problem that results from the problem (15), (16) by removing the state constraints $G_1^T Z(\tau) \geq \gamma,\ G_{2m+1}^T Z(\tau) \leq \gamma,\ \tau \in [0,1]$, and the constraints (16), namely in the problem

$$\begin{aligned} \min_{V(\cdot)}\ & f_0(z_{2m+1}(1)), \\ & \dot{Z}(\tau) = \mathcal{F}(Z(\tau), V(\tau)),\ \Phi(Z(0), Z(1)) = 0, \\ & G_i^T Z(\tau) \leq \gamma, i = 3,5,...,2m-1;\ G_i^T Z(\tau) = \gamma, i = 2,4,...,2m; \\ & |v_i(\tau)| \leq 1,\ i = 1,...,2m+1,\ 0 \leq \beta_i(\tau) \leq 1,\ i = 2,4,...,2m; \\ & \tau \in [0,1], \end{aligned} \tag{17}$$

with the control $V(\tau)$ (see (14)) and the state vector $Z(\tau)$ (see (13)).

The problem (17) is an optimal control problem with inequality and equality state constraints and boundary constraints $\Phi(Z(0), Z(1)) = 0$. Due to Assumptions 1-3 and the specific structure of this problem, regularity conditions (see [1]) are satisfied for the control $V^*(\tau)$ and the trajectory $Z^*(\tau)$, $\tau \in [0,1]$. Thus we can apply Theorems 4.1 and 12.1 from [1] and results from [7], according to which the control $V^*(\tau)$ and the trajectory $Z^*(\tau)$, $\tau \in [0,1]$, satisfy certain relations. Analyzing these relations and taking into account the structure of the vectors $V^*(\tau)$ and $Z^*(\tau)$ allow us to get the assertions of the lemma. ◇

Let us note that it follows from Lemma 1 that the continuity of the function $\psi(t|\mu), t \in (\tau_1, \tau^1)$, the relations (9), (10) and the assumption $\alpha^*(t) = 1$, $t \in (\tau_1, \tau_0)$, imply the inequalities

$$\psi^T(\bar{t}_i|\mu) a(x^*(\bar{t}_i)) \geq 0,\ \psi^T(\bar{t}_i|\mu) q_1^*(\bar{t}_i) \geq 0,\ i = 2,3,...,2m-1, \tag{18}$$

for each point set $\bar{t}_1, \bar{t}_2, ..., \bar{t}_{2m}$ satisfying (7).

Now we are ready to formulate and prove new necessary optimality conditions for problem (2) in the form of the maximum principle.

Theorem 1. *Let $u^*(\cdot)$, $\alpha^*(\cdot)$, $x^*(\cdot)$ be an optimal control and the corresponding trajectory of the problem (2), which satisfy Assumptions 1-3. Then there exist numbers λ_k, λ^k, $k = 1,...,p$, $y_0 \geq 0$, and a vector $y \in \mathbb{R}^{s_*}$, not all trivial,*

$$\sum_{k=1}^{p} (|\lambda_k| + |\lambda^k|) + y_0 + ||y|| > 0,$$

such that along a solution of the adjoint system

$$\dot{\psi}^T(t) = \begin{cases} -\psi^T(t)\frac{\partial(f^{\pm}(x^*(t))+b(x^*(t))u^*(t))}{\partial x}, & t \in \{t \in T : \\ & \pm(d^T x^*(t) - \gamma) > 0\}, \\ -\psi^T(t)\left(\frac{\partial F(x^*(t),u^*(t),\alpha^*(t))}{\partial x} - q_1^*(t)d^T\right), & t \in T_s^0 \cup T_s^1, \\ -\psi^T(t)\left(\frac{\partial F(x^*(t),u^*(t),\alpha^*(t))}{\partial x} - q_2^*(t)d^T\right), & t \in T_s^*, \end{cases}$$

$$\psi(t^*) = -y_0\frac{\partial f_0(x^*(t^*))}{\partial x} - \frac{\partial h^T(x^*(t^*))}{\partial x}y,$$
$$\psi(\tau_k - 0) = \psi(\tau_k + 0) + d\lambda_k,$$
$$\psi(\tau^k - 0) = \psi(\tau^k + 0) + d\lambda^k,\ k = 1, ..., p,$$

the following relations hold true:

$$\psi^T(t)b(x^*(t))u^*(t) = \max_{|u|\le 1}\psi^T(t)b(x^*(t))u,\ \ a.e.\ t \in T;$$

$$\psi^T(t)a(x^*(t))\alpha^*(t) = \max_{0\le\alpha\le 1}\psi^T(t)a(x^*(t))\alpha,\ \ a.e.\ t \in T_s; \tag{19}$$

$$\psi^T(t-0)\dot{x}^*(t-0) = \psi^T(t+0)\dot{x}^*(t+0),\ t = \tau_k,\ t = \tau^k,\ k = 1, ..., p,$$

$$\psi^T(t)q_1^*(t) \le 0,\ t \in int\ T_s^0, \quad \psi^T(t)q_1^*(t) \ge 0,\ t \in int\ T_s^1. \tag{20}$$

Here

$$q_i^*(t) := q_i(x^*(t), u^*(t), \alpha^*(t)); \quad i = 1, 2;$$
$$q_1(x,u,\alpha) := \left(\frac{\partial F(x,u,\alpha)}{\partial x}b(x) - \frac{\partial b(x)}{\partial x}F(x,u,\alpha)\right)/d^T b(x),$$
$$q_2(x,u,\alpha) := \left(\frac{\partial F(x,u,\alpha)}{\partial x}a(x) - \frac{\partial a(x)}{\partial x}F(x,u,\alpha)\right)/d^T a(x).$$

Proof. Again we will suppose for simplicity that relations (4) are satisfied. For an arbitrary $m \in \mathbb{N}$, $m \ge 2$, we consider the set of points

$$t_i^{(m)} = \tau_1 + (i-1)\frac{(\tau_0 - \tau_1)}{2m-2},\ i = 1, ..., 2m-1,\ \ t_{2m}^{(m)} = \tau^1. \tag{21}$$

For any $m \ge 2$ the set (21) satisfies relations (7). Hence Lemma 1 and the relation (18) imply that for a set (21) there exists a vector

$$\mu(m) = (y_0(m), y(m), \lambda_1(m), \lambda^1(m)),\ ||\mu(m)|| = 1,$$

such that the following relations hold true,

$$\psi^T(t|\mu(m))b(x^*(t))u^*(t) = \max_{|u|\le 1} \psi^T(t|\mu(m))b(x^*(t))u, \text{ a.e. } t \in T,$$
$$\psi^T(t|\mu(m))a(x^*(t))\alpha^*(t) = \max_{0\le\alpha\le 1} \psi^T(t|\mu(m))a(x^*(t))\alpha \tag{22}$$
$$\text{a.e. } t \in [\tau_0, \tau^1],$$
$$\psi^T(t-0|\mu(m))\dot{x}^*(t-0) = \psi^T(t+0|\mu(m))\dot{x}^*(t+0),\ t = \tau_1,\ t = \tau^1;$$

$$\psi^T(t_i^{(m)}|\mu(m))a(x^*(t_i^{(m)})) \ge 0,\ \psi^T(t_i^{(m)}|\mu(m))q_1^*(t_i^{(m)}) \ge 0, \tag{23}$$
$$i = 2, 3, ..., 2m-1.$$

Consider the sequence of the vectors $\mu(m),\ m = 2, 3, ...$ Since $||\mu(m)|| = 1$, $m = 2, 3, ...$, there exists a converging subsequence. Without loss of generality, we assume that the sequence $\mu(m),\ m = 2, 3, ...$ converges itself,

$$\mu^* = \lim_{m\to\infty} \mu(m).$$

Obviously, $||\mu^*|| = 1$.

It follows from (21) that for any point $t \in [\tau_1, \tau_0]$ there exists a sequence of indices

$$i(m) = i(m|t) \in \{2, 3, ..., 2m-1\}, m = 2, 3, 4, ...,$$

such that

$$t_{i(m)}^{(m)} \to t \text{ as } m \to \infty.$$

By construction (see (23)),

$$\psi^T(t_{i(m)}^{(m)}|\mu(m))a(x^*(t_{i(m)}^{(m)})) \ge 0,\ \psi^T(t_{i(m)}^{(m)}|\mu(m))q_1^*(t_{i(m)}^{(m)}) \ge 0, \tag{24}$$
$$m = 2, 3, 4, ...$$

For $m \to \infty$ in the last inequalities we get

$$\psi^T(t|\mu^*)a(x^*(t)) \ge 0,\ \psi^T(t|\mu^*)q_1^*(t) \ge 0,\ t \in [\tau_1, \tau_0]. \tag{25}$$

Similarly, for $m \to \infty$ in (22) we obtain

$$\psi^T(t|\mu^*)b(x^*(t))u^*(t) = \max_{|u|\le 1} \psi^T(t|\mu^*)b(x^*(t))u, \text{ a.e. } t \in T;$$

$$\psi^T(t|\mu^*)a(x^*(t))\alpha^*(t) = \max_{0\le\alpha\le 1} \psi^T(t|\mu^*)a(x^*(t))\alpha \text{ a.e. } t \in [\tau_0, \tau^1]; \tag{26}$$
$$\psi^T(t-0|\mu^*)\dot{x}^*(t-0) = \psi^T(t+0|\mu^*)\dot{x}^*(t+0),\ t = \tau_1,\ t = \tau^1.$$

The relations (25) and (26) are nothing but the assertions of Theorem 1 for the considered structure of the solution of the problem (2) (see the assumption (4)). Analogously we may prove the Theorem for other types of solution structure. ⋄

4 Discussion of the New Necessary Optimality Conditions

Because of the problem complexity, there are only few papers, e.g. [2,12,16,15], where necessary optimality conditions for optimal control problems with discontinuous dynamics are presented. The theorem formulated in this paper contains new crucial conditions (19) and (20).

Let us briefly discuss differences between our maximum principle and maximum principles known from the literature.

The necessary conditions in [2] are proved under the very strong assumption $T_s^0 = T_s^1 = \emptyset$, which is not needed in our theorem.

As in [12], the optimality conditions derived here are based on the so-called direct approach (see [9]). However, the necessary conditions in [12] contain the conditions (20) but not the condition (19).

The optimality conditions in [2,16] are formulated based on an so-called indirect approach (see [9]), which a priori leads to weaker results compared with the direct approach.

In [16], conditions (19), (20) are not considered. However, the conditions (19), (20) are essential and they are not a consequence of any other conditions mentioned in [12], [16].

The maximum principle in [15] is weaker than the Theorem 1. Indeed, one can easily show that for the problems of the form (2) the maximum principle from [15] is satisfied trivially for any feasible control. Furthermore, one can construct examples (one of them is presented below) where a non-optimal control satisfies the conditions of maximum principle from [15], but not the conditions of Theorem 1.

To finish the discussion of the new necessary conditions, we would like to stress the importance of the conditions (19).

Let us consider an optimal control problem with state constraints in the form

$$\begin{aligned} \min\ & f_0(x(t^*)), \\ & \dot{x} = f^-(x) + b(x)u,\ x(t_*) = x_0,\ h(x(t^*)) = 0, \\ & d^T x(t) \le \gamma,\ |u(t)| \le 1,\ t \in T = [t_*, t^*]. \end{aligned}$$

Suppose that this problem has an optimal control $u^*(t), t \in T$, and the corresponding trajectory $x^*(t)$, $t \in T$, such that mes $T_s > 0$, $T_s := \{t \in T : d^T x^*(t) = \gamma\}$. Then the control $u^*(t), t \in T$, and the function $\alpha^*(t) = 1$, $t \in T$, are feasible in the original problem (2) for any function $f^+(x)$ and satisfy the necessary optimality conditions from [12], [16]. However, one can easily construct functions $f^+(x)$ (and corresponding optimal control problems (2)), which together with the functions $u^*(t)$, $\alpha^*(t) = 1$, $t \in T$, violate the maximum condition (19). Hence, the control $u^*(t)$, $\alpha^*(t)$, $t \in T$, is not optimal in (2) according to Theorem 1.

5 Illustrative Examples

In order to demonstrate differences of our maximum principle from maximum principles known from the literature we have constructed several examples. The

aim of the examples is to show that the necessary conditions presented in this paper are stronger than known necessary conditions. Namely, we want to show that the new conditions (19) and (20) may be violated for controls that satisfy other necessary optimality conditions known from the literature, and hence our maximum principle guarantees that such controls are not optimal. Furthermore, we want to show that the new conditions are essential and does not follow from other known optimality conditions.

Example 1. Consider the optimal control problem depending on a parameter c,

$$
\begin{aligned}
\mathbf{P(c)}: \quad & \min\ x_1(t^*) - 2.5x_2(t^*),\\
& x(t_*) = x_0,\ x_3(t^*) = 1,\\
& |u(t)| \le 1,\ 0 \le \alpha(t) \le 1,\ t \in [t_*, t^*],\\
& \left.\begin{aligned} \dot{x}_1 &= x_2,\\ \dot{x}_2 &= x_3 + 5,\\ \dot{x}_3 &= u + 1/2, \end{aligned}\right\} \text{ if } x_3 < 0, \quad \left.\begin{aligned} \dot{x}_1 &= x_2 + c,\\ \dot{x}_2 &= x_3,\\ \dot{x}_3 &= u, \end{aligned}\right\} \text{ if } x_3 > 0,\\
& \left.\begin{aligned} \dot{x}_1 &= x_2 + c(1-\alpha),\\ \dot{x}_2 &= x_3 + 5\alpha,\\ \dot{x}_3 &= u + 1/2\alpha, \end{aligned}\right\} \text{ if } x_3 = 0,\\
& \text{with } x_0^T = (19/32, -37/16, -3/4),\ t_* = -0.5,\ t^* = 2
\end{aligned}
$$

and the control $u^*(\cdot)$, $\alpha^*(\cdot)$:

$$
u^*(t) = \begin{cases} 1, & t \in [-0.5, 0],\\ -0.5, & t \in [0, 1],\\ 1, & t \in [1, 2], \end{cases} \qquad \alpha^*(t) = \begin{cases} 1, & t \in [-0.5, 0],\\ 1, & t \in [0, 1],\\ 0, & t \in [1, 2]. \end{cases}
$$

For the control $u^*(\cdot)$, $\alpha^*(\cdot)$ we have $T_s = [0, 1]$, $T_s^1 = [0, 1]$, $T_s^0 = \emptyset$, $T_s^* = \emptyset$.

If we choose $c = c^0 = 0$, then in the problem $P(c^0)$ the control $u^*(\cdot)$, $\alpha^*(\cdot)$ is feasible, locally optimal and satisfies all necessary optimality conditions from Theorem 1.

If we choose $c = c^* = -5.5$, then in the problem $P(c^*)$ the control $u^*(\cdot)$, $\alpha^*(\cdot)$ is feasible and not optimal but it satisfies all necessary conditions from [12,16], and it satisfies all necessary conditions from Theorem 1 except for the condition (19). Hence, according to [12,16], this control may be locally optimal in the problem $P(c^*)$. On the other hand, following Theorem 1, it cannot be locally optimal in the problem $P(c^*)$.

Example 2. Consider the optimal control problem

$$
\begin{aligned}
& \min 2x_1(t^*) - 2x_2(t^*),\\
& x(t_*) = x_0,\ x_3(t^*) = -1,\ |u(t)| \le 1,\ 0 \le \alpha(t) \le 1,\ t \in [t_* = -0.5, t^* = 2],\\
& \left.\begin{aligned} \dot{x}_1 &= x_2,\\ \dot{x}_2 &= x_3,\\ \dot{x}_3 &= u, \end{aligned}\right\} \text{ if } x_3 < 0, \quad \left.\begin{aligned} \dot{x}_1 &= x_2 + 1,\\ \dot{x}_2 &= x_3,\\ \dot{x}_3 &= u, \end{aligned}\right\} \text{ if } x_3 > 0,
\end{aligned} \tag{27}
$$

$$\left.\begin{array}{l}\dot{x}_1 = x_2 + (1-\alpha),\\ \dot{x}_2 = x_3,\\ \dot{x}_3 = u,\end{array}\right\} \text{ if } x_3 = 0,$$
$$\text{with } x_0^T = (-23/48, -1/8, 1/2),$$

and the control $u^*(\cdot)$, $\alpha^*(\cdot)$

$$u^*(t) = -1, \alpha^*(t) = 0;\ t \in [-0.5, 0],$$
$$u^*(t) = 0, \alpha^*(t) = 1;\ t \in [0, 1],$$
$$u^*(t) = -1, \alpha^*(t) = 1;\ t \in [1, 2].$$

For the control $u^*(\cdot)$, $\alpha^*(\cdot)$ we have $T_s = [0, 1]$, $T_s^1 = [0, 1]$, $T_s^0 = T_s^* = \emptyset$.

The control $u^*(\cdot)$, $\alpha^*(\cdot)$ satisfies all necessary optimality conditions from [16], satisfies all necessary optimality conditions from Theorem 1, except for the condition (20), and is not locally optimal.

Note that results from [15] can not be applied to this example because Assumption 3) from [15] (namely the condition $d^T(\bar{f}^-(x, u^*(t)) - \bar{f}^+(x, u^*(t))) > 0$, $\forall\ x \in S_0(t)$, $t \in T_s$) is not satisfied.

Example 3. Consider the following problem,

$$\begin{aligned}\mathbf{P^*}: \quad & \min c^T x(t^*),\\ & x(t_*) = x_0,\ d^T x(t^*) = 1,\ |u(t)| \le 1,\ 0 \le \alpha(t) \le 1,\ t \in [t_*, t^*],\\ & \dot{x} = Ax + bu + g^-,\ \text{if } d^T x < 0,\\ & \dot{x} = Ax + bu + g^+,\ \text{if } d^T x > 0,\\ & \dot{x} = Ax + bu + \alpha g^- + (1-\alpha) g^+,\ \text{if } d^T x = 0,\end{aligned}$$

with $x \in \mathbb{R}^n$, $n = 4$, $t_* = -0.5$, $t^* = 2$,

$$c = \begin{pmatrix} -\frac{32}{7} \\ \frac{48}{7} \\ -5 \\ 0 \end{pmatrix},\ d = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix},$$

$$g^+ = \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix}, g^- = \begin{pmatrix} 0 \\ 0 \\ 5 \\ 1/2 \end{pmatrix},\ b = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The vector x_0 is uniquely defined by the condition that the trajectory $x(t), t \in [-0.5, 0]$, of the system $\dot{x} = Ax + b + g^-$, $x(-0.5) = x_0$, should satisfy the equality $x(0) = 0 \in \mathbb{R}^n$. Hence $d^T x_0 = -3/4$.

Consider the control $u^*(t), \alpha^*(t)$, $t \in [-0.5, 2]$,

$$\begin{aligned} & u^*(t) = 1,\ \alpha^*(t) = 1,\ t \in [-0.5, 0],\\ & u^*(t) = -1/2,\ \alpha^*(t) = 1,\ t \in [0, 1],\\ & u^*(t) = 1,\ \alpha^*(t) = 0,\ t \in [1, 2],\end{aligned} \tag{28}$$

and the corresponding trajectory $x^*(t) = (x_1^*(t), x_2^*(t), x_3^*(t), x_4^*(t)), t \in [t_*, t^*]$, with the function $d^T x^*(t) = x_4^*(t)$ satisfying

$$x_4^*(t) < 0, t \in [-0.5, 0);\ x_4^*(t) = 0, t \in [0, 1];\ x_4^*(t) > 0, t \in (1, 2].$$

For the control (28) we have $T_s = [0, 1],\ T_s^1 = [0, 1], T_s^0 = \emptyset$.

We can show that the control (28) is not locally optimal in the problem $P*$, satisfies all assumptions and all necessary optimality conditions from [15] and satisfies all conditions from Theorem 1 except for the condition (20).

6 Conclusions

We have presented a new maximum principle for optimal control problems in discontinuous systems, which takes into account a situation when a solution of the dynamic system lies on the switching surface. We have shown that the new maximum principle is stronger than known optimality conditions and contains new conditions which are essential and do not follow from other known optimality conditions.

References

1. Arutyunov, A.V.: Optimality Conditions: Abnormal and Degenerate Problems, Mathematics and its Applications, vol. 526. Kluwer (2000)
2. Ashchepkov, L.T.: Optimal Control of Discontinuous Systems. Nauka, Sibirsk. Otdel., Novosibirsk (1987)
3. Bock, H.G., Longman, R.W.: Optimal Control of Velocity Profiles for Minimization of Energy Consumption in the New York Subway System. In: Proc. 2nd IFAC Workshop on Control Applications of Nonlinear Programming and Optimization, pp. 34–43. IFAC (1980)
4. Boltyanskii, V.: The Maximum Principle for Systems with Discontinuous Right-Hand Side. In: Proceeding of the 6th IASTED International Conference on Intelligent Systems and Control, pp. 384–387 (2004)
5. Branicky, M.S.: Introduction to Hybrid Systems. In: Handbook of Networked and Embedded Control Systems, pp. 91–116. Birkhäuser (2005)
6. Donchev, T., Farkhi, E., Mordukhovich, B.S.: Discrete Approximations, Relaxation, and Optimization of One-Sided Lipschitzian Differential Inclusions in Hilbert Spaces. J. Differential Equations 243, 301–328 (2007)
7. Dubovitskii, A.Y., Dubovitskii, V.A.: Criterion of Existence of a Substantive Maximum Principle in a State Constraint Problem. Differential Equations 31(10), 1611–1616 (1995)
8. Filippov, A.F.: Differential Equations with Discontinuous Right-Hand Sides. Kluwer (1988)
9. Hartl, R., Sethi, S., Vickson, R.: A Survey of the Maximum Principles for Optimal Control Problems with State Constraints. SIAM Review 37(2), 181–218 (1995)
10. Kirches, C., Sager, S., Bock, H.G., Schlöder, J.P.: Time-Optimal Control of Automobile Test Drives with Gear Shifts. Optimal Control Applications and Methods 31(2), 137–153 (2010)

11. Kostina, E.A., Kostyukova, O.I., Schmidt, W.: New Maximum Principle for Optimal Control Problems in Discontinuous Dynamic Systems: Proof and Examples. Technical Report, University of Marburg (2012), http://www.uni-marburg.de/fb12/forschung/berichte/berichtemathe/pdfbfm/bfm12-03.pdf
12. Kugushev, E.I.: Necessary Optimality Conditions for Systems Described by Equations with Discontinuous Right-Hand Side. Vestn. Mosk. Gos. Univ., Ser. Math. Mechanics 2, 83–90 (1974)
13. Lee, E.B., Markus, L.: Foundations of Optimal Control Theory. Wiley (1967)
14. Mordukhovich, B.S.: Variational Analysis and Generalized Differentiation, II: Applications. Fundamental Principles of Mathematical Sciences, vol. 331. Springer, Heidelberg (2006)
15. Morozov, S.F., Sumin, M.I.: Optimal Control of Sliding Modes of Discontinuous Dynamical Systems. Izv. Vyssh. Uchebn. Zaved. Mat. 392(1), 53–61 (1990)
16. Oberle, H.J., Rosendahl, R.: Numerical Computation of a Singular-State Subarc in an Economic Optimal Control Model. Optimal Control Applications and Methods 27(4), 211–235 (2006)
17. Pontryagin, L.S., Boltyanskij, V.G., Gamkrelidze, R.V., Mishchenko, E.F.: Selected Works. In: Gamkrelidze, R.V. (ed.) Classics of Soviet Mathematics. The Mathematical Theory of Optimal Processes, vol. 4. Gordon and Breach Science Publishers, New York (1986)
18. Stewart, D.E., Anitescu, M.: Optimal Control of Systems with Discontinuous Differential Equations. Numerische Mathematik 114(4), 653–695 (2010)
19. Utkin, V.: Sliding Modes in Control and Optimization. Springer, Heidelberg (1992)
20. Vinter, R.B.: Optimal Control. Birkhäuser (2000)

Numerical Methods for the Optimal Control of Scalar Conservation Laws

Sonja Steffensen[1], Michael Herty[1], and Lorenzo Pareschi[2]

[1] RWTH Aachen University, Templergraben 55,
D-52065 Aachen, Germany
{herty,steffensen}@mathc.rwth-aachen.de
[2] University of Ferrara, Department of Mathematics,
Via Machiavelli 35, I-44121 Ferrara, Italy
lorenzo.pareschi@unife.it

Abstract. We are interested in a class of numerical schemes for the optimization of nonlinear hyperbolic partial differential equations. We present continuous and discretized relaxation schemes for scalar, one–conservation laws. We present numerical results on tracking type problems with nonsmooth desired states and convergence results for higher–order spatial and temporal discretization schemes.

Keywords: IMEX schemes, optimal control, conservation laws, Runge-Kutta methods.

1 Introduction

We consider an optimal control problem for scalar conservation laws of the type

$$\begin{aligned} &\text{minimize}_{u_0}\ J(u(T), u_0) \\ &\text{subject to } u_t + f(u)_x = 0, \qquad u(0,x) = u_0(x), \end{aligned} \tag{1}$$

Here, J and f are assumed to be smooth and possibly nonlinear functions. The initial value u_0 acts as control to the problem. It can be observed that the wave interactions that occur in the solution u in the case of a nonlinear flux function f pose the serious analytical challenges. Recently, the differentiability of J with respect to u_0 could be proven in the sense of shift–differentiability. We refer to [6,9,10,11,28,29,30,4,32] for more details.

Here, a class of numerical methods applied to the optimal control problem (1) is studied. We only consider the case of smooth initial data and smooth solutions u and refer to [4] for more details. For a numerical analysis including shock waves and in the case of the Lax–Friedrichs scheme we refer to [21,32] and the references therein.

1.1 Relaxation Method

As motiviation for a numerical scheme we follow the ideas of Jin and Xin [22]. Therein, a linear approximation (2) of the nonlinear hyperbolic equation

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 136–144, 2013.

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

has been discussed. For initial conditions $u(x,0) = u_0$ the approximation is

$$\begin{aligned} \partial_t u + \partial_x v &= 0\,, & u(x,0) &= u_0, \\ \partial_t v + a^2 \partial_x u &= \tfrac{1}{\epsilon}\left(f(u) - v\right), & v(x,0) &= f(u_0) \end{aligned} \tag{2}$$

where $\varepsilon > 0$ is the relaxation rate and a is a given constant satisfying the subcharacteristic condition $\max_u |f'(u)| \le a$. For ε being small, the solution u of (2) satisfies $\partial_t u + \partial_x f(u) = \varepsilon \partial_x((a^2 - f(u)^2)\partial_x u)$ (cf. [22]). Applying the relaxation to the optimal control problem (1), we obtain

$$\min_{u_0} J(u(\cdot,T),u_0) \quad \text{subject to} \quad \begin{cases} u_t + v_x = 0, \\ v_t + a^2 u_x = \tfrac{1}{\epsilon}\left(f(u) - v\right), \\ u(0,x) = u_0,\ v(0,x) = f(u_0) \end{cases} \tag{3}$$

The corresponding adjoint equations for (3) are given by (cf. [??])

$$\begin{aligned} -p_t - a^2 q_x &= \frac{q}{\epsilon} f'(u), \quad p(T,x) = p_T(x), \\ -q_t - p_x &= -\frac{q}{\epsilon}, \quad q(T,x) = q_T(x). \end{aligned}$$

For more information on the relaxation system, its limiting scheme for $\epsilon = 0$, further numerical analysis and extensions we refer to [1,2,5,14,3,8,22,25,27] and the references therein. Also, the computations are valid provided that all appearing functions are at least once differentiable. This is in general not the case for conservation laws.

2 IMEX-Runge-Kutta Discretization

Numerical discretization of the relaxation system using higher order temporal discretizations combined with higher order spatial discretization has been investigated in several recent publications as for example [22,27]. We apply so called implicit–explicit Runge-Kutta methods [26,27,3] as temporal discretization (IMEX RK). Here, the expliciti integration is used for the linear hyperbolic transport part and an implicit method is applied to the the stiff source term. Implicit-explicit Runge-Kutta method have been studied in the context of control problems for example in [4,19]. Define

$$\mathbf{y} = (u,v)^T, \qquad g(\mathbf{y}) = (v, a^2 u)^T \qquad \text{and} \qquad r(\mathbf{y}) := (0, -(v - f(u)))^T$$

then (2) becomes

$$\mathbf{y}_t + g(\mathbf{y})_x = \frac{1}{\varepsilon} r(\mathbf{y}), \qquad \text{and} \qquad \mathbf{y}(0,x) = (u^0, f(u^0))^T(x)$$

Applying a suitable discretization D_x of the spatial derivative yields the semi-discrete state equations

$$\mathbf{y}' = -D_x\, g(\mathbf{y}) + \frac{1}{\epsilon} r(\mathbf{y}), \qquad \mathbf{y}(0) = \mathbf{y}^0. \tag{4}$$

Remark 1. Spatial discretizations for the linear transport part are well–known. The simplest possible is a first–order Upwind method:

$$\frac{\partial}{\partial t}\mathbf{y}_j = -\frac{1}{\Delta x} \begin{pmatrix} 0 & 1 \\ a^2 & 0 \end{pmatrix} (\mathbf{y}_{j+1/2} - \mathbf{y}_{j-1/2}) + \frac{1}{\epsilon} r(\mathbf{y}_j)\,,$$

where $\mathbf{y}_{j+1/2}$ is obtained by applying the first-order upwind method to characteristic variables $v \pm au$. Higher order MUSCL schemes, WENO schemes or central schemes have also been studied in this context.

The resulting semi–discrete optimal control problem is then given by:

$$\begin{aligned} &\text{minimize} && j(\mathbf{y}(T), \mathbf{y}^0) \\ &\text{subject to} && \mathbf{y}' = -D_x\, g(\mathbf{y}) + \tfrac{1}{\epsilon} r(\mathbf{y}), \qquad \mathbf{y}(0) = \mathbf{y}^0. \quad t \in [0,T] \end{aligned} \tag{5}$$

In the context of relaxation schemes the semi–discrete problem is seen as a time–integration problem with stiff source which is discretized by an IMEX RK methods. For the numerical discretization we therefore consider the previous problem as an optimal control problem involving ordinary differential equations. Literature concerning the numerical analysis of Runge-Kutta methods for the optimality system of (5) have been studied in [17,7,24]. In [7,17] partitioned Runge-Kutta methods for the optimality system are obtained using the *discretize–then–optimize* approach. The derived partitioned Runge–Kutta methods have been analysed with regard to symplecticity and order of convergence. In [19], Herty and Schleper, moreover, analysed the associated adjoint imex Runge-Kutta method that one obtains if an explicit method is applied to $D_x g(y)$ and a (diagonally) implicit method to $\frac{1}{\epsilon} r(y)$. In the following, we will analyse general partitioned Runge-Kutta methods using IMEX RK methds. More details can be found in [20]. Therein, the following IMEX Runge-Kutta discretization of (4) is studied.

$$\begin{aligned} Y_n^{(i)} &= y_n + h \textstyle\sum_{j=1}^{i-1} \tilde{a}_{ij} D_x g(Y_n^{(j)}) + h \sum_{j=1}^{i} a_{ij} \frac{1}{\epsilon} r(Y_n^{(j)}) \qquad i = 1,..,s \\ y_{n+1} &= y_n + h \textstyle\sum_{i=1}^{s} \tilde{\omega}_i D_x g(Y_n^{(i)}) + h \sum_{i=1}^{s} \omega_i \frac{1}{\epsilon} r(Y_n^{(i)}), \qquad n = 0,1,2,. \end{aligned} \tag{6}$$

A nonlinear variable transformation and two intermediate states $\tilde{K}_n^{(i)}$ and $K_n^{(i)}$ give the equivalent system

$$\begin{aligned} \tilde{K}_n^{(i)} &= D_x g\left(y_n + h\sum_{j=1}^s \tilde{a}_{ij}\tilde{K}_n^{(j)} + h\sum_{j=1}^s a_{ij}K_n^{(j)}\right) \qquad i = 1,..,s \\ K_n^{(i)} &= \tfrac{1}{\epsilon} r\left(y_n + h\sum_{j=1}^s \tilde{a}_{ij}\tilde{K}_n^{(j)} + h\sum_{j=1}^s a_{ij}K_n^{(j)}\right) \qquad i = 1,..,s \\ y_{n+1} &= y_n + h\sum_{i=1}^s \tilde{\omega}_i\tilde{K}_n^{(i)} + h\sum_{i=1}^s \omega_i K_n^{(i)}, \qquad n = 0,1,2,. \end{aligned} \tag{7}$$

The associated optimality systems for the two previous optimization problems then coincide and we refer to [20] for mor details. It is proven that the adjoint schemes are equivalent to

$$\tilde{P}^{(i)} = p_n - h\sum_{j=1}^s \tilde{\alpha}_{ij}\, g'(Y_n^{(j)})^T \bar{D}_x \tilde{P}^{(j)} - h\sum_{j=1}^s \alpha_{ij}\,\frac{1}{\epsilon} r'(Y_n^{(j)})^T P^{(j)} \qquad i = 1,..,s$$

$$P^{(i)} = p_n - h\sum_{j=1}^s \tilde{\beta}_{ij}\, g'(Y_n^{(j)})^T \bar{D}_x \tilde{P}^{(j)} - h\sum_{j=1}^s \beta_{ij}\,\frac{1}{\epsilon} r'(Y_n^{(j)})^T P^{(j)} \qquad i = 1,..,s \tag{8}$$

$$p_{n+1} = p_n - h\sum_{i=1}^s \tilde{\omega}_i g'(Y_n^{(i)})^T \bar{D}_x \tilde{P}^{(i)} - h\sum_{i=1}^s \omega_i \frac{1}{\epsilon} r'(Y_n^{(i)})^T P^{(i)} \qquad n = 0,1,..,N-1$$

Here, the coefficients of the Runge-Kutta method $\tilde{\alpha}_{ij}, \alpha_{ij}, \tilde{\beta}_{ij}$ and β_{ij} are given by

$$\tilde{\alpha}_{ij} := \tilde{\omega}_j - \frac{\tilde{\omega}_j}{\tilde{\omega}_i}\tilde{a}_{ji}, \qquad \alpha_{ij} := \omega_j - \frac{\omega_j}{\tilde{\omega}_i}\tilde{a}_{ji}, \qquad \tilde{\beta}_{ij} := \tilde{\omega}_j - \frac{\tilde{\omega}_j}{\omega_i}a_{ji}, \qquad \beta_{ij} := \omega_j - \frac{\omega_j}{\omega_i}a_{ji}.$$

2.1 Properties of Discrete IMEX-RK Optimality System

For the resulting scheme (6),(8) order conditions can be stated [20]. To this end we add a suitable equation for $\tilde{p}$ to the previous system.

$$\tilde{p}_{n+1} = \tilde{p}_n - h\sum_{i=1}^s \tilde{\omega}_i\, f_y(Y_n^{(i)})^T \tilde{P}^{(i)} - h\sum_{i=1}^s \omega_i\, g_y(Y_n^{(i)})^T P^{(i)}. \tag{9}$$

The full method therefore is a standard additive Runge-Kutta scheme for

$$\begin{aligned} \mathbf{y}' &= -D_x\, g(\mathbf{y}) + \frac{1}{\epsilon} r(\mathbf{y}) \\ \tilde{\mathbf{p}}' &= g'(\mathbf{y})^T D_x \tilde{\mathbf{p}} + \frac{1}{\epsilon} r'(\mathbf{y})^T \mathbf{p} \\ \mathbf{p}' &= g'(\mathbf{y})^T D_x \tilde{\mathbf{p}} + \frac{1}{\epsilon} r'(\mathbf{y})^T \mathbf{p} \end{aligned}$$

If we define

$$c_i := \sum_{j=1}^s a_{ij}, \qquad \text{and} \qquad \tilde{c}_i := \sum_{j=1}^s \tilde{a}_{ij},$$

$$\gamma_i := \sum_{j=1}^{s} \alpha_{ij}, \qquad \text{and} \qquad \tilde{\gamma}_i := \sum_{j=1}^{s} \tilde{\alpha}_{ij},$$

$$\delta_i := \sum_{j=1}^{s} \beta_{ij}, \qquad \text{and} \qquad \tilde{\delta}_i := \sum_{j=1}^{s} \tilde{\beta}_{ij}$$

then (1) holds true.

Theorem 1. *Consider the Runge-Kutta scheme (6),(8),(9). This scheme is of*

- ***First-Order*** *: if (SRK1) is of first order*
- ***Second-Order*** *: if (SRK1) is of second order*
- ***Third-Order*** *: if (SRK1) is of third order and either*

$$\sum_{i=1}^{s} \omega_i \gamma_i^2 = \frac{1}{3}, \qquad \sum_{i=1}^{s} \omega_i \tilde{\gamma}_i^2 = \frac{1}{3}, \qquad \sum_{i=1}^{s} \omega_i \gamma_i \tilde{\gamma}_i = \frac{1}{3},$$

are satisfied or if

$$\sum_{i=1}^{s} \omega_i a_{ij} \gamma_i = \frac{1}{6}, \qquad \sum_{i=1}^{s} \omega_i \tilde{a}_{ij} \tilde{\gamma}_i = \frac{1}{6}$$

and if

$$\sum_{i=1}^{s} \omega_i a_{ij} \tilde{\gamma}_i = \frac{1}{6} \qquad \text{or} \qquad \sum_{i=1}^{s} \omega_i \tilde{a}_{ij} \gamma_i = \frac{1}{6}$$

are satisfied.

Note that the system (6) and (8) is not completely coupled, since the forward scheme (6) is solved independently of the adjoint scheme (8). General order conditions can be found e.g. in [23]. The proof of Theorem 1 and together with more details are discussed in [20].

3 Numerical Results

3.1 Scalar Example

As a simple example, we use a tracking type functional $J(u)$ together with Burgers' equation

$$u_t + \left(\frac{u^2}{2}\right)_x = 0,$$

and the desired state u_d at final time $T = 2.0$, that belongs to the initial condition $u_d(0, x) = \frac{1}{2} + \sin(x)$ and we start the optimization with the initial data

$u^s(0,x) \equiv 0.5$. Moreover, the spatial interval is given by $x \in [0, 2\pi]$, As discretization of the objective functional, we use

$$J(u(\cdot,T), u_0, u_d) = \frac{\Delta x}{2} \sum_{i=1}^{K} \|u_i - u_{d\,i}\|^2.$$

Moreover, the discrete gradient of the reduced cost functional is given by

$$\nabla_{u_{0,i}} \tilde{J} = p_{0,i} + (Df(u_0)^T \mathbf{q}_0)_i.$$

In order to solve the optimal control problem, we apply a steepest descent method (with respect to the reduced cost functional) with fixed stepsize $0 < \alpha < 1$, i.e. we set $u_0^{k+1} = u_0^k + \alpha \nabla_{u_{0,i}} \tilde{J}$. As stopping criterion for the optimization process we test $|\tilde{J}(u_0, u_d)| < tol$ where $tol = 1E-2$ denotes a predefined stopping tolerance. We observe grid independence in the case where u and u_0 are differentiable in space and time.

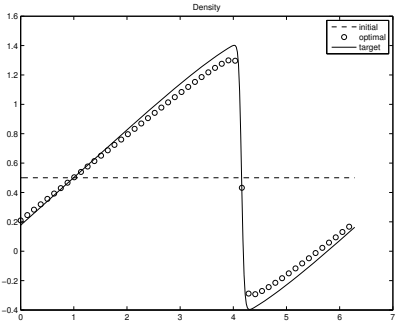
As first-order scheme, we test the Implicit-Explicit Euler scheme

$$\begin{aligned} u_i^* &= u_i^n \\ v_i^* &= v_i^n - \tfrac{\Delta t}{\epsilon}(v_i^* - f(u_i^*)) \\ u_i^{n+1} &= u_i^* - \Delta t\, D_x v_i^* \\ v_i^{n+1} &= v_i^* - \Delta t\, a^2 D_x v_i^* \end{aligned}$$

for the forward, as well as for the backward

$$\begin{aligned} q_i^* &= q_i^{n+1} - \Delta t D_x^* p_i^{n+1} \\ p_i^* &= p_i^{n+1} - \Delta t\, a^2 D_x^* q_i^{n+1} \\ q_i^n &= q_i^* - \tfrac{\Delta t}{\epsilon} q_i^n \\ p_i^n &= p_i^* + \tfrac{\Delta t}{\epsilon} q_i^n f'(u_i^n) \end{aligned}$$

The spatial gridsize is chosen to be $N_x = 300$, whereas the time discretization is done according to the CFL condition with constant $c_{CFL} = 0.5$.



N	Nr. of It.	CPU time (in sec.)
100	44	1.795572e+01
150	43	3.768984e+01
200	42	6.380441e+01
300	41	1.491838e+02

4 Summary

We briefly discussed a class of numerical methods applied to an optimal control problem for scalar, hyperbolic partial differential equations. Order conditions for the temporal numerical discretization in the case of differentiable functions have been stated. Future work includes the analysis of additional properties of the derived numerical discretizations as for example strong stability and asymptotic preservation properties.

Acknowledgments. This work has been supported by DFG HE5386/7-1, HE5386/8-1 and by DAAD 50727872, 50756459 and 54365630. We also acknowledge the support of Ateneo Italo-Tedesco (AIT) under the Vigoni project 2010-2012 'Adjoint implicit- -explicit methods for the numerical solution to optimization problems'.

Appendix

The discrete adjoint equations that correspond to the discrete optimization problem associated with (7) are

$$\tilde{\xi}_n^{(i)} = h\,\tilde{\omega}_i\,p_{n+1} + h\sum_{j=1}^{s}\tilde{a}_{ji}g'(Y_n^{(j)})^T\bar{D}_x\tilde{\xi}_n^{(j)} + h\sum_{j=1}^{s}\tilde{a}_{ji}\frac{1}{\epsilon}r'(Y_n^{(j)})^T\xi_n^{(j)}$$

$$\xi_n^{(i)} = h\,\omega_i\,p_{n+1} + h\sum_{j=1}^{s}a_{ji}g'(Y_n^{(j)})^T\bar{D}_x\tilde{\xi}_n^{(j)} + h\sum_{j=1}^{s}a_{ji}\frac{1}{\epsilon}r'(Y_n^{(j)})^T\xi_n^{(j)}$$

$$p_n = p_{n+1} + \sum_{i=1}^{s}g'(Y_n^{(i)})^T\bar{D}_x\tilde{\xi}_n^{(i)} + \sum_{i=1}^{s}\frac{1}{\epsilon}r'(Y_n^{(i)})^T\cdot\xi_n^{(i)}$$

$$p_N = j'(y_N, y^0)\,.$$

Moreover, the variable transformation that is needed to obtain (8) is given by

$$\tilde{P}_n^{(i)} := \frac{\tilde{\xi}_n^{(i)}}{h\,\tilde{\omega}_i} \qquad \text{and} \qquad P_n^{(i)} := \frac{\xi_n^{(i)}}{h\,\omega_i} \qquad (i = 1,..,s; \quad n = 0,..,N-1)\,.$$

On the other hand, using (6) the associated discrete adjoint equations are

$$\zeta_n^{(i)} = h\left(\tilde{\omega}_i f_y(Y_n^{(i)}) + \omega_i g_y(Y_n^{(i)})\right)^T p_{n+1} + \sum_{j=1}^{s}\tilde{a}_{ji}\,f_y(Y_n^{(i)})^T\zeta^{(j)}$$

$$+h\sum_{j=1}^{s} a_{ji}\, g_y(Y_n^{(i)})^T \zeta^{(j)} \qquad\qquad i=1,..,s$$

$$p_n = p_{n+1} + \sum_{i=1}^{s} \zeta_n^{(i)}, \qquad\qquad i=1,..,N-1, \qquad p_N = j'(y_N)$$

which can be transformed into the scheme (8) using the variable transformation

$$\tilde{P}_n^{(i)} := p_{n+1} + \sum_{j=1}^{s} \frac{\tilde{a}_{ji}}{\tilde{\omega}_i}\zeta_n^{(j)} \qquad \text{and} \qquad P_n^{(i)} := p_{n+1} + \sum_{j=1}^{s} \frac{a_{ji}}{\omega_i}\zeta_n^{(j)}\,.$$

References

1. Aregba-Driollet, D., Natalini, R.: Discrete kinetic schemes for multidimensional systems of conservation laws. SIAM J. Numer. Anal. 37, 1973–2004 (2000) (electronic)
2. Areba-Driollet, D., Natalini, R.: Convergence of relaxation schemes for conservation laws. Applicable Analysis 61(1), 163–193 (1996)
3. Ascher, U., Ruuth, S., Spiteri, R.: Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations. Applied Numerical Mathematics 25, 151–167 (1997)
4. Banda, M.K., Herty, M.: Adjoint imex–based schemes for control problems governed by hyperbolic conservation laws. Computational Optimization and Applications (2010)
5. Banda, M.K., Seaïd, M.: Higher-order relaxation schemes for hyperbolic systems of conservation laws. J. Numer. Math. 13(3), 171–196 (2005)
6. Bianchini, S.: On the shift differentiability of the flow generated by a hyperbolic system of conservation laws. Discrete Contin. Dynam. Systems 6, 329–350 (2000)
7. Bonnans, J.F., Laurent-Varin, J.: Computation of order conditions for symplectic partitioned Runge-Kutta schemes with application to optimal control. Numerische Mathematik 103, 1–10 (2006)
8. Boscarino, S., Pareschi, L., Russo, G.: Implicit-Explicit Runge-Kutta schemes for hyperbolic systems and kinetic equations in the diffusion limit (2011) (preprint)
9. Bressan, A., Guerra, G.: Shift-differentiability of the flow generated by a conservation law. Discrete Contin. Dynam. Systems 3, 35–58 (1997)
10. Bressan, A., Lewicka, M.: Shift differentials of maps in BV spaces. In: Nonlinear Theory of Generalized Functions (Vienna, 1997). Chapman & Hall/CRC Res. Notes Math., vol. 401, pp. 47–61. Chapman & Hall/CRC, Boca Raton, FL (1999)
11. Bressan, A., Marson, A.: A variational calculus for discontinuous solutions to conservation laws. Communications Partial Differential Equations 20, 1491–1552 (1995)
12. Bressan, A., Shen, W.: Optimality conditions for solutions to hyperbolic balance laws. In: Control Methods in PDE-Dynamical Systems. Contemp. Math., vol. 426, pp. 129–152 (2007)
13. Dimarco, G., Pareschi, L.: Asymptotic-Preserving IMEX Runge-Kutta methods for nonlinear kinetic equations (2011) (preprint)
14. Gottlieb, S., Shu, C.W., Tadmor, E.: Strong stability preserving high-order time discretization methods. SIAM Rev. 43, 89–112 (2001)

15. Dontchev, A.L., Hager, W.W.: The Euler approximation in state constrained optimal control. Math. Comp. 70, 173–203 (2001)
16. Dontchev, A.L., Hager, W.W., Veliov, V.M.: Second–order Runge–Kutta approximations in control constrained optimal control. SIAM J. Numer. Anal. 38, 202–226 (2000)
17. Hager, W.W.: Runge-Kutta methods in optimal control and the transformed adjoint system. Numerische Mathematik 87, 247–282 (2000)
18. Hairer, E., Nørsett, S.P., Wanner, G.: Solving Ordinary Differential Equations, Part I, Nonstiff Problems, 2nd edn. Springer Series in Computational Mathematics (1993)
19. Herty, M., Schleper, V.: Time discretizations for numerical optimization of hyperbolic problems. Applied Mathematics and Computation (2011)
20. Herty, M., Pareschi, L., Steffensen, S.: Implicit–Explicit Runge-Kutta schemes for numerical discretization of optimal control problems (preprint available at University Ferrara, 2012)
21. Giles, M., Ulbrich, S.: Convergence of linearized and adjoint approximations for discontinuous solutions of conservation laws. Part 2: Adjoint approximations and extensions. SIAM J. Numer. Anal. 48, 905–921 (2010)
22. Jin, S., Xin, Z.P.: The relaxation schemes for systems of conservation laws in arbitrary space dimensions. Comm. Pure Appl. Math. 48, 235–276 (1995)
23. Kennedy, C.A., Carpenter, M.H.: Additive Runge-Kutta schemes for convection-diffusion-reaction equations. Appl. Num. Math. 44, 139–181 (2003)
24. Lang, J., Verwer, J.: W-Methods in optimal control, TU Darmstadt (2011) (preprint)
25. Natalini, R., Terracina, A.: Convergence of a relaxation approximation to a boundary value problem for conservation laws. Comm. Partial Differential Equations 26(7-8), 1235–1252 (2001)
26. Pareschi, L., Russo, G.: Implicit-explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxation. J. Sci. Comput. 25, 129–155 (2005)
27. Pareschi, L., Russo, G.: Implicit-explicit Runge-Kutta schemes for stiff systems of differential equations. In: Brugnano, L., Trigiante, D. (eds.) Recent Trends in Numerical Analysis, vol. 3, pp. 269–289 (2000)
28. Ulbrich, S.: Optimal Control of Nonlinear Hyperbolic Conservation Laws with Source Terms, Technische Universitaet Muenchen (2001)
29. Ulbrich, S.: On the superlinear local convergence of a filer-sqp method. Technical Report (2002)
30. Ulbrich, S.: Adjoint-based derivative computations for the optimal control of discontinuous solutions of hyperbolic conservation laws. Syst. Control Lett. 48, 313–328 (2003)
31. Walther, A.: Automatic differentiation of explicit Runge–Kutta methods for optimal control. J. Comp. Opt. Appl. 36, 83–108 (2007)
32. Castro, C., Palacios, F., Zuazua, E.: An alternating descent method for the optimal control of the inviscid Burgers equation in the presence of shocks. Math. Models Methods Appl. Sci. 18, 369–416 (2008)

Necessary Conditions for Convergence Rates of Regularizations of Optimal Control Problems

Daniel Wachsmuth and Gerd Wachsmuth

Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenbergerstraße 69, A–4040 Linz, Austria, Chemnitz University of Technology, Department of Mathematics, D-09107 Chemnitz, Germany

Abstract. We investigate the Tikhonov regularization of control constrained optimal control problems. We use a specialized source condition in combination with a condition on the active sets. In the case of high convergence rates, these conditions are necessary and sufficient.

Keywords: optimal control problem, inequality constraints, Tikhonov regularization, source condition.

1 Introduction

In this article, we investigate regularization schemes for the following class of optimization problems:

$$\begin{aligned} &\text{Minimize} && \frac{1}{2}\|\mathcal{S}u - z\|_Y^2 + \beta\|u\|_{L^1(\Omega)} \\ &\text{such that} && u \in L^2(\Omega) \quad \text{and} \quad u_a \le u \le u_b \text{ a.e. on } \Omega. \end{aligned} \tag{P}$$

Here, Ω is a measurable subset of $\mathbb{R}^n$, $n \ge 1$, Y is a Hilbert space, $\mathcal{S} : L^2(\Omega) \to Y$ a bounded linear operator, and the function $z \in Y$ is given. The parameter β is assumed to be non-negative. The control constraints $u_a, u_b \in L^\infty(\Omega)$ satisfy $u_a \le 0 \le u_b$.

This model problem can be interpreted as an optimal control problem as well as an inverse problem. In the point of view of inverse problems, the unknown u has to be constructed in order to reproduce given measurements z. The inequality constraints on u reflect certain a-priori knowledge about the solution $u^\dagger$ of the linear ill-posed equation $\mathcal{S}u = z$. If the problem at hand is seen as an optimal control problem, then u is the control, $\mathcal{S}u$ the state of the system, which has to be close to a desired state z, the inequality constraints restrict the feasible set and may hinder the state $\mathcal{S}u$ to reach the target z. If the parameter β is positive, then the resulting optimal control will be sparse, that is, its support is a possibly small subset of Ω.

The resulting optimization problem (P) is nevertheless ill-posed if $\mathcal{S}$ is not continuously invertible. Due to the control constraints, problem (P) still possesses a solution, which is even unique if $\mathcal{S}$ is injective. However, the solution

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 145–154, 2013.

may be unstable with respect to perturbations in the problem data, for instance in the given state z. Here small perturbations due to measurement errors may lead to large changes in the solution. Consequently, any numerical approximation of (P) is challenging to solve and numerical approximations of solutions may converge arbitrarily slow. Let us note, that a positive value of β does not make the problem well-posed. This is due to the fact, that $L^1(\Omega)$ is not a dual space and hence bounded sets in $L^1(\Omega)$ are not compact w.r.t. the weak(-star) topology, see also the discussions in [7,8].

In order to overcome this difficulty, we apply common ideas from inverse problem theory. We will study a regularization of the type

$$\begin{aligned} &\text{Minimize} \quad \frac{1}{2}\|\mathcal{S}u - z\|_Y^2 + \beta\|u\|_{L^1(\Omega)} + \frac{\alpha}{2}\|u\|_{L^2(\Omega)}^2 \\ &\text{such that} \quad u \in L^2(\Omega) \quad \text{and} \quad u_a \le u \le u_b \text{ a.e. on } \Omega, \end{aligned} \tag{P_α}$$

where $\alpha > 0$ is given. Clearly, the problems (P_α) are uniquely solvable for $\alpha > 0$. Now, the question arises, whether their solutions u_α converge (weakly or strongly) to a solution u_0 of (P) for $\alpha \to 0$. Moreover, in the case of convergence, one is interested in proving convergence rates of $\|u_\alpha - u_0\|_{L^2(\Omega)}$ and $\|\mathcal{S}u_\alpha - \mathcal{S}u_0\|_Y$ under suitable assumptions.

In this work, we will prove necessary conditions for convergence rates. In some parts, the necessary conditions are similar to sufficient conditions found in earlier works [7,8]. Moreover, the result of Theorem 3 leads to a weakened sufficient condition for convergence rates.

1.1 Standing Assumptions and Notation

Let us fix the standing assumptions on the problem (P). We assume that $\mathcal{S} : L^2(\Omega) \to Y$ is linear and continuous. In many applications this operator $\mathcal{S}$ is compact. Furthermore, we assume that the Hilbert space adjoint operator $\mathcal{S}^\star$ maps into $L^\infty(\Omega)$, i.e., $\mathcal{S}^\star \in \mathcal{L}(Y, L^\infty(\Omega))$. These assumptions imply that the range of $\mathcal{S}$ is closed in Y if and only if the range of $\mathcal{S}$ is finite-dimensional, see [8, Prop. 2.1]. Hence, up to trivial cases, (P) is ill-posed. A typical example for $\mathcal{S}$ is the solution operator of the Poisson problem with homogeneous Dirichlet boundary conditions.

The set of feasible functions u is given by

$$U_{ad} := \{u \in L^2(\Omega) : \ u_a \le u \le u_b \text{ a.e. on } \Omega\}.$$

The problem (P_α) is uniquely solvable for $\alpha > 0$. We will denote its solution by u_α, with the corresponding state $y_\alpha := \mathcal{S}u_\alpha$ and adjoint state $p_\alpha := \mathcal{S}^\star(z - y_\alpha)$. There is a unique solution of (P) with minimal $L^2(\Omega)$ norm, see [8, Thm. 2.3, Lem. 2.7]. This solution and the associated state and adjoint state will be denoted by u_0, y_0 and p_0, respectively. Note that the weak convergence $u_\alpha \rightharpoonup u^\star$ in $L^2(\Omega)$, where $u^\star$ is a solution of (P) already implies $u^\star = u_0$, see [8, Rem. 3.3].

1.2 Optimality Conditions

As both problems (P) and (P_α) are convex, their solutions can be characterized by the following necessary and sufficient optimality conditions:

Theorem 1 ([7, Lemma 2.2]). *Let $\alpha \geq 0$ be given, and let u_α be a solution of (P_α) (or (P) in the case $\alpha = 0$).*

Then, there exists a subgradient $\lambda_\alpha \in \partial \|u_\alpha\|_{L^1(\Omega)}$, such that the variational inequality

$$(\alpha\, u_\alpha - p_\alpha + \beta\, \lambda_\alpha,\, u - u_\alpha) \geq 0 \quad \forall u \in U_{ad}, \tag{1}$$

is satisfied, where $p_\alpha = \mathcal{S}^\star(z - \mathcal{S}u_\alpha)$ is the associated adjoint state.

Here, $(\cdot, \cdot)$ refers to the scalar product in $L^2(\Omega)$.

Standard arguments (see [6, Section 2.8]) lead to a pointwise a.e. interpretation of the variational inequality, which in turn implies the following relation between u_α and p_α in the case $\alpha > 0$:

$$u_\alpha(x) = \begin{cases} u_a(x) & \text{if } p_\alpha(x) < \alpha\, u_a(x) - \beta \\ \frac{1}{\alpha}(p_\alpha(x) + \beta) & \text{if } \alpha\, u_a(x) - \beta \leq p_\alpha(x) \leq -\beta \\ 0 & \text{if } |p_\alpha(x)| < \beta \\ \frac{1}{\alpha}(p_\alpha(x) - \beta) & \text{if } \beta \leq p_\alpha(x) \leq \alpha\, u_b(x) + \beta \\ u_b(x) & \text{if } \alpha\, u_b(x) + \beta < p_\alpha(x) \end{cases} \quad \text{a.e. on } \Omega. \tag{2}$$

In the case $\alpha = 0$, we have

$$u_0(x) \begin{cases} = u_a(x) & \text{if } p_0(x) < -\beta \\ \in [u_a(x), 0] & \text{if } p_0(x) = -\beta \\ = 0 & \text{if } |p_0(x)| < \beta \\ \in [0, u_b(x)] & \text{if } p_0(x) = \beta \\ = u_b(x) & \text{if } \beta < p_0(x) \end{cases} \quad \text{a.e. on } \Omega. \tag{3}$$

Note that if $\beta = 0$, one obtains $u_0(x) \in [u_a(x), u_b(x)]$ where $p_0(x) = 0$ in (3). This implies that $u_0(x)$ is uniquely determined by $p_0(x)$ on the set, where it holds $|p_0(x)| \neq \beta$.

2 Sufficient Conditions for Convergence Rates

Let us first recall the sufficient conditions for convergence rates as obtained in [8]. We will work with the following assumption. There we denote by $\mathrm{proj}_{[a,b]}(v)$ the projection of the real number v onto the interval $[a, b]$.

Assumption 2. *Let u_0 be a solution of (P). Let us assume that there exist a measurable set $I \subset \Omega$, a function $w \in Y$, and positive constants κ, c such that it holds:*

1. **(source condition)** *$I \supset \{x \in \Omega : |p_0(x)| = \beta\}$, and for almost all $x \in I$*

$$u_0(x) = \begin{cases} \operatorname{proj}_{[u_a(x),0]}\big((\mathcal{S}^\star w)(x)\big) & \text{if } \beta > 0,\ p_0(x) \le -\frac{\beta}{2}, \\ \operatorname{proj}_{[0,u_b(x)]}\big((\mathcal{S}^\star w)(x)\big) & \text{if } \beta > 0,\ p_0(x) \ge \frac{\beta}{2}, \\ \operatorname{proj}_{[u_a(x),u_b(x)]}\big((\mathcal{S}^\star w)(x)\big) & \text{if } \beta = 0. \end{cases} \tag{4}$$

2. **(structure of active set)** *$A = \Omega \setminus I$ and for all $\epsilon > 0$*

$$\begin{aligned} \operatorname{meas}\big(\{x \in A :\ 0 < \big||p_0(x)| - \beta\big| < \epsilon\}\big) &\le c\,\epsilon^\kappa \quad \text{if } w \ne 0, \\ \operatorname{meas}\big(\{x \in A :\ 0 < \ |p_0(x)| - \beta\ < \epsilon\}\big) &\le c\,\epsilon^\kappa \quad \text{if } w = 0. \end{aligned} \tag{5}$$

Some remarks are in order. The first part of the assumption is analogous to source conditions in inverse problems: we assume that on the set $I \subset \Omega$ the solution u_0 is the restriction to I of a certain pointwise projection of an element in the range of $\mathcal{S}^\star$. This part of the condition is different from other conditions in the literature: in our earlier work [8] we used the assumption $u_0(x) = \operatorname{proj}_{[u_a(x),u_b(x)]}\big((\mathcal{S}^\star w)(x)\big)$ on I. However, in the light of the derivation of necessary conditions it turns out that such a condition can be weakened without losing anything with respect to convergence rates. In works on inverse problems [3,5], the source condition $u_0 = \operatorname{proj}_{U_{ad}}(\mathcal{S}^\star w)$ is used, which is retained as the special case $I = \Omega$ in Assumption 2.

The assumption (5) (without the second alternative) on the active sets was already employed to obtain regularization error estimates [7,8], error estimates for finite-element discretizations of (P) [2], as well as stability results of bang-bang controls [4]. Note that in the case $\beta = 0$, both conditions in (5) are equivalent. However, if $\beta > 0$ and $w = 0$ (in particular, if I has measure zero), the second alternative provides a weaker condition than the first one. Hence, condition (5) is weaker than the condition used in our earlier work [8].

Theorem 3. *Let Assumption 2 be satisfied.*
Let d be defined as

$$d = \begin{cases} \frac{1}{2-\kappa} & \text{if } \kappa \le 1, \\ 1 & \text{if } \kappa > 1 \text{ and } w \ne 0, \\ \frac{\kappa+1}{2} & \text{if } \kappa > 1 \text{ and } w = 0. \end{cases}$$

Then there is $\alpha_{max} > 0$ and a constant $c > 0$, such that

$$\begin{aligned} \|y_0 - y_\alpha\|_Y &\le c\,\alpha^d \\ \|p_0 - p_\alpha\|_{L^\infty(\Omega)} &\le c\,\alpha^d \\ \|u_0 - u_\alpha\|_{L^2(\Omega)} &\le c\,\alpha^{d-1/2} \end{aligned}$$

holds for all $\alpha \in (0, \alpha_{max}]$.

Under the assumptions of the theorem, one can prove also convergence rates for $\|u_\alpha - u_0\|_{L^1(A)}$ [8].

Proof. The proof is analogous to the proof of [8, Thm. 3.14]. We have to take into account the modification of the source condition (4) in the case $\beta > 0$ and the modification of (5) in the case $w = 0$. By [8, Lemma 2.12], we have

$$\|y_0 - y_\alpha\|_Y^2 + \alpha \|u_0 - u_\alpha\|_{L^2(\Omega)}^2 \le \alpha\,(u_0,\, u_0 - u_\alpha). \tag{6}$$

Since U_{ad} is bounded, we obtain $\|p_0 - p_\alpha\|_{L^\infty(\Omega)} \le c\,\alpha^{1/2}$ for some $c > 0$ independent of α.

Let now α be small enough such that $\|p_0 - p_\alpha\|_{L^\infty(\Omega)} < \beta/2$. This implies that p_0 and p_α have the same sign on the set $\{x \in I :\ |p_0(x)| \ge \beta/2\}$. Consequently, u_0 and u_α have the same sign on this set, too. Moreover, on the set $\{x \in I :\ |p_0(x)| < \beta/2\}$ it holds $|p_\alpha| < \beta$, and hence $u_\alpha = 0 = u_0$ on this set. This yields

$$(\chi_I u_0,\, u_0 - u_\alpha) \le (\chi_I \mathcal{S}^\star w,\, u_0 - u_\alpha)$$

for $\alpha > 0$ small enough. Note that in case of $w = 0$, the right-hand side in the previous estimate vanishes and it remains to estimate $(\chi_A\, u_0,\, u_0 - u_\alpha)$. Taking into account that $u_0(x) = 0$ whenever $|p_0(x)| < \beta$, the weakend estimate (5) is sufficient in this case. Arguing as in the proof of [8, Thm. 3.14] proves the claim. □

3 Necessary Conditions for Convergence Rates

3.1 Necessity of the Source Condition (4)

Theorem 4. *Let us suppose that $\|y_\alpha - y_0\|_Y = O(\alpha)$ with $y_0 = \mathcal{S}u_0$. Then there exists $w \in Y$ such that*

$$u_0(x) = \begin{cases} \operatorname{proj}_{[u_a(x),0]}\big((\mathcal{S}^\star w)(x)\big) & \text{if } \beta > 0,\ p_0(x) = -\beta, \\ \operatorname{proj}_{[0,u_b(x)]}\big((\mathcal{S}^\star w)(x)\big) & \text{if } \beta > 0,\ p_0(x) = +\beta, \\ \operatorname{proj}_{[u_a(x),u_b(x)]}\big((\mathcal{S}^\star w)(x)\big) & \text{if } \beta = 0,\ p_0(x) = 0. \end{cases}$$

If moreover $\|y_\alpha - y_0\|_Y = o(\alpha)$, then $u_0 = 0$ on $\{x \in \Omega :\ |p_0(x)| = \beta\}$, i.e. $w = 0$.

This result shows that the source condition (4) is necessary on the set $\{x \in \Omega :\ |p_0(x)| = \beta\}$.

Proof. Let us prove the claim in the case $\beta > 0$. The result in the case $\beta = 0$ can be proved with obvious modifications. Let us take a test function $u \in U_{ad}$ defined as

$$u(x) \begin{cases} = u_a(x) & \text{if } p_0(x) < -\beta, \\ \in [u_a(x), 0] & \text{if } p_0(x) = -\beta, \\ = 0 & \text{if } |p_0(x)| < \beta, \\ \in [0, u_b(x)] & \text{if } p_0(x) = \beta, \\ = u_b(x) & \text{if } p_0(x) > \beta. \end{cases}$$

Due to the relation

$$\lambda_0 = \mathrm{proj}_{[-1,1]}\Big(\frac{1}{\beta}p_0\Big), \tag{7}$$

which is a consequence of the necessary optimality condition, see [1] for a proof, we obtain $\lambda_0 = \pm 1$ where $p_0 = \pm\beta$. Hence it holds

$$(-p_0,\ u - u_0) = (-\beta\,\lambda_0,\ u - u_0) = \beta\,\|u_0\|_{L^1(\Omega)} - \beta\,\|u\|_{L^1(\Omega)} \tag{8}$$

for u as above.

Since $\lambda_0 \in \partial\|u_0\|_{L^1(\Omega)}$, we obtain

$$(\lambda_0,\ u_\alpha - u_0) \le \|u_\alpha\|_{L^1(\Omega)} - \|u_0\|_{L^1(\Omega)}. \tag{9}$$

Using the optimality of u_α and the relation $-p_\alpha = -p_0 + \mathcal{S}^\star\mathcal{S}(u_\alpha - u_0)$ we get

$$(-p_0 + \mathcal{S}^\star\mathcal{S}(u_\alpha - u_0) + \alpha\,u_\alpha,\ u - u_\alpha) + \beta\,\|u\|_{L^1(\Omega)} - \beta\,\|u_\alpha\|_{L^1(\Omega)} \ge 0.$$

Adding $(-p_0 + \beta\,\lambda_0,\ u_\alpha - u_0) \ge 0$ to the left-hand side yields

$$\begin{aligned}(\mathcal{S}^\star\mathcal{S}(u_\alpha - u_0) + \alpha\,u_\alpha,\ u - u_\alpha) + (-p_0,\ u - u_0) + (\beta\,\lambda_0,\ u_\alpha - u_0)\\ +\beta\,\|u\|_{L^1(\Omega)} - \beta\,\|u_\alpha\|_{L^1(\Omega)} \ge 0.\end{aligned}$$

Using (8) and (9) we obtain

$$(\mathcal{S}^\star\mathcal{S}(u_\alpha - u_0) + \alpha\,u_\alpha,\ u - u_\alpha) \ge 0.$$

Due to the assumptions of the theorem, the functions $\frac{1}{\alpha}(\mathcal{S}(u_\alpha - u_0)) = \frac{1}{\alpha}(y_\alpha - y_0)$ are uniformly bounded for $\alpha \searrow 0$. As a consequence, $\alpha \searrow 0$ implies

$$(\mathcal{S}^\star\dot{y}_0 + u_0,\ u - u_0) \ge 0$$

for any weak subsequential limit $\dot{y}_0$ of $\frac{1}{\alpha}(y_\alpha - y_0)$. Due to the construction of the test function u, we obtain

$$u_0 = \begin{cases}\mathrm{proj}_{[u_a,0]}(\mathcal{S}^\star\dot{y}_0) & \text{where } p_0 = -\beta,\\ \mathrm{proj}_{[0,u_b]}(\mathcal{S}^\star\dot{y}_0) & \text{where } p_0 = +\beta.\end{cases}$$

If $\|y_\alpha - y_0\|_Y = o(\alpha)$ then $\frac{1}{\alpha}(y_\alpha - y_0) \to 0$ strongly in Y for $\alpha \to 0$, hence $\dot{y}_0 = 0$, and $u_0 = 0$ on the set $\{|p_0| = \beta\}$. □

As can be seen from the proof, the element that realizes the source condition can be interpreted as the (weak) directional derivate of $\alpha \mapsto y_\alpha$ at $\alpha = 0$.

The result of the theorem resembles known results of necessity of the source condition in linear inverse problems, see e.g. [3,5].

3.2 Necessity of the Condition (5) on the Active Set

In this section, we want to prove the necessity of (5) in the case of high convergence rates $d > 1$. In this case, we have $w = 0$, see Theorem 4. It remains to show that the second condition in (5) is necessary to obtain convergence rates $d > 1$. Hence, we derive a bound on

$$\mu(\epsilon) := \left|\{x \in \Omega : \ 0 < |p_0(x)| - \beta < \epsilon\}\right|,$$

which is the measure of a subset of

$$A := \{x \in \Omega : \ \beta < |p_0(x)|\}.$$

For $\alpha > 0$ let $\tilde{u}_\alpha$ denote the unique solution of

$$\min_{u \in U_{ad}} -(u, p_0) + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)} + \beta\|u\|_{L^1(\Omega)}. \tag{$\mathrm{P}_\alpha^{\mathrm{aux}}$}$$

Analogous to (2), we have the representation

$$\tilde{u}_\alpha(x) = \begin{cases} u_a(x) & \text{if } p_0(x) < \alpha\, u_a(x) - \beta \\ \frac{1}{\alpha}(p_0(x) + \beta) & \text{if } \alpha\, u_a(x) - \beta \le p_0(x) \le -\beta \\ 0 & \text{if } |p_0(x)| < \beta \\ \frac{1}{\alpha}(p_0(x) - \beta) & \text{if } \beta \le p_0(x) \le \alpha\, u_b(x) + \beta \\ u_b(x) & \text{if } \alpha\, u_b(x) + \beta < p_0(x) \end{cases} \quad \text{a.e. on } \Omega. \tag{10}$$

Let us first prove a relation between the convergence rates of $\|u_0 - \tilde{u}_\alpha\|_{L^2(A)}$ and $\mu(\epsilon)$ for $\alpha \to 0$ and $\epsilon \to 0$, respectively.

Lemma 5. *Let us assume that there is $\sigma > 0$ such that $u_a \le -\sigma < 0 < \sigma \le u_b$ a.e. on Ω. Then it holds: If $\|u_0 - \tilde{u}_\alpha\|_{L^2(A)} = O(\alpha^d)$, $d > 0$, for $\alpha \to 0$, then $\mu(\epsilon) = O(\epsilon^{2d})$ for $\epsilon \to 0$.*

Proof. Due to the pointwise representations of $\tilde{u}_\alpha$ and u_0 in (10) and (3), respectively, it holds

$$\|u_0 - \tilde{u}_\alpha\|^2_{L^2(A)} = \int_{\{\beta < p_0 < \alpha u_b + \beta\}} (u_b - \alpha^{-1}(p_0 - \beta))^2 + \int_{\{\alpha u_a - \beta < p_0 < -\beta\}} (u_a - \alpha^{-1}(p_0 + \beta))^2.$$

Due to the assumption on the control constraints we have

$$\begin{aligned} \int_{\{\beta < p_0 < \alpha u_b + \beta\}} (u_b - \alpha^{-1}(p_0 - \beta))^2 &\ge \int_{\{\beta < p_0 < \alpha\sigma/2 + \beta\}} (u_b - \alpha^{-1}(p_0 - \beta))^2 \\ &\ge \int_{\{\beta < p_0 < \alpha\sigma/2 + \beta\}} (\sigma/2)^2 \\ &\ge (\sigma/2)^2 \left|\{x \in \Omega : 0 < p_0(x) - \beta < \alpha\sigma/2\}\right|. \end{aligned}$$

Similarly, we obtain

$$\int_{\{\alpha u_a-\beta<p_0<-\beta\}} (u_a-\alpha^{-1}(p_0+\beta))^2 \geq (\sigma/2)^2 \big|\{x\in\Omega : 0<-p_0(x)-\beta<\alpha\sigma/2\}\big|.$$

This implies

$$\|u_0-\tilde{u}_\alpha\|^2_{L^2(A)} \geq (\sigma/2)^2\mu(\alpha\sigma/2).$$

Hence if $\|u_0-\tilde{u}_\alpha\|_{L^2(A)} = O(\alpha^d)$ holds, then

$$\mu(\alpha\sigma/2) \leq O(\alpha^{2d}),$$

for $\alpha\to 0$, which proves the claim. □

Using the same arguments, we can prove the following result.

Corollary 6. *Let the requirements of Theorem 5 be satisfied. Let $p\in[1,\infty)$ be given. Then it holds*

$$\left(\frac{\sigma}{2}\right)^p \mu\left(\frac{\sigma}{2}\alpha\right) \leq \|u_0-\tilde{u}_\alpha\|^p_{L^p(A)} \leq M^p\,\mu(M\alpha)$$

with $M=\max(\|u_a\|_{L^\infty(\Omega)},\|u_b\|_{L^\infty(\Omega)})$.

Lemma 7. *Let $\tilde{u}_\alpha$ be defined as above. Then it holds*

$$\alpha\|\tilde{u}_\alpha-u_\alpha\|^2_{L^2(\Omega)} + \|y_0-y_\alpha\|^2_Y \leq (p_0-p_\alpha,\,\tilde{u}_\alpha-u_0).$$

Proof. Since u_α and $\tilde{u}_\alpha$ solve (P_α) and $(\mathrm{P}^{\mathrm{aux}}_\alpha)$, respectively, we have

$$\begin{aligned}(\alpha u_\alpha - p_\alpha + \beta\lambda_\alpha,\,\tilde{u}_\alpha-u_\alpha)\geq 0,\\ (\alpha\tilde{u}_\alpha - p_0 + \beta\tilde{\lambda}_\alpha,\,u_\alpha-\tilde{u}_\alpha)\geq 0,\end{aligned}$$

with some $\tilde{\lambda}_\alpha\in\partial\|\tilde{u}_\alpha\|_{L^1(\Omega)}$. Due to the monotonicity of the subdifferential we have $(\lambda_\alpha-\tilde{\lambda}_\alpha,\,u_\alpha-\tilde{u}_\alpha)\geq 0$. This gives

$$\alpha\|\tilde{u}_\alpha-u_\alpha\|^2_{L^2(\Omega)} \leq (p_0-p_\alpha,\,\tilde{u}_\alpha-u_\alpha).$$

The identity

$$\begin{aligned}(p_0-p_\alpha,\,\tilde{u}_\alpha-u_\alpha) &= (p_0-p_\alpha,\,\tilde{u}_\alpha-u_0+u_0-u_\alpha)\\ &= (p_0-p_\alpha,\,\tilde{u}_\alpha-u_0) - \|y_0-y_\alpha\|^2_Y\end{aligned}$$

finishes the proof. □

Theorem 8. *Let us assume that there is $\sigma>0$ such that $u_a\leq-\sigma<0<\sigma\leq u_b$ a.e. on Ω. The we have the following implication: If*

$$\|u_0-u_\alpha\|_{L^2(\Omega)} = O(\alpha^{d-1/2}),\qquad \|y_0-y_\alpha\|_Y = O(\alpha^d)\ \textit{for}\ \alpha\to 0$$

holds with $d>1$, then it follows

$$\mu(\epsilon)\leq O(\epsilon^{2d-1})\ \textit{for}\ \epsilon\to 0.$$

Proof. Let us begin with

$$\begin{aligned}\|u_0 - \tilde{u}_\alpha\|^2_{L^2(\Omega)} &\leq 2(\|u_0 - u_\alpha\|^2_{L^2(\Omega)} + \|u_\alpha - \tilde{u}_\alpha\|^2_{L^2(\Omega)}) \\ &\leq O(\alpha^{2d-1}) + \alpha^{-1}(p_0 - p_\alpha, \tilde{u}_\alpha - u_0) \\ &\leq O(\alpha^{2d-1}) + O(\alpha^{d-1})\|u_0 - \tilde{u}_\alpha\|_{L^2(\Omega)},\end{aligned}$$

which gives $\|u_0 - \tilde{u}_\alpha\|_{L^2(\Omega)} = O(\alpha^{d-1})$. Hence by Theorem 5, we obtain $\mu(\epsilon) = O(\epsilon^{2d-2})$. Let us note that the convergence rates imply $u_0(x) = 0$ if $|p_0(x)| = \beta$ by Theorem 4. Moreover, we have $u_0 = \tilde{u}_\alpha = 0$ on $\{x \in \Omega : |p_0(x)| \leq \beta\}$ by (3) and (10). This implies $u_0 = \tilde{u}_\alpha = 0$ on the set $\{x \in \Omega : \ |p_0(x)| \leq \beta\} = \Omega \setminus A$, cf. (10). Using the convergence rate $\|p_0 - p_\alpha\|_{L^\infty(\Omega)} = O(\alpha^d)$ and Theorem 6, we find

$$\begin{aligned}\alpha^{-1}|(p_0 - p_\alpha, \tilde{u}_\alpha - u_0)| &= O(\alpha^{d-1})\|\tilde{u}_\alpha - u_0\|_{L^1(\Omega)} \\ &= O(\alpha^{d-1})\|\tilde{u}_\alpha - u_0\|_{L^1(A)} \\ &\leq O(\alpha^{d-1})\mu(M\,\alpha).\end{aligned}$$

Since by the above considerations we already got $\mu(\epsilon) = O(\epsilon^{2d-2})$ this gives

$$\|u_0 - \tilde{u}_\alpha\|^2_{L^2(\Omega)} = O(\alpha^{2d-1}) + O(\alpha^{3(d-1)}).$$

Repeating this process k times until $k(d-1) \geq 2d-1$ yields

$$\|u_0 - \tilde{u}_\alpha\|^2_{L^2(\Omega)} = O(\alpha^{2d-1}),$$

which finishes the proof. □

Together with Theorem 4, this result shows that the requirements of Theorem 3 for convergence rates $d > 1$ are sharp. It is an open question, whether the requirement (5) on the active set is also necessary for convergence rates $d \leq 1$. In our opinion, this condition is too strong and has to be relaxed in order to obtain a characterization for convergence rates $d \leq 1$.

3.3 Necessary Conditions for Exact Reconstruction with $\alpha > 0$

Let us now investigate the case of exact reconstruction. That is, the solutions of the regularized problem u_α coincide with the (minimal L^2-norm) solution u_0 of the original problem.

Lemma 9. *Let us assume that $u_{\alpha^*} = u_0$ a.e. on Ω for some $\alpha^* > 0$. Then $u_\alpha = u_0$ a.e. on Ω for all $\alpha \in (0, \alpha^*)$.*

Proof. The claim follows from known monotonicity results: The mapping $\alpha \mapsto \|u_\alpha\|_{L^2}$ is monotonically decreasing, while $\alpha \mapsto \frac{1}{2}\|y_\alpha - y_d\|^2_Y + \beta\|u_\alpha\|_{L^1}$ is monotonically increasing from $(0, +\infty)$ to $\mathbb{R}$, see e.g. [8, Lemma 2.8]. □

Theorem 10. *Let us assume that there is $\sigma > 0$ such that $u_a \le -\sigma < 0 < \sigma \le u_b$ a.e. on Ω. Then the exact recovery $u_{\alpha^*} = u_0$ a.e. on Ω for some $\alpha^* > 0$ is equivalent to*

$$\left.\begin{aligned} u_0 = 0 \quad \text{on } \{x \in \Omega :\ |p_0(x)| = \beta\} \qquad \text{and} \\ \mu(\epsilon) = \big|\{x \in \Omega :\ 0 < |p_0(x)| - \beta < \epsilon\}\big| = 0 \end{aligned}\right\} \tag{11}$$

for some $\epsilon > 0$.

Proof. Let us assume $u_{\alpha^*} = u_0$ for some $\alpha^* > 0$. Theorem 9 and Theorem 4 imply $u_0(x) = 0$ for $x \in \{x \in \Omega :\ |p_0(x)| = \beta\}$. Moreover, due to $p_0 = p_{\alpha^*}$ we infer $u_0 = u_{\alpha^*} = \tilde{u}_{\alpha^*}$ from Theorem 7, where $\tilde{u}_{\alpha^*}$ is defined by (10). Hence, Theorem 6 implies $\mu(\sigma\,\alpha^*/2) = 0$.

To prove the converse, let (11) be satisfied for some $\epsilon > 0$. Using (6) we obtain

$$\alpha\,\|u_0 - u_\alpha\|^2_{L^2(\Omega)} \le \alpha\,(u_0, u_0 - u_\alpha) = \alpha\,(\chi_A\,u_0, u_0 - u_\alpha) \le C\,\alpha\,|A_\alpha|,$$

where $A = \{x \in \Omega :\ |p_0(x)| > \beta\}$ and $A_\alpha = \{x \in A :\ u_0(x) \ne u_\alpha(x)\}$. Arguing similarly as in [8, Corollary 3.13], we have $|A_\alpha| = 0$, and hence $\|u_0 - u_\alpha\|_{L^2(A)} = 0$ holds for $\alpha > 0$ small enough. □

In many applications, the adjoint state p_0 belongs to $C(\Omega)$. In this case, the result of Theorem 10 shows that an exact reconstruction is only possible if $|p_0(x)| \ne \beta$ for all $x \in \Omega$. This in turn implies either $u_0 \equiv u_a$ or $u_0 \equiv 0$ or $u_0 \equiv u_b$ on every connected component of Ω.

References

1. Casas, E., Herzog, R., Wachsmuth, G.: Optimality conditions and error analysis of semilinear elliptic control problems with L^1 cost functional. SIAM J. Optim. (to appear, 2012)
2. Deckelnick, K., Hinze, M.: A note on the approximation of elliptic control problems with bang-bang controls. Comput. Optim. Appl. 51(2), 931–939 (2012)
3. Engl, H.W., Hanke, M., Neubauer, A.: Regularization of inverse problems. Mathematics and its Applications, vol. 375. Kluwer Academic Publishers Group, Dordrecht (1996)
4. Felgenhauer, U.: On stability of bang-bang type controls. SIAM J. Control Optim. 41(6), 1843–1867 (2003)
5. Neubauer, A.: Tikhonov-regularization of ill-posed linear operator equations on closed convex sets. J. Approx. Theory 53(3), 304–320 (1988)
6. Tröltzsch, F.: Optimal Control of Partial Differential Equations. Graduate Studies in Mathematics, vol. 112. American Mathematical Society, Providence (2010); Theory, methods and applications, Translated from the 2005 German original by J. Sprekels
7. Wachsmuth, D., Wachsmuth, G.: Convergence and regularization results for optimal control problems with sparsity functional. ESAIM Control Optim. Calc. Var. 17(3), 858–886 (2011)
8. Wachsmuth, D., Wachsmuth, G.: On the regularization of optimization problems with inequality constraints. Control and Cybernetics (2011) (to appear)

Robustness Analysis of Stochastic Programs with Joint Probabilistic Constraints

Jitka Dupačová

Charles University in Prague, Faculty of Mathematics and Physics
Department of Probability and Mathematical Statistics
Sokolovská 83, 18675 Prague, Czech Republic

Abstract. Due to their frequently observed lack of convexity and/or smoothness, stochastic programs with joint probabilistic constraints have been considered as a hard type of constrained optimization problems, which are rather demanding both from the computational and robustness point of view. Dependence of the set of solutions on the probability distribution rules out the straightforward construction of the convexity-based global contamination bounds for the optimal value; at least local results for probabilistic programs of a special structure will be derived. Several alternative approaches to output analysis will be mentioned.

Keywords: Joint probabilistic constraints, contamination technique, output analysis.

1 Introduction

Consider the following abstract formulation of a stochastic program

$$\min_{x \in \mathcal{X}(P)} G_0(x, P) \tag{1}$$

where P is the probability distribution of a random vector ω with range $\Omega \subset \mathbb{R}^M$ and both the criterion G_0 and the set of feasible solutions $\mathcal{X}(P) \subset \mathbb{R}^N$ may depend on P. We assume that in (1)

$$\mathcal{X}(P) := \{x \in \mathcal{X} : G_j(x, P) \leq 0, j = 1, \ldots, J\} \tag{2}$$

where $G_j(x, P) \leq 0$ are joint probabilistic constraints such as

$$P(\omega : g(x, \omega) \leq 0) \geq 1 - \varepsilon \tag{3}$$

with $g : \mathbb{R}^N \times \Omega \to \mathbb{R}^K, K > 1$; Individual probabilistic constraints correspond to $K = 1$. Probability level $\varepsilon \in (0, 1)$ in (3) is fixed, prescribed by regulations or chosen by the decision maker.

Probabilistic constraints are sufficiently flexible and model well the intuitive requirements of system reliability or hedging against risk. Depending on the problem, multiple probabilistic constraints can be used. However, the set $\mathcal{X}(P)$

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 155–164, 2013.

is typically nonconvex, sometimes even disconnected, and functions $G_j(\bullet, P)$ need not be smooth. This is the reason why probabilistic programs have been recognized as hard optimization problems that are very demanding from the computational point of view. For a given probability distribution P, (1) with joint probabilistic constraints (3) is a nonlinear program and in principle, known algorithms can be adapted provided that checking feasibility is manageable and the set of feasible solutions is convex. In so doing, one has to cope with the fact that derivatives are expressed as surface or volume integrals, cf. Chapter 5 of [31] for an introductory survey and references.

The seminal results on convexity of problems with joint probabilistic constraints were proved by Prékopa, cf. [20], under assumptions concerning both the function g and the probability distribution P.

Definition 1 (α-concave functions). *A nonnegative function $f(x)$ defined on a convex set $\mathcal{C} \subset \mathbb{R}^N$ is α-concave with $\alpha \in [-\infty, \infty]$ if for all $x, y \in \mathcal{C}$ and $\lambda \in [0,1]$ the inequality*

$$f(\lambda x + (1-\lambda)y) \geq m_\alpha(f(x), f(y), \lambda)$$

holds true. The function $m_\alpha : \mathbb{R}_+ \times \mathbb{R}_+ \times [0,1] \to \mathbb{R}$ is defined as follows:

$$m_\alpha(a, b, \lambda) = 0 \text{ if } ab = 0$$

and for $a > 0$, $b > 0$, $0 \leq \lambda \leq 1$

$$m_\alpha(a,b,\lambda) = \begin{cases} a^\lambda b^{1-\lambda} & \text{if } \alpha = 0, \quad \textit{i.e. } f \textit{ log-concave} \\ \max[a,b] & \text{if } \alpha = \infty, \quad \textit{i.e. } f \textit{ quasi-convex} \\ \min[a,b] & \text{if } \alpha = -\infty, \ \textit{i.e. } f \textit{ quasi-concave} \\ (\lambda a^\alpha + (1-\lambda)b^\alpha)^{1/\alpha} & \textit{otherwise.} \end{cases}$$

If $f(x)$ is an α-concave function then it is locally Lipschitz continuous, directionally differentiable and Clarke regular, i.e. directional derivatives $f'(x,d)$ exist and

$$f'(x,d) = \lim_{y \to x, t \to 0} \frac{f(y+td) - f(y)}{t} \, \forall d \in \mathbb{R}.$$

One of the most general results about convexity of $\mathcal{X}(P)$ is the following extension of Prékopa's original theorem.

Theorem 1 (Theorem 4.39 in [31]). *Let the functions $g_k : \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R} \, \forall k$ be quasi-convex. Let $\omega \in \mathbb{R}^M$ be a random vector that has an α-concave probability distribution, then the function $P(\omega : g_k(x,\omega) \leq 0 \, \forall k)$ is α-concave on the set*

$$\mathcal{D} := \{x \in \mathbb{R}^N : \exists y \in \mathbb{R}^M \ \textit{s.t. } g_k(x,y) \leq 0 \, \forall k\}.$$

The required *joint* quasi-convexity of $g_k(x,\omega)$ is the main limitation for exploitation of this result. Theorem 1 is applicable e.g. for $g_k(x,\omega) = -g_k(x) + \omega_k \, \forall k$, i.e. for *separable* joint probabilistic constraints. We refer to [21], [22] and to Chapter 5 of [31] for details. Another favorable class are linear probabilistic constraints with Gaussian coefficients, see e.g. [22], [32].

To solve complex probabilistic programs one tries to simplify or reformulate the model, to approximate the probability distribution, etc. These approximations and simplifications ask for development of suitable validation techniques and for stability and robustness tests. See e.g. [15] for qualitative stability results under perturbations of all input data, including the probability distribution P, the set $\mathcal{X}$ and the probability level α.

Moreover, the probability distribution P itself need not be known completely. Nevertheless, the wish is to find a solution of (1) which is efficient and reliable enough to support sensible decisions. This gives a motivation for stability or robustness analysis of (1) with respect to perturbations of P. Dependence of the set of feasible solutions on P complicates the stability considerations substantially. We denote $\mathcal{X}^*(P)$ the set of optimal solutions, $\varphi(P)$ the optimal value of the objective function in (1) and we shall *assume that* $\varphi(P)$ *is finite.*

General stability results for (2) were proved by Römisch without any convexity assumptions; cf. Theorems 5 and 9 in [25]. Then the main stumbling block for their application is the requirement of the *metric regularity property* which is related with continuity of the set $\mathcal{X}(P)$ when some perturbations of P are considered; see e.g. [1] for the general theory and [16] for specific results for probabilistic constraints. When, in addition, the set of optimal solutions is nonempty and bounded, the perturbed probability distribution, say Q, is close to the true one and the objective function is locally Lipschitz continuous one gets a local Lipschitz property of the optimal value

$$|\varphi(P) - \varphi(Q)| \leq Ld(P, Q)$$

and upper semicontinuity of the set of optimal solutions. A proper selection of the probability distance d is crucial. These results were detailed mainly for separable linear probabilistic programs and α-concave probability distributions, see e.g. [16], [25], [26].

Similarly as in [11] we shall focus on quantitative stability properties of the optimal value with respect to perturbations of P. In Section 2 we shall apply relatively simple ideas of output analysis based on the contamination technique initiated in [4], [29] whose applications for stochastic programs with a fixed set of feasible decisions were elaborated e.g. in [8], [12]. The considered special type of perturbations reduces the stability analysis of (2) to that for parametric programs with one-dimensional real parameter. At the same time, it gets on with needs for what-if-analysis or stress testing.

For stochastic programs whose set of feasible decisions does not depend on P and the objective function $G_0(x, P)$ is linear or concave in P one obtains then global bounds for the optimal value function. Local contamination bounds for the optimal value function in (2) were derived in [11] under convexity of the set $\mathcal{X}$ and of functions $G_j(\bullet, P)\, \forall j$. We shall discuss possible extensions of these results to problems with probabilistic constraints for which one cannot rely on convexity properties. In Section 3 some alternative recent approaches will be indicated.

2 Robustness Analysis via Contamination

Contamination means to model the perturbations of P by its contamination by another *fixed* probability distribution Q, i.e. to use $P_t := (1-t)P + tQ$, $t \in [0,1]$ in stochastic program (1) – (2) at the place of P. Then the set of feasible solutions of (2) for the contaminated probability distribution P_t equals

$$\mathcal{X}(P_t) = \mathcal{X} \cap \{x : G_j(x, P_t) \leq 0, j = 1, \ldots, J\}. \tag{4}$$

For probabilistic programs $G_j(x,P) = 1 - \varepsilon - H_j(x,P)$ with $H_j(x,P) = P\{\omega : \omega \in \mathcal{H}_j(x)\}$ where $\mathcal{H}_j(x) = \{y \in \mathbb{R}^s : g_k(x,y) \leq 0 \text{ for } k \in K_j\}$ describes the j-th group of constraints depending on ω and on the decision vector x. Evidently, $G_j(x,P_t) = (1-t)G_j(x,P) + tG_j(x,Q) := G_j(x,t)\ \forall j$ are linear in t. We assume that the perturbed objective function $G_0(x,t)$ is also linear in t. The perturbed problem (2) is then the *linearly perturbed* parametric program

$$\min_{x \in \mathcal{X}} (1-t)G_0(x,0) + tG_0(x,1) \tag{5}$$

subject to

$$(1-t)G_j(x,0) + tG_j(x,1) \leq 0, j = 1, \ldots, J. \tag{6}$$

We denote $\mathcal{X}(t)$, $\varphi(t)$, $\mathcal{X}^*(t)$ the set of feasible solutions, the optimal value and the set of optimal solutions of (5)–(6). For $t = 0$, $\mathcal{X}(0)$, $\varphi(0)$, $\mathcal{X}^*(0)$ denote the set of feasible solutions, the optimal value and the set of optimal solutions of the initial unperturbed problem (1) with probabilistic constraints. We shall assume that $\mathcal{X}^*(0) \neq \emptyset$, i.e., that $\varphi(0)$ is finite.

Contamination technique was developed and applied for $\mathcal{X}(P)$ independent of P and for expectation type objective $G_0(x,P)$, cf. [8], [12]. Assume that such stochastic program

$$\min_{x \in \mathcal{X}} G_0(x,P) \tag{7}$$

was solved for P and that its optimal value $\varphi(P)$ is finite. Consider a contaminated distribution

$$P_t := (1-t)P + tQ,\ t \in [0,1]$$

with Q another *fixed* probability distribution such that $\varphi(Q)$ is finite. Via contamination, robustness analysis with respect to changes in P gets reduced to much simpler analysis of parametric program with scalar parameter t.
The objective function in (7) is linear in P so that the perturbed objective $G_0(x,t) := G_0(x,P_t) = (1-t)G_0(x,P) + tG_0(x,Q)$ is linear in t. For a fixed set of feasible solutions $\mathcal{X}(t) \equiv \mathcal{X}$ we get easily (see e.g. Theorem 4.16 of [1])

Theorem 2. *Assume that $\mathcal{X} \neq \emptyset$ and $\varphi(t)$ is finite for all $t \in [0,1]$. Then $\varphi(t)$ is a lower semicontinuous concave function on $[0,1]$.*

This result allows us to construct bounds for $\varphi(t)$

$$(1-t)\varphi(0) + t\varphi(1) \leq \varphi(t) \leq \varphi(0) + t\varphi'(0^+)\ \ \forall t \in [0,1], \tag{8}$$

i.e. the sought global contamination bounds for the perturbed optimal value $\varphi(P_t)$. They quantify change in optimal value due to considered perturbations of (7).

For parameter dependent sets of feasible solutions the optimal value function $\varphi(t)$ is concave only under rather strict assumptions such as $G_j(x,t)$, $j=1,\ldots,J$ jointly concave on $\mathcal{X}\times[0,1]$ (cf. Corollary 3.2 of [17].) We shall examine how to construct computable local upper and lower contamination bounds (8) for the perturbed optimal value $\varphi(t)$ for stochastic programs (1) with probabilistic constraints (3). These local bounds can be then exploited in robustness analysis of probabilistic programs with respect to small contamination of data, inclusion of additional scenarios, etc. The form of (8) suggests that we should concentrate on the existence and form of the directional derivatives and on assumptions under which for small t, the sets of feasible solutions $\mathcal{X}(t)$ remain fixed or the optimal value function $\varphi(t)$ is concave.

There exist formulas for directional derivative $\varphi(0^+)$ based on the Lagrange function $L(x,u,t)=G_0(x,P_t)+\sum_j u_j G_j(x,P_t)$ for the contaminated problem. The generic formula

$$\varphi'(0^+)=\min_{x\in\mathcal{X}^*(0)}\max_{u\in\mathcal{U}^*(x,0)}\frac{\partial}{\partial t}L(x,u,0)$$

simplifies thanks to linearity of the Lagrange function with respect to the parameter t. The derivations proceed in accordance with the assumed properties of problem (5)–(6); consult section 4.3.2 of [1]. The directional derivative $\varphi'(0^+)$ provides information about the influence of contamination on the optimal value $\varphi(t)$ for small t. It can be obtained without the second order sufficient condition, e.g. [14], [28], under assumptions which guarantee existence of a continuous trajectory $x^*(t)$ for a small contamination t. Besides of uniform compactness of $\mathcal{X}(t)$ for $t>0$ and small enough, the approach assumes that the unperturbed problem has unique optimal solution $x^*(0)$ for which the Mangasarian-Fromowitz constraint qualification holds. Multiple Lagrange multipliers, whose sets are bounded convex polyhedra, are not excluded and multiple optimal solutions may occur for $t>0$.

Classical stability results for *nonlinear parametric programs* with a *parameter dependent set of feasible solutions* such as (6), including directional differentiability of the optimal value function, were first obtained by applying the Implicit Function Theorem to the first-order optimality conditions under assumptions that imply existence and uniqueness of the optimal solution and of the corresponding Lagrange multipliers for the unperturbed problem, see e.g. [13]. For the Lagrange function

$$L(x,u,t)=G_0(x,t)+\sum\nolimits_j u_j G_j(x,t),$$

with differentiable functions $G_j(\bullet,t)$ and for $\mathcal{X}=\mathbb{R}^N$ the optimal solution and the vector of the corresponding Lagrange multipliers for (6) have to satisfy the first-order optimality condition

$$\nabla_x L(x,u,t)=\nabla_x G_0(x,t)+\sum\nolimits_j u_j\nabla_x G_j(x,t)=0.$$

Besides of the linear independence and the strict complementarity conditions valid at the optimal solution $x^*(0)$ of the unperturbed problem and at the corresponding vector of Lagrange multipliers $u^*(0)$, the derivation exploits also existence and nonsingularity of the Hessian matrix of the Lagrange function on the tangent space to the active constraints at $x^*(0), u^*(0)$; see e.g. [1], [13]. Then there exists $t_0 > 0$ and a smooth trajectory $[x^*(t), u^*(t)]$ emanating from $[x^*(0), u^*(0)]$ which satisfies the first-order optimality conditions for $0 \leq t \leq t_0$:

$$G_j(x^*(t), t) \leq 0,\ u_j^*(t) \geq 0,\ G_j(x^*(t), t)u_j^*(t) = 0, \quad j = 1, 2, \ldots, J,$$

$$\nabla_x G_0(x^*(t), t) + \sum\nolimits_j u_j^*(t) \nabla_x G_j(x^*(t), P) = 0$$

and the directional derivative

$$\varphi'(0^+) = L(x^*(0), u^*(0), 1) - L(x^*(0), u^*(0), 0).$$

This approach was applied in [5] for probabilistic programs under the second order sufficient condition. Having in mind the nonsmooth character of probabilistic constraints we wish to get bounds for the optimal value function $\varphi(t)$ under relaxed differentiability requirements. We shall see that thanks to the assumed structure of perturbations

- lower bound for $\varphi(t)$ can be derived for $G(x, P)$ linear (or concave) with respect to P without any smoothness or convexity assumptions with respect to x,
- further assumptions are needed for derivation of an upper bound.

The *lower bound* for the optimal value function was derived in [11] for the assumed structure of perturbations without any smoothness or convexity assumptions with respect to x. Let us consider first only one probability constraint and an objective G_0 *independent* of P, i.e. the unperturbed problem is

$$\min_{x \in \mathcal{X}} G_0(x) \text{ subject to } G(x, P) := 1 - \varepsilon - P(\omega : g(x, \omega) \leq 0) \leq 0. \tag{9}$$

Theorem 3 ([11]). *Let $\mathcal{X} \subset \mathbb{R}^N$ be a nonempty convex set, $G(x, t)$ be a linear function of $t \in [0, 1]$ and $\varphi(t)$ be finite for all $t \in [0, 1]$. Then the optimal value function*

$$\varphi(t) := \min_{x \in \mathcal{X}} G_0(x) \text{ subject to } G(x, t) \leq 0$$

is quasi-concave on $[0, 1]$ with the lower bound

$$\varphi(t) \geq \min\{\varphi(1), \varphi(0)\}. \tag{10}$$

When also the objective function *depends* on the probability distribution, i.e. on the contamination parameter t, the problem is

$$\min_{x \in \mathcal{X}} G_0(x, t) \text{ subject to } G(x, t) \leq 0. \tag{11}$$

For $G_0(x,P)$ linear in P, a lower bound can be obtained by application of the bound (10) separately to $G_0(x,P)$ and $G_0(x,Q)$:

$$\varphi(t) = \min_{x\in\mathcal{X}(t)} G_0(x,t) = \min_{x\in\mathcal{X}(t)} [(1-t)G_0(x,P) + tG_0(x,Q)] \geq$$
$$(1-t)\min\{\varphi(0), \min_{\mathcal{X}(Q)} G_0(x,P)\} + t\min\{\varphi(1), \min_{\mathcal{X}(P)} G_0(x,Q)\}. \tag{12}$$

The bound is more complicated but still computable. It requires solution of 4 problems two of which are the non-contaminated programs for probability distributions P, Q and the other ones use both P and Q alternating in the objective function and constraints. For *multiple constraints* and contaminated probability distributions it would be necessary to prove first the inclusion $\mathcal{X}(t) \subset \mathcal{X}(0) \cup \mathcal{X}(1)$. Then the lower bound (12) for the optimal value $\varphi(t) = \min_{x\in\mathcal{X}(t)} G_0(x,t)$ follows as in the case of one constraint.

Similarly as in [11], trivial *upper bounds* for $\varphi(t)$ can be obtained without any differentiability assumption if no constraint is active at $x^*(0)$ or if for all constraints active at $x^*(0)$, i.e. $G_j(x^*(0),0) = 0$, $j \in J_0$, inequalities $G_j(x^*(0),1) \leq 0$, $j \in J_0$ hold true. Then for t small enough, $x^*(0)$ is a feasible solution of (6), hence $G_0(x^*(0),t) \geq \varphi(t)$ for t small enough. Using linearity of G_0 with respect to t we obtain the upper bound

$$\varphi(t) \leq \varphi(0) + t(G_0(x^*(0),1) - \varphi(0));$$

compare with (8). An upper bound for $\varphi(t)$ can be also constructed whenever there is at disposal a feasible solution $\hat{x} \in \mathcal{X}(P_t)$ which may occur due to the structure of the solved problem. A direct search for $\hat{x} \in \mathcal{X}$ which satisfies constraints

$$G_j(x,0) \leq 0\,\forall j \text{ and } G_j(x,1) \leq 0\,\forall j$$

may be manageable, namely, when $Q = \delta_{\omega^*}$ is a degenerated probability distribution. Using it means to augment $\mathcal{X}$ by deterministic constraints $g_k(x,\omega^*) \leq 0$, $k \in K_j$, $j = 1,\ldots,J$. For problems with one joint probability constraint one may solve

$$\min_{x\in\mathcal{X}} G(x,1) \text{ subject to } G(x,0) \leq 0.$$

These ideas, however, do not exploit the parametric form of constraints in the definition of $\mathcal{X}(P_t)$. For problems with one joint probabilistic constraint solution of parametric program

$$\min_{x\in\mathcal{X}}[(1-t)G(x,0) + tG(x,1)] \tag{13}$$

for increasing values of t may lead to the sought solution $\hat{x} \in \mathcal{X}(P_t)$ and to the upper bound $\varphi(t) \leq G_0(\hat{x},t)$.

ILLUSTRATIVE EXAMPLE. In the jointly constrained probabilistic program

$$\begin{aligned} &\min x_1 + x_2 \\ &\text{subject to} \\ &P(\omega_1 x_1 + x_2 \geq 7, \omega_2 x_1 + x_2 \geq 4) \geq 1-\varepsilon, \\ &x_1 \geq 0,\ x_2 \geq 0 \end{aligned} \tag{14}$$

the random components (ω_1, ω_2) are independent and have uniform distributions on the intervals $[1, 4]$ and $[1/3, 1]$. It is a convex program and, thanks to the independence assumption, the explicit form of the optimal solution can be obtained directly: $x_1^*(P) \doteq 3.6735$, $x_2^*(P) \doteq 2.7755$ and $\varphi(P) \doteq 6.4480$ for $\varepsilon = .05$; cf. [18].

To stress the sample distribution we choose the extremal scenario $(\omega_1^*, \omega_2^*) = (1.02, 0.34)$. The optimal solution $x_1^*(P)$, $x_2^*(P)$ is infeasible for $t = 1$, $x_1^*(Q) \doteq 4.4118$, $x_2^*(Q) \doteq 2.5000$ and $\varphi(Q) \doteq 6.9118$. Hence, for all $0 \leq t \leq 1$ the lower bound (10) for $\varphi(t)$ is $\varphi(P) \doteq 6.4480$.

Solution $\hat{x}_1 = 4.4725$, $\hat{x}_2 = 2.4994$ of the "upper bound problem" (13) obtained for $t = 0$ is feasible for all contaminated problems $(7.0614 \geq 7, 4.02 > 4)$. Then, the value $6.9719 = \hat{x}_1 + \hat{x}_2$ is upper bound for $\varphi(t)\, \forall t$.

For *differentiable* functions G_j properties of the set $\mathcal{X}(t) = \mathcal{X}(P_t)$ for small t follow from results of [2], [23], [24]. Linear independence condition at $x^*(0)$ implies that $x^*(0)$ is a nondegenerate point, the vector $u^*(0)$ of Lagrange multipliers is unique and the problem (5)–(6) can be locally reduced to one with a *fixed set of feasible solutions:*

$$\min_z G_0(T(z,t),t) \text{ on a set } \mathcal{C} \tag{15}$$

where $T(z,t)$ is continuously differentiable and $T(0,0) = x^*(0)$. However, the cost for obtaining a fixed set of feasible solutions is that linearity of the objective function with respect to t gets lost. This can be compared to the situation described in detail in Example 1 of [3] for stochastic linear program with individual probabilistic constraints and random right-hand sides ω_k. Using quantiles of marginal probability distributions, the problem can be cast into the form of a linear program for which the dual feasible set is fixed, independent of P. However, the quantiles of the contaminated marginal probability distributions that appear as parameter dependent coefficients in the dual objective function are not linear in t.

3 Conclusions and Alternative Approaches

Whereas there exists a general lower bound, our discussion indicates that there are limited possibilities to construct local upper contamination bounds for nonconvex probabilistic programs when differentiability cannot be guaranteed.

In paper [3], an indirect approach was suggested: To apply contamination technique to a penalty reformulation of the probabilistic program. Then the set of feasible solutions does not depend on P and for the approximate problem, global bounds (8) follow. We refer to Example 4 of [3] for numerical results related with the illustrative example (14).

Another way how to get an upper bound for the optimal value of the probabilistic program is to apply the worst-case analysis with respect to a whole set $\mathcal{P}$ of considered probability distributions, cf. [19], [33]. This means to hedge against

all probability distributions belonging to the chosen ambiguity set and to solve the following problem:

$$\min_{x\in\mathcal{X}} \max_{P\in\mathcal{P}} G_0(x,P) \tag{16}$$

subject to

$$P(\omega : g(x,\omega) \le 0) \ge 1-\varepsilon \quad \forall P \in \mathcal{P}. \tag{17}$$

The problem (16)–(17) need not be more complicated than the underlying probabilistic program. Its tractability depends on function $g(x,\omega)$ and on the choice of the ambiguity set $\mathcal{P}$. In [19], $\mathcal{P}$ is the Prokhorov neighborhood of the true probability distribution P, whereas in [33], $\mathcal{P}$ contains probability distributions with a given mean, covariance matrix and support. In the last case, (16)–(17) can be solved via semidefinite optimization techniques. The results depend on the input information and similarly as in [10], their stability should be studied.

Acknowledgments. This research is supported by the grant 402/11/0150 of the Czech Science Foundation.

References

1. Bonnans, J.F., Shapiro, A.: Perturbation Analysis of Optimization Problems. Springer, New York (2000)
2. Bonnans, J.F., Shapiro, A.: Nondegeneracy and quantitative stability of parametrized optimization problems with multiple solutions. SIAM J. Optim. 8, 940–946 (1998)
3. Branda, M., Dupačová, J.: Approximation and contamination bounds for probabilistic programs. Ann. Oper. Res. 193, 3–19 (2012)
4. Dupačová, J.: Stability in stochastic programming with recourse – contaminated distributions. Math. Program. Study 27, 133–144 (1986)
5. Dupačová, J.: Stability in stochastic programming – probabilistic constraints. In: Arkin, V.I., Shiraev, A., Wets, R. (eds.) Stochastic Optimization. LNCIS, vol. 81, pp. 314–325. Springer, Berlin (1986)
6. Dupačová, J.: Stochastic programming with incomplete information: A survey of results on postoptimization and sensitivity analysis. Optimization 18, 507–532 (1987)
7. Dupačová, J.: Stability and sensitivity analysis in stochastic programming. Ann. Oper. Res. 27, 115–142 (1990)
8. Dupačová, J.: Scenario based stochastic programs: Resistance with respect to sample. Ann. Oper. Res. 64, 21–38 (1996)
9. Dupačová, J.: Reflections on robust optimization. In: Marti, K., Kall, P. (eds.) Stochastic Programming Methods and Technical Applications. LNEMS, vol. 437, pp. 111–127. Springer, Berlin (1998)
10. Dupačová, J.: Uncertainties in minimax stochastic programs. Optimization 60, 1235–1250 (2011)
11. Dupačová, J., Kopa, M.: Robustness in stochastic programs with risk constraints. Ann. Oper. Res. 200, 55–77 (2012), doi:10.1007/s10479-010-0824-9
12. Dupačová, J., Polívka, J.: Stress testing for VaR and CVaR. Quantitiative Finance 7, 411–421 (2007)

13. Fiacco, A.V.: Introduction to Sensitivity and Stability Analysis in Nonlinear Programming. Academic Press, New York (1983)
14. Gauvin, J., Dubeau, F.: Differential properties of the marginal function in mathematical programming. Math. Program. Study 19, 101–119 (1982)
15. Henrion, R.: Perturbation analysis of chance-constrained programs under variation of all constraint data. In: Marti, K., et al. (eds.) Dynamic Stochastic Optimization. LNEMS, vol. 532, pp. 257–274. Springer, Berlin (2004)
16. Henrion, R., Römisch, W.: Hölder and Lipschitz stability of solution sets in programs with probabilistic constraints. Math. Program. 100, 589–611 (2004)
17. Kyparisis, J., Fiacco, A.: Generalized convexity and concavity of the optimal value function in nonlinear programming. Math. Program. 39, 285–304 (1987)
18. Pagoncelli, B.K., Ahmed, S., Shapiro, A.: Sample average approximation method for chance constrained programming: Theory and applications. J. Optim. Theory Appl. 142, 399–416 (2009)
19. Pflug, G., Wozabal, D.: Ambiguity in portfolio selection. Quant. Fin. 7, 435–442 (2007)
20. Prékopa, A.: Logarithmic concave measures with application to stochastic programming. Acta Sci. Math. (Szeged) 32, 301–316 (1971)
21. Prékopa, A.: Stochastic Programming. Kluwer Acad. Publ., Dordrecht (1995)
22. Prékopa, A.: Probabilistic Programming. In: [27], ch. 5, pp. 267–351
23. Robinson, S.M.: Local structure of feasible sets in nonlinear programming, part II: Nondegeneracy. Math. Program. Study 22, 217–230 (1984)
24. Robinson, S.M.: Local structure of feasible sets in nonlinear programming, Part III: Stability and sensitivity. Math. Program. Study 30, 45–66 (1987)
25. Römisch, W.: Stability of stochastic programming problems. In: [27], ch. 8, pp. 483–554
26. Römisch, W., Schultz, R.: Stability analysis for stochastic programs. Ann. Oper. Res. 30, 241–266 (1991)
27. Ruszczyński, A., Shapiro, A. (eds.): Stochastic Programming. Handbooks in OR & MS, vol. 10. Elsevier, Amsterdam (2003)
28. Shapiro, A.: Sensitivity analysis of nonlinear programs and differentiability properties of metric projections. SIAM J. Control and Optimization 26, 628–645 (1988)
29. Shapiro, A.: On differential stability in stochastic programming. Math. Program. 47, 107–116 (1990)
30. Shapiro, A.: Monte Carlo sampling methods. In: [27], ch. 6, pp. 353–425
31. Shapiro, A., Dentcheva, D., Ruszczyński, A.: Lectures on Stochastic Programming. SIAM and MPS, Philadelphia (2009)
32. van Ackooij, W., Henrion, R., Möller, A., Zorgati, R.: On joint probabilistic constraints with Gaussian coefficient matrix. Operations Research Letters 39, 99–102 (2011)
33. Zymler, S., Kuhn, D., Rustem, B.: Distributionally robust joint chance constraints with second-order moment information. Math. Program., Ser. A (published online November 10, 2011)

State Estimation for Control Systems with a Multiplicative Uncertainty through Polyhedral Techniques

Elena K. Kostousova

Institute of Mathematics and Mechanics,
Ural Branch of the Russian Academy of Sciences,
16, S.Kovalevskaja street, Ekaterinburg, 620990, Russia
kek@imm.uran.ru

Abstract. The paper deals with polyhedral estimates for reachable tubes of differential systems with a multiplicative uncertainty, namely linear systems with set-valued uncertainties in initial states, additive inputs and coefficients of the system. We present nonlinear parametrized systems of ordinary differential equations (ODE) which describe the evolution of the parallelotope-valued estimates for reachable sets (time cross-sections of the reachable tubes). The main results are obtained for internal estimates. In fact, a whole family of the internal estimates is introduced. The properties of the obtained ODE systems (such as existence and uniqueness of solutions, nondegeneracy of estimates) are investigated. Using some optimization procedure we also obtain a differential inclusion which provides nondegenerate internal estimates. Examples of numerically constructed external and internal estimates are presented.

Keywords: Differential systems, reachable sets, set-valued state estimation, multiplicative uncertainty, polyhedral estimates, parallelepipeds, parallelotopes, interval analysis.

1 Introduction

The problem of constructing trajectory tubes (in particular, reachable tubes) is an essential theme in control theory [25]. Since practical construction of such tubes may be cumbersome, different numerical methods are devised for this cause, in particular, methods based on approximations of sets either by arbitrary polytopes with a large number of vertices or by unions of points [6], [3], [1] (here and below, we mention, as examples, only some references from numerous publications; see also references therein). Such methods, as well as the methods based on different schemes of discrete approximations of initial set-valued problems [2], [31] and numerical methods of solving the Hamilton-Jacobi-Bellman equation [30], are devised to obtain approximations as accurate as possible. But they may require much calculations, especially for large dimensional systems; also smaller step-sizes create a heavy computational load. It is appropriate to

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 165–176, 2013.

mention approximations by polytopes based on support functions or supporting points [4], [23].

Other techniques are based on estimates of sets by domains of some fixed shape such as ellipsoids, parallelepipeds, zonotopes [7], [11], [13], [15]-[17], [22], [24]-[29]. Fair results in this area were obtained for linear systems with set-valued initial states and set-valued additive uncertain inputs. The main advantage of the mentioned techniques is that they enable to obtain approximate solutions using relatively simple tools (up to explicit formulas). Note that more accurate approximations and even exact representations of the solutions may be obtained by using the whole families of such simple estimates (as it was proposed by A.B. Kurzhanski) [25]-[27], [22], [16]. The methods of interval analysis which use subpavings of interval vectors [14] serve the same purpose, but such methods may require much computations and memory for large dimensional systems.

It is also important to study linear systems when system matrices are uncertain too. This leads to the multiplicative uncertainty and additional difficulties due to nonlinearity of the problem (in particular, reachable sets — cross-sections of reachable tubes — can be non-convex). There are some results for such systems with different types of bounds on uncertainties [5], [8], [12], including constructing external ellipsoidal estimates [7], [29] and external interval (in other terms, coordinate-wise or box-valued) estimates [15], [24], [28].

We construct polyhedral (parallelepiped-valued and parallelotope-valued) estimates for reachable sets and reachable tubes of differential systems with parallelepiped-valued uncertainties in initial states and in additive uncertain inputs and with interval uncertainties in coefficients of the system. In contrast to interval analysis, faces of our estimates may be not parallel to coordinate planes. The main results are obtained for the internal estimates. Using constructions from [19], [20], we obtain nonlinear parametrized systems of ordinary differential equations (ODE) which describe the evolution of centers and matrices of the parallelotope-valued internal estimates for the reachable sets. So, in fact, the whole family of internal estimates is introduced (but, unfortunately, unlike the case of linear systems [25]-[27], [22], [16], this family does not ensure exact representations of the reachable sets in general). The properties of the obtained ODE systems (such as existence and uniqueness of solutions for fixed values of parameters, nondegeneracy of estimates) are investigated. Using some optimization procedure we also obtain a differential inclusion which provides nondegenerate internal estimates. ODE for external estimates were obtained earlier [18]. Here we remind these results for completeness of the exposition. Results of numerical simulations are presented.

The following notation is used below: $\mathbb{R}^n$ — the n-dimensional vector space; $\top$ — the transposition symbol; $\|x\|_1 = \sum_{i=1}^n |x_i|$, $\|x\|_2 = (x^\top x)^{1/2}$, $\|x\|_\infty = \max_{1\le i\le n} |x_i|$ — vector norms for $x = (x_1, x_2, \ldots, x_n)^\top \in \mathbb{R}^n$; e^i=$(0, \ldots, 0, 1, 0, \ldots, 0)^\top$ — the unit vector oriented along the axis $0x_i$ (the unit stands at i-position); $\mathrm{e} = (1, 1, \ldots, 1)^\top$; $\mathbb{R}^{n\times m}$ — the space of real $n\times m$-matrices $A = \{a_i^j\} = \{a^j\}$ (with columns a^j); I — the unit matrix; 0 — the zero matrix (vector); $\mathrm{Abs}\, A = \{|a_i^j|\}$ for $A = \{a_i^j\}$; $\mathrm{diag}\, \pi$, $\mathrm{diag}\, \{\pi_i\}$ — the diagonal matrix

A with $a_i^i = \pi_i$ (π_i — the components of the vector π); $\det A$ — the determinant of $A \in \mathbb{R}^{n\times n}$; $\operatorname{tr} A = \sum_{i=1}^n a_i^i$ — the trace of A; $\|A\| = \max_{1\le i\le n} \sum_{j=1}^m |a_i^j|$ — the matrix norm for $A\in\mathbb{R}^{n\times m}$ induced by the vector norm $\|x\|_\infty$; the notation of the type $k = 1,\dots,N$ is used instead of $k = 1,2,\dots,N$ for shortening.

2 Problem Formulation

Consider the following system ($x \in \mathbb{R}^n$ is the state):

$$\dot{x} = A(t)\,x + w(t), \quad t \in T = [0,\theta]. \tag{1}$$

Here the initial state $x(0) = x_0 \in \mathbb{R}^n$, the input (control/disturbance) $w(t) \in \mathbb{R}^n$ (which is assumed to be a Lebesgue measurable function) and the measurable matrix function $A(t) \in \mathbb{R}^{n\times n}$ are unknown but subjected to given set-valued constraints

$$x_0 \in \mathcal{X}_0, \quad w(t) \in \mathcal{R}(t), \quad \text{a.e. } t \in T, \tag{2}$$

$$A(t) \in \mathcal{A}(t) = \{A \in \mathbb{R}^{n\times n} |\, \underline{A}(t) \le A \le \overline{A}(t)\}, \quad \text{a.e. } t \in T, \tag{3}$$

where $\mathcal{X}_0$, $\mathcal{R}(t)$ are given convex compact sets in $\mathbb{R}^n$, the set-valued map $\mathcal{R}(t)$ is continuous, the matrix functions $\underline{A}(t)$, $\overline{A}(t)$ are continuous. Matrix and vector inequalities ($\le, <, \ge, >$) here and below are understood componentwise. The interval constraints (3) can be rewritten in the form

$$A(t) \in \mathcal{A}(t) = \{A |\, \operatorname{Abs}(A - \tilde{A}(t)) \le \hat{A}(t)\}, \;\; \tilde{A} = (\underline{A}+\overline{A})/2, \;\; \hat{A} = (\overline{A}-\underline{A})/2. \tag{4}$$

Let $\mathcal{X}(t)=\mathcal{X}(t,0,\mathcal{X}_0)$ be a *reachable set* of system (1)–(3) at time $t>0$ that is the set of all points $x\in\mathbb{R}^n$, for each of which there exist x_0, $w(\cdot)$, $A(\cdot)$ that satisfy (2)–(3) and generate a solution $x(\cdot)$ of (1) that satisfies $x(t)=x$. The multivalued function $\mathcal{X}(t)$, $t\in T$, is known as a *trajectory* (or *reachable*) *tube* $\mathcal{X}(\cdot)$.

We presume the sets $\mathcal{X}_0$, $\mathcal{R}(t)$ to be parallelotopes (then the sets $\mathcal{X}(t)$ are not obliged to be parallelotopes) and look for external and internal parallelepiped-valued or parallelotope-valued (shorter, *polyhedral*) estimates $\mathcal{P}^{\pm}(t)$ for $\mathcal{X}(t)$.

By a *parallelepiped* $\mathcal{P}(p,P,\pi)\subset\mathbb{R}^n$ we mean a set such that $\mathcal{P}=\mathcal{P}(p,P,\pi) = \{x \in \mathbb{R}^n |\, x=p+\sum_{i=1}^n p^i\pi_i\xi_i,\ \|\xi\|_\infty\le 1\}$, where $p\in\mathbb{R}^n$; $P=\{p^i\}\in\mathbb{R}^{n\times n}$ is such that $\det P\ne 0$, $\|p^i\|_2=1$[1]; $\pi \in \mathbb{R}^n$, $\pi \ge 0$. It may be said that p determines the center of the parallelepiped, P — the orientation matrix, p^i — the "directions" and π_i — the values of its "semi-axes". We call a parallelepiped *nondegenerate* if $\pi>0$.

By a *parallelotope* $\mathcal{P}[p,\bar{P}] \subset \mathbb{R}^n$ we mean a set $\mathcal{P} = \mathcal{P}[p,\bar{P}] = \{x \in \mathbb{R}^n |\ x = p+\bar{P}\zeta,\ \|\zeta\|_\infty \le 1\}$, where $p \in \mathbb{R}^n$ and the matrix $\bar{P} = \{\bar{p}^i\} \in \mathbb{R}^{n\times m}$, $m \le n$, may be singular. We call a parallelotope $\mathcal{P}$ *nondegenerate*, if $m = n$ and $\det \bar{P} \ne 0$.

Each parallelepiped $\mathcal{P}(p,P,\pi)$ is a parallelotope $\mathcal{P}[p,\bar{P}]$ with $\bar{P} = P\operatorname{diag}\pi$; each nondegenerate parallelotope is a parallelepiped with $P=\bar{P}\operatorname{diag}\{\|\bar{p}^i\|_2^{-1}\}$, $\pi_i = \|\bar{p}^i\|_2$ or, in a different way, with $P = \bar{P}$, $\pi = \mathrm{e}$, where $\mathrm{e} = (1,1,\dots,1)^\top$.

We call $\mathcal{P}$ an *external (internal) estimate for* $\mathcal{X} \subset \mathbb{R}^n$ if $\mathcal{P} \supseteq \mathcal{X}$ ($\mathcal{P} \subseteq \mathcal{X}$).

[1] The normality condition $\|p^i\|_2=1$ may be omitted to simplify formulas (it ensures the uniqueness of the representation of a nondegenerate parallelepiped).

Assumption 1. The set $\mathcal{X}_0 = \mathcal{P}_0 = \mathcal{P}[p_0, \bar{P}_0] = \mathcal{P}(p_0, P_0, \pi_0)$ is a parallelepiped, the sets $\mathcal{R}(t) = \mathcal{P}[r(t), \bar{R}(t)]$ are parallelotopes where $\bar{R}(t) \in \mathbb{R}^{n\times m}$, $m \le n$; $r(\cdot)$, $\bar{R}(\cdot)$ and $\underline{A}(\cdot)$, $\overline{A}(\cdot)$ are continuous vector and matrix functions.

Problem 1. Find some external $\mathcal{P}^+(t)$ and internal $\mathcal{P}^-(t)$ polyhedral estimates[2] for reachable sets $\mathcal{X}(t)$: $\mathcal{P}^-(t) \subseteq \mathcal{X}(t) \subseteq \mathcal{P}^+(t)$, $t \in T$.

3 Auxiliary Discrete Time Systems. Primary Estimates

Following arguments similar to [7], [25, Sec. 3.2] we obtain ODE for the estimates. The first step in this way is to construct estimates for reachable sets $\mathcal{X}[k]$ of auxiliary discrete time systems – the Euler approximations[3] of the initial system:

$$\begin{aligned} &x[k] = A[k-1]x[k-1] + w[k-1], \quad k=1,\dots,N; \quad x[0] \in \mathcal{P}_0; \\ &w[k] \in \mathcal{R}[k] = h_N \mathcal{R}(t_k); \quad A[k] \in \mathcal{A}[k] = \{I + h_N A \,|\, A \in \mathcal{A}(t_k)\}, \end{aligned} \tag{5}$$

$t_k{=}kh_N$, $h_N{=}\theta N^{-1}$. It is known that $\mathcal{X}[k]$ satisfy the relations $\mathcal{X}[k] = \mathcal{A}[k-1]\circ \mathcal{X}[k-1]+\mathcal{R}[k-1]$, $k=1,\dots,N$, $\mathcal{X}[0] = \mathcal{P}_0$, which involve two operations with sets — *multiplying an interval matrix* $\mathcal{A} = \{A \in \mathbb{R}^{n\times n} | \underline{A} \le A \le \overline{A}\}$ *on a set* $\mathcal{X} \subset \mathbb{R}^n$: $\mathcal{A}\circ\mathcal{X} = \{y \in \mathbb{R}^n \,|\, y = Ax,\ A \in \mathcal{A}, x \in \mathcal{X}\}$ and the *Minkowski sum* [25, p.93].

In [18], [19], [20], the ways of constructing primary polyhedral estimates for $\mathcal{A}\circ\mathcal{P}$ and $\mathcal{P}^1 + \mathcal{P}^2$ (where $\mathcal{P}$, $\mathcal{P}^1$, $\mathcal{P}^2$ are parallelepipeds or parallelotopes) are described; hence we have the corresponding recurrence relations for external and internal estimates $\mathcal{P}^{\pm}[k]$ for $\mathcal{X}[k]$. Passing to the limit as $N \to \infty$, we obtain the corresponding nonlinear ODE systems for parallelotopes/parallelepipeds $\mathcal{P}^{\pm}(t)$.

4 Internal Estimates

We come to the following ODE system for parallelotopes $\mathcal{P}^-(t){=}\mathcal{P}[p^-(t), \bar{P}^-(t)]$:

$$\frac{dp^-}{dt} = \tilde{A}(t)\, p^- + r(t), \quad p^-(0) = p_0; \tag{6}$$

$$\begin{aligned} &\frac{d\bar{P}^-}{dt} = \tilde{A}(t)\,\bar{P}^- + \operatorname{diag}\nu(t, \bar{P}^-; J(t))\cdot B(\bar{P}^-) + \bar{R}(t)\,\Gamma(t), \quad \bar{P}^-(0) = \bar{P}_0, \\ &\quad \nu_i(t, \bar{P}^-; J) = \hat{a}_i^{j_i}(t)\cdot \eta_{j_i}(t, \bar{P}^-), \quad i = 1,\dots,n, \\ &\quad \eta(t, \bar{P}^-) = \max\{0, \operatorname{Abs} p^-(t) - (\operatorname{Abs}\bar{P}^-)\mathrm{e}\}, \\ &\quad B = \operatorname{diag}\beta(\bar{P}^-)\cdot \bar{P}^-, \quad \beta_i(\bar{P}^-) = 1 \,/\, (\mathrm{e}^{i\top}(\operatorname{Abs}\bar{P}^-)\,\mathrm{e}), \quad i = 1,\dots,n, \end{aligned} \tag{7}$$

[2] Our estimates will satisfy the upper and lower semigroup properties and the superreachability and subreachability properties similarly to [25, Lemmas 3.3.2, 3.3.4] and [7, Remark 8.2] respectively; these properties are analogues to the semigroup property [25, p.9] for $\mathcal{X}(t)$. Also our set-valued estimates will be continuous.

[3] In connection with the Euler approximations, the papers [9], [32] may be mentioned which analyzed the numerical error of the set-valued method.

(the operation of maximum is understood componentwise). Here $\Gamma(t) \in \mathbb{R}^{m\times n}$ is an arbitrary Lebesgue measurable matrix function satisfying $\|\Gamma(t)\| = \max_{1\le i\le m} \sum_{j=1}^{n} |\gamma_i^j| \le 1$, a.e. $t \in T$, and $J = \{j_1, \dots, j_n\}$ is an arbitrary permutation of numbers $\{1, \dots, n\}$ or even a measurable vector function $J(\cdot)$ with values $J(t)$ being arbitrary permutations of numbers $\{1, \dots, n\}$. Let $\mathbb{G}$ and $\mathbb{J}$ be the sets of all such functions $\Gamma(\cdot)$ and $J(\cdot)$ respectively.

Later on it is useful to mark out the following case.

Assumption 2. Either $\mathcal{R}(t)$ are singletons (then the function $w(\cdot) \equiv r(\cdot)$ may be assumed to be measurable) or $\Gamma(\cdot) \in \mathbb{G}$ is such that $\bar{R}(t)\Gamma(t) \equiv 0$, $t \in T$.

Theorem 1. *Let the above assumptions about the system (1), (2), (4) be satisfied, $\mathcal{P}_0$ be a nondegenerate parallelotope ($\det \bar{P}_0 \ne 0$) and $J(\cdot) \in \mathbb{J}$, $\Gamma(\cdot) \in \mathbb{G}$. Then the system (6), (7) has a unique solution $(p^-(\cdot), \bar{P}^-(\cdot))$ at least on some subinterval $T_1 = [0, \theta_1] \subseteq T$, where $0<\theta_1\le\theta$, and we have $\det \bar{P}^-(t)\ne 0$, $t \in T_1$. The corresponding nondegenerate parallelotopes $\mathcal{P}^-(t)=\mathcal{P}[p^-(t), \bar{P}^-(t)]$, $t \in T_1$, are internal estimates for the reachable sets $\mathcal{X}(t)$ of the system (1), (2), (4): $\mathcal{P}^-(t)\subseteq\mathcal{X}(t)$, $t\in T_1$. Under Assumption 2, the subinterval T_1 coincide with T.*

Proof. Here and below we give mostly only sketches; more details can be found in [21]. The existence, uniqueness and extendability of the solution are obtained using the known results [10, pp. 7,8,10]. In particular, under Assumption 2, we verify that $P^-(t)$ can not leave the domain where $\det P^-(t) \ge \delta$ for some $\delta > 0$ and then use [10, p. 10, Theorem 4].

To prove $\mathcal{P}^-(t) \subseteq \mathcal{X}(t)$ we verify the subreachability property: $\mathcal{P}^-(t) \subseteq \mathcal{X}(t, s, \mathcal{P}^-(s))$, $\forall s, t : \ 0 \le s \le t \le \theta_1$. Fix $t \in T_1$. If $x^* \in \mathcal{P}^-(t)$, then there exists ξ such that $\operatorname{Abs}\xi \le \mathrm{e}$ and $x^* = p^-(t) + \bar{P}^-(t)\,\xi$. Consider $x^* = x^*(t)$ as a function of t (when ξ is fixed). Evidently, $x^*(s) \in \mathcal{P}^-(s)$ also for arbitrary $s \le t$. It remains to check that it is possible to find functions $A(\tau) = \tilde{A}(\tau) + \Delta A(\tau) \in \mathcal{A}(\tau)$ and $w(\tau) \in \mathcal{R}(\tau)$, $\tau \in [s, t]$, such that $x^*(\tau)$ will satisfy (1) for $\tau \in [s, t]$. Differentiating $x^*(\tau)$ with the account (6), (7) we have (the argument τ is omitted for short): $\dot{x}^* = \tilde{A}x^* + w + \operatorname{diag}\nu\, B\,\xi = Ax^* + w + q$, where $w = r + \bar{R}\Gamma\xi \in \mathcal{R}$ (because $\|\Gamma\,\xi\|_\infty \le \|\Gamma\|\,\|\xi\|_\infty \le 1$), $A = \tilde{A} + \Delta A$, $q = \operatorname{diag}\nu\, B\,\xi - \Delta A\,(p^- + \bar{P}^-\xi)$. The desired equality $q = 0$ is achieved if we take ΔA in the form $\Delta A = \operatorname{diag}\alpha\, D$, where $D = \{\mathrm{e}^{j_1} \cdots \mathrm{e}^{j_n}\}^\top$ is a matrix corresponding to a permutation $J=\{j_1, \dots, j_n\}=J(\tau)$ of rows of the unit matrix, and components of the vector $\alpha = \alpha(\tau)$ are calculated by formulas $\alpha_i = 0$, if $|p^-_{j_i}| \le \mathrm{e}^{j_i\top}(\operatorname{Abs}\bar{P}^-)\mathrm{e}$ (i.e. if $\nu_i = 0$), and $\alpha_i = \nu_i\, \mathrm{e}^{i\top} B\xi / (p^-_{j_i} + \mathrm{e}^{j_i\top}\bar{P}^-\xi)$ otherwise. Inequalities $|\alpha_i| \le \hat{a}_i^{j_i}$, $i = 1, \dots, n$ (which ensure $A \in \mathcal{A}$) are obtained by obvious estimates. □

Note that Theorem 1 describes the whole family of estimates $\mathcal{P}^-(\cdot)$ where $J(\cdot)$ and $\Gamma(\cdot)$ are parameters.

Remark 1. Obviously, we have $\mathcal{X}(t) \supseteq \mathcal{X}^0(t) \equiv \mathcal{P}^{0-}(t)$, $t \in T$, where $\mathcal{X}^0(t)$ are reachable sets of the system (1) under assumptions $x_0 \in \mathcal{P}_0$, $w(\cdot) \equiv r(\cdot)$ and $A(\cdot) \equiv \tilde{A}(\cdot)$, and parallelotopes $\mathcal{P}^{0-}(t)$ are determined by (6), (7) when $\nu \equiv 0$, $\Gamma \equiv 0$. We call these parallelotopes $\mathcal{P}^{0-}(t)$ *trivial internal estimates for* $\mathcal{X}(t)$.

The following corollary compares internal estimates $\mathcal{P}^-(t)$ for $\mathcal{X}(t)$ satisfying (6), (7) with trivial internal estimates $\mathcal{P}^{0-}(t)$ for $\mathcal{X}(t)$ in the sense of volume. We like to remind that volume of a nondegenerate parallelotope $\mathcal{P} = \mathcal{P}[p, \bar{P}] \subset \mathbb{R}^n$ is equal to $\operatorname{vol}\mathcal{P} = 2^n |\det \bar{P}|$.

Corollary 1. *Under conditions of Theorem 1, we have* $\operatorname{vol}\mathcal{P}^-(t) = \operatorname{vol}\mathcal{P}^{0-}(t) \cdot \exp(\psi_1(t) + \psi_2(t))$, $t \in T_1$, *where* $\psi_1(t) = \int_0^t \nu(\tau, \bar{P}^-(\tau); J(\tau))^\top \beta(\bar{P}^-(\tau))d\tau$, $\psi_2(t) = \int_0^t \operatorname{tr}(\Xi(\tau, \bar{P}^-(\tau))\Gamma(\tau))d\tau$, $\Xi(t, \bar{P}^-) = (\bar{P}^-)^{-1}\bar{R}(t)$.

Therefore under additional Assumption 2 we have:
(i) $\operatorname{vol}\mathcal{P}^-(t) \geq \operatorname{vol}\mathcal{P}^{0-}(t)$, *and* $\operatorname{vol}\mathcal{P}^-(t) > \operatorname{vol}\mathcal{P}^{0-}(t)$ *iff* $\psi_1(t) > 0$*;*
(ii) if it is turned out that $\mathcal{P}^-(t) \ni 0$ *for all* $t \in T$, *then* $\mathcal{P}^-(t) \equiv \mathcal{P}^{0-}(t)$, $t \in T$.

Proof. The expression for $\operatorname{vol}\mathcal{P}^-(t)$ follows from the equality $\det \bar{P}^-(t) = m_0(t) \cdot \exp(\psi_1(t) + \psi_2(t))$, where $m_0(t) = \det \bar{P}_0 \exp \int_0^t \operatorname{tr} \tilde{A}(\tau)\, d\tau$, which, in turn, can be obtained similarly to [17, p. 293]. Namely, we use the change of variables $\bar{P}^-(t) = \Phi(t)\, P(t)$, where Φ satisfies $\dot{\Phi} = \tilde{A}\,\Phi$, $\Phi(0) = I$, and obtain the relation

$$\frac{d \det P}{dt} \,/\, \det P = \nu(t, \bar{P}^-(t); J(t))^\top \beta(\bar{P}^-(t)) + \operatorname{tr}((\bar{P}^-)^{-1}\bar{R}\,\Gamma) \qquad (8)$$

on the base of the known relation $d \det P/dt = \det P \operatorname{tr}(P^{-1}\dot{P})$ and (7); then the Liouville formula $\det \Phi(t) = \exp \int_0^t \operatorname{tr} \tilde{A}(\tau)\, d\tau$ is used.

Assumption 2 yields $\psi_2(t) \equiv 0$. Thus (i) is evident, (ii) is true because we have $\psi_1(t) \equiv 0$ similarly to [20, Corollary 1]. □

In general case (without Assumption 2) we can obtain some differential inclusions which determine internal estimates on the whole time interval T. Consider two ways to do that. Following the first way, fix $J(\cdot) \in \mathbb{J}$ and $\Gamma(\cdot) \in \mathbb{G}$; if a denominator of some row of the matrix B vanishes, replace this row by a suitable set. The corresponding differential inclusion determines $\mathcal{P}^-(t)$ on T (see Theorem 2 below), but there is no guarantee for $\mathcal{P}^-(t)$ to be nondegenerate. The second way allows (using some considerations of "local" optimality of the estimate volume) to construct $\Gamma(\cdot)$ to ensure nondegenerate estimates on T (Theorem 3).

Following the first way, consider the matrix differential inclusion

$$\frac{d\bar{P}^-}{dt} \in \tilde{A}(t)\bar{P}^- + \operatorname{diag}\nu(t, \bar{P}^-; J(t)) \cdot \boldsymbol{B}(\bar{P}^-) + \bar{R}(t)\,\Gamma(t), \quad \bar{P}^-(0) = \bar{P}_0, \quad (9)$$

where $\nu(t, \bar{P}^-; J(t))$ is defined in (7), and $\boldsymbol{B}(\bar{P}^-)$ is the set of all matrices $B(\bar{P}^-)$ such that each row $\mathrm{e}^{i\top} B$ (i=1, …, n) satisfy the following conditions

$$\mathrm{e}^{i\top} B = \begin{cases} \mathrm{e}^{i\top}\bar{P}^- / (\mathrm{e}^{i\top}(\operatorname{Abs}\bar{P}^-)\,\mathrm{e}), & \text{if } \mathrm{e}^{i\top}(\operatorname{Abs}\bar{P}^-)\,\mathrm{e} \neq 0, \\ \text{arbitrary row such that } \|\mathrm{e}^{i\top} B\|_1 \leq 1, & \text{if } \mathrm{e}^{i\top}(\operatorname{Abs}\bar{P}^-)\,\mathrm{e} = 0. \end{cases} \qquad (10)$$

Theorem 2. *For arbitrary* $J(\cdot) \in \mathbb{J}$, $\Gamma(\cdot) \in \mathbb{G}$, *there exists a solution* $(p^-(\cdot), \bar{P}^-(\cdot))$ *of the system (6), (9)–(10) which is determined on the whole* T, *and all solutions of this system determine internal parallelotope-valued estimates* $\mathcal{P}^-(t)$ *for* $\mathcal{X}(t)$, $t \in T$.

Proof. Existence and extendability are obtained using [10, p. 66, Theorem 6]. The inclusions $\mathcal{P}^-(t) \subseteq \mathcal{X}(t)$ are proved similarly to Theorem 1. □

Following the second way under conditions of Theorem 1, assume, without loss of generality, that $\det \bar{P}_0 > 0$. The idea of local optimization arises from (8) and consists in finding the maximal possible velocity of increasing $\det \bar{P}^-(t)$ (therefore $\operatorname{vol} \mathcal{P}^-(t)$) at time t, by the choice of the value Γ, when the value $\bar{P}^- = \bar{P}^-(t)$ has already been found. Consider the set $\boldsymbol{\Gamma}(t, \bar{P}^-)$ of matrices $\Gamma(t, \bar{P}^-)$, which are solutions to the following optimization problem: $\max\{\operatorname{tr}(\Xi(t, \bar{P}^-)\Gamma) \mid \Gamma \in \mathbb{R}^{m\times n} \text{ s.t. } \|\Gamma\| \le 1\}$. This set $\boldsymbol{\Gamma}(t, \bar{P}^-)$ may be described in the following form:

$$\boldsymbol{\Gamma}(t, \bar{P}^-) = \{\Gamma(t, \bar{P}^-) = \{\gamma_k^i(t, \bar{P}^-)\} \mid \gamma_k^i(t, \bar{P}^-) = \operatorname{sign}(\xi_i^k(t, \bar{P}^-))\, l_k^i, \quad k = 1, \dots, m,\ i = 1, \dots, n,\ L = \{l_k^i\} \in \boldsymbol{L}\}. \tag{11}$$

Here $\Xi(t, \bar{P}^-) = \{\xi_i^k(t, \bar{P}^-)\} = (\bar{P}^-)^{-1}\bar{R}(t) \in \mathbb{R}^{n\times m}$, and $\boldsymbol{L}$ is a set of matrices $L = \{l_k^i\} \in \mathbb{R}^{m\times n}$ satisfying conditions $l_k^i \ge 0$, $k = 1, \dots, m$, $i = 1, \dots, n$; $l_k^i = 0$ if $i \notin I_k(t, \bar{P}^-)$, $k{=}1, \dots, m$, $i{=}1, \dots, n$; $\sum_{i=1}^n l_k^i = 1$, $k{=}1, \dots, m$; $I_k(t, \bar{P}^-) = \operatorname{Argmax}\{|\xi_i^k(t, \bar{P}^-)| \mid\mid i{=}1, \dots, n\}$, $k{=}1, \dots, m$, where $\operatorname{sign} z$ is equal to $-1, 0, 1$ for $z < 0$, $z = 0$, $z > 0$ respectively. Consider the matrix differential inclusion

$$\frac{d\bar{P}^-}{dt} \in \tilde{A}(t)\bar{P}^- + \operatorname{diag}\nu(t, \bar{P}^-; J(t)) \cdot B(\bar{P}^-) + \bar{R}(t)\,\boldsymbol{\Gamma}(t, \bar{P}^-),\ \bar{P}^-(0) = \bar{P}_0, \tag{12}$$

where $\nu(t, \bar{P}^-; J)$ and $B(\bar{P}^-)$ are the same as in (7).

Theorem 3. *Let the above conditions be satisfied and $\mathcal{P}_0$ be a nondegenerate parallelepiped with $\det \bar{P}_0 > 0$. Then, for each function $J(\cdot) \in \mathbb{J}$, there exists a solution $(p^-(\cdot), \bar{P}^-(\cdot))$ of system (6), (11)–(12), which is determined on the whole interval T, and all solutions of this system determine parallelotopes $\mathcal{P}^-(t)$ which turn out to be internal nondegenerate parallelepiped-valued estimates for $\mathcal{X}(t)$, $t \in T$.*

Proof. We use arguments similar to [17, Theorem 5.2]. Existence and extendability are obtained using [10, p. 66, Theorem 6]. In particular, (8) is used to see that the function $P = \Phi^{-1}\bar{P}^-$ is such that $\det P(t)$ is a nondecreasing function and therefore $P(t)$ can not leave the domain where $\det P(t) \ge \det P(0) > 0$; consequently, solutions to (12) are defined on the whole interval T and determine nondegenerate parallelepiped-valued estimates. The inclusions $\mathcal{P}^-(t) \subseteq \mathcal{X}(t)$ are proved similarly to Theorem 1. □

Remark 2. We can choose $J(\cdot)$ in (7), (9) and (12) in different ways, in particular as a constant. A simple way is also to apply a "local" optimization which arises from (8). Fix a natural number N and introduce a grid T_N of times $\tau_k = kh_N$, $k{=}0, \dots, N$, $h_N = \theta N^{-1}$. Let us, for each $\tau_k \in T_N$, solve the optimization problem which is to maximize $\nu(\tau_k, \bar{P}^-; J)^\top \beta(\bar{P}^-)$ over all possible permutations $J = \{j_1, \dots, j_n\}$ assuming that $\bar{P}^- = \bar{P}^-(\tau_k)$ has already been found. Then we can sequentially construct the piecewise constant function $J(t) \equiv J(\tau_k) \in \operatorname{Argmax}_J \nu(\tau_k, \bar{P}^-(\tau_k); J)^\top \beta(\bar{P}^-(\tau_k))$, $t \in [\tau_k, \tau_{k+1})$, $k = 0, \dots, N-1$, and find

$\bar{P}^-(\cdot)$. Note that the described procedure is not obliged to give the estimates $\mathcal{P}^-(t)$ with maximal volume even if $N \to \infty$.

5 External Estimates

In [18], the ODE systems of two types were obtained for external estimates for $\mathcal{X}(t)$ in the form of parallelepipeds $\mathcal{P}^+(t) = \mathcal{P}(p^+(t), P(t), \pi^+(t))$, where $P(t)$ is a fixed matrix function. Let us restate here, for completeness of the exposition, the ODE system for the more accurate estimates of the type II:

$$\frac{dp^+}{dt} = \dot{P}P^{-1}p^+ + P\,(\Phi^{(+)} - \Phi^{(-)})/2 + r, \quad p^+(0) = p_0; \tag{13}$$

$$\frac{d\pi^+}{dt} = (\Phi^{(+)} + \Phi^{(-)})/2 + \mathrm{Abs}\,(P^{-1}\bar{R})\,\mathrm{e}, \quad \pi^+(0) = \mathrm{Abs}\,(P(0)^{-1}P_0)\,\pi_0,$$
$$\text{where } \Phi_i^{(\pm)} = \max_{\xi \in \boldsymbol{\Xi}_i^{\pm}} \left(\pm P^{-1}(\tilde{A} - \dot{P}P^{-1})x + \mathrm{Abs}\,(P^{-1})\hat{A}\,\mathrm{Abs}\,x\right)_i, \tag{14}$$
$$x = p^+ + P\mathrm{diag}\,\pi^+\,\xi;\ \boldsymbol{\Xi}_i^{\pm} = \{\xi \,|\, \xi \in \boldsymbol{E}(\mathcal{P}(0, I, \mathrm{e})),\ \xi_i = \pm 1\},\ i = 1, \ldots, n,$$

the symbol $\boldsymbol{E}(\mathcal{P})$ denotes the set of all vertices of a parallelepiped $\mathcal{P} = \mathcal{P}(p, P, \pi)$, namely the set of points of the form $x = p + \sum_{j=1}^{n} p^j \pi_j \zeta_j$, $\zeta_j \in \{-1, 1\}$.

Theorem 4. *Let Assumption 1 be satisfied and $P(t) \in \mathbb{R}^{n\times n}$ be an arbitrary continuously differentiable function such that $\det P(t) \neq 0$, $t \in T$. Then the system (13),(14) has a unique solution $(p^+(\cdot), \pi^+(\cdot))$ on T, and the parallelepipeds $\mathcal{P}^+(t) = \mathcal{P}(p^+(t), P(t), \pi^+(t))$ are the external estimates for the reachable sets $\mathcal{X}(t)$ of the system (1), (2), (4): $\mathcal{X}(t) \subseteq \mathcal{P}^+(t)$, $t \in T$.*

Proof. The existence, uniqueness and extendability of the solution follow from [10, pp. 7,8,10], the inclusions — from [18, Theorem 1]. □

Remark 3. In fact, Theorem 4 describes the whole family of estimates where $P(\cdot)$ is a parameter. Some heuristic ways of choosing $P(\cdot)$ were indicated in [18] (in particular, (i) find $P(\cdot)$ from relations $\dot{P} = \tilde{A}(t)P$, $P(0) = P_0$, or (ii) put $P(t) \equiv I$).

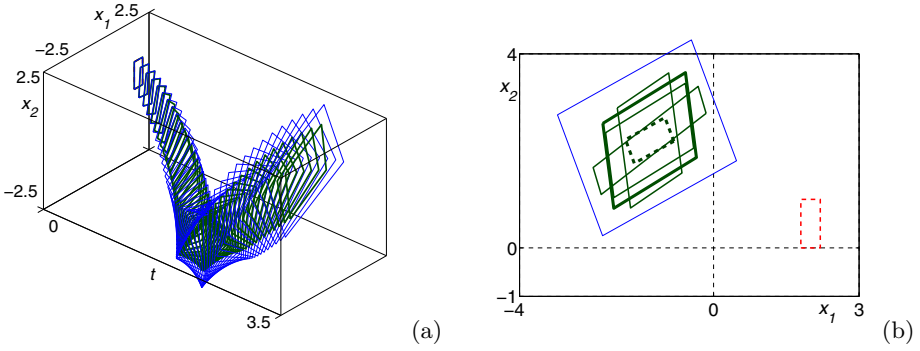


Fig. 1. External and internal estimates for $\mathcal{X}[\cdot]$ (a) and $\mathcal{X}[N]$ (b) in Example 1

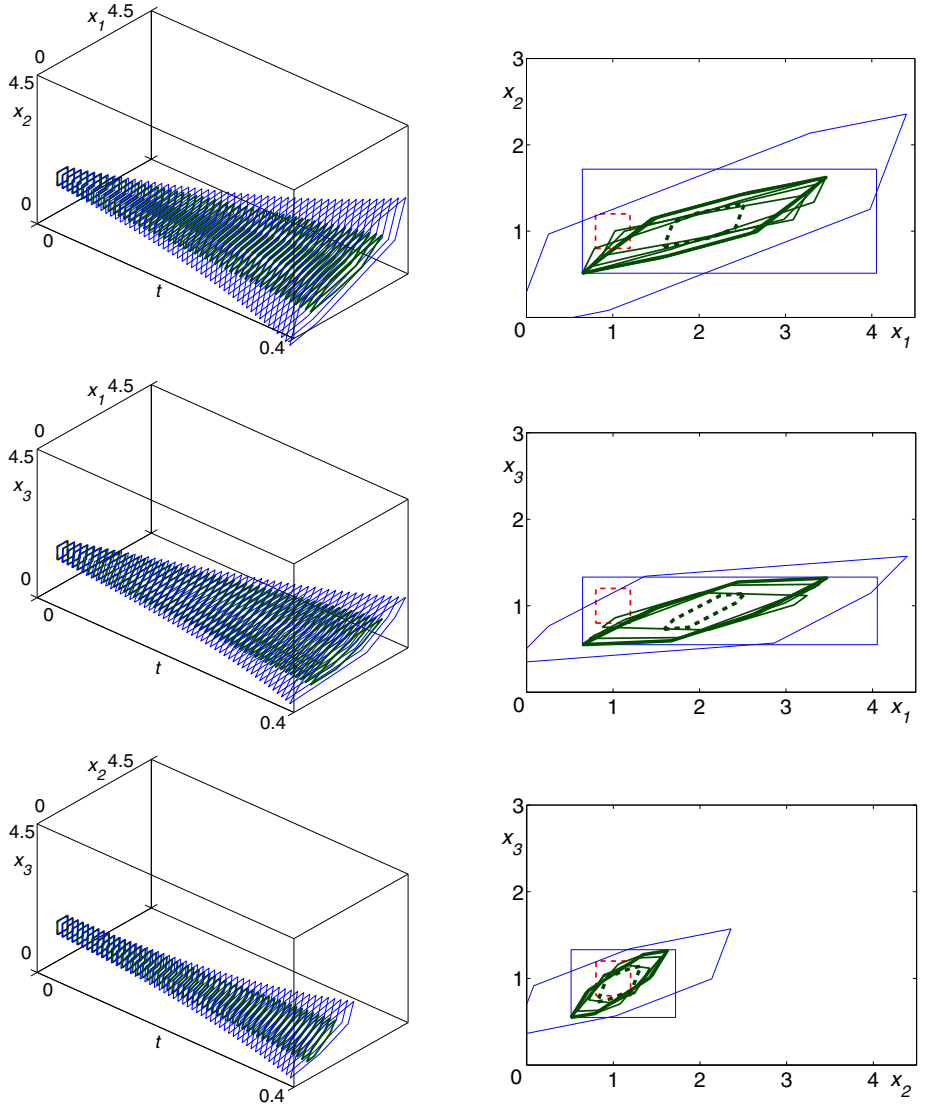


Fig. 2. Projections of external and internal estimates for $\mathcal{X}[\cdot]$ and $\mathcal{X}[N]$ in Example 2

6 Examples

Consider some examples. The estimates were calculated using the Euler approximations (5) with $N = 100$ (in fact, the estimates for $\mathcal{X}[k]$ are presented in figures below). But it would be emphasized that different schemes of approximation can be used for solving the obtained differential systems and finding the estimates.

Example 1. Let $\tilde{A} \equiv \begin{bmatrix} 0 & 1 \\ -1.5 & 0 \end{bmatrix}$, $\hat{A} \equiv \begin{bmatrix} 0 & 0 \\ 0.1 & 0 \end{bmatrix}$, $\mathcal{R} \equiv \mathcal{P}(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, I, \begin{bmatrix} 0 \\ 0.3 \end{bmatrix})$, $\mathcal{P}_0 = \mathcal{P}((2, 0.5)^\top, I, (0.2, 0.5)^\top)$, $\theta = 3.5$. Fig. 1(a) presents tubes formed by external (see Remark 3, (i)) and internal estimates for $\mathcal{X}[k]$. The internal ones are obtained by discrete analogous to Theorem 3 (similar to [17, Example 6.1]). Fig. 1(b) shows the initial set $\mathcal{P}_0$ (*dashed line*), the external estimate for $\mathcal{X}[N]$ (*thin line*) and four internal ones. Three of them correspond to "quasistationary" functions $\mathit{\Gamma}(\cdot)$ (similarly to [17, Example 6.1]), the last-named (*thick line*) corresponds to Theorem 3. For comparison, the trivial internal estimate $\mathcal{P}^{0-}[N]$ is shown too (*dashed thick line*); it is the "smallest" of the presented internal estimates.

Example 2. Let $\tilde{A} \equiv \begin{bmatrix} -1 & 0 & 5 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$, $\hat{A} \equiv \begin{bmatrix} 0 & 0 & 3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$, $\mathcal{R} \equiv \mathcal{P}(\begin{bmatrix} -0.6 \\ -0.4 \\ -0.2 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0.4 \end{bmatrix})$, $\mathcal{P}_0 = \mathcal{P}((1, 1, 1)^\top, I, (0.2, 0.2, 0.2)^\top)$, $\theta = 0.4$. Such system may be interpreted as a simple ecological model of dynamics of a number of microorganisms which have 3 stages of development, provide division at the last stage and produce from 2 to 8 descendants [33, p. 112]. The additive control describes injecting a preparation to reduce the population. Estimates for $\mathcal{X}[\cdot]$ and $\mathcal{X}[N]$ are shown in Fig. 2, where drawings are similar to Fig. 1. Since parallelotopes here are three-dimensional, we present their two-dimensional projections on coordinate plains. The reachable sets belong to the intersection of external estimates and contain the internal ones.

It must be admitted that the proposed estimates may turn out to be rather conservative. But we can calculate them easily via integration of the ODE, and they can give useful information, while it is hard to calculate exact reachable sets. Improved external (possibly nonconvex) estimates in the form of the union of parallelepipeds can be constructed for systems with constant coefficients [18].

Acknowledgments. The work was supported by the Program of the Presidium of the Russian Academy of Sciences No. 17 "Dynamic Systems and Control Theory" and by the Russian Foundation for Basic Research (grant 12-01-00043).

References

1. Artstein, Z., Raković, S.V.: Feedback and Invariance under Uncertainty via Set-Iterates. Automatica 44(2), 520–525 (2008)
2. Baier, R., Büskens, C., Chahma, I.A., Gerdts, M.: Approximation of Reachable Sets by Direct Solution Methods of Optimal Control Problems. Optim. Methods Softw. 22(3), 433–452 (2007)

3. Baier, R., Gerdts, M.: A Computational Method for Non-convex Reachable Sets Using Optimal Control. In: Proceedings of the European Control Conference (ECC), Budapest, Hungary, August 23-26, pp. 97–102 (2009)
4. Baier, R., Lempio, F.: Computing Aumann's Integral. In: Kurzhanski, A.B., Veliov, V.M. (eds.) Modeling Techniques for Uncertain Systems, Proc. of a Conferences held in Sopron, Hungary, July 6-10 (1992); Progress in Systems and Control Theory, vol. 18, pp. 71–92. Birkhäuser, Boston (1994)
5. Barmish, B.R., Sankaran, J.: The Propagation of Parametric Uncertainty via Polytopes. IEEE Trans. Automat. Control. AC 24(2), 346–349 (1979)
6. Bushenkov, V., Chernykh, O., Kamenev, G., Lotov, A.: Multi-dimensional Images Given by Mappings: Construction and Visualization. Pattern Recognition and Image Analysis 5(1), 35–56 (1995)
7. Chernousko, F.L., Rokityanskii, D.Y.: Ellipsoidal Bounds on Reachable Sets of Dynamical Systems with Matrices Subjected to Uncertain Perturbations. J. Optim. Theory Appl. 104(1), 1–19 (2000)
8. Digailova, I.A., Kurzhanski, A.B.: On the Joint Estimation of the Model and State of an Under-Determined System from the Results of Observations. Dokl. Math. 65(3), 459–464 (2002)
9. Dontchev, A.L., Farkhi, E.M.: Error Estimates for Discretized Differential Inclusions. Computing 41(4), 349–358 (1989)
10. Filippov, A.F.: Differential Equations with Discontinuous Right-Hand Sides. Nauka, Moscow (1985) (Russian)
11. Filippova, T.F.: Trajectory Tubes of Nonlinear Differential Inclusions and State Estimation Problems. J. Concr. Appl. Math. 8(3), 454–469 (2010)
12. Filippova, T.F., Lisin, D.V.: On the Estimation of Trajectory Tubes of Differential Inclusions. Proc. Steklov Inst. Math. Suppl. 2, S28–S37 (2000)
13. Gusev, M.I.: Estimates of Reachable Sets of Multidimensional Control Systems with Nonlinear Interconnections. Proc. Steklov Inst. Math. Suppl. 2, S134–S146 (2010)
14. Jaulin, L., Kieffer, M., Didrit, O., Walter, E.: Applied Interval Analysis. Springer, London (2001)
15. Kornoushenko, E.K.: Interval Coordinatewise Estimates for the Set of Accessible States of a Linear Stationary System. I–IV. Autom. Remote Control 41, 598–606 (1980), 41, 1633–1639 (1981), 43, 1266–1270 (1983), 44, 203–208 (1983)
16. Kostousova, E.K.: External and Internal Estimation of Attainability Domains by Means of Parallelotopes. Vychisl. Tekhnol. 3(2), 11–20 (1998) (Russian), http://www.ict.nsc.ru/jct/search/article?l=eng
17. Kostousova, E.K.: Control Synthesis via Parallelotopes: Optimization and Parallel Computations. Optim. Methods Softw. 14(4), 267–310 (2001)
18. Kostousova, E.K.: Outer Polyhedral Estimates for Attainability Sets of Systems with Bilinear Uncertainty. J. Appl. Math. Mech. 66(4), 547–558 (2002)
19. Kostousova, E.K.: On Polyhedral Estimates for Reachable Sets of Discrete-Time Systems with Bilinear Uncertainty. Autom. Remote Control. 72, 1841–1851 (2011)
20. Kostousova, E.K.: On Polyhedral Estimates for Trajectory Tubes of Dynamical Discrete-Time Systems with Multiplicative Uncertainty. In: Discrete Contin. Dyn. Syst., Dynamical Systems, Differential Equations and Applications. 8th AIMS Conference, Suppl., pp. 864–873 (2011)
21. Kostousova, E.K.: On Polyhedral Estimates for Reachable Sets of Differential Systems with a Bilinear Uncertainty. Trudy Instituta Matematiki i Mekhaniki UrO RAN 18(4) (to appear, 2012) (Russian)

22. Kostousova, E.K., Kurzhanski, A.B.: Guaranteed Estimates of Accuracy of Computations in Problems of Control and Estimation. Vychisl. Tekhnol. 2(1), 19–27 (1997) (Russian)
23. Krastanov, M., Kirov, N.: Dynamic Interactive System for Analysis of Linear Differential Inclusions. In: Kurzhanski, A.B., Veliov, V.M. (eds.) Modeling Techniques for Uncertain Systems, Proc. of a Conferences Held in Sopron, Hungary, July 6-10 (1992); Progress in Systems and Control Theory, vol. 18, pp. 123–130. Birkhäuser, Boston (1994)
24. Kuntsevich, V.M., Kurzhanski, A.B.: Calculation and Control of Attainability Sets for Linear and Certain Classes of Nonlinear Discrete Systems. J. Automation and Inform. Sci. 42(1), 1–18 (2010)
25. Kurzhanski, A.B., Vályi, I.: Ellipsoidal Calculus for Estimation and Control. Birkhäuser, Boston (1997)
26. Kurzhanski, A.B., Varaiya, P.: On Ellipsoidal Techniques for Reachability Analysis. Part I: External Approximations. Optim. Methods Softw. 17, 177–206 (2002)
27. Kurzhanski, A.B., Varaiya, P.: On Ellipsoidal Techniques for Reachability Analysis. Part II: Internal Approximations. Box-valued Constraints. Optim. Methods Softw. 17(2), 207–237 (2002)
28. Nazin, S.A., Polyak, B.T.: Interval Parameter Estimation Under Model Uncertainty. Math. Comput. Model. Dyn. Syst. 11(2), 225–237 (2005)
29. Polyak, B.T., Nazin, S.A., Durieu, C., Walter, E.: Ellipsoidal Parameter or State Estimation under Model Uncertainty. Automatica J. IFAC. 40(7), 1171–1179 (2004)
30. Taras'yev, A.M., Uspenskiy, A.A., Ushakov, V.N.: Approximation Schemas and Finite-Difference Operators for Constructing Generalized Solutions of Hamilton-Jacobi Equations. J. Comput. Systems Sci. Internat. 33(6), 127–139 (1995)
31. Veliov, V.M.: Second Order Discrete Approximations to Strongly Convex Differential Inclusions. Systems & Control Letters 13(3), 263–269 (1989)
32. Wolenski, P.R.: The Exponential Formula for the Reachable Set of a Lipschitz Differential Inclusion. SIAM J. Control Optim. 28(5), 1148–1161 (1990)
33. Zaslavskii, B.G., Poluektov, R.A.: Control of Ecological Systems. Nauka, Moscow (1988) (Russian)

An Algorithm for Two-Stage Stochastic Quadratic Problems

Eugenio Mijangos*

Department of Applied Mathematics
and Statistics and Operations Research
University of the Basque Country UPV/EHU, Spain
eugenio.mijangos@ehu.es

Abstract. An algorithm for solving quadratic, two-stage stochastic problems is developed. The algorithm is based on the framework of the Branch and Fix Coordination (BFC) method. These problems have continuous and binary variables in the first stage and only continuous variables in the second one. The objective function is quadratic and the constraints are linear. The nonanticipativity constraints are fulfilled by means of the twin node family strategy. On the basis of the BFC method for two-stage stochastic linear problems with binary variables in the first stage, an algorithm to solve these stochastic quadratic problems is designed. In order to gain computational efficiency, we use scenario clusters and propose to use either outer linear approximations or (if possible) perspective cuts. This algorithm is implemented in C++ with the help of the Cplex library to solve the quadratic subproblems. Numerical results are reported.

Keywords: Stochastic Programming, Mixed-Integer Quadratic Problems, Branch-and-Fix Coordination, Perspective Cuts.

1 Introduction

Two-stage stochastic mixed integer programs are among the most interesting problems since the complexity generated by the integrality of variables and the high dimensionality. Stochastic parameters can exist anywhere in the problem. In order to model the uncertainty a finite set of scenarios, Ω, is used, where each $\omega \in \Omega$ has an associated probability of ocurrence p^ω. In a two-stage program decisions on the first and second stage variables must be taken. First-stage variables are chosen before knowing the realization of the uncertain parameters. After having decided on first stage and having known each realization of the uncertain parameters, the second stage decision must be taken. The first-stage variables take the same value in each scenario, which yields *nonanticipativity constraints*. If we consider a finite number of scenarios, a general two-stage program can be expressed regarding the first-stage variables being equivalent to a large

* This work was partially supported by the Ministry of Science and Technology of Spain through MICINN Project DPI2008-02153.

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 177–187, 2013.

programming problem suggested in [10] and known as *deterministic equivalent model* (DEM). A general two-stage problem can include binary first-stage variables. The simplest form of two-stage stochastic integer problems have first-stage binary and second-stage continuous variables. In [7] a branch-and-cut method is used for those problems, which is based on the Benders decomposition method. An efficient branch-and-fix coordination (BFC) method for solving two-stage programs is provided in [1], where the first-stage has only binary variables, and where the uncertainty only appears in the coefficients of the objective function and in the right-hand-side of the constraints. If the first stage involves pure binary variables, finite termination is justified by using branching over the 0-1 first-stage variables, see among others [8] and [9]. Escudero *et al.* [3] study general two-stage stochastic mixed 0-1 problems, where the first stage only involves binary variables and continuous variables and the second stage continuous variables. They use a specialization of the BFC scheme and the twin-node-family (TNF) concept, which was introduced in [1]. Their scheme is specifically designed for coordinating the node branching selection and pruning and the 0-1 variable branching selection and fixing at each branch-and-fix (BF) tree. Also, they suggest to decompose the set of scenarios in clusters.

On the other hand, real problems with this structure exist and have a high dimensionality. They often need to be solved and it is important to find the procedure that will solve them with the highest efficiency. An example of this type is the Iberian Electricity Market (MIBEL), which comprises Spanish and Portuguese electricity systems, see [6].

In this paper we consider the two-stage mixed 0-1 quadratic problem

$$\begin{aligned} \text{minimize} \;\; & c^t\delta + q(x,y) \\ \text{subject to}: \;\; & l_a \le A \begin{bmatrix} \delta \\ x \end{bmatrix} \le u_a, \\ & l_t \le T \begin{bmatrix} \delta \\ x \\ y \end{bmatrix} \le u_t, \\ & x \ge 0, \; \underline{y} \le y \le \overline{y}, \; \delta \in \{0,1\}^{n_\delta}, \end{aligned}$$

where δ are first-stage-binary variables, $x \in \mathbf{R}^{n_x}$ are first-stage continuous variables, $y \in \mathbf{R}^{n_y}$ are second-stage continuous variables, c is the coefficient vector for δ, and q is the quadratic function defined as follows

$$q(x,y) = b^t \begin{bmatrix} x \\ y \end{bmatrix} + [x^t \;\; y^t] Q \begin{bmatrix} x \\ y \end{bmatrix},$$

where Q is a positive-definite matrix and b and Q are partitioned as $\begin{bmatrix} x \\ y \end{bmatrix}$; i.e.,

$$b = \begin{bmatrix} b_x \\ b_y \end{bmatrix} \quad \text{and} \quad Q = \begin{bmatrix} Q_{xx} \; Q_{xy} \\ Q_{yx} \; Q_{yy} \end{bmatrix}.$$

In addition, l_a and u_a are the bounds for the first-stage constraints and l_t and u_t are the bounds for the second-stage constraints.

Let us suppose that some of the coefficient in b_y, Q_{xy}, Q_{yx}, Q_{yy}, l_t, u_t and T are uncertain. The uncertainty is given by the scenarios ω in the finite set Ω and p^ω is the probability of that occurs $\omega \in \Omega$. Therefore, the initial problem given in a stochastic way can be written as the so-called Deterministic Equivalent Model **(DEM)**

$$\text{minimize } c^t\delta + \sum_{\omega\in\Omega} p^\omega q^\omega(x, y^\omega) \tag{1}$$

$$\text{subject to : } l_a \le A\begin{bmatrix}\delta\\ x\end{bmatrix} \le u_a, \tag{2}$$

$$l_t^\omega \le T^\omega \begin{bmatrix}\delta\\ x\\ y^\omega\end{bmatrix} \le u_t^\omega,\ \omega\in\Omega, \tag{3}$$

$$x \ge 0,\ \underline{y} \le y^\omega \le \overline{y},\ \omega\in\Omega, \tag{4}$$

$$\delta \in \{0,1\}^{n_\delta}, \tag{5}$$

As is shown by [3] the compact representation **DEM** can be written as a *splitting variable* representation; i.e., δ and x are respectively replaced by δ^ω and x^ω, for $\omega \in \Omega$. So, we have

$$\begin{aligned}
\textbf{(MIQP)}\quad &\text{minimize } \sum_{\omega\in\Omega} p^\omega(c^t\delta^\omega + q^\omega(x^\omega, y^\omega))\\
&\text{subject to : } l_a \le A\begin{bmatrix}\delta^\omega\\ x^\omega\end{bmatrix} \le u_a,\ \omega\in\Omega,\\
&\qquad l_t^\omega \le T^\omega\begin{bmatrix}\delta^\omega\\ x^\omega\\ y^\omega\end{bmatrix} \le u_t^\omega,\ \omega\in\Omega,\\
&\qquad x^\omega \ge 0,\ \underline{y} \le y^\omega \le \overline{y},\quad \delta^\omega \in \{0,1\}^{n_\delta},\ \omega\in\Omega,\\
&(\text{NAC}_\delta)\ \ \delta^\omega - \delta^{\omega'} = 0,\ \forall\omega,\omega'\in\Omega :\ \omega\ne\omega',\\
&(\text{NAC}_x)\ \ x^\omega - x^{\omega'} = 0,\ \forall\omega,\omega'\in\Omega :\ \omega\ne\omega',
\end{aligned}$$

where NAC_δ and NAC_x are the *nonanticipativity constraints.*

Note that the relaxation of the NACs in the model **MIQP** gives rise to $|\Omega|$ independent $\textbf{MIQP}^\omega$ submodels

$$\text{minimize } p^\omega(c^t\delta^\omega + q^\omega(x^\omega, y^\omega)), \tag{6}$$

$$\text{subject to : } l_a \le A\begin{bmatrix}\delta^\omega\\ x^\omega\end{bmatrix} \le u_a, \tag{7}$$

$$l_t^\omega \le T^\omega\begin{bmatrix}\delta^\omega\\ x^\omega\\ y^\omega\end{bmatrix} \le u_t^\omega, \tag{8}$$

$$x^\omega \ge 0,\ \underline{y} \le y^\omega \le \overline{y}, \tag{9}$$

$$\delta^\omega \in \{0,1\}^{n_\delta}, \tag{10}$$

and these models are linked by the NACs, which force the equality of the first-stage variables.

In this work to solve the original quadratic problem **DEM** a Branch-and-Fix-Coordination scheme (BFC) is used for each scenario $\omega \in \Omega$ to fulfill the integrality condition (IC) given by (10), so that the NAC_δ are also satisfied when selecting branching nodes and branching variables by the Twin-Node-Families concept (TNF), which was introduced by [1]. A similar approach to that suggested in [3] is used in this work to coordinate the selection of the branching node and branching variable for each scenario-related BF tree, such that the NAC_δ are satisfied when fixing δ^ω, $\forall \omega \in \Omega$, either to 1 or to 0. A *TNF integer set* is a set of integer BF nodes (i.e. they verify IC), one per BF tree, in which the NAC_δ are verified.

For each TNF integer set we use two quadratic submodels. The quadratic model $\mathbf{QP}^{TNF}$ obtained after fixing in **DEM** $\delta = \overline{\delta} \in \{0,1\}^{n_\delta}$ for a TNF integer set

$$
\begin{aligned}
(\mathbf{QP}^{TNF}) \quad Z^{TNF} = c^t\overline{\delta} + \ & \min \sum_{\omega\in\Omega} p^\omega q^\omega(x, y^\omega) \\
& \text{subject to}: \ l_a \le A \begin{bmatrix} \overline{\delta} \\ x \end{bmatrix} \le u_a, \\
& \qquad\qquad l_t^\omega \le T^\omega \begin{bmatrix} \overline{\delta} \\ x \\ y^\omega \end{bmatrix} \le u_t^\omega, \ \omega \in \Omega, \\
& \qquad\qquad x \ge 0, \ \underline{y} \le y^\omega \le \overline{y}, \ \omega \in \Omega,
\end{aligned}
$$

It gives a feasible solution and a possible incumbent solution.

The second quadratic submodel to solve at a TNF integer set corresponds to the case where not all the δ variables have been branched on in the current TNF, but all of them hold the integrality condition. Then, the other quadratic submodel $(\mathbf{QP}^f)$ is obtained from problem **DEM** with $\delta = \begin{pmatrix} \overline{\delta} \\ \delta^f \end{pmatrix}$, where $\overline{\delta}_j$, for $j \in \{1, \ldots, k\}$, are fixed to 0-1 values and the componentes of δ_j^f are in the interval $[0, 1]$.

$$
\begin{aligned}
(\mathbf{QP}^f) \quad Z^f = \ & \min c^t\delta + \sum_{\omega\in\Omega} p^\omega q^\omega(x, y^\omega) \\
& \text{subject to}: \ l_a \le A \begin{bmatrix} \delta \\ x \end{bmatrix} \le u_a, \\
& \qquad\qquad l_t^\omega \le T^\omega \begin{bmatrix} \delta \\ x \\ y^\omega \end{bmatrix} \le u_t^\omega, \ \omega \in \Omega, \\
& \qquad\qquad x \ge 0, \ \underline{y} \le y^\omega \le \overline{y}, \ \omega \in \Omega, \\
& \qquad\qquad \delta_j = \overline{\delta}_j \text{ fixed to 0-1, for } j \in \{1, \ldots, k\} \\
& \qquad\qquad \delta_j = \delta_j^f \in [0,1], \quad \text{for } j \in \{k+1, \ldots, n^\delta\}
\end{aligned}
$$

This model contributes strong lower bounds of the solution value of the descendent nodes from a given node, by satisfying the NAC_x.

1.1 Outline of BFC

This method branches on the δ-variables, obtaining the solution of the quadratic submodels $\mathbf{MIQP}^\omega$ and coordinating the selection of the branching node and branching variable for the BF trees, such that the NAC_δ constraints are fulfilled once fixed the suitable variables δ to 1 or to 0.

A sequence of lower bounds $\underline{Z}_i$ is computed, where $\underline{Z}_i = \sum_{\omega \in \Omega} z_i^\omega$ and z_i^ω is the solution to the quadratic relaxation $\mathbf{(QP}^\omega\mathbf{)}$ of $\mathbf{MIQP}^\omega$ once the previous variables δ have been fixed to 0 or to 1.

If the optimal solution obtained in each node of the TNF satisfies the IC (integrality constraints) and the NAC_δ, two cases can happen with respect to the NAC_x. If NAC_x are satisfied, the incumbent solution is updated and the TNF's branch is pruned; if the set of active nodes is empty, that solution is the optimum. Otherwise, to satisfy NAC_x we solve the TNF quadratic problem obtained by fixing the δ-variables that verified IC and NAC_δ; if this problem is feasible, the incumbent solution is updated, and if the TNF cannot be pruned, we continue with the tree examination. For more details about the BFC method for two-stage stochastic problems see [3].

1.2 Scenario Clusters

When the number of scenarios is very large, in order to gain computational efficiency we can take scenario clusters; see in [4] an information structuring for scenario cluster partitioning of nonsymmetric scenario trees.

Let $\widehat{p}$ be the number of clusters and $\Omega^1, \ldots, \Omega^{\widehat{p}}$, where $\Omega^p \cap \Omega^{p'} = \emptyset$ for $p, p' = 1, \ldots, \widehat{p}$, such that $p \neq p'$, and $\cup_{p=1}^{\widehat{p}} \Omega^p = \Omega$. So, instead of the submodel $\mathbf{MIQP}^\omega$ for $\omega \in \Omega$ we can consider the following submodel for the scenario cluster $p = 1, \ldots, \widehat{p}$

$$
\begin{aligned}
\mathbf{(MIQP}^p\mathbf{)} \quad \text{minimize} \quad & \sum_{\omega \in \Omega^p} p^\omega (c^t \delta^p + q^\omega(x^p, y^\omega)) \\
\text{subject to}: \quad & l_a \leq A \begin{bmatrix} \delta^p \\ x^p \end{bmatrix} \leq u_a, \\
& l_t^\omega \leq T^\omega \begin{bmatrix} \delta^p \\ x^p \\ y^\omega \end{bmatrix} \leq u_t^\omega, \omega \in \Omega^p \\
& x^p \geq 0, \ \underline{y} \leq y^\omega \leq \overline{y}, \ \omega \in \Omega^p, \quad \delta^p \in \{0,1\}^{n_\delta}
\end{aligned}
$$

These models are linked by the NACs $\delta^p - \delta^{p'} = 0$ and $x^p - x^{p'} = 0$, for all $p, p' \in \{1, \ldots, \widehat{p}\}$ such that $p \neq p'$.

However, since the number of branches to test can be huge, the BFC method has some troubles: the number of feasible solutions can be too high, a high number of quadratic problems QP^p, QP^{TNF}, and QP^f can exist to solve, and QP^{TNF} and QP^f can have very high dimensions. Hence, in order to gain computational efficiency, we propose to use either outer linear approximations or (if possible) perspective cuts to solve QP^p in each TNF (i.e., $\mathbf{MIQP}^p$ where the previous branching variables have been fixed and the rest is relaxed in $[0,1]$)

2 Outer Linear Approximations (OLA)

Let the problem $\min_{y\in Y} g(y)$, where g is convex and Y is a polyhedral set. The optimal value of that problem is not smaller than that of

$$\begin{aligned} &\text{minimize} && \eta \\ &\text{subject to :} && \eta \geq g(y_i) + \nabla g(y_i)^t (y - y_i), \\ & && y \in Y. \end{aligned}$$

Therefore, the value of η^* gives us an underestimate of $g(y^*)$ and, so, we can use it instead of $g(y^*)$ in the comparison with the current upper bound, in order to prune (or not) the current branch.

In our problem, we use as y_i the solution in the previous node.

3 Perspective Cuts

When Q_{xy} and Q_{yx} are zero matrices, the quadratic function q is defined as follows

$$q(x,y) = b_x^t x + b_y^t y + x^t Q_{xx} x + y^t Q_{yy} y.$$

This kind of model can be found in liberalized electricity markets [6] and [2].

For each scenario $\omega \in \Omega$ we can write the objective function of the submodel $\mathbf{MIQP}^\omega$ as follows

$$\begin{aligned} &p^\omega (c^t \delta^\omega + b_x^t x + (b_y^\omega)^t y^\omega + x^t Q_{xx} x + (y^\omega)^t Q_{yy}^\omega y^\omega) = \\ &p^\omega \left\{ \left(b_x^t x + x^t Q_{xx} x \right) + \left((y^\omega)^t Q_{yy}^\omega y^\omega + (b_y^\omega)^t y^\omega + c^t \delta^\omega \right) \right\} \end{aligned}$$

and, if $n := n^\delta = n^y$ and Q_{yy} is diagonal (as in [6]), we can write the last bracket as $\sum_{i=1}^n q_{ii}^\omega (y_i^\omega)^2 + b_i^\omega y_i^\omega + c_i \delta_i^\omega$.

For notational simplicity in this paragraph we drop the indices. The issue is then how to best represent the quadratic function $f(y,\delta) = qy^2 + by + c\delta$ by means of a piecewise-linear one. There is an effective way based on ideas developed by Frangioni and Gentile [5]. The function $f(y,\delta)$ is only relevant at points (y,δ) of its (disconnected) domain $\mathcal{D} = [0,0] \cup \left[[\underline{y}, \overline{y}] \times \{1\} \right]$. Standard branch-and-cut approaches typically solve the continuous relaxation of the mixed problem, where $\delta \in [0,1]$ instead of $\{0,1\}$, in order to obtain lower bounds on the optimal

value. This makes sense to use the *convex envelope* of $f(y,\delta)$ over $\mathcal{D}$, that is, the convex function with the smallest (in set-inclusion sense) epigraph containing that of $f(y,\delta)$. As is showed in [5] the convex envelope is

$$h(y,\delta) = \begin{cases} 0, & \text{if } (y,\delta) = (0,0) \\ \dfrac{qy^2}{\delta} + by + c\delta, & \left\{\begin{array}{l}\text{if } \delta\underline{y} \le y \le \delta\overline{y}, \\ \text{for } \delta \in (0,1]\end{array}\right\} \\ +\infty, & \text{otherwise.} \end{cases}$$

This function is strongly related with the *perspective-function* $\breve{f}(y,\delta) = \delta f(y/\delta)$ of $f(y) = qy^2 + by + c$, which is convex if $f(y)$ is convex.

$h(y,\delta) \ge f(y,\delta)$ for $0 < \delta \le 1$, i.e. h is a tighter objective function than f for the continuous relaxation. As is well-known, every convex function is the point-wise supremum of affine functions. In fact, the epigraph of h is composed of all and only triples (v,y,δ) satisfying $\delta\underline{y} \le y \le \delta\overline{y}$, $0 \le \delta \le 1$ and the infinite system of linear inequalities

$$v \ge (2q\widehat{y} + b)y + (c - q\widehat{y}^2)\delta$$

taking $\widehat{y} \in [\underline{y}, \overline{y}]$. For each $\widehat{y}$ we have an inequality so-called a *perspective cut* (PC), which is the unique supporting hyperplane to the function passing by $(0,0)$ and $(\widehat{y},1)$.

3.1 PC Formulation (PCF)

PC formulation (PCF) lies in choosing these supporting hyperplanes and using as an objective function the polyhedral function that is the point-wise maximum of the corresponding linear functions. A small set of initial PCs is chosen to solve the problem with the continuous relaxation. When $\delta^* > 0$, check whether the solution (v^*, y^*, δ^*) satisfies the PC for $\widehat{y} = y^*/\delta^*$; if not, the obtained cut can be added to PCF.

PCF starts with only two pieces, the ones corresponding with $\underline{y}$ and $\overline{y}$; additional cuts are then dynamically generated when needed as described in the previous paragraph.

Therefore, the objective function of $\mathbf{MIQP}^\omega$ for PCF becomes

$$p^\omega \left\{ \left(b_x^t x + x^t Q_{xx} x\right) + \left(\sum_{i=1}^{n} v_i^\omega\right) \right\},$$

and the initial PCs added to the constraints of MIQP^ω for each $i \in \{1,\ldots,n\}$

$$v_i^\omega \ge (2q_{ii}^\omega \underline{y}_i + b_i^\omega)y_i^\omega + (c_i - q_{ii}^\omega \underline{y}_i^2)\delta_i^\omega$$
$$v_i^\omega \ge (2q_{ii}^\omega \overline{y}_i + b_i^\omega)y_i^\omega + (c_i - q_{ii}^\omega \overline{y}_i^2)\delta_i^\omega$$

We can extend this formulation to scenario clusters obtaining the **MIQP**^{p} submodels for $p = 1, \ldots, \widehat{p}$ in this way

$$\begin{aligned}
\min \ & \sum_{\omega \in \Omega^p} p^\omega \left\{ \left(b_x^t x + x^t Q_{xx} x \right) + \left(\sum_{i=1}^{n} v_i^\omega \right) \right\} \\
\text{s.t.:} \ & v_i^\omega \geq (2q_{ii}^\omega \underline{y}_i + b_i^\omega) y_i^\omega + (c_i - q_{ii}^\omega \underline{y}_i^2)\delta_i^\omega, \ i \in \{1, \ldots, n\}, \ \omega \in \Omega^p \\
& v_i^\omega \geq (2q_{ii}^\omega \overline{y}_i + b_i^\omega) y_i^\omega + (c_i - q_{ii}^\omega \overline{y}_i^2)\delta_i^\omega, \ i \in \{1, \ldots, n\}, \ \omega \in \Omega^p \\
& l_a \leq A \begin{bmatrix} \delta^p \\ x^p \end{bmatrix} \leq u_a, \\
& l_t^\omega \leq T^\omega \begin{bmatrix} \delta^p \\ x^p \\ y^\omega \end{bmatrix} \leq u_t^\omega, \ \omega \in \Omega^p \\
& x^p \geq 0, \ y^\omega \in [\underline{y}, \overline{y}], \ \omega \in \Omega^p, \text{ and } \delta^p \in \{0,1\}^{n_\delta}
\end{aligned}$$

4 Implementation

These methods have been implemented in C++ with the help of Cplex 12.1 to solve only the quadratic subproblems QP^p in each node of the BF tree, for each $p \in \{1, \ldots, \widehat{p}\}$, and the QP^{TNF} and QP^f subproblems. These algorithmic alternatives have been considered:

- ▷ QBFC: coordination of δ in the TNF of the BF trees for clusters $p \in \{1, \ldots, \widehat{p}\}$ without using neither OLAs nor PCs, i.e. solving the quadratic subproblems QP^p.
- ▷ QBFC-PC: coordination of δ in the TNF of the BF trees for clusters $p \in \{1, \ldots, \widehat{p}\}$ using PCs.
- ▷ QBFC-OLA: coordination of δ in the TNF of the BF trees for clusters $p \in \{1, \ldots, \widehat{p}\}$ using OLAs.

For our instances the number of scenarios in each cluster is the same, $|\Omega^p| = |\Omega|/\widehat{p}$. Each cluster contains $|\Omega^p|$ consecutive scenarios, starting from the first one and following in natural order.

5 Numerical Tests

In order to obtain a computational comparison of the performance of the algorithmic alternatives QBFC, QBFC-PC, and QBFC-OLA some computational tests are carried out, which consist in solving two-stage stochastic problems, where the objective function is convex quadratic with linear constraints using QBFC code with those algorithmic choices. Therefore, these problems have a unique primal solution and the duality gap is zero. The tests have been performed on HP Compact with Intel Core 2 Quad Q9550 2.83GHz 4 CPU under Linux 2.6.38-8-generic-pae (x86_64).

The test problems have been randomly generated by using a C++ code developed by this author. This generator provides the scenarios set together with the associated probability of occurrence for two-stage stochastic mixed quadratic problems where $Q_{xy} = \mathbb{1}$, $Q_{xy} = \mathbf{0}$, and $Q_{yx} = \mathbf{0}$. Moreover, as can be seen in Table 1, in some problems $Q_{xx} = \mathbf{0}$ and in the rest of problems $Q_{xx} = \mathbb{1}$. Also, "# var" means the number of continuous variables, "# bin" the number of binary variables, "# constr" the number of constraints for DEM, see (1)-(5), and "dens" constraint matrix density %.

Table 1. Test problems

Prob.	$\lvert\Omega\rvert$	# var	# bin	# constr	Q_{xx}	dens.
P1	20	420	20	840	$\mathbf{0}$	38
P2	30	620	20	1240	$\mathbf{0}$	37
P3	40	820	20	1640	$\mathbf{0}$	42
P4	50	1020	20	2040	$\mathbf{0}$	43
P5	60	1220	20	2440	$\mathbf{0}$	39
P6	70	1420	20	2840	$\mathbf{0}$	37
P7	30	930	30	1860	$\mathbf{0}$	2
P8	40	1230	30	2460	$\mathbf{0}$	2
P9	50	1530	30	3060	$\mathbf{0}$	1
P10	60	1830	30	3660	$\mathbf{0}$	1
P11	70	2130	30	4260	$\mathbf{0}$	1
P12	100	3030	30	6060	$\mathbf{0}$	1
P13	30	930	30	1550	$\mathbb{1}$	13
P14	40	1230	30	2050	$\mathbb{1}$	10
P15	50	1530	30	2550	$\mathbb{1}$	10
P16	60	1830	30	3050	$\mathbb{1}$	9
P17	70	2130	30	3550	$\mathbb{1}$	9

Table 2 presents the main results of the computational experimentation for given values of the number of scenario clusters. Below the heading QBFC-1 are the times in CPU-seconds used for solving problems with 1 only scenario cluster (i.e. $\widehat{p} = 1$) and by solving the quadratic subproblem QP^p for each node using Cplex; the heading QBFC-5OLA indicates the CPU-times for 5 scenario cluster (i.e. $\widehat{p} = 5$) and by solving the quadratic subproblem QP^p for each node using outer linear approximations (OLA). Finally, PC means perspective cuts are used.

As can be observed in Table 2, the best efficiency is mainly obtained when the problems are solved with 5 clusters, except in the case of QBFC-5 because of the computational cost of solving a quadratic problem in each node of the BF tree for the different clusters. In addition, the outer linear approximations give us a higher efficiency than the perspective cuts.

6 Conclusions

An algorithm to solve two-stage stochastic quadratic problems based on the Twin Node Family concept involved in the Branch-and-Fix Coordination has

Table 2. Computational results: CPU-times

Prob.	QBFC-1	QBFC-1PC	QBFC-1OLA	QBFC-5	QBFC-5PC	QBFC-5OLA
P1	4.4	11.8	4.9	17.8	3.8	2.0
P2	9.7	14.2	6.7	6.6	8.0	3.0
P3	17.0	26.3	7.8	24.5	22.8	4.4
P4	21.4	43.3	16.3	14.5	18.6	6.3
P5	56.4	22.3	18.8	28.9	45.0	19.6
P6	33.9	96.0	28.7	10.6	20.0	9.7
P7	51.1	10.6	5.1	-	5.2	3.8
P8	5.2	9.3	3.9	-	5.0	2.3
P9	-	2.18	10.2	-	5.2	3.3
P10	135.4	28.5	14.2	-	18.8	10.5
P11	1052.7	39.1	14.5	-	17.2	3.7
P12	915.1	30.2	11.0	-	16.3	20.1
P13	10.5	7.1	7.8	-	5.3	6.8
P14	5.8	149.1	37.5	-	15.2	20.4
P15	94.4	69.5	58.4	-	33.2	16.1
P16	46.3	143.5	109.1	-	16.7	44.6
P17	1260.4	250.8	87.3	-	13.3	7.6

been implemented in C++ with the help of Cplex library to solve only the quadratic subproblems. When the problem's structure makes it possible, the algorithm uses perspective cuts or OLA to linearize the MIQ subproblems in each BF tree. The preliminary numerical results show a bit better efficiency with OLA than with PCF.

The path started from this work has the aim of solving nonlinear (non-quadratic) stochastic problems with nonlinear constraints and for two stage or more.

References

1. Alonso-Ayuso, A., Escudero, L.F., Ortuño, M.T.: BFC, a branch-and-fix coordination algorithm framework for solving some types of stochastic pure and mixed 0-1 programs. European Journal of Operational Research 151, 503–519 (2003)
2. Corchero, C., Mijangos, E., Heredia, F.J.: A new optimal electricity market bid model solved through perspective cuts. Top (2012), doi:10.1007/s11750-011-0240-6 (published online 2011)
3. Escudero, L.F., Garín, M., Merino, M., Pérez, G.: A general algorithm for solving two-stage stochastic mixed 0-1 first-stage problems. Computers & Operations Research 36, 2590–2600 (2009)
4. Escudero, L.F., Garín, M., Merino, M., Pérez, G.: An algorithmic framework for solving large-scale multistage stochastic mixed 0-1 problems with nonsymmetric scenario trees. Computers & Operations Research 39(5), 1133–1144 (2012)
5. Frangioni, A., Gentile, C.: Perspective cuts for a class of convex 0-1 mixed integer programs. Mathematical Programming 106, 225–236 (2006)

6. Heredia, F.J., Corchero, C., Mijangos, E.: Solving Electricity Market Quadratic Problems by Branch and Fix Coordination Methods. In: Hömberg, D., Tröltzsch, F. (eds.) CSMO 2011. IFIP AICT, vol. 391, pp. 516–525. Springer, Heidelberg (2013)
7. Laporte, G., Louveaux, F.: An integer L-shaped algorithm for the capacited vehicle routing problem with stochastic demands. Operations Research 50, 415–423 (2002)
8. Sen, S., Sherali, H.: Decomposition with branch-and-cut approaches for two-stage stochastic mixed-integer programming. Mathematical Programming Series A 105, 203–223 (2006)
9. Sherali, H.D., Fraticelli, B.M.P.: A modification of Bender's decomposition algorithm for discrete subproblems: an approach for stochastic programs with integer recourse. Journal of Global Optimization 22, 319–342 (2002)
10. Wets, R.J.-B.: Programming under uncertainty: the equivalent convex program. SIAM Journal on Applied Mathematics 14, 89–105 (1966)

Risk Minimizing Strategies for Tracking a Stochastic Target*

Andrzej Palczewski

Faculty of Mathematics, University of Warsaw, Banacha 2, 02-097 Warszawa, Poland
A.Palczewski@mimuw.edu.pl

Abstract. We consider a stochastic control problem of beating a stochastic benchmark. The problem is considered in an incomplete market setting with external economic factors. The investor preferences are modelled in terms of HARA-type utility functions and trading takes place in a finite time horizon. The objective of the investor is to minimize his expected loss from the outperformance of the benchmark compared to the portfolio terminal wealth, and to specify the optimal investment strategy. We prove that for considered loss functions the corresponding Bellman equation possesses a unique solution. This solution guaranties the existence of a well defined investment strategy. We prove also under which conditions the verification theorem for the obtained solution of the Bellman equation holds.

Keywords: optimal portfolios, stochastic target, benchmark tracking.

1 Introduction

We analyze the optimal portfolio and investment policy for an investor who is concerned about his wealth relative to the performance of a given benchmark. The benchmark evolves stochastically over time and the investor's objective is to minimize his loss with respect to this benchmark by investing in a portfolio of stochastically evolving financial instruments. Since the benchmark is not necessarily perfectly correlated with the investment opportunities, we are in the framework of an incomplete market, and there is no investment policy under which the investor can outperform the benchmark with certainty.

The portfolio problem where the objective is to exceed the performance of a selected target benchmark is sometimes called an active portfolio management. It is well known that many professional investors apply this benchmarking procedure. However, many small investors follow a benchmarking procedure as well, by trying to beat inflation, exchange rates, or other market indices.

The problem of an investment portfolio which outperforms a given benchmark has been studied for a long time. For objectives such as maximizing the probability that the investor's wealth achieves a certain performance goal relative to the benchmark, before falling below to a predetermined shortfall, or minimizing the expected time to reach the performance goal, the problem is studied by Browne [4], [5]. For the special case

* Financial support from MNiSzW grant no. NN-201-547838 is gratefully acknowledged.

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 188–196, 2013.

where the benchmark is perfectly correlated with the investment opportunities, these problems over a finite-horizon are analyzed in [4], and for a more general model that the benchmark is not perfectly correlated with the investment opportunity in [5]. The problem of finding the minimal initial data of a controlled process which guarantees to reach the benchmark with a given probability of success or, more generally, with a given level of expected loss was first introduced by Föllmer and Leukert [7] in the context of quantile hedging. This approach has been then extended to the stochastic target problem studied by Soner and Touzi [9, 10], and in a number of papers by Bouchard *et. al.* [1, 2, 3].

In opposition to the majority of previously mentioned papers, in this paper, we study a loss minimization objective when the prices of financial instruments are functions of external economic factors. A similar problem but without taking into account economic factors is solved by Browne [5]. The absence of economic factors makes the problem much simpler as the HJB equation is reduced in that case to an ODE. In the presence of external factors the HJB equation becomes a multidimensional nonlinear PDE for which the existence of solutions is a challenging problem. We solve this problem using the well developed theory of quasilinear parabolic equations. We also show that under suitable regularity assumptions the verification theorem holds. Hence, the obtained solution to the HJB equation is a solution to the optimization problem. The plan of the paper is as follows. In Section 2, we present the portfolio problem arising from the active portfolio management. In Section 3, we show that, under additional assumptions on the loss function, we can find a smooth solution to the HJB equation and construct effectively an optimal investment strategy. Section 4 is devoted to the formulation and proof of the verification theorem.

2 The Portfolio Problem

We consider the portfolio problem in which the prices of securities are functions of external state variables (economic factors). Our goal is to construct a portfolio which can outperform a stochastic benchmark. We consider a general setting of the problem. In particular, the risk factors which define the dynamics of the benchmark can be different from the risk factors in the dynamics of securities. Hence, the problem is an incomplete market problem.

The setting of the market is as follows: we have a market defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with the filtration $(\mathcal{F}_t)_{t\in[0,T]}$ generated by d-dimensional standard Wiener process $W(t) = (W_1, \ldots, W_d)$ (in what follows we treat W as a column vector). On that probability space we have N stochastic processes describing the prices of securities with the dynamics

$$\frac{dS_i(t)}{S_i(t)} = \mu_i(t, R)dt + \sum_{j=1}^{d} \sigma_{ij}(t, R)dW_j(t), \quad i = 1, 2, \ldots, N, \tag{1}$$

where μ_i and σ_{ij} depend on an M-dimensional vector of economic factors R.

We assume that the dynamics of factors R follow the Markovian diffusion process

$$dR_m(t) = \mu^r_m(t,R)dt + \sum_{i=1}^{d} b_{mi}(t,R)dW_i(t), \quad m = 1,2,\dots,M. \tag{2}$$

It is convenient to switch to vector notation and introduce the matrices $\sigma = (\sigma_{ij})$, $B = (b_{ij})$ and the column vectors $\mu = (\mu_1,\dots,\mu_N)'$, $\mu^r = (\mu^r_1,\dots,\mu^r_M)'$. (Here and in what follows x' denotes the transpose of the matrix or vector x.)

Assumption 2.1. *For the model of security prices we assume that coefficients $\mu(t,r)$ and $\sigma(t,r)$ are deterministic functions bounded and continuous for $t \in [0,T]$ and $r \in \mathbb{R}^M$. For the model of economic factors we make typical assumptions which guarantee the existence of strong solutions to equation (2), i.e., we assume that $\mu^r(t,r)$ and $B(t,r)$ are deterministic continuous functions of their arguments, which in addition fulfil the estimates*

$$\|\mu^r(t,r_1) - \mu^r(t,r_2)\| + \|B(t,r_1) - B(t,r_2)\| \le c\|r_1 - r_2\|, \tag{3}$$

$$\|\mu^r(t,r)\|^2 + \|B(t,r)\|^2 \le c^2(1 + \|r\|^2), \tag{4}$$

for $t \in [0,T]$, $r, r_1, r_2 \in \mathbb{R}^M$, where c is a positive constant.

We analyze the stochastic target problem in an incomplete market assuming that the dimension of risk factors is high, i.e. dimmension d of the Wiener process W is high, and the number of securities and economic factors much lower. This means that $d \gg N$ and $d \gg M$. To guarantee well-posedness and solvability of the optimization problem, we have to make additional assumptions.

Assumption 2.2. *About the model of securities dynamics we assume a "partial invertibility" of the model, i.e., the matrix $\Sigma = \sigma\sigma'$ is nonsingular and $\Sigma^{-1}(t,r)$ is bounded for $t \in [0,T]$ and $r \in \mathbb{R}^M$. This in fact means that securities are driven by N risk factors and limited to these N dimensions the security market is complete.*
About the model of dynamics of economic factors we assume that the matrix BB' is positive definite. Strictly speaking, we postulate that there exist positive constants ν_1, ν_2 such that for any $x \in \mathbb{R}^M$

$$0 < \nu_1\|x\|^2 \le x'BB'x \le \nu_2\|x\|^2.$$

The stochastic benchmark is modelled as a general log-normal stochastic process $H(t)$ which fulfils the equation

$$dH(t) := H(t)\left(\mu^H dt + \xi dW(t)\right),$$

where $\xi = (\xi_1,\dots,\xi_d)'$ is a column vector, and coefficients μ^H and ξ are deterministic functions of $t \in [0,T]$ and $r \in \mathbb{R}^M$.

We consider now a portfolio $V(t)$ consisting of assets $S_i(t)$, $i = 1,\dots,N$. Denoting by π_i the fraction of the total wealth $V(t)$ invested in the security S_i, we can write

$$dV^\pi(t) = V^\pi(t)\left(\mu^V dt + \theta^V dW(t)\right), \tag{5}$$

where after introducing the column vector $\pi = (\pi_1, \ldots, \pi_N)'$ we have $\mu^V = \mu'\pi$ and $\theta^V = \sigma'\pi$.

We define now the new process

$$X(t) = \frac{H(t)}{V^\pi(t)}. \tag{6}$$

This approach enables us to consider in the same framework losses and gains. Such an approach is not quite new in the financial literature. It is applied by Browne [4, 5]. A similar quotient is used by Dai Pra, Runggaldier and Tolotti [6] who optimize the quadratic loss in the benchmark tracking problem.

For the process $X(t)$ we obtain the following equation of evolution

$$\frac{dX}{X} = \mu^X dt + \theta dW, \tag{7}$$

where

$$\theta = \xi - \theta^V, \quad \mu^X = \mu^H - \mu^V - \theta'\theta^V.$$

We optimize the process $X(t)$ with respect to the vector of strategies π. Admissible strategies for our problem are defined as follows.

Definition 2.1. *Let U be a complete, separable metric space and $0 < T < \infty$. We define the set of admissible strategies $\Pi(t,x,r)$ as fulfilling the conditions:*

1. *$\pi : [t,T] \times \Omega \to U \subseteq \mathbb{R}^N$ is measurable, bounded and $\{\mathcal{F}_\tau\}_{\tau \geq t}$-adapted, for each $\pi \in \Pi(t,x,r)$,*
2. *$X(t) = x$ (budget constraint),*
3. *$R(t) = r$.*

We consider the optimization problem in the framework of utility theory. This means that we fix a utility function g and optimize the terminal value of process X measured by g. The optimization problem is of the form

$$\min_{\pi \in \Pi(t,x,r)} \mathbb{E}\big[g(X(T))|X(t) = x, R(t) = r\big]. \tag{8}$$

In fact, it is better to call g a loss function as our goal is to minimize losses and not to maximize gains.

Under Assumptions 2.1, 2.2 and Definition 2.1, equation (7) admits the unique solution and the value function

$$u(t,x,r) := \min_{\pi \in \Pi(t,x,r)} \mathbb{E}\big[g(X(T))|X(t) = x, R(t) = r\big] \tag{9}$$

is well defined.

With this value function we arrive at the following Hamilton-Jacobi-Bellman equation

$$\begin{aligned} &\partial_t u + \inf_{\pi \in \Pi} \Big(\mu^X x \partial_x u + (\mu^r)' \nabla_r u + \frac{1}{2}\theta'\theta x^2 \partial_{xx} u + \\ &\qquad + \frac{1}{2} BB' (\nabla_r \otimes \nabla_r u) + x\theta' B' \nabla_r \partial_x u \Big) = 0, \\ &u(T,x,r) = g(x), \end{aligned} \tag{10}$$

where ∇_r denotes the gradient operator with respect to vector variable r (a column vector), ∂_t and ∂_x denote differential operator with respect to scalar variables t and x, respectively. $\nabla_r \otimes \nabla_r$ has the following meaning: when x and y are n- dimensional column vectors then $x \otimes y$ denotes the $n \times n$ matrix xy', i.e. $\nabla_r \otimes \nabla_r u$ is a matrix of all second order derivatives of u with respect to variables r_i and r_j, $i, j = 1, \ldots, M$.

3 Smooth Solutions of the HJB Equation

To obtain smooth solutions to the HJB equation (10), we make the following assumption.

Assumption 3.1. *The loss function $g(x)$ is from the generalized HARA class and is given by the expression $g(x) = cx^\alpha$, for $\alpha > 1$ and $x \in [0, \infty)$.*

Remark 3.1. *In fact, from the technical point of view, we can assume only that g is such that $\alpha \neq -1$. The assumption $\alpha > 1$ is essential when we want to interpret $g(x)$ as a loss function.*

Under the above assumption, we postulate that the value function can be factorized in the form

$$u(t, x, r) = g(x)q(t, r). \tag{11}$$

Substituting the above factorization into equation (10), we obtain the following PDE problem for q:

$$\begin{aligned} &\partial_t q + \inf_{\pi \in \Pi} \Big(\alpha \mu^X q + (\mu^r)' \nabla_r q + \frac{\alpha(\alpha-1)}{2} \theta' \theta q + \\ &\qquad + \frac{1}{2} BB' (\nabla_r \otimes \nabla_r q) + \alpha \theta' B' \nabla_r q \Big) = 0, \\ &q(T, r) = 1. \end{aligned} \tag{12}$$

From equation (12) we can derive formally the optimal investment strategy

$$\pi^* = \pi^0 + \frac{\pi^1 \nabla_r q}{q}, \tag{13}$$

where

$$\begin{aligned} \pi^0 &= \frac{1}{1+\alpha} \Sigma^{-1} (\mu + \alpha \sigma \xi), \\ \pi^1 &= \frac{1}{1+\alpha} \Sigma^{-1} \sigma B'. \end{aligned}$$

Substituting expression (13) into the HJB equation (12) we obtain

$$\begin{aligned} &\partial_t q + \alpha \mu^* q + (\mu^r)' \nabla_r q + \frac{\alpha(\alpha-1)}{2} (\theta^*)' \theta^* q + \\ &\quad + \frac{1}{2} BB' (\nabla_r \otimes \nabla_r q) + \alpha (\theta^*)' B' \nabla_r q = 0. \end{aligned} \tag{14}$$

In this equation, μ^* denotes the value of μ^X, and θ^* the value of θ evaluated at the point of the optimal strategy π^*.

After rearrangements, the HJB equation (14) takes the form

$$\partial_t q + A_0(\nabla_r \otimes \nabla_r q) + A_1 \frac{\nabla_r q \otimes \nabla_r q}{q} + A_2 \nabla_r q + A_3 q = 0, \tag{15}$$

where

$$\begin{aligned}
A_0 =& \frac{1}{2} BB', \\
A_1 =& \frac{1}{2}\alpha(\alpha+1)(\pi^1)'\Sigma\pi^1 - \alpha B\sigma'\pi^1, \\
A_2 =& \mu^r + \alpha(\alpha+1)(\pi^1)'\Sigma\pi^0 - \alpha^2(\pi^1)'\sigma\xi + \alpha\big(B\xi - (\pi^1)'\mu - B\sigma'\pi^0\big), \\
A_3 =& \frac{1}{2}\alpha^2\big((\pi^0)'\Sigma\pi^0 - 2(\pi^0)'\sigma\xi + \xi'\xi\big) + \alpha\big(\mu^H - \mu'\pi^0 - \frac{1}{2}\xi'\xi\big).
\end{aligned}$$

Equation (15) has to be solved in the strip $0 \le t \le T$ with the terminal condition

$$q(T, r) = 1. \tag{16}$$

Equation (15) is a quasilinear parabolic equation which possesses a solution provided this solution is bounded away from zero. To find this solution we make the substitution

$$z = \ln q.$$

For the new function z we obtain the equation

$$\partial_t z + A_0(\nabla_r \otimes \nabla_r z) + (A_0 + A_1)\nabla_r z \otimes \nabla_r z + A_2 \nabla_r z + A_3 = 0, \tag{17}$$

with the terminal condition

$$z(T, r) = 0. \tag{18}$$

To solve equation (17) with condition (18), we use well known results in the theory of quasilinear parabolic equations.

Let us consider a boundary value problem for a n-dimensional quasilinear parabolic equation

$$\begin{aligned}
&\partial_t w - \sum_{i,j=1}^{n} a_{ij}(t,x)\partial_{x_i x_j} w + a(t, x, w, \partial_x w) = 0, \quad \text{for } (t,x) \in \mathcal{O}_T, \\
&w(t,x) = \psi(t,x), \quad \text{for } (t,x) \in \Gamma_T,
\end{aligned} \tag{19}$$

in a bounded domain $\mathcal{O}_T = [0,T] \times \mathcal{O}$, where $\mathcal{O}$ is a bounded domain in $\mathbb{R}^n$ with the boundary of class $H^{2+\beta}$, and $\Gamma_T = \partial\mathcal{O} \times [0,T] \cup \mathcal{O} \times \{t = 0\}$.

Theorem 7.4 in Chapter 6 of the book by Ladyzhenskaya, Solonnikov and Uraltseva [8] guarantees that, under suitable assumptions on coefficients a_{ij}, a and boundary function ψ, there exists a unique solution to the boundary value problem (19) in $H^{1+\beta/2,2+\beta}(\mathcal{O}_T)$.

To apply the above result to equation (17), we have to make additional assumptions.

Assumption 3.2. *Let functions μ, μ^r, μ^H, σ, B, ξ and Σ^{-1} be Hölder continuous functions of t with the Hölder exponent $\beta/2$, and Hölder continuous functions of r with the Hölder exponent β , for some $\beta > 0$.*

Now we can prove our main theorem.

Theorem 3.1. *Under Assumptions 2.1, 2.2, 3.1, 3.2 and the assumptions of Definition 2.1, there exists a unique solution $z(t,r)$ to the terminal problem (17)-(18) and $|z|$, $|\partial_t z|$, $|\partial_{x_i} z|$, $|\partial_{x_i x_j} z|$ are bounded in $[0,T] \times \mathbb{R}^M$. This solution belongs to $H^{1+\beta/2,2+\beta}(\mathcal{O}_T)$, where $\mathcal{O}_T = [0,T] \times \mathcal{O}$ and $\mathcal{O}$ is a bounded domain in $\mathbb{R}^M$.*

Proof: Let us consider the terminal problem (17)-(18) in $\mathcal{O}_T = [0,T] \times \mathcal{O}$, where $\mathcal{O}$ is a fixed bounded domain in $\mathbb{R}^M$. To solve this problem, we supplement equation (17) and terminal condition (18) with the boundary condition

$$z(t,r) = 0 \quad \text{for } (t,r) \in \partial\mathcal{O} \times [0,T]. \tag{20}$$

The above defined augmented problem fulfils already the assumptions of Theorem 7.4 in Chapter 6 of [8]. Due to this theorem, there exists a unique solution of equation (17) with terminal condition (18) and boundary condition (20). This solution together with its derivatives can be estimated in $\mathcal{O}_T$ with constants which depend only on constants present in the Assumptions, and not on the size of domain $\mathcal{O}$. Hence, the solution which exists in any bounded domain $\mathcal{O}_T$ belongs to $H^{1+\beta/2,2+\beta}(\mathcal{O}_T)$ and is uniformly bounded together with its derivatives independently of the size of the domain. Then we can consider a increased sequence of bounded smooth domains $\mathcal{O}^n$ that fill in the whole $\mathbb{R}^M$ and solutions z^n to problem (17), (18), (20) with $\mathcal{O}$ replaced by $\mathcal{O}^n$. By the standard Arzela-Ascoli theorem, we can choose a subsequence of z^n which converges to a function which is a solution to (17)-(18) on $[0,T] \times \mathbb{R}^M$. □

Remark 3.2. *In many situations, economic factors should be restricted to nonnegative values only. In that cases, Theorem 3.1 is still applicable as we can construct a sequence of bounded smooth domains $\mathcal{O}^n$ approximating the space $\mathbb{R}^M_+$.*

Corollary 3.1. *Let us observe that due to Theorem 3.1 fuction $z(t,r)$ and its derivatives are bounded. Returning back to the original function q, we conclude that $q(t,r)$ is bounded away from zero and the quotient q_{r_m}/q is bounded. It follows than that the optimal investment strategy given by expression (13) is bounded and admissible in accordance with Definition 2.1.*

4 Verification Theorem

Theorem 3.1 guaranties a smooth solution to the terminal problem (17)-(18). This solution is not necessarily a solution to the optimization problem. To prove the optimality, we need some additional results. First, we have to show that the solution to problem (17)-(18) is a function of class $C^{1,2}$ on $[0,T] \times \mathbb{R}^M$.

To this end, we can use Theorem 8.1 from Chapter 6 of [8] which guarantees the existence of a unique solution in $H^{1+\beta/2,2+\beta}(Q_T)$, where $Q_T = [0,T] \times \mathbb{R}^n$, to the Cauchy problem for a quasilinear parabolic equation

$$\begin{aligned} &\partial_t w - \sum_{i,j=1}^{n} a_{ij}(t,x)\partial_{x_i x_j} w + a(t,x,w,\partial_x w) = 0, \text{ in } (0,T]\times\mathbb{R}^n, \\ &w(0,x) = \psi(x), \text{ in } \mathbb{R}^n, \end{aligned} \tag{21}$$

if in addition to the assumptions of Theorem 7.4 from [8] the coefficients of the equation can be uniformly estimated on every bounded set with the bound independent of the size of this set.

To apply the above mentioned theorem, we make the following assumption

Assumption 4.1. *Let functions μ, μ^r, μ^H, σ, B, ξ and Σ^{-1}, in addition to being Hölder continuous, be uniformly bounded for all (t,r), $t\in[0,T]$, $r\in\mathbb{R}^M$.*

Theorem 4.1. *Under the assumptions of Theorem 3.1 and Assumption 4.1, there exists a unique solution to the terminal problem (17)-(18). This solution belongs to $H^{1+\beta,2+\beta/2}([0,T]\times\mathbb{R}^M)$ and the estimates of Theorem 3.1 hold for $(t,r)\in[0,T]\times\mathbb{R}^M$. In particular, the solution is a $C^{1,2}$ function on $[0,T]\times\mathbb{R}^M$.*

Proof: The proof follows straightforwardly from Theorem 8.1 in Chapter 6 of [8]. That theorem states that the solution to problem (21) is unique and belongs to $H^{1+\beta/2,2+\beta}([0,T]\times\mathbb{R}^M)$ for some $\beta>0$. It is obvious that such a solution is a function of class $C^{1,2}$. □

To use the classical stochastic verification theorem (cf. Theorem 5.1 in Chapter 5 of the book by Yong and Zhou [11]), we have to prove the following simple lemma.

Lemma 4.1. *Let $z(t,r)$ be the unique solution to the boundary value problem (17)-(18), which exists due to Theorem 4.1 in $[0,T]\times\mathbb{R}^M$. Let $q(t,r)=\exp\big(z(t,r)\big)$ and π^* be given by equation (13). Then*

$$\begin{aligned} &\inf_{\pi\in\Pi}\left(\alpha\mu^X q + (\mu^r)'\nabla_r q + \frac{\alpha(\alpha-1)}{2}\theta'\theta q + \frac{1}{2}BB'(\nabla_r\otimes\nabla_r q) + \alpha\theta' B'\nabla_r q\right) = \\ &= \alpha\mu^* q + (\mu^r)'\nabla_r q + \frac{\alpha(\alpha-1)}{2}(\theta^*)'\theta^* q + \frac{1}{2}BB'(\nabla_r\otimes\nabla_r q) + \alpha(\theta^*)' B'\nabla_r q, \end{aligned}$$

where μ^* denotes the value of μ^X, and θ^* the value of θ evaluated at the point of the optimal strategy π^*.

Proof: The proof is straightforward as the left and right hand sides of the equation in the Lemma are left hand sides of equations (12) and (14), respectively. But equation (14) has been obtained from equation (12) upon substitution (13). The fact that π^* is an admissible investment strategy has been already stated in Corollary 3.1. □

From Theorem 4.1 and Lemma 4.1, we easily obtain the verification theorem.

Theorem 4.2. *Under assumptions of Theorem 4.1, the function*

$$u(t,x,r) = g(x)q(t,r)$$

is the value function to the optimization problem (9), where $g(x)$ fulfils the conditions of Assumption 3.1 and $q(t,r)=\exp\big(z(t,r)\big)$ with $z(t,r)$ being the solution of the boundary value problem (17)-(18), which exists due to Theorem 4.1.

5 Conclusions

In this paper, we have solved the stochastic optimization problem for the loss minimization with the state variable being the ratio of a stochastic benchmark to an investment portfolio. The control parameter of this problem is the portfolio investment strategy. The problem is solved in a market model of N securities being log-normal stochastic processes and depending on M external economic factors. The stochastic benchmark is also a log-normal process but the set of risk factors on which this benchmark depends can be larger than the set of risk factors of the securities making the whole problem an incomplete market problem.

The stochastic optimization problem has been reduced to the HJB equation which, in this case, is a multidimensional quasilinear parabolic equation. Using the general theory of such equations, we have proved that under suitable regularity conditions the HJB equation possesses a unique solution which is sufficiently smooth to guarantee the fulfilment of the stochastic verification theorem. Hence, the solution to the HJB equation is a unique solution to the initial optimization problem. This is a natural extention of similar results obtained in a less general setting without the dependence of security prices on external economic factors.

A natural question which arises is the extension of the obtained results to a market with less restrictive assumpions. The most severe of these assumptions is the boundedness of the coefficients in the whole domain and the lifting of these restrictions will be the subject of future research.

References

[1] Bouchard, B., Elie, R., Touzi, N.: Stochastic target problems with controlled loss. SIAM J. Control Optim. 48, 3123–3150 (2009)
[2] Bouchard, B., Elie, R., Imbert, C.: Optimal control under stochastic target constraints. SIAM J. Control Optim. 48, 3501–3531 (2009)
[3] Bouchard, B., Vu, T.N.: A stochastic target approach for P&L matching problems, Preprint Ceremade, University Paris-Dauphine (2011)
[4] Browne, T.: Reaching Goals by a Deadline: Digital options and Continuous-Time Active Portfolio Management. Adv. Appl. Probab. 31, 551–577 (1999)
[5] Browne, T.: – Beating a moving target: optimal portfolio strategies for outperforming a stochastic benchmark. Fin. Stoch. 3, 275–294 (1999)
[6] Dai Pra, P., Runggaldier, W.J., Tolotti, M.: Pathwise optimality for benchmark tracking. IEEE Trans. Automat. Control. 49, 386–395 (2004)
[7] Föllmer, H., Leukert, P.: Quantile hedging. Fin. Stoch. 3, 251–273 (1999)
[8] Ladyzhenskaya, O.A., Solonnikov, V.A., Uraltseva, N.N.: Linear and Quasilinear Equations of Parabolic Type. Amer. Math. Soc. (1968)
[9] Soner, H.M., Touzi, N.: Stochastic target problems, dynamic programming and viscosity solutions. SIAM J. Control Optim. 41, 404–424 (2002)
[10] Soner, H.M., Touzi, N.: Dynamic programming for stochastic target problems and geometric flows. J. Europ. Math. Soc. 4, 201–236 (2002)
[11] Yong, J., Zhou, X.Y.: Stochastic Controls. Hamiltonian Systems and HJB Equations. Springer (1999)

Harvesting in Stochastic Environments: Optimal Policies in a Relaxed Model

Richard H. Stockbridge and Chao Zhu

Department of Mathematical Sciences
University of Wisconsin-Milwaukee
Milwaukee, WI 53201
{stockbri,zhu}@uwm.edu

Abstract. This paper examines the objective of optimally harvesting a single species in a stochastic environment. This problem has previously been analyzed in [1] using dynamic programming techniques and, due to the natural payoff structure of the price rate function (the price decreases as the population increases), no optimal harvesting policy exists. This paper establishes a relaxed formulation of the harvesting model in such a manner that existence of an optimal relaxed harvesting policy can not only be proven but also identified. The analysis imbeds the harvesting problem in an infinite-dimensional linear program over a space of occupation measures in which the initial position enters as a parameter and then analyzes an auxiliary problem having fewer constraints. In this manner upper bounds are determined for the optimal value (with the given initial position); these bounds depend on the relation of the initial population size to a specific target size. The more interesting case occurs when the initial population exceeds this target size; a new argument is required to obtain a sharp upper bound. Though the initial population size only enters as a parameter, the value is determined in a closed-form functional expression of this parameter.

Keywords: Singular stochastic control, linear programming, relaxed control.

AMS subject classification: 93E20, 60J60.

1 Introduction

This paper examines the problem of optimally harvesting a single species that lives in a random environment. Let X be the process denoting the size of the population and Z denote the cumulative amount of the species harvested. We assume $X(0-) = x_0 > 0$, $Z(0-) = 0$, and X and Z satisfy

$$dX(t) = b(X(t))dt + \sigma(X(t))dW(t) - dZ(t), \tag{1}$$

in which $W(\cdot)$ is a 1-dimensional standard Brownian motion that provides the random fluctuations in the population's size, and b and σ are real-valued continuous functions. We assume that b and σ are such that in the absence of harvesting

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 197–206, 2013.

the population process X takes values in $\mathbb{R}_+$ and that ∞ is a natural boundary so that the population will not explode to ∞ in finite time. The boundary 0 may be an exit or a natural boundary point but may not be an entrance point; this indicates that the species will not spontaneously reappear following extinction. Note that $X(0)$ may not equal $X(0-)$ due to an instantaneous harvest $Z(0)$ at time 0 and the process Z is restricted so that $\Delta Z(t) := Z(t) - Z(t-) \leq X(t-)$ for all $t \geq 0$. This latter condition indicates that one cannot harvest more of the species than exists. Let $r > 0$ denote the discount rate and f denote the marginal yield for harvesting. The objective is to select a harvesting strategy Z so as to maximize the expected discounted revenue

$$J(x_0, Z) := \mathbf{E}_{x_0}\left[\int_0^\tau e^{-rs} f(X(s-))dZ(s)\right], \tag{2}$$

where $\tau = \inf\{t \geq 0 : X(t) = 0\}$ denotes the extinction time of the species.

As a result of developments in stochastic analysis and stochastic control techniques, there has been a resurgent interest in determining the optimal harvesting strategies in the presence of stochastic fluctuations (see, e.g., [1,6]). In particular, [1] examines the current problem using dynamic programming techniques and determines the value function. The paper indicates the lack of an optimal policy in the admissible class of (strict) harvesting policies by commenting that a "chattering" policy will be optimal. The problem of optimal harvesting of a single species in a random environment is also studied in [8] in which the model is extended to regime-switching diffusions so as to capture different dynamics such as for drought and non-drought conditions. The paper also adopts a dynamic programming solution approach to determine the value function while at the same time exhibiting ϵ-optimal harvesting policies since, as in the static environment of [1], no optimal harvesting policy exists. In light of the complexities of the regime-switching model, it further identifies a condition under which the value function is shown to be continuous and a viscosity solution to the variational inequality.

The focus of this paper is on developing a relaxed formulation for the harvesting problem under which an optimal harvesting control exists and on establishing optimality using a linear programming formulation instead of dynamic programming. In addition, it is sufficient to have a weak solution to (1) rather than placing Lipschitz and polynomial growth conditions on the coefficients b and σ that guarantee existence of a strong solution. Intuitively, relaxation completes the space of admissible harvesting rules by allowing measure-valued policies. A benefit of the linear programming solution methodology is the analysis concentrates on the optimal value for a single, fixed initial condition, rather than seeking the value *function* and thus no smoothness properties need to be established about the value as a function of the initial position.

To set the stage for the relaxed singular control formulation of the model, let $\mathcal{D} = C_c^2(\mathbb{R}_+)$ and for a function $g \in \mathcal{D}$, define the operators A and B by

$$Ag(x) = \frac{1}{2}\sigma^2(x)g''(x) + b(x)g'(x), \text{ and} \tag{3}$$

$$Bg(x,z) = \begin{cases} \frac{g(x-z)-g(x)}{z}, & \text{if } z > 0, \\ -g'(x), & \text{if } z = 0, \end{cases} \tag{4}$$

where $x, z \in \mathbb{R}_+$. Itô's formula then implies

$$g(X(t)) = g(x_0) + \int_0^t Ag(X(s))\,ds + \int_0^t Bg(X(s), \Delta Z(s))\,dZ(s) + \int_0^t \sigma(X(s))g'(X(s))\,dW(s), \quad \forall g \in \mathcal{D}.$$

It therefore follows that for any $g \in \mathcal{D}$

$$g(X(t)) - g(x_0) - \int_0^t Ag(X(s))\,ds - \int_0^t Bg(X(s), \Delta Z(s))\,dZ(s) \tag{5}$$

is a mean 0 martingale. In fact, requiring (5) to be a martingale for a sufficiently large collection of functions g is a way to characterize the processes (X, Z) which satisfy (1). We turn now to a precise formulation of the model in which the processes are relaxed solutions of a controlled martingale problem for the operators (A, B).

1.1 Formulation of the Relaxed Model

For a complete and separable metric space S, we define $M(S)$ to be the space of Borel measurable functions on S, $B(S)$ to be the space of bounded, measurable functions on S, $C(S)$ to be the space of continuous functions on S, $\overline{C}(S)$ to be the space of bounded, continuous functions on S, $\mathcal{M}(S)$ to be the space of finite Borel measures on S, and $\mathcal{P}(S)$ to be the space of probability measures on S. $\mathcal{M}(S)$ and $\mathcal{P}(S)$ are topologized by weak convergence.

Recall, the amount of harvesting is limited by the size of the population. Define $\mathcal{R} = \{(x,z) : 0 \le z \le x, x \ge 0\}$; $\mathcal{R}$ denotes the space on which the paired process (X, Z) evolves when considering solutions of (1).

The formulation of the population model in the presence of "relaxed" harvesting policies adapts the relaxed formulation for singular controls given in [5] to the particulars of the harvesting problem. This adaptation sets the state space E to be $\mathbb{R}_+$ and the control space $U = \mathbb{R}_+$, with $\mathcal{U} = \mathcal{R} \subset \mathbb{R}_+ \times \mathbb{R}_+$.

Let X be an $\mathbb{R}_+$-valued process and Γ be an $\mathcal{L}(\mathcal{R})$-valued random variable. Let Γ_t denote the restriction of Γ to $\mathcal{R} \times [0,t]$. Then (X, Γ) is a *relaxed solution* of the harvesting model if there exists a filtration $\{\mathcal{F}_t\}$ such that (X, Γ_t) is $\{\mathcal{F}_t\}$-progressively measurable, $X(0-) = x_0$, and for every $g \in \mathcal{D}$,

$$g(X(t)) - g(x_0) - \int_0^t Ag(X(s))\,ds - \int_{\mathcal{R}\times[0,t]} Bg(x,z)\,\Gamma(dx \times dz \times ds) \tag{6}$$

is an $\{\mathcal{F}_t\}$-martingale, in which the operators A and B are given by (3) and (4), respectively. Throughout the paper we assume that a relaxed solution (X, Γ) exists and is strong Markov. Let $\mathcal{A}$ denote the set of measures Γ for which there is some X such that (X, Γ) is a relaxed solution of the harvesting model.

We turn now to the extension of the reward criterion (2) to the relaxed framework. Specifically, $f : \mathbb{R}_+ \mapsto \mathbb{R}_+$ represents the instantaneous marginal yield accrued from harvesting. Assume f is continuous and non-increasing with respect to x. Thus $f(x) \geq f(y)$ whenever $x \leq y$; this assumption indicates that the price when the species is plentiful is smaller than when it is rare. Moreover, we assume $0 < f(0) < \infty$. Let (X, Γ) be a solution to the harvesting model (6). Let $S = (0, \infty)$ denote the survival set of the species and $\tau = \inf\{t \geq 0 : X(t) \notin S\}$. Then the expected total discounted value from harvesting is

$$J(x_0, \Gamma) := \mathbf{E}\left[\int_{\mathcal{R}\times[0,\tau]} e^{-rs} f(x)\, \Gamma(dx \times dz \times ds)\right]. \tag{7}$$

The goal is to maximize the expected total discounted value from harvesting over relaxed solutions (X, Γ) of the harvesting model and to find an optimal harvesting strategy Γ^*. Thus, we seek

$$V(x_0) = J(x_0, \Gamma^*) := \sup_{\Gamma\in\mathcal{A}} J(x_0, \Gamma). \tag{8}$$

We emphasize that the initial position x_0 is merely a parameter in the problem and that V is not to be viewed as a function with any particular properties but merely is the value of the harvesting problem when the initial population size is x_0. We do, however, obtain the value in functional form for x_0 in two regions.

2 Linear Programming Formulation and Main Result

Throughout this paper, we assume the equation $(A - r)u(x) = 0$ has two fundamental solutions ψ and ϕ, where ψ is strictly increasing and ϕ is strictly decreasing; without loss of generality we may assume $\psi(0) = 0$ (see [1]).

The main result of this paper is summarized in the following theorem.

Theorem 1. *Assume that there exists some $\tilde{b} \geq 0$ such that*

(i) $\frac{f(x)}{\psi'(x)} \leq \frac{f(\tilde{b})}{\psi'(\tilde{b})}, \quad \forall x \geq 0,$
(ii) *the function f/ψ' is nonincreasing on $[\tilde{b}, \infty)$, and*
(iii) *the function f is continuously differentiable on $(\tilde{b}, \infty)$.*

Put $b^* = \inf\{\tilde{b} \geq 0 : \tilde{b}$ satisfies (i)–(iii)$\}$. *Then the value is given by*

$$V(x_0) = \frac{f(b^*)}{\psi'(b^*)}\psi(x_0 \wedge b^*) + \int_{b^*}^{x_0\vee b^*} f(y)dy \tag{9}$$

and an optimal relaxed harvesting policy is given by

$$\Gamma^*(dx \times dz \times dt) = I_{(b^*,\infty)}(x_0)\lambda_{[b^*,\,x_0]}(dx)\delta_{\{0\}}(dz)\delta_{\{0\}}(dt) + \Gamma_{b^*}(dx \times dz \times dt), \tag{10}$$

where $\lambda_{[b^*,\,x_0]}(\cdot)$ *denotes Lebesgue measure on* $[b^*, x_0]$ *and* Γ_{b^*} *is defined in Proposition 6.*

We begin the task of reformulating the harvesting problem with the following observation. Let $\tilde{\tau}$ be any $\{\mathcal{F}_t\}$-stopping time. The optional sampling theorem along with the requirement that (6) be a mean 0 martingale for each $g \in \mathcal{D}$ implies

$$\begin{aligned} e^{-r(t\wedge\tilde{\tau})}g(X(t\wedge\tilde{\tau})) - g(x_0) - \int_0^{t\wedge\tilde{\tau}} e^{-rs}[A-r]g(X(s))\,ds \\ - \int_{\mathcal{R}\times[0,t\wedge\tilde{\tau}]} e^{-rs}Bg(x,z)\,\Gamma(dx\times dz\times ds) \end{aligned}$$

is also a martingale. Recall $g \in \mathcal{D}$ means g has compact support and hence is bounded. So taking expectations and letting $t \to \infty$ yields

$$\begin{aligned} g(x_0) = \mathbf{E}\left[e^{-r\tilde{\tau}}I_{\{\tilde{\tau}<\infty\}}g(X(\tilde{\tau}))\right] - \mathbf{E}\left[\int_0^{\tilde{\tau}} e^{-rs}[A-r]g(X(s))\,ds\right] \\ - \mathbf{E}\left[\int_{\mathcal{R}\times[0,\tilde{\tau}]} e^{-rs}Bg(x,z)\,\Gamma(dx\times dz\times ds)\right]. \end{aligned} \tag{11}$$

The initial analysis takes $\tilde{\tau} = \tau$; later we will need (11) for a different stopping time.

The measures involved in the infinite-dimensional linear program are expected discounted occupation measures corresponding to relaxed solutions (X, Γ) of the harvesting model. Indeed, for any Borel measurable $G_1 \subset S$ and $G \subset \mathcal{R}$, we define

$$\begin{aligned} &\mu_\tau(G_1) = \mathbf{E}\left[e^{-r\tau}I_{G_1}(X(\tau))I_{\{\tau<\infty\}}\right], \quad \mu_0(G_1) = \mathbf{E}\left[\int_0^\tau e^{-rs}I_{G_1}(X(s))ds\right], \\ &\mu_1(G) = \mathbf{E}\left[\int_{\mathcal{R}\times[0,\tau]} e^{-rs}I_G(x,z)\Gamma(dx\times dz\times ds)\right]. \end{aligned} \tag{12}$$

Using these measures, the singular control problem of maximizing (7) over relaxed solutions of the harvesting problem (6) can be written in the form

$$\begin{cases} \text{Maximize} & \displaystyle\int f d\mu_1, \\ \text{subject to} & \displaystyle\int g d\mu_\tau - \int (A-r)g d\mu_0 - \int Bg d\mu_1 = g(x_0), \quad \forall g \in \mathcal{D}, \\ & \mu_\tau, \mu_0, \mu_1 \in \mathcal{M}(S), \mu_\tau(S) \le 1, \mu_0(S) \le \frac{1}{r}. \end{cases} \tag{13}$$

Since each relaxed solution (X, Γ) defines measures μ_τ, μ_0 and μ_1 by (12), the harvesting problem is embedded in (13). There might be feasible measures

which do not arise in this manner. Consequently, letting $V_{lp}(x_0)$ denote the value of the LP problem (13) with initial condition $X(0-) = x_0 > 0$, we have $V(x_0) \leq V_{lp}(x_0)$.

3 The Proof of Theorem 1

This section is devoted to the proof of Theorem 1 and involves two steps.

3.1 Step 1: Universal Upper Bound

The proof follows along the lines of the arguments used in [4]. The general argument involves finding an upper bound for $V_{lp}(x_0)$ by reducing the number of constraints in the linear program (13). We state the results and leave the proofs to the reader.

Proposition 2. *Let b^* be defined as in Theorem 1. Then for every $x_0 \geq 0$,*

$$V(x_0) \leq \frac{f(b^*)}{\psi'(b^*)}\psi(x_0). \tag{14}$$

Notice the bound in (14) holds for all initial positions x_0. The following result shows that this bound is sharp for $x_0 \leq b^*$.

Proposition 3. *For $x_0 \leq b^*$, let L_{b^*} denote the local time process of X at b^*. Define the random measure Γ_{b^*} for Borel measurable $G \subset \mathcal{R}$ and $t \geq 0$ by*

$$\Gamma_{b^*}(G \times [0,t]) = \int_0^t I_G(X(s-), \Delta L_{b^*}(s))\, dL_{b^*}(s). \tag{15}$$

Then $J(x_0, L_{b^*}) = J(x_0, \Gamma_{b^*}) = \frac{f(b^*)}{\psi'(b^*)}\psi(x_0)$.

Since $\Delta L_{b^*}(s) = 0$ for every $s \geq 0$, an optimal strategy is to harvest just enough of the population (using the local time of X^* at b^*) so that the population size "reflects" at b^*.

The value function has been determined for intial population sizes x_0 that are smaller than b^*. It therefore remains to prove the validity of (9) when $x_0 > b^*$.

3.2 Step 2: Return of Stochasticity for a Refined Upper Bound

This step is the more interesting of the two and requires a new argument and also a different type of harvesting policy than appears in the literature.

When dealing with singular control problems, one usually takes the so-called reflection strategy, namely, $Z(t) = (x_0 - b^*)^+ + L_{b^*}(t)$, where one follows an immediate jump from x_0 to b^* by using the local time process L_{b^*} at b^*. Such a reflection strategy is used in [2], [7] and others. The corresponding income is

$$J(x_0, Z) = f(x_0)(x_0 - b^*) + \frac{f(b^*)}{\psi'(b^*)}\psi(b^*).$$

When f is strictly decreasing, the reflection strategy is not optimal. Our purpose is to find an optimal relaxed harvesting strategy.

To develop a sharp upper bound, it is beneficial to revisit the definitions of the occupation measures in (12) so that the connection between the measures, the initial position and the harvesting strategy is more clearly displayed. Let $x_0 \in \mathbb{R}_+$ and (X, Γ) be a relaxed solution of the harvesting model. Modify the notations of the measures to indicate their dependence on x_0 and Γ by writing $\mu_\tau(G; x_0, \Gamma)$, $\mu_0(G; x_0, \Gamma)$ and $\mu_1(G; x_0, \Gamma)$.

Proposition 4. *For $x_0 > b^*$,*

$$V(x_0) \leq \int_{b^*}^{x_0} f(y)\, dy + \frac{f(b^*)}{\psi'(b^*)} \cdot \psi(b^*). \tag{16}$$

Proof. The proof of (16) is broken into two parts, with a technical lemma between the parts.

Part 1: Define the stopping time $\tau_{b^*} = \inf\{t \geq 0 : X(t) \leq b^*\}$ to be the first time the process X takes value at most b^* and note that $\tau_{b^*} \leq \tau$. For the harvesting measure Γ, define $\Gamma_{\tau_{b^*}}$ by

$$\Gamma_{\tau_{b^*}}(G \times [0,t]) = I_{\{\tau_{b^*} < \tau\}} \Gamma(G \times [\tau_{b^*}, \tau_{b^*} + t]), \qquad G \in \mathcal{B}(\mathcal{R}),\ t \geq 0.$$

Notice that $\Gamma_{\tau_{b^*}}$ captures all harvesting using the measure Γ from time τ_{b^*} onwards. Also define the measures $\mu_{0,\tau_{b^*}}$ and $\mu_{1,\tau_{b^*}}$ by

$$\mu_{0,\tau_{b^*}}(G; x_0, \Gamma) = \mathbf{E}_{x_0}\left[\int_0^{\tau_{b^*}} e^{-rs} I_G(X(s))\, ds\right],$$

$$\mu_{1,\tau_{b^*}}(G; x_0, \Gamma) = \mathbf{E}_{x_0}\left[\int_{\mathcal{R}\times[0,\tau_{b^*})} e^{-rs} I_G(x,z)\, \Gamma(dx \times dz \times ds)\right].$$

Note carefully that any harvesting at the time τ_{b^*} is excluded from the measure $\mu_{1,\tau_{b^*}}$. Also observe that the total mass of $\mu_{0,\tau_{b^*}}$ equals $r^{-1}\left(1 - \mathbf{E}_{x_0}\left[e^{-r\tau_{b^*}}\right]\right)$.

Using the strong Markov property of (X, Γ), for each $G \in \mathcal{B}(\mathcal{R})$ it follows that

$$\begin{aligned}
&\mathbf{E}_{x_0}\left[\int_{\mathcal{R}\times[0,\tau]} e^{-rs} I_G(x,z)\, \Gamma(dx \times dz \times ds)\right] \\
&= \mathbf{E}_{x_0}\left[\int_{\mathcal{R}\times[0,\tau_{b^*})} e^{-rs} I_G(x,z)\, \Gamma(dx \times dz \times ds)\right] \\
&\quad + \mathbf{E}_{x_0}\left[\mathbf{E}_{x_0}\left[I_{\{\tau_{b^*}<\tau\}} \int_{\mathcal{R}\times[\tau_{b^*},\tau]} e^{-rs} I_G(x,z)\, \Gamma(dx \times dz \times ds)\right] \middle| \mathcal{F}_{\tau_{b^*}}\right] \\
&= \mathbf{E}_{x_0}\left[\int_{\mathcal{R}\times[0,\tau_{b^*})} e^{-rs} I_G(x,z)\, \Gamma(dx \times dz \times ds)\right] \\
&\quad + \mathbf{E}_{x_0}\left[e^{-r\tau_{b^*}} I_{\{\tau_{b^*}<\tau\}} \mathbf{E}_{X(\tau_{b^*})}\left[\int_{\mathcal{R}\times[0,\tau]} e^{-rs} I_G(x,z)\, \Gamma_{\tau_{b^*}}(dx \times dz \times ds)\right]\right].
\end{aligned}$$

As a result, for each $G \in \mathcal{B}(\mathcal{R})$, this identity can be written in terms of the measures as

$$\mu_1(G; x_0, \Gamma) = \mu_{1,\tau_{b^*}}(G; x_0, \Gamma) + \mathbf{E}_{x_0}\left[e^{-r\tau_{b^*}} I_{\{\tau_{b^*}<\tau\}} \mu_1(G; X(\tau_{b^*}), \Gamma_{\tau_{b^*}})\right].$$

Notice, in particular, that the expectation term involves the measure μ_1 evaluated at the random initial position $X(\tau_{b^*})$. Hence

$$\begin{aligned}
&\int f(y)\,\mu_1(dy; x_0, \Gamma) \\
&= \int f(y)\,\mu_{1,\tau_{b^*}}(dy; x_0, \Gamma) + \mathbf{E}_{x_0}\left[e^{-r\tau_{b^*}} I_{\{\tau_{b^*}<\tau\}} \int f(y)\,\mu_1(dy; X(\tau_{b^*}), \Gamma_{\tau_{b^*}})\right] \\
&\le \int f(y)\,\mu_{1,\tau_{b^*}}(dy; x_0, \Gamma) + \mathbf{E}_{x_0}\left[e^{-r\tau_{b^*}} I_{\{\tau_{b^*}<\tau\}}\right] \frac{f(b^*)}{\psi'(b^*)} \cdot \psi(b^*), \qquad (17)
\end{aligned}$$

in which the inequality follows from Step 1.

This concludes Part 1 of the proof. Part 2 concentrates on estimating the first term of the right-hand side of (17); a technical lemma is required.

Lemma 5 *Assume the conditions in Theorem 1. Define the function h by $h(x) := \int_{b^*}^{x} f(y)dy$ for $x \ge 0$. Then the following estimates hold:*

$$(A - r)h(x) \le r\frac{f(b^*)}{\psi'(b^*)}\psi(b^*), \qquad \text{for } x \ge b^* \text{ and} \qquad (18)$$

$$-Bh(x, z) \ge f(x), \qquad \text{for all } (x, z) \in \mathcal{R}. \qquad (19)$$

Proof. Since by assumption the function f/ψ' is nonincreasing and differentiable on (b^*, ∞), we have

$$0 \ge \frac{d}{dx}\left(\frac{f(x)}{\psi'(x)}\right) = \frac{f'(x)\psi'(x) - f(x)\psi''(x)}{(\psi'(x))^2}, \quad x > b^*.$$

But ψ is strictly increasing and so $\psi'(x) > 0$. Hence it follows that $f'(x)\psi'(x) - f(x)\psi''(x) \le 0$, or equivalently $f'(x) \le \frac{f(x)}{\psi'(x)}\psi''(x)$, for $x > b^*$. It then follows that for each $x > b^*$

$$\begin{aligned}
(A - r)h(x) &\le \frac{1}{2}\sigma^2(x)\frac{f(x)}{\psi'(x)}\psi''(x) + b(x)f(x) - r\frac{f(x)}{\psi'(x)}(\psi(x) - \psi(b^*)) \\
&= \frac{f(x)}{\psi'(x)}\left[\frac{1}{2}\sigma^2(x)\psi''(x) + b(x)\psi'(x) - r\psi(x)\right] + r\frac{f(x)}{\psi'(x)}\psi(b^*) \\
&= r\frac{f(x)}{\psi'(x)}\psi(b^*) \le r\frac{f(b^*)}{\psi'(b^*)}\psi(b^*).
\end{aligned}$$

Turning to a consideration of (19), since f is nonincreasing, for any $0 \le x_1 < x_2$, we have

$$f(x_2)[x_2 - x_1] \le \int_{x_1}^{x_2} f(y)dy = h(x_2) - h(x_1).$$

Hence it follows that for $(x,z) \in \mathcal{R}$, we have

$$-Bh(x,z) = \begin{cases} h'(x) = f(x), & \text{if } z = 0 \\ \frac{h(x)-h(x-z)}{z} \geq f(x), & \text{if } z > 0. \end{cases}$$

The relation (19) is therefore established.

Part 2: The goal is of this part of the proof is to estimate $\int f(y)\,\mu_{1,\tau_{b^*}}(dy \times dz)$ of (17). Using Itô's formula, one obtains for each $t > 0$,

$$\begin{aligned} &- \mathbf{E}\left[\int_{\mathcal{R}\times[0,t\wedge\tau_{b^*})} e^{-rs} Bh(x,z)\,\Gamma(dx \times dz \times ds)\right] \\ &= h(x_0) - \mathbf{E}\left[e^{-r(t\wedge\tau_{b^*})} h(X((t\wedge\tau_{b^*})-))\right] + \mathbf{E}\left[\int_0^{t\wedge\tau_{b^*}} e^{-rs}[A-r]h(X(s))\,ds\right], \end{aligned}$$

in which the deliberate choice of the half-open interval $[0, t\wedge\tau_{b^*})$ in the integral with respect to Γ leads to the use of $X((t\wedge\tau_{b^*})-)$ for the location of the process just before any harvest occurs at time τ_{b^*}. This is extremely important since $h(X((t\wedge\tau_{b^*})-)) \geq 0$ and hence the right-hand side is not decreased by dropping the first expectation. Letting $t \to \infty$ yields

$$\begin{aligned} &- \mathbf{E}\left[\int_{\mathcal{R}\times[0,\tau_{b^*})} e^{-rs} Bh(x,z)\,\Gamma(dx \times dz \times ds)\right] \\ &\qquad \leq h(x_0) + \mathbf{E}\left[\int_0^{\tau_{b^*}} e^{-rs}[A-r]h(X(s))\,ds\right]. \end{aligned}$$

Using the estimates (18) and (19) and the definition of the measures $\mu_{1,\tau_{b^*}}$ and $\mu_{0,\tau_{b^*}}$, we obtain

$$\begin{aligned} \int_{\mathcal{R}} f(y)\,\mu_{1,\tau_{b^*}}(dy \times dz; x_0, \Gamma) &\leq -\int_{\mathcal{R}} Bh(y,z)\,\mu_{1,\tau_{b^*}}(dy \times dz; x_0, \Gamma) \\ &\leq h(x_0) + \int r \cdot \frac{f(b^*)\psi(b^*)}{\psi'(b^*)}\,\mu_{0,\tau_{b^*}}(dx; x_0, \Gamma) \\ &= h(x_0) + \left(1 - \mathbf{E}_{x_0}\left[e^{-r\tau_{b^*}}\right]\right)\frac{f(b^*)}{\psi'(b^*)} \cdot \psi(b^*), \quad (20) \end{aligned}$$

in which the last equality follows from the mass of $\mu_{0,\tau_{b^*}}$. Combining (17) and (20) produces the desired relation

$$\int f(y)\,\mu_1(dy \times dz; x_0, \Gamma) \leq \int_{b^*}^{x_0} f(y)\,dy + \frac{f(b^*)}{\psi'(b^*)} \cdot \psi(b^*).$$

We have derived an upper bound for the value $V(x_0)$ in Proposition 4. The following proposition exhibits an optimal relaxed harvesting policy. The proof is left to the reader.

Proposition 6. *Let $\lambda_{[b^*,x_0]}(\cdot)$ denote Lebesgue measure on $[b^*, x_0]$. Also let L_{b^*} denote the local time process of Proposition 3 with x_0 taken to be b^* and denote by Γ_{b^*} the random measure defined in (15). Finally, define the relaxed harvesting strategy by*

$$\Gamma^*(dx \times dz \times dt) = \lambda_{[b^*,x_0]}(dx)\delta_{\{0\}}(dz)\delta_{\{0\}}(dt) + \Gamma_{b^*}(dx \times dz \times dt).$$

Then

$$V(x_0) = J(x_0, \Gamma^*) = \int_{b^*}^{x_0} f(y)dy + \frac{f(b^*)}{\psi'(b^*)}\psi(b^*). \tag{21}$$

We observe that the manner in which this optimal harvesting policy differs from the typical "reflection" strategy occurs at the initial time. Whereas the reflection strategy has the process X instantaneously jump from x_0 to b^*, the optimal relaxed harvesting policy obtains this relocation in an instantaneous *but continuous* manner.

Finally we note that the combination of Propositions 2 and 6 establishes Theorem 1. Moreover, the optimal relaxed harvesting policy in (10) unifies the two cases.

Acknowledgements. This research was supported in part by the U.S. National Security Agency under Grant Agreement Number H98230-09-1-0002, the National Science Foundation under DMS-1108782, a grant from the UWM Research Growth Initiative, and City University of Hong Kong (SRG) 7002677. The United States Government is authorized to reproduce and distribute reprints notwithstanding any copyright notation herein.

References

1. Alvarez, L.H.R.: Singular stochastic control in the presence of a state-dependent yield structure. Stochastic Process. Appl. 86(2), 323–343 (2000)
2. Choulli, T., Taksar, M., Zhou, X.Y.: A diffusion model for optimal dividend distribution for a company with constraints on risk control. SIAM J. Control Optim. 41(6), 1946–1979 (2003)
3. Ethier, S.N., Kurtz, T.G.: Markov processes: Characterization and Convergence. John Wiley & Sons Inc., New York (1986)
4. Helmes, K.L., Stockbridge, R.H.: Thinning and harvesting in stochastic forest models. J. Econom. Dynam. Control 35(1), 25–39 (2011)
5. Kurtz, T.G., Stockbridge, R.H.: Stationary solutions and forward equations for controlled and singular martingale problems. Electron. J. Probab. 6, 1–52 (2001), Paper No. 14
6. Lungu, E., Øksendal, B.: Optimal harvesting from a population in a stochastic crowded environment. Math. Biosci. 145(1), 47–75 (1997)
7. Pham, H.: Continuous-time stochastic control and optimization with financial applications. Springer, Berlin (2009)
8. Song, Q.S., Stockbridge, R.H., Zhu, C.: On optimal harvesting problems in random environments. SIAM J. Control Optim. 49(2), 859–889 (2011)

Estimation of Loan Portfolio Risk on the Basis of Markov Chain Model

Nikolay Timofeev[1] and Galina Timofeeva[2,⋆]

[1] Ural State University of Railway Transport,
Kolmogorov str. 64, 620034, Ekaterinburg, Russia
[2] Ural Federal University,
Mira str., 19, 62002, Ekaterinburg, Russia

Abstract. A change of shares of credits portfolio is described by Markov chain with discrete time. A credit state is determined on as an accessory to some group of credits depending on presence of indebtedness and its terms. We use a model with discrete time and fix the system state through identical time intervals - once a month. It is obvious that the matrix of transitive probabilities is known incompletely. Various approaches to the matrix estimation are studied and methods of forecast the portfolio risk are proposed. The portfolio risk is set as a share of problematic loans. We propose a method to calculate necessary reserves on the base of the considered model.

Keywords: Loan portfolio, Markov chain, incomplete information.

1 Introduction

Markov chain models [1] are widely used to explain the dynamics of state changes for different systems. Often they are used as a mathematical model for some random physical process.

Markov chains are used in finance and economics to model a variety of different phenomena, including asset prices, market crashes and credit portfolio dynamics [2,3].

If the transition probability matrix of Markov chain is known then dynamics of the system states probabilities is completely described by a system of difference equations. As a rule the transition probabilities are unknown and estimated during the system evolution.

The most important indicator of a bank loan portfolio quality is a probability of default which is closely connected with a share of the problematic loans [5,6]. A value of necessary reserves depends on quality and structure of the portfolio. On the one hand reserves should provide low probability of default, on the other hand they impact on profitability of the portfolio.

Let's assume that a change of shares of credits portfolio is described by Markov chain with discrete time. In this case the credit state is determined on as an accessory to some group of credits depending on presence of indebtedness and its

⋆ The investigations are partially supported by the Russian Foundation for Basic Research, project 10-01-00672a.

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 207–216, 2013.

terms. We will use a model with discrete time and fix the system state through identical time intervals – once a month. It is proposed that the transition probabilities vary a little. Thus, we consider a stable economic situation when the transition probabilities are constant.

It is obvious that the matrix of transitive probabilities is known incompletely. Its values are estimated using a data on changes the quality of loans (a migration analysis of the portfolio). Criterion of a choice is accuracy of the forecast of a share of problematic loans.

2 Mathematical Model

2.1 Dynamics of System States

We consider a system with k states, the probability that the system is in i-th state at moment t denote by $x_i(t)$, $i = 1, \ldots, k$. Thus the following conditions hold:

$$0 \leq x_i(t) \leq 1, \quad x_1(t) + \ldots + x_k(t) = 1. \tag{1}$$

The dynamics of the system states probabilities is described by the discrete Markov chain model:

$$x_j(t+1) = \sum_{i=1}^{k} p_{ij} x_i(t), \quad t = 0, 1, \ldots, T, \tag{2}$$

where p_{ij} is the probability of transition from state i to state j in one step.

The first-order stationary Markov model for credit transitions is somewhat restrictive as a credit quality responds to changes in economics. Using a higher-order Markov process or a nonstationary transition probability matrix may be more appropriate, but in such models one should estimates too many parameters and the requirements to statistical data increase quite substantially. Thus the simple Markov chain is usually used in stable economic situation and with not longer time horizon [2,3].

Let's denote by $x(t)$ a vector of states probabilities $x(t) = \{x_1(t), \ldots, x_k(t)\}^\top$, by P a matrix of the transition probabilities $P = \{p_{ij}\}$ and rewrite equation (2) in the vector form:

$$x(t+1) = P^\top x(t), \quad t = 0, 1, \ldots, T. \tag{3}$$

When the transition probability matrix P is known incompletely there is a problem to estimate $x(T)$. It is assumed that we have information about the number of transitions from i-th state to j-th on t step, $t = 1, \ldots, m$.

2.2 Ways to Select Groups

We consider two ways of describing a credit portfolio dynamics: a regular Markov chain in which we do not take into account repaid loans and renovation of the

portfolio and a scheme which included "new loan" and "repaid loan" as possible states of a loan. In the first way we investigate a steady-state behavior of the loan portfolio shares, in the second way we study the profitability of loans from delivery to its repayment.

Let us consider a fist way of describing the states of loan. For beginning we consider a simplified scheme with three groups of loans:

1. Loans without delay, including the new ones (S_1);
2. Overdue loans with $1-65$ days delay (S_2);
3. Non-performing (problematic) loans (S_3).

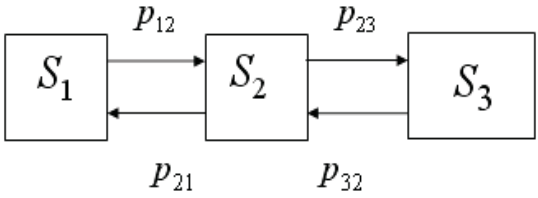


Fig. 1. Graph of the Markov chain for the simplified scheme

A graph of the system state is in Fig.1. An investigation of this scheme allows to define the most important features of the model. As a rule risk-managers use more detailed schemes which takes into account a number of overdue days on a loan. For example let consider the expanded scheme with 5 groups of loans [7]: loans without delay, including a new one; overdue loans with 1 to 35 days delay; overdue loans with 35 to 65 days delay; more than 65 days overdue loans (problematic loans); reconstructed loans.

In these schemes the renovation and repayment of credits are considered without allocating a separate state. New credits are included into the first group together with credits without delay.

A loan is "reconstructed" in case it was problematic in the previous period and a borrower has made partial payments under the credit.

In schemes with the repayments(amortization) there is the category "repaid credits". The scheme on the basis of the expanded scheme of loans has 6 groups of loans: loans without delay, including a new one; overdue loans with 1 to 35 days delay; overdue loans with 35 to 65 days delay; more than 65 days overdue loans (problematic loans); reconstructed loans; repaid loans.

3 Estimation of Transition Probabilities

For the estimation of the probability p_{ij} one usually use the statistical data about transitions from one state to another. Let't denote by $n_i(t)$ the number of individuals who are in state i in period t, and by $n_{ij}(t)$ the number of individuals

who were in state i in period $t-1$ and are in state j in period t. We can estimate the probability p_{ij} of an individual being in state j in period t given that they were in state i in period $t-1$.

The probability of transition $p_{ij}(t)$ from any given state i is approximated by a proportion of individuals that started in state i and ended in state j as a proportion of all individuals in that started in state i:

$$w_{ij}(t) = \frac{n_{ij}(t)}{n_i(t-1)}. \tag{4}$$

If the Markov chain is stationary (i.e. $p_{ij}(t) \equiv p_{ij}$) then one can use another estimate:

$$\tilde{w}_{ij} = \frac{\sum\limits_{t=1}^{T} n_{ij}(t)}{\sum\limits_{t=0}^{T-1} n_i(t)} \tag{5}$$

Using the methods described above, it is possible to estimate a transition matrix using count data.

Anderson and Goodman [4] showed that the estimator w_{ij} given by equation (5) is a maximum-likelihood estimator and calculated statistical moments of random values $\xi_{ij}(t) = n_{ij} - p_{ij}n_i(t-1)$.

On the t-th step $n_i(t-1) \triangleq n_i$, $i = 1, \ldots, k$ are known. The statistical moment of w_{ij} are following [8]:

$$E(w_{ij}(t)) = p_{ij}, \tag{6}$$

$$\begin{aligned} Var(w_{ij}(t)) &= \tfrac{p_{ij}(1-p_{ij})}{n_i}, \\ Cov(w_{ij}(t), w_{il}(t)) &= -\tfrac{p_{ij}p_{il}}{n_i}, \quad j \neq l, \\ Cov(w_{ij}(t), w_{cl}(t)) &= 0, \quad i \neq c. \end{aligned} \tag{7}$$

Here $E(\xi)$ is a mathematical expectation of a random value ξ, $Var(\xi)$ is its variance, $Cov(\xi, \zeta)$ is a covariance between ξ and ζ.

From relations (7) follows that w_{ij} and w_{cl} are noncorrelated if $i \neq c$.

Suppose that instead of observing the actual count of transitions from the different states, we only observe the aggregate proportions $y_i(t)$, which represent the proportion of observations with the state i. The aggregate proportions $y_i(t)$ estimate the system state probabilities: $y_i(t) \approx Nx_i(t)$, $i = 1, \ldots, N$, $t = 1, \ldots, T$, where N is a number of all individuals.

If the time series of observations T are sufficiently long, it is possible to estimate a transition matrix P from aggregate data using the least square method [9] and its generalizations [10].

The maximum-likelihood estimates in asset prices model is used to estimate transition matrices in credit risk modeling with a decades-old methodology that uses aggregate proportions data [2].

To specificate an order and a number of states in Markov chain model one may use criteria [4], but they are based on the detailed analysis of the data such as $n_{ijm}(t)$, where $n_{ijm}(t)$ is a number of individuals in state i at $t-2$, in j at $t-1$ and in m at t.

4 Approaches to Estimation the Share of Problematic Loans

We considered two ways to forecast the share of problem loans taking into account the uncertainty of the transition probabilities matrix. There are a confidence estimation method and a simulation method.

4.1 Confidence Estimation

The confidence estimation method consists of two stages: construction a confidence set Z_α for the transition probabilities matrix based on statistical data and relations (6)–(7) and the analysis of all possible trajectories of the system taking into account that the probabilities are constant but uncertain.

Let's estimate the elements of the transition probability matrix P. Denote the confidence region for $\{p_{ij} = z_s,\ i = 1, \ldots, k,\ j = 1, \ldots, k,\ j \neq i\}$ on m-th step by $Z_\alpha \subset \mathbb{R}^K$, $s = 1, \ldots, K$, $K = k(k-1)$.

Thus p_{ij} are the transition probabilities then $Z_\alpha \subset Z_+ \subset \mathbb{R}^K$, where Z_+ is the set of all possible values of transition probabilities $\{p_{ij},\ j \neq i\}$

$$Z_+ = \{z_s : \quad 0 \leq z_s \leq 1, \quad \sum_{s=1}^{K} z_s \leq 1\} \subset \mathbb{R}^K.$$

Estimation for p_{ii} follows from the equalities

$$p_{i1} + \ldots + p_{ik} = 1, \quad i = 1, \ldots, k. \tag{8}$$

In the considered model $p_{ij} = z_s$ are distributed approximately normal [4] with mean values equal to $w_{ij} \triangleq \bar{z}_s$ and a covariance matrix G defined by relations (7) with substitution p_{ij} by w_{ij}.

Therefore we may use the confidence set Z_α defined by joint restrictions:

$$Z_\alpha = \{z \in Z_+ : (z - \bar{z})^\top G (z - \bar{z}) \leq b_K^{(\alpha)}\}, \tag{9}$$

where $b_K^{(\alpha)}$ is the α-quantile of χ^2 distribution with K degrees of freedom.

The next step is to solve the state estimation problem of a multistage deterministic system with uncertain matrix P:

$$\begin{aligned} &x(t+1) = P^\top x(t), \quad t = m, \ldots, T, \\ &x(m) = x^*, \quad P \subset Z. \end{aligned} \tag{10}$$

and to find an information set

$$X(t,Z) = \{x \in R_+^k : x = (P^\top)^{T-m} x^*, \quad P \in Z\}. \tag{11}$$

We may construct information sets for system (10) using approaches proposed by Kurzanski and Tanaka [11].

This method is very time-consuming because the number of estimated elements of the matrix is large. We can perform calculations only for the scheme with 3 groups of loans, in which only 4 probabilities should be estimated. For schemes 2 and 3 the number of estimated probabilities are 9 and 17 respectively.

Example 1. Let us consider the estimation problem for the scheme with 3 groups of loans (Fig.1). For this scheme the matrix of transition probabilities has a form

$$P = \begin{pmatrix} 1-p_{12} & p_{12} & 0 \\ p_{21} & 1-p_{21}-p_{23} & p_{23} \\ 0 & p_{32} & 1-p_{32} \end{pmatrix}, \tag{12}$$

and only 4 transition probabilities p_{ij} should be estimated. Denote them as z_s, $s = 1, \ldots, 4$:

$$p_{12} = z_1, \quad p_{23} = z_2, \quad p_{32} = z_3, \quad p_{21} = z_4. \tag{13}$$

From a statistical data we estimate the mean values $\bar{z}_s = w_{ij}$, where w_{ij} defined by (4). A covariance matrix is calculated using relations (7), where estimates w_{ij} substituted instead of values p_{ij}:

$$G = \begin{pmatrix} \frac{\bar{z}_1(1-\bar{z}_1)}{n_1} & 0 & 0 & 0 \\ 0 & \frac{\bar{z}_2(1-\bar{z}_2)}{n_2} & 0 & -\frac{\bar{z}_2\bar{z}_4}{n_2} \\ 0 & 0 & \frac{\bar{z}_3(1-\bar{z}_3)}{n_3} & 0 \\ 0 & -\frac{\bar{z}_2\bar{z}_4}{n_2} & 0 & \frac{\bar{z}_4(1-\bar{z}_4)}{n_2} \end{pmatrix}. \tag{14}$$

Then find a confidence set for transition probabilities p_{ij} in ellipsoidal form (9). For the considered scheme with 3 groups of loans we obtain

$$Z_\alpha = \{z_s \in R_+^4 : (z-\bar{z})^\top G (z-\bar{z}) \le b_4^{(\alpha)}\}. \tag{15}$$

Then we find an information set for system (5) for a given $T = 12$ using ellipsoidal calculus [12]. Thus we get the confidence set for the portfolio shares after 12 months $x(12)$ and for the share of problematic loans $x_3(12)$ in particularly. For our data we obtain $x_3(12) \in [0.03; 0.152]$ with probability $\alpha = 0.95$.

4.2 Simulation Method

This method has 4 stages:

1. We determine statistical moments (4) and (5) for the transition probabilities based on statistical data on transitions between loan states;
2. Generate the random vector of unknown probabilities as the Gaussian vector with the given statistical moments;

3. Simulate the dynamics of system (2) with the large number of runs;
4. On the basis of the calculations determine the mean value and the confidence interval for the share of problem loans.

A histogram of the predicted share of problem loans after 12 months one can see in Fig. 2. On the base of modeling we find the expected mean of value of the share

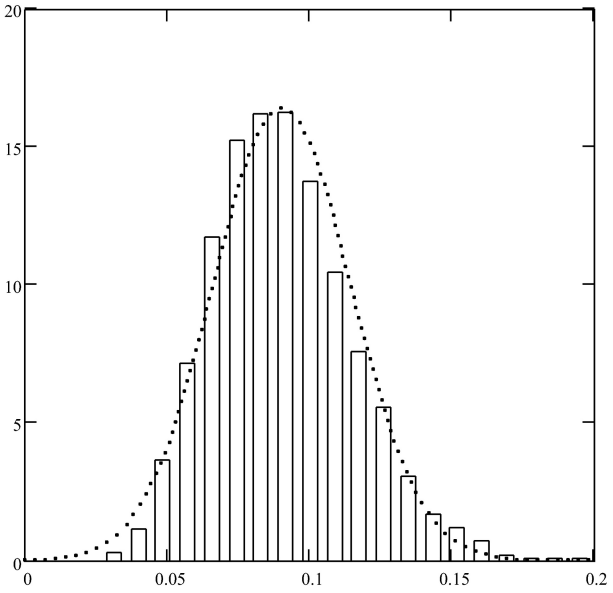


Fig. 2. Histogram of the share of problematic loans and normal density

of problem loans after 12 month: $x(t+12) = 0.09$ and quantile $q_{0.95} = 0.135$, i.e. $x(t+12) < 0.135$ with the probability $\alpha = 0.95$.

This result is more precise than the confidence interval obtained by the confidence estimation, because the confidence interval for the share of problematic loans, obtained in section 4.1, depends on the form of confidence set for unknown parameters p_{ij} and the ellipsoidal form is not optimal in this case.

5 Estimation of a Reserve

According recommendations of the International Committee [13] bank portfolio managers should estimate a risk of the portfolio and form reserves in a proportion of the its value.

There are different approaches to estimate the necessary reserves. Some of them based on probability of "nonreconstruction" of the problematic credits. Let's denote by $p_{ij}^{[t]}$ a probability of transition from the i-th state to j-th state by t steps. It is an element of the t power of the transition probability matrix $P^{[t]} = P^t$. Denote by $D^{[t]} = p_{mm}^{[t]}$, where m-th state is "problematic loans". The value $1 - D^{[T]}$ is called a probability of reconstruction during a period T.

The value of risk for j-th group of loans is defined as

$$\hat{r}_j = p_{jm}^{[\tau_1]} D^{[\tau_2]}, \text{if } \ j \neq m,$$

$$\hat{r}_m = D^{[\tau_1]}.$$

The reserve W is equals to: $W = W_1\hat{r}_1 + \ldots + W_k\hat{r}_k$, where W_j is a sum of loans in j-th group.

Risk managers take different τ_1 and τ_2, such as 6, 12 or 24 months. The problem is how to choose the period of transition to the problematic state and the reconstruction period. In some approaches value of τ_1 depends on j and a period of an overdue.

We propose another approach taking into account time structure of possible losses. Let's define a risk of loans in j-th group as maximum of a sum of problematic loans in future for loans of this group. Thus, the risk for j-th loans group equals

$$r_j = \max_{t \in 0,\ldots,T} \frac{1}{(1+\rho)^t} p_{jm}^{[t]}, \tag{16}$$

where ρ is a month discount factor.

For a new loan we have

$$r_0 = \max_{t \in 0,\ldots,T} \frac{1}{(1+\rho)^t} p_{0m}^{[t]}. \tag{17}$$

We can take into account that transition probabilities are known incompletely and instead of (16) use its quantile:

$$q_j(\alpha) : \mathcal{P}\{\max_{t \in 0,\ldots,T} \frac{1}{(1+\rho)^t} p_{jm}^{[t]} \leq q_j(\alpha)\} \geq \alpha, \tag{18}$$

where $\mathcal{P}(A)$ is a probability of a random event A. The quantile may be calculated using the confidence approach or the simulation method (see Sect. 4).

Example 2. Let's consider a scheme with repayment (Scheme 3) and calculate the value of reserves for new loans using the proposed approach. Is proposed that we may estimate statistical moments (6) of the transition probabilities p_{ij} based on a previous data and estimates (4), (5) or their modifications.

We take a possible value of transitions probabilities matrix $P = \{p_{ij}\}$, calculate

$$h(t) = \frac{1}{(1+\rho)^t} p_{0m}^{[t]}, \quad t = 0, \ldots, T,$$

and find their maximum $r_0(t)$ which depended on matrix P.

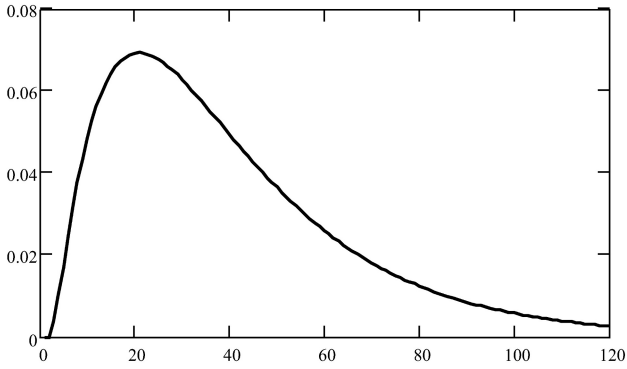


Fig. 3. Dependence of the discounted sum of problem loans $h(t)$ on time

In Fig. 3 one can see changes of $h(t)$ (the discounted sum of problem loans) for a fixed matrix P. Values of $h(t)$ decreasing for $t \geq 20$.

We generate many times transition probabilities according normal distribution with the given statistical moments, calculated the discounted sum of problem loans $h(t) = h(t, P)$, and its maximum $r_0 = r_0(P)$, then we obtain estimates of a mean value and quantile (18).

In Fig.4 one can see a histogram for risk estimates $r_0 = r_0(P)$. For considered data mean value of risk $\bar{r}_0 = 0.039$, $q_{0.95} = 0.065$.

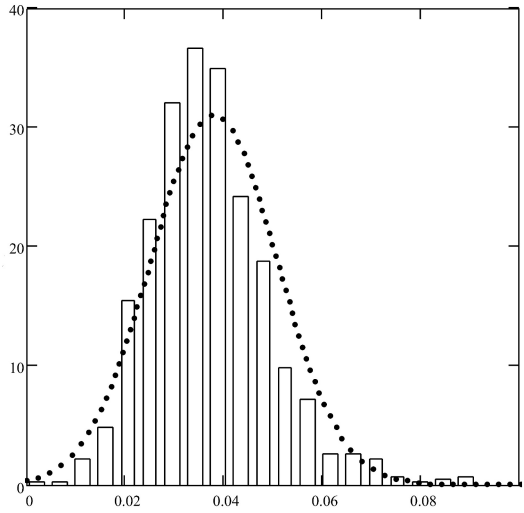


Fig. 4. Histogram of risk $r_0(P)$ and normal density

6 Conclusion

We studied the discrete Markov chain model for loan portfolio. It is proposed that transition probabilities are unknown and estimated during the process. We proposed methods to estimate system state probabilities in future. Obtained results apply to forecast credit portfolio shares and to define necessary reserves.

References

1. Markov, A.A.: Rasprostranenie zakona bol'shih chisel na velichiny, zavisyaschie drug ot druga. Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete. Ser. 2, vol. 15, pp. 135–156 (1906); Reprinted in: Markov, A.A.: Extension of the limit theorems of probability theory to a sum of variables connected in a chain. Appendix B of: Howard, R. Dynamic Probabilistic Systems. 1: Markov Chains. John Wiley and Sons (1971)
2. Jones, M.T.: Estimating Markov Transition Matrices Using Proportions Data: An Application to Credit Risk. IMF Working Paper, WP-05-219 (2005)
3. Thyagarajan, V., Saiful, M.: Retail Banking Loan Portfolio Equilibrium Mix: A Markov Chain Model Analysis. Amer. J. of Applied Sciences 2(1), 410–419 (2005)
4. Anderson, T.W., Goodman, L.A.: Statistical Inference About Markov Chains. Annals of Mathematical Statistics 28, 89–110 (1957)
5. Hanson, S., Schuermann, T.: Confidence intervals for probabilities of default. Journal of Banking & Finance 30(8), 2281–2301 (2006)
6. De Andrade, F.W.M., Thomas, L.: Structural models in consumer credit. European Journal of Operational Research 183, 1569–1581 (2007)
7. Timofeev, N.A.: Mathematical Model of the Vintage Analysis of Banking Credit Portfolio. Gerald of USURT 9(1), 61–69 (2011)
8. Timofeeva, G.A., Timofeev, N.A.: Predicting the Components of Credit Portfolio Based on a Markov Chain Model. Automation and Remote Control 73(4), 637–651 (2012)
9. Kalbfleisch, J.D., Lawless, J.F.: Least-Squares Estimation of Transition Probabilities From Aggregate Data. Canadian J. of Statistics 12(3), 169–182 (1984)
10. MacRae, E.C.: Estimation of Time-Varying Markov Processes with Aggregate Data. Econometrica 45(1), 183–198 (1977)
11. Kurzanski, A.B., Tanaka, M.: On a unified framework for deterministic and stochastic treatment of identification problem. IIASA, Laxenburg (1989)
12. Kurzhanskiy, A.A., Varaiya, P.: Ellipsoidal Techniques for Reachability Analysis of Discrete-Time Linear Systems. IEEE Transactions on Automatic Control 52(1), 26–38 (2007)
13. Basel Committee on Banking Supervision, International Convergence of Capital Measurement and Capital Standards: A Revised Framework (2004), http://www.bis.org/publ/bcbs107.html

MPC/LQG for Infinite-Dimensional Systems Using Time-Invariant Linearizations

Peter Benner[1] and Sabine Hein[2]

[1] Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstr. 1, 39106 Magdeburg, Germany
benner@mpi-magdeburg.mpg.de
http://www.mpi-magdeburg.mpg.de/mpcsc/benner/
[2] Voith Engineering Services GmbH, Aue 23–27, 09112 Chemnitz
sabine.hein@mathematik.tu-chemnitz.de

Abstract. We provide a theoretical framework for model predictive control of infinite-dimensional systems, like, e.g., nonlinear parabolic PDEs, including stochastic disturbances of the input signal, the output measurements, as well as initial states. The necessary theory for implementing the MPC step based on an LQG design for infinite-dimensional linear time-invariant systems is presented. We also briefly discuss the necessary ingredients for the numerical computations using the derived theory.

1 Introduction

The control of nonlinear processes is a fundamental problem in engineering. A usual strategy for computer-aided control consists in pre-computing, in an *offline* phase, an optimal trajectory and control input, and in the implementation (*online* phase) to endow the system with a feedback controller in order to compensate for external disturbances and deviations from the optimized trajectory. A successful strategy for designing a nonlinear control scheme for complex dynamical systems, whose global optimization is impossible in real time, is model predictive control (MPC), see, e.g., [10,14]. In this approach, the behavior of the dynamical process is predicted on a small (local) time horizon and then optimized for a certain time interval using an auxiliary problem for which the computational solution of the local optimization problem is feasible in real-time. The control strategy is then applied for a small time step, the process is advanced in time, and for the next time step, prediction and optimization are repeated for the new state of the system based on new available measurements. Under certain conditions, this process converges to the optimal solution of the global control problem if the time steps and prediction/optimization horizons tend to zero, see, e.g., [10,14] and references therein. If used as a feedback control scheme, the "optimization" goal is to minimize the deviation from the desired trajectory so that *stabilization*, i.e., convergence to zero, becomes the goal.

If the state is not fully available in the prediction step, one is faced with the problem of incomplete observations. This requires to include a state estimator in

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 217–224, 2013.

the prediction/optimization step. If the local optimization problem is solved via an auxiliary linear-quadratic optimal control (LQ) problem (based on a suitable linearization of the nonlinear system) without control and state constraints, the optimal state estimator in a least-squares sense is given by the Kalman(-Bucy) filter [19], and the solution of the local LQ problem is obtained by the linear-quadratic Gaussian (LQG) design, consisting of a combination of the Kalman filter and the linear-quadratic regulator (LQR). This requires the solution of two Riccati equations. Depending on whether the linearization is time-invariant or time-varying, these are the algebraic or differential Riccati equations (ARE or DRE, respectively) — see, e.g., [12] or any other textbook on linear control design. In [18], this MPC/LQG approach was suggested for finite-dimensional problems. Here, we will extend this idea to infinite-dimensional systems.

In the following we consider the control problem

$$\min \int_0^{T_f} \langle y(t), Q(t)y(t)\rangle_{\mathcal{Y}} + \langle u(t), R(t)u(t)\rangle_{\mathcal{U}}\, dt + G(x(T_f)), \tag{1}$$

subject to the semi-linear stochastic system with additive unmodeled disturbance

$$\dot{x}(t) = f(x(t)) + Bu(t) + Fv(t), \quad t > 0, \quad x(0) = x_0 + \eta, \tag{2}$$

$u(t) \in \mathcal{U}$, $x(t) \in \mathcal{X}$, where $v(t)$ is an unknown Gaussian disturbance process with covariance V and η denotes the noise in the initial condition. Since in many applications the state is not completely available we introduce the output function (simulating measurements)

$$y(t) = Cx(t) + w(t), \quad y \in \mathcal{Y},$$

where $w(t)$ is a measurement noise process which will also be assumed to be Gaussian with covariance W. If (2) is an ordinary differential equation then we have a finite-dimensional problem with $\mathcal{X} = \mathbb{R}^n$, $\mathcal{Y} = \mathbb{R}^p$ and $\mathcal{U} = \mathbb{R}^m$. In the case of a partial differential equation (PDE), the problem is infinite-dimensional and $\mathcal{X}, \mathcal{Y}, \mathcal{U}$ are appropriate Hilbert spaces. Here, B, F, Q, R are linear operators on these Hilbert spaces, f is a nonlinear map, and $\langle\,.\,,\,.\,\rangle$ are inner products on the respective Hilbert spaces.

The outline of the paper is as follows: in the next section, we will briefly sketch the MPC/LQG control design. In Section 3, we will then provide the necessary theoretical background to solve the local LQG problems for infinite-dimensional systems and briefly discuss a framework for the numerical approximation of the solution of the AREs to be solved in a practical implementation of the infinite-dimensional controller. Note that we have demonstrated the efficiency of the suggested MPC/LQG approach for the stabilization of the noisy Burgers equation in [4] and for a 3D reaction-diffusion system in [6]. Concluding remarks are provided in Section 4.

2 The MPC/LQG Controller

Given a reference trajectory and control $(\bar{x}(t), \bar{u}(t))$ obtained, e.g., from an offline optimization procedure, in the following we design a model predictive feedback control strategy. This MPC/LQG approach is based on a linearization of (2) on small intervals to obtain a linear time-invariant (LTI) or time-varying (LTV) problem. Due to space restrictions, we will concentrate here on the LTI case. For the general strategy in the LTV case, we refer to [5,15] We solve this linear problem on a small interval by using an LQG design. Note that we write M^* to denote the adjoint operator corresponding to the linear operator M and the derivative of f from (2) is to be understood as the Fréchet derivative. With these preliminaries, the strategy is the following:

(1) **Prediction and optimization step on** $[t_i, t_i + T_p]$**,** $t_i + T_p \le T_f$**:**
linearize (2) around a given set point $\bar{x}$ to obtain $A = f'(\bar{x}(t_i))$ and the linear state equation

$$\dot{z}(t) = Az(t) + B\tilde{u}(t) + Fv(t), \quad z(t_i) = x(t_i) - \bar{x}(t_i), \quad y(t) = Cx(t) + w(t),$$

with $z(t) = x(t) - \bar{x}(t)$ and $\tilde{u}(t) = u(t) - \bar{u}(t)$. Then solve the ARE

$$0 = XA + A^*X - XBR^{-1}B^*X + C^*QC \tag{3}$$

in order to obtain $X_\star$ and $K = -R^{-1}B^*X_\star$.

(2) **Implementation step on** $[t_i, t_i + T_c]$**,** $T_c \le T_p$**:**
solve the filter ARE (FARE)

$$0 = A\Sigma + \Sigma A^* - \Sigma C^*W^{-1}C\Sigma + FVF^*, \tag{4}$$

where V, W are the covariance matrices of the noise processes. Feed the real system on $[t_i, t_i + T_c]$ with

$$u(t) = \bar{u}(t) - K(\hat{x}(t) - \bar{x}(t)),$$

and obtain the "measurement" $y(t)$ by solving the nonlinear system on $[t_i, t_i + T_c]$. Estimate the state by $\hat{x}(t)$ by solving

$$\dot{\hat{z}}(t) = A\hat{z}(t) + B\tilde{u}(t) + L\big(y(t) - C\hat{x}(t)\big), \quad \hat{z}(t) = \hat{x}(t) - x^*(t),$$

using the estimator gain $L = \Sigma_\star C^*W^{-1}$.

(3) **Receding Horizon Step:**
update $t_{i+1} = t_i + T_c$ and go to the first step.

Note that the solutions of the AREs (3) and (4) are linear selfadjoint operators on $\mathcal{D}(A)$, the domain of A, and $\mathcal{D}(A^*)$, respectively.

Some remarks are in order:

Remark 1. For the finite-dimensional case, if G in (1) is selected as a control Lyapunov function, Ito and Kunisch established the asymptotic stability and estimated the performance for the receding horizon synthesis in [16]. Analogous results for the LTV case in the slightly more general MPC setting are obtained in [5,15].

Remark 2. Solving the AREs (3) and (4) yields an LQG controller for an infinite time-horizon. Therefore, we can also consider this scheme as an MPC scheme with infinite prediction and optimization horizon.

In the following, we will discuss an appropriate setting in which this procedure is well-posed and can be approximated using an appropriate numerical scheme. Note that using efficient numerical algorithms, large-scale AREs resulting from discretized PDE control problems can be solved in a reasonable time-scale, see [7]. Whether or not this is real-time feasible depends on the control horizon T_c of the process. Further advances in computer hardware and improvements of the numerical algorithms will certainly allow real-time solution of AREs for moderately fast processes with medium-fine granularity of the discretization in the near future.

3 Infinite-Dimensional LQG Theory

Consider the following nonlinear optimal control problem:

$$\min \mathcal{J}(u) := \langle x_{T_f}, G x_{T_f} \rangle_{\mathcal{X}} + \int_0^\infty \langle x(t), C^* Q C x(t) \rangle_{\mathcal{X}} + \langle u(t), R u(t) \rangle_{\mathcal{U}} \, dt, \tag{5}$$

$$\begin{aligned} \text{subject to} \quad \dot{x}(t) &= f(x(t)) + Bu(t) + Fv(t), \quad t > 0, \\ y(t) &= Cx(t) + w(t), \quad t > 0, \\ x(0) &= x_0 + \eta. \end{aligned}$$

Following the infinite-dimensional LQG theory derived in [11] and denoting the set of linear maps from $\mathcal{M}$ to $\mathcal{N}$ by $\mathcal{L}(\mathcal{M}, \mathcal{N})$, we will assume the following:

Assumption 1

- $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{U}$ *are Hilbert spaces,* $f : \mathcal{D}(f) \subseteq \mathcal{X} \to \mathcal{X}$ *is a nonlinear map;*
- $B \in \mathcal{L}(\mathcal{U}, \mathcal{X})$, $F \in \mathcal{L}(\mathcal{U}, \mathcal{X})$, $C \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$, $G \in \mathcal{L}(\mathcal{X})$*;*
- $Q \in \mathcal{L}(\mathcal{Y})$, $R, R^{-1} \in \mathcal{L}(\mathcal{U})$, *all self-adjoint and nonnegative and* $\langle \nu, R\nu \rangle \geq \alpha ||\nu||^2$ *for all* $\nu \in \mathcal{U}$ *and some* $\alpha > 0$*;*
- $x_0 \in \mathcal{X}$ *and* η *is a zero mean Gaussian random variable on* $\mathcal{X}$ *with covariance* Σ_0*,*

– *$v(t)$ and $w(t)$ are Wiener processes (Gaussian with zero mean) on the Hilbert spaces $\mathcal{U}$ and $\mathcal{Y}$ with incremental covariance operators $V \in \mathcal{L}(\mathcal{U})$ and $W, W^{-1} \in \mathcal{L}(\mathcal{Y})$, respectively.*

Assuming that $f(x)$ is Fréchet-differentiable and linearizing on small intervals $[t_i, t_i + T_p]$ around a stationary operating point $\bar{x}$, we obtain the stochastic LTI problem in differential form on $[t_i, t_i + T_p]$:

$$\begin{aligned} dz(t) &= Az(t)dt + B\tilde{u}(t)dt + Fdv(t), \quad t_i < t < t_i + T_p, \\ d\tilde{y}(t) &= Cz(t)dt + dw(t), \quad t_i < t < t_i + T_p, \\ z(t_i) &= z_{t_i}, \end{aligned} \tag{6}$$

with $z(t) := h(t) = x(t) - \bar{x}(t)$, $\tilde{u}(t) = u(t) - \bar{u}(t)$ and

$$f(\bar{x} + h)(t) \approx f(\bar{x}(t)) + Ah(t),$$

where $A := f'(\bar{x}(t_i))$ is the Fréchet derivative of f, evaluated at $\bar{x}(t_i)$.

Note that the linear system (6) is only a local approximation to the original nonlinear system, but nevertheless we will use it to solve the LQG problem on an infinite horizon. The so obtained control is then only applied locally on the control horizon $[t_i, t_i + T_c]$, then the prediction horizon is shifted by T_c, and a new linearization based on $\bar{x}(t_i + T_c)$, leading to a new LQG problem, is used.

To avoid problems of existence and uniqueness of the stochastic evolution equation (6), we use its integral form on $[t_i, t_i + T_p]$:

$$\begin{aligned} z(t) &= S_{t-t_i} z(t_i) + \int_{t_i}^{t} S_{t-s} B\tilde{u}(s)\, ds + \int_{t_i}^{t} S_{t-s} F\, dv(s), \\ &\qquad t_i \le s \le t \le t_i + T_p, \\ \tilde{y}(t) &= \int_{t_i}^{t} Cz(s)\, ds + w(t), \quad t_i < t \le t_i + T_p, \\ z(t_i) &= z_{t_i}, \end{aligned} \tag{7}$$

where S_t is a strongly continuous semigroup on $\mathcal{X}$ generated by A on $[t_i, t_i + T_p]$ (see, e.g., [13] for the notion of semigroups and their properties).

A direct consequence of results from [11] is then the following theorem which yields the solution to the MPC/LQG/LTI problem on $[t_i, t_i + T_p]$ for $T_p = \infty$:

Theorem 1. *Under the Assumptions 1, the optimal control and corresponding estimated state for the minimization problem (5) subject to (7) on $[t_i, t_i + T_p]$ are given by*

$$\begin{aligned} u_*(t) &= u_r(t) - R^{-1} B^* \Pi_\infty \big(\hat{x}_*(t) - \bar{x}(t)\big), \\ \hat{x}_*(t) &= S_{t_i} \hat{x}(t_i) + \int_{t_i}^{t} S_{t-s} \Sigma_\infty C^* W^{-1}\, dy(s) + \int_{t_i}^{t} S_{t-s} (f(\bar{x}(s)) - A\bar{x}(s))\, ds, \end{aligned}$$

where S_t is the strongly continuous semigroup generated by

$$A - BR^{-1}B^*\Pi_\infty - \Sigma_\infty C^*W^{-1}C,$$

and Π_∞ and Σ_∞ are the unique nonnegative, self-adjoint solutions of the ARE

$$0 = A^*\Pi + \Pi A - \Pi BR^{-1}B^*\Pi + C^*QC,$$

and the FARE

$$0 = A\Sigma + \Sigma A^* - \Sigma C^*W^{-1}C\Sigma + FVF^*.$$

Note again that the global solution of the LQG problem which is computed for $T_p = \infty$ in the above theorem is only used locally on $[t_i, t_i + T_c]$. In this way, we can implement the infinite-dimensional MPC/LQG controller for any infinite dimensional system satisfying the assumptions made in this section.

Checking the Assumptions 1 for a practical problem is tedious. In [15], as an example the control problem for the Burgers equation

$$\begin{aligned} x_t(t,\xi) &= \nu\, x_{\xi\xi}(t,\xi) - x(t,\xi)\, x_\xi(t,\xi) + b(\xi)u(t), \quad \text{on } (0,T_f] \times (0,1) \\ x(t,0) &= x(t,1) = 0, \quad t \in (0,T_f], \\ x(0,\xi) &= x_0(\xi), \quad \xi \in (0,1), \end{aligned}$$

is considered and it is shown that linearization about a set point leads to an infinite-dimensional LTI system satisfying the Assumptions 1.

For a practical implementation, it is of course necessary to discretize the infinite-dimensional system and to work with a finite-dimensional approximation. If the infinite-dimensional problem is defined via a PDE, this can be achieved using a spatial semi-discretization based on finite differences or finite elements. As both A and its adjoint A^* appear in the formulation of the problem and their discretizations A_h and A_h^T are used to define the finite-dimensional AREs that need to be solved to obtain the approximate feedback and estimator gain matrices K_h and L_h, standard convergence results for finite element discretizations are not sufficient. The necessary conditions for *dual convergence* are stated in [2] for linear parabolic equations with distributed control and are generalized in [8] to boundary control problems. Both papers only consider the LQR problem, the extension to the LQG case is rather straightforward and is executed in [15, Section 8.2.6]. The numerical solution of the resulting large-scale AREs is the computational bottleneck of the suggested control approach. The effective solution of large-scale AREs and associated LQR problems is discussed, e.g., in [1,3,7,20]. The main idea is to apply a Newton-type method to the quadratic nonlinear systems of equations defined by the AREs and to solve the Newton steps by effective iterative methods. It should be noted that for LQG design as discussed here, the actual solution operators/matrices are not necessary as one is only interested in the feedback and estimator gain matrices K and L. Note that the Newton iteration for AREs can be re-written in such a way that one directly iterates on approximations to these operators rather than on approximations to

the ARE solutions. This saves a significant amount of workspace and computational effort and is thus recommended in the context of the suggested MPC/LQG scheme. For an efficient variant of the Newton-Kleinman iteration suitable for large-scale AREs, see [9]. Numerical examples demonstrating the effectiveness of the proposed MPC/LQG feedback control design for PDE-constrained optimization problems are shown in [4] for the Burgers equation and in [5] for a bilinear 3D reaction-diffusion system.

4 Conclusions

We have presented a framework for model predictive control of infinite-dimensional nonlinear systems subject to stochastic perturbations based on an LQG design to implement the optimization step. This includes the state estimation using a Kalman filter. Linearization about the set point leads to an LTI system. We have focused here on the necessary theoretical ingredients to render this step well-posed. Sufficient conditions for convergence of a numerical approximation scheme to implement the LQG design in a computational procedure can be derived, but are not detailed here due to space restrictions. Though stabilization properties of the nonlinear infinite-dimensional MPC/LQG controller have not been shown yet, numerical experiments in [4,5] illustrate the good performance of this control scheme. Further improvements can be obtained if one allows for time-varying linearizations in the optimization step, i.e., linearization around the reference trajectory. The treatment of this case is similar to the LTI case and will be described, together with further numerical experiments, in a forthcoming detailed publication. A convergence and stabilization proof of the infinite-dimensional design based on ideas presented in [16,17,18] is in progress. Further investigations are necessary in order to make the approach real-time feasible. This may require algorithmic improvements in the Riccati solvers or the inclusion of a model reduction strategy in the prediction and optimization step.

References

1. Banks, H., Ito, K.: A numerical algorithm for optimal feedback gains in high dimensional linear quadratic regulator problems. SIAM J. Cont. Optim. 29(3), 499–515 (1991)
2. Banks, H., Kunisch, K.: The linear regulator problem for parabolic systems. SIAM J. Cont. Optim. 22, 684–698 (1984)
3. Benner, P.: Solving large-scale control problems. IEEE Control Systems Magazine 14(1), 44–59 (2004)
4. Benner, P., Görner, S.: MPC for the Burgers equation based on an LGQ design. Proc. Appl. Math. Mech. 6(1), 781–782 (2006)
5. Benner, P., Hein, S.: Model predictive control based on an LQG design for time-varying linearizations. Preprint CSC/09–07, Chemnitz Scientific Computing Preprints, Fakultät für Mathematik, TU Chemnitz (2009), http://nbn-resolving.de/urn:nbn:de:bsz:ch1-201000221

6. Benner, P., Hein, S.: Model predictive control for nonlinear parabolic differential equations based on a linear quadratic Gaussian design. Proc. Appl. Math. Mech. 9(1), 613–614 (2009)
7. Benner, P., Li, J.R., Penzl, T.: Numerical solution of large Lyapunov equations, Riccati equations, and linear-quadratic control problems. Numer. Lin. Alg. Appl. 15(9), 755–777 (2008)
8. Benner, P., Saak, J.: Linear-quadratic regulator design for optimal cooling of steel profiles. Tech. Rep. SFB393/05-05, Sonderforschungsbereich 393 Parallele Numerische Simulation für Physik und Kontinuumsmechanik, TU Chemnitz, 09107 Chemnitz, FRG (2005), http://www.tu-chemnitz.de/sfb393
9. Benner, P., Saak, J.: A Galerkin-Newton-ADI Method for Solving Large-Scale Algebraic Riccati Equations. Preprint SPP1253-090, DFG Priority Program "Optimization with Partial Differential Equations" (SPP1253) (January 2010), http://www.am.uni-erlangen.de/home/spp1253/wiki/index.php/Preprints
10. Camacho, E., Bordons, C.: Model Predictive Control, 2nd edn. Advanced Textbooks in Control and Signal Processing. Springer, London (2004)
11. Curtain, R., Pritchard, T.: Infinite Dimensional Linear System Theory. LNCIS, vol. 8. Springer, New York (1978)
12. Datta, B.: Numerical Methods for Linear Control Systems. Elsevier Academic Press (2004)
13. Engel, K.L., Nagel, R.: One-Parameter Semigroups for Linear Evolution Equations. Springer, New York (2000)
14. Grüne, L., Pannek, J.: Nonlinear Model Predictive Control: Theory and Algorithms. Springer, London (2011)
15. Hein, S.: MPC/LQG-Based Optimal Control of Nonlinear Parabolic PDEs. Ph.D. thesis, TU Chemnitz, Department of Mathematics (March 2010), http://nbn-resolving.de/urn:nbn:de:bsz:ch1-201000134
16. Ito, K., Kunisch, K.: On asymptotic properties of receding horizon optimal control. SIAM J. Cont. Optim. 40, 1455–1472 (2001)
17. Ito, K., Kunisch, K.: Receding horizon optimal control for infinite dimensional systems. ESAIM: Control Optim. Calc. Var. 8, 741–760 (2002)
18. Ito, K., Kunisch, K.: Receding horizon control with incomplete observations. SIAM J. Cont. Optim. 45(1), 207–225 (2006)
19. Kalman, R., Bucy, R.: New results in linear filtering and prediction theory. Trans. ASME, Series D 83, 95–108 (1961)
20. Morris, K., Navasca, C.: Solution of algebraic Riccati equations arising in control of partial differential equations. In: Control and Boundary Analysis. Lect. Notes Pure Appl. Math, vol. 240, pp. 257–280. Chapman & Hall/CRC, Boca Raton (2005)

On an Algorithm for Dynamic Reconstruction in Systems with Delay in Control

Marina Blizorukova*

Ural Federal University and Institute of Mathematics and Mechanics,
Ural Branch of the Russian Academy of Sciences,
Ekaterinburg, 620990 Russia
msb@imm.uran.ru

Abstract. We discuss a problem of the dynamic reconstruction of unknown input controls in nonlinear vector equations. A regularizing algorithm is proposed for reconstructing these controls simultaneously with the processes. The algorithm is stable with respect to informational noises and computational errors.

Keywords: dynamic reconstruction, method of auxiliary models.

1 Introduction Problem Statement

Consider a controlled system described by the following equation

$$\dot{x}(t) = f_1(t, u_t(s), x_t(s)) + f_2(t, x_t(s))u(t) \tag{1}$$

with the initial state

$$u_{t_0}(s) = u_0(s) \in C([-\tau_m^u, 0]; R^{n_1}), \qquad x_{t_0}(s) = x_0(s) \in C([-\tau_n^x, 0]; R^{n_2}). \tag{2}$$

Here t is time from a fixed interval $T = [t_0, \vartheta]$ $(t_0 < \vartheta < +\infty)$; $x(t) = (x_1(t), \ldots, x_{n_2}(t))$ is the phase state of the system; $u(t) = (u_1(t), \ldots, u_{n_1}(t))$ is a control; the symbols $x_t(s)$ and $u_t(s)$ mean the functions $x_t(s) = x(t+s)$ for $s \in [-\tau_n^x, 0]$ and $u_t(s) = u(t+s)$ for $s \in [-\tau_m^u, 0]$, respectively. We assume that initial state (2) is Lipschitz. For simplicity, we assume also that the initial state $x_0(s)$, $u_0(s)$ is fixed and known. The control $u = u(t) = (u_1(t), \ldots, u_{n_1}(t))$ is called an admissible control if its components $u_i(t)$, $i \in [1 : n_1]$, are Lebesgue measurable functions on the interval T and values $u(t)$ belong to a given compact set P from Euclidean space R^{n_1} for almost all $t \in T$. The set of all admissible controls is denoted by $P(\cdot)$. Therefore, $P(\cdot) = \{u(\cdot) \in L_2(T; R^{n_1}) : u(t) \in P$ for a. a. $t \in T\}$. By the trajectory (or the solution) $x(\cdot)$ of equation (1) with initial state (2) corresponding to some admissible control $u(\cdot)$, we call absolutely continuous on T function $x = x(t)$ satisfying (1) for a.a. $t \in T$.

* This work was supported by the Russian Foundation for Basic Research (12-01-00175-a), by the Ural-Siberian Integration Project (12-C-1-1017), and by the Program for support of leading scientific schools of Russia (6512.2012.1).

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 225–234, 2013.

Condition 1. The elements of matrix function

$$f_{2ij}(t, x_t(s)) = f_{2ij}(t, x(t), x(t-\tau_1^x), \ldots, x(t-\tau_n^x)), \quad i \in [1:n_2], \quad j \in [1:n_1],$$

and vector-valued function

$$f_{1i}(t, u_t(s), x_t(s)) =$$

$$= f_{1i}(t, u(t-\tau_1^u), \ldots, u(t-\tau_m^u), x(t), x(t-\tau_1^x), \ldots, x(t-\tau_n^x)), \quad i \in [1:n_2]$$

satisfy the Lipschitz conditions

$$|f_{2ij}(t_1, x_0^{(1)}, x_1^{(1)}, \ldots, x_n^{(1)}) - f_{2ij}(t_2, x_0^{(1)}, x_1^{(2)}, \ldots, x_n^{(2)})| \leq \tag{3}$$

$$\leq C_1(|t_2 - t_1| + \sum_{j=0}^{n} |x_j^{(1)} - x_j^{(2)}|),$$

$$|f_{1i}(t_1, u_1^{(1)}, \ldots, u_m^{(1)}, x_0^{(1)}, x_1^{(1)}, \ldots, x_n^{(1)}) - f_{1i}(t_2, u_1^{(2)}, \ldots, u_m^{(2)}, x_0^{(2)}, x_1^{(2)}, \ldots, x_n^{(2)})|$$

$$\leq d_1(|t_2 - t_1| + \sum_{i=1}^{m} |u_i^{(1)} - u_i^{(2)}| + \sum_{j=0}^{n} |x_j^{(1)} - x_j^{(2)}|). \tag{4}$$

In this case, under this condition for any pair, i.e., for initial state (2) and the control $u(\cdot) \in P(\cdot)$, there exists a unique solution of equation (1).

Let $u(\cdot)$ be an admissible control realizing during the given time interval T; $x(\cdot)$ be the real motion generated by this control. We assume that the phase states $x(\tau_i)$ of the system are inaccurately measured at frequent enough time moments $\tau_i \in T$ in the process. Measurement results $\xi^h(\tau_i) \in R^{n_2}$ satisfy the inequalities

$$|\xi^h(\tau_i) - x(\tau_i)| \leq h. \tag{5}$$

Here, the quantity $h \in (0, 1)$ specifies the measurement error.

In the present paper, we construct an algorithm that reconstructs the control $u(\cdot)$ on the basis of the current information $\xi^h(\cdot)$ in real time. Since the exact reconstruction is impossible due to the error of measurements $\xi^h(\cdot)$ we require that the algorithm should generate some approximation. Namely, it is required to construct an algorithm allowing us, on the basis of the inaccurate measurements $\xi^h(\cdot)$, and in real time, to form the admissible control $v^h(\cdot)$ such that the mean-square deviation of $v^h(\cdot)$ from $u(\cdot)$:, i.e.,

$$|v^h(\cdot) - u(\cdot)|^2_{L_2(T)} = \int_{t_0}^{\vartheta} |v^h(t) - u(t)|^2 \, dt, \tag{6}$$

is arbitrarily small for the sufficiently small measurement error h. Since the measurements are inaccurate it is in general impossible to identify $u(t)$ precisely, therefore the problem is to approximate the input by some function $v^h(t)$.

Here and below, the symbol $|\cdot|$ stands for both the Euclidean norm and the corresponding matrix norm and for the modulo of a number. In what follows, we set $\tau_m^u = \tau_n^x = \tau$ for simplicity, and by $\xi^h(\cdot)$ we denote the function $\xi^h(t)$, $t \in [t_0 - \tau, \vartheta]$ such that $\xi^h(t) = x_0(t - t_0)$ for $t \in [t_0 - \tau, t_0)$, $\xi^h(t) = \xi^h(\tau_i)$ for $t \in [\tau_i, \tau_{i+1})$, $i \in [0 : d-1]$, where $\tau_i = \tau_{h,i}$, $d = d_h$, $\xi^h(\tau_i)$ satisfies (5).

The suggested solution outline is the following ([1–6]). An auxiliary control system (model M) described by equation of the form

$$\dot{w}(t) = F(t, \xi_t^h(s), v_t^h(s)), \quad w_{t_0}(s) = w_0(s), \quad t \in T \tag{7}$$

is associated with the real dynamical system (1). Here the vector $w \in R^{n_2}$ characterizes state of the model, the form of function F is corrected below, vector v^h is control action. After that, the problem of reconstruction of input $u(\cdot)$ is replaced by the problem of positional control of the model. This process is realized on the time interval T in such a way that control $v^h(\cdot)$ "approximates" appropriately $u(\cdot)$. First, one takes a uniform net $\Delta = \{\tau_i\}_{i=0}^m$, $\tau_{i+1} = \tau_i + \delta$, $\delta > 0$, $i \in [0 : m]$, $\tau_0 = 0$, $\tau_m = T$ with the step δ. Then, on the interval $t \in [\tau_i, \tau_{i+1})$ the model is acted upon the controls

$$v_i^h = V_h(\tau_i, w_{\tau_i}(s), \xi_{\tau_i}^h(s)) \tag{8}$$

calculated at the moment τ_i by use of some rule, which hereinafter we shall identify with mapping V_h. Thus, the controls in the model are realized by the method of feedback control. Its value on the interval $[\tau_i, \tau_{i+1}]$ depends on the measurement results $\xi^h(\cdot)$ corresponding to the phase state $x(\cdot)$ of the system (1) and state w of the model (7). The described process forms the piece-wise function

$$v^h(t) = v_i^h, \qquad t \in [\tau_i, \tau_{i+1})$$

in real time synchro with the motion of real system (1). Thus, to solve the problem above, we should specify a model and a control law for this model.

2 Algorithm for Solving the Problem

As a model, we take the following system of linear ordinary differential equation

$$\dot{w}(t) = f_1(\tau_i, v_{\tau_i}^h(s), \xi_{\tau_i}^h(s)) + f_2(\tau_i, \xi_{\tau_i}^h(s))v_i^h + 2(\xi^h(\tau_i) - w(\tau_i)), \tag{9}$$

$$w \in R^{n_2}, \quad t \in [\tau_i, \tau_{i+1}), \quad \tau_i = \tau_{h,i}, \quad v_{t_0}^h(s) = u_0(s),$$

with the initial state $w(t_0) = \xi^h(t_0)$. The solution of this equation $w(\cdot) = w(\cdot; t_0, w_{t_0}(s), v^h(\cdot))$ is understood in the sense of Caratheodory. So, the right-hand side of equation of the model (7) has the form

$$F(t, \xi_t^h(s), v_t^h(s)) = f_1(\tau_i, v_{\tau_i}^h(s), \xi_{\tau_i}^h(s)) + f_2(\tau_i, \xi_{\tau_i}^h(s))v_i^h +$$

$$+2(\xi^h(\tau_i) - w(\tau_i)), \quad t \in [\tau_i, \tau_{i+1}).$$

Introduce the following notation: $\Delta^{(j)} = [t_j, t_{j+1}]$, $t_j = t_0 + \tau_1^x j$; the symbol l stands for the integer part of the number τ/τ_1^x; $j_* = \max\{j : t_j < \vartheta\}$,

$$g_j(h) = h^{(1/3)^j}, \quad j \in [1 : j_*].$$

Fix a partition of the interval T with a step $\delta = \delta(h)$ depending on the measurement error h, i.e.,

$$\Delta_h = \{\tau_{h,i}\}_{i=0}^{d_h}, \quad \tau_i = \tau_{h,i}, \quad \tau_{h,0} = t_0, \quad \tau_{h,d_h} = \vartheta, \tag{10}$$

(for simplicity, we assume that $\tau_i - \tau_{i-1} = \delta = \delta(h)$). Without loss of generality, we can suppose that the partition Δ_h is chosen in such a way that $t_j \in \Delta_h$. Define the law of forming the control v_i^h in the model (for $\tau_i \in [t_j, t_{j+1}) \cap T$) by the relations

$$V_h(\tau_i, w_{\tau_i}(s), \xi_{\tau_i}^h(s)) = V_j(\tau_i, w_{\tau_i}(s), \xi_{\tau_i}^h(s))$$

$$= \arg\min\{2(l_i, f_2(\tau_i, \xi_{\tau_i}^h(s))v) + \alpha_j|v|^2 : v \in P\}. \tag{11}$$

Here α_j is a parameter, $j \in [0 : j_*]$, $l_i = w(\tau_i) - \xi^h(\tau_i)$.

Condition 2. Let $n_2 \geq n_1$, and let there exists a number $c_* > 0$ such that the matrix $f_2(t, x_t(s))$ has a minor of order n_1 with the property: the $n_1 \times n_1$-dimensional matrix $\bar{f}_2(t) = \bar{f}_2(t, x_t(s))$ corresponding to this minor satisfies the inequality

$$|\bar{f}_2(t)u| \geq c_*|u|$$

for each $t \in T$ and all $u \in R^{n_1}$.

We choose the parameter α_j which plays the role of the regularizer, as follows:

$$\alpha_0 = Ch^{2/3}, \quad \alpha_j = Cg_j^{2/3}(h), \quad j \geq 1, \quad C = \text{const} > 0. \tag{12}$$

Let us describe the algorithm for solving the problem above.

Before the initial moment the value h and the partition $\Delta = \Delta_h$ with diameter $\delta = \delta(h)$ are fixed. The work of the algorithm starting at time $t = 0$ is decomposed into $m_h - 1$ steps. At the i-th step carried out during the time interval $\delta_i = [\tau_i, \tau_{i+1})$, $\tau_i = \tau_{h,i}$, the following actions take place. First, at time moment τ_i vector v_i^h is calculated by formula (11). Then the control $v^h(t) = v_i^h$ is fed onto the input of the model (9). After that, we transform the state $w_{\tau_i}(s)$ of the model into $w_{\tau_{i+1}}(s)$. The procedure stops at time ϑ.

The following theorem is true.

Theorem 1. *Let $\delta = \delta(h) \leq h$. Then the inequalities*

$$\nu^{(j)} \equiv |v^h(\cdot) - u(\cdot)|^2_{L_2(\Delta^{(j-1)}; R^{n_1})} \leq c_j g_j(h), \quad j \in [1 : j_*],$$

are valid. Here, $v^h(t) = u(t)$ for $t \in [t_0 - \tau, t_0]$, $v^h(t) = u_0(-\tau)$ for $t \in [t_0 - \tau - \tau_1^u, t_0 - \tau)$.

The proof of the theorem is based on auxiliary statements, which are used in forthcoming considerations. Introduce two systems

$$\dot{p}(t) = f_1(t) + f_2(t)u_1(t), \quad t \in T,$$
$$\dot{q}(t) = F_1(t) + F_2(t)u_2(t),$$

where $p(t)$, $q(t) \in R^n$, $f_1(\cdot)$, $F_1(\cdot) \in L_2(T; R^n)$, $f_2(\cdot) \in L_2(T; R^{n\times r})$, $F_2(\cdot) \in L_2(T; R^{n\times r})$, $u_1(\cdot)$, $u_2(\cdot) \in L_2(T; R^r)$, $|u_l(\cdot)|_{L_\infty(T;R^r)} \le K$, $l = 1, 2$.

Introduce the notation: $\Delta_*^{(j)} = [t_j^*, t_{j+1}^*] \cap T$, $t_j^* = t_0 + \tau_* j$, $j \in [0 : j_0]$, $\Delta^{(-1)} = [t_0 - \tau_*, t_0]$, $\tau_* = \text{const} \in (0, \vartheta - t_0)$, $j_0 = \max\{j : t_j^* \le \vartheta\}$. Let $r \le n$ and let there exists a number $c > 0$ such that the matrix $f_2(t)$ has a minor of order r such that the $r \times r$-matrix $\bar{f}_2(t)$ corresponding to this minor satisfies the following inequality: $|\bar{f}_2(t)u| \ge c|u|$ for each $t \in T$ and all $u \in R^r$.

It is easy to verify the following lemmas.

Lemma 1. *Let the function $t \to (\bar{f}_2(t))^{-1}u_1(t)$ be a function of bounded variation on T and let the conditions*

$$|f_1(\cdot) - F_1(\cdot)|^2_{L_2(\Delta_*^{(j)};R^n)} \le a_1^{(j)}, \quad |f_2(\cdot) - F_2(\cdot)|^2_{L_2(\Delta_*^{(j)};R^{n\times r})} \le a_2^{(j)},$$

$$|p(t) - q(t)|^2 + \tilde{\alpha}_j \int_{t_j^*}^{t} \{|u_2(\nu)|^2 - |u_1(\nu)|^2\}\, d\nu \le a_3^{(j)} \qquad t \in [t_j^*, t_{j+1}^*],$$

$$|p(t_j^*) - q(t_j^*)|^2 \le a_4^{(j)}, \quad \tilde{\alpha}_j = \text{const} \in (0, +\infty)$$

be true. Then the inequality

$$|u_1(\cdot) - u_2(\cdot)|^2_{L_2(\Delta_*^{(j)};R^r)} \le K_j\{\sum_{l=1}^{4}(a_l^{(j)})^{1/2} + \tilde{\alpha}_j^{1/2}\} + a_3^{(j)}/\tilde{\alpha}_j$$

is valid.

Lemma 2. *The bunches of solutions of systems (1) and (9) are bounded in the space $W^{1,\infty}(T; R^{n_2}) = \{x(\cdot) \in L_2(T; R^{n_2}); \dot{x}(\cdot) \in L_2(T; R^{n_2})\}$.*

We use the relation

$$\varepsilon_j(t) = |x(t) - w(t)|^2 + \alpha_j \int_{t_j}^{t} \{|v^h(\nu)|^2 - |u(\nu)|^2\}\, d\nu, \quad j \in [0 : j_*], \quad t \in T.$$

Lemma 3. *The following inequalities*

$$\varepsilon_j(t) \le b_j, \quad t \in \Delta^{(j)} \cap T, \quad j \in [0 : j_*],$$

are valid, where

$$b_j = |x(t_j) - w(t_j)|^2 + c_j^{(1)}(h + \delta) + c_j^{(2)} \sum_{k=j-l}^{j} \nu^{(k)},$$

$c_j^{(1)}$, $c_j^{(2)}$ are some constants, which can be explicitly written.

Proof. Fix $\tau_i \in \Delta^{(j)}$. Then for $t \in \Delta^{(j)} \cap \delta_i = [\tau_i, \tau_{i+1}]$, we obtain

$$\varepsilon_j(t) \le \varepsilon_j(\tau_i) + \sum_{j=1}^{4} \Lambda_{ji}(t), \tag{13}$$

where

$$\Lambda_{1i}(t) = 2(s_i, \int_{\tau_i}^{t} \{f_1(\nu, u_\nu(s), x_\nu(s)) - f_1(\tau_i, v_\nu^h(s), \xi_{\tau_i}^h(s))\}\, d\nu), \;\; s_i = x(\tau_i) - w(\tau_i),$$

$$\Lambda_{2i}(t) = 2(s_i, \int_{\tau_i}^{t} \{f_2(\nu, x_\nu(s))u(\nu) -$$

$$- f_2(\tau_i, \xi_{\tau_i}^h(s))v_i^h\}\, d\nu) + \alpha_j \int_{\tau_i}^{t} \{|v^h(\nu)|^2 - |u(\nu)|^2\}\, d\tau,$$

$$\Lambda_{3i}(t) = -2(t-\tau_i)(s_i, \xi^h(\tau_i) - w(\tau_i)), \quad \Lambda_{4i}(t) = (t-\tau_i) \int_{\tau_i}^{t} |\dot{w}(\tau) - \dot{x}(\tau)|^2\, d\tau.$$

By virtue of lemma 2, we have

$$\Lambda_{4i}(t) \le K_*^{(j)} (t-\tau_i)^2, \quad t \in \delta_i. \tag{14}$$

Note that $v^h(\tau_i + s) = v^h(t+s)$ for $s \ge t_0 - \tau_i$, $t \in [\tau_i, \tau_{i+1}]$ and in addition

$$|\xi^h(\tau_i + s) - x(t+s)| \le K_*(h + t - \tau_i) \quad \text{for} \quad \tau_i + s \ge t_0 - \tau. \tag{15}$$

Taking into account lemma 2, as well as the Lipschitz property of the functions $u_0(s)$ and $x_0(s)$, inequalities (4) and the relation

$$|\xi^h(\tau_i + s) - x(t+s)| \le K_*(h + t - \tau_i) \quad \text{for} \quad \tau_i + s \ge t_0 - \tau, \tag{16}$$

we obtain for $t \in \delta_i$ the estimate

$$\int_{\tau_i}^{t} |f_1(\nu, u_\nu(s), x_\nu(s)) - f_1(\tau_i, v_\nu^h(s), \xi_{\tau_i}^h(s))|\, d\nu \le$$

$$\le K_1^{(j)}(t-\tau_i)(h + t - \tau_i) + K_2^{(j)}(t-\tau_i)^{1/2} \sum_{k=1}^{m} \Big(\int_{\tau_i - \tau_k^u}^{t-\tau_k^u} |u(\nu) - v^h(\nu)|^2\, d\nu \Big)^{1/2}.$$

Here, $\tau_0^x = 0$. In this case, the inequality

$$\Lambda_{1i}(t) \le 2(t-\tau_i)|x(\tau_i) - w(\tau_i)|^2 + K_3^{(j)} \{(t-\tau_i)(h + t - \tau_i)^2 +$$

$$+\sum_{k=1}^{m}\int_{\tau_i-\tau_k^u}^{t-\tau_k^u}|u(\nu)-v^h(\nu)|^2\,d\nu\} \tag{17}$$

holds for $t\in\delta_i$. In view of (5), we have

$$\Lambda_{3i}(t)\le -2(t-\tau_i)|x(\tau_i)-w(\tau_i)|^2+K_4^{(j)}h(t-\tau_i),\quad t\in\delta_i. \tag{18}$$

Moreover, from (5), (3), and (16), we derive

$$|f_2(\nu,x_\nu(s))u(\nu)-f_2(\tau_i,\xi^h_{\tau_i}(s))u(\nu)|\le K_0(h+\nu-\tau_i)$$

for $\nu\in[\tau_i,\tau_{i+1}]$. In this case,

$$\Lambda_{2i}(t)\le K_5^{(j)}(t-\tau_i)(h+t-\tau_i)+$$

$$+\int_{\tau_i}^{t}\{2(l_i,f_2(\tau_i,\xi^h_{\tau_i}(s))\{v_i^h-u(\nu)\}+\alpha_j\{|v_i^h|^2-|u(\nu)|^2\}\}\,d\nu.$$

The rule for forming the control v_i^h (11) and the last inequality imply

$$\Lambda_{2i}(t)\le K_5^{(j)}(t-\tau_i)(h+t-\tau_i). \tag{19}$$

Finally, taking into account (13)–(19), we conclude that for $t\in\Delta^{(j)}\cap\delta_i$

$$\varepsilon_j(t)\le\varepsilon_j(\tau_i)+K_6^{(j)}\delta(h+\delta)+K_3^{(j)}\sum_{k=1}^{m}\int_{\tau_i-\tau_k^u}^{t-\tau_k^u}|u(\nu)-v^h(\nu)|^2\,d\nu,$$

i.e., for $t\in\Delta^{(j)}=[t_j,t_{j+1}]$,

$$\varepsilon_j(t)\le\varepsilon_j(t_j)+K_7^{(j)}(h+\delta)+K_8^{(j)}\int_{t_j-\tau}^{t_{j+1}-\tau_1^u}|u(\nu)-v^h(\nu)|^2\,d\nu.$$

Note that $\tau=l\tau_1^u+\gamma$, $\gamma\ge 0$. Therefore, $t_{j+1}-\tau_1^u=t_j$, $t_{j-l-1}\le t_j-\tau\le t_{j-l}$. In this case, for $t\in\Delta^{(j)}$ we have

$$\varepsilon_j(t)\le\varepsilon_j(t_j)+K_7^{(j)}(h+\delta)+K_9^{(j)}\sum_{k=j-l}^{j}\nu^{(k)}.$$

Here, constants $K_k^{(j)}$, $k\in[0:9]$ are written explicitly. Thus, one can assume that $c_j^{(1)}=K_7^{(j)}$ and $c_j^{(2)}=K_9^{(j)}$. The lemma is proved.

Lemma 4. *Let $\delta\le h$ and values α_j be given by (12). Then the inequalities*

$$\nu^{(j)}\le c_j g_j(h), \tag{20}$$

$$b_j\le c_j^{(0)}g_j(h) \tag{21}$$

are valid.

Proof. For simplicity, set $t_{j_*+1} = \vartheta$. By virtue of lemma 3, we have for $t \in \Delta^{(j)}$

$$|x(t)-w(t)| \le \Big(\varepsilon_j(t)+\alpha_j \int\limits_{t_j}^{t} \{|v^h(\nu)|^2+|u(\nu)|^2\}\, d\nu\Big)^{1/2} \le \Big(b_j+\alpha_j\rho_A\Big)^{1/2}, \quad (22)$$

where $\rho_A = 2\tau_* d^2(P)$ and $d(P) = \sup\{|u| : u \in P\}$. Taking into account the inclusion $t_j \in \Delta_h$, we conclude that for any $j \in [0 : j_*]$, one can specify the number $i = i_j(h)$ such that $t_j = \tau_{i_j(h)}$. Introduce the notation $\varrho_j \equiv |f_1(\cdot) - F_1(\cdot)|^2_{L_2(\Delta^{(j)};R^{n_2})}$. In this case, by virtue of lemma 2, as well as of (4) and (16), we obtain

$$\varrho_j \le d_j^{(1)} \sum_{i=i_j(h)}^{i=i_{j+1}(h)-1} \int\limits_{\tau_i}^{\tau_{i+1}} \{\delta^2 + h^2 + \gamma^h(\nu) + \gamma_i^h(\nu) + |\xi^h(\tau_i) - w(\tau_i)|^2\}\, d\nu,$$

where

$$\gamma^h(\nu) = \sum_{k=1}^{m} |u(\nu-\tau_k^u) - v^h(\nu-\tau_k^u)|^2, \quad \gamma_i^h(\nu) = \sum_{k=0}^{n} |x(\nu-\tau_k^x) - \xi^h(\tau_i-\tau_k^x)|^2.$$

Note that

$$\int\limits_{t_j}^{t_{j+1}} \gamma^h(\nu)\, d\nu \le d_j^{(2)} \int\limits_{t_{j-l-1}}^{t_j} |u(\nu) - v^h(\nu)|^2\, d\nu = d_j^{(2)} \sum_{k=j-l}^{j} \nu^{(k)}, \quad (23)$$

$$\int\limits_{t_j}^{t_{j+1}} \gamma_i^h(\nu)\, d\nu \le d_j^{(3)}(h^2+\delta^2). \quad (24)$$

In addition,

$$\nu^{(k)} = 0 \qquad k \in [-l : 0]. \quad (25)$$

Therefore, combining inequalities (22)–(24), we obtain the estimates

$$\varrho_j \le d_j^{(5)}\{h^2+\delta^2 + \sum_{k=j-l}^{j} \nu^{(k)} + b_j + \alpha_j\}, \quad j \in [0 : j_*]. \quad (26)$$

One can easily see that the following estimates also hold:

$$|f_2(\cdot) - F_2(\cdot)|^2_{L_2(\Delta^{(j)};R^{n_2\times n_1})} \le d_j^{(5)}(h^2+\delta^2), \quad j \in [0 : j_*]. \quad (27)$$

Here $d_j^{(1)}$–$d_j^{(5)}$ are some constants, which can be explicitly written. By lemma 3, (22), and (25), for $\delta \le h$, we have the inequalities

$$\varepsilon_0(t) \le b_0 \le c_0^* h, \quad t \in \Delta^{(0)}, \quad (28)$$

$$|x(t_1) - w(t_1)|^2 \le \rho_A \alpha_0 + c_0^* h \le c_* h^{2/3}. \tag{29}$$

Taking into account (25)–(28), for $h \in (0,1)$, we obtain

$$\varrho_0 \le d_0^{(1)}\{h^2+\delta^2+b_0+h^{2/3}\} \le d_0^* h^{2/3}, \qquad |f_2(\cdot)-F_2(\cdot)|^2_{L_2(\Delta^{(0)};R^{n_2\times n_1})} \le c_j^{(*)} h^2.$$

By virtue of condition 1, one can use lemma 1. Set $p = x$, $q = w$, $u_1 = u$, $u_2 = v^h$, $f_1(t) = f_1(t, u_t(s), x_t(s))$, $f_2(t) = f_2(t, x_t(s))$, $F_1(t) = f_1(\tau_i, v^h_{\tau_i}(s), \xi^h_{\tau_i}(s)) + 2(\xi^h(\tau_i) - w(\tau_i))$, $F_2(t) = f_2(\tau_i, \xi^h_{\tau_i}(s))$ $t \in [\tau_i, \tau_{i+1})$. Then, assuming $a_1^{(0)} = d_0^* h^{2/3}$, $a_2^{(0)} = c_j^{(*)} h^2$, $a_3^{(0)} = c_0^* h$, $a_4^{(0)} = c_* h^{2/3}$, $\tilde\alpha_0 = \alpha_0 = ch^{2/3}$, we have

$$\nu^{(1)} = |u(\cdot) - v^h(\cdot)|^2_{L_2(\Delta^{(0)};R^{n_1})} \le \tilde c_1 h^{1/3} = c_1 g_1(h). \tag{30}$$

It means that inequality (20) holds for $j = 1$. Further, by using (29) and (30), we deduce that

$$b_1 = |x(t_1) - w(t_1)|^2 + c_1^{(1)}(h+\delta) + c_1^{(2)} \sum_{k=1-l}^{1} \nu^{(k)} \le \tilde c_1^{(0)} h^{1/3} = c_1^{(0)} g_1(h).$$

Inequality (21) for $j = 1$ is also verified. It follows from (22) that

$$|x(t_j) - w(t_j)|^2 \le b_{j-1} + \rho_A \alpha_{j-1}, \quad j \in [1 : j_* - 1]. \tag{31}$$

Consequently, in view of relations (31), as well as of the rule for definition b_j, we have the inequality

$$b_j \le b_{j-1} + d_j\Big(h + \alpha_{j-1} + \sum_{k=j-l}^{j} \nu^{(k)}\Big), \quad d_j = \text{const} \in (0, +\infty). \tag{32}$$

Setting $a_1^{(j)} = d_j^{(4)}\{h^2 + \delta^2 + \sum_{k=j-l}^{j} \nu^{(k)} + a_3^{(j)} + \alpha_j\}$, $a_3^{(j)} = b_j$, $a_2^{(j)} = d_j^{(5)}(h^2 + \delta^2)$, $a_4^{(j)} = b_{j-1} + \rho_A \alpha_{j-1}$, $j \in [1 : j_*]$ for $j \ge 1$ in lemma 1 and taking into account inequalities (32), we obtain

$$\nu^{(j+1)} \le c^{(j)}\{h^{1/2} + \Big(\sum_{k=j-l}^{j} \nu^{(k)}\Big)^{1/2} + b_{j-1}^{1/2} + \alpha_{j-1}^{1/2} + \alpha_j^{1/2}\} + b_j \alpha_j^{-1}, \ j \in [1 : j_*].$$

Here, we used lemma 3 and inequalities (27), (28), and (31)) for choosing values $a_i^{(j)}$. Now, to proof inequalities (20) and (21), one can use the proof by induction. The lemma is proved.

3 Example

The algorithm was tested by a model example. The following system

$$\begin{aligned} \dot x_1(t) &= 2x_1(t-1) + u(t) \\ \dot x_2(t) &= x_2(t-1) + x_1(t) + u(t-1), \quad t \in T = [0, 2], \end{aligned} \tag{33}$$

with initial conditions $x_0(s) = y_0(s) = 1$, $u(s) = 0$ for $s \in [-1, 0]$ and control $u(t) = t$ was considered. The solution $x(t) = \{x_1(t), x_2(t)\}$ of system (33) was calculated analytically. During the experiment, we assumed that $\xi^h(\tau_i) = x_1(\tau_i) + h$. As a model, we took the system (9), which has the form

$$\begin{aligned} \dot{w}^{(0)}(t) &= 2\xi_1^h(\tau_i - 1) + v_i^h + 2(\xi_1^h(\tau_i) - w^{(0)}(\tau_i)) \quad \text{for} \quad t \in [\tau_i, \tau_{i+1}) \\ \dot{w}^{(1)}(t) &= \xi_2^h(\tau_i - 1) + \xi_1^h(\tau_i) + v^h(\tau_i - 1) + 2(\xi_2^h(\tau_i) - w^{(1)}(\tau_i)), \end{aligned} \tag{34}$$

with the initial condition $w^{(0)}(s) = w^{(1)}(s) = 1$, for $s \in [-1, 0]$. Here $v^h(\tau_i) = v_i^h$ for $t \in [\tau_i, \tau_{i+1})$, $i \geq 0$, $v^h(s) = 0$ for $s \in [-1, 0)$. The controls v_i^h in model (34) were calculated by the following formula (see (11))

$$v_i^h = \arg\min\{2l_i v + \alpha_j |v|^2 : |v| \leq K\},$$

where $l_i = w^{(0)}(\tau_i) - \xi_1^h(\tau_i)$.

In figures 1 and 2 the results of calculations are presented for the case when $\delta = 10^{-4}$, $\alpha_0 = Ch^{2/3}$, $\alpha_1 = Ch^{2/9}$, $C = 0.2$, $K = 10$. Fig. 1 corresponds to the case when $h = 0.001$, fig. 2 — $h = 0.02$. In these figures the solid (dashed) lines represent the model control $v^h(\cdot)$ (the real control $u(\cdot)$). The equations were solved by the Euler method with step δ.

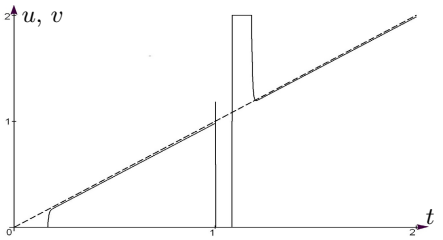


Fig. 1.

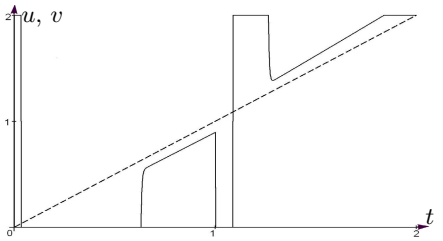


Fig. 2.

References

1. Osipov, Y.S., Kryazhimskii, A.V.: Inverse Problems for Ordinary Differential Equations: Dynamical Solutions. Gordon and Breach, London (1995)
2. Maksimov, V.I.: Dynamical Inverse Problems of Distributed Systems. VSP, Utrecht–Boston (2002)
3. Maksimov, V., Pandolfi, L.: Dynamical reconstruction of inputs for constraction semigroup systems: the boudary input case. J. Optim. Theor. Appl. 103, 401–420 (1999)
4. Maksimov, V., Troltzsch, F.: Dynamical state and control reconstruction for a phase field model. Dynamics of Continuous, Discrete and Impulsive Systems. A: Mathematical Analysis 13(3-4), 419–444 (2006)
5. Osipov, Y.S., Kryazhimskii, A.V., Maksimov, V.I.: Methods of Dinamical Reconstruction of Inputs of Controlled Systems. Ekaterinburg (2011) (in Russian)
6. Maksimov, V.: Lyapunov function method in input reconstruction problems of systems with aftereffect. J. Math. Sci. 140(6), 832–849 (2007)

Computation of Value Functions in Nonlinear Differential Games with State Constraints

Nikolai Botkin, Karl-Heinz Hoffmann, Natalie Mayer, and Varvara Turova*

Technische Universität München, Center for Mathematics,
Boltzmannstr. 3, 85748 Garching, Germany
{botkin,hoffmann,turova}@ma.tum.de,
mayer@lcc.mw.tum.de

Abstract. Finite-difference schemes for the computation of value functions of nonlinear differential games with non-terminal payoff functional and state constraints are proposed. The solution method is based on the fact that the value function is a generalized viscosity solution of the corresponding Hamilton-Jacobi-Bellman-Isaacs equation. Such a viscosity solution is defined as a function satisfying differential inequalities introduced by M. G. Crandall and P. L. Lions. The difference with the classical case is that these inequalities hold on an unknown in advance subset of the state space. The convergence rate of the numerical schemes is given. Numerical solution to a non-trivial three-dimensional example is presented.

Keywords: Differential games, non-terminal payoff functionals, state constraints, value functions, viscosity solutions, finite-difference schemes.

1 Introduction

Numerical methods for solving differential games (see [1,2,3] for concepts) are intensively developed during two or three last decades. We consider control systems with nonlinear dynamics, non-terminal payoff functionals, and state constraints. Our approach is based on the approximation of viscosity solutions of the Hamilton-Jacobi-Bellman-Isaacs equation associated with the considered differential game.

In [4], a pair of differential inequalities determining the value function of nonlinear differential games with non-terminal payoff functionals was introduced. Additionally, the directional differentiability of the value function was required. In [5], such a requirement was relaxed, and the results were stated in terms of upper and lower directional derivatives. At the same time, the concept of viscosity solutions for Hamilton-Jacobi equations was proposed in [6] and [7]. Further investigations [8] showed that the inequalities for the upper and lower directional derivatives are equivalent to the inequalities defining viscosity solutions.

* Corresponding author.

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 235–244, 2013.

Grid methods based on vanishing viscosity techniques for finding viscosity solutions of Hamilton-Jacobi equations were suggested in [9]. In [10], an abstract operator that generates approximate solutions was introduced, and the uniform convergence of approximate solutions to a viscosity solution was proved. A representation of this operator in terms of differential game theory was given in [11]. The results of [10] and [11] cover differential games with the payoff functional

$$\gamma_1(x(\cdot)) = \chi(T, x(T)). \tag{1}$$

In [12], the approach of [10] and [11] was extended to differential games with more general (non-terminal) payoff functionals of the form

$$\gamma_2(x(\cdot)) = \min_{t\in[t_0,T]} \chi(t, x(t)), \tag{2}$$

where t_0 is the starting time, T the termination time, $x(\cdot)$ a trajectory of the controlled system, and χ a given function.

In the present paper, differential games with payoff functionals of the form

$$\gamma_3(x(\cdot)) = \max\{\min_{t\in[t_0,T]} \chi(t, x(t)), \max_{t\in[t_0,T]} \theta(t, x(t))\}, \tag{3}$$

where χ and θ are given functions, satisfying the relation $\chi(t,x) \geq \theta(t,x)$ for all t and x, are considered. As it will be seen later, the first part of functional (3), $\min_{t\in[t_0,T]} \chi(t, x(t))$, is responsible for the quality of the process, and the second part, $\max_{t\in[t_0,T]} \theta(t, x(t))$, accounts for state constraints. In the following, differential inequalities defining viscosity solutions in the case of payoff functional (3) will be formulated and compared with those related to payoff functionals (2) and (1). A finite difference scheme based on a modified abstract operator that generates approximations of viscosity solutions is presented, and an example of computation of value function for a three-dimensional problem originated from the famous isotropic rocket game introduced in [1] is given.

2 Statement of the Problem

Consider a collision-avoidance differential game with the dynamics

$$\dot{x} = f(t, x, \alpha, \beta), \quad t \in [0, T], \quad x \in \mathbb{R}^n, \quad \alpha \in A \subset \mathbb{R}^\mu, \quad \beta \in B \subset \mathbb{R}^\nu, \tag{4}$$

where t is time; $x = (x_1, ..., x_n)$ the state vector; α, β are control parameters of the players; and A, B are given compacts. The game starts at $t = t_0$ and finishes at $t = T$. The first player, control parameter α, strives to bring the trajectories of system (4) to a target set given by

$$M := \{(t, x) : t \in [0, T], \chi(t, x) \leq 1\}$$

within the time $[t_0, T]$. The objective of the second player, control parameter β, is opposite. Besides, the trajectories should remain in a state constraint set given by

$$N := \{(t, x) : t \in [0, T], \theta(t, x) \leq 1\}.$$

Here, $\chi : [0,T] \times \mathbb{R}^n \to \mathbb{R}$ and $\theta : [0,T] \times \mathbb{R}^n \to \mathbb{R}$ are some given functions such that $\chi(t,x) \geq \theta(t,x)$ for all t and x so that $M \subset N$ holds.

We extend this differential game by considering the payoff functional (3) being minimized by the first player and maximized by the second one. It is easily seen that the value function of such an extended problem gives a solution to the collision-avoidance differential game. In fact, if the value function of the differential game (4) and (3) is less than or equal to 1 at the starting position of the extended game, then there exists a strategy of the first player such that, for all strategies of the second player and all trajectories $x(\cdot)$, two conditions hold:

(a) $\min_{t \in [t_0,T]} \chi(t,x(t)) \leq 1$ (the position $(t,x(t))$ arrives at the target set M at some time instant $t \leq T$),

(b) $\max_{t \in [t_0,T]} \theta(t,x(t)) \leq 1$ (the position $(t,x(t))$ remains in the state constraint set N for all $t \in [t_0,T]$).

Let the extended game is formalized as in [2,3,4]. That is, the players use feedback strategies which are arbitrary functions

$$\mathcal{A} : [0,T] \times \mathbb{R}^n \to A, \quad \mathcal{B} : [0,T] \times \mathbb{R}^n \to B.$$

For any initial position $(t_0,x_0) \in [0,T] \times \mathbb{R}^n$ and any strategies $\mathcal{A}$ and $\mathcal{B}$, two functional sets $X_1(t_0,x_0,\mathcal{A})$ and $X_2(t_0,x_0,\mathcal{B})$ are defined. These sets consist of the limits of the step-by-step solutions of (4) generated by the strategies $\mathcal{A}$ and $\mathcal{B}$, respectively (see [2,3,4]).

We assume that the function f is uniformly continuous, bounded and Lipschitzian in t and x on $[0,T] \times \mathbb{R}^n \times A \times B$; the functions χ and θ are bounded and Lipschitzian in t,x; and the following saddle point condition holds:

$$H(t,x,p) := \max_{\beta \in B} \min_{\alpha \in A} \langle p, f(t,x,\alpha,\beta) \rangle = \min_{\alpha \in A} \max_{\beta \in B} \langle p, f(t,x,\alpha,\beta) \rangle$$

for any $p \in \mathbb{R}^n$, $(t,x) \in [0,T] \times \mathbb{R}^n$.

It is proved in [4,13] that the differential game (3,4) has a value function $c : (t,x) \to c(t,x)$ defined by the relation

$$c(t,x) = \min_{\mathcal{A}} \max_{\substack{x(\cdot) \in \\ X_1(t,x,\mathcal{A})}} \gamma_3(x(\cdot)) = \max_{\mathcal{B}} \min_{\substack{x(\cdot) \in \\ X_2(t,x,\mathcal{B})}} \gamma_3(x(\cdot)).$$

Thus, the upper value of the game coincides with the lower one for all $(t,x) \in [0,T] \times \mathbb{R}^n$. The value function is bounded and Lipschitzian in t,x on $[0,T] \times \mathbb{R}^n$.

3 Viscosity Solutions

We formulate differential inequalities defining the value function of the differential game (3,4) and compare them with corresponding differential inequalities for the value functions of differential games (2,4) and (1,4).

Proposition 1. *A Lipschitz function c is the value function of differential game (3) and (4) if and only if:*

(i) for any $(t,x) \in [0,T] \times \mathbb{R}^n$, $c(T,x) = \chi(T,x)$ and $\theta(t,x) \le c(t,x) \le \chi(t,x)$;

(ii) for any point $(s_0,y_0) \in [0,T] \times \mathbb{R}^n$ such that $c(s_0,y_0) \le \chi(s_0,y_0)$ and any function $\varphi \in \mathbb{C}^1$ such that $c-\varphi$ attains a local minimum at (s_0,y_0), the following inequality holds

$$\frac{\partial \varphi}{\partial t}(s_0,y_0) + H(s_0,y_0,\frac{\partial \varphi}{\partial y}(s_0,y_0)) \le 0; \tag{5}$$

(iii) for any point $(s_0,y_0) \in [0,T] \times \mathbb{R}^n$ such that $c(s_0,y_0) \ge \theta(s_0,y_0)$ and any function $\varphi \in \mathbb{C}^1$ such that $c-\varphi$ attains a local maximum at (s_0,y_0), the following inequality holds

$$\frac{\partial \varphi}{\partial t}(s_0,y_0) + H(s_0,y_0,\frac{\partial \varphi}{\partial y}(s_0,y_0)) \ge 0. \tag{6}$$

The proof of Proposition 1 is given in [14].

Remark 1. If the relation $\theta(t,x) \le c(t,x)$ in (i) and the condition $c(s_0,y_0) \ge \theta(s_0,y_0)$ in (iii) are omitted, relations (i)-(iii) define the value function of differential game (2,4) (see [12]). If, additionally, the relation $c(t,x) \le \chi(t,x)$ in (i) and the condition $c(s_0,y_0) \le \chi(s_0,y_0)$ in (ii) are omitted, relations (i)-(iii) define the value function of differential game (1,4).

Remark 2. We call a Lipschitz function c satisfying relations (i)-(iii) of Proposition 1 a generalized viscosity solution of the Hamilton-Jacobi equation

$$c_t + H(t,x,c_x) = 0.$$

Thus, a generalized solution exists and is unique.

4 Finite-Difference Schemes

In this section, an upwind operator (see [15] for the idea) is introduced, and finite-difference schemes based on this operator are described.

Let $\rho, h_1, ..., h_n$ be time and space discretization step sizes. The upwind operator F is defined as follows:

$$F(c;t,\rho,h_1,...,h_n)(x) = c(x) + \rho \max_{\beta \in B} \min_{\alpha \in A} \sum_{i=1}^{n} (p_i^R f_i^+ + p_i^L f_i^-),$$

where $f_i = f_i(t,x,\alpha,\beta)$ are the right hand sides of the control system, and

$$a^+ = \max\{a,0\}, \quad a^- = \min\{a,0\},$$

$$p_i^R = [c(x_1,...,x_i+h_i,...,x_n) - c(x_1,...,x_i,...,x_n)]/h_i,$$

$$p_i^L = [c(x_1,...,x_i,...,x_n) - c(x_1,...,x_i-h_i,...,x_n)]/h_i.$$

Remark 3. Note that, if ρ is fixed, the time step operator can be restricted to functions defined on rectangular grids with the step size h_i in ith coordinate, $i = \overline{1,n}$. Therefore, this operator will yield fully discrete finite difference schemes when used in the approximation procedure considered below.

Let $\mathcal{M} = T/\rho + 1$. Denote $t_m = m\rho$, $m = 0, ..., \mathcal{M}$, and introduce the following notation:

$$c^m(x_{i_1}, ..., x_{i_n}) = c(t_m, i_1h_1, ..., i_nh_n),$$
$$\chi^m(x_{i_1}, ..., x_{i_n}) = \chi(t_m, i_1h_1, ..., i_nh_n),$$
$$\theta^m(x_{i_1}, ..., x_{i_n}) = \theta(t_m, i_1h_1, ..., i_nh_n).$$

In the case of functional (1), the finite-difference scheme be

$$c^{m-1} = F(c^m; t_m, \rho, h_1, ..., h_n),\ c^{\mathcal{M}} = \chi^{\mathcal{M}}. \tag{7}$$

In the case of functional (2), it is modified as follows:

$$c^{m-1} = \min\{F(c^m; t_m, \rho, h_1, ..., h_n), \chi^m\},\ c^{\mathcal{M}} = \chi^{\mathcal{M}}. \tag{8}$$

When the state constraint is presented, i.e. the functional (3) is considered, the numerical scheme be

$$c^{m-1} = \max\left\{\min\{F(c^m; t_m, \rho, h_1, ..., h_n), \chi^m\}, \theta^m\right\},\ c^{\mathcal{M}} = \chi^{\mathcal{M}}. \tag{9}$$

The following convergence result holds.

Theorem 1. *Let M be a bound of the right hand side of system (4). If $\frac{\rho}{h_i} \leq \frac{1}{M\sqrt{n}}$, then the grid functions obtained by the procedures (7), (8), and (9) converge point-wise to the value functions of games (1,4), (2,4), and (3,4), respectively, as $\rho \to 0$, $h_i \to 0$, and the convergence rate is $\max(\sqrt{\rho}, \max_i \sqrt{h_i})$.*

The proof of the Theorem is given in [14] and [16].

5 Example

It should be noted that high-dimensional computation ($n \geq 3$) of value functions of nonlinear differential games with state constraints is a very difficult problem. Since about fifteen years, several groups are working on appropriate numerical methods (see e.g. [17,18,19,20,21]), but only few three-dimensional problems are solved numerically. The following example deals with a very famous unsolved problem.

In the PhD thesis by Pierre Bernhard [22] and in paper [23] by Joseph Lewin and Geert Jan Olsder, a pursuit-evasion game deduced from the game of isotropic rockets [1] is considered:

$$\begin{aligned} \dot{x} &= -\frac{Wy}{V_p}\sin\phi + V_e \sin\psi, \\ \dot{y} &= \frac{Wx}{V_p}\sin\phi + V_e\cos\psi - V_p, \\ \dot{V}_p &= W\cos\phi. \end{aligned} \tag{10}$$

Here, x and y are the coordinates of the evader (E) in the moving reference system whose origin is at the position of the pursuer (P), and the axis y is directed along the velocity of P; W is the magnitude of the acceleration of P; V_p the magnitude of the velocity of P; ϕ, control of P, the angle between the vectors of the acceleration and velocity of P (we assume that $-\pi/2 \le \phi \le \pi/2$, i.e. $\dot{V}_p \ge 0$); V_e the magnitude of the velocity of E; ψ, control of E, the angle between the velocity vector of E and the direction of y-axis ($0 \le \psi \le 2\pi$).

The target set is a a cylinder

$$M = \{(x, y, V_p) : x^2 + y^2 \le 0.3^2\}, \tag{11}$$

and the state constraint set is given by

$$N = \{(x, y, V_p) : a \le V_p \le b\}, \tag{12}$$

where a and b are positive numbers, which will be specified later.

It is observed in [23] that the classical homicidal chauffeur game [1] can be deduced from (10). In fact, permitting only bang-bang controls of the pursuer, $\phi = \pm\pi/2$, implies that $\cos\phi = 0$, and therefore $V_p = const$. Introduce a new control parameter $u = \sin\phi$ of the pursuer and allow it to assume values from the interval $[-1, 1]$ because the control system is linear with respect to u. Moreover, set $W \equiv 1$, $V_p \equiv 1$, $V_e \equiv 0.3$, and introduce new control parameters, $v_1 = V_e \sin\psi$ and $v_2 = V_e \cos\psi$, of the evader. Reduce the target set (11) to $M = \{(x, y) : x^2 + y^2 \le 0.3^2\}$. This yields the classical homicidal chauffeur game

$$\dot{x} = -yu + v_1, \quad \dot{y} = xu + v_2 - 1, \quad |u| \le 1, \quad \sqrt{v_1^2 + v_2^2} \le 0.3 \tag{13}$$

whose numerical solutions are known and can be used for the verification of computations applied to problem (10)–(12). Namely, solutions of (10)–(12) have to converge to the solution of (13) as a and b go to 1 in (12).

The value function of differential game (10)–(12) is computed using the numerical scheme (9). The spatial region and the grid size are chosen as $[-10, 10] \times [-10, 10] \times [0.1, 2]$ and $300 \times 300 \times 60$, respectively. The time horizon T is equal to 7, and the time step equals 0.01. The computation time is about 15 minutes on a Linux SMP-computer with 8xQuad-Core AMD Opteron processors (Model 8384, 2.7 GHz) and shared 64 Gb memory.

Figures 1 and 2 show the computed three-dimensional set

$$\{(x, y, V_p) : c(0, x, y, V_p) \le 0.3\}. \tag{14}$$

In the case of Figure 1, state constraint (12) is specified by $a = 0.8$ and $b = 1.2$, whereas $a = 0.5$ and $b = 1.5$ for Figure 2.

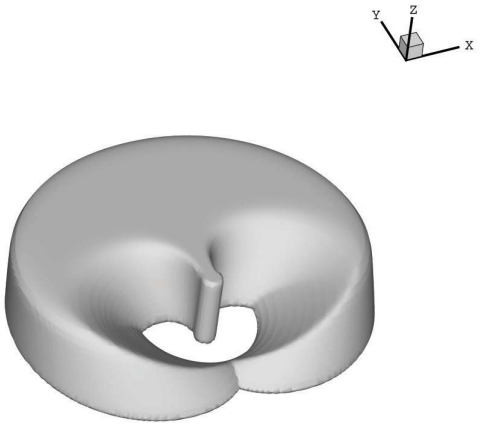


Fig. 1. Level set (14) corresponding to the state constraint $0.8 \leq V_p \leq 1.2$ (z-axis measures V_p)

Figure 3 shows the comparison of solutions of problems (10)–(12) and (13). Curve 1 bounds the solvability set of problem (13) without any state constraints. Curves 2 and 3 bound the sets

$$\{(x, y) : c(0, x, y, 1) \leq 0.3\}, \tag{15}$$

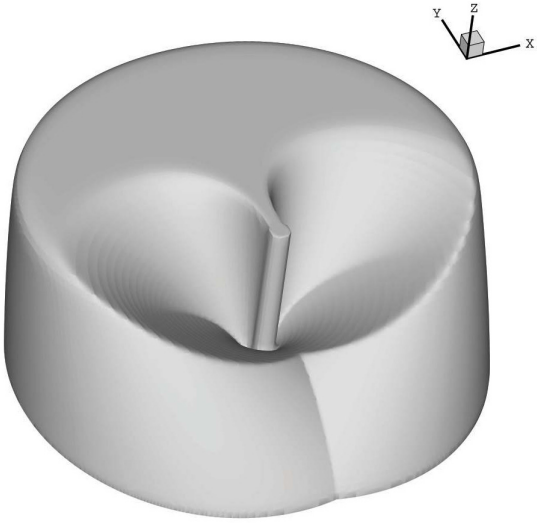


Fig. 2. Level set (14) corresponding to the state constraint $0.5 \leq V_p \leq 1.5$ (z-axis measures V_p)

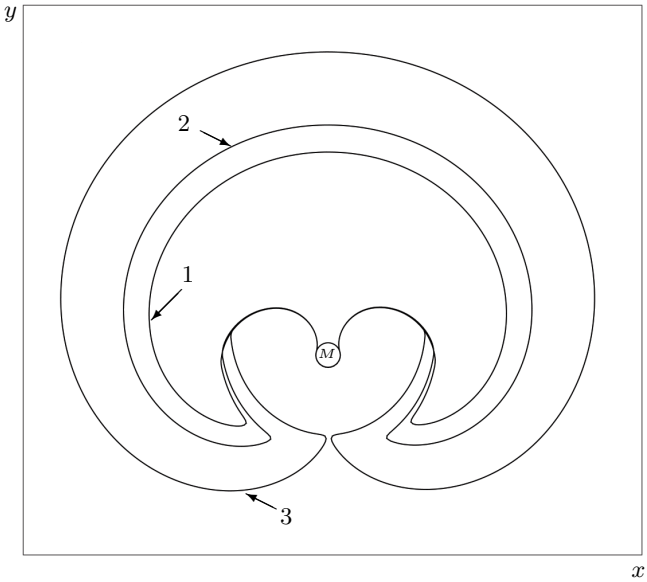


Fig. 3. Comparison of solutions of problems (10)–(12) and (13). Curves 2 and 3 show level set (15) in the case of the state constraints $0.8 \leq V_p \leq 1.2$ and $0.5 \leq V_p \leq 1.5$, respectively. Curve 1 shows the solvability set of problem (13).

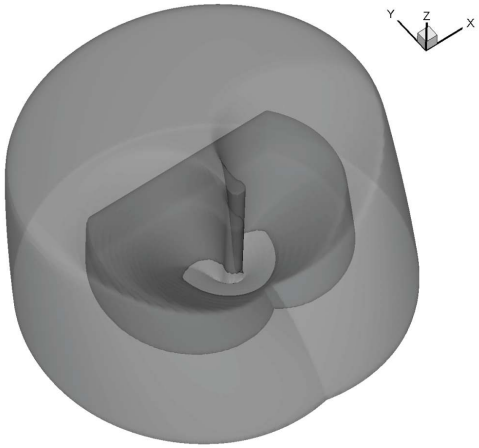


Fig. 4. Comparison of the level set (14) corresponding to the state constraints $0.5 \leq V_p \leq 1.5$ and $|y| \leq 3$ (z-axis measures V_p) with the set from Fig. 2

where c is as before the value function of problem (10)–(12) computed with $a = 0.8$ and $b = 1.2$ in the case of curve 2, and $a = 0.5$ and $b = 1.5$ in the case of curve 3. It is seen that the closer a and b are to 1, the closer the corresponding curve is to curve 1.

Figure 4 shows the case when the state constraint $|y| \leq 3$ is additionally imposed. The obtained set (14) is compared with the set given in Fig. 2.

6 Conclusion

Our experience shows that the numerical method outlined in this paper is appropriate for solving three- and even four-dimensional nonlinear problems with state constraints. Next steps are to be aimed towards dimensions five and six, which demands sparse representation of grid functions and operations on them, bearing in mind supercomputing systems available now. Such results will allow us to consider e.g. aircraft applications related to essentially nonlinear take-off and landing problems with complex state constraints inherent for them.

Acknowledgements This work was supported by the German Research Society (Deutsche Forschungsgemeinschaft) in the framework of the intention "Optimization with partial differential equations" (SPP 1253) and by Award No KSA-C0069/UK-C0020, made by King Abdullah University of Science and Technology (KAUST).

References

1. Isaacs, R.: Differential Games. John Wiley, New York (1965)
2. Krasovskii, N.N., Subbotin, A.I.: Positional Differential Games. Nauka, Moscow (1974) (in Russian)
3. Krasovskii, N.N., Subbotin, A.I.: Game-Theoretical Control Problems. Springer, New York (1988)
4. Subbotin, A.I., Chentsov, A.G.: Optimization of Guaranteed Result in Control Problems. Nauka, Moscow (1981)
5. Subbotin, A.I.: Generalization of the Main Equation of Differential Game Theory. J. Optimiz. Theory and Appl. 43, 103–133 (1984)
6. Crandall, M.G., Lions, P.L.: Viscosity Solutions of Hamilton-Jacobi Equations. Trans. Amer. Math. Soc. 277, 1–47 (1983)
7. Crandall, M.G., Evans, L.C., Lions, P.L.: Some Properties of Viscosity Solutions of Hamilton-Jacobi Equations. Trans. Amer. Math. Soc. 282, 487–502 (1984)
8. Subbotin, A.I., Taras'yev, A.M.: Stability Properties of the Value Function of a Differential Game and Viscosity Solutions of Hamilton-Jacobi Equations. Problems of Control and Information Theory 15, 451–463 (1986)
9. Crandall, M.G., Lions, P.L.: Two Approximations of Solutions of Hamilton-Jacobi Equations. Math. Comp. 43, 1–19 (1984)
10. Souganidis, P.E.: Approximation Schemes for Viscosity Solutions of Hamilton-Jacobi Equations. J. of Differential Equations 59, 1–43 (1985)
11. Souganidis, P.E.: Max - min Representation and Product Formulas for the Viscosity Solutions of Hamilton-Jacobi Equations with Applications to Differential Games. Nonlinear Analysis, Theory, Methods and Applications 9, 217–257 (1985)
12. Botkin, N.D.: Approximation Schemes for Finding the Value Functions for Differential Games with Nonterminal Payoff Functional. Analysis 14(2), 203–220 (1994)

13. Subbotin, A.I.: Generalized Solutions of First Order PDEs: The Dynamical Optimization Perspective. Birkhäuser, Basel (1995)
14. Botkin, N.D., Hoffmann, K.-H., Mayer, N., Turova, V.L.: Approximation Schemes for Solving Disturbed Control Problems with Non-Terminal Time and State Constraints. Analysis 31, 355–379 (2011)
15. Malafeyev, O.A., Troeva, M.S.: A Weak Solution of Hamilton-Jacobi Equation for a Differential Two-Person Zero-Sum Game. In: Eighth International Symposium on Differential Games and Applications, pp. 366–369. Université de Genève, Maastricht (1998)
16. Botkin, N.D., Hoffmann, K.-H., Turova, V.L.: Stable Numerical Schemes for Solving Hamilton-Jacobi-Bellmann-Isaacs Equations. SIAM J. on Scientific Computing 33(2), 992–1007 (2011)
17. Bardi, M., Koike, S., Soravia, P.: Pursuit-Evasion Games with State Constraints: Dynamic Programming and Discrete Time Approximations. Discrete and Continuous Dynamical Sytems, Series A 2(6), 361–380 (2000)
18. Grigor'eva, S.V., Pakhotinskikh, V.Y., Uspenskii, A.A., Ushakov, V.N.: Construction of Solutions in Certain Differential Games with Phase Constraints. Mat. Sbornik 196(4), 51–78 (2005)
19. Cardaliaguet, P., Quincampoix, M., Saint-Pierre, P.: Differential Games through Viability Theory: Old and Recent Results. In: Jorgensen, S., Quincampoix, M., Vincent, T.L. (eds.) Advances in Dynamic Game Theory: Numerical Methods and Applications to Ecology an Economics. Annals of the Int. Society of Dynamic Games IX, pp. 3–35. Birkhäuser, Basel (2007)
20. Cristiani, E., Falcone, M.: Fully-Discrete Schemes for Value Function of Pursuit-Evasion Games with State Constraints. In: Bernhard, P., Gaitsgory, V., Pourtallier, O. (eds.) Advances in Dynamic Games and Their Applications. Annals of the Int. Society of Dynamic Games X, pp. 177–206. Birkhäuser, Basel (2009)
21. Bokanowski, O., Forcadel, N., Zidani, H.: Reachability and Minimal Times for State Constrained Nonlinear Problems without Any Controllability Assumption. SIAM J. on Control and Optimization 48(7), 4292–4316 (2010)
22. Bernhard, P.: Linear Pursuit-Evasion Games and the Isotropic Rocket. PhD Thesis, Stanford University (1971)
23. Lewin, J., Olsder, G.J.: The Isotropic Rocket Surveillance Game. Comput. Math. Appl. 18(1-3), 15–34 (1989)

Geometric Conditions for Regularity of Viscosity Solution to the Simplest Hamilton-Jacobi Equation*

Vladimir V. Goncharov** and Fátima F. Pereira

Universidade de Évora, CIMA-UE, Rua Romão Ramalho 59,
7000-671, Évora, Portugal
{goncha,fmfp}@uevora.pt

Abstract. Continuing research in [13] and [14] on well-posedness of the optimal time control problem with a constant convex dynamics in a Hilbert space we adapt one of the regularity conditions obtained there to a slightly more general problem, where nonaffine additive term appears. We prove existence and uniqueness of a minimizer in this problem as well as continuous differentiability of the value function, which can be seen as the viscosity solution to a Hamilton-Jacobi equation, near the boundary.

Keywords: optimal time control problem, viscosity solution, eikonal equation, duality mapping, proximal normals, proximal regularity, Hölder continuity.

1 Introduction

Let us start with the first order partial differential equation in finite dimensions

$$\Gamma\left(x, u\left(x\right), \nabla u\left(x\right)\right) = 0 \tag{1}$$

where $\Gamma : \Omega \times \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ is a continuous function, nonlinear with respect to (w.r.t.) the third variable; $\Omega \subset \mathbb{R}^n$ is an open bounded region. Due to applications in optimal control and dynamical systems (1) is traditionally called (stationary) *Hamilton-Jacobi equation*. There are various notions of solutions to this equation. For instance, a function $u : \overline{\Omega} \to \mathbb{R}$ of class $C\left(\overline{\Omega}\right) \cap C^1\left(\Omega\right)$ satisfying (1) for all $x \in \Omega$ is said to be *classical solution*, while a Lipschitz continuous function $u : \overline{\Omega} \to \mathbb{R}$ such that (1) holds for almost each (a.e.) $x \in \Omega$ is usually called *generalized* (or *almost everywhere*) *solution*. Speaking about solutions of (1) we always have in mind some prescribed boundary condition

$$u\left(x\right) = \theta\left(x\right), \quad x \in \partial\Omega, \tag{2}$$

* Work is realized in framework of the project "Variational Analysis: Theory and Applications" (PTDC/MAT/111809/2009) financially supported by Fundação para Ciência e Tecnologia (FCT), the Portuguese institutions COMPETE, QREN and the European Regional Development Fund (FEDER).

** Corresponding author.

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 245–254, 2013.

where $\theta : \overline{\Omega} \to \mathbb{R}$ is a (continuous) given function. Since in practice a classical solution to the boundary value problem (1)-(2) often fails to exist while generalized solution may not be unique, another physically reasonable concept (so named *viscosity solution*) was introduced by M. Crandall and P.-L. Lions in 1983 (see [6]) while similar constructions under different names were known earlier (see, e.g., [16], [12], [15]). This concept was mainly based on the idea of "vanishing viscosity" in the sense that each viscosity solution is the uniform limit of the sequence of solutions $u^{\varepsilon}(\cdot)$ to the respective boundary value problems for the nonlinear elliptic equations

$$\Gamma(x, u, \nabla u) - \varepsilon \Delta u = 0 \tag{3}$$

as $\varepsilon \to 0+$ where $\Delta = \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2}$ is the Laplace operator (notice that (3) has a unique classical solution for each $\varepsilon > 0$ small enough due to Theorem 3.2 [15]). The exact definition of viscosity solution can be given either in terms of the suitable test functions (similarly as the notion of the generalized solutions of linear PDE in the sense of distributions), or by involving a Fréchet generalization of the gradient of a function at the point of nondifferentiability. It turned out that for each suitable boundary data $\theta(\cdot)$ a (continuous) viscosity solution to the problem (1)-(2) exists, is unique and stable w.r.t. both $\theta(\cdot)$ and $\Gamma(\cdot)$. Furthermore, it is agreed with other types of solutions. In particular, each viscosity solution belonging to $C^1(\Omega)$ is classical. For the main results of Theory of Viscosity Solutions, very developed and powerful field of the modern mathematics, we refer to [1]−[2] and the bibliography therein. For a concise survey of viscosity solutions in finite dimensions see also the excellent tutorial lessons by A. Bressan [3].

Afterwards, the concept and the main results concerning viscosity solutions were generalized to Banach spaces with the *Radon-Nikodym property* (see [7], [8]), in particular, to Hilbert spaces. The gradient ∇u in (1) is understood then in the sense of Fréchet. Notice that although the definition based on the Fréchet sub- and superdifferentials remains the same, the interpretation of viscosity solutions via "vanishing viscosity" is no longer valid in infinite dimensions. The motivation, however, comes now from the Theory of Differential Games.

In our paper we deal only with the case when the hamiltonian Γ in (1) does not depend of x neither u, and is convex w.r.t. the third variable. Already S. N. Kružkov studied in [15] such Hamilton-Jacobi equations arizing from the geometric optics. For instance, when $n = 3$ and $\Gamma(x, u, \xi) = |\xi| - a$, with a constant $a > 0$, one has the so called *eikonal equation* describing the propagation of a light wave from a point source placed at the origin in homogeneous medium with refraction index $1/a$. If, instead, this medium is anisotropic and has constant coefficients of refraction of light rays parallel to the coordinate axes (say c_i) then the propagation of light can be described by the (more general) elliptic equation

$$\sum_{i=1}^{n} c_i^2 u_{x_i}^2 - 1 = 0. \tag{4}$$

If, besides that, the medium moves with a constant velocity $\overrightarrow{v}$ then the equation contains already a linear additive term and admits the form

$$\sum_{i=1}^{n} c_i^2 u_{x_i}^2 + \frac{2}{c} \langle \overrightarrow{v}, \nabla u \rangle - 1 = 0, \tag{5}$$

where c means the speed of the light in a vacuum.

In general, denoting by F the *closed convex hull* of the set of zeros

$$\{\xi \in \mathbb{R}^n : \Gamma(\xi) = 0\}$$

and assuming F to be bounded with $\text{int}F \neq \emptyset$ (the hamiltonians in (4) and (5) satisfy these conditions), the equation (1) can be reduced to

$$\rho_F(\nabla u(x)) - 1 = 0, \tag{6}$$

where $\rho_F(\cdot)$ is the *Minkowski functional* (*gauge function*) associated to F,

$$\rho_F(\xi) := \inf\{\lambda > 0 : \xi \in \lambda F\}.$$

More precisely, it was proved in [4] that under appropriate conditions involving a kind of geometric compatibility of F, $\theta(\cdot)$ and the domain Ω the (unique) viscosity solution $\hat{u}(\cdot)$, $\hat{u}|_{\partial\Omega} = \theta$, of (6) is the viscosity solution of the problem (1)-(2) (belonging to the space $W^{1,\infty}(\Omega)$) and vice versa. Furthermore, this viscosity solution can be given by the formula

$$\hat{u}(x) = \inf_{y \in C} \{\rho_{F^0}(x - y) + \theta(y)\} \tag{7}$$

whenever $\theta : \mathbb{R}^n \to \mathbb{R}$ is a Lipschitz continuous function such that

$$\nabla\theta(x) \in \text{int}F \quad \text{for a.e. } x \in \mathbb{R}^n. \tag{8}$$

Here $C := \mathbb{R}^n \backslash \Omega$ and F^0 is the *polar set* for F.

Let now H be a Hilbert space with the norm $\|\cdot\|$ and the inner product $\langle \cdot, \cdot \rangle$. Then the convolution (7) remains the unique viscosity solution to the equation (6) with the boundary data $u(x) = \theta(x)$, $x \in C$, whenever the slope condition

$$\theta(x) - \theta(y) < \rho_{F^0}(x - y) \quad \forall x, y \in C \tag{9}$$

holds. Notice that the inequality (9) follows from (8) in finite dimensions while in an arbitrary Hilbert space from the inclusion $\partial^c\theta(x) \subset F$, $x \in H$, where ∂^c is the *Clarke generalized gradient* of a Lipschitz continuous function.

So, we are interested in regularity properties of the function (7), which were well studied when $\theta \equiv 0$ (see [10], [13], [14]). In the latter case let us notice the following:

1) existence and regularity of the (Fréchet) gradient $\nabla\hat{u}(x)$ depends on uniqueness (in infinite dimensions also on existence) of a minimizer in (7);

2) it is not possible that $\nabla \hat{u}(x)$ exists everywhere out of C unless some special situations;
3) the function $\hat{u}(x)$ can be interpreted as the minimal time necessary to achieve the closed set C from $x \in H \backslash C$ by trajectories of the differential inclusion $\dot{x}(t) \in -F^0$.

Taking into account 1) and 2) it is natural to study the regularity only in an (open) neighbourhood of C (*target set* due to 3). If $F = \overline{B}$ is the closed unit ball centred in the origin then $\hat{u}(\cdot)$ is nothing else than the *distance* from C, and the minimizers in (7) are the usual *metric projections* onto C. In this case the (necessary and sufficient) condition guaranteeing well-posedness of the problem and the (Lipschitz) continuity of the gradient $\nabla \hat{u}(x)$ near C is so named φ-*convexity* (or *proximal smoothness*) of the set C well studied up to now (see survey [9] and the bibliography therein).

As concerns an arbitrary gauge F (and $\theta \equiv 0$) then in [13], [14] two different hypotheses are given, under which both a unique minimizer in (7) (that is a point on the boundary ∂C attained from x for the minimal time) and the gradient $\nabla \hat{u}(x)$ are (Hölder) continuous in a neighbourhood of C. It turns out that one of these hypotheses (based on certain ballance between external normals to the sets C and F) can be adapted to the case of a Lipschitz continuous perturbation $\theta(\cdot)$.

We start in Section 2 with the basic definitions and an auxiliary statement. Then, in Section 3, we study the mathematical programming problem (7) from the viewpoint of the existence, uniqueness and the (Lipschitz) regularity of minimizers near the set C. The geometric condition ensuring such well-posedness is emphasized here. Finally, in Section 4 we examine the (Fréchet) differentiability of the value function $\hat{u}(\cdot)$ and justify the (Hölder) continuity of its gradient also under the assumption that either F^0 or the restriction $\theta|_C$ is smooth.

2 Preliminaries

Given a convex closed bounded set $F \subset H$ with $0 \in \mathrm{int} F$ we consider the so called *duality mapping* $\mathfrak{J}_F : \partial F^0 \to \partial F$, which associates to each $\xi^* \in \partial F^0$ the set of (normalized) linear functionals that support F^0 at ξ^*,

$$\mathfrak{J}_F(\xi^*) := \{\xi \in \partial F : \langle \xi, \xi^* \rangle = 1\}.$$

In other words, $\mathfrak{J}_F(\xi^*) = \mathbf{N}_{F^0}(\xi^*) \cap \partial F$ where $\mathbf{N}_{F^0}(\xi^*)$ is the *normal cone* to the polar F^0 at ξ^*. It can be interpreted also as the subdifferential $\partial \rho_{F^0}(\xi^*)$ in the sense of Convex Analysis. For each dual pair (ξ, ξ^*), i.e., such that $\xi \in \partial F$, $\xi^* \in \partial F^0$ and $\langle \xi, \xi^* \rangle = 1$ let us define the modulus of rotundity (see [13])

$$\widehat{\mathfrak{C}}_F(r, \xi, \xi^*) := \inf\{\langle \xi - \eta, \xi^* \rangle : \eta \in F, \|\xi - \eta\| \geq r\}, \quad r > 0.$$

If the set F is *strictly convex* (*rotund*) at ξ w.r.t. ξ^*, i.e., $\widehat{\mathfrak{C}}_F(r, \xi, \xi^*) > 0 \quad \forall r > 0$ then ξ is an *exposed point* of F and, in particular, ξ is the unique element of

$\mathfrak{J}_F(\xi^*)$. So, in this case ξ is well defined whenever ξ^* is fixed. Furthermore, given a set $U \subset \partial F^0$ we say that F is *uniformly rotund* w.r.t. U if

$$\inf\left\{\widehat{\mathfrak{C}}_F(r,\xi,\xi^*) : \xi^* \in U\right\} > 0 \quad \forall r > 0.$$

By [14, Proposition 2.1] this property is equivalent to the uniform continuity of $\mathfrak{J}_F(\cdot)$ in the following sense

$$\sup_{\eta\in\mathfrak{J}_F(\eta^*)} \|\mathfrak{J}_F(\xi^*) - \eta\| \to 0 \quad \text{as} \quad \|\xi^* - \eta^*\| \to 0,\ \ \xi^* \in U,\ \ \eta^* \in \partial F^0 \tag{10}$$

(we clearly identify $\mathfrak{J}_F(\xi^*)$ with its element whenever it is a singleton). Uniform rotundity implies also the existence and the uniform continuity on U of the *Fréchet gradient* $\nabla\rho_{F^0}(\xi^*)$.

Besides the concepts of Convex Analysis above we will use the following notations. For a lower semicontinuous function $\varphi : H \to \mathbb{R}\cup\{+\infty\}$ we denote by $\partial^p\varphi(x)$, $\partial^l\varphi(x)$, $\partial^-\varphi(x)$ and $\partial^c\varphi(x)$ the *proximal*, *limiting* (*Mordukhovich*), *Fréchet* and *Clarke subdifferential*, respectively, at a point x, $\varphi(x) < +\infty$. All the definitions and the basic facts of the calculus for non convex sets can be found, e.g., in [5]. Here we observe only that the inclusions

$$\partial^p\varphi(x) \subset \partial^-\varphi(x) \subset \partial^l\varphi(x) \subset \partial^c\varphi(x) \tag{11}$$

always hold, while one of the reverse inclusions takes place whenever $\varphi(\cdot)$ is regular at x in some sense. For instance, $\varphi(\cdot)$ is said to be *proximal* (*Clarke*) *regular* at x if $\partial^p\varphi(x) = \partial^l\varphi(x)$ (respectively, $\partial^-\varphi(x) = \partial^c\varphi(x)$).

If $C \subset H$ is a nonempty closed set then the notion of some kind of *normal cone* to C at a point $x \in C$ can be given as the respective subdifferential of the indicator function $\mathbf{I}_C(\cdot)$ equal to 0 on C and to $+\infty$ elsewhere. In particular, the *proximal normal cone* $\mathbf{N}^p_C(x) := \partial^p\mathbf{I}_C(x)$. Further on we denote by $\partial^*C := \{x \in \partial C : \mathbf{N}^p_C(x) \neq \{0\}\}$ the *effective boundary*, which is dense in ∂C.

Returning to the problem of minimization in (7) let us formulate first an approximation result, which is crucial for what follows. It can be proved similarly as Lemma 5.1 [13] by using the Ekeland's variational principle as well as the fuzzy sum rule for the proximal subdifferentials (see [5, Theorem 1.8.3]).

Lemma 1. *Let $C \subset H$ be a nonempty closed set, and $\theta : H \to \mathbb{R}\cup\{+\infty\}$ be a lower semicontinuous function, lipschitzean on C. If $x \in H\backslash C$ and $\{x_n\} \subset C$ is a minimizing sequence for the function $y \mapsto \rho_{F^0}(x-y) + \theta(y)$ on C then there exist another minimizing sequence $\{x'_n\} \subset C$ and sequences $\{x''_n\}$, $\{v_n\}$ and $\{\xi_n\}$ such that $v_n \in \partial^p(\theta|_C)(x'_n)$, $\xi_n \in \partial\rho_{F^0}(x - x''_n)$ and $\|x'_n - x_n\| + \|x''_n - x_n\| \to 0$, $\|v_n - \xi_n\| \to 0$ as $n \to \infty$. Here $\theta|_C := \theta + \mathbf{I}_C$.*

Notice that if the points x'_n are such that $\partial^p(\theta|_C)(x'_n) = \partial^p\theta(x'_n) + \mathbf{N}^p_C(x'_n)$ and $\partial^p\theta(x'_n) \subset \gamma F$ for some $0 < \gamma < 1$ then without loss of generality we can assume that $x'_n \in \partial^*C$ and $v_n \in \partial F$.

3 Existence, Uniqueness and Regularity of Minimizers

Our standing hypothesis in what follows will be a slightly strengthened condition than (9):

(**H**) there exists $0 < \gamma < \frac{1}{\|F\|\|F^0\|}$ such that

$$\theta(x) - \theta(y) \leq \gamma \rho_{F^0}(x - y) \quad \forall x, y \in C,$$

where $\|F\| := \sup\{\|\xi\| : \xi \in F\}$.

Hence $\theta(\cdot)$ is lipschitzean on C with the constant $\gamma \|F\|$.

Given now an arbitrary point $x_0 \in \partial C$ let us emphasize the main local assumptions, under which the well-posedness results hold:

($\mathbf{H}_1$) the mapping $x \mapsto \mathfrak{J}_{F^0}(\partial^p(\theta|_C)(x) \cap \partial F)$ is **single-valued** and **lipschitzean** (with a constant $L = L(x_0) > 0$) on the set

$$C_\delta(x_0) := \{x \in \partial^* C : \|x - x_0\| \leq \delta\}, \ \ \delta > 0;$$

($\mathbf{H}_2$) in the δ-neighbourhood of x_0 the **sum rule** $\partial^p(\theta|_C)(x) = \partial^p\theta(x) + \mathbf{N}_C^p(x)$ takes place;

($\mathbf{H}_3$) F^0 is **uniformly rotund** w.r.t. the set

$$U_\delta(x_0) := \partial F \cap \bigcup_{x \in C_\delta(x_0)} \partial^p(\theta|_C)(x). \tag{12}$$

For each $x \in H$ we denote by $\pi_C^{F,\theta}(x)$ the (possibly empty) set of all minimizers of the function $y \mapsto \rho_{F^0}(x - y) + \theta(y)$ on C.

Theorem 1. *Under the hypotheses* ($\mathbf{H}_1$) – ($\mathbf{H}_3$) *there exists a neighbourhood* $\mathcal{U}(x_0)$ *such that the mapping* $x \mapsto \pi_C^{F,\theta}(x)$ *is single-valued and continuous on* $\mathcal{U}(x_0)$.

Proof. Let us give a sketch of the proof. Taking without loss of generality $\delta > 0$ such that $\delta\gamma\|F\| < \left(1 - \gamma\|F\|\left\|F^0\right\|\right)/L$, let us set

$$\mathcal{U}(x_0) := \left\{ x \in H : \|x - x_0\| < \frac{\left(1 - \gamma\|F\|\left\|F^0\right\|\right)\delta}{2\|F\|\|F^0\|}, \right.$$
$$\left. \hat{u}(x) < \hat{u}(x_0) + \frac{1 - \gamma\|F\|\left\|F^0\right\|}{L} - \delta\gamma\|F\| \right\}. \tag{13}$$

Fix $x \in \mathcal{U}(x_0) \setminus C$ and a minimizing sequence $\{x_n\} \subset C$ of $y \mapsto \rho_{F^0}(x - y) + \theta(y)$. Let us choose $\{x_n'\} \subset C$, $\{x_n''\} \subset H$, $v_n \in \partial^p(\theta|_C)(x_n')$ and $\xi_n \in \partial\rho_{F^0}(x - x_n'')$ as in Lemma 1. Our goal is to prove that $\{x_n'\}$ (hence $\{x_n\}$ as well) is a Cauchy sequence.

To this end we show, first, that $\|x'_n - x_0\| \leq \delta$. It follows then from $(\mathbf{H}_2)$ and from the remark after Lemma 1 that $x'_n \in C_\delta(x_0)$ and $v_n \in \partial F$ for all n large enough. Consider a (nonincreasing) sequence $\nu_n \to 0+$ such that

$$\|x'_n - x_n\| + \|x''_n - x_n\| \leq \nu_n; \tag{14}$$

$$\rho_{F^0}(x - x'_n) + \theta(x'_n) \leq \hat{u}(x) + \nu_n; \tag{15}$$

$$\|v_n - \xi_n\| \leq \nu_n,$$

$n = 1, 2, \ldots$ (see Lemma 1). Then using the hypothesis $(\mathbf{H}_3)$ together with the property (10) (applied to the gauge F^0) gives that

$$\beta_n := \sup_{\|\xi-\eta\| \leq \nu_n} \sup\{\|\mathfrak{J}_{F^0}(\xi) - \mathfrak{J}_{F^0}(\eta)\| : \xi \in \partial F,\ \eta \in U_\delta(x_0)\} \to 0 \quad \text{as} \quad n \to \infty.$$

Taking into account that $v_n \in U_\delta(x_0)$ (see (12)) and $\xi_n \in \partial F$ we obtain $\|\mathfrak{J}_{F^0}(\xi_n) - \mathfrak{J}_{F^0}(v_n)\| \leq \beta_n$, and, consequently, by $(\mathbf{H}_1)$

$$\|\mathfrak{J}_{F^0}(\xi_n) - \mathfrak{J}_{F^0}(\xi_m)\| \leq 2\beta_n + L\|x'_n - x'_m\| \tag{16}$$

for all $m \geq n \geq 1$.

On the other hand, by the elementary properties of the convex subdifferentials and duality mappings we find that $(x - x''_n)/\rho_{F^0}(x - x''_n) = \mathfrak{J}_{F^0}(\xi_n)$, and hence

$$\begin{aligned}\|x''_n - x''_m\| &= \|\rho_{F^0}(x - x''_n)\mathfrak{J}_{F^0}(\xi_n) - \rho_{F^0}(x - x''_m)\mathfrak{J}_{F^0}(\xi_m)\| \leq \\ &\leq \rho_{F^0}(x - x''_n)\|\mathfrak{J}_{F^0}(\xi_n) - \mathfrak{J}_{F^0}(\xi_m)\| + \\ &\quad + |\rho_{F^0}(x - x''_n) - \rho_{F^0}(x - x''_m)|\left\|F^0\right\|. \end{aligned} \tag{17}$$

The terms $|\rho_{F^0}(x - x''_n) - \rho_{F^0}(x - x''_m)|$ and $\rho_{F^0}(x - x''_n)$ can be approximately estimated by $\|x'_n - x'_m\|$ and by $\hat{u}(x) - \hat{u}(x_0)$, respectively (we use for that the inequalities (14) and (15)). Hence, taking into account also (16) we deduce from (17) that

$$\begin{aligned}&\|x'_n - x'_m\| \leq \|x''_n - x''_m\| + 2\nu_n \leq \mu_n + \\ &+ \left[L(\mu'_n + \hat{u}(x) - \hat{u}(x_0) + \delta\gamma\|F\|) + \gamma\|F\|\left\|F^0\right\|\right]\|x'_n - x'_m\|\end{aligned}$$

for all $m \geq n \geq 1$, where $\{\mu_n\}$ and $\{\mu'_n\}$ are some sequences, converging to zero. We conclude the proof recalling the definition of the neighbourhood $\mathcal{U}(x_0)$. Thus, the limit $\bar{x} := \lim_{n\to\infty} x'_n$ will be (unique) minimizer from $\pi_C^{F,\theta}(x)$. The continuous dependence of this singleton on $x \in \mathcal{U}(x_0)$ also follows.

Remark 1. In fact, adapting the proof of Theorem 3.1 [14] we can show that the mapping $x \mapsto \pi_C^{F,\theta}(x)$ is locally lipschitzean on the same neighbourhood (13) with the Lipschitz constant tending to $+\infty$ as x tends to the boundary $\partial\mathcal{U}(x_0)$.

4 Differentiability of the Viscosity Solution

We announce, first, a result on subdifferential regularity of the function (7) at a fixed point $x \notin C$, similar to Proposition 5.1 [14]. We see that the regularity relies upon well-posedness of the minimizers studied in the previous section.

Theorem 2. *Let us fix $x \in H \backslash C$ and assume that*

- $\pi_C^{F,\theta}(y)$ *is a singleton for each* y, $\|x-y\| \leq \delta$, $\delta > 0$;
- *the restriction* $\theta|_C$ *is proximally regular at* $\bar{x} := \pi_C^{F,\theta}(x)$;
- *the following "centred" Hölder property*

$$\left\| \pi_C^{F,\theta}(x) - \pi_C^{F,\theta}(y) \right\| \leq K \|x-y\|^{\beta} \quad \forall y,\ \|x-y\| \leq \delta,$$

holds with an exponent $1/2 < \beta \leq 1$ *and a constant* $K = K(x) > 0$.

Then the function $\hat{u}(\cdot)$ *is Clarke regular at* x. *More precisely,*

$$\partial^c \hat{u}(x) = \partial^- \hat{u}(x) = \partial \rho_{F^0}(x-\bar{x}) \cap \partial^-(\theta|_C)(\bar{x}) \neq \emptyset. \tag{18}$$

Recalling Theorem 1 and Remark 1 we immediately obtain from the statement above that under the hypotheses $(\mathbf{H}_1)-(\mathbf{H}_3)$ the viscosity solution $\hat{u}(\cdot)$ is Clarke regular and (18) holds for all $x \in \mathcal{U}(x_0)$ where the neighbourhood $\mathcal{U}(x_0)$ is defined by (13). Thus, for the (Fréchet) continuous differentiability it suffices to require that the intersection $\partial \rho_{F^0}(x-\bar{x}) \cap \partial^-(\theta|_C)(\bar{x})$ is a singleton continuously depending on $x \in \mathcal{U}(x_0)$. However, this is difficult to verify directly because the mapping (18) (which is nothing else than the Fréchet gradient $\nabla \hat{u}(x)$) depends on the point x through *a priori* unknown function $\pi_C^{F,\theta}(\cdot)$. On the other hand, this condition splits into two different hypotheses regarding the smoothness either of both the function $\theta(\cdot)$ and the set C, or of the polar gauge F^0. Moreover, such hypotheses can be given plainly in terms of boundary points of the sets C and F^0. Notice that C is said to be *with smooth boundary* near x_0 if $\mathbf{N}_C^l(x) \cap \partial \overline{B}$ is a singleton (say $\{\mathfrak{n}_C(x)\}$) continuously depending on $x \in \partial C$, $\|x-x_0\| \leq \delta$, $\delta > 0$. Also the smoothness of F^0 can be equivalently substituted by the rotundity assumption for F.

Thus, we arrive at the following result.

Theorem 3. *Given* $x_0 \in \partial C$ *and* $\delta > 0$ *let us assume the hypotheses* $(\mathbf{H}_1)-(\mathbf{H}_3)$. *Suppose also that at least one of the following two conditions holds:*

(**i**) C *has smooth boundary, and* $\theta(\cdot)$ *is of class* $\mathcal{C}^1$ *near* x_0;
(**ii**) F *is rotund w.r.t. each* $\xi^* \in \mathfrak{J}_{F^0}(\partial^p(\theta|_C)(x))$, $x \in \partial C$ *with* $\|x-x_0\| \leq \delta$.

Then $\hat{u}(\cdot)$ *is Fréchet continuously differentiable on a neighbourhood* $\mathcal{U}(x_0)$. *Furthermore, in the first case we have*

$$\nabla \hat{u}(x) = \nabla \theta(\bar{x}) + \lambda \mathfrak{n}_C(\bar{x}),$$

where $\lambda = \lambda(\bar{x}) > 0$ *is the unique positive root of the equation*

$$\rho_F(\nabla \theta(\bar{x}) + \lambda \mathfrak{n}_C(\bar{x})) = 1,$$

while in the second

$$\nabla \hat{u}(x) = \nabla \rho_{F^0}(x-\bar{x}).$$

Here $\bar{x} := \pi_C^{F,\theta}(x)$, $x \in \mathcal{U}(x_0)$, *as before.*

Proof. If the condition (**i**) is fulfilled then taking into account that $\pi_C^{F,\theta}(\cdot)$ is single-valued and continuous on $\mathcal{U}(x_0)$, and that

$$\partial^l(\theta|_C)(\bar{x}) = \nabla\theta(\bar{x}) + \mathbf{N}_C^l(\bar{x}) = \{\nabla\theta(\bar{x}) + \lambda\mathfrak{n}_C(\bar{x}) : \lambda \geq 0\}$$

whenever $x \in \mathcal{U}(x_0)$, we obtain that the intersection in (18) reduces to the singleton $\{\nabla\theta(\bar{x}) + \lambda(\bar{x})\,\mathfrak{n}_C(\bar{x})\}$ (see also (11)). The continuity of $\nabla\hat{u}(\cdot)$ can be shown now by the standart implicit function argument.

Under the alternative assumption (**ii**) it sufices to observe that due to a *necessary condition of optimality* (in the proximal form) the (unique) minimizer $\bar{x} = \pi_C^{F,\theta}(x)$ must satisfy the relationship

$$\partial^p(\theta|_C)(\bar{x}) \cap \mathbf{N}_{F^0}\left(\frac{x-\bar{x}}{\rho_{F^0}(x-\bar{x})}\right) \cap \partial F \neq \emptyset. \tag{19}$$

Then, it follows from (19) that $\frac{x-\bar{x}}{\rho_{F^0}(x-\bar{x})} \in \mathfrak{J}_{F^0}(\xi)$ for some $\xi \in \partial^p(\theta|_C)(\bar{x})$. Therefore, if $x \in \mathcal{U}(x_0)$ then $\bar{x}$ is closed to x_0 as well, and taking $\xi^* = \frac{x-\bar{x}}{\rho_{F^0}(x-\bar{x})}$ we deduce from (**ii**) that $\rho_{F^0}(\cdot)$ is (Fréchet) continuously differentiable at ξ^*. So, the intersection in (18) reduces to $\{\nabla\rho_{F^0}(x-\bar{x})\}$, and the continuity w.r.t. x also follows.

Remark 2. If in addition to the hypothesis (**i**) in Theorem 3 we assume that both unit normal vector $\mathfrak{n}_C(\cdot)$ and the gradient $\nabla\theta(\cdot)$ are Hölder continuous with an exponent $0 < \alpha \leq 1$ on a δ-neighbourhood of x_0 then $\nabla\hat{u}(\cdot)$ will be also Hölder continuous near x_0 with the same exponent (we say that $\hat{u}(\cdot)$ is of class $\mathcal{C}^{1,\alpha}_{loc}$ on $\mathcal{U}(x_0)$). One can derive the Hölder inequality for $\nabla\hat{u}(\cdot)$ by using Theorem 3 and the estimates for the Hausdorff distance between the polars for convex solids (see Lemma 2 [11]).

References

[1] Bardi, M., Capuzzo-Dolcetta, I.: Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations. Birkhäuser, Boston (1997)

[2] Barles, G.: Solutions de Viscosité des Équations de Hamilton-Jacobi. Springer, Berlin (1994)

[3] Bressan, A.: Hamilton-Jacobi equations and optimal control. An illustrated tutorial. Trondheim, NTNU (2001)

[4] Cardaliaguet, P., Dacorogna, B., Gangbo, W., Georgy, N.: Geometric restrictions for the existence of viscosity solutions. Ann. Inst. Henri Poincaré 16, 189–220 (1999)

[5] Clarke, F.H., Ledyaev, Y.S., Stern, R.J., Wolenski, P.R.: Nonsmooth Analysis and Control Theory. Springer, New York (1998)

[6] Crandall, M., Lions, P.-L.: Viscosity solutions of Hamilton-Jacobi equations. Trans. Amer. Math. Soc. 277, 1–42 (1983)

[7] Crandall, M., Lions, P.-L.: Hamilton-Jacobi equations in infinite dimensions. I: Uniqueness of viscosity solutions. J. Funct. Anal. 62, 379–396 (1985)

[8] Crandall, M., Lions, P.-L.: Hamilton-Jacobi equations in infinite dimensions. II: Existence of viscosity solutions. J. Funct. Anal. 65, 368–405 (1986)

[9] Colombo, G., Thibault, L.: Prox-regular sets and applications. In: Gao, D.Y., Motreano, D. (eds.) Handbook on Nonconvex Analysis. International Press, Boston (2010)
[10] Colombo, G., Wolenski, P.R.: Variational Analysis for a class of minimal time functions in Hilbert spaces. J. Convex Anal. 11, 335–361 (2004)
[11] Dal Maso, G., Goncharov, V.V., Ornelas, A.: A Lipschitz selection from the set of minimizers of a nonconvex functional of the gradient. Nonlin. Anal.: Theory, Meth. and Appl. 37, 707–717 (1999)
[12] Douglis, A.: The continuous dependence of generalized solutions of non-linear partial differential equations upon initial data. Comm. Pure Appl. Math. 14, 267–284 (1961)
[13] Goncharov, V.V., Pereira, F.F.: Neighbourhood retractions of nonconvex sets in a Hilbert space via sublinear functionals. J. Convex Anal. 18, 1–36 (2011)
[14] Goncharov, V.V., Pereira, F.F.: Geometric conditions for regularity in a time-minimum problem with constant dynamics. J. Convex Anal. 19, 631–669 (2012)
[15] Kružkov, S.N.: Generalized solutions of the Hamilton-Jacobi equation of the Eikonal type. I. Math. USSR Sbornik 27, 406–446 (1975)
[16] Oleinik, O.A.: On discontinuous solutions of non-linear differential equations. Uspehi Mat. Nauk 12, 3–73 (1957)

Stabilization of the Gas Flow in Star-Shaped Networks by Feedback Controls with Varying Delay

Martin Gugat, Markus Dick*, and Günter Leugering**

Lehrstuhl 2 für Angewandte Mathematik
Friedrich-Alexander-Universität Erlangen-Nürnberg
Cauerstr. 11, 91058 Erlangen, Germany
{gugat,dick,leugering}@math.fau.de
http://www.mso.math.fau.de

Abstract. We consider the subcritical gas flow through star-shaped pipe networks. The gas flow is modeled by the isothermal Euler equations with friction. We stabilize the isothermal Euler equations locally around a given stationary state on a finite time interval. For the stabilization we apply boundary feedback controls with time-varying delays. The delays are given by C^1-functions with bounded derivatives. In order to analyze the system evolution, we introduce an L^2-Lyapunov function with delay terms. The boundary controls guarantee the exponential decay of the Lyapunov function with time.

Keywords: boundary feedback stabilization, Euler equations, gas network, Lyapunov function, star-shaped network, time-varying delay.

1 Introduction

Recently, there has been intense research on the system dynamics in gas networks (see e.g. [1,2,5,7,8,10,13]). Due to the pipe wall friction, there is a loss of pressure along the pipe. A common model for the gas flow in pipes is the isothermal Euler equations with friction, a 2×2 PDE system of balance laws (see (1)). We study the isothermal Euler equations on a star-shaped network of N ($N \geq 2$) pipes that meet at a central node ω. The flow at the node ω is governed by the continuity of the density and conservation of mass (see (5)).

Our main result, stated in Theorem 1, is a method to stabilize the gas flow locally around a given stationary subcritical state on a finite time interval. To do so, we use boundary feedback controls with varying delays at the pipe ends which are not at the node ω (see (16)). In order to measure the system evolution, we

* Corresponding author.

** This work was supported by the DFG SPP 1253 and DAAD D/0811409 (Procope 2009/10). The authors like to thank the Institut Henri Poincaré (Paris, France) for providing a very stimulating environment during the "Control of Partial and Differential Equations and Applications" program in the Fall 2010.

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 255–265, 2013.

introduce a network Lyapunov function (see (22)) which is the sum of a weighted and squared L^2-norm (see (20)) and a delay term (see (21)) for each pipe. The feedback controls guarantee the exponential decay of the Lyapunov function with time (see (36)) and, hence, the exponential stability of the system.

In contrast to a previous work which studies only the case of *constant* delays (see [7]), the novelty of this paper is that we consider *nonconstant*, i.e. time-dependent, delays. This is very important for the daily operation of real gas networks. E.g., in the control of such networks via electrical and mechanical systems, nonconstant time delays often appear (see [3]).

This paper is structured as follows: In Sect. 2 we give the network notation, consider the isothermal Euler equations in terms of the physical and characteristic variables and present the coupling conditions at the node ω. In Sect. 3 stationary and nonstationary states are studied. The stabilization method, i.e. the feedback controls, the Lyapunov function and the exponential stability result (Theorem 1), are stated in Sect. 4. In Sect. 5 we prove Theorem 1.

The weighted and squared L^2-norm from (20) for the Euler equations has first been presented in [8]. It is an extension of the Lyapunov function introduced in [4]. Delay terms of the form (21) have been presented in [12] for the time-delayed stabilization of the wave equation. Related questions of the stabilization of the wave equation are e.g. studied in [6,9,14].

2 Gas Flow in a Star-Shaped Pipe Network

In this section we consider the gas flow in a star-shaped pipe network. First, we give the network notation. Then, we present the isothermal Euler equations in terms of the physical variables (see (1)) and in terms of the characteristic variables (see (3)). The coupling conditions at the central node of the network are stated in (5).

2.1 Network Notation

We consider a star-shaped network of N $(N \geq 2)$ cylindrical pipes with the same diameter $\delta > 0$ that meet at a central node ω. We define the index set $I = \{1, ..., N\}$ and number the pipes from *pipe* 1 to *pipe* N. Variables referring to pipe i $(i \in I)$ are denoted with a superscript (i). We model the pipes by a one-dimensional space model and parameterize the length $L^{(i)} > 0$ of pipe i by the space interval $[0, L^{(i)}]$ such that the end $x = 0$ is at the node ω. We consider the system on a finite time interval $[0, T]$ with $T > 0$.

2.2 Isothermal Euler Equations in Physical Variables

A common model for the system dynamics in gas pipes is the isothermal Euler equations with friction, a hyperbolic 2×2 system of balance laws (see [1,2,10,13]): For pipe i $(i \in I)$ they have the following form on $[0, T] \times [0, L^{(i)}]$:

$$\begin{cases} \partial_t \rho^{(i)}(t,x) \;+\; \partial_x q^{(i)}(t,x) = 0 \,, \\ \partial_t q^{(i)}(t,x) \;+\; \partial_x \left(\frac{(q^{(i)}(t,x))^2}{\rho^{(i)}(t,x)} + a^2 \rho^{(i)}(t,x) \right) = -\frac{\theta}{2} \frac{q^{(i)}(t,x)\,|q^{(i)}(t,x)|}{\rho^{(i)}(t,x)} \end{cases} \tag{1}$$

where $\rho^{(i)}(t,x) > 0$ is the density of the gas and $q^{(i)}(t,x) \neq 0$ the mass flux. The sign of $q^{(i)}$ depends on the direction of the gas flow. It is positive if the gas in pipe i flows away from the node ω. The constant $a > 0$ is the sonic speed in the gas and the constant $\theta = \nu/\delta$ is the quotient of the friction factor $\nu > 0$ and the pipe diameter $\delta > 0$. The first equation in (1) states the conservation of mass and the second equation is the momentum equation. In this paper we study *subsonic* or *subcritical* C^1-states, i.e. C^1-states with $|q^{(i)}|/\rho^{(i)} < a$ for all $i \in I$. The equations (1) have the same form for each pipe. However, our calculations would also be true if we had different pressure laws on different pipes.

2.3 Isothermal Euler Equations in Characteristic Variables

The equations (1) can be transformed to the Riemann invariants (characteristic variables)

$$R_{\pm}^{(i)} \;=\; -q^{(i)}/\rho^{(i)} \mp\, a\, \ln(\rho^{(i)}) \,. \tag{2}$$

For the calculation of the Riemann invariants see [5,8]. In terms of $R_{\pm}^{(i)}$ the system (1) has the form

$$\partial_t \begin{pmatrix} R_+^{(i)} \\ R_-^{(i)} \end{pmatrix} + \begin{pmatrix} \lambda_+^{(i)} & 0 \\ 0 & \lambda_-^{(i)} \end{pmatrix} \partial_x \begin{pmatrix} R_+^{(i)} \\ R_-^{(i)} \end{pmatrix} = -\frac{\theta}{8} \left(R_+^{(i)} + R_-^{(i)}\right) \left|R_+^{(i)} + R_-^{(i)}\right| \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{3}$$

with the eigenvalues

$$\lambda_{\pm}^{(i)} \;=\; \frac{q^{(i)}}{\rho^{(i)}} \pm a \;=\; -\frac{R_+^{(i)} + R_-^{(i)}}{2} \pm a \,. \tag{4}$$

In the subsonic case, $\lambda_+^{(i)}$ is strictly positive and $\lambda_-^{(i)}$ strictly negative.

2.4 Coupling Conditions

At the node ω, i.e. at the ends $x = 0$ of the pipes, we have the following coupling conditions in terms of the physical variables ($t \in [0,\, T]$) (see [1,11]):

$$\rho^{(1)}(t,0) \;=\; \rho^{(i)}(t,0) \quad (i \in I \backslash \{1\}) \qquad \text{and} \qquad \sum_{i \in I} q^{(i)}(t,0) \;=\; 0 \,. \tag{5}$$

The first equation in (5) says that the density is continuous at the node ω. Due to the parameterization of the pipes, the second equation in (5) means that

the total ingoing mass flux is equal to the total outgoing mass flux at ω. The conditions (5) can equivalently be stated in terms of $R_{\pm}^{(i)}$ as ($t \in [0, T]$)

$$\left(R_+^{(1)}(t,0), ..., R_+^{(N)}(t,0)\right) = \left(R_-^{(1)}(t,0), ..., R_-^{(N)}(t,0)\right) A_\omega \tag{6}$$

with the orthogonal, symmetric $(N \times N)$-matrix $A_\omega = (a_{kl})_{k,l=1}^N$ with the entries

$$a_{kk} = (N-2)/N \quad (k \in I) \quad \text{and} \quad a_{kl} = -2/N \quad (k, l \in I, k \neq l)\,. \tag{7}$$

3 Stationary and Nonstationary States

3.1 Stationary States

The existence and behavior of stationary solutions of the isothermal Euler equations, i.e. solutions which do not explicitly depend on the time t, is studied in [5,8]. We denote the stationary variables as $\bar{\rho}(x)$, $\bar{q}(x)$, $\bar{R}_{\pm}(x)$ and $\bar{\lambda}_{\pm}(x)$. In [5,8] it is shown that $\bar{q}(x)$ is constant along a pipe and $\bar{\rho}(x)$ is strictly monotonically decreasing in the direction of the gas flow. Furthermore, a stationary subsonic C^1-solution of the isothermal Euler equations exists on the whole pipe if the pipe length is shorter than a critical length (see (8)). The critical length depends on the inflow density, the mass flux, the friction factor and the pipe diameter. For typical high-pressure gas pipes the critical length is around 180km (see [5]).

For the system (1) on a star-shaped network with the conditions (5), the existence and behavior of stationary subsonic C^1-solutions is in detail discussed in [7]. In the following we summarize the main results from [7]: The stationary mass fluxes $\bar{q}^{(i)} \neq 0$ are constant and have to satisfy $\sum_{i \in I} \bar{q}^{(i)} = 0$. At the node ω, we have a constant density $\bar{\rho}_\omega$ with $\bar{\rho}^{(i)}(0) = \bar{\rho}_\omega$ and $|\bar{q}^{(i)}|/\bar{\rho}_\omega < a$ for all $i \in I$. Furthermore, the lengths of the pipes with positive mass flux, i.e. with $\bar{q}^{(i)} > 0$, have to satisfy

$$L^{(i)} < \frac{1}{\theta}\left(a^2 \frac{\bar{\rho}_\omega^2}{(\bar{q}^{(i)})^2} - 1 + 2\ln\left(\frac{\bar{q}^{(i)}}{a\bar{\rho}_\omega}\right)\right)\,. \tag{8}$$

3.2 Nonstationary States

Assume that on the star-shaped network we have a given stationary subsonic state $(\bar{\rho}^{(i)}(x), \bar{q}^{(i)})$ with the corresponding Riemann invariants $(\bar{R}_+^{(i)}(x), \bar{R}_-^{(i)}(x))$ $\in (C^1([0, L^{(i)}]))^2$ and the eigenvalues $\bar{\lambda}^{(i)}(x)$ ($i \in I$) (see (2), (4)). We define the numbers (see (4))

$$\sigma^{(i)} = \operatorname{sign}(\bar{q}^{(i)}) = -\operatorname{sign}(\bar{R}_+^{(i)} + \bar{R}_-^{(i)}) \in \{-1, 1\}\,. \tag{9}$$

Now we consider a nonstationary state $(\bar{R}_+^{(i)}(x) + r_+^{(i)}(t,x), \bar{R}_-^{(i)}(x) + r_-^{(i)}(t,x))$ on $[0, T] \times [0, L^{(i)}]$ in a local neighborhood of $(\bar{R}_+^{(i)}, \bar{R}_-^{(i)})$. That is we assume the given stationary state $\bar{R}_{\pm}^{(i)}(x)$ to be slightly perturbed by $r_{\pm}^{(i)}(t,x)$ where the

C^1-norms $||r_\pm^{(i)}||_{C^1([0,\,T]\times[0,\,L^{(i)}])}$ are small. In particular, we suppose that the mass flux directions of the nonstationary state are the same as of the stationary state, i.e. $(i \in I)$ (see (9))

$$\sigma^{(i)} \;=\; -\mathrm{sign}(\bar R_+^{(i)} + \bar R_-^{(i)} + r_+^{(i)} + r_-^{(i)})\,. \tag{10}$$

The stabilization method presented in Sect. 4 guarantees that the absolute values of $r_\pm^{(i)}$ are small enough (see (24), (35)), such that (10) holds and such that the direction of the mass fluxes does not change during the stabilization process. From (3) we obtain the following quasilinear system for $r_\pm^{(i)}(t,x)$ $(i \in I)$

$$\begin{cases} \partial_t\, r_+^{(i)} + (\bar\lambda_+^{(i)} - \frac{r_+^{(i)}+r_-^{(i)}}{2})\,\partial_x\, r_+^{(i)} = (r_+^{(i)} + r_-^{(i)})(-K_+^{(i)} + \sigma^{(i)}\frac{\theta}{8}(r_+^{(i)} + r_-^{(i)}))\,, \\ \partial_t\, r_-^{(i)} + (\bar\lambda_-^{(i)} - \frac{r_+^{(i)}+r_-^{(i)}}{2})\,\partial_x\, r_-^{(i)} = (r_+^{(i)} + r_-^{(i)})(-K_-^{(i)} + \sigma^{(i)}\frac{\theta}{8}(r_+^{(i)} + r_-^{(i)})) \end{cases} \tag{11}$$

on $[0,\,T]\times[0,\,L^{(i)}]$ with the strictly positive C^1-functions

$$K_\pm^{(i)}(x) \;=\; \frac{\theta}{8}\,|\bar R_+^{(i)}(x) + \bar R_-^{(i)}(x)|\,\frac{4a \mp (\bar R_+^{(i)}(x) + \bar R_-^{(i)}(x))}{2a \mp (\bar R_+^{(i)}(x) + \bar R_-^{(i)}(x))} \;>\; 0\,. \tag{12}$$

The linearity of the equation (6) implies that for $r_\pm^{(i)}(t,0)$ at the node ω we have the equation $(t \in [0,\,T])$

$$\Big(r_+^{(1)}(t,0), ..., r_+^{(N)}(t,0)\Big) \;=\; \Big(r_-^{(1)}(t,0), ..., r_-^{(N)}(t,0)\Big)\,A_\omega \tag{13}$$

with the matrix A_ω as in (7).

4 Feedback Stabilization with Time-Varying Delay

In this section we present a method for the stabilization of the system (11) on $[0,\,T]\times[0,\,L^{(i)}]$ with time-delayed feedbacks. The corresponding boundary controls are given in (16). In order to measure the system evolution, we define the Lyapunov function $\mathcal{F}(t)$ in (22) with the weighted and squared L^2-norms $\mathcal{E}^{(i)}(t)$ in (20) and the delay terms $\mathcal{D}^{(i)}(t)$ in (21). Theorem 1 states the existence of a unique C^1-solution of (11) with small C^1-norm (see (35)) for which $\mathcal{F}(t)$ decays exponentially with time (see (36)). Hence, Theorem 1 gives the exponential stability of the system (11) around $(r_+^{(i)}, r_-^{(i)}) = (0,0)$.

4.1 Boundary Feedback Controls with Time-Varying Delay

Let a finite time $T > 0$ and a stationary subsonic state $(\bar R_+^{(i)}(x), \bar R_-^{(i)}(x)) \in (C^1([0,\,L^{(i)}]))^2$ on the star-shaped network be given with the eigenvalues $\bar\lambda_\pm^{(i)}(x)$ as in (4) $(i \in I)$. Define the numbers $\sigma^{(i)} \in \{-1, 1\}$ as in (9) and the functions

$K_{\pm}^{(i)}(x) \in C^1([0,\, L^{(i)}])$ as in (12). Let functions $\tau^{(i)}(t) \in C^1([0,\, T])$ $(i \in I)$ be given that satisfy $(t \in [0,\, T])$:

$$0 \,<\, \tau^{(i)}(t) \,<\, \frac{T}{2} \quad \text{and} \quad |\frac{d}{dt}\tau^{(i)}(t)| \,<\, 1\,. \tag{14}$$

Define the constants

$$\overline{\tau}^{(i)} = \max_{t\in[0,\,T]} \tau^{(i)}(t), \quad \hat{\tau}^{(i)} = \max_{t\in[0,\,T]} |\frac{d}{dt}\tau^{(i)}(t)| \quad \text{and} \quad \tau_{max} = \max_{i\in I} \left\{\overline{\tau}^{(i)}\right\}\,. \tag{15}$$

For a nonstationary state $\bar{R}_{\pm}^{(i)}(x) + r_{\pm}^{(i)}(t,x)$ on $[0,\, T] \times [0,\, L^{(i)}]$ we consider the system (11) with the condition (13) with the matrix A_ω as in (7). At the end $x = L^{(i)}$ of pipe i $(i \in I)$ we apply the controls

$$r_{-}^{(i)}(t, L^{(i)}) \,=\, \begin{cases} \vartheta^{(i)}(t) & (t \in [0,\, \overline{\tau}^{(i)}])\,, \\ (1-\varsigma^{(i)}(t))\,\vartheta^{(i)}(t) + \varsigma^{(i)}(t)\,k^{(i)}\,r_{+}^{(i)}(t-\tau^{(i)}(t), L^{(i)}) & (t \in (\overline{\tau}^{(i)},\, 2\overline{\tau}^{(i)}])\,, \\ k^{(i)}\,r_{+}^{(i)}(t-\tau^{(i)}(t), L^{(i)}) & (t \in (2\overline{\tau}^{(i)},\, T]) \end{cases} \tag{16}$$

with feedback constants $k^{(i)} \in (-1, 1)$ and functions $\vartheta^{(i)}(t) \in C^1([0,\, 2\overline{\tau}^{(i)}])$ and $\varsigma^{(i)}(t) \in C^1([\overline{\tau}^{(i)},\, 2\overline{\tau}^{(i)}])$ that have the following properties:

$$\varsigma^{(i)}(\overline{\tau}^{(i)}) \,=\, \frac{d}{dt}\varsigma^{(i)}(\overline{\tau}^{(i)}) \,=\, \frac{d}{dt}\varsigma^{(i)}(2\overline{\tau}^{(i)}) \,=\, 0 \quad \text{and} \quad \varsigma^{(i)}(2\overline{\tau}^{(i)}) \,=\, 1\,. \tag{17}$$

4.2 Network Lyapunov Function with Delay Terms

In order to define the network Lyapunov function, we define the numbers $(i \in I)$

$$\mu^{(i)} \,=\, \left(\int_0^{L^{(i)}} \frac{1}{\bar{\lambda}_{+}^{(i)}(x)} + \frac{1}{|\bar{\lambda}_{-}^{(i)}(x)|}\,dx\right)^{-1} \,>\, 0 \tag{18}$$

and the functions $(i \in I,\, x \in [0,\, L^{(i)}])$

$$h_{\pm}^{(i)}(x) \,=\, \exp\left(-\mu^{(i)} \int_0^x \frac{1}{\bar{\lambda}_{\pm}^{(i)}(\xi)}\,d\xi\right) > 0\,. \tag{19}$$

For constants $A_{\pm}^{(i)} > 0$ we define the weighted and squared L^2-norms $(i \in I,$ $t \in [0,\, T])$

$$\mathcal{E}^{(i)}(t) \,=\, \int_0^{L^{(i)}} \frac{A_{+}^{(i)}}{\bar{\lambda}_{+}^{(i)}(x)}\,h_{+}^{(i)}(x)\,(r_{+}^{(i)}(t,x))^2 \,+\, \frac{A_{-}^{(i)}}{|\bar{\lambda}_{-}^{(i)}(x)|}\,h_{-}^{(i)}(x)\,(r_{-}^{(i)}(t,x))^2\,dx \tag{20}$$

and the delay terms $(i \in I,\, t \in [\overline{\tau}^{(i)},\, T])$

$$\mathcal{D}^{(i)}(t) \,=\, \int_0^{\tau^{(i)}(t)} A_{+}^{(i)}\,h_{+}^{(i)}(L^{(i)})\,\exp(-\mu^{(i)}s)\,(r_{+}^{(i)}(t-s, L^{(i)}))^2\,ds\,. \tag{21}$$

The network Lyapunov function $\mathcal{F}(t)$ is defined as ($t \in [\tau_{max}, T]$)

$$\mathcal{F}(t) = \sum_{i \in I} \mathcal{E}^{(i)}(t) + \mathcal{D}^{(i)}(t) . \tag{22}$$

Constants of the form $\mu^{(i)}$ in (18) and functions of the form $h_{\pm}^{(i)}(x)$ and $\mathcal{E}^{(i)}(t)$ in (19) and (20) have been introduced in [8]. Delay terms of the form (21) have been presented in [12].

4.3 Main Result: Exponential Stability

Theorem 1 states the existence of a unique C^1-solution of the system (11) on $[0, T] \times [0, L^{(i)}]$ ($i \in I$) with the boundary conditions (13) and (16) and initial data of the form ($i \in I$, $x \in [0, L^{(i)}]$)

$$(r_+^{(i)}(0,x), r_-^{(i)}(0,x)) = (\varphi_+^{(i)}(x), \varphi_-^{(i)}(x)) \tag{23}$$

with functions $\varphi_{\pm}^{(i)} \in C^1([0, L^{(i)}])$. For this solution, the function $\mathcal{F}(t)$ decays exponentially on $[2\tau_{max}, T]$ (see (36)). The decay rate is $\eta = \min_{i \in I} \alpha^{(i)} \beta^{(i)} \mu^{(i)}$ with numbers $\alpha^{(i)} \in (0, 1)$ and $\beta^{(i)} \in (0, 1)$. The C^1-norm of the solution $r_{\pm}^{(i)}(t,x)$ is bounded by a constant $\varepsilon_1 > 0$ (see (35)) which has to satisfy ($i \in I$)

$$\varepsilon_1 < \min_{x \in [0, L^{(i)}]} |\bar{\lambda}_{\pm}^{(i)}(x)| , \quad 2\varepsilon_1 < \min_{x \in [0, L^{(i)}]} |\bar{R}_+^{(i)}(x) + \bar{R}_-^{(i)}(x)| \tag{24}$$

and

$$\varepsilon_1 \left(\frac{\theta}{4} + \frac{1}{2} \right) \left(3 + \max \left\{ \exp(1) \frac{A_-^{(i)}}{A_+^{(i)}} \frac{\bar{\lambda}_+^{(i)}(L^{(i)})}{|\bar{\lambda}_-^{(i)}(L^{(i)})|}, \frac{A_+^{(i)}}{A_-^{(i)}} \frac{|\bar{\lambda}_-^{(i)}(0)|}{\bar{\lambda}_+^{(i)}(0)} \right\} \right) < (1 - \beta^{(i)}) \alpha^{(i)} \mu^{(i)} . \tag{25}$$

The C^1-norms of the initial data $\varphi_{\pm}^{(i)}$ and the functions $\vartheta^{(i)}$ in the controls (16) have to be sufficiently small. More precisely, there exists a number $\varepsilon_2 \in (0, \varepsilon_1]$ such that the following inequalities have to hold:

$$||\varphi_{\pm}^{(i)}||_{C^1([0, L^{(i)}])} \leq \varepsilon_2 \tag{26}$$

and

$$||\vartheta^{(i)}||_{C^1([0, \overline{\tau}^{(i)}])} \leq \varepsilon_2 , \quad ||(1 - \varsigma^{(i)}) \vartheta^{(i)}||_{C^1([\overline{\tau}^{(i)}, 2\overline{\tau}^{(i)}])} \leq \frac{\varepsilon_2}{2} . \tag{27}$$

Note that the second inequality in (27) holds for any $\varsigma^{(i)}$ if the C^1-norm of $\vartheta^{(i)}$ on $[\overline{\tau}^{(i)}, 2\overline{\tau}^{(i)}]$ is small enough. For Theorem 1 we define the positive real numbers ($i \in I$)

$$U_{\pm}^{(i)} = \max_{x \in [0, L^{(i)}]} \left| \frac{\bar{\lambda}_{\pm}^{(i)}(x)}{\bar{\lambda}_{\mp}^{(i)}(x)} \right| \frac{K_{\mp}^{(i)}(x)}{K_{\pm}^{(i)}(x)} > 0 \tag{28}$$

and

$$V_{\pm}^{(i)} = \min_{x \in [0, L^{(i)}]} \left| \frac{\bar{\lambda}_{\pm}^{(i)}(x)}{\bar{\lambda}_{\mp}^{(i)}(x)} \right| \frac{K_{\mp}^{(i)}(x)}{K_{\pm}^{(i)}(x)} \left(1 + \frac{(1 - \alpha^{(i)}) \mu^{(i)}}{K_{\mp}^{(i)}(x)} \right) > 0 . \tag{29}$$

Theorem 1. *Consider a star-shaped network of pipes as described in Sect. 2.1. Let a finite time $T > 0$ and functions $\tau^{(i)}(t) \in C^1([0,T])$ with (14) be given ($i \in I$). Define the constants $\overline{\tau}^{(i)}$, $\hat{\tau}^{(i)}$ and τ_{max} as in (15). Let real numbers $\alpha^{(i)} \in (0,1)$, $\beta^{(i)} \in (0,1)$ and a stationary subsonic state $(\bar{R}_+^{(i)}(x), \bar{R}_-^{(i)}(x)) \in (C^1([0,L^{(i)}]))^2$ with the eigenvalues $\bar{\lambda}_\pm^{(i)}(x)$ as in (4) be given ($i \in I$) that satisfies the coupling conditions (6). Define the numbers $\sigma^{(i)} \in \{-1,1\}$ and $\mu^{(i)} > 0$ as in (9) and (18) and the functions $K_\pm^{(i)} > 0$ and $h_\pm^{(i)} > 0$ as in (12) and (19). Define the numbers $U_\pm^{(i)} > 0$ and $V_\pm^{(i)} > 0$ as in (28) and (29) and assume that we have*

$$\exp(1)\, U_+^{(i)} \;\le\; V_+^{(i)} \quad \text{or} \quad \exp(1)\, U_-^{(i)} \;\le\; V_-^{(i)} \,. \tag{30}$$

Choose constants $A_\pm^{(i)} > 0$ that satisfy

$$A_+^{(i)} \;\le\; 1 \;\le\; A_-^{(i)} \tag{31}$$

and assume that we have

$$A_+^{(i)}/A_-^{(i)} \in [\exp(1)\, U_+^{(i)},\, V_+^{(i)}] \quad \text{or} \quad A_-^{(i)}/A_+^{(i)} \in [U_-^{(i)},\, \exp(-1)\, V_-^{(i)}] \,. \tag{32}$$

Choose a real number $\varepsilon_1 > 0$ that satisfies (24) and (25) for all $i \in I$. Then there exists $\varepsilon_2 \in (0, \varepsilon_1]$ such that the following statements hold:

Choose functions $\vartheta^{(i)}(t) \in C^1([0, 2\overline{\tau}^{(i)}])$, $\varsigma^{(i)}(t) \in C^1([\overline{\tau}^{(i)}, 2\overline{\tau}^{(i)}])$ and $\varphi_\pm^{(i)}(x) \in C^1([0, L^{(i)}])$ ($i \in I$) that satisfy (17), (26) and (27) and such that the C^1-compatibility conditions are satisfied at the points $(t,x) = (0,0)$ and $(t,x) = (0, L^{(i)})$ (see Remark 1). Choose constants $k^{(i)} \in (-1,1)$ ($i \in I$) that satisfy

$$\exp(1)\,(k^{(i)})^2\, A_-^{(i)} \;\le\; A_+^{(i)}\, \exp(-\mu^{(i)}\overline{\tau}^{(i)})\,(1 - \hat{\tau}^{(i)}) \tag{33}$$

and

$$|k^{(i)}| \;\le\; \varepsilon_2 / (8\varepsilon_1 ||\varsigma^{(i)}||_{C^1([\overline{\tau}^{(i)},\, 2\overline{\tau}^{(i)}])}) \,. \tag{34}$$

Then the initial-boundary value problem (11), (13), (16), (23) has a unique solution $(r_+^{(i)}, r_-^{(i)}) \in (C^1([0,T] \times [0, L^{(i)}]))^2$ that satisfies

$$||r_\pm^{(i)}||_{C^1([0,T]\times[0,L^{(i)}])} \;\le\; \varepsilon_1 \,. \tag{35}$$

For this solution define the functions $\mathcal{E}^{(i)}(t)$, $\mathcal{D}^{(i)}(t)$ and $\mathcal{F}(t)$ as in (20), (21) and (22). Then the Lyapunov function $\mathcal{F}(t)$ satisfies the following inequality with $\eta = \min_{i \in I} \alpha^{(i)} \beta^{(i)} \mu^{(i)}$:

$$\mathcal{F}(t) \;\le\; \mathcal{F}(2\tau_{max})\, \exp\left(-\eta(t - 2\tau_{max})\right) \quad \text{for} \quad t \in [2\tau_{max}, T] \,. \tag{36}$$

Remark 1. The C^1-compatibility conditions guarantee that the initial data (23) and the boundary conditions (13) and (16) and their first derivatives fit together at the points $(t,x) = (0,0)$ and $(t,x) = (0, L^{(i)})$ ($i \in I$). They can be calculated from $\varphi_\pm^{(i)}$, $\vartheta_\pm^{(i)}$, A_ω and the equations (11) (see [7]).

Remark 2. The inequalities (31) and the second assumption in (32) hold if, e.g., $V_-^{(i)}$ is sufficiently large such that

$$\max\{1, U_-^{(i)}\} \le \exp(-1)\, V_-^{(i)} . \tag{37}$$

More precisely, if (37) is satisfied, we can first choose the quotient $A_-^{(i)}/A_+^{(i)}$ such that

$$A_-^{(i)}/A_+^{(i)} \in [\max\{1, U_-^{(i)}\}, \exp(-1)\, V_-^{(i)}] .$$

Then, without changing the quotient $A_-^{(i)}/A_+^{(i)}$, we choose $A_\pm^{(i)} > 0$ such that (31) holds. The conditions (32) and (37) are satisfied if the number $\mu^{(i)} > 0$ from (18) is sufficiently large which is the case if the length $L^{(i)}$ is not too long.

5 Proof of the Main Result Stated in Theorem 1

In this section we prove Theorem 1. The existence of a unique solution of (11) follows from Theorem 2.1 in [15] where initial-boundary value problems for first order quasilinear hyperbolic systems are studied (see also [5,8]). For the proof of (36) we use the estimates (38), (40) and (41) for $\frac{d}{dt}\mathcal{E}^{(i)}(t)$, $\frac{d}{dt}\mathcal{D}^{(i)}(t)$ and $\frac{d}{dt}\mathcal{F}(t)$. The estimate (38) is the same as in [7] where $\alpha^{(i)} = \beta^{(i)} = 1/2$ ($i \in I$). The calculation of (40) is more complicated than in [7] where only constant delays are considered. Using integration by parts, Young's Inequality and the conditions (25) and (32), we obtain the following estimate for $\frac{d}{dt}\mathcal{E}^{(i)}(t)$ ($t \in [0, T]$):

$$\frac{d}{dt}\mathcal{E}^{(i)}(t) \le -\alpha^{(i)}\beta^{(i)}\mu^{(i)}\mathcal{E}^{(i)}(t) + \Big[A_-^{(i)} h_-^{(i)}(x)(r_-^{(i)}(t,x))^2 - A_+^{(i)} h_+^{(i)}(x)(r_+^{(i)}(t,x))^2\Big]_{x=0}^{L^{(i)}} . \tag{38}$$

For a detailed calculation of (38) see [7]. For the derivative $\frac{d}{dt}\mathcal{D}^{(i)}(t)$ we get ($t \in [\overline{\tau}^{(i)}, T]$)

$$\begin{aligned}\tfrac{d}{dt}\mathcal{D}^{(i)}(t) = {} & 2\int_0^{\tau^{(i)}(t)} A_+^{(i)}\, h_+^{(i)}(L^{(i)})\, \exp(-\mu^{(i)} s)\, r_+^{(i)}(t-s, L^{(i)})\, \partial_t r_+^{(i)}(t-s, L^{(i)})\, ds \\ & + A_+^{(i)}\, h_+^{(i)}(L^{(i)})\, \exp(-\mu^{(i)}\tau^{(i)}(t))\, (r_+^{(i)}(t-\tau^{(i)}(t), L^{(i)}))^2\, \tfrac{d}{dt}\tau^{(i)}(t) .\end{aligned} \tag{39}$$

Using the equation $\partial_s r_+^{(i)}(t-s, L^{(i)}) = -\partial_t r_+^{(i)}(t-s, L^{(i)})$ and integration by parts, from (39) we obtain ($t \in [\overline{\tau}^{(i)}, T]$)

$$\begin{aligned}\tfrac{d}{dt}\mathcal{D}^{(i)}(t) = {} & -\mu^{(i)}\mathcal{D}^{(i)}(t) + A_+^{(i)} h_+^{(i)}(L^{(i)})(r_+^{(i)}(t, L^{(i)}))^2 \\ & - A_+^{(i)} h_+^{(i)}(L^{(i)}) \exp(-\mu^{(i)}\tau^{(i)}(t))(r_+^{(i)}(t-\tau^{(i)}(t), L^{(i)}))^2 (1 - \tfrac{d}{dt}\tau^{(i)}(t)) .\end{aligned}$$

Hence, the inequalities (14) and the definition of $\overline{\tau}^{(i)}$ and $\hat{\tau}^{(i)}$ in (15) imply ($t \in [\overline{\tau}^{(i)}, T]$)

$$\begin{aligned}\tfrac{d}{dt}\mathcal{D}^{(i)}(t) \le {} & -\mu^{(i)}\mathcal{D}^{(i)}(t) + A_+^{(i)} h_+^{(i)}(L^{(i)})(r_+^{(i)}(t, L^{(i)}))^2 \\ & - A_+^{(i)} h_+^{(i)}(L^{(i)}) \exp(-\mu^{(i)}\overline{\tau}^{(i)})(r_+^{(i)}(t-\tau^{(i)}(t), L^{(i)}))^2 (1 - \hat{\tau}^{(i)}) .\end{aligned} \tag{40}$$

From (38) and (40) we obtain the following estimate for $\frac{d}{dt}\mathcal{F}(t)$ with $\eta = \min_{i\in I} \alpha^{(i)}\beta^{(i)}\mu^{(i)}$ $(t \in [\tau_{max}, T])$:

$$\frac{d}{dt}\mathcal{F}(t) \;\leq\; -\,\eta\,\mathcal{F}(t) \;+\; B_0(t) \;+\; B_L(t) \tag{41}$$

with the boundary terms

$$\begin{aligned} B_0(t) &= \sum_{i\in I} A_+^{(i)}(r_+^{(i)}(t,0))^2 \;-\; A_-^{(i)}(r_-^{(i)}(t,0))^2\,, \\ B_L(t) &= \sum_{i\in I} \Big[A_-^{(i)}\,h_-^{(i)}(L^{(i)})\,(r_-^{(i)}(t,L^{(i)}))^2 \\ &\quad -\; A_+^{(i)}\,h_+^{(i)}(L^{(i)})\,\exp(-\mu^{(i)}\overline{\tau}^{(i)})\,(1-\hat{\tau}^{(i)})\,(r_+^{(i)}(t-\tau^{(i)}(t),L^{(i)}))^2\Big]\,. \end{aligned}$$

The equation (13), the orthogonality of the matrix A_ω from (7) and the inequalities (31) yield that we have $B_0(t) \leq 0$ for all $t \in [\tau_{max}, T]$. Furthermore, the boundary controls (16) and the inequality (33) guarantee $B_L(t) \leq 0$ for all $t \in [2\tau_{max}, T]$. Thus, from (41) we obtain $(t \in [2\tau_{max}, T])$

$$\frac{d}{dt}\mathcal{F}(t) \;\leq\; -\,\eta\,\mathcal{F}(t)$$

which implies the inequality (36).

References

1. Banda, M.K., Herty, M., Klar, A.: Coupling conditions for gas networks governed by the isothermal Euler equations. Netw. Heterog. Media 1, 295–314 (2006)
2. Banda, M.K., Herty, M., Klar, A.: Gas flow in pipeline networks. Netw. Heterog. Media 1, 41–56 (2006)
3. Bogel, G.D.: Method and apparatus for control of pipeline compressors. United States Patent Number 4526513 (1985)
4. Coron, J.-M., d'Andréa-Novel, B., Bastin, G.: A strict Lyapunov function for boundary control of hyperbolic systems of conservation laws. IEEE Trans. Automat. Control 52, 2–11 (2007)
5. Dick, M., Gugat, M., Leugering, G.: Classical solutions and feedback stabilization for the gas flow in a sequence of pipes. Netw. Heterog. Media 5, 691–709 (2010)
6. Gugat, M.: Boundary feedback stabilization by time delay for one-dimensional wave equations. IMA J. Math. Control Inform. 27, 189–203 (2010)
7. Gugat, M., Dick, M.: Time-delayed boundary feedback stabilization of the isothermal Euler equations with friction. Math. Control Relat. Fields 1, 469–491 (2011)
8. Gugat, M., Herty, M.: Existence of classical solutions and feedback stabilization for the flow in gas networks. ESAIM Control Optim. Calc. Var. 17, 28–51 (2011)
9. Gugat, M., Sigalotti, M.: Stars of vibrating strings: switching boundary feedback stabilization. Netw. Heterog. Media 5, 299–314 (2010)
10. Herty, M., Mohring, J., Sachers, V.: A new model for gas flow in pipe networks. Math. Methods Appl. Sci. 33, 845–855 (2010)

11. Marigo, A.: Entropic solutions for irrigation networks. SIAM J. Appl. Math. 70, 1711–1735 (2009/2010)
12. Nicaise, S., Valein, J., Fridman, E.: Stability of the heat and of the wave equations with boundary time-varying delays. Discrete Contin. Dyn. Syst. Ser. S 2, 559–581 (2009)
13. Osiadacz, A.: Simulation and Analysis of Gas Networks. Gulf Publishing Company, Houston (1987)
14. Wang, J.-M., Guo, B.-Z., Krstic, M.: Wave equation stabilization by delays equal to even multiples of the wave propagation time. SIAM J. Control Optim. 49, 517–554 (2011)
15. Wang, Z.: Exact controllability for nonautonomous first order quasilinear hyperbolic systems. Chinese Ann. Math. Ser. B 27, 643–656 (2006)

Real-Time Nonlinear Model Predictive Control of a Glass Forming Process Using a Finite Element Model

Janko Petereit and Thomas Bernard

Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB, Fraunhoferstr. 1, 76131 Karlsruhe, Germany
{Janko.Petereit,Thomas.Bernard}@iosb.fraunhofer.de

Abstract. The control of complex forming processes (e.g., glass forming processes) is a challenging topic due to the mostly strongly nonlinear behavior and the spatially distributed nature of the process. In this paper a new approach for the real-time control of a spatially distributed temperature profile of an industrial glass forming process is presented. As the temperature in the forming zone cannot be measured directly, it is estimated by the numerical solution of the partial differential equation for heat transfer by a finite element scheme. The numerical solution of the optimization problem is performed by the solver HQP (Huge Quadratic Programming). In order to meet real-time requirements, in each sampling interval the full finite element discretization of the temperature profile is reduced considerably by a spline approximation. Results of the NMPC concept are compared with conventional PI control results. It is shown that NMPC stabilizes the temperature of the forming zone much better than PI control. The proposed NMPC scheme is robust against model mismatch of the disturbance model. Furthermore, the allowed parameter settings for a real-time application (i.e., control horizon, sampling period) have been determined. The approach can easily be adapted to other forming processes where the temperature profile shall be controlled.

1 Introduction

The control of complex forming processes (e.g., glass forming processes) is a challenging topic due to the mostly strongly nonlinear behavior and the spatially distributed nature of the process. The nonlinearity is caused on the one hand by the physics of the process (e.g., radiation) and on the other hand by nonlinear material properties (e.g., memory effect). In recent years considerable progress has been made in the field of simulation tools (e.g., finite element based simulation of complex rheological processes [11], [1]) and model order reduction (MOR) methods [4], [3]. Consequently, the use of these (reduced) models for process control and optimization is in the focus of actual research activities [2]. Nonlinear Model Predictive Control (NMPC) concepts play an important role in this context, as NMPC can be applied for nonlinear and spatially distributed dynamic systems. Main advantages of NMPC approaches are that the

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 266–275, 2013.

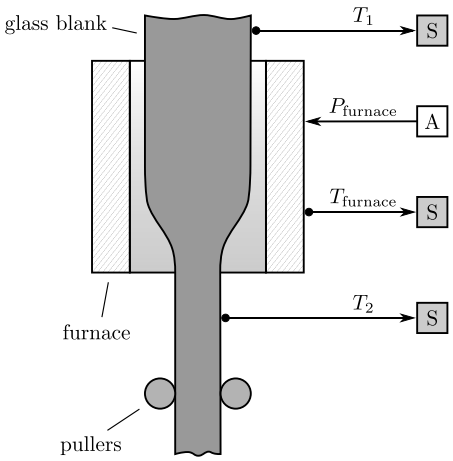


Fig. 1. Glass forming process overview

performance criteria can be designed in a transparent way and that even time dependent constraints can be applied to manipulated, output and state variables of the controlled system. Drawbacks of NMPC are that computational costs for the solution of the nonlinear programming problems often are high, hence powerful optimization solvers are needed [8], [9]. A suited solver for large scale dynamic optimization problems is HQP (Huge Quadratic Programming, [10]). HQP has been used e.g., for NMPC appplications in process industry [12], power plants [7] as well as for optimal management of water resources [5].

For complex forming processes in many cases an optimal spatially distributed (and sometimes time-dependent) viscosity profile of the material has to be assured. As viscosity is difficult to measure on-line, temperature often is used as an auxiliary controlled variable. Hence the problem consists in controlling a spatially distributed temperature profile. In [6] a first NMPC concept has been investigated. In this work a predictive functional control scheme (PFC) has been applied in order to optimize the forming control loops, but not the temperature control. Only a very short prediction horizon is assumed in the mentioned simulation studyy. In the present paper a new approach for the real-time control of a spatially distributed temperature profile is presented. As the temperature in the forming zone cannot be measured directly, it is estimated by the numerical solution of the partial differential equation (PDE) for heat transfer by a finite element (FE) scheme. As the dimension of the state space model, which is result of the FE algorithm, is too large for real-time optimization, in each sampling interval the full finite element discretization of the temperature profile is reduced considerably by a spline approximation. The numerical solution of the optimization problem is performed by HQP. The approach can easily be adapted to other forming processes where the temperature profile shall be controlled.

The paper is organized as follows. Section 2 describes an industrial glass forming process to which the NMPC approach is applied. The details of the proposed

NMPC concept are presented in section 3. In section 4 the results of the NMPC concept are discussed. Section 5 gives some conclusions.

2 Optimal Temperature Profile of an Industrial Glass Forming Process

The NMPC concept which will be introduced in the next section has been applied to an industrial glass forming process (Fig. 1). A vertically hung glass cylinder is fed into a ring-shaped furnace where it begins to liquefy. It then starts flowing downwards viscously and is taken up by pullers which pull the resulting glass tube with an appropriate speed. It is essential for the process to stabilize the viscosity profile along the cylinder. The viscosity can be controlled by the furnace temperature.

Figure 1 depicts the relevant sensors and actuators regarding the temperature control: As the furnace tightly encloses the glass cylinder, there are only three temperature sensors available – one each right above (T_{cyl}) and below (T_{tube}) the furnace, and one which measures the furnace temperature (T_{furnace}) itself. There is no possibility to measure the temperature in the forming zone (T_{form}) directly. Thus, a strategy is needed to reconstruct this control variable using the measurements of the only three available sensors. This is achieved by a finite element model. The forming temperature is controlled by increasing or decreasing the heating power and hence furnace temperature.

In order to ensure optimal product quality there are three major variables which must be controlled in an optimal way. The first two are the feeding and pulling speed (respectively force) which are accounted for by a common control loop. We call this one the *geometry control*. The third variable is the (spatially distributed) temperature of the glass in the forming zone, whose control is done by a seperate loop, which we call the *temperature control*. In the scope of this paper we will concentrate on the latter, i.e., the temperature control.

For glass the system's behavior is in general strongly nonlinear due to the impact of radiation. The spatio-temporal temperature distribiution can be calculated solving the heat transfer PDE. If a symmetric cylinder is assumed and radial temperature distribution the following one dimensional PDE describes the temperature distribution $T(z,t)$ along the vertical z-axis:

$$\rho c_p(T)\frac{\partial T}{\partial t} = \lambda(T)\frac{\partial^2 T}{\partial z^2} + \dot{q}_{\mathrm{conv}}(T,z) + \dot{q}_{\mathrm{rad,oven}}(T,z) + \dot{q}_{\mathrm{rad,dist}}(z,t) \qquad (1)$$

Radiative heat transfer exchange between the oven and the glass is described by the nonlinear term $\dot{q}_{\mathrm{rad,oven}}$ (Stefan-Boltzmann law). Radiation inside the glass is considered by means of an effective heat conduction coefficient $\lambda(T)$ which is nearly exponential increasing with temperature and hence introduces a second nonlinearity in the model. The term $\dot{q}_{\mathrm{conv}}$ describes the convective heat transfer due to the movement of the glass cylinder. A further nonlinearity is introduced by temperature dependence of the specific heat capacity $c_p(T)$.

At the end of a batch production (i.e., the when the length of the cylinder which is fed on the oven tends to zero), disturbances arise due to radiation effects which are caused by the end of the cylinder. This radiative disturbance are described by the space and time dependent term $\dot{q}_{\mathrm{rad,dist}}(z,t)$. As the disturbance cannot be measured directly, it is estimated online by a special parameter estimation approach.

Due to the transport length and the corresponding dead time it is important that the disturbances are suppressed as early as possible by an optimal strategy of furnace temperature during process end phase. As the steadiness of the viscosity in the forming zone exerts the main impact upon the quality of the final product, an optimal control must stabilize the spatially distributed temperature (and hence viscosity) profile of the forming zone as good as possible.

To accommodate for the deficiencies of the PI controller we developed a control concept which explicitly takes into account the large dead time of the process and the disturbances arising during the end phase. As a PI controller can only react to changes in the control variable which are already in effect, we switched over to a predictive control strategy, which is able to take corrective action before any impact of the disturbances is visible to the temperature sensors. This behavior is necessary since the delay time of the furnace is too large to lower the temperature in sufficient time. Otherwise, overshooting of the forming temperature would be unavoidable.

As the prediction model for the NMPC scheme we exploit the finite element model of the 1D temperature distribution $T(z,t)$ along the vertical z-axis by solving the PDE (1) online. The finite element model considers heat conduction, convection by the movement of the glass, radiation and the time dependent disturbances caused by the end of the cylinder.

3 Concept for Real-Time Optimization of a Spatially Distributed Temperature Profile

In our solution concept (Fig. 2) we propose a real-time nonlinear model predictive control of the temperature of a forming zone (here for glass forming processes). In most cases the temperature profile cannot be measured directly, hence it has to be estimated by a spatially distributed model. The one dimensional heat transfer PDE (1) has to be solved numerically for the temperature distribution $T(z,t)$, where z denotes the height and t the time. The time-dependent disturbances at the end of a cylinder are estimated by a disturbance model using the knowledge from previous productions to determine the model's parameters in a predictive way.

The NMPC approach is formulated as follows. The predicted values of the controlled variable $y(t)$ are collected in the vector $\hat{\boldsymbol{y}}$ for the timesteps from $t_k ... t_{k+n_\mathrm{p}}$, where t_k is the actual time and n_p the prediction horizon. The predicted values are calculated by the model

$$\hat{\boldsymbol{y}} = \boldsymbol{F}(\boldsymbol{x}_k, \Delta \boldsymbol{u}) \tag{2}$$

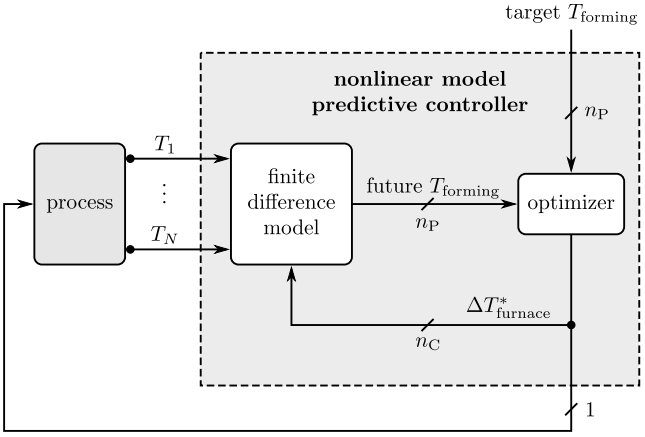


Fig. 2. NMPC temperature control scheme

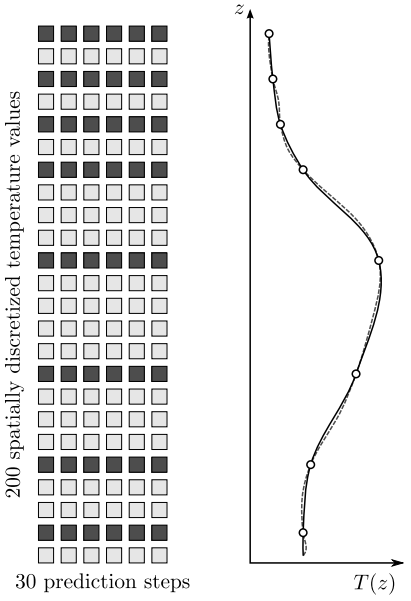


Fig. 3. Spline approximation

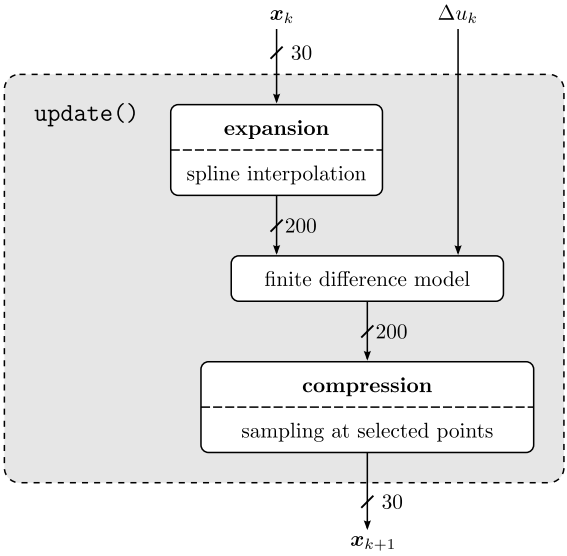


Fig. 4. HQP update step

where $\boldsymbol{x}_k$ denotes the actual state of the model and the vector $\boldsymbol{\Delta u}$ denotes all changes of the controlled variables from the actual time step t_k to the prediction horizon (we set prediction horizon n_p equal control horizon n_p). The optimization

$$J = (\boldsymbol{w} - \hat{\boldsymbol{y}})^\top \boldsymbol{Q}\,(\boldsymbol{w} - \hat{\boldsymbol{y}}) + \boldsymbol{\Delta u}^\top \boldsymbol{R}\,\boldsymbol{\Delta u} \tag{3}$$

$$\boldsymbol{\Delta u}_\mathrm{opt} = \underset{\boldsymbol{\Delta u}}{\operatorname{argmin}}\, J \tag{4}$$

is performed subject to

$$\begin{aligned} \Delta u_\mathrm{min} \le \Delta u(k+j|k) \le \Delta u_\mathrm{max} \qquad & j = 0, \ldots, n_\mathrm{C} - 1 \\ \hat{y}_\mathrm{min} \le \hat{y}(k+j|k) \le \hat{y}_\mathrm{max} \qquad & j = 1, \ldots, n_\mathrm{P} \end{aligned} \tag{5}$$

In this formulation the *changes* of the manipulated variable u are calculated. If necessary, also the absolute value u could be considered directly in the optimization problem.

The proposed control concept has been implemented in a C++ application by utilizing the software suite HQP (Huge Quadratic Programming), a solver for large scale nonlinear programming problems, to solve the arising nonlinear optimal control problems. In order to meet real-time requirements, in each sampling interval the full finite element discretization of the temperature profile is

reduced considerably by a spline approximation (see Fig. 4 and 3). The spline approximation is admissible due to the temperature profile's smoothness. In our application the number of states could be reduced from 200 to 30.

4 Results

We evaluated the control performance resulting with the proposed NMPC concept and compared it to a PI controller using previously recorded production data. It is shown that NMPC stabilizes the temperature of the forming zone much better than PI control. As can be seen in Fig. 5 with very small prediction and control horizon ($n_\mathrm{p} = n_\mathrm{c} = 2$) the performance of NMPC and PI control do not differ very much. In subplot 1 of Fig. 5 the deviation e_forming from the reference value is shown for a production time of 40 minutes. In subplot 2 the change of the furnace temperature (manipulated variable) compared to the value at starting time $t = 0$ min. can be seen (i.e., $\bar{T}_\mathrm{furnace}(t) = T_\mathrm{furnace}(t) - T_\mathrm{furnace}(0)$). Subplot 3 shows the changes of furnace temperature in each sample time. These changes have been restricted to $\pm 5\,\mathrm{K}$ during the optimization in order to guarantee a smooth change of the furnace temperature. Finally, in subplot 4 the computation time for each sample interval (30 s) is shown. It can be seen that the computation time does not differ very much (about 2 s). Fig. 6 shows the results with a much larger prediction and control horizon ($n_\mathrm{p} = n_\mathrm{c} = 30$). From subplot 1 it can be observerd that the control deviation with NMPC is much smaller compared to the results with PI control. Accordingly, the decrease of the furnace temperature (subplot 2) with NMPC starts much earlier. This is caused by the internal disturbance model which predicts the disturbances early enough. From subplot 3 it can be seen that the defined threshold for changes of the furnace temperature are reached for a period of about 10 minutes. In subplot 4 it can be seen that the computation is always smaller than the sample time of 30 s, hence the real-time condition is satisfied. The computation time decreases strongly in the last 15 minutes of the production due to the shortened horizon.

The allowed parameter settings for a real-time application (e.g., control horizon, sampling period) have been determined. In Fig. 7 the computation time is plotted versus the prediction and control horizon (n_p, n_c). The internal time discretization $d\tau$ of the finite element model ($n_\mathrm{FDM} = T_\mathrm{s}/d\tau$) is used as parameter ($T_\mathrm{s}$: sample time). From Fig. 7 it is obvious that for a wide range of parameter values n_p, n_c and n_FDM real-time application (computation time $<$ sample time $= 30\,\mathrm{s}$) is possible.

Furthermore, simulation based application of the proposed NMPC scheme to a large number of historical runs showed that the concept is robust against model mismatch of the disturbance model. This property is very important for the industrial implementation.

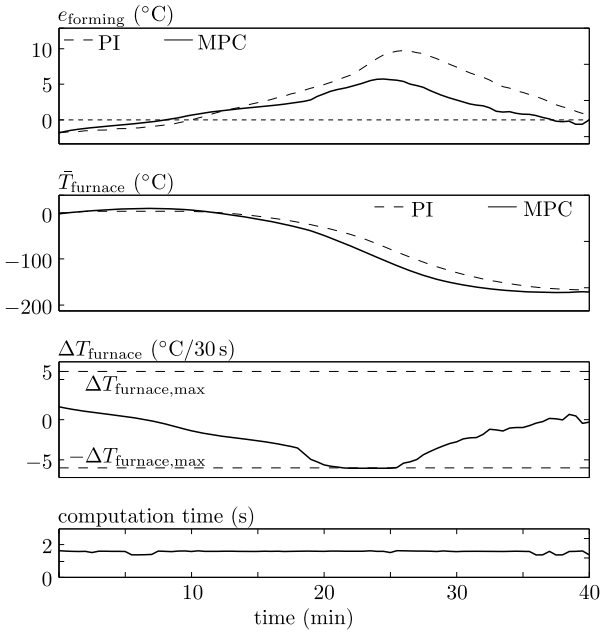


Fig. 5. Comparison NMPC and PI control ($n_{\text{P}} = n_{\text{C}} = 2$)

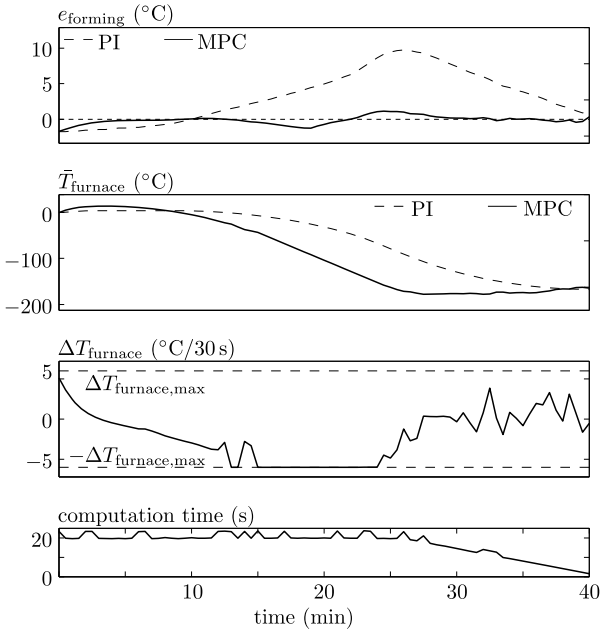


Fig. 6. Comparison NMPC and PI control ($n_{\text{P}} = n_{\text{C}} = 30$)

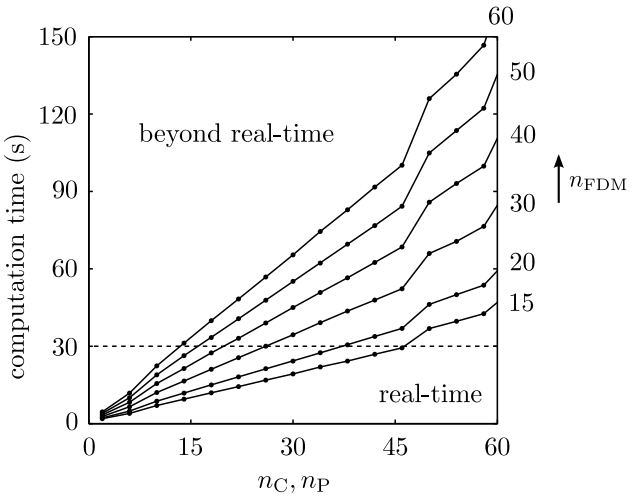


Fig. 7. Computation time vs. prediction horizon

5 Conclusion

In this paper a new approach for the real-time control of a spatially distributed temperature profile of an industrial glass forming process has been presented. A numerically efficient solution of the partial differential equation has been implemented. By means of the powerful solver HQP a real-time application could be achieved. It has been shown that with the proposed NMPC concept much better disturbance rejection and hence stabilization of the temperature profile than conventional PI control is achieved. The approach can easily be adapted to other forming processes where the temperature profile shall be controlled.

References

1. Glass Days 2005 – Analysis and Simulation of Processes in Glass Production and Processing, Kaiserslautern, Germany (April 14-15, 2005)
2. EU project PROMATCH (PROmoting and structuring Multidisciplinary Academic-industrial collaboration in research and Training trough SME teCHnology developers) (2008), http://promatch.ele.tue.nl/
3. Antoulas, A.C.: Approximation of Large-Scale Dynamical Systems. SIAM Press, Philadelphia (2005)
4. Benner, P., Sorensen, D.C., Mehrmann, V.: Dimension Reduction of Large-Scale Systems. Springer, Berlin (2005)
5. Bernard, T., Krol, O., Linke, H., Rauschenbach, T.: Optimal management of regional water supply systems using a reduced finite-element groundwater model. at – Automatisierungstechnik 57(12), 593–600 (2008)
6. Bernard, T., Moghaddam, E.E.: Nonlinear model predictive control of a glass forming process based on a finite element model. In: IEEE Conference on Control Applications (CCA 2006), Munich, Germany (October 2006)

7. D'Amato, F.J.: Industrial application of a model predictive control solution for power plant startups. In: Proc. 2006 IEEE International Conference on Control Applications (2006)
8. Diehl, M., et al.: An efficient algorithm for nonlinear model predictive control of large-scale systems part 1. at – Automatisierungstechnik 50(12) (2002)
9. Diehl, M., et al.: An efficient algorithm for nonlinear model predictive control of large-scale systems part 2. at – Automatisierungstechnik 51(1) (2003)
10. Franke, R., Arnold, E.: The solver Omuses/HQP for structured largescale constrained optimization: algorithm, implementation and example application. In: Sixth SIAM Conference on Optimization (1999), http://hqp.sourceforge.net/index.html
11. Krause, D.: Mathematical Simulation in Glass Technology. Springer (2002)
12. Nagy, Z.K., Franke, R., Mahn, B., Allgöwer, F.: Real-time Implementation of Nonlinear Model Predictive Control of Batch Processes in an Industrial Framework. In: Proc. NMPC 2005, Freudenstadt, Germany, August 26-30 (2005)

Exponential Stability of the System of Transmission of the Wave Equation with a Delay Term in the Boundary Feedback

Salah-Eddine Rebiai

Laboratoire des Techniques Mathématiques, Faculté des Sciences,
Université de Batna, 05000 Batna, Algeria
rebiai@hotmai.com

Abstract. We consider a system of transmission of the wave equation with Neumann feedback control that contains a delay term and that acts on the exterior boundary. First, we prove under some assumptions that the closed-loop system generates a C_0-semigroup of contractions on an appropriate Hilbert space. Then, under further assumptions, we show that the closed-loop system is exponentially stable. To establish this result, we introduce a suitable energy function and use multiplier method together with an estimate taken from [3] (Lemma 7.2) and compactness-uniqueness arguments.

Keywords: Wave equation, transmission problem, time delays, boundary stabilization, exponential stability.

1 Introduction

It is by now well-known that certain infinite-dimensional second-order systems are not robust with respect to arbitarily small delays in the damping. This lack of stability robustness was first shown to hold for the one-dimensional wave equation ([2]). Later, further examples illustrating this phenomenon were considered in [1]: the two-dimensional wave equation with damping introduced through Neumann-type boundary conditions on one edge of a square boundary and the Euler-Bernoulli beam equation in one dimension with damping introduced through a specific set of boundary conditions on the right end point.

Recently, Xu et al [9] established sufficient conditions that guarantee the exponential stability of the one-dimensional wave equation with a delay term in the boundary feedback. Nicaise and Pignotti [5] extended this result to the multi-dimensional wave equation with a delay term in the boundary or internal feedbacks. The same type of result was obtained by Nicaise and Rebiai [6] for the Schrödinger equation.

Motivated by the references [9], [5] and [6]; we investigate in this paper the problem of exponential stability for the system of transmission of the wave equation with a delay term in the boundary feedback.

Let Ω be an open bounded domain of $\mathbb{R}^n$ with a boundary Γ of class C^2 which consists of two non-empty parts Γ_1 and Γ_2 such that $\overline{\Gamma_1} \cap \overline{\Gamma_2} = \emptyset$. Let Γ_0 with

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 276–285, 2013.

$\overline{\Gamma_0} \cap \overline{\Gamma_1} = \overline{\Gamma_0} \cap \overline{\Gamma_2} = \emptyset$ be a regular hypersurface of class C^2 which separates Ω into two domains Ω_1 and Ω_2 such that $\Gamma_1 \subset \partial\Omega_1$ and $\Gamma_2 \subset \partial\Omega_2$. Furthermore, we assume that there exists a real vector field $h \in (C^2(\overline{\Omega}))^n$ such that:
$(H.1)$ The Jacobian matrix J of h satisfies

$$\int_\Omega J(x)\zeta(x).\zeta(x)d\Omega \geq \alpha \int_\Omega |\zeta(x)|^2 \, d\Omega,$$

for some constant $\alpha > 0$ and for all $\zeta \in L^2(\Omega; \mathbb{R}^n)$;
$(H.2)$ $h(x).\nu(x) \leq 0$ on Γ_1;

$(H.3)$ $h(x).\nu(x) \geq 0$ on Γ_0.
where ν is the unit normal on Γ or Γ_0 pointing towrds the exterior of Ω or Ω_1.

Let $a_1, a_2 > 0$ be given. Consider the system of transmission of the wave equation with a delay term in the boundary conditions:

$$y''(x,t) - a(x)\Delta y(x,t) = 0 \qquad \text{in } \Omega \times (0,+\infty), \tag{1}$$

$$y(x,0) = y^0(x), y'(x,0) = y^1(x,0) \qquad \text{in } \Omega, \tag{2}$$

$$y_1(x,t) = 0 \qquad \text{on } \Gamma_1 \times (0,+\infty), \tag{3}$$

$$\frac{\partial y_2(x,t)}{\partial \nu} = -\mu_1 y_2'(x,t) - \mu_2 y_2'(x,t-\tau) \quad \text{on } \Gamma_2 \times (0,+\infty), \tag{4}$$

$$y_1(x,t) = y_2(x,t), \qquad \text{on } \Gamma_0 \times (0,+\infty), \tag{5}$$

$$a_1 \frac{\partial y_1(x,t)}{\partial \nu} = a_2 \frac{\partial y_2(x,t)}{\partial \nu} \qquad \text{on } \Gamma_0 \times (0,+\infty), \tag{6}$$

$$y_2'(x,t-\tau) = f_0(x,t-\tau) \qquad \text{on } \Gamma_2 \times (0,\tau). \tag{7}$$

where:

- $a(x) = \begin{cases} a_1, & x \in \Omega_1 \\ a_2, & x \in \Omega_2 \end{cases}$
- $y(x,t) = \begin{cases} y_1(x,t), (x,t) \in \Omega_1 \times (0,+\infty) \\ y_1(x,t), (x,t) \in \Omega_2 \times (0,+\infty) \end{cases}$
- $\frac{\partial .}{\partial \nu}$ is the normal derivative.
- μ_1 and μ_2 are positive real numbers.
- τ is the time delay

- y^0, y^1, f_0 are the initial data which belong to suitable spaces.

In the absence of delay, that is $\mu_2 = 0$, Liu and Williams [4] have shown that the solution of (1)-(6) decays exponentially to zero in the energy space $H^1_{\Gamma_1}(\Omega) \times L^2(\Omega)$ provided that

$$a_1 > a_2 \tag{8}$$

and $\{\Omega, \Gamma_0, \Gamma_1, \Gamma_2\}$ satisfies $(H.1), (H.2), (H.3)$, and
$(H.4)$ $h(x).\nu(x) \geq \gamma > 0$.

The purpose of this paper is to investigate the stability of problem (1) – (7) in the case where both μ_1 and μ_2 are different from zero. To this end, assume as in [5] that

$$\mu_1 > \mu_2. \tag{9}$$

and define the energy of a solution of (1) – (7) by

$$E(t) = \frac{1}{2}\int_\Omega \left[|y'(x,t)|^2 + a(x)\,|\nabla(y(x,t)|^2\right] dx + \frac{\xi}{2}\int_{\Gamma_2}\int_0^1 |y(x,t-\tau\rho)|^2\, d\rho\, d\sigma(x), \tag{10}$$

where

$$a_2\tau\mu_2 < \xi < a_2\tau(2\mu_1 - \mu_2), \tag{11}$$

We show that if $\{\Omega, \Gamma_0, \Gamma_1, \Gamma_2\}$ satisfies $(H.1), (H.2)$ and $(H.3)$, then there is an exponential decay rate for $E(t)$.The proof of this result combines multipliers technique and compactness-uniqueness arguments.
The main result of this paper can be stated as follows.

Theorem 1. *Assume* $(H1)$, $(H.2), (H.3)$, (8) *and* (9). *Then there exist constants* $M \geq 1$ *and* $\omega > 0$ *such that*

$$E(t) \leq Me^{-\omega t}E(0).$$

Theorem 1 is proved in Section 3. In Section 2, we investigate the well-posedness of system (1) – (7) using semigroup theory.

2 Well-Poseness of Problem (1) – (7)

Inspired from [5] and [6], we introduce the auxilliary variable $z(x,\rho,t) = y(x, t-\tau\rho)$. With this new unknown, problem (1) – (7) is equivalent to

$$y''(x,t) - a(x)\Delta y(x,t) = 0 \qquad \text{in } \Omega\times(0,+\infty), \tag{12}$$
$$y(x,0) = y^0(x), y'(x,0) = y^1(x) \qquad \text{in } \Omega, \tag{13}$$
$$y(x,t) = 0 \qquad \text{on } \Gamma_1\times(0,+\infty), \tag{14}$$
$$\frac{\partial z(x,\rho,t)}{\partial t} + \frac{1}{\tau}\frac{\partial z(x,\rho,t)}{\partial \rho} = 0 \qquad \text{on } \Gamma_2\times(0,+\infty) \tag{15}$$
$$\frac{\partial y_2(x,t)}{\partial \nu} = -\mu_1 y_2'(x,t) - \mu_2 z(x,1,t) \quad \text{on } \Gamma_2\times(0,+\infty), \tag{16}$$
$$y_1(x,t) = y_2(x,t) \qquad \text{on } \Gamma_0\times(0,+\infty), \tag{17}$$
$$a_1\frac{\partial y_1(x,t)}{\partial \nu} = a_2\frac{\partial y_2(x,t)}{\partial \nu} \qquad \text{on } \Gamma_0\times(0,+\infty), \tag{18}$$
$$z(x,0,t) = y'(x,t) \qquad \text{on } \Gamma_2\times(0,+\infty) \tag{19}$$
$$z(x,\rho,0) = f_0(x,-\tau\rho) \qquad \text{on } \Gamma_2\times(0,1) \tag{20}$$

Now, we endow the Hilbert space

$$\mathcal{H} = H^1_{\Gamma_1}(\Omega)\times L^2(\Omega)\times L^2(\Gamma_2; L^2(0,1))$$

with the inner product

$$\left\langle \begin{pmatrix} u \\ v \\ z \end{pmatrix}; \begin{pmatrix} \overline{u} \\ \overline{v} \\ \overline{z} \end{pmatrix} \right\rangle = \int_\Omega (a(x)\nabla u(x)\nabla \overline{u}(x) + v(x)\overline{v}(x))\, dx + \xi \int_{\Gamma_2} \int_0^1 z(x,\rho)\overline{z}(x,\rho) d\rho\, d\sigma(x)$$

and define a linear operator in $\mathcal{H}$ by

$$\begin{aligned} D(A) = \{(u,v,z)^T \in H^2(\Omega_1,\Omega_2,\Gamma_1) \times H^1_{\Gamma_1}(\Omega) \times L^2(\Gamma_2; H^1(0,1)); \\ \frac{\partial u}{\partial \nu} = -\mu_1 v - \mu_2 z(.,1), v = z(.,0) \text{ on } \Gamma_2\} \end{aligned} \tag{21}$$

$$A \begin{pmatrix} u \\ v \\ z \end{pmatrix} = \begin{pmatrix} v \\ a(x)\Delta u \\ -\tau^{-1}\frac{\partial z}{\partial \rho} \end{pmatrix} \tag{22}$$

The spaces used for the definition of $\mathcal{H}$ and $D(A)$ are

$$H^1_{\Gamma_1}(\Omega) = \{u \in H^1(\Omega) : u = 0 \text{ on } \Gamma_1\}$$

$$H^2(\Omega_1,\Omega_2,\Gamma_1) = \{u_i \in H^2(\Omega_i) : u = 0 \text{ on } \Gamma_1,\ u_1 = u_2 \text{ and } a_1\frac{\partial u_1}{\partial \nu} = a_2\frac{\partial u_2}{\partial \nu} \text{ on } \Gamma_0\}$$

Then we can rewrite (12) − (20) as an abstract Cauchy problem in $\mathcal{H}$

$$\begin{cases} \frac{d}{dt}Y(t) = AY(t) \\ Y(0) = Y_0 \end{cases} \tag{23}$$

where

$$Y(t) = (y, y', z)^T \text{ and } Y_0 = (y_0, y_1, f_0(.,-.\tau))^T$$

Proposition 1. *The operator A defined by (21) and (22) generates a strongly continuous semigroup on $\mathcal{H}$. Thus, for every $Y_0 \in \mathcal{H}$, problem (23) has a unique solution Y whose regularity depends on the the initial datum Y_0 as follows:*

$$Y(.) \in C([0,+\infty); \mathcal{H}) \text{ if } Y_0 \in \mathcal{H},$$
$$Y(.) \in C([0,+\infty); D(A)) \cap C^1([0,+\infty); \mathcal{H}) \text{ if } Y_0 \in D(A).$$

Proof. Let $Y = \begin{pmatrix} u \\ v \\ z \end{pmatrix} \in D(A)$. Then

$$\begin{aligned} \langle AY, Y\rangle = \int_\Omega a(x)\nabla u(x).\nabla v(x)dx + \int_\Omega (a(x)\Delta u(x))v(x)dx - \\ \frac{\xi}{\tau}\int_{\Gamma_2}\int_0^1 z_\rho(x,\rho)z(x,\rho)d\rho d\Gamma \end{aligned} \tag{24}$$

Applying Green's first theorem, we obtain

$$\int_{\Omega}(a(x)\Delta u(x))v(x)dx = a_1\int_{\Gamma_1} v(x)\frac{\partial u(x)}{\partial \nu}d\Gamma - a_1\int_{\Omega_1}\nabla u(x).\nabla v(x)dx + a_2\int_{\Gamma_2} v(x)\frac{\partial u(x)}{\partial \nu}d\Gamma - a_2\int_{\Omega_2}\nabla u(x).\nabla v(x)dx$$
$$= a_2\int_{\Gamma_2} v(x)\{-\mu_1 v(x) - \mu_2 z(x,1)\}d\Gamma - \int_{\Omega} a(x)\nabla u(x).\nabla v(x)dx \tag{25}$$

Integrating by parts in ρ, we get

$$\int_{\Gamma_2}\int_0^1 z_\rho(x,\rho)z(x,\rho)d\rho d\Gamma = \frac{1}{2}\int_{\Gamma_2}\{z^2(x,1) - z^2(x,0)\}d\Gamma \tag{26}$$

Inserting (25) and (26) into (24) results in

$$\langle AY, Y\rangle = -a_2\mu_1\int_{\Gamma_2} v^2(x)d\Gamma - a_2\mu_2\int_{\Gamma_2} v(x)z(x,1)d\Gamma - \frac{\xi}{2\tau}\int_{\Gamma_2} z^2(x,1)d\Gamma + \frac{\xi}{2\tau}\int_{\Gamma_2} v^2(x)d\Gamma$$

from which follows using the Cauchy-Schwarz inequality

$$\langle AY, Y\rangle \leq -(a_2\mu_1 - \frac{a_2\mu_2}{2} + \frac{\xi}{2\tau})\int_{\Gamma_2} v^2(x)d\Gamma - (\frac{\xi}{2\tau} - \frac{a_2\mu_2}{2})\int_{\Gamma_2} z^2(x,1)d\Gamma \tag{27}$$

(27) implies that

$$\langle AY, Y\rangle \leq 0$$

Thus A is dissipative.
Now we show that for a fixed $\lambda > 0$ and $(g,h,k)^T \in \mathcal{H}$, there exists $Y = (u,v,z)^T \in D(A)$ such that

$$(\lambda I - A)Y = (g,h,k)^T$$

or equivalently

$$\lambda u - v = g \tag{28}$$
$$\lambda v - a(x)\Delta u = h \tag{29}$$
$$\lambda z + \frac{1}{\tau}z_\rho = k \tag{30}$$

Suppose that we have found u with the appropriate regularity, then we can determine z. Indeed, from (21) and (30) we have

$$\begin{cases} z_\rho(x,\rho) = -\lambda\tau z(x,\rho) + \tau k(x,\rho) \\ z(x,0) = v(x) \end{cases}$$

The unique solution of the above initial value problem is

$$z(x,\rho) = e^{-\lambda\tau\rho}v(x) + \tau e^{-\lambda\tau\rho}\int_0^\rho e^{\lambda\tau s}k(x,s)ds$$

and in particular

$$z(x,1) = \lambda e^{-\lambda\tau}u(x) + z_0(x), \quad x \in \Gamma_2$$

where

$$z_0(x) = -e^{-\lambda\tau}g(x) + \tau e^{-\lambda\tau}\int_0^1 e^{\lambda\tau s}k(x,s)ds$$

By (28) and (29), the function u satisfies

$$\lambda^2 u - a(x)\Delta u = h + \lambda g \tag{31}$$

Problem (31) can be reformulated as

$$\int_\Omega (\lambda^2 u - a(x)\Delta u)w dx = \int_\Omega (h+\lambda g)w dx, \quad w \in H^1_{\Gamma_1}(\Omega) \tag{32}$$

Using Green's first theorem and recalling (21), we express the right-hand side of (32) as follows

$$\int_\Omega (\lambda^2 u - a(x)\Delta u)w dx = \int_\Omega (\lambda^2 uw + a(x)\nabla u.\nabla w)dx + a_2\int_{\Gamma_2}\{\mu_1(\lambda u - g)w$$
$$+\mu_2(\lambda e^{-\lambda\tau}u(x) + z_0(x))w\}d\Gamma$$

Therefore (32), can be rewritten as

$$\int_\Omega (\lambda^2 uw + a(x)\nabla u.\nabla w)dx + a_2\int_{\Gamma_2}(\mu_1 + \mu_2 e^{-\lambda\tau})\lambda uw d\Gamma = \int_\Omega (h+\lambda g)w d\Gamma$$
$$+a_2\mu_1\int_{\Gamma_2} g w d\Gamma - a_2\mu_2\int_{\Gamma_2} z_0 w d\Gamma, \qquad \forall w \in H^1_{\Gamma_1}(\Omega). \tag{33}$$

Since the left-hand side of (33) is coercive on $H^1_{\Gamma_1}(\Omega)$, the Lax-Milgram Theorem guarantees the existence and uniqueness of a solution $y \in H^1_{\Gamma_1}(\Omega)$ of (31). If we consider $w \in \mathcal{D}(\Omega)$ in (28), then y is a solution in $\mathcal{D}'(\Omega)$ of

$$\lambda^2 u - a(x)\Delta u = h + \lambda g \tag{34}$$

and thus $\Delta u \in L^2(\Omega)$.
Combining (33) together with (34), we obtain after using Green's first theorem

$$a_2\int_{\Gamma_2}(\mu_1 + \mu_2 e^{-\lambda\tau})\lambda uw d\Gamma + a_2\int_\Omega \frac{\partial u}{\partial \nu}w d\Gamma = a_2\mu_1\int_{\Gamma_2} g w d\Gamma - a_2\mu_2\int_{\Gamma_2} z_0 w d\Gamma$$

which implies that

$$\frac{\partial u(x)}{\partial \nu} = -\mu_1 v(x) - \mu_2 z(x,1)$$

So, we have found $(u,v,z)^T \in D(A)$ which satisfies (28) – (30). Thus, by the Lumer-Phillips Theorem (see for instance [8], Theorem 1.4.3), generates a strongly continuous semigroup of contractions on $\mathcal{H}$.

3 Proof of Theorem 1

We prove Theorem 1 for smooth initial data. The general case follows by a standard density argument.
We proceed in several steps.
Step 1.
Since

$$E(t) = \frac{1}{2} \|(y, y', z)\|^2_{\mathcal{H}}$$

Then, we deduce from the proof of Proposition 1 that $E(t)$ is non-increasing and

$$\frac{d}{dt}E(t) \leq -C \int_{\Gamma_2} \{y^2(x,t) + y'^2(x,t)\} d\Gamma \tag{35}$$

where

$$C = \min\{a_2\mu_1 - \frac{a_2\mu_2}{2} + \frac{\xi}{2\tau}, \frac{\xi}{2\tau} - \frac{a_2\mu_2}{2}\}$$

Step 2.
Set

$$E(t) = \mathcal{E}(t) + E_d(t)$$

where

$$\mathcal{E}(t) = \frac{1}{2} \int_\Omega \{a(x) |\nabla y(x,t)|^2 + |y'(x,t)|^2\} dx$$

and

$$E_d(t) = \frac{\xi}{2\tau} \int_{\Gamma_2} \int_0^1 |y'(x, t - \tau\rho)|^2 \, d\rho d\Gamma$$

$E_d(t)$ can be rewritten via a change of variable as

$$E_d(t) = \frac{\xi}{2\tau^2} \int_t^{t+\tau} \int_{\Gamma_2} y'^2(x, s - \tau) d\Gamma ds \tag{36}$$

From (36), we obtain

$$E_d(t) \leq C_1 \int_0^T \int_{\Gamma_2} y'^2(x, s - \tau) d\Gamma ds \tag{37}$$

for $0 \leq t \leq T$ and T large enough.
Step 3.
By applying energy methods (multiplier $2h.\nabla y + (divh - \alpha)y$) (see the appendix) to problem (1) − (7), we obtain for all $T > 0$.

$$\int_0^T \mathcal{E}(t)dt \leq C_2(\mathcal{E}(0) + \mathcal{E}(T)) + C_3 \int_0^T \int_{\Gamma_2} \{(\frac{\partial y(x,t)}{\partial \nu})^2 + y'^2(x,t)\} d\Gamma dt +$$
$$C_4 \int_0^T \int_{\Gamma_2} |\nabla_\sigma y(x,t)|^2 \, d\Gamma dt + C_5 \int_0^T \int_\Omega |y(x,t)|^2 \, d\Omega dt \tag{38}$$

where $\nabla_\sigma y$ is the tangential gradient of y.

Step 4.

We eliminate the tangential gradient from (38) by using the following estimate due to Lasiecka and Triggiani (Lemma 7.2 in [3])

$$\int_\epsilon^{T-\epsilon}\int_{\Gamma_2}|\nabla_\sigma y(x,t)|^2\,d\Gamma dt \le C_6\{\int_0^T\int_{\Gamma_2}\{(\frac{\partial y(x,t)}{\partial\nu})^2+y'^2(x,t)\}d\Gamma dt + \|y\|^2_{L^2(0,T;H^{1/2+\delta}(\Omega))}\}$$

where ϵ and δ are arbitrary positive constants. We obtain

$$\int_0^T\mathcal{E}(t)dt \le C_2(\mathcal{E}(0)+\mathcal{E}(T))+C_7\int_0^T\int_{\Gamma_2}\{(\frac{\partial y(x,t)}{\partial\nu})^2+y'^2(x,t)\}d\Gamma dt + C_8\|y\|^2_{L^2(0,T;H^{1/2+\delta}(\Omega))} \tag{39}$$

Step 5.

We differentiate $\mathcal{E}(t)$ with respect to t and apply Green's first theorem. We obtain

$$\frac{d}{dt}\mathcal{E}(t)=a_2\int_{\Gamma_2}y'(x,t)\frac{\partial y(x,t)}{\partial\nu}d\Gamma dt \tag{40}$$

From (40), we get via the Cauchy-Schwarz inequality

$$\mathcal{E}(0)\le\mathcal{E}(T)+\frac{a_2}{2}\int_0^T\int_{\Gamma_2}\{y'^2(x,t)+(\frac{\partial y(x,t)}{\partial\nu})^2\}d\Gamma dt \tag{41}$$

Insertion of (41) into (39) yields

$$\int_0^T\mathcal{E}(t)dt \le 2C_2\mathcal{E}(T)+C_9\int_0^T\int_{\Gamma_2}\{(\frac{\partial y(x,t)}{\partial\nu})^2+y'^2(x,t)\}d\Gamma dt + C_8\|y\|^2_{L^2(0,T;H^{1/2+\delta}(\Omega))} \tag{42}$$

Step 6.

Since $E(t)$ is non-increasing and $E(t)=\mathcal{E}(t)+E_d(t)$, then (42) together with (37) implies that

$$TE(T)\le 2C_2\mathcal{E}(T)+C_9\int_0^T\int_{\Gamma_2}\{(\frac{\partial y(x,t)}{\partial\nu})^2+y'^2(x,t)\}d\Gamma dt + C_8\|y\|^2_{L^2(0,T;H^{1/2+\delta}(\Omega))}+TC_1\int_0^T\int_{\Gamma_2}y'^2(x,t-\tau)d\Gamma dt \tag{43}$$

Thus invoking again the identity $E(t)=\mathcal{E}(t)+E_d(t)$ and recalling the boundary condition (4), we obtain from (43)

$$E(T)\le C_{10}\int_0^T\int_{\Gamma_2}\{y'^2(x,t)+y'^2(x,t-\tau)\}d\Gamma dt+C_{11}\|y\|^2_{L^2(0,T;H^{1/2+\delta}(\Omega))} \tag{44}$$

for T large enough.

Step 7.
We drop the lower order term on the right-hand side of (44) by a compactness-uniqueness argument to obtain

$$E(T) \leq C_{12} \int_0^T \int_{\Gamma_2} \{y'^2(x,t) + y'^2(x,t-\tau)\} d\Gamma dt \tag{45}$$

Step 8.
The estimate (45) together with (35) yields

$$E(T) \leq \frac{C_{13}}{1+C_{13}} E(0) \tag{46}$$

The desired conclusion follows now from (46) since the system (1) – (7) is invariant by translation.

References

1. Datko, D.: Not all feedback stabilized hyperbolic systems are robust with respect to small time delays in their feedbacks. SIAM J. Control Optim. 26, 697–713 (1988)
2. Datko, R., Lagnese, J., Polis, M.P.: An example on the effect of time delays in boundary feedback stabilization of wave equations. SIAM J. Control Optim. 24, 152–156 (1986)
3. Lasiecka, I., Triggiani, R.: Uniform stabilization of the wave equation with Dirichlet or Neumann feedback control without geometrical conditions. Appl. Math. Optim. 25, 189–244 (1992)
4. Liu, W., Williams, G.H.: The exponential stability of the problem of transmission of the wave equation. Bull. Austra. Math. Soc. 97, 305–327 (1998)
5. Nicaise, S., Pignotti, C.: Stability and instability results of the wave equation with a delay term in the boundary or internal feedbacks. SIAM J. Control Optim. 45, 1561–1585 (2006)
6. Nicaise, N., Rebiai, S.E.: Stabilization of the Schrödinger equation with a delay term in boundary feedback or internal feedback. Portugal. Math. 68, 19–39 (2011)
7. Nicaise, S., Valein, J.: Stabilization of second order evolution equations with unbounded feedback with delay. ESAIM Control Optim. Calc. Var. 16, 420–456 (2010)
8. Pazy, A.: Semigroups of linear operators and applications to partial differential equations. Springer, New York (1983)
9. Xu, G.Q., Yung, S.P., Li, L.K.: Stabilization of wave systems with input delay in the boundary control. ESAIM Control Optim. Calc. Var. 12, 770–785 (2006)

Appendix: Sketch of Proof of (38)

We multiply both sides of (1) by $2h.\nabla y+(divh-\alpha)y$ and integrate over $(0,T)\times\Omega$. We obtain

$$
\begin{aligned}
&2\int_0^T\int_\Omega a(x)J\nabla y.\nabla y d\Omega dt+\alpha\int_0^T\int_\Omega\{y'^2-a(x)\left|\nabla y\right|^2\}d\Omega dt=\\
&-\int_\Omega\{2y'h.\nabla y+(divh-\alpha)y'y\}_0^T d\Omega-\int_0^T\int_\Omega a(x)y\nabla y.\nabla(divh-\alpha)d\Omega dt+\\
&a_1\int_0^T\int_{\Gamma_1}\left|\frac{\partial y_1}{\partial\nu}\right|^2 h.\nu d\Gamma dt-(a_1-a_2)\int_0^T\int_{\Gamma_0}\left|\nabla y_1\right|^2 h.\nu d\Gamma dt-\\
&\frac{(a_1-a_2)^2}{a_2}\int_0^T\int_{\Gamma_0}\left|\frac{\partial y_1}{\partial\nu}\right|^2 h.\nu d\Gamma dt+\int_0^T\int_{\Gamma_2}\left|y_2'\right|^2 h.\nu d\Gamma dt+\\
&2a_2\int_0^T\int_{\Gamma_2}\left|\frac{\partial y_2}{\partial\nu}\right|^2 h.\nabla y_2 d\Gamma dt-a_2\int_0^T\int_{\Gamma_2}\left|\nabla y_2\right|^2 h.\nu d\Gamma dt+\\
&a_2\int_0^T\int_{\Gamma_2}\left|\frac{\partial y_2}{\partial\nu}\right|^2(divh-\alpha)d\Gamma dt \qquad (47)
\end{aligned}
$$

after using the boundary conditions (3) and (5). Identity (47) is used together with $(H.1),(H.2),(H.3)$ and (8) to obtain estimate (38).

Nonlinear Stabilizers in Optimal Control Problems with Infinite Time Horizon

Alexander Tarasyev and Anastasia Usova*

Institute of Mathematics and Mechanics
of the Ural Branch of the Russian Academy of Sciences,
S.Kovalevskaja Str. 16, 620990, Ekaterinburg, Russia
tam@imm.uran.ru, tarasiev@iiasa.ac.at,
anastasy.ousova@gmail.com
http://www.imm.uran.ru/engl.asp

Abstract. In optimal control problems with infinite time horizon, arising in models of economic growth, there are essential difficulties in analytical and even in numerical construction of solutions of Hamiltonian systems. The problem is in stiff properties of differential equations of the maximum principle and in non-stable character of equilibrium points connected with corresponding transversality conditions. However, if a steady state exists and meets several conditions of regularity then it is possible to construct a nonlinear stabilizer for the Hamiltonian system. This stabilizer inherits properties of the maximum principle, generates a nonlinear system with excluded adjoint variables and leads its trajectories to the steady state. Basing on the qualitative theory of differential equations, it is possible to prove that trajectories generated by the nonlinear stabilizer are close to solutions of the original Hamiltonian system, at least locally, in a neighborhood of the steady state. This analysis allows to create stable algorithms for construction of optimal solutions.

Keywords: optimal control, nonlinear control system, nonlinear stabilizer, economic systems.

Introduction

This paper deals with optimal control problems with infinite time horizon basing on economical growth models which is relied on classical constructions of growth theory (see [10], [11]). Also it includes ideas of a SEDIM model [9] describing

* The research is supported by the Russian Fund of Basic Research (Grants 11-01-0042-a, 11-01-12088-ofi-m-2011, 11-01-12112-ofi-m-2011), by the Program for the Sponsorship of Leading Scientific Schools (Grant NSCH-64508.2010.1), by the Program of the Presidium of RAS "Dynamic Systems and Control Theory", by the Program of the Presidium of RAS No. 38Π (Project 12-Π-1-1038), by the Project of the Ural Branch of RAS "Socio-Economic Development of Regions: Forecasting and Optimal Control" (Grant 12-Π-7-1001), and the International Institute for Applied System Analysis (IIASA).

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 286–295, 2013.

the role of different economic factors such as the demographic ones in a country's economic development. Another technique in the background (see [2], [5]) considers capital and useful work as the key drivers of economic growth and uses optimal control theory to design past and future growth trajectories.

The research of optimal control problems uses as basis the Pontryagin's maximum principle [8] for the problem with infinite time horizon (see [1], [3], [5]). We investigate properties of the maximized Hamiltonian function and provide analysis of existence of steady states in domains of specific control regimes and focus attention on the domain corresponding to the transient control regimes of investment. We consider linearized Hamiltonian system in this domain. Special attention is given to the Jacobi matrix which has two negative and two positive eigenvalues that is the steady state has the saddle character. According to the results of the qualitative theory of differential equations [4] the trajectory of the nonlinear Hamiltonian dynamics converges to the steady state tangentially to the plane generated by eigenvectors corresponding to negative eigenvalues of the Jacobi matrix. This analysis provides the important information about the growth rates of optimal synthetic trajectories.

A novelty of the proposed solution is based on the idea of creating of nonlinear stabilizers built on the feedback principle (see [6], [7]) which lead the system from any current position to a steady state. The constructed nonlinear stabilizer generates the dynamic system closed in phase variables and having the property of local stability. Also we construct solutions of the Hamiltonian system and the stabilized Hamiltonian system in a steady state neighborhood and compare behavior of these trajectories. Simulated optimal trajectories of nonlinear Hamiltonian systems are obtained numerically by the implicit Runge–Kutta method.

1 Two-Sectors Economical Growth Model and Optimal Control Problem

The Model. The model is based on analysis of the Gross Domestic Product (GDP) dynamics which is denoted by symbol Y. It is supposed that changes of GDP depend on three production factors: capital stock K, labor L (or it can be named as human capital) and useful work U. The production function F describes the relation between these factors and GDP (Y), that is $Y = F[K, L, U]$. It is assumed that the production function F has the property of homogeneity of degree one, i.e.

$$F[\alpha K, \alpha L, \alpha U] = \alpha F[K, L, U] \quad \forall \alpha > 0.$$

This model includes also a parameter $P(t)$ denoting the number of workers in a country at time t. According to the Sanderson model [9] we assume that labor L is proportional to the number of workers P with coefficient E. This coefficient has the sense of labor efficiency of one worker. Hence, we have the following equality: $L(t) = E(t)P(t)$. Due to the homogeneous property of the production function we introduce relative variables: $k = K/P$, $l = E = L/P$, $u = U/P$, $y = Y/P$.

It is supposed that the total number of workers P has exponential growth trend

$$\dot{P}(t) = \rho P(t), \quad \rho > 0. \tag{1}$$

Here ρ is a positive predefined constant denoting the relative growth rate. It should be mentioned that the considered dynamics for the labor force is quite adequate for the US statistical data in the period from 1900 to 2005. Parameter ρ is small enough and equal to (approx.) 10^{-2}.

The dynamics of the capital stock $K(t)$ is determined by the Solow model in which changes of capital depend on investment level $S(t)$ with the depreciate rate δ, i.e.

$$\dot{K}(t) = S(t) - \delta K(t). \tag{2}$$

Investments in capital constitute a part of GDP (Y). Hence, it can be written as follows: $S(t) = s(t)Y(t)$, where function $s(t)$ may take any value in the range from zero to the positive constant a_s which is less than one, i.e. $0 \leqslant s(t) \leqslant a_s < 1$.

Changes in labor are described by the equation:

$$\dot{L}(t) = bR(t), \tag{3}$$

where function $R(t)$ denotes investments in growth of the labor efficiency. Investments $R(t)$ is also a share of GDP (Y), i.e. $R(t) = r(t)Y(t)$. It is assumed that function $r(t)$ takes any values from zero to the predefined constant a_r which is less than one. The positive parameter b stands for the marginal effectiveness of investment in human capital. It is supposed that *the relative useful work (per one worker)* $u(t) = U(t)/P(t)$ *is constant with an average value* $\widetilde{u}, \quad \forall t \geq t_0$. Due to this assumption the production function $F[k, l, u]$ can be rewritten as follows: $F[k, l, u] = F[k, l, \widetilde{u}] = f(k, l)$.

Based on equations (1), (2) and (3), one can evaluate dynamics of relative variables k and l.

Let functions $C(t)$ and $c(t)$ describe the total consumption level in a country and the consumption level per one worker, respectively. It is assumed that *the closed economical system is considered in which GDP* (Y) *is spent on consumption* (C) *and investments in capital stock* (S) *and human capital* (R): $Y(t) = C(t) + S(t) + R(t)$, *or in relative variables:* $y(t) = c(t) + (s(t) + r(t))y(t)$. Hence one can easily calculate consumption per one worker

$$c(t) = (1 - s(t) - r(t))y(t) \approx (1 - s(t))(1 - r(t))y(t). \tag{4}$$

Optimal Control Problem. Let us consider investments s and r as control variables. It is supposed that the utility function of the growth process is described by an integral consumption index discounted on the infinite horizon. We use the consumption index of the logarithmic type, rather common for the theory of endogenous growth (see [12]). Let us note that the utility of such type is closely related to the notion of entropy in thermodynamics, mechanics and dynamic systems $J = \int\limits_{t_0}^{+\infty} e^{-\lambda t} \ln c(t)\, dt$. Here, parameter λ is the discounting factor.

It should be mentioned that the following equality $d\ln c(t) = \dfrac{dc(t)}{c(t)}$ determines relative growth of the consumption $c(t)$ (4) per one worker. In fact, the introduced utility function presents the summary growth of the relative consumption adjusted to the value of money depreciation.

Problem 1. The optimal control problem presumes maximization of the utility function

$$J = \int\limits_{t_0}^{+\infty} e^{-\lambda t}\left(\ln\left(1-s(t)\right) + \ln\left(1-r(t)\right) + \ln f(k(t),l(t))\right)\,dt$$

over trajectory $(k(\cdot), l(\cdot), s(\cdot), r(\cdot))$ of the system

$$\begin{cases} \dot{k}(t) = s(t)f(k(t),l(t)) - (\delta+\rho)k(t) \\ \dot{l}(t) = br(t)f(k(t),l(t)) - \rho l(t) \end{cases}$$

with control parameters $(s(\cdot), r(\cdot))$ subject to constraints

$$0 \le s(t) \le a_s < 1, \quad 0 \le r(t) \le a_r < 1, \quad 0 \le a_s + a_r < 1, \tag{5}$$

and phase variables $(k(\cdot), l(\cdot))$ satisfying initial conditions $k(t_0) = k^0$, $l(t_0) = l^0$.

The production function $y = f(k,l)$ meets the following conditions

PF_1. For all positive values of phase variables k and l function $f(k,l)$ is positive with its partial derivatives, i.e. $f(k,l) > 0, \quad f_k > 0, \quad f_l > 0$.

PF_2. For all positive values of phase variables k and l function $f(k,l)$ is a strictly concave function in phase variables, i.e. $f_{kk} < 0, \quad f_{kk}f_{ll} - f_{kl}^2 > 0$.

Here we use the following notations for the first and second order derivatives of the production function $f = f(k,l)$

$$f_k = \frac{\partial f(k,l)}{\partial k},\ f_l = \frac{\partial f(k,l)}{\partial l},\ f_{kl} = \frac{\partial^2 f(k,l)}{\partial k \partial l},\ f_{kk} = \frac{\partial^2 f(k,l)}{\partial k^2},\ f_{ll} = \frac{\partial^2 f(k,l)}{\partial l^2}.$$

Let us note that the problem 1 can be solved within the optimal control theory for problems with infinite horizon (see [1], [5]).

2 Model Analysis

Model analysis is based on the Pontryagin maximum principle [8] for problems with infinite time horizon [1].

Hamiltonian Function. We investigate properties of the Hamiltonian function $\widetilde{H} = \widetilde{H}(t;k,l;s,r;\widetilde{\psi}_1,\widetilde{\psi}_2)$ which is defined by the equality:

$$\begin{aligned}\widetilde{H}(t;k,l;s,r;\widetilde{\psi}_1,\widetilde{\psi}_2) = e^{-\lambda t}(\ln(1-s) + \ln(1-r) + \ln f(k,l)) \ + \\ + \ \widetilde{\psi}_1(sf(k,l) - (\delta+\rho)k) + \widetilde{\psi}_2(brf(k,l) - \rho l).\end{aligned} \tag{6}$$

Let us formulate the main property of the Hamiltonian function (6).

Proposition 1. *The Hamiltonian function* $\widetilde{H}(t;k,l;s,r;\widetilde{\psi}_1,\widetilde{\psi}_2)$ *is concave in control variables* s *and* r.

It is convenient to introduce new variables for excluding the exponential time term: $\psi_1 = \widetilde{\psi}_1 e^{\lambda t}$, $\psi_2 = \widetilde{\psi}_2 e^{\lambda t}$ and $\widehat{H} = \widetilde{H} e^{\lambda t}$. Substituting new variables to the Hamiltonian function (6) we get the expression:

$$\widehat{H}(k,l;s,r;\psi_1,\psi_2) = \ln(1-s) + \ln(1-r) + \ln f(k,l) + \\ + \psi_1(sf(k,l) - (\delta+\rho)k) + \psi_2(brf(k,l) - \rho l). \qquad (7)$$

Since control variables s and r satisfy to restrictions (5), the optimal control has the following structure:

$$s^0 = \begin{cases} 0, & (k,l,\psi_1) \in \Delta_s^1 = \{(k,l,\psi_1) : \psi_1 f(k,l) \leqslant 1\}; \\ 1 - \dfrac{1}{\psi_1 f(k,l)}, & (k,l,\psi_1) \in \Delta_s^2 = \left\{(k,l,\psi_1) : 1 \leqslant \psi_1 f(k,l) \leqslant \dfrac{1}{1-a_s}\right\}; \\ a_s, & (k,l,\psi_1) \in \Delta_s^3 = \left\{(k,l,\psi_1) : \psi_1 f(k,l) \geqslant \dfrac{1}{1-a_s}\right\}; \end{cases}$$

$$r^0 = \begin{cases} 0, & (k,l,\psi_2) \in \Delta_r^1 = \{(k,l,\psi_2) : b\psi_2 f(k,l) \leqslant 1\}; \\ 1 - \dfrac{1}{b\psi_2 f(k,l)}, & (k,l,\psi_2) \in \Delta_r^2 = \left\{(k,l,\psi_2) : 1 \leqslant b\psi_2 f(k,l) \leqslant \dfrac{1}{1-a_r}\right\}; \\ a_r, & (k,l,\psi_2) \in \Delta_r^3 = \left\{(k,l,\psi_2) : b\psi_2 f(k,l) \geqslant \dfrac{1}{1-a_r}\right\}. \end{cases} \qquad (8)$$

Substituting values of optimal control to the Hamiltonian function $\widehat{H}(\cdot)$ in (7) we obtain the maximized Hamiltonian: $H(k,l;\psi_1,\psi_2) = \widehat{H}(k,l;s^0,r^0;\psi_1,\psi_2)$. There exist nine domains $D_{ij} = \Delta_s^i \cap \Delta_r^j \quad (i,j = 1,2,3)$ of definition of the maximized Hamiltonian function. These domains are determined by the structure of optimal controls. Let us discuss important properties of the maximized Hamiltonian function.

Proposition 2. *The maximized Hamiltonian function* $H(k,l;\psi_1,\psi_2)$ *is a smooth function in variables* k,l *and* ψ_1,ψ_2 *in domains* D_{ij} $(i,j = \overline{1,3})$ *and on boundaries between these domains.*

Proposition 3. *The maximized Hamiltonian is a strictly concave function in phase variables* k,l *for all positive values of conjugate variables* ψ_1 *and* ψ_2, *if the following matrix is negatively defined:*

$$\partial f(k,l) = \begin{pmatrix} -f & f_k & f_l \\ f_k & f_{kk} & f_{kl} \\ f_l & f_{lk} & f_{ll} \end{pmatrix}, \quad \forall \quad (k,l,\psi_1,\psi_2) \in D_{22},\ \psi_1 > 0, \psi_2 > 0.$$

Necessary and Sufficient Conditions of Optimality. Let us mention that for the control problem 1 all conditions of the existence theorem (see [1], [3]) are fulfilled. Moreover, one can formulate necessary [1] and sufficient [5] conditions of optimality for problems with infinite horizon in the form of the Pontryagin maximum principle. It should be noted that properties 2 and 3 ensure sufficiency of necessary optimality conditions [5].

Qualitative Analysis. Firstly, we construct the Hamiltonian system and investigate the existence of steady states. Due to the structure of the optimal control $(s^0(t), r^0(t))$ in (8) the Hamiltonian system has different form in each domain $D_{ij} \quad (i, j = \overline{1,3})$. The special attention is given to domain D_{22} with the transient control regime, where both controls are not constant.

In the domain $D_{22} = \Delta_s^2 \cap \Delta_r^2$ the Hamiltonian system has the following form:

$$\begin{cases} \dot{k} &= f(k,l) - (\delta+\rho)k - \dfrac{k}{z_1} = H_1, \\ \dot{l} &= bf(k,l) - \rho l - \dfrac{l}{z_2} = H_2, \\ \dot{z}_1 &= \left(\lambda - f_k(k,l) + \dfrac{f(k,l)}{k}\right) z_1 - b\dfrac{k}{l} f_k(k,l) z_2 + \dfrac{k}{f(k,l)} f_k(k,l) - 1 = H_3, \\ \dot{z}_2 &= -\dfrac{l}{k} f_l(k,l) z_1 + \left(\lambda - bf_l(k,l) + b\dfrac{f(k,l)}{l}\right) z_2 + \dfrac{l}{f(k,l)} f_l(k,l) - 1 = H_4, \end{cases} \tag{9}$$

where new adjoint variables z_1 and z_2 are defined as follows: $z_1 = k\psi_1$ and $z_2 = l\psi_2$ and symbols H_i denotes functions $H_i = H_i(k, l, z_1, z_2)$, $i = \overline{1,4}$.

Let us suppose that the Hamiltonian system has a steady state P^* with coordinates $P^* = (k^*, l^*, z_1^*, z_2^*)$. In this case conjugate coordinates z_1^* and z_2^* of the steady state can be found from the first two equations of the Hamiltonian system, namely

$$z_1^* = \frac{k^*}{f(k^*, l^*) - (\delta+\rho)k^*}, \quad z_2^* = \frac{l^*}{bf(k^*, l^*) - \rho l^*}. \tag{10}$$

Further, we construct the linearized Hamiltonian system in a neighborhood of the steady state. Let symbol $A = \{\alpha_{ij}\}_{i,j=1}^4$ denotes the matrix of the linearized Hamiltonian system, where

$$\alpha_{i1} = \frac{\partial H_i(P^*)}{\partial k}, \; \alpha_{i2} = \frac{\partial H_i(P^*)}{\partial l}, \; \alpha_{i3} = \frac{\partial H_i(P^*)}{\partial z_1}, \; \alpha_{i4} = \frac{\partial H_i(P^*)}{\partial z_2}, \; i = \overline{1,4}.$$

3 Nonlinear Stabilizer

A nonlinear stabilizer is constructed under the following assumptions

A_1. *It is assumed that matrix A has two real negative λ_1 and λ_2 and two real positive λ_3 and λ_4 eigenvalues.*

This assumption means that the steady state $P^* = (k^*, l^*, z_1^*, z_2^*)$ has the saddle character.

Let the symbols $h_i = \{h_{ij}\}_{j=1}^4, \quad i = \overline{1,4}$ denote eigenvectors corresponding to eigenvalues $\lambda_i, i = \overline{1,4}$, respectively.

A_2. *It is supposed that first two coordinates of eigenvectors h_1 and h_2 corresponding to negative eigenvalues λ_1 and λ_2 meet the restriction $h_{11}h_{22} \neq h_{12}h_{21}$.*

Construction of Nonlinear Stabilizer. Idea of construction of the nonlinear stabilizer is based on results of the qualitative theory of differential equations

(see [4]). Namely, the trajectory of the nonlinear Hamiltonian dynamics converges to the steady state tangentially to the plane generated by eigenvectors corresponding to negative eigenvalues of the Jacobi matrix. Let us describe the algorithm of construction of the nonlinear stabilizer.

1. To build the plane π generated by two eigenvectors h_1 and h_2 corresponding to two negative eigenvalues λ_1 and λ_2, so that the steady state P^* belongs to this plane π.
2. To extract conjugate variables z_1 and z_2 from equations of the plane.
3. To substitute the obtained relations of extraction instead of conjugate variables into control functions $s^0(t)$ and $r^0(t)$ corresponding to domain D_{22}.

As a result, the algorithm provides construction of control $\widehat{s}(t)$ and $\widehat{r}(t)$ which is called nonlinear stabilizer. Let us consider each step in details.

Plane Construction. Any vector v located in the plane π can be expressed through eigenvectors h_1 and h_2 in the following way: $v = \nu_1 h_1 + \nu_2 h_2$. Hence, the plane π generated by two eigenvectors h_1 and h_2 and containing the equilibrium point P^* can be written as follows:

$$\begin{aligned} k - k^* &= \nu_1 h_{11} + \nu_2 h_{21}, \quad l - l^* = \nu_1 h_{12} + \nu_2 h_{22}, \\ z_1 - z_1^* &= \nu_1 h_{13} + \nu_2 h_{23}, \; z_2 - z_2^* = \nu_1 h_{14} + \nu_2 h_{24}. \end{aligned} \tag{11}$$

Due to assumption A_2 coefficients ν_1 and ν_2 can be found from the first two equations (11).

Extraction of Conjugate Variables. Conjugate variables z_1 and z_2 can be extracted from the equations (11) of the plane π. As a result, we obtain

$$\begin{aligned} z_1 &= z_1(k,l) = z_1^* + \gamma_{11}(k - k^*) + \gamma_{12}(l - l^*), \\ z_2 &= z_2(k,l) = z_2^* + \gamma_{21}(k - k^*) + \gamma_{22}(l - l^*), \end{aligned} \tag{12}$$

where $\gamma_{11} = -\dfrac{\begin{vmatrix} h_{12} & h_{13} \\ h_{22} & h_{23} \end{vmatrix}}{\begin{vmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{vmatrix}}$, $\gamma_{12} = \dfrac{\begin{vmatrix} h_{11} & h_{13} \\ h_{21} & h_{23} \end{vmatrix}}{\begin{vmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{vmatrix}}$, $\gamma_{21} = -\dfrac{\begin{vmatrix} h_{12} & h_{14} \\ h_{22} & h_{24} \end{vmatrix}}{\begin{vmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{vmatrix}}$, $\gamma_{22} = \dfrac{\begin{vmatrix} h_{11} & h_{14} \\ h_{21} & h_{24} \end{vmatrix}}{\begin{vmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{vmatrix}}$.

It should be mentioned that the following equalities take place

$$z_1(k^*, l^*) = z_1^*, \quad z_2(k^*, l^*) = z_2^*. \tag{13}$$

Nonlinear Stabilizer. The only thing left is to substitute expressions (12) into relations (8) for optimal controls in the domain D_{22}. Finally, we get the following structure of the nonlinear stabilizer:

$$\widehat{s}(k,l) = 1 - \frac{k}{z_1(k,l) f(k,l)}, \quad \widehat{r}(k,l) = 1 - \frac{l}{b z_2(k,l) f(k,l)}. \tag{14}$$

Substituting expressions for conjugate variables (12) to the first two equations of the Hamiltonian system (9) we get the stabilized Hamiltonian system:

$$\dot{k} = f(k,l) - (\delta + \rho)k - \frac{k}{z_1(k,l)}, \quad \dot{l} = b f(k,l) - \rho l - \frac{l}{z_2(k,l)}. \tag{15}$$

Properties of the Nonlinear Stabilizer. Let us indicate main properties of the constructed nonlinear stabilizer.

Proposition 4. *The nonlinear stabilizer* (14) *generates the nonlinear system* (15) *having the steady state with coordinates* (k^*, l^*) *which are the same as the first two coordinates at the steady state of the original Hamiltonian system* (9).

Proof of this propositions is based on the property (13) of the representation of adjoint variables $z_1 = z_1(k,l), z_2 = z_(k,l)$ in the plane π and relations (10) for conjugate coordinates z_1^*, z_2^* of the steady state.

Let us consider the linearized Hamiltonian system with Jacobi matrix A projected on subspace π. We substitute representation $z_1 = z_1(k,l)$ and $z_2 = z_2(k,l)$ (12) of conjugate variables into the first two equations of the linearized dynamics and collect similar terms

$$\dot{k} = \overline{a}_{11}(k-k^*) + \overline{a}_{12}(l-l^*), \quad \dot{l} = \overline{a}_{21}(k-k^*) + \overline{a}_{22}(l-l^*), \tag{16}$$

where $\overline{a}_{ij} = \alpha_{ij} + \alpha_{13}\gamma_{1j} + \alpha_{14}\gamma_{2j}$, $i,j = 1,2$.

Proposition 5. *The matrix of the linearized stabilized system is the same as the matrix* $\overline{A} = \{\overline{a}_{ij}\}_{i,j=1}^2$ *of the linearized Hamiltonian system projected on plane* π.

In order to prove this proposition it is necessary to linearized stabilized Hamiltonian system (15) at the steady state (k^*, l^*) neighborhood.

Let the symbol $\overline{A}$ denote the matrix of the linearized stabilized system (16). The next important question deals with eigenvalues of the stabilized Hamiltonian system (15).

Proposition 6. *The linearized stabilized Hamiltonian system* (15) *has two real negative eigenvalues coinciding with eigenvalues* λ_1 *and* λ_2 *and the following eigenvectors*

$$\overline{h}_1 = (h_{11}, h_{12}), \quad \overline{h}_2 = (h_{21}, h_{22}). \tag{17}$$

Proof. Basing on property 5 one can assert that the linearized stabilized Hamiltonian system coincides with the linearized Hamiltonian system (16) projected on plane π. For the Jacobi matrix A evaluated at the steady state P^* the following equalities are fulfilled $A\ h_i = \lambda_i\ h_i$, $i = \overline{1,4}$. Moreover eigenvectors h_1, h_2 are located at the plane π. Thus, for coordinates of these vectors are valid relations $h_{i3} = \gamma_{11}h_{i1} + \gamma_{12}h_{i2}, \quad h_{i4} = \gamma_{21}h_{i1} + \gamma_{22}h_{i2}, \quad i = 1,2$. Using these facts let us check the following equalities $\overline{A}\,\overline{h}_i = \lambda_i\overline{h}_i$, $i = 1,2$.

$$\overline{A}\,\overline{h}_i = \begin{pmatrix} \overline{a}_{11} & \overline{a}_{12} \\ \overline{a}_{21} & \overline{a}_{22} \end{pmatrix} \begin{pmatrix} h_{i1} \\ h_{i2} \end{pmatrix} =$$

$$= \begin{pmatrix} \alpha_{11}h_{i1} + \alpha_{12}h_{i2} + \alpha_{13}(\gamma_{11}h_{i1} + \gamma_{12}h_{i2}) + \alpha_{14}(\gamma_{21}h_{i1} + \gamma_{22}h_{i2}) \\ \alpha_{21}h_{i1} + \alpha_{22}h_{i2} + \alpha_{23}(\gamma_{11}h_{i1} + \gamma_{12}h_{i2}) + \alpha_{24}(\gamma_{21}h_{i1} + \gamma_{22}h_{i2}) \end{pmatrix} =$$

$$= \begin{pmatrix} \alpha_{11}h_{i1} + \alpha_{12}h_{i2} + \alpha_{13}h_{i3} + \alpha_{14}h_{i4} \\ \alpha_{21}h_{i1} + \alpha_{22}h_{i2} + \alpha_{23}h_{i3} + \alpha_{24}h_{i4} \end{pmatrix} = \begin{pmatrix} \lambda_i h_{i1} \\ \lambda_i h_{i2} \end{pmatrix} = \lambda_i\overline{h}_i, \quad i = 1,2.$$

□

The following theorem collects all obtained results.

Theorem 1. *Under assumptions A_1 and A_2 for the Hamiltonian system (9) constructed in domain D_{22} and linearized in a neighborhood of the steady state P^* the nonlinear stabilizer (14) exists and generates the nonlinear dynamical system (15) which is closed with respect to the phase variables k, l and has the following properties*
1. the steady state of the closed system (15) has coordinates (k^*, l^*) coinciding with the phase coordinates k and l of the steady state P^* of the original Hamiltonian system (9);
2. the system (15) is stabilized at the steady state P^*;
3. the eigenvectors $\overline{h}_1$ and $\overline{h}_2$ of the linearized closed system (16) generated by the nonlinear stabilizer are evaluated by formulas (17).

The proof of the theorem follows directly from properties of the nonlinear stabilizer.

Remark 1. The constructed nonlinear stabilizer generates the nonlinear system which is closed with respect to phase variables. The solution of the obtained stabilized system approximates optimal trajectories of the original Hamiltonian system in a neighborhood of the steady state, since a trajectory of the nonlinear Hamiltonian system that tends to the equilibrium point is tangent to the plane formed by two eigenvectors corresponding to negative eigenvalues. One can use this fact to estimate the growth rates of optimal trajectories The growth rates are determined by values of negative eigenvalues.

Numerical Simulations. The calculations are carried out on the basis of the data on the US economy in the period of 1900 to 2005. The data values are normalized with respect to the data values of 1900. The production function of the Cobb–Douglas type is used: $f(k, l) = \mu\, k^{\alpha}\, l^{\beta}$. The calibration procedure for the model parameters provides the following values: $\mu = 2.19942$, $\alpha = 0.31$, $\beta = 0.09$, $\lambda = 0.03$, $\delta = 0.2$, $\rho = 0.013$, $b = 0.31$, $a_s = 0.3$, $a_r = 0.2$, $k^0 = 1$, $l^0 = 1$. The Hamiltonian system has the equilibrium point P^* with coordinates $k^* = 5.75$, $l^* = 5.2$, $z_1^* = 1.8188$, and $z_2^* = 2.9684$. The control parameters at the equilibrium point take the values $s^* = 27.95$ and $r^* = 3.79$. All four eigenvalues of the matrix A calculated at the equilibrium point $P^* = (k^*, l^*, z_1^*, z_2^*)$ are real numbers; two of them are positive, and the other two are negative: $\lambda_1 = -0.268$, $\lambda_2 = -0.094$, $\lambda_3 = 0.124$, $\lambda_4 = 0.298$.

Trajectories of the system (15) generated by the nonlinear stabilizer (14) and the original Hamilton system (9) are calculated numerically by the Runge–Kutta method. Figure 1.(a) demonstrates phase trajectories $k(l)$ as a solutions of the stabilized (15) and Hamiltonian (9) dynamics. One can see that these trajectories almost coincide with each other especially at the vicinity of the steady state. Optimal trajectories of the capital stock $k(t)$ and labor efficiency $l(t)$ and its stabilized solutions are depicted at figures 1.(b) and 1.(c) respectively. In the steady state neighborhood optimal trajectories are very close to its stabilized solutions.

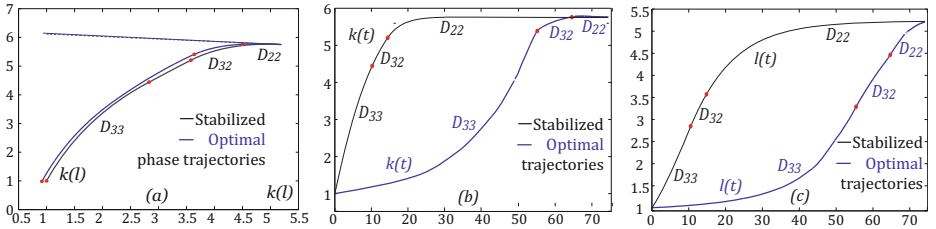


Fig. 1. Stabilized and optimal graphs (a) of phase trajectories, $k(l)$; (b) of the capital stock, $k(t)$; (c) of the labor efficiency, $l(t)$

References

1. Aseev, S.M., Kryazhimskiy, A.V.: The Pontryagin Maximum Principle and Optimal Economic Growth Problems. In: Proceedings of the Steklov Institute of Mathematics, vol. 257. Pleiades Publishing (2007)
2. Ayres, R.U., Warr, B.: The Economic Growth Engine: How Energy and Work Drive Material Prosperity. Edward Elgar Publishing, Cheltenham UK (2009)
3. Balder, E.J.: An existance result for optimal economic growth problems. J. Math. Anal. Appl. 95, 195–213 (1983)
4. Hartman, P.: Ordinary Differential Equations. J. Wiley and Sons, N.Y. (1964)
5. Krasovskii, A.A., Tarasyev, A.M.: Conjugation of Hamiltonian Systems in Optimal Control Problems. Preprints of the 17th World Congress of the International Federation of Automatic Control, IFAC, Seoul, Korea, pp. 7784–7789 (2008)
6. Krasovskii, A.N., Krasovskii, N.N.: Control under Lack of Information. Birkhauser, Boston (1995)
7. Krasovskii, N.N., Subbotin, A.I.: Game-Theoretical Control Problems. Springer, Berlin (1988)
8. Pontryagin, L.S., Boltyanskii, V.G., Gamkrelidze, R.V., Mishchenko, E.F.: The Mathematical Theory of Optimal Processes. Interscience, New York (1962)
9. Sanderson, W.: The SEDIM Model: Version 0.1. IIASA Interim Report IR-04-041, 42 pages (2004)
10. Shell, K.: Applications of Pontryagin's Maximum Principle to Economics. Mathematical Systems Theory and Economics 1, 241–292 (1969)
11. Solow, R.M.: Growth Theory: An Exposition. Oxford University Press, New York (1970)
12. Tarasyev, A.M., Watanabe, C.: Optimal Dynamics of Innovation in Models of Economic Growth. Journal of Optimization Theory and Applications 108, 175–203 (2001)

Combined Feedforward/Model Predictive Tracking Control Design for Nonlinear Diffusion-Convection-Reaction-Systems

Tilman Utz[1,⋆], Knut Graichen[1], and Andreas Kugi[2]

[1] Institute of Measurement, Control, and Microtechnology, Ulm University
Albert-Einstein-Allee 41, 89081 Ulm, Germany
[2] Automation and Control Institute, Vienna University of Technology
Gußhausstraße 27–29, 1040, Vienna, Austria
tilman.utz@uni-ulm.de

Abstract. The tracking control design for setpoint transitions of a quasi-linear diffusion-convection-reaction system with boundary control is considered. For this a suitable model-based feedforward control is determined that relies on the flatness-based parametrization of the control input. A receding horizon feedback control is added within a two-degrees-of-freedom control scheme to account for disturbances, model inaccuracies, and input constraints. The tracking performance of this control scheme is shown by means of simulation studies.

A large class of chemical reactors with an interaction of diffusive, convective, and reactive effects leads to infinite-dimensional mathematical models in the form of nonlinear boundary-controlled parabolic partial differential equations (PDEs) [6]. The control design for setpoint transitions of chemical reactors, e. g., for ignition, extinction, or grade-transitions constitutes a challenging problem. In this contribution, the well-known two-degrees-of-freedom (2DOF) control scheme is applied in order to tackle this control task. The basic idea consists in first designing a feedforward control to steer the system along prescribed trajectories. The trajectory planning and feedforward control are complemented with a state feedback tracking control stabilizing the system about the desired trajectories.

In the literature, there exists a variety of concepts for the design of both feedforward and feedback tracking controllers. For the feedforward control design, approaches using the flatness concept [2] have found widespread attention. The flatness property allows for a parametrization of the state and input in terms of a so-called flat output and its time derivatives and therefore provides a systematic approach for feedforward control design. Originally proposed for finite-dimensional systems, generalizations of the flatness concept have been successfully carried over to certain classes of PDEs, see, e. g., [7,10,12]. In these so-called late lumping approaches the parametrization is directly solved for the underlying PDE. In contrast, the early lumping approach to control design is based on a

⋆ Corresponding author.

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 296–305, 2013.

finite-dimensional approximation of the system. Using suitable finite difference schemes to approximate the spatial derivatives results in a semi-discretization which is differentially flat, and the equivalence of the respective feedforward controls has been shown under certain conditions for some types of PDEs, see, e. g., [14,19].

In practice, the feedforward part has to be amended by an additional feedback tracking control in order to compensate for model uncertainties or disturbances. For the design of such stabilizing feedback controllers for infinite-dimensional systems, receding horizon optimal control, e. g., [9,15], constitutes a promising tool. This method is particularly attractive since it provides, in contrast to most state-of-the-art tracking control design methods for DCRSs, see, e. g., [12,13], the possibility to systematically include constraints as they are frequently encountered in technical systems.

The paper is structured as follows: In Section 1, the diffusion-convection-reaction system (DCRS) and the control task is introduced. Section 2 is devoted to the control design based on a semi-discretization of the considered infinite-dimensional system. The paper provides simulation results in Section 3 and conclusions are drawn in Section 4.

1 Problem Formulation

The quasilinear, scalar DCRS, described by the (suitably scaled) parabolic PDE

$$p(\theta(z,t))\partial_t\theta(z,t) = \partial_z(q(\theta(z,t))\partial_z\theta(z,t)) - \nu\partial_z\theta(z,t) + r(\theta(z,t))\theta(z,t) \ , \quad (1)$$

$z \in (0,1)$, $t > 0$ is considered. The storage coefficient $p(\theta(z,t)) = p_0 + p_1\theta(z,t)$, the diffusion coefficient $q(\theta(z,t)) = q_0 + q_1\theta(z,t)$, and the reaction coefficient $r(\theta(z,t)) = r_0 + r_1\theta(z,t)$ depend on the state $\theta(z,t)$ in an affine way, whereas the convection parameter $\nu \geq 0$ is constant. The boundary conditions

$$\partial_z\theta(z,t)\big|_{z=0} = d(t) \ , \quad (2a)$$

$$q(\theta(1,t))\ \partial_z\theta(z,t)\big|_{z=1} = \tilde{q}(u(t) - \theta(1,t)) \ , \quad (2b)$$

$t > 0$, $\tilde{q} > 0$, and the initial condition

$$\theta(z,0) = \theta_{\text{init}}(z) \ , \quad (3)$$

$z \in [0,1]$, complete the infinite-dimensional system. In order for (1) to be parabolic it has to be assured that $q(\theta(z,t))$ is positive in the considered range of the state variable. The exogenous input variable $d(t)$ in (2a) represents an additional sink or source term that will be considered as an unknown disturbance and is assumed to be zero for the feedforward controller design. Additionally, the control input $u(t)$ entering via the boundary condition (2b) is supposed to be subject to so-called box constraints, i. e.,

$$u(t) \in [u^-, u^+] \ . \quad (4)$$

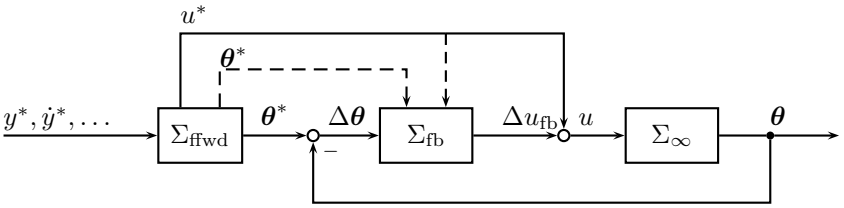


Fig. 1. 2DOF control scheme consisting of the infinite-dimensional system Σ_∞, a trajectory planning and feedforward control Σ_{ffwd}, which provides nominal state and control input trajectories $\boldsymbol{\theta}^*$ and u^* based on desired output trajectories y^*, and a tracking controller Σ_{fb}, which provides a correction to the nominal control input trajectories Δu_{fb} based on the state tracking error $\Delta\boldsymbol{\theta}$

Denoting the state evaluated at $z = 0$ as the output of the system, i. e. $y(t) = \theta(0, t)$, the control task consists in stably and robustly carrying out transitions between setpoint values and the associated steady-state profiles of (1)–(3), within the transition time $T = 1$. The 2DOF control scheme used to accomplish these transitions is schematically depicted in Figure 1.

2 Tracking Control Based on Finite Difference Semi-Discretization

The design of both the flatness-based feedforward control and the receding horizon tracking control depends on a semi-discretization of the infinite-dimensional system (1)–(3), see also [17]. The methodology to obtain the semi-discretized system pursued in this contribution is to discretize the spatial coordinate z using finite differences on an equidistant grid with N grid elements and the nodes $z_0 = 0, z_1 = \Delta z, \ldots, z_k = k\Delta z, \ldots, z_N = 1$ where $\Delta z = 1/N$. Applying the central finite difference schemes

$$\partial_z \theta_k = \frac{1}{2\Delta z}\left(\theta_{k+1} - \theta_{k-1}\right) + \mathcal{O}(\Delta z^2) \,, \tag{5a}$$

$$\partial_z^2 \theta_k = \frac{1}{\Delta z^2}\left(\theta_{k+1} - 2\theta_k + \theta_{k-1}\right) + \mathcal{O}(\Delta z^2) \,, \tag{5b}$$

$$(\partial_z \theta_k)^2 = \frac{1}{\Delta z^2}\big(\theta_{k+1} - \theta_k\big)\big(\theta_k - \theta_{k-1}\big) + \mathcal{O}(\Delta z^2) \tag{5c}$$

to the PDE (1) and the boundary conditions (2), leads to the following system of $N + 1$ ODEs for the discretized states[1] $\theta_k(t) = \theta(z_k, t)$:

$$(p_0 + p_1\theta_0)\,\dot{\theta}_0 = (q_0 + q_1\theta_0)\,\frac{2(\theta_1 - \theta_0 - \Delta z d)}{\Delta z^2} + q_1 d^2 - \nu d + (r_0 + r_1\theta_0)\theta_0 \,, \tag{6a}$$

[1] For the sake of readability, time-dependencies as in $\theta_k(t)$ are omitted whenever they are clear from the context.

$$(p_0 + p_1\theta_k)\,\dot{\theta}_k = (q_0 + q_1\theta_k)\,\frac{\theta_{k+1} - 2\theta_k + \theta_{k-1}}{\Delta z^2} + q_1\frac{(\theta_{k+1} - \theta_k)(\theta_k - \theta_{k-1})}{\Delta z^2}$$
$$- \nu\frac{\theta_{k+1} - \theta_{k-1}}{2\Delta z} + (r_0 + r_1\theta_k)\theta_k\ , \quad k = 1, \ldots, N-1\ , \tag{6b}$$

$$(p_0 + p_1\theta_N)\,\dot{\theta}_N = \frac{2(q_0 + q_1\theta_N)}{\Delta z^2}\left(\frac{\Delta z\tilde{q}}{q_0 + q_1\theta_N}(u - \theta_N) - \theta_N + \theta_{N-1}\right)$$
$$+ \frac{q_1}{\Delta z^2}\left(\frac{2\Delta z\tilde{q}}{q_0 + q_1\theta_N}(u - \theta_N) + \theta_{N-1} - \theta_N\right)(\theta_N - \theta_{N-1})$$
$$- \frac{\nu\tilde{q}}{q_0 + q_1\theta_N}(u - \theta_N) + (r_0 + r_1\theta_N)\theta_N\ . \tag{6c}$$

Note that the scheme (5c) for the squared first derivative with respect to z is only applied to nodes with $k > 0$, which is advantageous for the parametrization considered in the subsequent section. The initial conditions obtained by evaluating (3) at the nodes

$$\theta_k(0) = \theta_{\text{init}}(z_k)\ , \tag{7}$$

$k = 0, \ldots, N$ complete the finite-dimensional semi-discretized approximation of the infinite-dimensional system (1)–(3).

2.1 Flatness-Based State and Input Parametrization

The semi-discretization (6) has the property of being flat, which is shown in the following. The k-th equation of (6a) and (6b) is affine in $\theta_{k+1}(t)$, $k = 0, \ldots, N-1$ and (6c) is affine in the control input $u(t)$. Thus, solving these equations for $\theta_{k+1}(t)$ and $u(t)$, respectively, yields

$$\theta_1 =: \Psi_0(\theta_0, \dot{\theta}_0)\ , \tag{8a}$$
$$\theta_{k+1} =: \Psi_k(\theta_k, \dot{\theta}_k, \theta_{k-1})\ , \quad k = 1, \ldots, N-1, \tag{8b}$$
$$u =: \Psi_N(\theta_N, \dot{\theta}_N, \theta_{N-1})\ . \tag{8c}$$

Considering $y(t) = \theta_0(t)$ it follows from (8a) that $\theta_1(t)$ can be parametrized in terms of $y(t)$ and $\dot{y}(t)$. Differentiating (8a) with respect to time

$$\dot{\theta}_1 = \frac{\partial\Psi_0}{\partial\theta_0}\dot{\theta}_0 + \frac{\partial\Psi_0}{\partial\dot{\theta}_0}\ddot{\theta}_0 \tag{9}$$

and inserting this result into (8b) for $k = 1$ directly yields a parametrization of $\theta_2(t)$ by $y(t)$, $\dot{y}(t)$ and $\ddot{y}(t)$. Obviously, this procedure can be analogously continued for $k = 2, \ldots, N-1$ with

$$\dot{\theta}_{k+1} = \frac{\partial\Psi_k}{\partial\theta_k}\dot{\theta}_k + \frac{\partial\Psi_k}{\partial\dot{\theta}_k}\ddot{\theta}_k + \frac{\partial\Psi_k}{\partial\theta_{k-1}}\dot{\theta}_{k-1}\ , \tag{10}$$

such that every state $\theta_k(t)$, $k = 1, \ldots, N$ as well as the control input $u(t)$ are recursively parametrized by $y(t)$ and its first $N+1$ time derivatives. Hence $y(t)$ constitutes a flat output.

As a consequence, nominal state trajectories $\theta_k^*(t)$ and a control input $u^*(t)$ can be calculated by recursively evaluating (8) using a sufficiently smooth trajectory for the flat output $y^*(t)$, such that the output of (6)–(7) for an exact model and in the absence of disturbances exactly tracks $y^*(t)$. Under certain conditions on the growth of the time-derivatives of the prescribed flat output $y^*(t)$ (defined by its so-called Gevrey-class) and given a suitable set of system parameters p_0, p_1, q_0, q_1, ν, r_0, and r_1, it can be shown that the control input $u^*(t)$ converges for $N \to \infty$ to a suitable control input for the DCRS (1)–(3), see, e. g., [10,19].

2.2 Receding Horizon Tracking Control

In the case of model uncertainties or in order to account for disturbances or control constraints, the feedforward control Σ_{ffwd} in Figure 1 has to be extended by a feedback controller Σ_{fb}. The receding horizon control strategy thereby is formulated for the same semi-discretization that is used for the flatness-based parametrization.

In view of the 2DOF control structure in Figure 1, the feedback controller Σ_{fb} is designed to stabilize the system Σ_∞ along the reference trajectories $\theta_k^*(t)$, $k = 0, \dots, N$ provided by the flatness-based trajectory planning Σ_{ffwd}. This means that the tracking errors

$$\Delta\theta_k(t) = \theta(z_k, t) - \theta_k^*(t) \ ,$$

$k = 0, \dots, N$ have to be suppressed by the control action $\Delta u_{\text{fb}}(t)$, which corrects the feedforward control $u^*(t)$, see Figure 1. Using $\boldsymbol{\theta}(t) = [\theta_0(t), \dots, \theta_N(t)]^{\mathsf{T}} \in \mathbb{R}^{N+1}$ to summarize the differential equations (6) in the form

$$\dot{\boldsymbol{\theta}}(t) = \mathbf{f}(\boldsymbol{\theta}(t), u(t)) \ ,$$

the tracking error $\Delta\boldsymbol{\theta}(t) = [\Delta\theta_0(t), \dots, \Delta\theta_N(t)]^{\mathsf{T}} \in \mathbb{R}^{N+1}$ satisfies the error dynamics

$$\Delta\dot{\boldsymbol{\theta}}(t) = \mathbf{f}\Big(\boldsymbol{\theta}^*(t) + \Delta\boldsymbol{\theta}(t), u^*(t) + \Delta u_{\text{fb}}(t)\Big) - \dot{\boldsymbol{\theta}}^*(t) =: \mathbf{F}(\Delta\boldsymbol{\theta}(t), \Delta u_{\text{fb}}(t), t) \ . \quad (11)$$

The receding horizon controller design accounts for the nonlinear and time-varying error dynamics (11) by solving the following optimal control problem (OCP) in a discrete-time fashion for each instant of time $t_i = i\Delta t$ with the given sampling time Δt:

$$\min_{\Delta u(\cdot)} J(\Delta u(\cdot), \Delta\boldsymbol{\theta}^i) = ||\Delta\boldsymbol{\theta}(t_{i,\text{f}})||_P^2 + \int_{t_i}^{t_{i,\text{f}}} ||\Delta\boldsymbol{\theta}(t)||_Q^2 + R\Delta u(t)^2 \, \mathrm{d}t \quad (12a)$$

$$\text{s.t.} \quad \Delta\dot{\boldsymbol{\theta}}(t) = \mathbf{F}(\Delta\boldsymbol{\theta}(t), \Delta u(t), t) \ , \quad \Delta\boldsymbol{\theta}(t_i) = \Delta\boldsymbol{\theta}^i \quad (12b)$$

$$\Delta u(t) \in [\Delta u^-(t), \Delta u^+(t)] \ , \quad t \in [t_i, t_{i,\text{f}}] \cdot \quad (12c)$$

Starting from the tracking error $\Delta\boldsymbol{\theta}^i = \boldsymbol{\theta}(t_i) - \boldsymbol{\theta}^*(t_i)$ at time t_i, the error dynamics in (12b) are used to predict the error trajectory $\Delta\boldsymbol{\theta}(t)$ over a finite

prediction horizon $t \in [t_i, t_{i,\mathrm{f}}]$ with the final time $t_{i,\mathrm{f}} = t_i + t_\mathrm{f}$, where $t_\mathrm{f} \geq \Delta t$ denotes the (constant) horizon length. The cost functional (12a) to be minimized penalizes the tracking error $\Delta\boldsymbol{\theta}(t)$ as well as the feedback control action $\Delta u(t)$ with respect to the positive definite matrices[2] P, Q and the positive scalar R. The time-varying input constraints (12c) follow from the original constraints (4) of the feedforward trajectory $u^*(t)$ in the form

$$\Delta u^{\pm}(t) := u^{\pm} - u^*(t)\,, \quad t \in [t_i, t_{i,\mathrm{f}}]\,. \tag{13}$$

In the following, it is assumed that the OCP (12) possesses an optimal solution $\Delta\bar{u}(t; \Delta\boldsymbol{\theta}^i)$, $\Delta\bar{\boldsymbol{\theta}}(t; \Delta\boldsymbol{\theta}^i)$, $t \in [t_i, t_{i,\mathrm{f}}]$. Note that this assumption is not very restrictive due to the absence of terminal and state constraints.

The OCP (12) is solved in each time step t_i of the receding horizon scheme and only the first part of the control trajectory $\Delta\bar{u}(t; \Delta\boldsymbol{\theta}^i)$ is used as the feedback control in Figure 1, i. e.

$$\Delta u_\mathrm{fb}(t) = \Delta\bar{u}(t; \Delta\boldsymbol{\theta}^i)\,, \qquad t \in [t_i, t_i + \Delta t)\,, \qquad i \in \mathbb{N}_0^+\,. \tag{14}$$

In the next time step t_{i+1}, the OCP (12) is solved again with respect to the new tracking error $\Delta\boldsymbol{\theta}^{i+1}$ that is used as initial condition in (12b).

If the system exactly follows the reference trajectory at time t_i, i. e., if $\Delta\boldsymbol{\theta}^i = \mathbf{0}$, and in the absence of disturbances and model inaccuracies the optimal solution of the OCP (12) is

$$\Delta\bar{u}(t; \Delta\boldsymbol{\theta}^i) = 0\,, \qquad \Delta\bar{\boldsymbol{\theta}}(t; \Delta\boldsymbol{\theta}^i) = \mathbf{0}\,, \qquad t \in [t_i, t_{i,\mathrm{f}}] \tag{15}$$

with $J(\Delta\bar{u}(\cdot), \Delta\boldsymbol{\theta}^i) = 0$. Hence, if the system exactly tracks the nominal trajectories the control action of the feedback controller Σ_fb is zero, see Figure 1.

Important design parameters of the receding horizon control scheme are the choice of the weighting matrices $P \in \mathbb{R}^{(N+1)\times(N+1)}$ and $Q \in \mathbb{R}^{(N+1)\times(N+1)}$, of the scalar weight R, and the horizon length t_f. Receding horizon formulations in model predictive control often use terminal set or equality constraints to achieve stability. In the case of a free end point formulation as it is the case in (12), stability can be shown, e. g., if the terminal cost function $||\Delta\boldsymbol{\theta}(t_{i,\mathrm{f}})||_P^2$ represents a (local) control Lyapunov function [1,11,8] or if the horizon length t_f is sufficiently large [5]. For the error dynamics (12b), which is time-dependent due to the feedforward trajectories, the rigorous proof of stability [3] as well as the consistency of this finite-dimensional control with the original infinite-dimensional system is subject of current research. In this contribution, the stability and performance of the receding horizon tracking controller are demonstrated by means of simulation studies in the following section.

3 Simulation Example

With the flat output $y(t) = \theta_0(t)$, the control task under consideration consists in realizing the finite-time transition between two setpoints $y(0) = y_0$ and $y(1) =$

[2] Here $||\mathbf{x}||_S^2 = \mathbf{x}^\mathsf{T} S\mathbf{x}$ denotes the weighted Euclidean norm of the vector $\mathbf{x} \in \mathbb{R}^n$ with the matrix $S \in \mathbb{R}^{n\times n}$ being positive definite.

y_1, which correspond to the (infinite-dimensional) steady-state profiles $\theta(z,0)$ and $\theta(z,1)$. A suitable reference trajectory $y^*(t)$ for the trajectory planning has to comply with the desired steady-state output values at the beginning ($t=0$) and at the end ($t=1$) of the setpoint transition, i. e.

$$y^*(0) = y_0 \ , \qquad y^*(1) = y_1 \ , \qquad \frac{\mathrm{d}^l}{\mathrm{d}t^l} y(t)\bigg|_{t=0} = \frac{\mathrm{d}^l}{\mathrm{d}t^l} y(t)\bigg|_{t=1} = 0 \tag{16}$$

$l = 1, \ldots, N+1$. The temporal path of the transition can be provided either by a polynomial of suitable order or, especially with regard to the continuous limit of the parametrization, by any smooth function of an appropriate Gevrey-class, see, e. g., [10].

In all simulation results shown in this contribution, the control design relies on a semi-discretization with $N = 10$ grid elements. The following set of system parameters $p_0 = 1.2$, $p_1 = 0.05$, $q_0 = 1$, $q_1 = -0.05$, $\tilde{q} = 1$, $\nu = 0.1$, $r_0 = 1$, and $r_1 = 0.2$ is used and the desired initial and final output values are $y_0 = 1$ and $y_1 = 2$, respectively. In addition, the values $u^{\pm} = \pm 4$ are used as box constraints (4) and a non-zero disturbance

$$d(t) = \begin{cases} -0.4 & \text{for } t \in [0.4, 0.6] \\ 0 & \text{else} \end{cases} \tag{17}$$

is considered in the simulations. The receding horizon control design is based on the time discretization $t_i = i\Delta t$ with the sampling time $\Delta t = 0.005$ and the horizon length $t_\mathrm{f} = 0.3$. In each time step t_i, the OCP (12) is numerically solved with a tailored gradient projection method [4], such that the OCP may be solved in a computationally very efficient way, see also [16]. The weighting matrix Q is set to a diagonal matrix with the diagonal element values interpolated between 200 for the first error state corresponding to the output and the value 2 for the last error state. The terminal weighting matrix P is set to $0.1Q$, while R is chosen as 0.3. The overall feedback controller Σ_fb is implemented as a Cmex function in MATLAB, and for the simulations, the standard MATLAB-solver `ode15s` is used to solve the semi-discretized system (6)–(7) on a grid with N_sim nodes, where $N_\mathrm{sim} \gg N$.

In Figure 2, the setpoint transition in the nominal case, i. e., without disturbance or input constraints, as well as a transition with disturbance (17) and a transition with input constraints (4) is shown. The time behaviour of the feedforward control $u^*(t)$ as well as the evolution of the cost $J(\Delta\bar{u}(\cdot), \Delta\boldsymbol{\theta}^i)$ show that the contribution of the tracking controller $\Delta u_\mathrm{fb}(t)$ vanishes in the nominal case. For the non-zero disturbance (17), this is of course no longer the case. However, it can be seen that after restoring the nominal conditions, the tracking error is reduced very fast, the transition is completed as prescribed and the simulation remains stable. At the same time, the cost $J(\Delta\bar{u}(\cdot), \Delta\boldsymbol{\theta}^i)$ decreases monotonically to zero. In the case of input constraints (4), deviations from the desired behaviour are also inevitable, since they are not considered in the feedforward control design. It can also be observed that due to the prediction horizon the

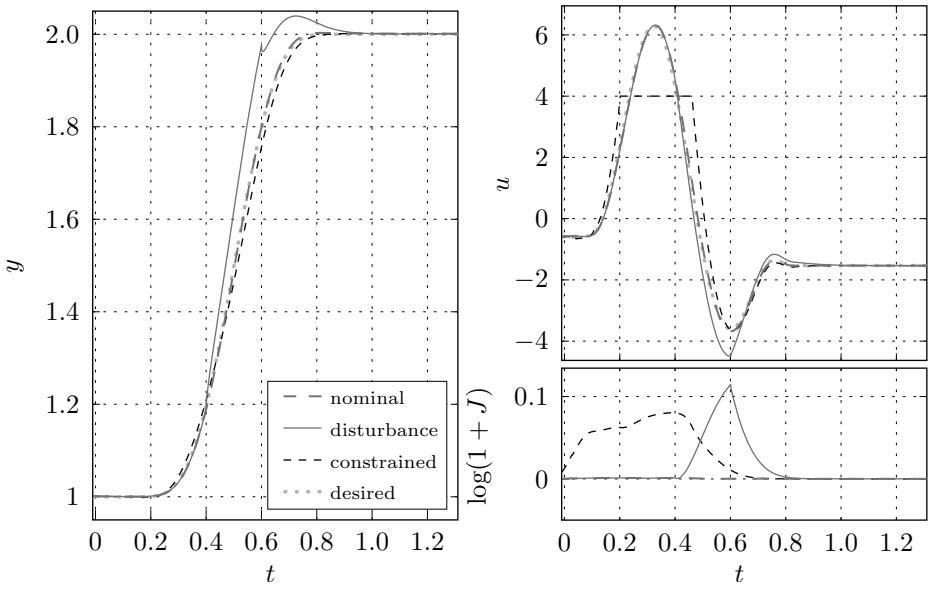


Fig. 2. Output $y(t)$, control input $u(t)$ and cost $J(\Delta\bar{u}(\cdot), \Delta\boldsymbol{\theta}^i)$ for the tracking control of the DCRS in the nominal case, in the case of non-zero disturbance, and under input constraints

feedback controller becomes aware of the impending violation of the constraints before it actually occurs. This results in a rise of the cost and subsequently in the pre-steering action visible in the control input time behaviour, see Figure 2.

The use of a 2DOF control scheme offers some benefits for the realization of setpoint transition tracking control, which are pointed out in the following. In Figure 3 the same setpoint transition as before is considered with non-zero disturbance (17) and the input constraints (4). However, the flatness-based trajectory generation and feedforward control are replaced by a pure feedforward of the steady-state reference values of the state variables $\theta_k(t)$, $k = 0, \ldots, N$ and of the input $u(t)$ at $t = 0.1$. Furthermore, a shorter prediction horizon $t_\mathrm{f} = 0.1$ is considered for the receding horizon controller. This could be motivated by the need to reduce the computational cost for the solution of the OCP (12). It can be observed on the one hand that with the prediction horizon $t_\mathrm{f} = 0.3$, the receding horizon control carries out the transition within a transition time comparable to the one observed for the 2DOF control scheme. However, this also results in significant control action especially at the beginning of the transition. On the other hand, the transition time is increased in the case of the short prediction horizon $t_\mathrm{f} = 0.1$ and the desired final output value is not reached within the simulation time. The largely different tracking behaviour seems comprehensible since the transition time is an important tuning parameter of the receding horizon control. This is in contrast to the simulation results obtained if the flatness-based trajectory generation and feedforward control are used in the

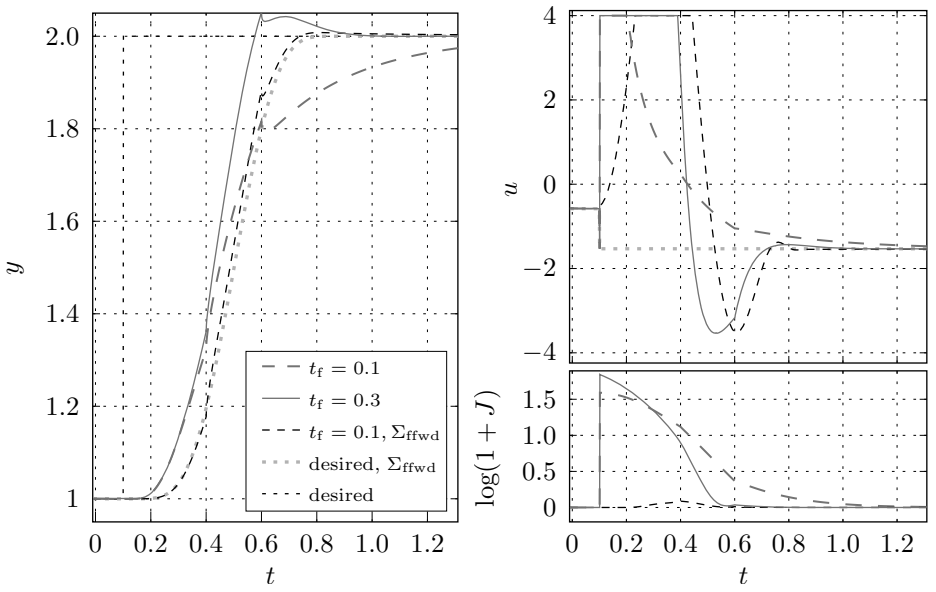


Fig. 3. Output $y(t)$, control input $u(t)$ and cost $J(\Delta\bar{u}(\cdot), \Delta\boldsymbol{\theta}^i)$ for the tracking control of the DCRS with non-zero disturbance and under input constraints both without flatness-based feedforward control and prediction horizon $t_f \in \{0.1, 0.3\}$, and with flatness-based feedforward control (denoted by Σ_{ffwd} in the figure legend) and prediction horizon $t_f = 0.1$

2DOF control scheme. A slight deterioration of the tracking behaviour can also be seen for a reduced prediction horizon. However, as the main control action for the transition is provided by the feedforward control, this deterioration remains comparatively small. This confirms the observation [18] that the disturbance rejection may be designed nearly independently from the setpoint transition in the 2DOF control scheme.

4 Conclusion

In this contribution, a 2DOF control scheme is presented for setpoint transition tracking control of a quasilinear scalar DCRS. A flatness-based feedforward controller and a receding horizon feedback tracking controller are designed in an early lumping approach using a finite-difference semi-discretization of the DCRS. Thereby, input constraints are systematically incorporated into the receding horizon control design. In the simulation studies, the 2DOF control scheme shows a good performance for both trajectory tracking and disturbance rejection. Furthermore, the 2DOF control scheme allows for a nearly independent tuning of the tracking performance and the disturbance rejection. Stability of the feedback tracking control scheme as well as the further decrease of the computational costs are subject to current research activities.

References

1. Chen, H., Allgöwer, F.: A quasi-infinite horizon nonlinear model predictive control scheme with guaranteed stability. Automatica 34(10), 1205–1217 (1998)
2. Fliess, M., Lévine, J., Martin, P., Rouchon, P.: Flatness and defect of non–linear systems: Introductory theory and examples. Int. J. Control 61(6), 1327–1361 (1995)
3. Graichen, K., Kugi, A.: Stability and incremental improvement of suboptimal MPC without terminal constraints. IEEE Trans. Autom. Control 55(11), 2576–2580 (2010)
4. Graichen, K., Käpernick, B.: A real-time gradient method for nonlinear model predictive control. In: Zheng, T. (ed.) Frontiers of Model Predictive Control, pp. 9–28 (2012), http://www.intechopen.com/books/frontiers-of-model-predictive-control/a-real-time-gradient-method-for-nonlinear-model-predictive-control
5. Jadbabaie, A., Hauser, J.: On the stability of receding horizon control with a general terminal cost. IEEE Trans. Autom. Control 50(5), 674–678 (2005)
6. Jensen, K.F., Ray, W.H.: The bifurcation behavior of tubular reactors. Chem. Eng. Sci. 17(2), 199–222 (1982)
7. Laroche, B., Martin, P., Rouchon, P.: Motion planning for the heat equation. Int. J. Robust Nonlinear Control 10, 629–643 (2000)
8. Limon, D., Alamo, T., Salas, F., Camacho, E.F.: On the stability of constrained MPC without terminal constraint. IEEE Trans. Autom. Control 51(5), 832–836 (2006)
9. Lions, J.L.: Optimal Control of Systems Governed by Partial Differential Equations. Springer, Berlin (1970)
10. Lynch, A.F., Rudolph, J.: Flatness-based boundary control of a class of quasilinear parabolic distributed parameter systems. Int. J. Control 75(15), 1219–1230 (2002)
11. Mayne, D.Q., Rawlings, J.B., Rao, C.V., Scokaert, P.O.M.: Constrained model predictive control: stability and optimality. Automatica 36(6), 789–814 (2000)
12. Meurer, T.: Feedforward and feedback tracking control of diffusion–convection–reaction systems using summability methods. Fortschritt-Berichte VDI Nr. 8/1081. VDI Verlag, Düsseldorf (2005)
13. Meurer, T., Kugi, A.: Tracking control for boundary controlled parabolic pdes with varying parameters: Combining backstepping and differential flatness. Automatica 45, 1182–1194 (2009)
14. Ollivier, F., Sedoglavic, A.: A generalization of flatness to nonlinear systems of partial differential equations. Application to the command of a flexible rod. In: Proc. 5th IFAC NOLCOS, St. Petersburg, Russia, pp. 196–200 (2001)
15. Tröltzsch, F., Wachsmuth, D.: On convergence of a receding horizon method for parabolic boundary control. Optim. Methods Soft. 19(2), 201–216 (2004)
16. Utz, T., Graichen, K.: Computationally efficient receding horizon trajectory tracking control for a tubular reactor example. Proc. Appl. Math. Mech. 12, pp. 721–722 (2012)
17. Utz, T., Graichen, K., Kugi, A.: Trajectory planning and receding horizon tracking control of a quasilinear diffusion-convection-reaction system. In: Proc. 8th IFAC NOLCOS, Bologna, Italy, pp. 587–592 (2010)
18. Utz, T., Hagenmeyer, V., Mahn, B.: Comparative evaluation of nonlinear model predictive and flatness-based two-degree-of-freedom control design in view of industrial application. J. Process Control 17(2), 129–141 (2007)
19. Utz, T., Meurer, T., Kugi, A.: Trajectory planning for quasilinear parabolic distributed parameter systems based on finite-difference semi-discretizations. Int. J. Control 83(6), 1093–1106 (2010)

Temporal and One-Step Stabilizability and Detectability of Time-Varying Discrete-Time Linear Systems

L. Gerard Van Willigenburg[1,*] and Willem L. De Koning[2]

[1] Systems & Control Group of Wageningen University,
P.O. Box 17, 6700 AA Wageningen, The Netherlands
gerard.vanwilligenburg@wur.nl
[2] Department of Mathematics of Delft
University of Technology, Kroeskarper 6, Leiden, The Netherlands
wilros@planet.nl

Abstract. Time-varying discrete-time linear systems may be temporarily uncontrollable and unreconstructable. This is vital knowledge to both control engineers and system scientists. Describing and detecting the temporal loss of controllability and reconstructability requires considering discrete-time systems with variable dimensions and the j-step, k-step Kalman decomposition. In this note for linear discrete-time systems with variable dimensions measures of temporal and one-step stabilizability and detectability are developed. These measures indicate to what extent the temporal loss of controllability and reconstructability may lead to temporal instability of the closed loop system when designing a static state or dynamic output feedback controller. The measures are calculated by solving specific linear quadratic cheap control problems.

Keywords: Temporal system properties, linear discrete-time systems, cheap LQ control problems, j-step k-step Kalman decomposition.

1 Introduction

Feedback control design and stability analysis of nonlinear systems along trajectories is often performed using the linearized dynamics about the trajectory [1], [2]. If the trajectory is time-varying the linearized model is *time-varying*. If in addition the nonlinear dynamics or the controls are non-smooth, i.e. in the case of bang-bang or digital control, the *structure* of the time-varying linearized system may change. Even if the nonlinear dynamics and the controls are smooth the structure of the time-varying linearized system may almost change. For control system design this is vital information since this structure reveals the *temporal loss* of controllability and reconstructability of the linearized system. They in turn may lead to *temporal instability* of a closed-loop control system [3], [4]. Recently we investigated these issues for continuous-time systems assuming continuous-time control. This investigation lead to the introduction of the properties temporal and differential

* Corresponding author.

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 306–317, 2013.

stabilizability and detectability for continuous-time linear systems [5]. In addition *measures* of these properties were introduced and calculated by solving specific *linear quadratic cheap control problems* [5], [6], [7].

Associated with computer control are digital control problems (sampled-data control problems). They concern the control of continuous-time systems by means of piecewise constant controls using sampled measurements. A common approach is to transform such control problems into equivalent discrete-time control problems [8], [9], [10]. Following this approach feedback control system design is performed in discrete-time. This motivates the discrete-time development in this paper that on the one hand parallels, but on the other is also very different from the one in continuous-time. The fact that discrete-time is not dense, as opposed to continuous-time, causes some major differences. In continuous-time our investigation required the introduction of piecewise constant rank systems and the differential Kalman decomposition [3], [4]. In discrete-time their counterparts are discrete-time linear systems with variable state dimensions and the j-step, k-step Kalman decomposition [11].

This paper develops *measures* of temporal stability of time-varying linear discrete-time systems over arbitrary finite time intervals, notably intervals where controllability or reconstructability is lost temporarily. Associated to this, measures of *temporal and one-step stabilizability and detectability* are developed. These measures can for instance be used to analyse temporal instability of a closed loop control system design using LQG output feedback.

Temporal stability may sound as a contradiction because formally stability relates to behavior when time tends to infinity. However, in one of his early seminal papers [12] Kalman together with Bertram already proposed measures of stability over finite time intervals (page 386). Intuitively stability relates to growth of the system state. Intuitively over intervals where the state grows we call the system temporal unstable and over intervals where the state decays, we call the system temporal stable. This intuition is formalized by the temporal stability property proposed in this note. This property is derived from a *measure* of temporal stability also proposed in this note that measures the maximum growth of the state over an arbitrary interval. Our concept of stability over a finite time interval differs from what is called finite-time stability [13], [14]. The reason we make a different choice is that our measures, their computation and the associated control system designs, come down to solving standard LQ problems. The standard LQ problems are of a special type called cheap control LQ problems [6], [7] . They are characterized by a control penalty that tends to zero. Computations and control system design associated to finite-time stability concern matrix inequalities [13], [14]. Generally these are much more difficult to solve.

2 Temporal and One-Step Stability, Stabilizability and Detectability

Temporal uncontrollability/unreachability and temporal unreconstructability/ unobservability of linear time-varying systems was introduced and investigated in continuous-time [3], [4] and in discrete-time [11]. Intuitively, temporal stabilizability

and temporal detectability are associated properties that apply over intervals where the system is temporal uncontrollable/unreachable and temporal unreconstructable/unobservable respectively. In continuous-time this was formalized in [5]. In this section we formalize the discrete-time case. This requires considering variable dimension discrete-time linear systems (VDD systems) [11], [15], [16], [17] as well as j-step controllability, j-step reachability, k-step reconstructability, k-step observability and the associated j-step, k-step Kalman decomposition. All these are introduced in [11] that relies partly on [18]. In this section we consider VDD systems with a time domain $[i_0, i_N]$ where i_0 may tend to $-\infty$ and i_N may tend to $+\infty$. Intervals where the VDD system is temporal uncontrollable/unreachable or temporal unreconstructable/unobservable are denoted by $[i_s, i_f]$.

Definition 1.

A VVD system is called *j-step unreachable over the interval* $[i_s, i_f + j]$*/ j-step uncontrollable over the interval* $[i_s - j, i_f]$, $i_0 + j \le i_s < i_f \le i_N - j$ if $\forall i \in [i_s, i_f + j]$ the system is not j-step reachable at time i / not j-step controllable from time $i - j$.

Lemma 1.

If i_s, i_f satisfy the conditions in Definition 1 then over the interval $[i_s, i_f]$ the VDD system is 1) not j-step reachable at each time and 2) not j-step controllable from each time.

Proof:
Follows immediately from [11] and Definition 1.

Definition 2.

A VDD system that satisfies the conditions in Definition 1 is called *j-step uncontrollable/unreachable over the interval* $[i_s, i_f]$.

Definition 3 (Dual of Definition 1)**.**

A VVD system is called *k-step unobservable over the interval* $[i_s - k, i_f]$*/ k-step unreconstructable over the interval* $[i_s, i_f + k]$, $i_0 + k \le i_s < i_f \le i_N - k$ if $\forall i \in [i_s - k, i_f]$ the system is not k-step observable at time i / not k-step reconstructable from time $i + k$.

Lemma 2 (dual of Lemma 1)**.**

If i_s, i_f satisfy the conditions in Definition 3 then over the interval $\left[i_s, i_f\right]$ the VDD system is 1) not k-step observable at each time and 2) not k-step reconstructable from each time.

Definition 4 (dual of Definition 2)**.**

A VDD system that satisfies the conditions in Definition 3 is called *k-step unreconstructable/unobservable over the interval* $\left[i_s, i_f\right]$.

Application of the j-step k-step Kalman decomposition [11], [19] at each time $i \in \left[i_0, i_N\right]$, reveals all closed intervals (i.e. consisting of at least two consecutive discrete-time instants) where the system is j-step uncontrollable/unreachable and dually all closed intervals where the system is k-step unreconstructable/unobservable. As in Definition 2 and Definition 4 such intervals will be denoted by $\left[i_s, i_f\right]$. These closed intervals are precisely the intervals where stability of the closed loop system may be lost temporarily when designing static state and dynamic output feedback controllers.

Stabilizability is a property that relates entirely to the uncontrollable part of a system. A general approach to determine stabilizability is to extract this uncontrollable part, that is autonomous, by means of a Kalman decomposition, and to determine its stability. It will become clear in this section that application of a state basis transformation changes temporal stability and stabilizability properties. To recover them we therefore need to transform back to the original state basis. As opposed to this general approach, the stabilizability analysis presented in this section is much more straightforward and simple. It does not require transformation of the state basis because it relies fully on well established standard LQ theory applied to the original system representation. Therefore the associated numerical computations are also very efficient.

The stabilizability analysis in this section is *unconventional* in the sense that stability, stabilizability and detectability over *finite time intervals* is required. Stability over an interval relates to growth of the magnitude of the state over this interval. Throughout this paper $\|\bullet\|$ denotes the matrix 2 norm. For vectors this amounts to the L2 norm. In the next section we will demonstrate how to compute numerically the temporal and one-step stabilizability and detectability measures presented in this section, using only evaluations of the system matrices.

Definition 5.

An autonomous VVD system is called *temporal stable over the interval* $\left[i_s, i_f\right]$ if for any $x_{i_s} \neq 0$, $\left\|x_{i_f}\right\| \Big/ \left\|x_{i_s}\right\| < 1$.

Loosely speaking, according to Definition 5 an autonomous VDD system is called temporal stable over $\left[i_s, i_f \right]$ if for any initial state the magnitude of the associated terminal state is smaller than that of the initial state. An important difference between our definition and other finite-time stability concepts [13], [14] is that ours does not impose any restrictions on the magnitude of the state inside the interval. The advantage of Definition 5 is that it matches LQ control design as opposed to finite-time stability that relates to control system design using matrix inequalities [13] that is generally much more complicated.

Definition 6.

Associate to Definition 5 the following *temporal stability measure*,

$$\rho\left(i_s, i_f\right) = \max_{x_{i_s} \neq 0} \left(\frac{\left\| x_{i_f} \right\|^2}{\left\| x_{i_s} \right\|^2} \right) \geq 0 . \tag{1}$$

Observe that $\rho\left(i_s, i_f\right)$ in Definition 6 is the largest possible ratio $\left\| x_{i_f} \right\|^2 \Big/ \left\| x_{i_s} \right\|^2$. This ratio matches the largest possible ratio $\left\| x_{i_f} \right\| \Big/ \left\| x_{i_s} \right\|$ in Definition 5. Therefore $\rho\left(i_s, i_f\right)$ is indeed a measure of temporal stability associated to Definition 5. The smaller $\rho\left(i_s, i_f\right)$ the larger temporal stability. It will become clear that the squares in equation (1) are needed to achieve compatibility with LQ control computations.

Theorem 1.

An autonomous VDD system is temporal stable over the time interval $\left[i_s, i_f \right]$ if and only if,

$$\rho\left(i_s, i_f\right) = \left\| \Phi^T_{i_s, i_f} \Phi_{i_s, i_f} \right\| < 1, \tag{2}$$

where Φ_{i_s, i_f} represents the state transition matrix of the associated autonomous system from time i_s to i_f.

Proof:
Because Theorem 1 applies to autonomous systems,

$$x_{i_f} = \Phi_{i_s, i_f} x_{i_s} . \tag{3}$$

Using equation (3) the temporal stability measure (1) becomes,

$$\rho\left(i_s, i_f\right) = \max_{x_{i_s} \neq 0} \left(\frac{\left\| \Phi_{i_s, i_f} x_{i_s} \right\|^2}{\left\| x_{i_s} \right\|^2} \right) = \max_{x_{i_s} \neq 0} \left(\frac{x_{i_s}^T \Phi_{i_s, i_f}^T \Phi_{i_s, i_f} x_{i_s}}{x_{i_s}^T x_{i_s}} \right) = \left\| \Phi_{i_s, i_f}^T \Phi_{i_s, i_f} \right\|. \tag{4}$$

The last equality in equation (4) holds because $\Phi_{i_s, i_f}^T \Phi_{i_s, i_f}$ is nonnegative symmetric. Theorem 1 now follows from (4), Definition 5 and Definition 6 and,

$$\left\| x_{i_f} \right\| / \left\| x_{i_s} \right\| < 1 \Leftrightarrow \left\| x_{i_f} \right\|^2 / \left\| x_{i_s} \right\|^2 < 1. \tag{5}$$

Stabilizability over a finite time-interval relates to the ability to stabilize the system over that interval by means of control.

Definition 7.

Associate to Definition 5 and Definition 6 the following *temporal stabilizability measure* that applies to VDD systems considered over the interval i_s, i_f,

$$\rho_{\min}\left(i_s, i_f\right) = \max_{x_{i_s} \neq 0} \left(\frac{\min_{u_i \mid x_{i_s}} \left\| x_{i_f} \right\|^2}{\left\| x_{i_s} \right\|^2} \right) \geq 0, \tag{6}$$

where $u_i \mid x_{i_s}$ indicates a control law dependent on x_{i_s}.

Definition 8.

A VDD system is called *temporal stabilizable over* $\left[i_s, i_f \right]$ if $\rho_{\min}\left(i_s, i_f\right) < 1$.

Theorem 2.

A VDD system is temporal controllable over $\left[i_s, i_f \right] \Rightarrow \rho_{\min}\left(i_s, i_f\right) = 0 \Rightarrow$ the VDD system is temporal stabilizable over $\left[i_s, i_f \right]$.

Proof:

If a VDD system is temporal controllable over $\left[i_s, i_f \right]$, then according to Definition 1 and [11], any state x_{i_s} can be controlled to $x_{i_f} = 0$. This implies $\rho_{\min}\left(i_s, i_f\right) = 0$ and, according to Definition 8, temporal stabilizability over $\left[i_s, i_f \right]$.

Remark 1.

As with ordinary controllability and stabilizability, temporal controllability is a stronger property than temporal stabilizability.

To state the main theorem in this section consider the following parameterized discrete-time LQ problem. Given the system,

$$x_{i+1} = \Phi_i x_i + \Gamma_i u_i,\ i \in \left[i_s, i_f - 1 \right], \tag{7}$$

with initial state,

$$x_{i_s}, \tag{8}$$

find the control $u_i,\ i \in \left[i_s, i_f - 1 \right]$ that minimizes the cost function,

$$J_{LQ}(\varepsilon) = x_{i_f}^T H x_{i_f} + \sum_{i=i_s}^{i_f - 1} \left[x_i^T Q_i x_i + u_i^T R_i^\varepsilon u_i \right], \tag{9}$$

with,

$$H = I_n,\ Q_i = 0,\ R_i^\varepsilon = \varepsilon I_m,\ 0 \le \varepsilon \ll 1. \tag{10}$$

If $\varepsilon > 0$ the Linear Quadratic control problem (7), (8)-(10) satisfies $H \ge 0,\ Q_i \ge 0,\ R_i^\varepsilon > 0$. In this standard case it is well known that the optimal control is given by,

$$u_i = -L_i^\varepsilon x_i,\ L_i^\varepsilon = \left(\Gamma_i^T S_{i+1}^\varepsilon \Gamma_i + R_i^\varepsilon \right)^{-1} \Gamma_i^T S_{i+1}^\varepsilon \Phi_i, \tag{11}$$

and the minimum cost by,

$$J_{LQ}^*(\varepsilon) = x_{i_s}^T S_{i_s}^\varepsilon x_{i_s}, \tag{12}$$

where $S_i^\varepsilon,\ i \in \left[i_s, i_f \right]$ is the solution of the matrix Riccati difference equation,

$$S_i^\varepsilon = \Phi_i^T S_{i+1}^\varepsilon \Phi_i - L_i^T \left(\Gamma_i^T S_{i+1}^\varepsilon \Gamma_i + R_i^\varepsilon \right) L_i + Q_i,\ S_{i_f}^\varepsilon = H. \tag{13}$$

Theorem 3.

$$S_i^* = \lim_{\varepsilon \downarrow 0} S_i^\varepsilon, \tag{14}$$

exists, where $S_i^\varepsilon, \varepsilon \downarrow 0$ satisfies the matrix Riccati difference equation (13) with data as specified by equation (10). Furthermore,

$$\rho_{\min}\left(i,i_f\right)=\left\|S_i^*\right\|,\ i\in\left[i_s,i_f-1\right]. \tag{15}$$

As a special case of (15),

$$\rho_{\min}\left(i_s,i_f\right)=\left\|S_{i_s}^*\right\|. \tag{16}$$

Proof:
First observe that in the parameterized LQ problem (7)-(10) we may replace the initial time i_s by $i'\in\left[i_s,i_f-1\right]$. This also hold for the stabilizability measure $\rho_{\min}$. Next from equations (9), (10) observe that

$$\min_{u_i|x_{i'}} J_{LQ}\left(0\right)=\min_{u_i|x_{i'}}\left(x_{i_f}^T x_{i_f}\right)=\min_{u_i|x_{i'}}\left\|x_{i_f}\right\|^2 \tag{17}$$

Now the key to proving (14), (15) is to prove that,

$$\min_{u_i|x_{i'}} J_{LQ}\left(0\right)=\lim_{\varepsilon\downarrow 0} J_{LQ}^*\left(\varepsilon\right)=x_{i'}^T S_{i'}^* x_{i'} \tag{18}$$

Suppose equation (18) holds. Then from equations (6), (17), (18),

$$\begin{aligned}\rho_{\min}\left(i',i_f\right)&=\max_{x_{i'}\neq 0}\left(\frac{\min_{u_i|x_{i'}}\left\|x_{i_f}\right\|^2}{\left\|x_{i'}\right\|^2}\right)=\\ \max_{x_{i'}}&\left(\frac{x_{i'}^T S_{i'}^* x_{i'}}{x_{i'}^T x_{i'}}\right)=\left\|S_{i'}^*\right\|\end{aligned} \tag{19}$$

The last equality in equation (19) holds because $S_{i'}^*$ is nonnegative symmetric. So we are left to prove equation (18). Consider the j-step, k-step Kalman decomposition at time i_f with $j=i_f-i'$. According to this decomposition the linear system (7) can be decomposed into a part that is j-step controllable from time i' and a part that is autonomous. The contribution of the j-step controllable part to $\min_{u_i|x_{i'}} J_{LQ}\left(0\right)$ is zero. The contribution to $J_{LQ}^*\left(\varepsilon\right)$ tends to zero as $\varepsilon\downarrow 0$. The contribution of the autonomous part to both $\min_{u_i|x_{i'}} J_{LQ}\left(0\right)$ and $J_{LQ}^*\left(\varepsilon\right)$ is fixed and independent of ε. Because the system matrices are bounded this contribution is also finite. This proves the existence of the limit (16) and the equality (18).

Remark 2.

There are three reasons for considering $0<\varepsilon \ll 1$ in equation (10), instead of $\varepsilon = 0$. Taking $0<\varepsilon \ll 1$, ε may be used to 1) keep the control within certain bounds that apply in practice and 2) as a numerical tolerance to prevent ill-conditioning of the computation of equation (11) when $\Gamma_i^T S_{i+1}^{\varepsilon} \Gamma_i$ is not full rank and $L_i \to \infty$ as $\varepsilon \downarrow 0$. In practice the selection of $0<\varepsilon \ll 1$ will be a compromise and S_i^{ε} will *approximate* S_i^*, $i \in \left[i_s, i_f - 1 \right]$. As a result all computations in this paper involving S_i^{ε} will be approximations, although generally very good ones. Thirdly $\varepsilon = 0$ leads to a singular LQ problem that is generally much more difficult to solve and the solution of which need not be unique.

When analyzing control systems the state behavior over the entire interval $\left[i_s, i_f \right]$ is generally of interest, not just the behavior at the initial time i_s and the final time i_f. This behavior is partly considered by equation (15) of Theorem 3 that determines the stabilizability measure for each sub interval $\left[i, i_f \right]$, $i \in \left[i_s, i_f - 1 \right]$. The following theorem introduces a one-step stabilizability measure that applies to individual time instants.

Theorem 4.

$\left\| S_i^* \right\| - \left\| S_{i+1}^* \right\|$ is *a one-step stabilizability measure (os-stabilizability measure) at time* $i \in \left[i_s, i_f - 1 \right]$.

Proof:
From (15),

$$\rho_{\min}\left(i_s, i_f\right) = \left\| S_{i_s}^* \right\| = \left\| S_{i_f}^* \right\| + \sum_{i=i_s}^{i_f-1} \left(\left\| S_i^* \right\| - \left\| S_{i+1}^* \right\| \right) \tag{20}$$

so $\left\| S_i^* \right\| - \left\| S_{i+1}^* \right\|$, $i \in \left[i_s, i_f - 1 \right]$ is the *one-step contribution at time* i to the temporal stabilizability measure $\rho_{\min}\left(i_s, i_f\right)$. If this contribution is negative $\rho_{\min}\left(i_s, i_f\right)$ decreases and temporal stabilizability increases.

Definition 9.

A VDD system is called *one-step stabilizable (os-stabilizable) at time* $i \in \left[i_s, i_f - 1 \right]$ if $\left\| S_i^* \right\| - \left\| S_{i+1}^* \right\| < 0$.

Because for a VDD system temporal and one-step detectability are dual to temporal and one-step stabilizability, the following definitions and theorems are stated without further explanation and proof.

Theorem 5 (dual of Theorem 3)**.**

$$P_i^* = \lim_{\varepsilon \downarrow 0} P_i^\varepsilon \,, \tag{21}$$

exists, where $P_i^\varepsilon, \varepsilon \downarrow 0$ satisfies the matrix Riccati difference equation that is dual to (13),

$$P_{i+1}^\varepsilon = \Phi_i P_i^\varepsilon \Phi_i^T - L_i^\varepsilon \left(C_i P_i^\varepsilon C_i^T + R_i \right) L_i^{\varepsilon T} + Q_i,\ P_{i_s}^\varepsilon = H \,, \tag{22}$$

with,

$$L_i^\varepsilon = \Phi_i P_i^\varepsilon C_i^T \left(C_i P_i^\varepsilon C_i^T + R_i^\varepsilon \right)^{-1}, \tag{23}$$

with data as specified by equation (10). Furthermore,

$$\sigma_{\min}\left(i_s, i\right) = \left\| P_i^* \right\|,\ i \in \left[i_s + 1, i_f \right]. \tag{24}$$

where $\sigma_{\min}\left(i, i_f\right)$ is a *temporal detectability measure over the interval* $\left[i, i_f \right]$. As a special case,

$$\sigma_{\min}\left(i_s, i_f\right) = \left\| P_{i_f}^* \right\|. \tag{25}$$

Definition 10 (dual of Definition 8)**.**

A VDD system is called *temporal detectable over* $\left[i_s, i_f \right]$ if $\sigma_{\min}\left(i_s, i_f\right) < 1$.

Theorem 6 (dual of Theorem 4)**.**

$\left\| P_{i+1}^* \right\| - \left\| P_i^* \right\|$ is *a one-step detectability measure (os-detectability measure) at time* $i \in \left[i_s, i_f - 1 \right]$.

Definition 11 (dual of Definition 9)**.**

A VDD system is called *one-step detectable (os-detectable) at time* $i \in \left[i_s, i_f - 1 \right]$ if $\left\| P_{i+1}^* \right\| - \left\| P_i^* \right\| < 0$.

3 Conclusions

New temporal properties and associated measures for control system design concerning time-varying linear discrete-time systems were introduced in this paper. The properties and associated measures concern temporal and one-step stabilizability and detectability. They indicate to what extent control system design is problematic when discrete-time linear time-varying systems are temporal uncontrollable or temporal unreconstructable. Temporal uncontrollability and unreconstructability are detected by the j-step, k-step Kalman decomposition. As demonstrated in this paper, after introduction of a suitable, simple stability property, that applies over finite time intervals, application of ordinary standard LQ theory and algorithms enables the computation of associated temporal and one-step stabilizability and detectability *measures*. These determine to what extent a static or dynamic feedback control system becomes temporal unstable. A major application concerns the temporal stability analysis of digital perturbation output feedback controllers for nonlinear systems tracking control and state trajectories that may be optimal [1], [10].

As an alternative to LQ theory, temporal stabilizability may be determined by extracting the temporal uncontrollable or temporal unreconstructable subsystems and analyzing their temporal stability. In principle, the j-step, k-step Kalman decomposition is able to extract these subsystems. The extraction employs state basis transformations that generally change temporal stability properties. The approach presented in this paper is more simple and direct because it applies standard LQ theory to the original, untransformed system.

Although the LQ problems in this paper are singular in principle, it is advantageous to approximate them by non-singular LQ problems, as demonstrated in this paper. The interpretation of $\left\|S_i^*\right\|$ as a temporal stabilizability measure is new and highly interesting. The same applies to the interpretation of $\left\|S_i^*\right\| - \left\|S_{i+1}^*\right\|$ as a one-step stabilizability measure that measures the contribution to stabilizability of each single time-step.

Along the lines of this paper we are also currently exploring temporal properties of time-varying linear systems with white stochastic parameters [20]. Among others these enable robust digital optimal perturbation feedback design for nonlinear systems.

References

[1] Athans, M.: The role and use of the Linear- Quadratic-Gaussian problem in control system design. IEEE Trans. Aut. Contr. 16, 529–552 (1971)

[2] van Willigenburg, L.G., De Koning, W.L.: On the synthesis of time-varying LQG weights and noises along optimal control and state trajectories. Optimal Control Applications and Methods 27, 137–160 (2006)

[3] Van Willigenburg, L.G., De Koning, W.L.: A Kalman decomposition to detect temporal linear system structure. In: Proceedings European Control Conference, Kos, Greece, July 2-7, Paper nr. 78, 6 p. (2007)
[4] Van Willigenburg, L.G., De Koning, W.L.: Temporal linear system structure. IEEE Trans. Aut. Contr. 53(5), 1318–1323 (2008)
[5] Van Willigenburg, L.G., De Koning, W.L.: Temporal and differential stabilizability and detectability of piecewise constant rank systems. Optimal Control Application & Methods, published online in Wiley Online Library (wileyonline library.com) (2011), doi:10.1002/oca.997
[6] Jameson, A., O'Malley, R.E.: Cheap control of the time-invariant regulator. Appl. Math. & Optimization 1(4), 337–354 (1975)
[7] Kokotovic, P.V., O'Malley, R.E., Sannuti, P.: Singular perturbations and order-reduction in control theory – an overview. Automatica 12, 123–132 (1976)
[8] Levis, A.H., Schlueter, R.A., Athans, M.: On the behavior of optimal linear sampled-data regulators. International Journal of Control 13, 343–361 (1971)
[9] Van Willigenburg, L.G., De Koning, W.L.: The digital optimal regulator and tracker for stochastic time-varying systems. International Journal of Systems Science 12, 2309–2322 (1992)
[10] Van Willigenburg, L.G.: Digital optimal control and LQG compensation of asynchronous and aperiodically sampled nonlinear systems. In: Proceedings 3rd European Control Conference, Rome, Italy, vol. 1, pp. 496–500 (September 1995)
[11] van Willigenburg, L.G., De Koning, W.L.: Temporal linear system structure: The discrete-time case. In: Proceedings of the ECC 2009, Budapest, August 23-26, pp. 225–230 (2009)
[12] Kalman, R.E., Bertram, J.E.: Control system design via the "Second Method" of Lyapunov, I Continuous-time systems. Transactions of the ASME, Journal of Basic Engineering, 371–393 (June 1960)
[13] Amato, F., Ariola, M., Carbone, M., Cosentiono, C.: Finite-Time Control of Linear Systems: A Survey. In: Current Trends in Nonlinear Systems and Control. Systems and Control: Foundations & Applications, Part II, pp. 195–213 (2006)
[14] Amato, F., Ambrosino, R., Ariola, M., Cosentino, C.: Finite-time stability of linear time-varying systems with jumps. Automatica 45, 1354–1358 (2009)
[15] Gohberg, I., Kaashoek, M.A., Lerer, L.: Minimality and realization of discrete time-varying systems. Operator Theory: Advances and Applications 56, 261–296 (1992)
[16] Van Willigenburg, L.G., De Koning, W.L.: Minimal and non-minimal optimal fixed-order compensators for time-varying discrete-time systems. Automatica 38, 157–165 (2002)
[17] Sandberg, H., Rantzer, A.: Balanced truncation of linear time-varying systems. IEEE Trans. Aut. Contr. 49(2), 217–229 (2004)
[18] Van Willigenburg, L.G., De Koning, W.L.: Linear systems theory revisited. Automatica 44, 1686–1696 (2008)
[19] Boley, D.: Computing the Kalman decomposition: An optimal method. IEEE Trans. Aut. Contr. 29(1), 51–53 (1984)
[20] van Willigenburg, L.G., De Koning, W.L.: Compensatability and optimal compensation of systems with white parameters in the delta domain. International Journal of Control 83(12), 2546–2563 (2010)

Optimal Control of Unsteady Flows Using a Discrete and a Continuous Adjoint Approach

Angelo Carnarius[1], Frank Thiele[2], Emre Özkaya[3], Anil Nemili[3], and Nicolas R. Gauger[3]

[1] Institut für Strömungmechanik und Technische Akustik, TU Berlin, Müller-Breslau-Str. 8, Berlin, 10623, Germany
angelo.carnarius@cfd.tu-berlin.de
[2] CFD Software Entwicklungs- und Forschungsgesellschaft mbH, Wolzogenstr. 4, Berlin, 14163, Germany
frank.thiele@cfd-berlin.com
[3] Computational Mathematics Group, CCES, RWTH Aachen University, Schinkelstr. 2, Aachen, 52062, Germany
{ozkaya,nemili,gauger}@mathcces.rwth-aachen.de

Abstract. While active flow control is an established method for controlling flow separation on vehicles and airfoils, the design of the actuation is often done by trial and error. In this paper, the development of a discrete and a continuous adjoint flow solver for the optimal control of unsteady turbulent flows governed by the incompressible Reynolds-averaged Navier-Stokes equations is presented. Both approaches are applied to testcases featuring active flow control of the blowing and suction type and are compared in terms of accuracy of the computed gradient.

Keywords: optimal control, active flow control, discrete adjoint, continuous adjoint, unsteady turbulent flows, URANS.

1 Introduction

For many aerodynamic applications in aviation and automotive industry, flow separation has to be taken into account. The lift of an airfoil at a high angle of attack, for instance, decreases drastically, if the flow separates on the suction side.

Many studies in the past decades have shown that the aerodynamic behaviour of a body can be improved by using active flow control [4]. However, the choice of the control parameters is very case-specific and not trivial. An efficient method of finding the optimal set of actuation parameters is the gradient-based optimisation, which requires the calculation of the gradient of the cost function with respect to the control parameters. The control variables are then updated in an iterative manner according to a descent direction, which can be obtained from the gradient vector.

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 318–327, 2013.

A very efficient way of computing the gradient is by using adjoint methods. Compared to simpler approaches such as Finite Differences or the Complex Taylor Series Expansion (CTSE) [13], adjoint-based methods compute the gradient vector at a fixed expense independent of the number of actuation parameters. Adjoint methods are commonly divided into the continuous and the discrete approach.

In the continuous adjoint method [10], first the optimality system for a given objective function is derived and the resulting PDEs are then discretised and solved numerically. This procedure is called *first optimise then discretise*. The continuous approach is numerically efficient but it is known to suffer from consistency problems. The gradient can become inaccurate for insufficient time steps and grid spacing, which can be disadvantageous for complex configurations. Furthermore, most statistical turbulence models required for the unsteady Reynolds-averaged Navier-Stokes equations (URANS) are non-differentiable. The common approach is to use the so-called *constant eddy viscosity* or *frozen turbulence* assumption, i.e. the eddy viscosity is treated as independent of the control parameters and therefore taken from the primal solution. This assumption can lead to significant errors in the computed gradient [1].

The concept of the discrete adjoint method [3,6,12] is to *first discretise then optimise*, i.e. the discretised governing equations are used to derive the optimality system. This approach allows the generation of a fully consistent optimality system independent of the grid size, time step and turbulence model, as it does not require analytical differentiability [11]. Furthermore, Automatic Differentiation (AD) techniques [8] can be used to develop the discrete adjoint solver for a given simulation code in a semi-automatic fashion.

In this paper, we present the development of a continuous and a discrete adjoint solver for the optimal control of unsteady turbulent flows governed by the incompressible URANS equations. Both approaches, which are presented in more detail in sections 3 and 4, are based on the same flow solver ELAN [16]. For the current study, the adjoint solvers are applied to testcases which feature active flow control of the blowing and suction type.

2 Flow Model

Governing equations For this study, the unsteady, incompressible, turbulent flow in the domain Ω is described by the Reynolds-averaged Navier-Stokes equations[1]

$$\frac{\partial u_i}{\partial x_i} = 0 \tag{1}$$

$$\frac{\partial \varrho u_i}{\partial t} + \frac{\partial \varrho u_i u_j}{\partial x_j} + \frac{\partial p}{\partial x_i} - \frac{\partial}{\partial x_j}\left[(\mu + \mu_t)\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)\right] = 0\,, \tag{2}$$

[1] In the following, the Einstein summation convention is used, which implies summation from 1 to 3 over indices which appear twice in a single term. Indices, which appear only once take the value 1, 2 and 3 individually.

where u_i and p are the Reynolds-averaged velocity and pressure, respectively. The density ϱ and the dynamic viscosity μ are constant for the cases shown here. The eddy viscosity μ_t is obtained from the Wilcox-k-ω-model [15], which consists of transport equations for the turbulent kinetic energy k and the turbulent frequency ω.

Boundary Conditions At the farfield boundaries (Γ_i), u_i, k and ω were prescribed. On the body surface (Γ_b), u_i and $\partial k/\partial n$ were set to zero, whereas a high-Re boundary condition [16] was used for the turbulent frequency. At the outflow (Γ_o), the gradient of the turbulent quantities normal to the boundary was set to zero. For the Navier-Stokes equations, we want the sum of normal and friction forces to vanish at the outlet, i.e.

$$-pn_i + (\mu + \mu_t)\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) n_j = 0\,. \tag{3}$$

At the control segment (Γ_c), the actuation velocity $c_i(b_j)$ as well as the turbulent quantities were prescribed. The dependence of $c_i(b_j)$ on the vector of the actuation parameters, b_j, is case-specific and is given in the testcase descriptions.

3 Continuous Adjoint Approach

Let J be the objective function to be minimised. Then the optimisation problem can be stated as

$$J(u_i, p, c_i(b_j)) \;\rightsquigarrow\; \min \text{ over } (u_i, p, c_i(b_j)) \text{ subject to } R(u_i, p, c_i(b_j)) = 0\,, \tag{4}$$

where R represents the state equations including the boundary conditions. In the cases presented here, the objective function can be written as[2]

$$J = -\frac{1}{T}\int_0^T \int_{\Gamma_{b,c}} \left[(\mu + \mu_t)\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) n_j - pn_i - \varrho u_i u_j n_j\right] e_i \, dA\, dt$$
$$+ \frac{\gamma}{u_\infty}\frac{1}{T}\int_0^T \int_{\Gamma_c} \varrho u^2 \sqrt{(u_i n_i)^2 + \epsilon}\, dA\, dt\,. \tag{5}$$

If the unity vector e_i is parallel to the mean flow, eq. 5 is the time-averaged drag. If e_i is oriented normal to the mean flow, eq. 5 represents the time-averaged downforce. The second integral is a penalty term which accounts for the energy consumption of the actuation and can be scaled by the factor γ. The parameter ϵ is only required for the differentiability of the penalty term, i.e. $1 \gg \epsilon > 0$.

[2] Note, that the negative sign of the first integral is a result of the convention that the normal vector n_i is directed out of the wall-adjacent control volume, i.e. into the body surface.

To solve the minimisation problem, one first introduces the Lagrange function

$$L(u_i, p, c_i(b_j), v_i, q) = C\, J(u_i, p, c_i(b_j)) + \int_0^T \int_\Omega q R_\varrho \, dV \, dt + \int_0^T \int_\Omega v_i R_u \, dV \, dt \,, \quad (6)$$

where the Lagrange multipliers v_i and q are the adjoint velocity and pressure, respectively, and C is a scaling factor to fix the units. By setting the variation of L with respect to the state variables, $\frac{\partial L}{\partial u_k}\delta u_k$ and $\frac{\partial L}{\partial p}\delta p$, to zero, one can obtain the adjoint equations. First the variations δu_k and δp have to be separated from other terms by using integration by parts and the boundary conditions have to be applied to the boundary integrals. As the resulting equations have to be fulfilled for any variation δu_k and δp, all integrals have to vanish individually, which gives the adjoint equations and boundary conditions. Setting the variation of L with respect to the control to zero and using the same procedure gives the equation for the gradient calculation. Due to the page limitation, a detailed derivation has to be omitted and the adjoint system can only be summarised. The adjoint PDEs read

$$\begin{aligned}
\frac{\partial v_i}{\partial x_i} &= 0 && \Omega \\
-\frac{\partial \varrho v_i}{\partial t} + \varrho v_j \frac{\partial u_j}{\partial x_i} - \frac{\partial \varrho u_j v_i}{\partial x_j} + \frac{\partial q}{\partial x_i} - \frac{\partial}{\partial x_j}\left[(\mu + \mu_t)\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right)\right] &= 0 && \Omega \\
v_i &= 0 && \Gamma_i \\
\varrho v_i u_j n_j - q n_i + (\mu + \mu_t)\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right) n_j &= 0 && \Gamma_o \\
v_i + \frac{C}{T} e_i &= 0 && \Gamma_{b,c} \,,
\end{aligned} \quad (7)$$

with the initial condition $v_i = 0$ at $t = T$. The gradient w.r.t. the actuation parameters can be evaluated from

$$\begin{aligned}
\frac{dJ}{db_n} = &\int_0^T \int_{\Gamma_c} \left[-\varrho u_j v_j n_m - q n_m + (\mu + \mu_t)\left(\frac{\partial v_m}{\partial x_j} + \frac{\partial v_j}{\partial x_m}\right) n_j\right] \frac{\partial c_m}{\partial b_n} \, dA \, dt \\
&+ \frac{\gamma}{u_\infty}\frac{C}{T} \int_0^T \int_{\Gamma_c} \left[\frac{\varrho c_k c_k}{\sqrt{c_i n_i c_j n_j + \epsilon}} c_l n_l n_m + 2\varrho c_m \sqrt{c_i n_i c_j n_j + \epsilon}\right] \frac{\partial c_m}{\partial b_n} \, dA \, dt \,.
\end{aligned} \quad (8)$$

Note, that the frozen turbulence assumption has been used, i.e. an adjoint turbulence model is not required.

4 Discrete Adjoint Approach

If we consider the discrete implementations of the objective function J and the state equations R, the discrete optimisation problem can be stated as:

$$J_d(y, b_i) \;\rightsquigarrow\; \min \text{ over } (y, b_i) \text{ subject to } R_d(y, b_i) = 0 \,, \quad (9)$$

where $y = (u_i, p)$ is the discrete state vector and J_d, R_d denote the discrete implementations of J and R. Note, that in the discrete realisation the actuation variables b_i are chosen as independent variables. The gradient of J_d with respect to the actuation parameters b_i can be computed from

$$\frac{dJ_d}{db_i} = \frac{\partial J_d}{\partial b_i} - \psi^\top \frac{\partial R_d}{\partial b_i}\,, \tag{10}$$

where the adjoint vector ψ can be determined by solving the adjoint system

$$\left(\frac{\partial R_d}{\partial y}\right)^\top \psi = \frac{\partial J_d}{\partial y}\,. \tag{11}$$

One way of constructing the adjoint system is by computing $\partial R_d/\partial y$ and $\partial J_d/\partial y$ using Finite Differences. The linear system of equations is then hand-coded and solved by an iterative method (e.g. GMRES). The resulting adjoint variables are used to calculate the gradient vector in eq. 10. A more promising way of developing the adjoint system is by employing the reverse mode of AD, which has the major advantage that it constructs the adjoint system consistently and computes the gradient vector dJ_d/db_i accurate to machine precision. In the present work, the discrete adjoint solver is developed by employing the AD tool TAPENADE [9] in reverse mode of differentiation.

If the reverse mode of AD is applied in a black-box fashion, the resulting adjoint code will have tremendous memory requirements. In order to reduce the excessive memory demands, we apply the reverse accumulation and checkpointing strategies for the Automatic Differentiation of the underlying flow solver. The solution strategy for the incompressible URANS equations mainly consists of two iterative loops: the time evolution step and the iterations for the velocity-pressure coupling scheme. Inside each time step, the velocity-pressure coupling iterations are performed, which are commonly known as *outer iterations* in the CFD community. It may be noted, that the outer iterations for the velocity-pressure coupling scheme converge to a fixed point in each time step. If AD is applied to these outer iterations in a black-box fashion, the flow solutions at each outer iteration of the primal solver must be saved for the adjoint part. However, the adjoint iterations require only the converged primal solution. Therefore, a lot of memory and run time can be saved, if we make use of the iterative structure and store only the converged flow solution in each physical time step. This can be achieved by employing the reverse accumulation approach [2,5], the details of which are presented in the following.

Consider the total derivative of a discrete objective function J_d with respect to the control b_i at the converged state solution y^* for any time step:

$$\frac{dJ_d(y^*, b_i)}{db_i} = \frac{\partial J_d(y^*, b_i)}{\partial b_i} + \frac{\partial J_d(y^*, b_i)}{\partial y^*}\frac{dy^*}{db_i}\,. \tag{12}$$

On the other hand, if we have a fixed point for the state solution $y^* = G(y^*, b_i) \Leftrightarrow R_d\,(y^*, b_i) = 0$, we get

$$\frac{dy^*}{db_i} = \frac{\partial G(y^*, b_i)}{\partial b_i} + \frac{\partial G(y^*, b_i)}{\partial y^*}\frac{dy^*}{db_i} = \left(I - \frac{\partial G(y^*, b_i)}{\partial y^*}\right)^{-1} \frac{\partial G(y^*, b_i)}{\partial b_i}\,. \tag{13}$$

Multiplying on both sides with $\frac{\partial J_d(y^*,b_i)}{\partial y^*}^\top$, we obtain

$$\left(\frac{\partial J_d(y^*,b_i)}{\partial y^*}\right)^\top \frac{dy^*}{db_i} = \underbrace{\left(\frac{\partial J_d(y^*,b_i)}{\partial y^*}\right)^\top \left(I - \frac{\partial G(y^*,b_i)}{\partial y^*}\right)^{-1}}_{:=\overline{y}^{*\top}} \frac{\partial G(y^*,b_i)}{\partial b_i}. \quad (14)$$

From the definition of $\overline{y}^{*\top}$ in equation (14) and making use of equation (13), the adjoint fixed point iteration can be written as

$$\overline{y}^{*\top} = \overline{y}^{*\top}\frac{\partial G(y^*,b_i)}{\partial y^*} + \left(\frac{\partial J_d(y^*,b_i)}{\partial y^*}\right)^\top. \quad (15)$$

The first term on the right hand side of the above equation is the adjoint of a single outer iteration. This can be generated by applying the reverse mode of AD to the wrapper subroutine G, which combines all the steps done within one outer iteration of the flow solver. The gradient vectors $\partial J_d/\partial y^*$ and $\partial J_d/\partial b_i$ come from the adjoint of the post-processor, which is computed only once for each time iteration.

We now focus our attention on adjoining the time iterations. In general, the computation of the unsteady adjoint solution over the time interval $[0,T]$ with N time steps requires the storage of flow solutions at time steps T_0 to T_{N-1}. The stored solutions are then used in solving the adjoint equations from T_N to T_0. For many practical aerodynamic configurations with millions of grid points and a large number of unsteady time steps, the storage costs may become prohibitively expensive.

One way of circumventing the excessive storage cost is by employing a checkpointing strategy [8], where the flow solutions are stored only at selective time steps known as checkpoints. These are then used to recompute the intermediate states that have not been stored. In the present example, we chose r ($r \ll N$) checkpoints. We then have $0 = T_0 = T_{C_1} < T_{C_2} < \cdots < T_{C_{r-1}} < T_{C_r} < T_N = T$. Here, T_{C_r} represents the time step at r^{th} checkpoint. During the adjoint computation over the subinterval $[T_{C_r}, T_N]$, required flow solutions at intermediate time steps are recomputed by using the stored solution at T_{C_r} as the initial condition. The above procedure is then repeated over other subintervals $\left[T_{C_{r-1}}, T_{C_r}\right]$ until all adjoints are computed. It may be noted, that the checkpoints can be reused when they become free. We designate them as intermediate checkpoints.

Various checkpointing strategies have been proposed based on the storage criteria. If all the checkpoints are stored in main memory, it is called single-stage checkpointing. In yet another approach called multi-stage checkpointing [14], the checkpoints are stored both in main memory and on hard-disk, thus reducing the number of flow recomputations. In the present work, we have used the single-stage binomial checkpointing strategy, which is implemented in the algorithm `revolve` [7] and generates the checkpointing schedules in a binomial fashion, so that the number of flow recomputations is proven to be optimal.

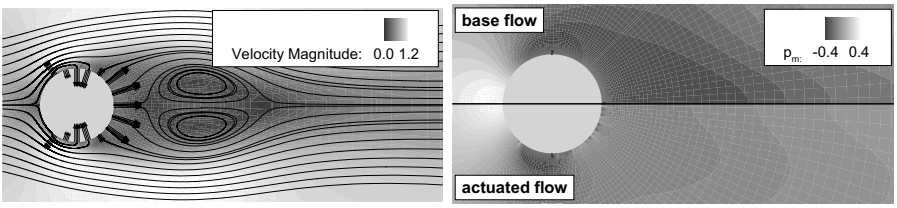
5 Numerical Results

5.1 Cylinder with Pulsed Blowing and Suction

The first application is the unsteady laminar flow around a circular cylinder at a Reynolds-number of $Re = 100$, based on the cylinder diameter D and the freestream velocity u_∞. The objective is to reduce the drag by applying pulsed blowing or suction according to

$$c_n = u_a \sin\left[2\pi f\,(t - t_0)\right] - u_a \tag{16}$$

on 15 slits, which are equidistantly distributed in 75% of the cylinder surface, see fig. 1(a). In eq. 16, c_n is the actuation velocity normal to the slit surface



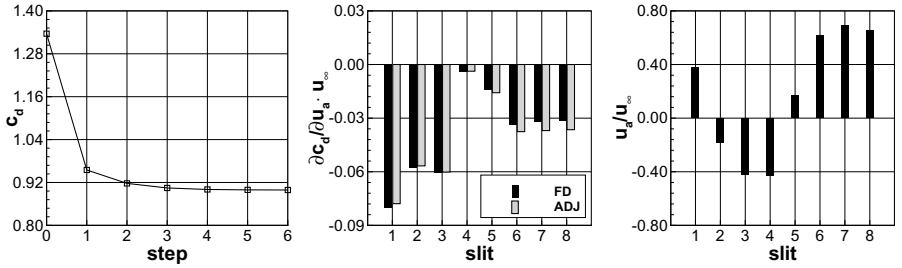
(a) Snapshot of actuated flow (b) Time-averaged pressure field

Fig. 1. Contour plots for the cylinder flow

and u_a, f and t_0 are the amplitude, frequency and phase shift, respectively. The actuation mode, i.e. blowing or suction, is set by the sign of the amplitude. For the case studied here, the actuation amplitudes at all slits are the parameters to be optimised, while the frequency and phase shift were fixed to $f = 1\,u_\infty/D$ and $t_0 = 0\,D/u_\infty$, respectively.

Only the continuous adjoint flow solver was applied to this testcase in order to test its accuracy in the unsteady laminar mode, i.e. without the influence of the frozen turbulence assumption. A numerical mesh consisting of about 25000 control volumes (CV) and a time step of $\Delta t = 0.04\,D/u_\infty$ was used for the computations. In every iteration of the optimisation, which was performed with the steepest descent method, the primal solution was integrated over 15000 time steps. For the calculation of the objective function and the gradient, the first 5000 time steps were neglected to remove the initial transient. The optimisation was terminated when all sensitivities had dropped by two orders of magnitude.

As can be seen from fig. 2(a), the drag coefficient of the cylinder decreases from $c_d = 1.336$ to $c_d = 0.899$ when actuated with the optimal control parameters, which is a reduction of more than 30%. The comparison of the sensitivities at the first optimisation step shows a good agreement of the adjoint-based gradient with Finite Differences, see fig. 2(b). There are only small deviations, which can be attributed to the insufficient grid spacing and time step. Note, that only the slits on the upper half of the cylinder are shown, as the optimisation leads to



(a) Drag coefficient over optimisation step (b) Comparison of gradient at step 1 (c) Final amplitude distribution

Fig. 2. Optimisation results for the cylinder flow

a symmetric actuation. The same holds for the optimal amplitude distribution, which is presented in fig. 2(c). The blowing and suction at slits one to four generates a symmetric vortex pair which is pushed away from the rear of the cylinder by the blowing at slits five to eight. As is obvious in fig. 1(b), this increases the pressure level behind the cylinder, thus reducing the pressure drag.

5.2 NACA0015 Airfoil with Synthetic Jet Actuation

The second application is the lift maximisation for the unsteady turbulent flow around a NACA0015 airfoil at $Re = 10^6$, based on the cord length c and the freestream velocity u_∞. The angle of attack (AoA) is $\alpha = 20^o$, leading to a massive separation on the suction side. In this case, sinusoidal blowing and suction (also called synthetic jet), which can be modelled according to

$$c_i = u_a r_i \sin\left[2\pi f\,(t - t_0)\right], \quad r_i = \begin{pmatrix} \cos(\beta - \theta) \\ \sin(\beta - \theta) \end{pmatrix}, \tag{17}$$

is applied at four slits on the suction side of the airfoil with a constant frequency of $f = 1.28\,u_\infty/c$. Compared to the pulsed actuation (eq. 16) the blowing angle β can now be varied. The angle of the slit surface, θ, is fixed by the geometry of the airfoil. Computations with the discrete and the continuous adjoint solver were performed on a coarse mesh with 9500 CV and $\Delta t = 0.005\,c/u_\infty$ over 100 time steps, including the initial transient.

The comparison of the sensitivity gradients, summarised in tab. 1, reveals an excellent agreement between the forward and reverse mode AD, giving only very small differences of approx. 1×10^{-5}.

Compared to the AD-based solver, the results of the continuous adjoint code are significantly less accurate. One reason for this is the insufficient grid spacing, which is known to cause consistency problems with the continuous adjoint approach [11]. Furthermore, this can also be attributed to the frozen turbulence assumption. The active flow control modifies the separation on the suction side of the airfoil considerably, which has a strong impact on the turbulence field. This is completely neglected by the frozen turbulence assumption.

Table 1. Comparison of the sensitivities for the NACA0015 testcase

control parameter	forward mode AD	reverse mode AD	continuous adjoint
amplitude slit 1	0.132843488475446	0.132869414677547	0.070114751633607
amplitude slit 2	0.167065662623720	0.167070460770784	0.091718158272631
amplitude slit 3	0.181252126166289	0.181247271988999	0.103029635268416
amplitude slit 4	0.155843813170431	0.155844489164031	0.078944639252318
angle slit 1	0.005209720130677	0.005212791392634	0.000849278587241
angle slit 2	0.006705398122871	0.006697219503597	0.000654244527370
angle slit 3	0.006841527973356	0.006841492789784	0.002024334722777
angle slit 4	0.007246750396135	0.007246751418587	0.001287773996372
phase slit 1	0.204178681978258	0.204246051913213	0.278962118275295
phase slit 2	0.244693324906123	0.244791572866917	0.295707942189874
phase slit 3	0.244819168327026	0.244817849004966	0.304905513643976
phase slit 4	0.125955476539906	0.125957535080716	0.150374244523809

6 Summary and Outlook

In this paper, the development of a continuous and a discrete adjoint flow solver for the optimal control of unsteady, turbulent flows governed by the incompressible URANS equations was presented. For the continuous adjoint approach, the wide-spread frozen turbulence assumption was used, while the AD-based discrete approach is fully consistent independent of the grid size, time step and turbulence model, as it does not require analytical differentiability. The numerical efficiency of the discrete solver has been improved by employing the reverse accumulation technique and the binomial checkpointing, which allows the application of the discrete adjoint solver to practical configurations.

The numerical results of the drag reduction of the cylinder flow showed, that the continuous adjoint method works well for unsteady laminar flows. However, it gives fairly inaccurate sensitivity gradients when applied to the turbulent flow around a NACA0015 airfoil at a high Re-number due to the frozen turbulence assumption and insufficient grid spacing. In contrast to this, the sensitivities obtained from the AD-based adjoint solver are of excellent accuracy and match the forward mode AD nearly perfectly.

In future studies, the different approaches will be applied to more complex geometries such as multi-element high-lift configurations or simplified car models, aiming at a more detailed comparison of the adjoint methods in terms of accuracy and numerical efficiency.

Acknowledgments. This research was partly funded by the German Science Foundation (DFG) within the scope of the project *Instationäre Optimale Strömungskontrolle aerodynamischer Konfigurationen.*

References

1. Carnarius, A., Thiele, F., Özkaya, E., Gauger, N.R.: Adjoint approaches for optimal flow control. AIAA Paper 2010-5088 (2010)
2. Christianson, B.: Reverse accumulation and attractive fixed points. Optimization Methods and Software 3, 311–326 (1994)
3. Elliot, J., Peraire, J.: Practical 3D aerodynamic design and optimization using unstructured meshes. AIAA Journal 35(9), 1479–1485 (1997)
4. Gad-el-Hak, M.: The Taming of the Shrew: Why Is It so Difficult to Control Turbulence. In: King, R. (ed.) Active Flow Control. NNFM, vol. 95, pp. 1–24. Springer, Heidelberg (2007)
5. Gauger, N.R., Walther, A., Moldenhauer, C., Widhalm, M.: Automatic Differentiation of an Entire Design Chain for Aerodynamic Shape Optimization. In: Tropea, C., Jakirlic, S., Heinemann, H.-J., Henke, R., Hönlinger, H. (eds.) New Results in Numerical and Experimental Fluid Mechanics VI. NNFM, vol. 96, pp. 454–461. Springer, Heidelberg (2007)
6. Giles, M.B., Ghate, D., Duta, M.C.: Algorithm Developments for Discrete Adjoint Methods. AIAA Journal 41(2), 198–205 (2003)
7. Griewank, A., Walther, A.: Algorithm 799: Revolve: An implementation of checkpointing for the reverse or adjoint mode of computational differentiation. ACM Trans. Math. Software 26(1), 19–45 (2000)
8. Griewank, A., Walther, A.: Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation, 2nd edn. SIAM, Other Titles in Applied Mathematics, Philadelphia (2008)
9. Hascoët, L., Pascual, V.: TAPENADE 2.1 user's guide. Technical Report, INRIA (2004)
10. Jameson, A.: Aerodynamic Design via Control Theory. Journal of Scientific Computing 3, 233–260 (1988)
11. Nemili, A., Özkaya, E., Gauger, N.R., Carnarius, A., Thiele, F.: Optimal Control of Unsteady Flows Using Discrete Adjoints. AIAA-Paper 2011-3720 (2011)
12. Nielsen, E., Anderson, W.K.: Aerodynamic design optimization on unstructured meshes using the Navier-Stokes equations. AIAA Journal 37(11), 957–964 (1999)
13. Squire, W., Trapp, G.: Using complex variables to estimate derivatives of real functions. SIAM Rev. 10(1), 110–112 (1998)
14. Stumm, P., Walther, A.: Multistage approaches for optimal offline checkpointing. SIAM J. Scientific Computing 31(3), 1946–1967 (2009)
15. Wilcox, D.C.: Reassesment of the scale determing equation for advanced turbulence models. AIAA Journal 26(11) (1988)
16. Xue, L.: Entwicklung eines effizienten parallelen Lösungsalgorithmus zur dreidimensionalen Simulation komplexer turbulenter Strömungen. PHD thesis, Institut für Strömungmechanik und Technische Akustik, Technische Universität Berlin (1998)

Well-Posedness and Long Time Behavior for a Class of Fluid-Plate Interaction Models

Igor Chueshov and Iryna Ryzhkova

Department of Mathematics and Mechanics,
Kharkov National University,
4 Svobody square, Kharkov, 61077, Ukraine
{chueshov,ryzhkova}@univer.kharkov.ua
http://mathphys.univer.kharkov.ua

Abstract. We deal with well-posedness and asymptotic dynamics of a class of coupled systems consisting of linearized 3D Navier–Stokes equations in a bounded domain and a classical (nonlinear) elastic plate/shell equation. We consider three models for plate/shell oscillations: (a) the model which accounts for transversal displacement of a flexible flat part of the boundary only, (b) the model for in-plane motions of a flexible flat part of the boundary, (c) the model which accounts for both transversal and longitudinal displacements. For all three cases we present well-posedness results and prove existence of a compact global attractor. In the first two cases the attractor is of finite dimension and possesses additional smoothness. We do not assume any kind of mechanical damping in the plate component in the case of models (a) and (b). Thus our results means that dissipation of the energy in the fluid due to viscosity is sufficient to stabilize the system in the latter cases.

Keywords: Fluid–structure interaction, nonlinear shell/plate, linearized 3D Navier–Stokes equations, global attractor, finite dimension.

1 Introduction

We consider a coupled (hybrid) system, which describes interaction of a homogeneous viscous incompressible fluid that occupies a bounded domain $\mathcal{O}$ with elastic plate/shell. Boundary $\partial\mathcal{O}$ of $\mathcal{O}$ consists of the (solid) walls of the container S and a horizontal (flat) part Ω, on which a thin (nonlinear) elastic shell or plate is placed. The motion of the fluid is described by linearized 3D Navier–Stokes equations. To describe deformations of the shell/plate we use several elastic models. In all cases we assume that large deflections of the shell produce small effect on the fluid. This corresponds to the case when the fluid fills the container which is large in comparison with the size of the shell/plate.

We note that the mathematical studies of the problem of interaction of viscous fluids and elastic plates/bodies have a long history. We refer to [3,12,14] for the case of plates/membranes, to [10] in the case of moving elastic bodies, and to [1,2,11] in the case of elastic bodies with fixed interface; see also the literature cited in these papers.

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 328–337, 2013.

2 Mathematical Description of the Models

Fluid. Let $\mathcal{O} \subset \mathbb{R}^3$ be a bounded domain with sufficiently smooth boundary $\partial\mathcal{O}$. We assume that $\partial\mathcal{O} = \Omega \cup S$, where $\Omega \subset \{x = (x_1; x_2; 0) : x' \equiv (x_1; x_2) \in \mathbb{R}^2\}$ with a smooth contour $\Gamma = \partial\Omega$ and S is a surface lying in $\mathbb{R}^3_- = \{x_3 \leq 0\}$. The exterior normal on $\partial\mathcal{O}$ is denoted by n. We have that $n = (0; 0; 1)$ on Ω. The surface S corresponds to solid walls of the container with a fluid, and Ω models an elastic shell or plate placed over the fluid.

To describe the fluid we consider the following *linear* Navier–Stokes equations in $\mathcal{O}$ for the fluid velocity field $v = v(x,t) = (v^1(x,t); v^2(x,t); v^3(x,t))$ and the pressure $p(x,t)$:

$$v_t - \nu\Delta v + \nabla p = G_f \quad \text{in} \quad \mathcal{O} \times (0, +\infty)\,, \tag{1}$$

$$\mathrm{div} v = 0 \quad \text{in} \quad \mathcal{O} \times (0, +\infty)\,, \tag{2}$$

where $\nu > 0$ is the dynamical viscosity and G_f is a volume force.

We denote by $T_f(v)$ the surface force exerted by the fluid on the shell which is equal to $Tn|_\Omega$, where n is an outer unit normal to $\partial\mathcal{O}$ at Ω and $T = \{T_{ij}\}_{i,j=1}^3$ is the stress tensor of the fluid,

$$T_{ij} \equiv T_{ij}(v) = \nu\left(v^i_{x_j} + v^j_{x_i}\right) - p\delta_{ij}, \quad i,j = 1,2,3\,.$$

Since $n = (0; 0; 1)$ on Ω, we have that

$$T_f(v) = (\nu(v^1_{x_3} + v^3_{x_1});\ \nu(v^2_{x_3} + v^3_{x_2}); 2\nu v^3_{x_3} - p)\,.$$

A specific form of this force as well as boundary conditions on Ω for the fluid depend on elastic plate/shell model we choose.

General Model (GM). We start with the full von Karman shallow shell model which accounts for both transversal and in-plane displacements (see [20,15,13,17] and the references therein). To this end we equip (1) and (2) with the (non-slip) boundary conditions imposed on the velocity field $v = v(x,t)$:

$$v = 0 \ \text{ on } S, \quad v \equiv (v^1; v^2; v^3) = (u^1_t; u^2_t; w_t) \ \text{ on } \Omega\,, \tag{3}$$

where $u = u(x,t) \equiv (u^1; u^2; w)(x,t)$ is the displacement of the shell occupying Ω. Here w stands for the transversal displacement, $\bar{u} = (u^1; u^2)$ — for the lateral (in-plane) displacements.

To describe the shell motion we use the full von Karman model which takes into account rotational inertia of the filaments and possible presence of in-plane acceleration terms (see the literature cited above):

$$M_\alpha(w_{tt} + \gamma w_t) + \Delta^2 w + \mathrm{trace}\,\{K\mathcal{N}(u)\} - \mathrm{div}\,\{\mathcal{N}(u)\nabla w\} = G_3 - 2\nu v^3_{x_3} + p\,, \tag{4}$$

and

$$\varrho\bar{u}_{tt} = \mathrm{div}\,\{\mathcal{N}(u)\} + \left(G_1 - \nu(v^1_{x_3} + v^3_{x_1}); G_2 - \nu(v^2_{x_3} + v^3_{x_2})\right)\,, \tag{5}$$

where $M_\alpha = 1 - \alpha\Delta$, $K = \operatorname{diag}(k_1, k_2)$, and

$$\mathcal{N}(u) \equiv \begin{pmatrix} N_{11} & N_{12} \\ N_{12} & N_{22} \end{pmatrix} = \mathcal{C}(\epsilon_0(\bar{u}) + wK + f(\nabla w))$$

with $\bar{u} = (u_1; u_2)$, $\mathcal{C}(\epsilon) = D\,[\mu \operatorname{trace}\epsilon \cdot I + (1-\mu)\epsilon]$, and

$$\epsilon_0(\bar{u}) = \frac{1}{2}(\nabla\bar{u} + \nabla^T\bar{u}), \quad f(s) = \frac{1}{2}s \otimes s,\ s \in \mathbb{R}^2\ .$$

Here $D = Eh/(1-\mu^2)$, E is Young's modulus, $0 < \mu < 1/2$ is Poisson's ratio, h is the thickness of the shell, $\alpha > 0$ and $\varrho \geq 0$ are constants taking into account rotational inertia and in-plane inertia of the shell, $\gamma \geq 0$ is a parameter which describes intensity of the viscous damping of the shell material. We denote by $G_{sh} \equiv (G_1; G_2; G_3)$ a (given) body force applied to the shell.

We impose the clamped boundary conditions on the shell:

$$u^1|_{\partial\Omega} = u^2|_{\partial\Omega} = 0 \tag{6}$$

and

$$w|_{\partial\Omega} = \left.\frac{\partial w}{\partial n}\right|_{\partial\Omega} = 0\ . \tag{7}$$

We supply (1)–(7) with the initial data for the velocity field $v = (v^1; v^2; v^3)$ and the shell displacement vector $u = (u^1; u^2; w)$ of the form[1]

$$v\big|_{t=0} = v_0, \quad u\big|_{t=0} = u_0, \quad w_t\big|_{t=0} = w_1, \quad \varrho\left[\bar{u}_t\big|_{t=0} - \bar{u}_1\right] = 0\ , \tag{8}$$

where $\bar{u} = (u^1; u^2)$. Here $v_0 = (v_0^1; v_0^2; v_0^3)$, $u_0 = (u_0^1; u_0^2; w_0)$, w_1, and $\bar{u}_1 = (u_1^1; u_1^2)$ are given vector functions which we will specify later.

From (2) and (3) we can also derive the compatibility condition

$$\int_\Omega w(x', t)dx' = const \quad \text{for all} \quad t \geq 0\ , \tag{9}$$

which can be interpreted as preservation of the volume of the fluid.

Simplified Model 1 (SM1). This kind of models arises in the study of the problem of blood flows in large arteries (see, e.g., [12] and the references therein). The model assumes additional hypothesis that the transversal displacement w of the plate is negligible relatively to the in-plane displacement $(u^1; u^2)$. Thus we consider only longitudinal deformations of the plate and take into account only tangential shear forces which fluid exerts on the plate. Formally this means that we omit equation (4) and put $w \equiv 0$ in (3) and (5). Thus we arrive at the following boundary conditions imposed on the velocity field $v = v(x, t)$:

$$v = 0 \ \text{ on } S, \quad v \equiv (v^1; v^2; v^3) = (u_t; 0) \equiv (u_t^1; u_t^2; 0) \ \text{ on } \Omega\ , \tag{10}$$

[1] We put the multiplier ϱ in the fourth relation of (8) to emphasize that this relation is not needed in the case of negligibly small in-plane inertia ($\varrho = 0$).

where $u = u(x,t) \equiv (u^1(x,t); u^2(x,t))$ is the in-plane displacement vector of the plate placed on Ω satisfying (5) with $\mathcal{N}_0(u) = \mathcal{C}(\epsilon_0(u))$ instead of $\mathcal{N}$.

We assume that for this case the external (in-plane) force $(G^1; G^2)$ in (5) is a nonlinear feedback force represented by a potential Φ:

$$G^i = f^i(u^1, u^2) \equiv \frac{\partial \Phi(u^1, u^2)}{\partial u^i}, \quad i = 1, 2 \, .$$

Since $v^3(x_1; x_2; 0) = 0$ for $(x_1; x_2) \in \Omega$ due to the second relation in (10), we have $v^3_{x_i} = 0$ on Ω, $i = 1, 2$. Thus after rescaling of the elastic constants we arrive at the following equations for the in-plane displacement $u = (u^1; u^2)$:

$$u^i_{tt} - \Delta u^i - \lambda \partial_{x_i} [\operatorname{div} u] + \nu v^i_{x_3}|_{x_3=0} + f^i(u) = 0, \quad i = 1, 2 \, , \tag{11}$$

where λ is a nonnegative parameter. We impose the clamped boundary conditions (6) for the displacement $u = (u^1; u^2)$ on $\Gamma = \partial\Omega$. Thus we obtain

Problem (SM1): Find vector functions $v = (v^1; v^2; v^3)$ and $u = (u^1; u^2)$ satisfying (in some sense) equations (1), (2), (10), (11), (6) and the initial data

$$v|_{t=0} = v_0, \quad u|_{t=0} = u_0, \quad u_t|_{t=0} = u_1 \, .$$

This problem with $\lambda = 0$ and $f^i(u) \equiv 0$ was considered in [12] (see also the literature cited there) with the additional strong (Kelvin-Voight type) damping force applied to the interior of the plate. In contrast with [12] we do not assume the presence of mechanical damping terms in the plate component of the system and consider a nonlinearly forced model.

Simplified Model 2 (SM2). This model is concerned to dynamics of the transversal displacement w. The corresponding model assumes a special structure of the in-plane displacements $\bar{u} = \bar{u}(x,t) \equiv (u^1(x,t); u^2(x,t))$ in (4) as a function of the transversal displacement w only. Hence we neglect the equation in (5) (see, e.g., [15,20] and the references therein). We also assume that $\alpha = 0$. This formal procedure leads to the following boundary conditions imposed on the velocity field $v = v(x,t)$:

$$v = 0 \ \text{ on } S; \quad v \equiv (v^1; v^2; v^3) = (0; 0; w_t) \ \text{ on } \Omega \, . \tag{12}$$

This and also (2) imply that $v^3_{x_3} = 0$ and therefore the third (transversal) component $T_f(v)$ on $\partial\Omega$ is exactly the pressure p of the fluid. Thus the transversal displacement $w = w(x,t)$ of the plate satisfies the following equation:

$$w_{tt} + \Delta^2 w + \mathcal{F}(w) = G_{pl} + p|_\Omega \ \text{ in } \ \Omega \times (0, \infty) \, , \tag{13}$$

where G_{pl} is a given body force on the plate, $\mathcal{F}(u)$ is a nonlinear feedback force which will be specified later. As a result we obtain

Problem (SM2): Find the fluid velocity field $v = (v^1; v^2; v^3)$, the pressure p, and the transversal displacement of the plate w satisfying (in some sense) equations (1), (2), (12), (13) and also compatibility condition (9), boundary conditions (7), and initial conditions of the form

$$v(0) = v_0, \quad w(0) = w_0, \quad w_t(0) = w_1 \, .$$

Remark 1. We emphasize that even in the linear case due to the structure of the surface fluid forces $T_f(v)$ we cannot split system (1)–(8) into two sets of equations describing longitudinal and transversal plate movements separately, i.e., we cannot reduce the model in **(GM)** to the cases considered in the models in **(SM1)** and **(SM2)**. The point is that in the case **(SM1)** equation (11) for longitudinal plate deformations does not contain the terms $v^3_{x_i}$ and the model does not require any compatibility conditions like (9) because the volume of the fluid obviously preserves. In the case of purely transversal displacements (see **(SM2)**) the force exerted on the plate by the fluid contains the pressure only. See also [9] for a further discussion.

Spaces and Notations. To describe fluid velocity fields we introduce the following spaces. Let $\mathcal{C}(\mathcal{O})$ be the class of C^∞ vector-valued solenoidal (i.e., divergence-free) functions on $\overline{\mathcal{O}}$ which vanish in a neighborhood of S. We denote by X the closure of $\mathcal{C}(\mathcal{O})$ with respect to the L_2-norm and by V the closure of $\mathcal{C}(\mathcal{O})$ with respect to the $H^1(\mathcal{O})$-norm. One can see that

$$X = \left\{v = (v^1; v^2; v^3) \in [L_2(\mathcal{O})]^3 \,:\, \operatorname{div} v = 0, \gamma_n v \equiv (v, n) = 0 \text{ on } S\right\}\,,$$
$$V = \left\{v = (v^1; v^2; v^3) \in [H^1(\mathcal{O})]^3 \,:\, \operatorname{div} v = 0,\ v = 0 \text{ on } S\right\}.$$

We refer to [19], for instance, for the details concerning spaces of this type.

To describe shell/plate displacements we use the Sobolev spaces $H^s(\Omega)$ and $H^s_0(\Omega)$. We also denote $\widehat{H}^s(\Omega) = H^s(\Omega) \cap \widehat{L}_2(\Omega)$ for $s \geq 0$, where $\widehat{L}_2(\Omega)$ is the subspace in $L_2(\Omega)$ consisting of functions with zero average over Ω.

For $D =$ either $\mathcal{O}$ or Ω we denote by $\|\cdot\|_D$ the norm in $L_2(D)$ and by $\|\cdot\|_{s,D}$ the norm in $H^s(D)$ and keep the corresponding notations for the inner products.

3 Results: Well-Posedness and Long-Time Dynamics

General Model. We deal with weak (variational) solutions to (1)-(9) and consider the cases $\varrho > 0$ and $\varrho = 0$ simultaneously. This is possible due to an additional regularity estimate for the shell velocities, which follows from (3) and from the standard trace theorem. Even in the case $\varrho = 0$ we have that

$$||w_t(t)||^2_{H^{1/2}(\Omega)} + ||u^1_t(t)||^2_{H^{1/2}(\Omega)} + ||u^2_t(t)||^2_{H^{1/2}(\Omega)} \leq C||\nabla v(t)||^2_{\mathcal{O}} \tag{14}$$

for every weak solution $(v(t); u(t))$. We use this observation to suggest unified way to prove a well-posedness result not distinguishing the cases $\varrho > 0$ and $\varrho = 0$ in contrast with [20] (see also [17]). We also note that in the case we neglect the inertia of longitudinal deformations ($\varrho = 0$) the equations in (5) become elliptic. However, we keep the initial data for the in-plane displacement $(u^1; u^2)$. The point is that the first order evolution for $(u^1; u^2)$ goes from the boundary condition for the fluid velocity in (3).

As a phase space we use

$$\mathcal{H} = \begin{cases} \{(v_0; u_0; u_1) \in X \times W \times Y :\ v_0 = u_1 \text{ on } \Omega\}\,, & \varrho > 0\,, \\ \left\{(v_0; u_0; w_1) \in X \times W \times \widehat{H}^1_0(\Omega) :\ (v_0)^3 = w_1 \text{ on } \Omega\right\}, & \varrho = 0\,, \end{cases}$$

where w_1 is the third component of the initial displacement velocity u_1 and

$$W = H_0^1(\Omega) \times H_0^1(\Omega) \times \widehat{H}_0^2(\Omega), \quad Y = L_2(\Omega) \times L_2(\Omega) \times \widehat{H}_0^1(\Omega) \, .$$

Our main result concerning **(GM)** is the following well-posedness theorem.

Theorem 1. *Assume that $\alpha > 0$, $\gamma \geq 0$, $\varrho \geq 0$, and*

$$U_0 \in \mathcal{H}, \quad G_f \in V', \quad G_{sh} \in \left[H^{-1/2}(\Omega)\right]^2 \times H^{-1}(\Omega)$$

(in the case $\varrho = 0$ the data $\bar{u}_1$ are not fixed). Then for any interval $[0, T]$ there exists a unique weak solution $(v(t); u(t))$ to (1)–(9) with the initial data U_0. This solution satisfies an energy balance equality and generates a continuous (both in strong and weak sense) evolution semigroup S_t in the space $\mathcal{H}$. The evolution operator S_t is defined as follows

- **case $\varrho > 0$:** $S_t(v_0; u_0; u_1) \equiv U(t) = (v(t); u(t); u_t(t))$, *where the couple $(v(t); u(t))$ solves (1)–(9);*
- **case $\varrho = 0$:** $S_t(v_0; u_0; w_1) \equiv \bar{U}(t) = (v(t); u(t); w_t(t))$, *where $v(t)$ and $u(t) = (u^1(t); u^2(t); w(t))$ solves (1)–(9) with $\varrho = 0$.*

Proof. We use the compactness method with Galerkin's approximations, which was inspired by the method developed in [3] for the case of a linear plate interacting with nonlinear Navier-Stokes equations. Uniqueness relies on the same idea as in [17] and involves Brésis–Gallouet type inequality. We follow the scheme of [13] in the proof of continuity properties of the solution and the energy equality. The details of the proof can be found in [9].

Remark 2. In the case $\alpha = 0$ we can prove existence of weak solutions which satisfy an energy inequality using the same type of argument. Uniqueness of the solutions is still an open question. Sedenko's method does not work here because the nonlinearity is strongly supercritical when $\alpha = 0$.

Our next result deals with global attractors. We recall (see, e.g., [5,18]) that *global attractor* of the dynamical system $(S_t, \mathcal{H})$ is defined as a bounded closed set $\mathfrak{A} \subset \mathcal{H}$ which is invariant ($S_t \mathfrak{A} = \mathfrak{A}$ for all $t > 0$) and uniformly attracts all other bounded sets:

$$\lim_{t \to \infty} \sup\{\mathrm{dist}_{\mathcal{H}}(S_t y, \mathfrak{A}) : \ y \in B\} = 0 \quad \text{for any bounded set } B \text{ in } \mathcal{H}.$$

Theorem 2. *Assume that $\alpha > 0$, $\gamma > 0$, and the external forces satisfy*

$$G_f \equiv 0, \quad G_{sh}^1 = G_{sh}^2 \equiv 0, \quad \text{and} \quad G_{sh}^3 \equiv g \in H^{-1}(\Omega) \, .$$

Let the set of the stationary points in $\mathcal{H}$ of the problem (1)–(9) is bounded. Then the corresponding evolution semigroup S_t possesses a compact global attractor.

To prove this theorem we apply J.Ball's method (in the form presented in [16]). To this end we need the property $\gamma > 0$, i.e., assume a presence of mechanical damping in the transversal component of displacement. The question whether the system under consideration demonstrates compact long-time behavior without mechanical damping in the shell component is still open. The main obstacle in this case is that the dissipation of the energy in the fluid leads to the mechanical damping of order $1/2$ in the shell components (see (14)). This is not enough to stabilize kinetic energy of the plate, which is of the first order, (at least, uniformly). The same effect is valid for the model **(SM2)** with rotational inertia (see Remark 3 below). However, in the case of rotational inertia neglected ($\alpha = 0$) the matter differs. One can compare Remark 3 with the results for models **(SM1)** and **(SM2)**, in which rotational inertia is not accounted for. We *do not require* any mechanical damping in these models and compact asymptotic dynamics of the corresponding systems is guaranteed by viscous dissipation in the fluid. Moreover, in these cases we can establish finite-dimensionality of the the corresponding attractors and also their smoothness. See [4,8] for details.

Simplified Model 1. We assume that the plate force potential $\Phi(u) \in C^2(\mathbb{R}^2)$ is a nonnegative polynomially bounded function,

$$\left| \frac{\partial \Phi(u)}{\partial u^i \partial u^j} \right| \le C\,(1 + |u|^p)\,, \quad i,j = 1,2, \quad u = (u^1; u^2) \in \mathbb{R}^2\,,$$

and the following dissipativity condition holds: for any $\delta > 0$ there exist $c_1(\delta) > 0$ and $c_2(\delta) \ge 0$ such that

$$\sum_{i=1,2} u^i f^i(u) - c_1(\delta)\Phi(u) + \delta|u|^2 \ge -c_2(\delta) \quad \text{with} \quad f^i(u) = \frac{\partial \Phi(u)}{\partial u^i}\,. \tag{15}$$

We can consider as examples

$$\Phi(u) = \psi_0(|u^1|^2 + |u^2|^2) \quad \text{or} \quad \Phi(u) = \psi_1(u^1) + \psi_2(u^2),$$

where $\psi_i(s)$ are nonnegative polynomials.

We use the following phase space for this model:

$$\mathcal{H} = \{v \in X : \ (v,n) = 0 \text{ on } \Omega\} \times \left[H_0^1(\Omega)\right]^2 \times \left[L_2(\Omega)\right]^2\,.$$

Theorem 3 (Well-Posedness). *Let $U_0 \in \mathcal{H}$. Then for any interval $[0,T]$ there exists a unique weak solution $(v(t); u(t))$ to problem* **(SM1)** *for which an energy balance equality holds. Moreover, this solution defines a continuous evolution operator $S_t : \mathcal{H} \mapsto \mathcal{H}$ by the formula*

$$S_t(v_0; u_0; u_1) = (v(t); u(t); u_t(t))\,,$$

where the couple $(v(t); u(t))$ solves problem **(SM1)**, *and there exists a constant $a_{R,T} > 0$ such that for any couple of initial data possessing the property $\|U\|_{\mathcal{H}}, \|\hat{U}\|_{\mathcal{H}} \le R$ we have*

$$\|S_t U - S_t \hat{U}\|_{\mathcal{H}}^2 + \int_0^t \|\nabla(v - \hat{v})\|_{\mathcal{O}}^2 d\tau \le a_{R,T}\|U - \hat{U}\|_{\mathcal{H}}^2, \quad \forall\, t \in [0,T]\,,$$

where $S_tU = (v(t); u(t); u_t(t))$ *and* $S_t\hat{U} = (\hat{v}(t); \hat{u}(t); \hat{u}_t(t))$. *In the linear case* $\Phi(u) \equiv 0$ *problem* **(SM1)** *generates an exponentially stable* C_0*-semigroup of contractions* $e^{-t\mathcal{A}}$ *in* $\mathcal{H}$*: there exist* $C, \alpha > 0$ *such that*

$$\|e^{-t\mathcal{A}}U\|_{L(\mathcal{H},\mathcal{H})} \leq Ce^{-\alpha t} \quad \textit{for all } t > 0\,. \tag{16}$$

We refer to [4] for the details . We note that property (16) improves the result in [12] which states the strong stability only.

The following result shows that model **(SM1)** demonstrates finite dimensional long-time dynamics.

Theorem 4 (Global Attractor). *The dynamical system* $(\mathcal{H}, S_t)$ *generated by* **(SM1)** *possesses a compact global attractor* $\mathfrak{A}$ *of finite fractal dimension*[2]. *If relation* (15) *holds with* $c_2(\delta) \equiv 0$, *then the global attractor* $\mathfrak{A}$ *consists of a single point,* $\mathfrak{A} = \{(0; 0; 0)\}$, *which is exponentially attractive, i.e. there exist* $c_R > 0$ *and* $\alpha > 0$ *such that*

$$\|S_tU\|_{\mathcal{H}} \leq c_R e^{-\alpha t} \quad \textit{for any } U \in \mathcal{H} \textit{ such that } \|U\|_{\mathcal{H}} \leq R\,.$$

It is well-known (see, e.g., [18]), that to prove the existence of a compact global attractor it is sufficient to show that the system is dissipative and asymptotically smooth. To prove dissipativity we use an appropriate Lyapunov function. As for asymptotic smoothness of the system and finite-dimensionality of the attractor, we rely on recently developed approach based on stabilizability estimates (see [6,7] and the references therein). We refer to [4] for further details.

Simplified Model 2. We impose the following hypotheses concerning the nonlinear feedback force $\mathcal{F}(u)$ in the plate equation (13).

(F1) $\mathcal{F}(u)$ is locally Lipschitz from $H_0^{2-\epsilon}(\Omega)$ into $H^{-1/2}(\Omega)$ for some $\epsilon > 0$.
(F2) There exists a C^1-functional $\Pi(u)$ on $H_0^2(\Omega)$ such that $\mathcal{F}(u) = \Pi'(u)$.
(F3) The plate force potential Π is bounded on bounded subsets of $H_0^2(\Omega)$, and there exist $\eta < 1/2$ and $c \geq 0$ such that

$$\eta\|\Delta w\|_\Omega^2 + \Pi(w) \geq -c, \quad \eta\|\Delta w\|_\Omega^2 + (w, \mathcal{F}(w))_\Omega \geq -c \quad \forall\, w \in H_0^2(\Omega)\,.$$

We can consider Kirchhoff, von Karman, or Berger nonlinearities as examples of nonlinear feedback (elastic) force $\mathcal{F}(u)$, see [8] for the details.

We use

$$\mathcal{H} = \left\{(v_0; w_0; w_1) \in X \times \widehat{H}_0^2(\Omega) \times \widehat{L}_2(\Omega) :\ v_0^3 = w_1 \text{ on } \Omega\right\}$$

as a phase space and deal with weak (variational) solutions.

Theorem 5 (Well-Posedness). *Assume that* $U_0 = (v_0; w_0; w_1) \in \mathcal{H}$, $G_f \in V'$, *and* $G_{pl} \in H^{-1/2}(\Omega)$. *Then for any interval* $[0, T]$ *there exists a unique weak*

[2] For the definition and basic properties of the fractal dimension see, e.g., [18].

solution $(v(t); w(t))$ *to problem* **(SM2)** *with the initial data* U_0. *The solution possesses the property*

$$U(t) \equiv (v(t); w(t); w_t(t)) \in C(0, T; \mathcal{H})$$

and satisfies the energy balance equality

$$\mathcal{E}(v(t), w(t), w_t(t)) + \nu \int_0^t ||\nabla v||^2_{\mathcal{O}} d\tau = \mathcal{E}(v_0, w_0, w_1) + \int_0^t (G_f, v)_{\mathcal{O}} d\tau$$

for every $t > 0$, *where the energy functional* $\mathcal{E}$ *is defined by the relation*

$$\mathcal{E}(v, w, w_t) = \frac{1}{2}\left(\|v\|^2_{\mathcal{O}} + \|w_t\|^2_{\Omega} + \|\Delta w\|^2_{\Omega}\right) + \int_{\Omega} \Pi(w(x))dx - (G_{pl}, w)_{\Omega}\,.$$

Moreover, there exists a constant $a_{R,T} > 0$ *such that for any couple of weak solutions* $U(t) = (v(t); w(t); w_t(t))$ *and* $\hat{U}(t) = (\hat{v}(t); \hat{w}(t); \hat{w}_t(t))$ *with the initial data possessing the property* $\|U_0\|_{\mathcal{H}}, \|\hat{U}_0\|_{\mathcal{H}} \le R$ *we have*

$$\|U(t) - \hat{U}(t)\|^2_{\mathcal{H}} + \int_0^t \|\nabla(v - \hat{v})\|^2_{\mathcal{O}} d\tau \le a_{R,T}\|U_0 - \hat{U}_0\|^2_{\mathcal{H}}, \quad t \in [0, T]\,. \tag{17}$$

In the case $G_f \equiv 0$, $G_{pl} \equiv 0$, $\mathcal{F}(u) \equiv 0$ *the problem generates a strongly continuous exponentially stable contraction semigroup* T_t *on* $\mathcal{H}$.

Using the same approach as in the proof of Theorem 4 we can establish the following result (see [8] for details).

Theorem 6 (Global Attractor). *The dynamical system* $(S_t, \mathcal{H})$ *generated by* **(SM2)** *possesses a compact global attractor* $\mathfrak{A}$. *Moreover,*

(1) $\mathfrak{A}$ *is the unstable set emanating from the set of equilibria* $\mathcal{N}$, $\mathfrak{A} = \mathbb{M}_+(\mathcal{N})$;
(2) *the attractor has finite fractal dimension;*
(3) *any trajectory* $\gamma = \{(v(t); w(t); w_t(t)) : t \in \mathbb{R}\}$ *from the attractor* $\mathfrak{A}$ *possesses the property* $(v_t; w_t; w_{tt}) \in L_\infty(\mathbb{R}; X \times \widehat{H}^2_0(\Omega) \times \widehat{L}_2(\Omega))$, *and there is* $R > 0$ *such that* $\left(\|v_t\|^2_{\mathcal{O}} + \|w_t\|^2_{2,\Omega} + \|w_{tt}\|^2_{\Omega}\right) \le R^2$ *for all* $t \in \mathbb{R}$ *and* $\gamma \subset \mathfrak{A}$.

We recall (see, e.g., [18]) that the *unstable set* $\mathbb{M}_+(\mathcal{N})$ emanating from some set $\mathcal{N} \subset \mathcal{H}$ is a subset of $\mathcal{H}$ such that for each $z \in \mathbb{M}_+(\mathcal{N})$ there exists a full trajectory $\{y(t) : t \in \mathbb{R}\}$ satisfying $y(0) = z$ and $\operatorname{dist}(y(t), \mathcal{N}) \to 0$ as $t \to -\infty$.

Remark 3. We can consider model **(SM2)** with the rotational inertia accounted for (i. e., with the additional inertial term $-\alpha\Delta w_{tt}$ in the plate equation). In this case the phase space is $\{(v_0; w_0; w_1) \in \mathcal{H} : w_1 \in H^1_0(\Omega)\}$ and the corresponding analog of Theorem 5 remains true, except the property of exponential stability of the linear semigroup T_t (the case $G_f \equiv 0$, $G_{pl} \equiv 0$, $\mathcal{F}(u) \equiv 0$). To prove the existence of the attractor for **(SM2)** with the rotational inertia as in the case of **(GM)**, we need to assume presence of rotational mechanical damping in the plate equation (see Theorem 2 and the comments after its statement). Whether model **(SM2)** with rotational inertia term and without mechanical dissipation demonstrates a compact long-time dynamics is still an open question.

References

1. Avalos, G., Triggiani, R.: Semigroup well-posedness in the energy space of a parabolic hyperbolic coupled Stokes–Lamé PDE system of fluid-structure interaction. Discr. Contin. Dyn. Sys. Ser.S 2, 417–447 (2009)
2. Barbu, V., Grujić, Z., Lasiecka, I., Tuffaha, A.: Smoothness of weak solutions to a nonlinear fluid–structure interaction model. Indiana Univ. Math. J. 57, 1173–1207 (2008)
3. Chambolle, A., Desjardins, B., Esteban, M., Grandmont, C.: Existence of weak solutions for the unsteady interaction of a viscous fluid with an elastic plate. J. Math. Fluid Mech. 7, 368–404 (2005)
4. Chueshov, I.: A global attractor for a fluid-plate interaction model accounting only for longitudinal deformations of the plate. Math. Meth. Appl. Sci. 34, 1801–1812 (2011)
5. Chueshov, I.: Introduction to the Theory of Infinite-Dimensional Dissipative Systems. Acta, Kharkov (1999) (in Russian); English translation: Acta, Kharkov (2002); http://www.emis.de/monographs/Chueshov/
6. Chueshov, I., Lasiecka, I.: Long-Time Behavior of Second Order Evolution Equations with Nonlinear Damping. In: Chueshov, I., Lasiecka, I. (eds.) Memoirs of AMS, vol. 195(912). AMS, Providence (2008)
7. Chueshov, I., Lasiecka, I.: Von Karman Evolution Equations. Springer, New York (2010)
8. Chueshov, I., Ryzhkova, I.: A global attractor for a fluid-plate interaction model. Preprint ArXiv:1109.4324v1 (September 2011), to appear in Comm. Pure Appl. Anal.
9. Chueshov, I., Ryzhkova, I.: Unsteady interaction of a viscous fluid with an elastic shell modeled by full von Karman equations. Preprint ArXiv:1112.6094v1 (December 2011), To appear in J. Differ. Equat.
10. Coutand, D., Shkoller, S.: Motion of an elastic solid inside an incompressible viscous fluid. Arch. Ration. Mech. Anal. 176, 25–102 (2005)
11. Du, Q., Gunzburger, M.D., Hou, L.S., Lee, J.: Analysis of a linear fluid–structure interaction problem. Discrete Contin. Dyn. Syst. 9, 633–650 (2003)
12. Grobbelaar-Van Dalsen, M.: Strong stability for a fluid-structure model. Math. Methods Appl. Sci. 32, 1452–1466 (2009)
13. Koch, H., Lasiecka, I.: Hadamard well-posedness of weak solutions in nonlinear dynamic elasticity-full von Karman systems. In: Prog. Nonlinear Differ. Equ. Appl., vol. 50, pp. 197–216. Birkhäuser, Basel (2002)
14. Kopachevskii, N., Pashkova, Y.: Small oscillations of a viscous fluid in a vessel bounded by an elastic membrane. Russian J. Math. Phys. 5(4), 459–472 (1998)
15. Lagnese, J.: Boundary Stabilization of Thin Plates. SIAM, Philadelphia (1989)
16. Moise, I., Rosa, R., Wang, X.: Attractors for non-compact semigroups via energy equations. Nonlinearity 11, 1369–1393 (1998)
17. Sedenko, V.I.: On the uniqueness theorem for generalized solutions of initial-boundary problems for the Marguerre–Vlasov vibrations of shallow shells with clamped boundary conditions. Appl. Math. Optim. 39, 309–326 (1999)
18. Temam, R.: Infinite-Dimensional Dynamical Dystems in Mechanics and Physics. Springer, New York (1988)
19. Temam, R.: Navier-Stokes Equations: Theory and Numerical Analysis, 1984th edn. AMS Chelsea Publishing, Providence (2001)
20. Vorovich, I.I.: On some direct methods in nonlinear oscillations of shallow shells. Izvestiya AN SSSR, Matematika 21(6), 142–150 (1957) (in Russian)

On the Normal Semilinear Parabolic Equations Corresponding to 3D Navier-Stokes System

Andrei Fursikov*

Department of Mechanics & Mathematics, Moscow State University,
119991 Moscow, Russia

Abstract. The semilinear normal parabolic equations corresponding to 3D Navier-Stokes system have been derived. The explicit formula for solution of normal parabolic equations with periodic boundary conditions has been obtained. It was shown that phase space of corresponding dynamical system consists of the set of stability (where solutions tends to zero as time $t \to \infty$), the set of explosions (where solutions blow up during finite time) and intermediate set. Exact description of these sets has been given.

Keywords: Equations of normal type, Navier-Stokes system, structure of dynamical flow.

1 Introduction

As well known (see e.g. [1],[2]), existence of weak solution to 3D Navier-Stokes equations is proved with help of energy estimate, which is true because the image $B(v)$ of nonlinear operator from Navier-Stokes equations consists of vectors tangent to sphere in the L_2-space with the centrum in origin. If these vectors would be tangent to sphere in Soblev H^1-space, one could prove existence of strong solution to 3D Navier-Stokes system by the methods similar to ones used to prove existence of a weak solution. But this is not the matter: in this case $B(v) = B_\tau(v)+B_n(v)$ where $B_\tau(v)$ is the component tangent to sphere in H^1 and $B_n(v)$ is normal component. In this paper we change nonlinear operator $B(v)$ of input system on its normal part $B_n(v)$. Obtained equations, which we call Normal Parabolic Equations do not satisfy to analog in H^1 of energy estimate "'in the most degree"'. We hope that investigation of these equations can help to understand better the problems connected with solvability of 3d Navier-Stokes system in the class of strong solutions.

In this paper we study Normal Parabolic Equations (NPE) corresponding to 3D Navier-Stokes system. In Section 2 we derive NPE. In Section 3 we study some properties of NPE. The key property obtained there is existence of explicit formula for solution to NPE. In section 4 the structure of dynamical flow corresponding to NPE is investigated.

* The work has been fulfilled by RAS program "Theoretical problems of modern mathematics"', project "'Optimization of numerical algorithms of Mathematical Physics problems"'.

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 338–347, 2013.

Note that NPE has been introduced in [3] where normal parabolic equation corresponding to Burgers equation was studied. Here we generalize results from [3] on the case of NPE corresponding to Navier-Stokes system and, besides, continuer to develop the theory of NPE.

2 Semilinear Parabolic Equations of Normal Type

Our aim is to try to understand better how to investigate 3D Navier-Stokes system in phase space of one time differentiable vector fields where energy estimate is not true. To this end we derive some semilinear parabolic equation.

2.1 Navier-Stokes System and Helmholtz Equations

Let consider 3D Navier-Stokes equations with periodic boundary conditions:

$$\partial_t v(t,x) - \Delta v + (v,\nabla)v + \nabla p(t,x) = 0, \quad \operatorname{div} v = 0, \tag{1}$$

$$v(t,\dots,x_i,\dots) = v(t,\dots,x_i+2\pi,\dots), \qquad i = 1,2,3, \tag{2}$$

$$v(t,x)|_{t=0} = v_0(x) \tag{3}$$

where $t \in \mathbb{R}_+, x = (x_1,x_2,x_3) \in \mathbb{R}^3$, $v(t,x) = (v_1,v_2,v_3)$ is the velocity vector field of fluid flow, ∇p is the gradient of pressure, Δ is Laplace operator, $(v,\nabla)v = \sum_{j=1}^3 v_j \partial_{x_j} v$. Periodic boundary conditions (2) mean in fact that Navier-Stokes equations (1) and initial conditions (3) are defined on torus $\mathbb{T}^3 = (\mathbb{R}/2\pi\mathbb{Z})^3$.

We transform problem (1)-(3) for velocity to the problem for curl of velocity as unknown function:

$$\omega(t,x) = \operatorname{curl} v(t,x) = (\partial_{x_2} v_3 - \partial_{x_3} v_2,\ \partial_{x_3} v_1 - \partial_{x_1} v_3,\ \partial_{x_1} v_2 - \partial_{x_2} v_1) \tag{4}$$

Recall the following well-known formulas of vectorial analysis:

$$(v,\nabla)v = \omega \times v + \nabla \frac{|v|^2}{2}, \tag{5}$$

$$\operatorname{curl}(\omega \times v) = (v,\nabla)\omega - (\omega,\nabla)v, \quad \text{if} \quad \operatorname{div} v = 0, \quad \operatorname{div}\omega = 0 \tag{6}$$

where $\omega \times v = (\omega_2 v_3 - \omega_3 v_2,\ \omega_3 v_1 - \omega_1 v_3,\ \omega_1 v_2 - \omega_2 v_3)$ is vector product and $|v|^2 = v_1^2 + v_2^2 + v_3^2$. Let substitute (5) into the first equality of (1) and apply to both parts of obtained equality operator curl. Then in virtue of (4),(6), and formula $\operatorname{curl}\nabla F = 0$ we obtain the Helmholtz equations

$$\partial_t \omega(t,x) - \Delta\omega + (v,\nabla)\omega - (\omega,\nabla)v = 0 \tag{7}$$

We add these equations with initial conditions

$$\omega(t,x)|_{t=0} = \omega_0(x) \tag{8}$$

where $\omega_0 = \operatorname{curl} v_0$.

2.2 Normal Parabolic Equations (NPE) and Their Derivation

For each $m \in \mathbb{Z}_+ = \{j \in \mathbb{Z} : j \geq 0\}$ we define the space

$$V^m = V^m(\mathbb{T}^3) = \{v(x) \in (H^m(\mathbb{T}^3))^3 : \mathrm{div} v = 0, \int_{\mathbb{T}^3} v(x)dx = 0\} \tag{9}$$

where $H^m(\mathbb{T}^3)$ is Sobolev space.

Multiplying Navier-Stokes system (1) on v scalarly in $L_2(\mathbb{T}^3)$ we obtain after integration by parts (on x) and integration on t well-known energy estimate

$$\int_{\mathbb{T}^3} |v(t,x)|^2 dx + 2\int_0^t \int_{\mathbb{T}^3} |\nabla_x v(\tau,x)|^2 dx d\tau \leq \int_{\mathbb{T}^3} |v_0(x)|^2 dx \tag{10}$$

that gives opportunity to prove existence of weak solutions to problem (1)-(3). Unfortunately, this solution is not smooth enough to establish its uniqueness. If in a hope to get existence of smooth solution to (1) we would try to get analog of energy estimate in phase space V^1, multiplying (1) on v scalarly in $V^1(\mathbb{T}^3)$ we will not get analog of bound (10). Let try to understand situation passing from Navier-Stokes to Helmholtz equations.

Using decomposition in Fourier series

$$v(x) = \sum_{k\in\mathbb{Z}^3} \widehat{v}(k) e^{ix\cdot k}, \quad \text{where} \quad \widehat{v}(k) = (2\pi)^{-3} \int_{\mathbb{T}^3} v(x) e^{-ix\cdot k} dx,$$

$x \cdot k = \sum_{j=1}^3 x_j k_j$, $k = (k_1, k_2, k_3)$, and well-known formula curl curl $v = -\Delta v$ if div $v = 0$, we see that on space V^m inverse operator to curl is well-defined and is determined by the formula

$$\mathrm{curl}^{-1}\omega(x) = i \sum_{k\in\mathbb{Z}^3} \frac{k \times \widehat{\omega}(k)}{|k|^2} e^{ix\cdot k} \tag{11}$$

That is why operator

$$\mathrm{curl} \ : \ V^1 \longrightarrow V^0$$

realized isomorphism of the spaces. Therefore sphere in V^1 for problem (1), (3) is equivalent to sphere in V^0 for problem (7), (8).

Let denote nonlinear term of Helmholtz equation by B:

$$B(\omega) = (v, \nabla)\omega - (\omega, \nabla)v \tag{12}$$

(we did not indicate dependence B on v because it can be expressed via ω by (11)).

Multiplying equality (12) on $\omega = (\omega_1, \omega_2, \omega_3)$ scalarly in V^0 and integrating by parts we get

$$(B(\omega), \omega)_{V^0} = -\int_{\mathbb{T}^3} \sum_{j,k=1}^3 \omega_j \partial_j v_k \omega_k dx \tag{13}$$

that, generally saying, is not equal to zero. Just because of this 3D Helmholtz equations do not possess energy estimate. In other words, operator B admits the decomposition

$$B(\omega) = B_n(\omega) + B_\tau(\omega) \tag{14}$$

where vector $B_n(\omega)$ is orthogonal to the sphere $\Sigma_\omega = \{u \in V^0 : \|u\|_{V^0} = \|\omega\|_{V^0}\}$ at the point ω, and vector $B_\tau(\omega)$ is tangential to Σ_ω at ω. Generally saying, both operators B_n, B_τ in (14) are not equal to zero. Note that just the component $B_n \neq 0$ prevents to derivation of energy bound and therefore it is quite possible that the main difficulties obstructing to investigation of Navier-Stokes equations are connected just with this operator. That is why there is reason to omit in Helmholtz equations the component B_τ and to study on the first stage analog of equations (7) in which nonlinear operator $B(\omega)$ is changed on its normal component $B_n(\omega)$. Such equations we call Normal Parabolic Equations (NPE).

Let construct now normal parabolic equations with respect to sphere in V^0, corresponding to problem (7), (8).

Since summand $(v, \nabla)\omega$ from (7) is tangential operator:

$$\int_{\mathbb{T}^3} (v, \nabla)\omega \cdot \omega dx = 0,$$

normal part of nonlinear operator from (7) is defined by nonlinear term $(\omega, \nabla)v$. We look it for in the form $\Phi(\omega)\omega$ where Φ is unknown functional that is found by equation

$$\int_{\mathbb{T}^3} \Phi(\omega)\omega(x) \cdot \omega(x)dx = \int_{\mathbb{T}^3} (\omega(x), \nabla)v(x) \cdot \omega(x)dx \tag{15}$$

Relation (15) implies desired formula for Φ:

$$\Phi(\omega) = \begin{cases} \int_{\mathbb{T}^3} (\omega(x), \nabla)\mathrm{curl}^{-1}\omega(x) \cdot \omega(x)dx \,/ \int_{\mathbb{T}^3} |\omega(x)|^2 dx, & \omega \neq 0, \\ 0, & \omega \equiv 0 \end{cases} \tag{16}$$

where $\mathrm{curl}^{-1}\omega(x)$ is defined in (11).

So, normal parabolic equations corresponding to system (7) are defined as follows:

$$\partial_t \omega(t, x) - \Delta\omega - \Phi(\omega)\omega = 0, \quad \mathrm{div}\,\omega = 0 \tag{17}$$

where functional Φ is defined in (16). These equations supplied with initial conditions (8) and periodic boundary conditions are the main object of our investigation in this paper.

3 Properties of Normal Parabolic Equations

3.1 Explicit Formula for Solution of NPE

In this subsection we derive explicit formula for NPE solution. This is the key result because it gives the possibility to establish many important properties on NPE. Some of them will be obtained below in next sections. The following assertion is true:

Lemma 1. *Let $S(t,x,y_0)$ be resolving operator of the following Stokes system with periodic boundary conditions:*

$$\partial_t y(t,x) - \Delta y(t,x) = 0, \quad div\, y = 0, \quad y(t,x)|_{t=0} = y_0(x), \tag{18}$$

i.e. $S(t,x,y_0) = y(t,x)$ (we assume, of course, that $div\, y_0 = 0$). The solution of problem (17),(18) has the form

$$\omega(t,x;\omega_0) = \frac{S(t,x;\omega_0)}{1 - \int_0^t \Phi(S(\tau,\cdot;\omega_0))d\tau} \tag{19}$$

The proof of this lemma is reduced to substitution (19) into (17) and direct checking of obtained equality.

3.2 Properties of Functional Φ

Let $s \in \mathbb{R}$. Recall that Sobolev space $H^s(\mathbb{T}^3)$ is the space of periodic real-valued distributions $z(x)$ possessing with the finite norm

$$\|z\|^2_{H^s(\mathbb{T}^3)} \equiv \|z\|^2_s = \sum_{k\in\mathbb{Z}^3\setminus\{0\}} |k|^{2s}|\widehat{z}(k)|^2 < \infty \tag{20}$$

where $\widehat{z}(k)$ are Fourier coefficients of z.

We will use the following generalization of spaces (9) of solenoidal vector fields:

$$V^s \equiv V^s(\mathbb{T}^3) = \{v(x) \in (H^s(\mathbb{T}^3))^3 : \quad \mathrm{div} v(x) = 0, \quad \int_{\mathbb{T}^3} v(x)dx = 0\}, \quad s \in \mathbb{R} \tag{21}$$

Lemma 2. *Let $\Phi(u)$ be functional (16). There exists a constant $c > 0$ such that for each $u \in V^{3/2}$*

$$|\Phi(u)| \le c\|u\|_{3/2} \tag{22}$$

This lemma is proved similarly to analogous bound from [3].

Lemma 3. *Let Φ be functional (16). For each $\beta < 1/2$ there exists a constant $c_1 > 0$ such that for each $y_0 \in V^{-\beta}(\mathbb{T}^3)$, $t > 0$*

$$\left|\int_0^t \Phi(S(\tau,\cdot,y_0))d\tau\right| \le c_1\|y_0\|_{-\beta} \tag{23}$$

where $S(t,\cdot,y_0)$ is resolving operator of problem (18).

Proof. Using (22) we get

$$\left|\int_0^t \Phi(S(\tau,\cdot,y_0))d\tau\right| \le c\int_0^t e^{-\tau/2}\left(\sum_{k\neq 0}(|\widehat{y_0}(k)|^2|k|^{-2\beta})|k|^{3+2\beta}e^{-(k^2-1)\tau}\right)^{1/2} d\tau \tag{24}$$

where $\widehat{y}_0(k)$ are Fourier coefficients of y_0. Solution $\hat{\rho} = \rho(t)$ of extremal problem

$$f(t,\rho) = \rho^{3+2\beta} e^{-(\rho^2-1)t} \to \max, \quad \rho \geq 1$$

is defined with expression $\rho(t) = \sqrt{\frac{3+2\beta}{2t}}$, and

$$f(t,\rho(t)) = \begin{cases} \left(\frac{3+2\beta}{2t}\right)^{\frac{3+2\beta}{2}} e^{-(3+2\beta-2t)/2}, & t \leq \frac{3+2\beta}{2}, \\ 1, & t \geq \frac{3+2\beta}{2} \end{cases} \tag{25}$$

Substitution (25) into (24) implies (23).

Remark 1. Lemma 3 implies that the functional from left side of bound (23) is well defined for $y_0 \in V^{-\beta}(\mathbb{T}^3)$ with $\beta < 1/2$. In particular, in virtue of this Lemma and (19) solution of problem (17),(8) is well defined for each initial condition $\omega_0 \in V^0$, and therefore our choice V^0 as phase space for corresponding dynamical system is correct. Note also that simple modification of Lemma 3 proof gives continuity of the functional from left side in (23) with respect to $y_0 \in V^{-\beta}$, $\beta < 1/2$.

4 The Structure of NPE Dynamics

The aim of this section is to find out the main feature of dynamical flow corresponding to NPE. We decompose the phase space of the dynamical system on three sets with different behavior of dynamical flow inside each of them.

4.1 Distinctive Sets of Phase Space

Let give definitions of three subsets of phase space for NPE. Recall that we take $V^0(\mathbb{T}^3) \equiv V^0$ as the phase space for problem (17),(8).

Definition 1. *The set $M_- \equiv M_-(\alpha) \subset V^0$ of initial conditions ω_0 such that the solution $\omega(t,x;\omega_0)$ of problem (17),(8) exists and satisfies inequality*

$$\|\omega(t,\cdot;\omega_0)\|_0 \leq \alpha \|\omega_0\|_0 e^{-t} \qquad \forall t > 0 \tag{26}$$

is called the set of stability. Here $\alpha > 1$ is a certain fixed number.

The following simple sufficient condition for belonging to $M_-(\alpha)$ is true in virtue of (19): If $\omega_0 \in V^0$ satisfies the bound

$$\sup_{t\in\mathbb{R}_+} \int_0^t \Phi(S(\tau,\cdot;\omega_0))d\tau \leq \frac{\alpha-1}{\alpha} \tag{27}$$

then $\omega_0 \in M_-(\alpha)$.

Definition 2. *The set $M_+ \subset V^0$ of initial conditions ω_0 from (17), (8) such that corresponding solution $\omega(t,x;\omega_0)$ exists only on a finite interval $t \in (0,t_0)$ with $t_0 > 0$ depending on ω_0, and blows up at $t = t_0$ is called the set of explosions.*

In virtue of formula (19) for solution $\omega(t,x;\omega_0)$

$$M_+ = \{\omega_0 \in V^0 : \exists t_0 > 0 \int_0^{t_0} \Phi(S(\tau,\cdot;\omega_0))d\tau = 1\} \tag{28}$$

The minimal magnitude from the set $\{t_0\}$ for which equality in (28) holds is called the time of explosion.

Definition 3. *The collection*

$$M_I(\alpha) = V^0 \setminus \{M_-(\alpha) \cup M_+\} \tag{29}$$

is called intermediate set.

Remark 2. Definitions of stability and intermediate sets include parameter $\alpha > 1$ and from this point of view they are not absolute. Nevertheless they are convenient for using.

We study below the properties of these sets and, in particular, we show that all these sets are nonempty. We begin from the set of stability. This set is the most important for us.

4.2 Subsets Belonging to the Set of Stability

Let $\rho > 0, \beta < 1/2$. Introduce the set

$$\begin{aligned} El^\beta_\rho &= \{v \in V^0(\mathbb{T}^3) : \|v\|^2_{-\beta} = \sum_{k\in\mathbb{Z}^3\setminus\{0\}} \frac{|\widehat{v}(k)|^2}{|k|^{2\beta}} \le \rho^2\} \\ &= \{v \in V^0(\mathbb{T}^3) : \sum_{k\in\mathbb{Z}^3\setminus\{0\}}^{\infty} \frac{|\widehat{v}(k)|^2}{\rho^2|k|^{2\beta}} \le 1\}, \end{aligned} \tag{30}$$

which we can interpret as ellipsoid in $V^0(\mathbb{T}^3)$ with length of axes directed along functions $e^{ik\cdot x}, e^{-ik\cdot x}$ equal to $\rho|k|^\beta$. Since $\rho|k|^\beta \to \infty$ as $|k| \to \infty$, this ellipsoid is unbounded in V^0.

Lemma 4. *Let $c_1\rho < 1$ and $\rho \le (\alpha-1)/(\alpha c_1)$ where c_1 is the constant from (23). Then*

$$El^\beta_\rho \subset M_-(\alpha) \tag{31}$$

where the set El^β_ρ is defined in (30).

Proof. Note that solution $S(t,x,\omega_0)$ of problem (18) with $y_0=\omega_0$ satisfies inequality:

$$\|S(t,\cdot,\omega_0)\|_0^2=\sum_{k\neq 0}e^{-2|k|^2t}|\widehat{\omega}_0(k)|^2=e^{-2t}\sum_{k\neq 0}e^{-2(|k|^2-1)t}|\widehat{\omega}_0(k)|^2\le e^{-2t}\|\omega_0\|_0^2 \tag{32}$$

Let $\omega_0\in El_\rho^\beta$, i.e. $\|\omega_0\|_{-\beta}\le\rho$. Formula (19) and inequalities (32), (23) imply (26) if $\alpha\ge 1/(1-c_1\rho)$. But the last inequality is equivalent to $\rho\le(\alpha-1)/(\alpha c_1)$.

This Lemma is analog of local existence theorem for 3D Navier-Stokes equations obtained in [4] with help of Fujita-Kato approach [5] and of local existence theorem for NPE connected with Burgers equation (see [3]). The proof of Lemma 4 is essentially easier then proofs of aforementioned results because here we use explicit formula (19) for solutions.

We show in this subsection that, actually, $M_-(\alpha)$ is essentially wider than El_ρ^β. For this goal we consider one infinite-dimensional subspace and show that it belongs to $M_-(\alpha)$

Let introduce the following subset U_L of $\mathbb{Z}^3\setminus\{0\}$:

$$U_L=\{\xi\in\mathbb{Z}^3\setminus\{0\}:\quad \xi+\eta-\zeta\neq 0\quad \forall\xi,\eta,\zeta\in U_L\} \tag{33}$$

An example of the subset belonging to U_L is the following set:

$$\{k=(k_1,k_2,k_3)\in\mathbb{Z}^3\setminus\{0\}:k_1+k_2+k_3\quad\text{is odd number }\}$$

Lemma 5. *The subspace*

$$L=\{\omega_0=\sum_{k\in U_L}(z_ke^{ik\cdot x}+\overline{z}_ke^{-ik\cdot x}),\quad z_k\in\mathbb{C}^3,\quad z_k\cdot k=0\}\subset V^0(\mathbb{T}^3) \tag{34}$$

belongs to $M_-(\alpha)$ if U_L is the set (33). Moreover

$$\forall\omega_0\in L\quad \Phi(S(t,\cdot,\omega_0))\equiv 0\quad \forall t\ge 0 \tag{35}$$

The proof of this Lemma is similar to analogous Lemma from [3].

Lemmas 4, 5 imply that $M_-(\alpha)\neq\emptyset$.

4.3 Certain Sets of Unit Sphere of V^0

Let denote the unit sphere of the phase space V^0 as follows:

$$\Sigma=\{v\in V^0:\quad \|v\|_0=1\} \tag{36}$$

To understand better the structure of phase flow corresponding to problem (17),(8) we introduce on Σ several sets. Define

$$A_-(t)=\{v\in\Sigma:\int_0^t\Phi(S(\tau,\cdot,v))d\tau\le 0\},$$

$$A_+(t)=\{v\in\Sigma:\int_0^t\Phi(S(\tau,\cdot,v))d\tau\ge 0\},$$

$$A_0(t)=\{v\in\Sigma:\int_0^t\Phi(S(\tau,\cdot,v))d\tau=0\},$$

and

$$A_- = \cap_{t\geq 0} A_-(t), \quad A_+ = \cap_{t\geq 0} A_+(t), \quad A_0 = \cap_{t\geq 0} A_0(t) \tag{37}$$

All these sets are closed and nonempty. For instance, $A_0 \neq \emptyset$ in virtue of Lemma 5. Sets $A_\pm(t), A_\pm$ possess nonempty interior in topology of Σ, i.e. in the topology induced on Σ by topology of the space V^0. This assertion follows from continuity of the functional $V^{-\beta} \ni v \to \int_0^t \Phi(S(\tau,\cdot,v))d\tau$ with $\beta < 1/2$ and in particular for $\beta = 0$ (see Remark 1). Evidently, $A_0 = A_- \cap A_+$.

Linearity on v of operator $S(t,x,v)$ and oddness of $\Phi(v)$ with respect to v imply

Lemma 6.

$$v \in A_- \quad \textit{if and only if} \quad -v \in A_+$$

Introduce also the sets

$$B_+ = \Sigma \setminus A_- \equiv \{v \in \Sigma : \exists t_0 > 0 \int_0^{t_0} \Phi(S(\tau,\cdot,v))d\tau > 0\}, \quad B_- = \Sigma \setminus A_+ \tag{38}$$

It is easy to see that the set B_+ is open in topology of Σ. Moreover, the boundary ∂B_+ of set B_+ is defined by the relation

$$\partial B_+ = \{v \in \Sigma : \forall t > 0 \int_0^t \Phi(S(\tau,\cdot,v))d\tau \leq 0, \exists t_0 > 0 : \int_0^{t_0} \Phi(S(\tau,\cdot,v))d\tau = 0\}$$

It is clear that $A_0 \subset \partial B_+$ and $\partial B_+ \setminus A_0 \neq \emptyset$.

4.4 On Structure of Phase Space V^0

Let us introduce the following function defined on the set B_+ of sphere Σ:

$$B_+ \ni v \to b(v) = \max_{t\geq 0} \int_0^t \Phi(S(\tau,v))d\tau \tag{39}$$

Evidently, $b(v) > 0$ and $b(v) \to 0$ as $v \to \partial B_+$. Let $\rho \in (0,1]$. We define the following map $\Gamma_\rho(v)$ that plays the key role in description of structure of phase flow generated by boundary value problem (17),(8):

$$B_+ \ni v \to \Gamma_\rho(v) = \frac{\rho}{b(v)} v \in V^0 \tag{40}$$

where $b(v)$ is function (39). Note that $\|\Gamma_\rho(v)\|_0 \to \infty$ as $v \to \partial B_+$.

Theorem 1. *Let $\alpha > 1$ be a parameter from definition of the stability set $M_-(\alpha)$ and $\rho = (\alpha-1)/\alpha$. Then the image $\Gamma_\rho(v)$, $v \in B_+$ of the map Γ_ρ divides the space V^0 on two separate parts. The part containing origin coincides with the set of stability $M_-(\alpha)$. The part of V^0 between $\Gamma_\rho(v)$, $v \in B_+$ and $\Gamma_1(v)$, $v \in B_+$ coincides with intermediate space $M_I(\alpha)$, and the rest part of V^0 coincides with the set of explosions M_+.*

The proof of this theorem will be given in some other place.

Theorem 1 implies that $M_+ \neq \emptyset$ and $M_I(\alpha) \neq \emptyset$.

References

[1] Ladyzhenskaya, O.A.: The Mathematical Theory of Viscous Incompressible Flow. Gordon and Breach, New York (1969)
[2] Temam, R.: Navier-Stokes Equations – Theory and Numerical Analysis. AMS Chelsea Publishing, Providence (2001)
[3] Fursikov, A.V.: On one semilinear parabolic equation of normal type. Mathematics and life sciences. De Gruyter 1, 147–160 (2012)
[4] Fursikov, A.V.: Local Existence Theorems with Unbounded Set of Input Data and Unboundedness of Stable Invariant Manifolds for 3D Navier-Stokes Equations. J. Discr. and Cont. Dyn. Syst. Series S 3(2), 269–290 (2010)
[5] Fujita, H., Kato, T.: On the Navier-Stokes initial value problem. J. Arch. for Rat. Mech. and Anal. 16, 269–315 (1964)

A Nonlinear Model Predictive Concept for Control of Two-Phase Flows Governed by the Cahn-Hilliard Navier-Stokes System

Michael Hinze and Christian Kahle

Fachbereich Mathematik,
Optimierung und Approximation
Universität Hamburg, Bundesstraße 55, 20146 Hamburg, Germany
{Michael.Hinze,Christian.Kahle}@uni-hamburg.de

Abstract. We present a nonlinear model predictive framework for closed-loop control of two-phase flows governed by the Cahn-Hilliard Navier-Stokes system. We adapt the concept for instantaneous control from [6,12,16] to construct distributed closed-loop control strategies for two-phase flows. It is well known that distributed instantaneous control is able to stabilize the Burger's equation [16] and also the Navier-Stokes system [6,12]. In the present work we provide numerical investigations which indicate that distributed instantaneous control also is well suited to stabilize the Cahn-Hilliard Navier-Stokes system.

Keywords: Flow control, Navier-Stokes, Cahn-Hilliard, Model Predictive Control, Instantaneous Control, Adaptive Finite Elements.

1 Introduction

The aim of this work is the development of numerical methods for closed-loop control of two-phase flows governed by the Cahn-Hilliard Navier-Stokes system. The approach is based on an inexact variant of model predictive control called instantaneous control. Instantaneous control in the context of flow control is proposed in e.g. [5,6,14], and for distributed control of the Navier-Stokes system is analyzed in [12], where among other things it is shown that the method is able to exponentially stabilize given solution states supposing certain smallness assumptions on the initial conditions. For an overview in the field of nonlinear model predicitve control we refer to [20,21] and also to the monograph [9], where also further references can be found.

The outline of this paper is as follows. In section 2 we describe the concept of nonlinear model predictive control as it is used in the present work, and also introduce instantaneous control. In section 3 we present a brief introduction to the Cahn-Hilliard Navier-Stokes system, including its numerical treatment. In section 4 we discribe the instantaneous control strategy for the Cahn-Hilliard Navier-Stokes system and demonstrate its performance at morphing a circle into a square in section 5. We end with some conclusions formulated in section 6.

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 348–357, 2013.

2 Nonlinear Model Predictive Control

The aim of model predictive control consist in steering or keeping the state of a dynamical system to or at a given desired trajectory. To fix the concept, we are going to apply it to the Cahn-Hilliard Navier-Stokes system. Let us first consider an abstract dynamical system with initial condition x_0, state $x(t)$, observation $y(t)$ and control $u(t)$:

$$\begin{aligned} \dot{x}(t) + Ax(t) &= b(x,t) + \mathcal{B}u(t),\ (t>0) && \text{state,} \\ y(t) &= \mathcal{C}x(t) && \text{observation,} \\ x(0) &= x_0 && \text{initial condition.} \end{aligned} \tag{1}$$

Our aim consists in constructing a nonlinear feedback control law K with $\mathcal{B}u(t) = K(x(t))$ which steers the dynamical system to the desired trajectory $\bar{x}(t)$, i.e. $x(t) \overset{!}{\to} \bar{x}(t)$ $(t \to \infty)$. To simplify notations we from here onwards use $\mathcal{B} = id$ and $\mathcal{C} = id$, i.e. we do not distinguish between state and observation and we allow fully distributed controls.

To prepare for model predicitve control, system (1) is discretized in time using the semi-implicit Euler method on an equidistant time grid $0 = t_0 \le t_1 \le \dots$ with $t_{k+1} - t_k = \tau$ for $k = 0, 1, \dots$. Here x^k denotes the state at time t_k and b^k denotes the nonlinearity $b(x^k, t_k)$. We obtain the time discrete model

$$(I + \tau A)x^{k+1} = x^k + \tau b^k + u^{k+1}, \quad k = 0, 1, \dots. \tag{2}$$

For $L \in \mathbb{N}$ and x^j given, we consider the optimal control problem

$$\begin{aligned} &\min J(x^{j+1}, \dots, x^{j+L}, u^{j+1}, \dots, u^{j+L}) \\ &\quad \text{s.t. (2) for } j = k, \dots, k+L-1, \end{aligned} \tag{$\mathcal{P}_k$}$$

where

$$J(x^{j+1}, \dots, x^{j+L}, u^{j+1}, \dots, u^{j+L}) := \sum_{i=1}^{L} \left(\frac{1}{2}\|x^{k+i} - \bar{x}^{k+i}\|^2 + \frac{\alpha}{2}\|u^{k+i}\|^2 \right)$$

Let us note that for $L = 1$ problem ($\mathcal{P}_k$) admits a unique solution. However, in the case $L > 1$ the solution might not be unique due to the nonlinear character of the transition constraints (2). In this case we assume that ($\mathcal{P}_k$) admits a solution.

Now we define the abstract model predictive control strategy using a computing oracle called $RECIPE$.

Model Predicitve Control:

1. Initialization: Specify time grid (t_j) and discrete state $\bar{x}$, set $k = 0$ and specify $L_0 > 0$.
2. Given u^k, x^k, set
   $u^{k+1} = RECIPE(u^k, x^k, L_k)$.

3. Solve (2) with u^{k+1}, i.e. compute x^{k+1} according to $(I + \tau A)x^{k+1} = x^k + \tau b(x^k, t_k) + u^{k+1}$.
4. Set $k = k + 1$, goto 2.

In the classical model predictive control context [20,21] the oracle $RECIPE(u, x, L)$, for given u, x, L, solves the optimal control problem ($\mathcal{P}_k$) with $x^k \equiv x, u^k \equiv u$ and L the length of the control horizon. From the solution $(x^{k+1}, \ldots, x^{k+L}, u^{k+1}, \ldots, u^{k+L})$ only x^{k+1} and u^{k+1} are actually used to steer the discrete system to the next time instance t_{k+1}.

In practical applications often quick response to system changes is necessary. In such cases it may be too time consuming to solve problem ($\mathcal{P}_k$) exactly. Instead, an approximate solution could be used. This leads to the so called concept of instantaneous control [6,12], whose oracle is described next.

Instantaneous Control:
Given L, x, v, then $u = RECIPE(v, x, L)$ iff $u = u^{k+1}$, where $(x^{k+1}, \ldots, x^{k+L}, u^{k+1}, \ldots, u^{k+L})$ is the result of a steepest descent step applied to the solution of ($\mathcal{P}_k$) with $x^k \equiv x$ and $u^k \equiv v$.

In the case $L = 1$, instantaneous control realizes the steps:

- Solve $(I + \tau A)z = x + \tau b(x, t_k) + v$,
- solve $(I + \tau A^*)\lambda = -(x - z)$,
- set $d = \alpha v - \lambda$,
- determine $\rho > 0$ (step size for steepest descent),
- set $RECIPE = v - \rho d$,

where we have used the adjoint calculus to expose the derivations of the functional J, see e.g. [15].

Instantaneous control with $L = 1$ is analytically investigated in [16] for the control of Burger's equation and in [12] for control of the two-dimensional Navier-Stokes system. Among other things it is shown in [12, Thm. 4.4,4.5] that

$$\|x^k - \bar{x}^k\|_{H^1(\Omega)} \leq c\kappa^k \quad \text{for some } \kappa \in (0, 1),$$

and also that instantaneous control may be regarded as the discrete realization of a nonlinear feedback operator K which is able to steer $x(t)$ exponentially fast to the desired trajectory $\bar{x}(t)$, i.e. with $u(t) = K(x(t))$ in (1) there holds

$$\|x(t) - \bar{x}(t)\|_{H^1(\Omega)} \leq c \exp\left(-\frac{\rho}{\tau} t\right),$$

where ρ denotes the step size in the steepest descent algorithm and τ is the time step in the discretization, see [12, Thm. 4.1,4.2] for the details.

3 The Cahn-Hilliard Navier-Stokes System

The Cahn-Hilliard Navier-Stokes equations are a diffuse interface model for describing two-phase flows. In comparison to sharp interface models (see e.g. [8])

which model the interface between the two components as a sharp line, diffuse interface models allow a partial mixing of the components yielding a small diffuse interface. A derivation of the model used here can be found e.g. in [1]. It is related to the model 'H' for two-phase flows in the classification of Hohenberg and Halperin [17]. We note that there are also models for flows with three components, see e.g [3].

For $\Omega \subset \mathbb{R}^n, (n = 2, 3)$ and $T > 0$ we here consider the following weak form of the Cahn-Hilliard Navier-Stokes system with double-obstacle free energy according to [2].

Find $(c(t), w(t), y(t), p(t))$ in $\mathcal{K} \times H^1(\Omega) \times H^1_0(\Omega) \times L^2_{(0)}(\Omega)$ such that

$$
\begin{aligned}
&(\partial_t y, v) + \frac{1}{Re}(\nabla y : \nabla v) \\
&+((y \cdot \nabla)y, v) - (p, \operatorname{div} v) + (c\nabla w, v) = 0 && \forall v \in H^1_0(\Omega),\ a.e.\ t \in (0,T], && (3)\\
&(-\operatorname{div} y, v) = 0 && \forall v \in L^2_{(0)}(\Omega),\ a.e.\ t \in (0,T], && (4)\\
&c(t) \in \mathcal{K} && a.e.\ t \in (0,T], && (5)\\
&(\partial_t c, v) + \frac{1}{Pe}(\nabla w, \nabla v) - (cy, \nabla v) = 0 && \forall v \in H^1(\Omega),\ a.e.\ t \in (0,T], && (6)\\
&\gamma^2(\nabla c, \nabla(v-c)) - (w + c, v - c) \geq 0 && \forall v \in \mathcal{K},\ a.e.\ t \in (0,T], && (7)\\
&c(x,0) = c^0(x) && \forall x \in \Omega, && (8)\\
&\partial_\nu c = 0, \quad \partial_\nu w = 0 && \text{on } \partial\Omega \times (0,T], && (9)\\
&y(x,0) = y^0(x) && \forall x \in \Omega, && (10)\\
&y = 0 && \text{on } \partial\Omega \times (0,T]. && (11)
\end{aligned}
$$

Here, for $v, w \in H^1(\Omega)$

$$(\nabla v : \nabla w) := \int_\Omega \nabla v : \nabla w \, dx = \int_\Omega \sum_{i,j=1}^{n} (\nabla v)_{ij} (\nabla w)_{ij} \, dx,$$

and $c^0 \in \mathcal{K} := \{v \in H^1(\Omega) \,|\, |v| \leq 1 \text{ a.e. in } \Omega\})$, $y^0 \in H^1_0(\Omega)$.

The function c is called order parameter and satisfies $c = c(t,x) \in [-1,1]$, with $c \equiv 1$ on the pure A-phase and $c \equiv -1$ on the pure B-phase region, respectively, where A and B are the two components of the fluid. Initially, i.e. for $t = 0$, we assume that the concentration equals c^0. The quantity w represents the chemical potential, y denotes the mean flow velocity field, i.e. $y = \frac{1+c}{2} y_A + \frac{1-c}{2} y_B$, where y_A and y_B denote the fluid velocities in the fluid phases A and B, respectively, and $p \in L^2_{(0)}(\Omega) = \{v \in L^2(\Omega) \,|\, (v,1) = 0\}$ denotes the mean free pressure of the fluid. The flow profile at $t = 0$ is given by y^0. The Péclet number Pe and the Reynolds number Re are given physical constants. The parameter γ is related to the width of the diffuse interface region which is of size $\mathcal{O}(\gamma^2)$ [4].

For an analytical treatment of the above system we refer to [1, Chapter 6.5]. Especially in two space dimensions there exists a unique solution (c, w, y, p) to this system and we have

$$y \in C^0([0,T], \mathbf{H}_0^1(\Omega)), \quad c \in BC_\omega([0,T], H^1(\Omega)), \quad \nabla w \in L^2(0,T; L^2(\Omega)).$$

Here, $BC_\omega([0,T], H^1(\Omega))$ is the space of bounded and weakly continuous functions from $[0,T]$ with values in $H^1(\Omega)$.

3.1 Time Discretization and Treatment of the Variational Inequality in (7)

For the discretization of (3)–(11) we use the semi-implicit approach proposed in e.g. [7,19] with constant step size $\tau > 0$. The variational inequality in (7) according to [10] is relaxed using Moreau-Yosida regularization. At time instance t this results in finding $(y, p, c, w) \in H^1(\Omega)^n \times L^2_{(0)}(\Omega) \times H^1(\Omega) \times H^1(\Omega)$ such that

$$\begin{aligned}
(y - y_{old}, v) + \frac{\tau}{Re}(\nabla y : \nabla v) + \tau((y_{old} \cdot \nabla) y_{old}, v) & \\
+\tau(c_{old} \nabla w_{old}, v) - \tau(p, \operatorname{div} v) = 0 \quad & \forall v \in H_0^1(\Omega), \qquad (12)\\
(-\operatorname{div} y, v) = 0 \quad & \forall v \in L^2_{(0)}(\Omega), \qquad (13)\\
(c, v) + \frac{\tau}{Pe}(\nabla w, \nabla v) - \tau(c_{old} y, \nabla v) - (c_{old}, v) = 0 \quad & \forall v \in H^1(\Omega), \qquad (14)\\
\gamma^2(\nabla c, \nabla v) - (w, v) + (\lambda_s(c), v) - (c_{old}, v) = 0 \quad & \forall v \in H^1(\Omega). \qquad (15)
\end{aligned}$$

is satisfied. Here the subscript $_{old}$ refers to the value of the respective function at time $t_{old} = t - \tau$ and $\lambda_s(c) = \lambda_s^+(c) + \lambda_s^-(c) = s(\max(0, c-1) + \min(0, c+1))$ stems from Moreau-Yosida regularization of (7), see e.g. [10,11] for details. Let us emphasize that (12)–(15) is decoupled in the sense that using y_{old}, c_{old} and w_{old}, the flow y and the pressure p at time t can be computed from (12)–(13), and then using this flow vector y, the concentration c and the chemical potential w are obtained from (14)–(15). Furthermore, normalizing p by $(p, 1) = 0$, it can be shown that (12)–(15) admits a unique solution (y, p, c, w), compare [11, Thm 4.1]. The system (14)–(15) is nonlinear and can be treated by semi-smooth Newton methods, see [11].

3.2 Spatial Discretization

The spatial discretization is performed by linear finite elements for both the concentration and the chemical potential to obtain finite element approximations c^h, w^h. For the flowfield and the pressure we use the LBB-stable Taylor-Hood $P^2 - P^1$ finite elements, see e.g. [18,22], to obtain finite element approximations y^h, p^h. For the spatial treatment of the Cahn-Hilliard part (14)–(15) we use the adaptive approach presented in [10,11]. We emphasize that we use different spatial meshes for the Cahn-Hilliard and the Navier-Stokes part.

4 Instantaneous Control of the Cahn-Hilliard Navier-Stokes System

The aim of this section is to develop a simple distributed closed-loop control strategy for the Cahn-Hilliard Navier-Stokes system. It uses instantaneous control with $L = 1$ on a control horizon which coincides with the time interval of one time step. We consider the idealized situation that the flow can be controlled by a vector field which is distributed over the whole spatial domain and that the concentration c can be measured in the whole spatial domain. The control goal consists in steering the concentration c towards a prescribed concentration trajectory c_d by applying volume forces to the flow field.

Now let $t_0 = 0$ and $k \in \mathbb{N}$. At time instance $t = t_k$, for $\alpha > 0$, we consider the minimization problem

$$\min J_k(u) := \frac{1}{2}\|c - c_d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)} \qquad (\mathcal{P}_k)$$

s.t.

$$(y - y_{old}, v) + \frac{\tau}{Re}(\nabla y : \nabla v) + \tau((y_{old} \cdot \nabla)y_{old}, v)$$
$$+\tau(c_{old}\nabla w_{old}, v) - \tau(p, \operatorname{div} v) = u \quad \forall v \in H^1_0(\Omega), \qquad (16)$$
$$(-\operatorname{div} y, v) = 0 \quad \forall v \in L^2_{(0)}(\Omega), \qquad (17)$$
$$(c, v) + \frac{\tau}{Pe}(\nabla w, \nabla v) - \tau(c_{old}y, \nabla v) - (c_{old}, v) = 0 \quad \forall v \in H^1(\Omega), \qquad (18)$$
$$\gamma^2(\nabla c, \nabla v) - (w, v) + (\lambda_s(c_{old}), v) - (c_{old}, v) = 0 \quad \forall v \in H^1(\Omega). \qquad (19)$$

where now $_{old}$ refers to the time instance t_{k-1}.

It is not hard to show that problem ($\mathcal{P}_k$) admits a unique solution $u \in L^2(\Omega)^n$, which together with (y, p, c, w) satisfies the adjoint system

$$(p_1, v) + \gamma^2(\nabla p_2, \nabla v) = (c - c_d, v) \quad \forall v \in H^1(\Omega), \qquad (20)$$
$$(p_2, v) = \frac{\tau}{Pe}(\nabla p_1, \nabla v) \quad \forall v \in H^1(\Omega), \qquad (21)$$
$$(p_3, v) + \frac{\tau}{Re}(\nabla p_3, \nabla v) - \tau(p_4, \operatorname{div} v) = \tau(p_1 \nabla c_{old}, v) \quad \forall v \in (H^1_0(\Omega))^n, \qquad (22)$$
$$(\operatorname{div} p_3, v) = 0 \quad \forall v \in L^2_{(0)}(\Omega), \qquad (23)$$
$$\alpha u + p_3 = 0. \qquad (24)$$

i.e. there exists a uniquely determined adjoint $(p_1, p_2, p_3, p_4) \in H^1(\Omega) \times H^1(\Omega) \times H^1(\Omega)^n \times L^2_{(0)}(\Omega)$ which solves (20)–(24). The gradient of J_k is given by $\nabla J_k(v) = \alpha v + p_3$. Its evaluation for a given $v \in L^2(\Omega)^n$ amounts to first solving (16)–(19) for (y, p, c, w) and then solving (20)–(23) for (p_1, p_2, p_3, p_4).

Let us note that (16)–(19) differ from (12)–(15) in the explicit treatment of the nonlinearity λ_s, which in (19) is frozen at t_{old}. We emphasize that (16)–(19) is not used to simulate the controlled Cahn-Hilliard Navier-Stokes system, but to construct a feedback operator K such that $u = K(y, p, c, w)$. With this

feedback operator available, system (3)–(11) then is controlled through inserting $u = K(y, p, c, w))$ as right hand in (3). The resulting system then is treated on the time discrete level as in (12)–(15), where K is evaluated at t_{old}, i.e.

$$u = K(y_{old}, p_{old}, c_{old}, w_{old}).$$

Next we describe the construction of K from (16)–(24). For this pupose let us denote by B the solution operator associated to the quasi-Stokes problem (12)–(13) and by $\mathcal{C}$ the linear, fourth order solution operator of the Cahn-Hilliard system (18)–(19). Then the system to obtain p_3 can be written in the form

$$\begin{aligned} y &= B\left(y_{old} - \tau(y_{old}\nabla y_{old})) - \tau c_{old}\nabla w_{old} + u\right), \\ c &= \mathcal{C}\left(c_{old} - \tau y\nabla c_{old} + \frac{\tau}{Pe}\Delta(\lambda_s(c_{old}) - c_{old})\right), \\ p_1 &= \mathcal{C}(c - c_d), \text{ and} \\ p_3 &= -\tau B(p_1\nabla c_{old}). \end{aligned}$$

Using these abbreviations, the control u obtained by the instantaneous control strategy for $u_0^k = 0$ is given by

$$\begin{aligned} \tilde{y} &= B(y_{old} + \tau b(y_{old}) - \tau c_{old}\nabla w_{old}), \\ \tilde{c} &= \mathcal{C}\left(c_{old} - \tau\nabla c_{old}\tilde{y} + \frac{\tau}{Pe}\Delta(\lambda_s(c_{old}) - c_{old})\right), \\ u &= \rho\tau B\nabla c_{old}\mathcal{C}\left(\tilde{c} - c_d\right) =: K(y_{old}, p_{old}, w_{old}, w_{old}) \end{aligned} \tag{25}$$

and is inserted in (12).

Note that this system does not depend on α since $u_0^k = 0$.

The spatially discrete treatment of (12)–(15) with this control is similar as in the uncontrolled case. We note that (24) motivates to use Taylor-Hood finite elements for the discretization of the control, see the concept of variational discretization proposed in [13].

5 Numerical Results - Circle2Square

To demonstrate the effectiveness of our control method we now morph a circle into a square with the following setup. As domain we use $\Omega := (0,1)^2$, the initial concentration c_0 is chosen as 1 in $B_{\frac{1}{4}}(\frac{1}{2},\frac{1}{2})$ and as -1 in $\Omega \setminus B_{\frac{1}{4}}(\frac{1}{2},\frac{1}{2})$, see Fig. 1 (left). Control is applied to the flow field with control gain of steering c to the desired state c_d with values 1 in the square centered at $(\frac{7}{20},\frac{7}{20})$, see Fig. 1 (right), with edge length such that $(c_0, 1) = (c_d, 1)$. This requirement is meaningful, since our time discretization scheme is mass-conserving. We choose $Re = 10$, $Pe = 100$, $\gamma = 1/(40\pi)$, $\alpha = 1e-4$, $\tau = 0.01$ and use $\rho = 1$ as step size in the steepest descent method.

Fig. 2 presents snapshots of the concentration after 40 time steps (left) and after 500 time steps (right, where also the controlled flow is depicted). Clearly, the corners formed in c_d cannot be reached by the controlled concentration, since

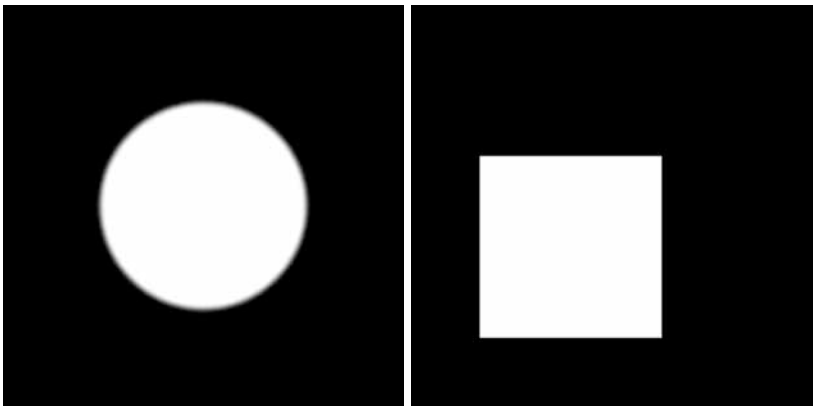


Fig. 1. Initial state c_0 (left) and desired state c_d (right). Black indicates -1, white indicates $+1$.

the phasefield approach used here always delivers a smooth diffuse interface in the present situation. Fig. 3 presents the evaluation of $d(t) := \|c(t) - c_d(t)\|_{L^2(\Omega)}$ for various Reynolds numbers ranging in the interval $[10, 200]$. We observe a clear decrease in $d(t)$, till a certain value around $1e-1$ is reached. The method is not able to further reduce $d(t)$ due to the unreachability of c_d. The oscillations of d for larger Reynolds numbers can be explained by the indirect control method (flow is controlled, concentration should be steerd to c_d), which becomes more sensible with increasing Reynolds number.

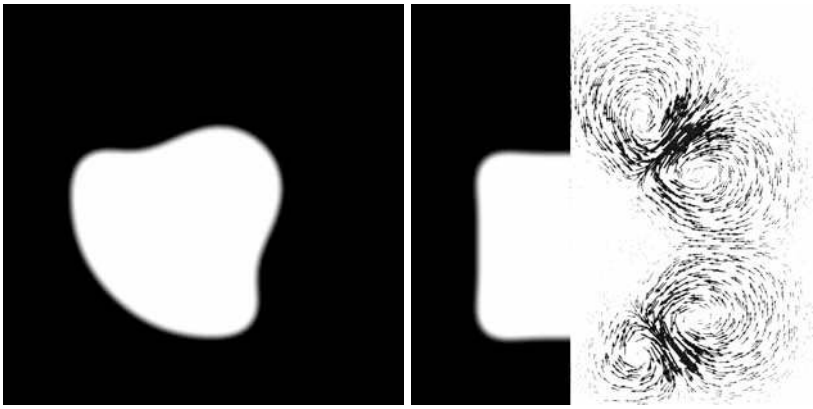


Fig. 2. Controlled state at $t = 40\tau$ and $t = 500\tau$

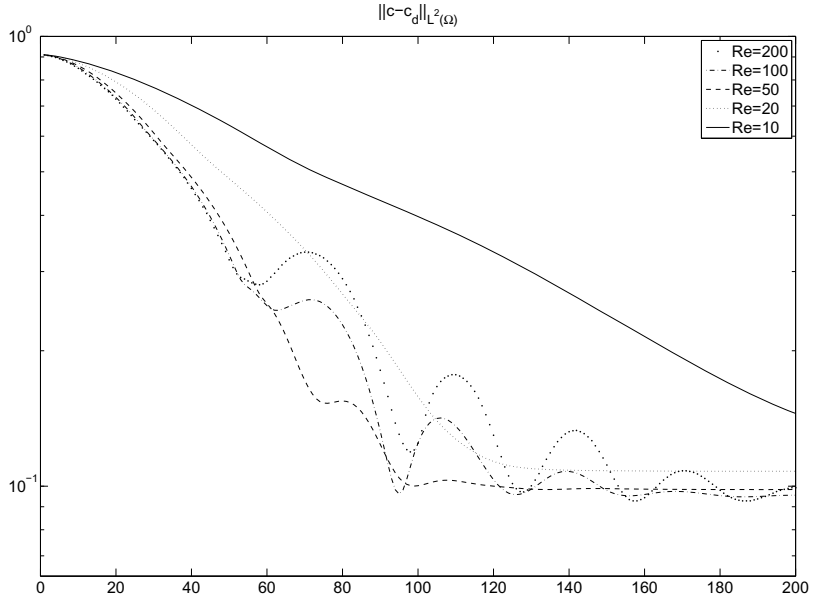


Fig. 3. The reduction of $\|c - c_d\|$ for various Reynolds numbers

6 Conclusion

We have presented a general inexact model predictive control concept called instantaneous control and sketched its interpretation as closed loop controller. For the Cahn-Hilliard Navier-Stokes system we have derived a nonlinear feedback $u = K(y, p, c, w)$ which realizes instantaneous control on the continuous level. With morphing the circle into a square we have numerically demonstrated the scope and effectiveness of our approach in control of two-phase flows.

References

1. Abels, H.: Diffuse interface models for two-phase flows of viscous incompressible fluids. Max-Planck Institut für Mathematik in den Naturwissenschaften, Leipzig, Lecture Note no.: 36 (2007)
2. Blowey, J.F., Elliott, C.M.: The Cahn-Hilliard gradient theory for phase separation with non-smooth free energy. Part I: Mathematical analysis. European Journal of Applied Mathematics 2, 233–280 (1991)
3. Boyer, F., Lapuerta, C., Minjeaud, S., Piar, B., Quintard, M.: Cahn-Hilliard/Navier-Stokes model for the simulation of three-phase flows. Transp. Porous Media 82(3), 463–483 (2010)
4. Cahn, J.W., Hilliard, J.E.: Free energy of a nonuniform system-I: Interfacial free energy. J. Chem. Phys. 28, 258–267 (1958)

5. Choi, H.: Suboptimal control of turbulent flow using control theory. In: Proceedings of the International Symposium on Mathematical Modelling of Turbulent Flows, Tokyo, Japan (1995)
6. Choi, H., Hinze, M., Kunisch, K.: Instantaneous control of backward-facing step flows. Applied Numerical Mathematics 31(2), 133–158 (1999)
7. Eyre, D.J.: Unconditionally gradient stable time marching the Cahn-Hilliard equation. In: Computational and Mathematical Models of Microstructural Evolution. MRS Proceedings, vol. 529 (1998)
8. Gross, S., Reusken, A.: Numerical methods for two-phase incompressible flows. Springer Series in Computational Mathematics, vol. 40. Springer (2011)
9. Grüne, L., Pannek, J.: Nonlinear Model Predictive Control. Communications and Control Engineering. Springer (2011)
10. Hintermüller, M., Hinze, M., Kahle, C.: An adaptive finite element Moreau-Yosida-based solver for a coupled Cahn-Hilliard/Navier-Stokes system (preprint, submitted for publication)
11. Hintermüller, M., Hinze, M., Tber, M.H.: An adaptive finite element Moreau-Yosida-based solver for a non-smooth Cahn-Hilliard problem. Optimization Methods and Software 25(4-5), 777–811 (2011)
12. Hinze, M.: Instantaneous closed loop control of the Navier-Stokes system. SIAM J. Contr. Optim. 44(2), 564–583 (2005)
13. Hinze, M.: A variational discretization concept in control constrained optimization: the linear quadratic case. Computational Optimization and Applications 30(1), 45–61 (2005)
14. Hinze, M., Kunisch, K.: Three control methods for time - dependent Fluid Flow. Flow, Turbulence and Combustion, vol. 60, pp. 273–298. Springer (2000)
15. Hinze, M., Pinnau, R., Ulbrich, M., Ulbrich, S.: Optimization with PDE constraints. Springer, Heidelberg (2009)
16. Hinze, M., Volkwein, S.: Instantaneous control for the Burgers equation: Convergence analysis and numerical implementation. Nonlinear Analysis 50(1), 1–26 (2002)
17. Hohenberg, P.C., Halperin, B.I.: Theory of dynamic critical phenomena. Rev. Mod. Phys. 49(3), 435–479 (1977)
18. Hood, P., Taylor, G.: Navier-Stokes equations using mixed interpolation. Finite Element Methods in Flow Problems. UAH Press (1974)
19. Kay, D., Styles, V., Welford, R.: Finite element approximation of a Cahn-Hilliard-Navier-Stokes system. Interfaces and Free Boundaries 10(1), 15–43 (2008)
20. Nevistic, V., Primbs, J.A.: Finite Receding Horizon Control: A General Framework for Stability and Performance Analysis. Technical Report 6, Automatic control laboratory, ETH Zürich (1997)
21. Qin, S.J., Badgwell, T.A.: A survey of industrial model predictive control technology. Control Engineering Practice 11, 733–764 (2003)
22. Verfürth, R.: A posteriori error analysis of space-time finite element discretizations of the time-dependent Stokes equations. Calcolo 47, 149–167 (2010)

Embedding Domain Technique for a Fluid-Structure Interaction Problem

Cornel Marius Murea[1] and Andrei Halanay[2]

[1] Laboratoire de Mathématiques, Informatique et Applications,
Université de Haute Alsace, France
cornel.murea@uha.fr
http://www.edp.lmia.uha.fr/murea/
[2] Department of Mathematics 1,
University Politehnica of Bucharest, Romania
halanay@mathem.pub.ro

Abstract. We present a weak formulation for a steady fluid-structure interaction problem using an embedding domain technique with penalization. Except of the penalizing term, the coefficients of the fluid problem are constant and independent of the deformation of the structure, which represents an advantage of this approach. A second advantage of this model is the fact that the continuity of the stress at the fluid-structure interface does not appear explicitly. Numerical results are presented.

Keywords: fluid-structure interaction, embedding domain.

1 A Steady Fluid-Structure Interaction Problem

The present paper is devoted to the study of the numerical behavior of an elastic structure immersed in a viscous incompressible fluid. We use Stokes equation to model the flow motion. The displacement of the structure under the flow motion will be modeled by linear elasticity equations, under the small deformations assumption. In this paper, we study the steady case.

Let $D \subset \mathbb{R}^2$ be a bounded open domain with boundary ∂D. Let Ω_0^S be the undeformed structure domain, and suppose that its boundary admits the decomposition $\partial\Omega_0^S = \Gamma_D \cup \Gamma_0$, where Γ_0 is a relatively open subset of the boundary. On Γ_D we impose zero displacement for the structure. We assume that $\Omega_0^S \subset D$.

Suppose that the structure is elastic and denote by $\mathbf{u} = (u_1, u_2) : \Omega_0^S \to \mathbb{R}^2$ its displacement. A particle of the structure with initial position at the point $\mathbf{X}$ will occupy the position $\mathbf{x} = \varphi(\mathbf{X}) = \mathbf{X} + \mathbf{u}(\mathbf{X})$ in the deformed domain $\Omega_u^S = \varphi\left(\Omega_0^S\right)$.

We assume that $\Omega_u^S \subset D$ and the fluid occupies $\Omega_u^F = D \setminus \overline{\Omega}_u^S$. We set $\Gamma_u = \varphi(\Gamma_0)$ and we suppose that Γ_u does not touch the container wall, i.e. $\partial D \cap \Gamma_u = \emptyset$. We recall that Γ_0 is a relatively open subset. The boundary Γ_u represents the moving fluid-structure interface. The boundary of the deformed

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 358–367, 2013.

structure is $\partial\Omega_u^S = \Gamma_D \cup \Gamma_u$. In the case when $\overline{\Omega}_u^S \subset D$, the fluid-structure geometrical configuration is represented in Figure 1 and the boundary of the fluid domain admits the decomposition $\partial\Omega_u^F = \partial D \cup \Gamma_D \cup \Gamma_u$.

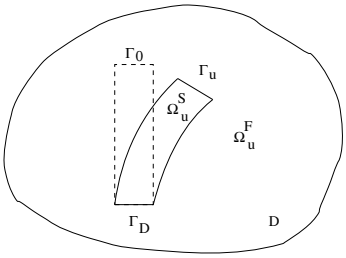


Fig. 1. Geometrical configuration

Also, for the first numerical test, we consider the case when $\Gamma_D \subset \partial D$ as in Figure 2 and the boundary of the fluid domain admits the decomposition $\partial\Omega_u^F = (\partial D \setminus \Gamma_D) \cup \Gamma_u$.

2 Weak Formulation Using an Embedding Domain Technique with Penalization

We introduce the tensor $\epsilon(\mathbf{w}) = \frac{1}{2}\left(\nabla\mathbf{w} + (\nabla\mathbf{w})^T\right)$ and we assume that the fluid is Newtonian and the Cauchy stress tensor is given by $\sigma^F(\mathbf{v}, p) = -p\,\mathbf{I} + 2\mu^F\epsilon(\mathbf{v})$, where $\mu^F > 0$ is the viscosity of the fluid and $\mathbf{I}$ is the unit matrix. We assume that the structure verifies the linear elasticity equation, under the assumption of small deformations. The stress tensor of the structure written in the Lagrangian framework is $\sigma^S(\mathbf{u}) = \lambda^S(\nabla\cdot\mathbf{u})\,\mathbf{I} + 2\mu^S\epsilon(\mathbf{u})$, where $\lambda^S, \mu^S > 0$ are the Lamé coefficients.

We present in an informal and intuitive manner the ideas behind our approximation approach using embedding domain technique with penalization. In the fluid domain, Stokes equations are solved:

$$-\nabla\cdot\sigma^F(\mathbf{v}, p) = \mathbf{f}^F, \quad \text{in } \Omega_u^F \tag{1}$$

$$\nabla\cdot\mathbf{v} = 0, \quad \text{in } \Omega_u^F \tag{2}$$

We introduce two more equations concerning the fluid fields, but written on the deformed structure domain:

$$-\nabla\cdot\sigma^F(\mathbf{v}, p) + \frac{1}{\varepsilon}\mathcal{P}(\mathbf{v}) = \mathbf{f}^F, \quad \text{in } \Omega_u^S \tag{3}$$

$$\nabla\cdot\mathbf{v} = 0, \quad \text{in } \Omega_u^S \tag{4}$$

where $\varepsilon > 0$ is a penalization parameter,

$$\mathcal{P}(\mathbf{v}) = \left(|v_1|^{\alpha-1} sgn(v_1), |v_2|^{\alpha-1} sgn(v_2) \right) \tag{5}$$

where $\mathbf{v} = (v_1, v_2)$ and $1 < \alpha < 2$ is a real number.

Remark 1. This choice of the penalization term is justified in [2], in order to obtain existence of the fluid-structure interaction problem. For the steady case, the role of the penalization term is to obtain very small values of the fluid velocity in the structure domain. If we take other values for α or for a H^1 penalization, we do not get more regularity of the solution.

Let χ_u^S be the characteristic function of Ω_u^S. Combining (1) and (3), it follows that

$$-\nabla \cdot \sigma^F(\mathbf{v}, p) + \frac{1}{\varepsilon}\chi_u^S \mathcal{P}(\mathbf{v}) = \mathbf{f}^F, \quad \text{in } D. \tag{6}$$

Similarly, we have from (2) and (4)

$$\nabla \cdot \mathbf{v} = 0, \quad \text{in } D. \tag{7}$$

In view of the equation (3), the "fictitious" fluid velocity and pressure defined on the structure domain Ω_u^S depend on ϵ. In the following, we denote by $\mathbf{v}_\varepsilon$ and p_ε the fluid velocity and pressure defined all over the domain D.

Let us introduce the bi-linear forms

$$a_S(\mathbf{u}, \mathbf{w}^S) = \int_{\Omega_0^S} \left(\lambda^S (\nabla \cdot \mathbf{u}) (\nabla \cdot \mathbf{w}^S) + 2\mu^S \epsilon(\mathbf{u}) : \epsilon(\mathbf{w}^S) \right) d\mathbf{X}$$

$$a_F(\mathbf{v}, \mathbf{w}) = \int_D 2\mu^F \epsilon(\mathbf{v}) : \epsilon(\mathbf{w})\, d\mathbf{x}$$

$$b_F(\mathbf{w}, p) = -\int_D (\nabla \cdot \mathbf{w})\, p\, d\mathbf{x}$$

and the Hilbert spaces

$$W^S = \left\{ \mathbf{w}^S \in \left(H^1(\Omega_0^S)\right)^2;\ \mathbf{w}^S = 0 \text{ on } \Gamma_D \right\},$$

$$W = \left(H_0^1(D)\right)^2,$$

$$Q = L_0^2(D) = \{ q \in L^2(D);\ \int_D q\, dx = 0 \}.$$

We assume for the moment that $\mathbf{f}^F \in \left(L^2(D)\right)^2$, $\mathbf{f}^S \in \left(L^2(\Omega_0^S)\right)^2$ and $\mathbf{g} \in \left(H^{1/2}(\partial D)\right)^2$, such that $\int_{\partial D} \mathbf{g} \cdot \mathbf{n}^F ds = 0$.

For a given $\mathbf{u} \in \left(W^{1,\infty}(\Omega_0^S)\right)^2$, such that $\|\mathbf{u}\|_{1,\infty,\Omega_0^S} < 1$ and $\mathbf{u} = 0$ on Γ_D, we define:

- fluid velocity $\mathbf{v}_\varepsilon \in \left(H^1(D)\right)^2$, $\mathbf{v}_\varepsilon = \mathbf{g}$ on ∂D,
- fluid pressure $p_\varepsilon \in Q$,
- structure displacement $\mathbf{u}_\varepsilon \in W^S$,

as the solution of the following weakly coupled system of PDE's:

$$a_F(\mathbf{v}_\varepsilon, \mathbf{w}) + b_F(\mathbf{w}, p_\varepsilon) + \frac{1}{\varepsilon}\int_D \widetilde{H}_u \mathcal{P}(\mathbf{v}_\varepsilon)\cdot \mathbf{w}\, d\mathbf{x} = \int_D \mathbf{f}^F \cdot \mathbf{w} d\mathbf{x}, \quad \forall \mathbf{w} \in W \tag{8}$$

$$b_F(\mathbf{v}_\varepsilon, q) = 0, \quad \forall q \in Q \tag{9}$$

$$a_S(\mathbf{u}_\varepsilon, \mathbf{w}^S) = \int_{\Omega_0^S} \mathbf{f}^S \cdot \mathbf{w}^S\, d\mathbf{X} + \int_{\Omega_0^S} J\left(\sigma^F(\mathbf{v}_\varepsilon, p_\varepsilon)\circ\varphi\right)\mathbf{F}^{-T} : \nabla_{\mathbf{X}}\mathbf{w}^S\, d\mathbf{X} + \frac{1}{\varepsilon}\int_{\Omega_0^S} J\widetilde{H}_u \mathcal{P}(\mathbf{v}_\varepsilon\circ\varphi)\cdot\mathbf{w}^S\, d\mathbf{X} - \int_{\Omega_0^S} J\left(\mathbf{f}^F\circ\varphi\right)\cdot\mathbf{w}^S\, d\mathbf{X}, \ \forall \mathbf{w}^S \in W^S \tag{10}$$

where $\varphi(\mathbf{X}) = \mathbf{X} + \mathbf{u}(\mathbf{X})$, $\mathbf{F}(\mathbf{X}) = \mathbf{I} + \nabla_{\mathbf{X}}\mathbf{u}(\mathbf{X})$, $J(\mathbf{X}) = \det \mathbf{F}(\mathbf{X})$.

The equations (8) and (9) are obtained from (6) and (7). The coefficient $\widetilde{H}_u$ in (8) is a regularization of the characteristic function of Ω_u^S, which is necessary in order to prove the continuity of the solution with respect to the structure displacement.

Remark 2. From the structure equation $-\nabla\cdot\sigma^S(\mathbf{u}_\varepsilon) = \mathbf{f}^S$, in Ω_0^S using Green's formula, we obtain for all $\mathbf{w}^S = 0$ on Γ_D that

$$a_S(\mathbf{u}_\varepsilon, \mathbf{w}^S) = \int_{\Omega_0^S} \mathbf{f}^S\cdot\mathbf{w}^S\, d\mathbf{X} + \int_{\Gamma_0} \sigma^S(\mathbf{u}_\varepsilon)\,\mathbf{n}^S\cdot\mathbf{w}^S dS.$$

We can prove (see [2]) that the sum of the last three terms in (10) is equal to the fluid forces acting on the structure which is also equals to $\int_{\Gamma_0}\sigma^S(\mathbf{u}_\varepsilon)\,\mathbf{n}^S\cdot\mathbf{w}^S dS$. In fact, from (10) and the above weak formulation of the structure, we can get that the boundary condition at the interface concerning the continuity of the stress is verified in a weak sense (see [2]). The second boundary condition at the interface is the continuity of the velocity, i.e. $\mathbf{v} = 0$ on Γ_u in the steady case. This is obtained by using the penalization term in the structure domain.

Define the nonlinear operator

$$T_\varepsilon : \left\{\mathbf{u}\in\left(W^{1,\infty}(\Omega_0^S)\right)^2;\ \|\mathbf{u}\|_{1,\infty,\Omega_0^S} < 1,\ \mathbf{u} = 0 \text{ on } \Gamma_D\right\} \to \left(W^{1,\infty}(\Omega_0^S)\right)^2$$

by $T_\varepsilon(\mathbf{u}) = \mathbf{u}_\varepsilon$. A solution of the penalized fluid-structure interaction problem will be, by definition, a fixed point of T_ε. In [2], we discuss the existence of a solution of the penalized fluid-structure interaction problem. The convergence of $\mathbf{u}_\varepsilon,\ \mathbf{v}_\varepsilon,\ p_\varepsilon$ when ε goes to 0 is also analyzed.

3 Partitioned Procedures Based on Fixed Point Iterations

The penalized term $\mathcal{P}(\mathbf{v})$ is non-linear in $\mathbf{v}$ for $\alpha \neq 2$. But for $\alpha = 2$, we have $\mathcal{P}(\mathbf{v}) = \mathbf{v}$. Now, the fluid problem, at the **Step 2** of the algorithm below, becomes linear and, for a given $\mathbf{u}_\epsilon^k$, it has a unique solution.

For $1 < \alpha < 2$, we can prove the existence of a fixed point for the nonlinear operator T_ε defined at the end of the previous section, but not for $\alpha = 2$. We can also replace $\widetilde{H}_u$ in (8) by χ_u^S the characteristic function of Ω_u^S in order to simplify the computation. The regularization of the the characteristic function was necessary in order to prove the continuity of the solution with respect to the structure displacement.

Under the assumption of small displacements for the structure, we can approach the Jacobian determinant J by 1 and the gradient of the deformation $\mathbf{F}$ by the identity matrix $\mathbf{I}$. The structure problem at the **Step 3** is linear and, for given $\mathbf{v}_\epsilon^k$ and p_ϵ^k, it has a unique solution.

Algorithm.

Step 1. Given the initial displacement of the structure $\mathbf{u}^0 \in W^S$, compute the characteristic function $\chi_{u^0}^S$, put $k := 0$.

Step 2. Find the velocity $\mathbf{v}_\varepsilon \in \left(H^1(D)\right)^2$, $\mathbf{v}_\varepsilon = \mathbf{g}$ on ∂D and the pressure $p_\varepsilon^k \in Q$ by solving the fluid problem

$$a_F\left(\mathbf{v}_\varepsilon^k, \mathbf{w}\right) + b_F\left(\mathbf{w}, p_\varepsilon^k\right) + \frac{1}{\varepsilon}\int_D \chi_{u_\varepsilon^k}^S \mathbf{v}_\varepsilon^k \cdot \mathbf{w}\, d\mathbf{x} = \int_D \mathbf{f}^F \cdot \mathbf{w}\, d\mathbf{x},\ \forall \mathbf{w} \in W$$
$$b_F\left(\mathbf{v}_\varepsilon^k, q\right) = 0,\ \forall q \in Q.$$

Step 3. Find the new displacement of the structure $\mathbf{u}_\varepsilon^{k+1} \in W^S$ by solving

$$a_S\left(\mathbf{u}_\varepsilon^{k+1}, \mathbf{w}^S\right) = \int_{\Omega_0^S} \left(\mathbf{f}^S - \mathbf{f}^F\right) \cdot \mathbf{w}^S\, d\mathbf{x} + \int_{\Omega_0^S} 2\mu^F \epsilon\left(\mathbf{v}_\varepsilon^k\right) : \epsilon\left(\mathbf{w}^S\right)\, d\mathbf{x}$$
$$- \int_{\Omega_0^S} \left(\nabla \cdot \mathbf{w}^S\right) p_\varepsilon^k\, d\mathbf{x} + \frac{1}{\varepsilon}\int_{\Omega_0^S} \left(\mathbf{v}_\varepsilon^k \circ \varphi_\varepsilon^k\right) \cdot \mathbf{w}^S\, d\mathbf{x} \quad \forall \mathbf{w}^S \in W^S$$

where $\varphi_\varepsilon^k(\mathbf{X}) = \mathbf{X} + \mathbf{u}_\varepsilon^k(\mathbf{X})$.

Step 4. Stopping test: if $\left\|\mathbf{u}_\varepsilon^k - \mathbf{u}_\varepsilon^{k+1}\right\|_{0,\Omega_0^S} \leq tol$, then **Stop**.

Step 5. Compute the characteristic function $\chi_{u_\varepsilon^{k+1}}^S$, put $k := k + 1$ and **Go to Step 2.**

It is possible to consider Navier-Stokes equations for the fluid domain, then at the **Step 2** we have to solve a non-linear system. In this case we can use the non-linear penalization term for $\alpha \neq 2$. If we use a non-linear model for the structure, too, we have to use more accurate approximations for the the Jacobian determinant J and the gradient of the deformation $\mathbf{F}$.

4 Numerical Tests

The numerical tests have been produced using the software *FreeFem++* [3].

4.1 Test 1. Shell in Steady-State Cross Flow

First, we have performed numerical simulation using a 2D model adapted from [1] (see Figure 2) where we have changed the physical parameters of the fluid and of the structure.

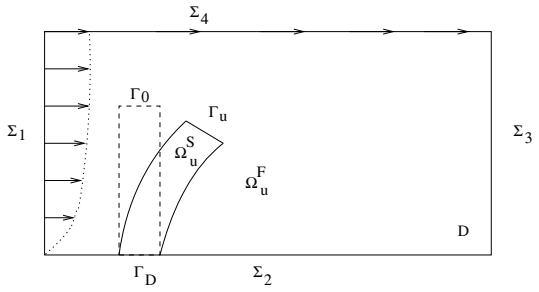


Fig. 2. Geometrical configuration for the Test 1

The dimensions of a rectangular elastic structure are: height $\ell = 3\ m$, thickness $h = 0.125\ m$. The computational domain of the fluid D is a rectangle of height $H = 5\ m$ and length $L = 12\ m$. The distance between the left side of the fluid and the left side of the structure is $2\ m$. The lower left corner of Figure 2 is $(x_1 = 0, x_2 = 0)$.

We denote by Σ_1, Σ_3 the left and the right vertical boundaries of D and by Σ_2, Σ_4 the bottom and the top boundaries of D, respectively.

The mechanical proprieties of the structure (polybutadiene) are: Young modulus $E^S = 1.6 \times 10^6\ N/m^2$, Poisson's ratio $\nu^S = 0.49$, the applied volume forces on the structure $\mathbf{f}^S : \Omega_0^S \to \mathbb{R}^2$, $\mathbf{f}^S = (0, 0)\ N/m^3$.

The dynamic viscosity of the fluid (glycerin) is $\mu^F = 1.14\ N \cdot s/m^2$.

The inflow velocity profile on Σ_1 is

$$v_1(x_1, x_2) = V \times 1.5 \frac{\left(2H\, x_2 - x_2^2\right)}{H^2}\ m/s,\ V = 1, \quad v_2(x_1, x_2) = 0.$$

The other boundary conditions are: $\mathbf{v} = \mathbf{0}$ (no-slip) on Σ_2, $\mathbf{v} \cdot \mathbf{n}^F = 0$ (slip) on Σ_4 and $\mathbf{v} \times \mathbf{n}^F = 0$, $\mathbf{n}^F \cdot \left(\sigma^F\left(\mathbf{v}, p\right)\mathbf{n}^F\right) = 0$ (the tangential velocity and the normal traction are zero) on Σ_3.

We use a fixed mesh for the fluid domain of 13096 triangles and 6719 vertices. The mesh of the structure domain has 188 triangles and 145 vertices. The fluid and structure meshes are not compatible, for example, a vertex on the structure boundary is not necessary a vertex on the fluid mesh. For the approximation of the fluid velocity and pressure we have employed the triangular finite elements $\mathbb{P}_1$+*bubble* and $\mathbb{P}_1$ respectively, also called "mini" finite elements. The finite element $\mathbb{P}_1$ was used in order to solve the structure problem. The characteristic function was approached by $\mathbb{P}_0$ finite elements.

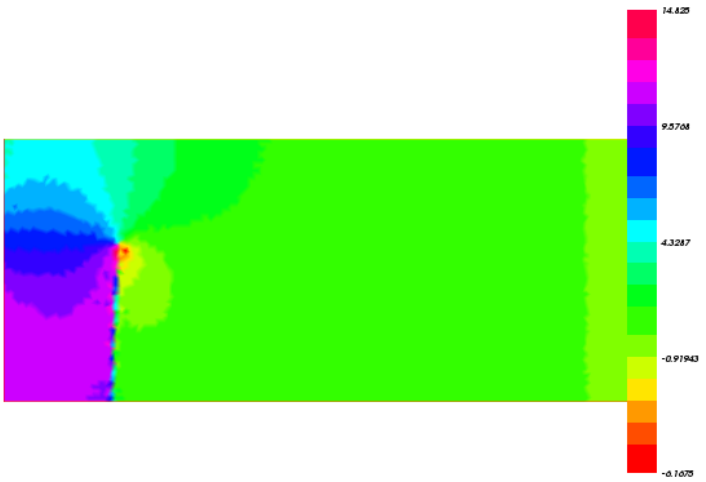


Fig. 3. The fluid pressure [Pa]

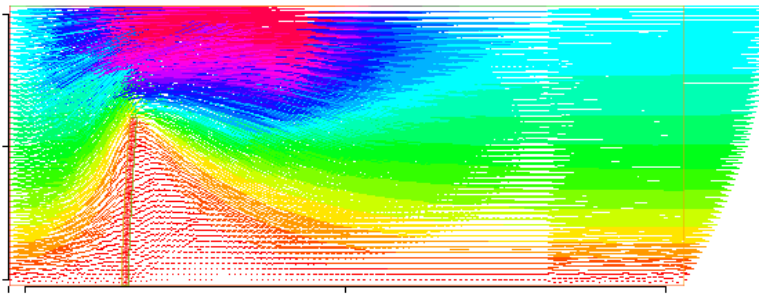


Fig. 4. The fluid velocity around the final position of the structure. In each point of the grid, there is an arrow giving the direction of the velocity and the length of the arrow is proportional to the euclidean norm of the velocity. The maximal value for the horizontal component v_1 is 2.82 m/s and for the vertical component v_2 is 1.18 m/s.

We have performed the simulation using the **Algorithm** described in the previous section. For the stopping criterion at the **Step 4**, we have used the tolerance $tol = 0.2 \times 10^{-4}$. The penalization parameter is $\varepsilon = 10^{-3}$. The stopping criterion holds after 5 iterations of the fixed point algorithm. The maximal structural displacement is 0.14 m.

The fluid velocity into the fictitious domain is very small

$$\|\mathbf{v}_\varepsilon\|_{0,\Omega^S_{u_\varepsilon}} = \sqrt{\int_D \chi^S_{u_\varepsilon} \mathbf{v}_\varepsilon \cdot \mathbf{v}_\varepsilon \, d\mathbf{x}} = 0.0703384.$$

4.2 Test 2. Flexible Appendix in a Flow

We have adapted the benchmark from [4]. Originally, the structure was placed horizontally, parallel to the flow, but the displacements in this case are very small. We have placed the structure vertically, transversely to the flow, see Figure 5.

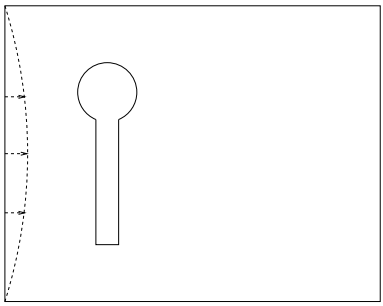


Fig. 5. Geometrical configuration

The structure is composed by a rectangular flexible appendix attached to a fixed circle. The circle center is positioned at $(0.2, 0.2)\ m$ measured from the left top corner of the channel. The circle has the radius $r = 0.5\ m$ and the rectangular appendix is of length $\ell = 0.35\ m$, thickness $h = 0.02\ m$. The Young modulus is $E^S = 1.6 \times 10^6\ N/m^2$ and Poisson's ratio is $\nu^S = 0.49$ (polybutadiene).

The channel has the length $L = 2.5\ m$ and the width $H = 0.75\ m$. The fluid dynamic viscosity is $\mu^F = 1.420\ N \cdot s/m^2$ (glycerin).

We have used the following boundary conditions: at the inflow the velocity is

$$v_1(x_1, x_2) = V \times 1.5 \frac{(H\, x_2 - x_2^2)}{(H/2)^2}\ m/s,\ V = 1, \quad v_2(x_1, x_2) = 0;$$

at the bottom and the top we have imposed the no-slip boundary condition $\mathbf{v} = \mathbf{0}$ and at the outflow the traction free $\sigma^F(\mathbf{v}, p)\,\mathbf{n}^F = 0$.

We use a fixed mesh for the fluid domain of 30330 triangles and 15461 vertices and a structure mesh of 128 triangles and 97 vertices. We have employed the same finite elements as for the Test 1. We have treated by the embedding domain technique only the flexible part of the structure.

The penalization parameter is $\varepsilon = 10^{-4}$ and for the stopping criterion $tol = 10^{-8}$. The fixed point algorithm stops after 8 iterations.

The maximal horizontal displacement of the structure is $0.10886\ m$. The pressure and the velocity of the fluid are presented in Figures 6 and 7. The fluid velocity in the fictitious domain is very small

$$\|\mathbf{v}_\varepsilon\|_{0,\Omega^S_{u_\varepsilon}} = \sqrt{\int_D \chi^S_{u_\varepsilon} \mathbf{v}_\varepsilon \cdot \mathbf{v}_\varepsilon\, d\mathbf{x}} = 0.080920.$$

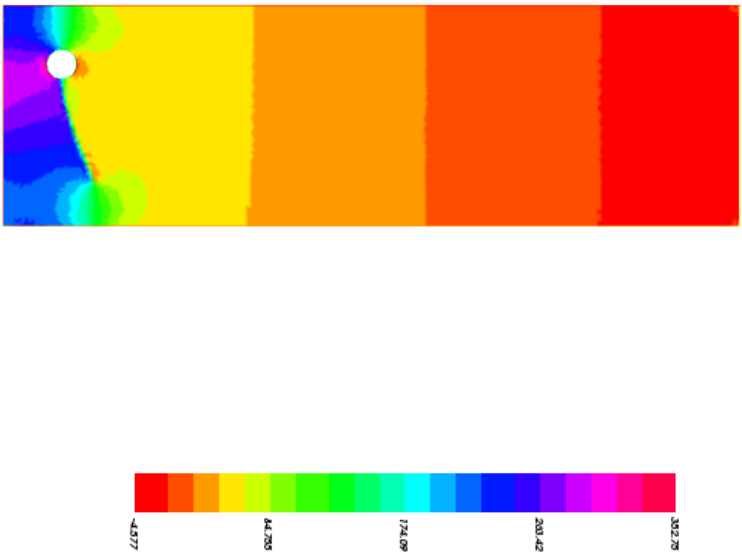


Fig. 6. The fluid pressure [Pa]

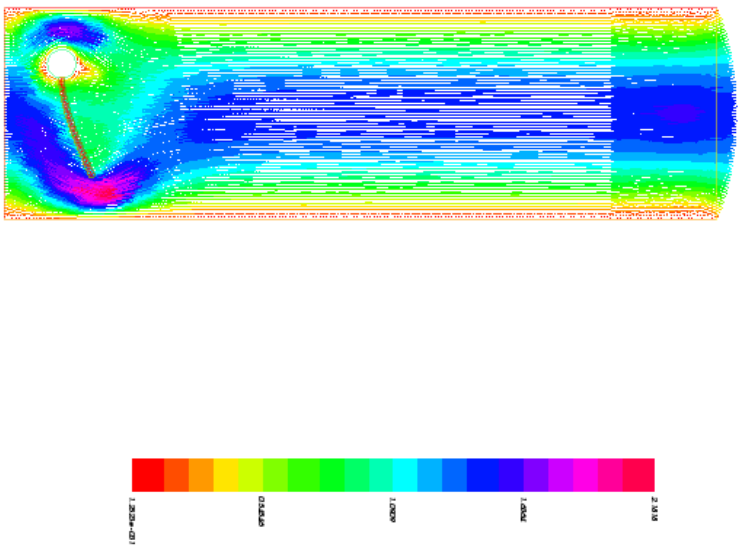


Fig. 7. The fluid velocity [m/s] around the final position of the structure. In each point of the grid, there is an arrow giving the direction of the velocity. The length of the arrow is proportional to the euclidean norm of the velocity which is represented in the color bar.

5 Conclusions

We have presented a fixed point algorithm for solving steady fluid-structure interaction problem. Using the embedding domain technique with penalization, the fluid equations as well as the structure equations are solved in fixed meshes. The fluid and structure meshes could be generated independently. The algorithm can be used for the three dimensions problems.

Acknowledgment. The second author gratefully acknowledges support by Grant CNCS Romania 145/2011.

References

1. Bathe, K.-J., Ladezma, G.: Benchmark problems for incompressible fluid flows with structural interactions. Comput. & Structures 85, 628–644 (2007)
2. Halanay, A., Murea, C.M., Tiba, D.: Existence and approximation for a steady fluid-structure interaction problem using fictitious domain approach with penalization, accepted for publication in Mathematics and its Applications
3. Hecht, F.: FreeFem++, http://www.freefem.org
4. Turek, S., Hron, J.: Proposal for numerical benchmarking of fluid-structure interaction between an elastic object and laminar incompressible flow. In: Bungartz, H.-J., Schfer, M. (eds.) Fluid-Structure Interaction - Modelling, Simulation, Optimization. Lect. Notes Comput. Sci. Eng., vol. 53, pp. 371–385. Springer, Berlin (2006)

Note on Level Set Functions

Piotr Fulmański and Alicja Miniak-Górecka

Faculty of Mathematics and Computer Science, University of Łódź
Banacha 22, 90-238 Łódź, Poland
fulmanp@math.uni.lodz.pl, alicja_miniak@wp.pl

Abstract. In this note a concept of ε-level set function is introduced, i.e. a function which approximates a level set function satisfying the Hamilton-Jacobi inequality. We prove that each Lipschitz continuous solution of the Hamilton-Jacobi inequality is an ε-level set function. Next, a numerical approximation of the level set function is presented, i.e. method for the construction of an ε-level set function.

Keywords: level set function, numerical approximation, shape optimization.

1 Introduction

The goal of shape optimization is to deform and modify the admissible shapes in order to comply with a given cost function that needs to be optimized.

Let $D \subset R^n$ be a given bounded domain and $\Omega_t \subset D$, be a sets from a family of admissible shapes Θ, indexed by t from some set of indexes. Assume that a certain functional $J(\cdot)$ reaches its minimum value on the set Ω_t for a x_t^{min} function.

Consider the following shape optimization problem: find a set $\Omega_{opt} \in \Theta$, for which there exists a function x_{opt}^{min} such that the following formula holds

$$J_{\Omega_{opt}}(x_{opt}^{min}) \leq J_{\Omega_t}(x_t^{min}) \qquad \Omega_t \in \Theta,$$

that is

$$J(\Omega_{opt}) = \inf_{\Omega \in \Theta} J(\Omega).$$

Problem formulated this way is difficult to solve – the crucial part is the construction of the family Θ so a known mathematical methods could be used. While solving this problem we were inspired by the approach we found in paper [1], where minimization over a family of sets is turned into a minimization over functions. Following this idea e.g. the level set function could be used to connect sets with functions – it allows us to manipulate boundary of the given shape through the level set function. A very brief sketch of this approach (transformation from optimization over domains into optimization over functions) is given in section 2.1. Notice that whenever a computation is mentioned, it means that, due to numerical computations limits, we are able to find only an approximate

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 368–377, 2013.

solution to a given problem. This is why, in practice, when solving shape optimization problem with the help of level set functions, only an approximation of it could be used and this is the main aim of this paper: to present a numerical approximation of the level set function, i.e. we want to present a method for the construction of an ε-level set function.

2 Level Set Method

Let Ω be an open and connected subset of D for which there exists a continuous function $\Psi(x) : D \to R$ such that $\Omega = \{x \in D : \Psi(x) < 0\}$. In consequence, the boundary Γ of Ω is a set of all points $x \in D$, such that $\Psi(x) = 0$. Let $\phi : (t, x) \in [0, 1] \times D \to R$ be any function of class C^1, such that

$$\phi(0, x) = \Psi(x), \quad x \in D.$$

If Ω is a subject to changes in time we can describe Ω and its boundary Γ at time t (denoted as Ω_t and Γ_t) as

$$\Omega_t(\phi) = \{x \in D : \phi(t, x) < 0\}$$

and

$$\Gamma_t(\phi) = \{x \in D : \phi(t, x) = 0\}.$$

Let $x : [0, 1] \times \Gamma(0) \to D$ be a continuous function, which for every point $x_0 \in \Gamma(0)$ assigns its location at time t, $t \in [0, 1]$, i.e. $x(t, x_0) = x \in \Gamma(t)$. Function $x(\cdot, x_0)$ represents the location of the point x_0 at successive time steps t, determining a trajectory starting from the point $x_0 \in \Gamma_0$. For fixed starting point x_0, a trajectory represents the movement of this point. Taking all points $x_0 \in \Gamma_0$ into account we have the movement of a given boundary of Ω. This is why we call a trajectory starting at x_0 a *deformation of the point* x_0. We call the family of trajectories for all points $x_0 \in \Gamma_0$ the *deformation of the initial domain* Ω.

Let $V_n(x(t, x_0))$, $t \in [0, 1]$, $x_0 \in \Gamma_0(\phi)$ be a Lipschitz mapping assigning to every point $x(y, x_0)$ its speed of movement in a normal direction to the boundary $\Gamma_t(\phi)$. A well known level set formula (e.g. [4]) according to which the changes of the function $\phi(t, \cdot)$ affect the boundary Γ_t takes the following form

$$\frac{\partial \phi}{\partial t}(t, x(t, x_0)) + |\nabla \phi(t, x(t, x_0))| \, V_n(x(t, x_0)) = 0 \tag{1}$$

Thus ϕ has to satisfy the following equation of Hamilton-Jacobi type

$$\frac{\partial \phi}{\partial t}(t, x) + |\nabla \phi(t, x)| \, V_n(x) = 0, \qquad (t, x) \in (0, 1) \times D$$

with initial condition

$$\phi(0, x) = \Psi(x), \quad x \in D. \tag{2}$$

2.1 Problem Reformulation

Denote by

$$\mathit{\Gamma} = \left\{\Psi :\ \Psi \in C(\bar{\Omega}) \text{ and } \Psi = 0 \text{ on } \partial\Omega_\Psi,\ \bar{\Omega}_\Psi \subset \bar{\Omega},\ \partial\Omega_\Psi\text{-smooth}\right\}$$

Put $\Omega_t(\phi_t) = (x \in \Omega \mid \phi_t < 0)$. The family Θ of sets over which our shape optimization problem is considered can be defined as

$$\Theta = \{\Omega_t(\phi_t) : t \in [0,1],\ \phi_t = \phi(t,\cdot),\ \phi \in Lips([0,1], \bar{\Omega}_\Psi)\}$$

where ϕ satisfies (1) in Ω_Ψ, $\Psi \in \mathit{\Gamma}$, with boundary condition (2) on $\partial\Omega_\Psi$, $\Psi \in \mathit{\Gamma}$. Define a new family

$$\Phi = \{\phi_t : \phi_t = \phi(t,\cdot),\ \Omega_t(\phi_t) \in \Theta,\ t \in [0,1]\}$$

Now we can reformulate the shape optimization problem to the following problem

$$J(\Omega_{opt}) = \inf_{\phi_t \in \Phi} J(\phi_t).$$

2.2 ε - Level Set Function

However, from the practical point of view only an approximate solution to (1) is considered, i.e. a solution $\phi_\varepsilon(\cdot,\cdot)$, which instead of an equality satisfies an inequality

$$-\varepsilon \le \frac{\partial \phi_\varepsilon}{\partial t}(t, x(t,x_0)) + |\nabla \phi_\varepsilon(t, x(t,x_0))|\, V_n(x(t,x_0)) \le 0. \tag{3}$$

Therefore, instead of a level set function we have its approximation. We call a function $(t,x) \to \phi_\varepsilon(t,x)$, defined in $[0,1] \times \Omega$, an $\varepsilon-$ level set function if

$$-\varepsilon \le \phi_\varepsilon(t, x(t,x_0)) \le 0, \quad (t,x_0) \in [0,1] \times \partial\Omega, \tag{4}$$

$$\Psi(x) \le \phi_\varepsilon(0,x) \le \Psi(x) + \varepsilon/2, \quad x \in \bar{\Omega}. \tag{5}$$

It is also well known that there exists a Lipschitz continuous $\varepsilon-$ level set function and that it satisfies the Hamilton - Jacobi inequality

$$-\varepsilon \le \frac{\partial \phi_\varepsilon}{\partial t}(t,x) + |\nabla \phi_\varepsilon(t,x)|\, V_n(t,x) \le 0 \tag{6}$$

and initial condition (5). We have the following theorem, which is very important from the numerical point of view.

Theorem 1. *Each element of the set* W_ε,

$$W_\varepsilon = \Big\{ w(t,x) \text{ is Lipschitz: } -\frac{\varepsilon}{2} \le w(0,x) \le 0,\ x \in \partial\Omega;$$
$$-\frac{\varepsilon}{2} \le \frac{\partial}{\partial t} w(t,x) + |\nabla w(t,x)|\, V_n(t,x) \le 0,\ \text{a.a. } (t,x) \in [0,1] \times \Omega \Big\}$$

is an $\varepsilon-$ *level set function, i.e. it satisfies* (4)-(5).

Proof. We use the ideas of a proof from [2]. Let t_0, $0 < t_0 \leq 1$ and $\delta > 0$, be such that the interval $[\delta, t_0 - \delta]$ is nonempty; let x_0, $x_0 \in \partial\Omega$, be an arbitrary initial value and let the $x(t), t \in [0, t_0 - \delta]$ start at x_0. Of, course, by assumptions the values of $x(t)$ are then bounded on $[\delta, t_0 - \delta]$, i.e. there is some compact set Q such that $x(t) \in Q$, $t \in [\delta, t_0 - \delta]$. Let $B_\tau(R^n)$ be a ball in R^n (with Euclidean norm) with radius $\tau \in R$ and center at 0. Denote $Q_1 = Q + B_1(R^n)$. Take $w \in W_\varepsilon$ and $\alpha \in R$ such that $0 < \alpha < \varepsilon/4$ and define

$$w_1(t,x) = w(t,x) + \alpha(t-1),\ (t,x) \in [0,1] \times \Omega.$$

Then, for a.a. $(t,x) \in [0,1] \times \Omega$, w_1 satisfies

$$\alpha - \frac{\varepsilon}{2} \leq \frac{\partial}{\partial t} w_1(t,x) + |\nabla w_1(t,x)|\, V_n(t,x)\} \leq \alpha.$$

Let us choose $0 < \beta_0 < \min\{1, \delta\}$ and define a function $(t,x) \to w_2^{\beta_0}(t,x)$ on $[\delta, t_0 - \delta] \times Q$ by the convolution $w_2^{\beta_0}(t,x) = (w_1 * \rho_{\beta_0})(t,x)$ where

$$\rho_{\beta_0}(t,x) = \frac{1}{\beta_0^{n+1}} \rho_1\left(\frac{t}{\beta_0}, \frac{x}{\beta_0}\right),$$

$$\int_{R^{n+1}} \rho_1(t,x)dtdx = 1,\ \text{supp}\ \rho_1 \subset B_1(R^{n+1}).$$

We claim that there exists $\beta' > 0$ such that for $\beta \leq \beta'$ and $(t,x) \in [\delta, t_0 - \delta] \times Q$,

$$\frac{1}{2}\alpha - \frac{\varepsilon}{2} \leq \frac{\partial}{\partial t} w_2^\beta(t,x) + \left|\nabla w_2^\beta(t,x)\right| V_n(t,x) \leq \frac{3}{2}\alpha. \tag{7}$$

Indeed, since $w_1(t,x)$ is Lipschitz continuous, there exists M, such that $\left|\frac{\partial}{\partial x} w_1\right| \leq M$ and

$$\begin{aligned}
&\left|\left|\nabla w_2^\beta(t,x)\right| V_n(t,x) - \left(\left(|\nabla w_1(\cdot,\cdot)|\, V_n(\cdot,\cdot)\right) * \rho_\beta\right)(t,x)\right| \\
&\leq \int\limits_{B_\beta(R^{n+1})} |\nabla w_1(t-s, x-y)|\, |V_n(t,x) - V_n(t-s, x-y, u)|\, \rho_\beta(s,y)dsdy \\
&\leq M \sup_{\substack{(t,x)\in[\delta,t_0-\delta]\times Q \\ (s,y)\in B_\beta(R^{n+1})}} |V_n(t,x) - V_n(t-s, x-y)|\,.
\end{aligned}$$

The right-hand side of the inequality presented above tends to zero as $\beta \to 0$ and that on $[\delta, t_0 - \delta] \times Q$, there is $\beta_2 > 0$ such that for $\beta \leq \beta_2$,

$$\left|\left|\nabla w_2^\beta(t,x)\right| V_n(t,x) - \left(\left(|\nabla w_1(\cdot,\cdot)|\, V_n(\cdot,\cdot)\right) * \rho_\beta\right)(t,x)\right| < \frac{\alpha}{2}.$$

Let us put on $[\delta, t_0 - \delta] \times Q$,

$$\begin{aligned}
F(t,x) &= \frac{\partial}{\partial t} w_2^\beta(t,x) + \left|\nabla w_2^\beta(t,x)\right| V_n(t,x) \\
&= \left(\left(\frac{\partial w_1}{\partial t}(\cdot,\cdot) + |\nabla w_1(\cdot,\cdot)|\, V_n(\cdot,\cdot)\right) * \rho_\beta\right)(t,x) \\
&+ \left|\nabla w_2^\beta(t,x)\right| V_n(t,x) - \left(\left(|\nabla w_1(\cdot,\cdot)|\, V_n(\cdot,\cdot)\right) * \rho_\beta\right)(t,x).
\end{aligned}$$

Considering the above estimations, we can find $0 < \beta' \leq \min\{\beta_0, \beta_1, \beta_2\}$ such that for $\beta \leq \beta'$,

$$\begin{aligned}\frac{1}{2}\alpha - \frac{\varepsilon}{2} &\leq \left(\left(\alpha - \frac{\varepsilon}{2}\right) * \rho_\beta\right)(t,x) - \frac{\alpha}{2} \leq F(t,x) \\ &\leq (\alpha * \rho_\beta)(t,x) + \frac{\alpha}{2} = \frac{3}{2}\alpha, \text{ for } (t,x) \in [\delta, t_0 - \delta] \times Q.\end{aligned}$$

It is clear that $w_2^\beta(\cdot,\cdot)$ is $C^\infty([\delta, t_0 - \delta] \times Q)$ and the function $(t,x) \to F(t,x)$, is continuous on $[\delta, t_0 - \delta] \times Q$. After integrating the inequalities (7) in $[\delta, t_0 - \delta]$ and considering the definition of V_n,

$$\begin{aligned}&\int_\delta^{t_0-\delta} \left(\frac{1}{2}\alpha - \frac{\varepsilon}{2}\right) dt \\ &\leq \int_\delta^{t_0-\delta} \left(\frac{\partial}{\partial t} w_2^\beta(t,x) + \left\{\left|\nabla w_2^\beta(t,x)\right| V_n(t,x)\right\}\right) dt \\ &\leq \int_\delta^{t_0-\delta} \frac{3}{2}\alpha dt.\end{aligned} \tag{8}$$

As a consequence of (8), the following are obtained:

$$\left(\frac{1}{2}\alpha - \frac{\varepsilon}{2}\right)(t_0 - 2\delta) \leq \left(\int_\delta^{t_0-\delta} \frac{d}{dt} w_2^\beta(t, x(t)) dt\right) \leq \frac{3}{2}\alpha(t_0 - 2\delta)$$

and

$$\begin{aligned}\left(\frac{1}{2}\alpha - \frac{\varepsilon}{2}\right)(t_0 - 2\delta) &\leq w_2^\beta(t_0 - \delta, x(t_0 - \delta)) \\ -w_2^\beta(\delta, x(\delta)) &\leq \frac{3}{2}\alpha(t_0 - 2\delta).\end{aligned} \tag{9}$$

By the property of convolution, we see that $w_2^\beta \to w_1$ uniformly on $[\delta, t_0 - \delta] \times Q$ and thus (9) leads to

$$\begin{aligned}\left(\frac{1}{2}\alpha - \frac{\varepsilon}{2}\right)(t_0 - 2\delta) &\leq w_1(t_0 - \delta, x(t_0 - \delta)) \\ -w_1(\delta, x(\delta)) &\leq \frac{3}{2}\alpha(t_0 - 2\delta).\end{aligned}$$

Taking the limit with $\alpha \to 0$, we obtain

$$-\frac{\varepsilon}{2}(t_0 - 2\delta) \leq w(t_0 - \delta, x(t_0 - \delta)) - w(\delta, x(\delta)) \leq 0$$

Since δ was chosen arbitrarily and

$$-\frac{\varepsilon}{2} \leq w(0, x_0) \leq 0,$$

we infer further that

$$-\varepsilon \leq w(t_0, x(t_0) \leq 0.$$

Since t_0 and $w \in W_\varepsilon$ were chosen arbitrarily, the theorem is proved.

In this section, we proved that each Lipschitz continuous solution of the Hamilton-Jacobi inequality is an ε-level set function. As a direct conclusion of the theorem we infer the following corollary.

Corollary 1. *Let $\varepsilon_n > 0$, $\varepsilon_n \to 0$. Then each sequence of ε_n-level set functions $w_{\varepsilon_n} \in W_{\varepsilon_n}$ tends uniformly to the level set function $\phi(t,x)$ on $[0,1] \times \Omega$.*

3 Numerical Approximation

In the second part of this paper we want to present a numerical approximation of the level set function for the equation (1), i.e. we want to present a method for the construction a ε-level set function for the equation (3), which satisfies (4) and (5). In order to achieve that, an adaptation of the method developed by J. Pustelnik in his Ph.D. thesis [3] is used.

Let $T \subset [0,1] \times D$ be a compact set and $(t,x) \to w(t,x)$ be a function defined on set T', $T \subset T'$ of class $C^2(T')$ such that

$$-\frac{\varepsilon}{2} \leq w(0,x) \leq 0,\ x \in \partial\Omega.$$

For $w(\cdot,\cdot)$ define now on the set T a new function $(t,x) \to F_w(t,x)$, corresponding to the left hand side of the formula (1)

$$F_w(t,x) := \frac{\partial w}{\partial t}(t,x) + |\nabla w(t,x)|\, V_n(x). \tag{10}$$

Function $(t,x) \to F_w(t,x)$ is a continuous function on T. Moreover it is also a Lipschitz function on T let M_{F^w} be a Lipschitz constant for the function $F_w(\cdot,\cdot)$. Owing to the compactness of T, function $F_w(\cdot,\cdot)$ reaches its lower and upper limits denoted respectively as k_l and k_u.

Let $\eta > 0$ be any fixed real number and $\{y_j^\eta\}_{j\in\mathbb{Z}}$ a sequence of numbers such that $y_0^\eta = 0$ and $y_{j+1}^\eta - y_j^\eta = \eta$ for $j \in \mathbb{Z}$. Define a new set J

$$J := \{j \in \mathbb{Z} :\ \exists_{(t,x)\in T}\ y_j^\eta < F_w(t,x) \leq y_{j+1}^\eta\}.$$

and let $P_T = \{P_j^{\eta,w}\}_{j\in J}$ be a family of sets covering the set T where

$$P_j^{\eta,w} := \left\{(t,x) \in T : y_j^\eta < F_w(t,x) \leq y_{j+1}^\eta\right\}$$

As a consequence of the definition of the family P_T and uniform continuity of the function $F_w(\cdot,\cdot)$ on the set T we have the following proposition

Proposition 1. *There exists a real number $\varepsilon > 0$, such that for every point $(t,x) \in T$ a ball with radius ε centered in (t,x) is covered either by one set $P_j^{\eta,w}$, $j \in J$ or by two sets $P_{j_1}^{\eta,w}$, $P_{j_2}^{\eta,w}$, $j_1, j_2 \in J$ and $|j_1 - j_2| = 1$.*

Let $h^{\eta,w}(\cdot,\cdot)$ be a function defined on T as follows

$$h^{\eta,w}(t,x) := -y_{j+1}^\eta \text{ for } (t,x) \in P_j^{\eta,w}, \quad j \in J. \tag{11}$$

As a consequence of the above definition we have

$$\underset{(t,x)\in T}{\forall} -\eta \leq F_w(t,x) + h^{\eta,w}(t,x) \leq 0. \tag{12}$$

Lemma 1. *Let $x_w(\cdot, x_0)$ be a deformation of any point x_0. There exists an increasing sequence of m points $\{t_i\}_{i=1,\ldots,m}$, $t_1 = 0$ and $t_m = 1$ such that*

$$\underset{t\in[t_i,t_{i+1}]}{\forall} |F_w(t_i, x_w(t_i, x_0)) - F_w(t, x_w(t, x_0))| \leq \frac{\eta}{2},\ i = 1, \ldots, m-1. \tag{13}$$

Proof. This is a simple consequence of the absolute continuity of $x_w(\cdot, \cdot)$.

Notice, that Lemma 1 holds for any $\tau \in [0, 1]$, since for any $\tau \in [0, 1]$ there exists an increasing sequence of m_τ points $\{t_i^\tau\}_{i=1,\ldots,m_\tau}$, where $t_1^\tau = 0$ and $t_{m_\tau}^\tau = \tau$, for which the following formula holds

$$\underset{t\in[t_i,t_{i+1}]}{\forall} |F_w(t_i, x_w(t_i, x_0)) - F_w(t, x_w(t, x_0))| \leq \frac{\eta}{2},\ i = 1, \ldots, m_\tau - 1.$$

Moreover, having the aforementioned sequence for $\tau = 1$, we can easily determine a sequence for any $\tau = \{t_i^1\}_{i=1,\ldots,m_1}$. As a consequence of the formula (13) we have that for any $i \in \{1, \ldots, m_\tau - 1\}$, if $(t_i, x_w(t_i, x_0)) \in P_j^{\eta,w}$ for some $j \in J$, than for every $x_0 \in \Gamma_0$ the following property holds

$$\underset{t\in[t_i,t_{i+1}]}{\forall} (t, x_w(t, x_0)) \in P_{j-1}^{\eta,w} \cup P_j^{\eta,w} \cup P_{j+1}^{\eta,w}.$$

From the above and Definition (11) for all $t \in [t_i, t_{i+1}]$ we get that

$$[h^{\eta,w}(t_i, x_w(t_i, x_0)) - \eta \leq h^{\eta,w}(t, x_w(t, x_0)) \leq h^{\eta,w}(t_i, x_w(t_i, x_0)) + \eta]\,. \tag{14}$$

Particularly for every $i \in \{2, \ldots, m_\tau - 1\}$

$$h^{\eta,w}(t_i, x_w(t_i, x_0)) - h^{\eta,w}(t_{i-1}, x_w(t_{i-1}, x_0)) = \eta^i_{x_w(\cdot,x_0)}, \tag{15}$$

where $\eta^i_{x_w(\cdot,\cdot)} \in \{-\eta, 0, \eta\}$. Integration of (14) results, for any $i \in \{1, \ldots, m_\tau - 1\}$, in the following double inequality

$$\begin{aligned}
&[h^{\eta,w}(t_i, x_w(t_i, x_0)) - \eta]\,(t_{i+1} - t_i) \\
&\leq \int_{t_i}^{t_{i+1}} h^{\eta,w}(t, x_w(t, x_0))dt \leq [h^{\eta,w}(t_i, x_w(t_i, x_0)) + \eta]\,(t_{i+1} - t_i)
\end{aligned}$$

and in consequence

$$\begin{aligned}
&\sum_{i\in 1,\ldots,m_\tau-1} [h^{\eta,w}(t_i, x_w(t_i, x_0))(t_{i+1} - t_i)] - \eta\tau \\
&\leq \int_0^\tau h^{\eta,w}(t, x_w(t, x_0))dt \\
&\leq \sum_{i\in 1,\ldots,m_\tau-1} [h^{\eta,w}(t_i, x_w(t_i, x_0))(t_{i+1} - t_i)] + \eta\tau.
\end{aligned} \tag{16}$$

Owing to the fact, that by simple calculation the expression

$$\sum_{i\in 1,\ldots,m_\tau-1} [h^{\eta,w}(t_i, x_w(t_i, x_0))(t_{i+1} - t_i)]$$

can be substituted by some sum of differences (15), finally formula (16) takes the following form

$$\begin{aligned}&\sum_{i\in 2,\ldots,m_\tau-1}\eta^i_{x_w(\cdot,x_0)}(\tau-t_i)+h^{\eta,w}(0,x_w(0,x_0))\tau-\eta\tau\\&\leq\int_0^\tau h^{\eta,w}(t,x_w(t,x_0))dt\\&\leq\sum_{i\in 2,\ldots,m_\tau-1}\eta^i_{x_w(\cdot,x_0)}(\tau-t_i)+h^{\eta,w}(0,x_w(0,x_0))\tau+\eta\tau.\end{aligned}\tag{17}$$

Notice that inequality (17) is very useful in computation. It allows estimation of an integral of function $h^{\eta,w}(\cdot,\cdot)$ along deformation $x_w(\cdot,x_0)$ as a finite sum of values from the set $\{-\eta,0,\eta\}$. Moreover for any two deformations of two different points $x_0^1\in\Gamma_0$ and $x_0^2\in\Gamma_0$ values

$$\sum_{i\in 2,\ldots,m_\tau-1}\eta^i_{x_w(\cdot,x_0^1)}(\tau-t_i)+h^{\eta,w}(0,x_w(0,x_0^1))\tau$$

and

$$\sum_{i\in 2,\ldots,m_\tau-1}\eta^i_{x_w(\cdot,x_0^2)}(\tau-t_i)+h^{\eta,w}(0,x_w(0,x_0^2))\tau$$

are equal if the following conditions hold

$$\eta^i_{x_w(\cdot,x_0^1)}=\eta^i_{x_w(\cdot,x_0^2)}\text{ for every }i\in\{2,\ldots,m_\tau-1\},\tag{18}$$

$$x_0^1\in P_j^{\eta,w}\text{ i }x_0^2\in P_j^{\eta,w},j\in J.\tag{19}$$

In consequence, in the set K of all deformations $x_w(\cdot,x_0)$, $x_0\in\Gamma_0$ an equivalence relation E can be introduced, taking as an equivalent any two deformations $x(\cdot,x_0^1)$ and $x(\cdot,x_0^2)$, $x_0^1,x_0^2\in\Gamma_0$ fulfills (18) and (19). The cardinality of a set K_E of all disjoint equivalence class of relation E is finite and limited from above by value $3^{m_\tau-1}$. Now define a set X of $m_\tau-1$-dimensional vectors $x=(x_1,\ldots,x_{m_\tau-1})$, where $x_1=0$ and $x_i=\eta^i_{x_w^j},\quad i=2,\ldots,m_\tau-1$, while $x_w^j\in X_E$ is any element of j-th equivalence class, $i=1,\ldots,||K_E||$. Inequality (17) can be rewritten as

$$\begin{aligned}&\sum_{i\in 1,\ldots,m_\tau-1}x^i_{x_w(\cdot,x_0)}(\tau-t_i)+h^{\eta,w}(0,x_w(0,x_0))\tau-\eta\tau\\&\leq\int_0^\tau h^{\eta,w}(t,x_w(t,x_0))dt\\&\leq\sum_{i\in 1,\ldots,m_\tau-1}x^i_{x_w(\cdot,x_0)}(\tau-t_i)+h^{\eta,w}(0,x_w(0,x_0))\tau+\eta\tau.\end{aligned}\tag{20}$$

Thus infinite space of all deformation can be reduced to the finite set.

Lemma 2. *If x_0 is any point from Γ_0 and $\tau \in [0,1]$ then we have the following inequality*

$$\begin{aligned}
&-\sum_{i\in 2,\ldots,m_\tau-1} x^i_{x_w(\cdot,x_0)}(\tau - t_i) - h^{\eta,w}(0, x_w(0,x_0))\tau - 2\eta\tau \\
&\leq w(\tau, x(\tau)) - w(0,x_0) \\
&\leq -\sum_{i\in 2,\ldots,m_\tau-1} x^i_{x_w(\cdot,x_0)}(\tau - t_i) - h^{\eta,w}(0, x_w(0,x_0))\tau + \eta\tau.
\end{aligned}$$

Proof. Integration of (12) along any deformation $x(\cdot, x_0)$ on interval $[0,\tau]$ gives

$$-\eta\tau - \int_0^\tau h^{\eta,w}(t,x)dt \leq \int_0^\tau F_w(t,x)dt \leq -\int_0^\tau h^{\eta,w}(t,x)dt,$$

and in consequence

$$\begin{aligned}
&-\eta\tau - \int_0^\tau h^{\eta,w}(t,x)dt \\
&\leq \int_0^\tau \left(\frac{\partial w}{\partial t}(t, x(t,x_0)) + |\nabla w(t, x(t,x_0))|\, V_n(t, x(t,x_0))\right) dt \\
&\leq -\int_0^\tau h^{\eta,w}(t,x)dt.
\end{aligned}$$

Considering equation (17) we have

$$\begin{aligned}
&-\sum_{i\in 2,\ldots,m_\tau-1} x^i_{x_w(\cdot,x_0)}(\tau - t_i) - h^{\eta,w}(0, x_w(0,x_0))\tau - 2\eta\tau \\
&\leq \int_0^\tau \left(\frac{\partial w}{\partial t}(t, x(t,x_0)) + |\nabla w(t, x(t,x_0))|\, V_n(t, x(t,x_0))\right) dt \\
&\leq -\sum_{i\in 2,\ldots,m_\tau-1} x^i_{x_w(\cdot,x_0)}(\tau - t_i) - h^{\eta,w}(0, x_w(0,x_0))\tau + \eta\tau,
\end{aligned}$$

and finally because

$$\begin{aligned}
&\int_0^\tau \left(\frac{\partial w}{\partial t}(t, x(t,x_0)) + |\nabla w(t, x(t,x_0))|\, V_n(t, x(t,x_0))\right) dt \\
&= \int_0^\tau \frac{d}{dt} w(t, x(t,x_0))dt
\end{aligned}$$

we have

$$\begin{aligned}
&-\sum_{i\in 2,\ldots,m_\tau-1} x^i_{x_w(\cdot,x_0)}(\tau - t_i) - h^{\eta,w}(0, x_w(0,x_0))\tau - 2\eta\tau \\
&\leq w(\tau, x(\tau)) - w(0,x_0) \\
&\leq -\sum_{i\in 2,\ldots,m_\tau-1} x^i_{x_w(\cdot,x_0)}(\tau - t_i) - h^{\eta,w}(0, x_w(0,x_0))\tau + \eta\tau.
\end{aligned}$$

Theorem 2. *Set a real number $\eta > 0$, point $x_0 \in \partial\Omega$ and $\tau \in [0,1]$. Then*

$$w(\tau, x(\tau)) - w(0, x_0) + \sum_{i \in 2,\ldots,m_\tau - 1} x^i_{x_w(\cdot,x_0)}(\tau - t_i) + h^{\eta,w}(0, x_w(0, x_0))\tau - \eta\tau$$

is a value of some ε-level set function at the point $(\tau, x(\tau, x_0))$ for $\varepsilon = 3\eta\tau$.

Proof. From the Theorem 2 we obtain

$$\begin{aligned} -3\eta\tau \leq\ & w(\tau, x(\tau)) - w(0, x_0) \\ & + \sum_{i \in 2,\ldots,m_\tau - 1} x^i_{x_w(\cdot,x_0)}(\tau - t_i) + h^{\eta,w}(0, x_w(0, x_0))\tau - \eta\tau \\ & \leq 0. \end{aligned}$$

From the above we infer that it is enough to take into account a finite number of points from $\partial\Omega$ to get the approximation of the level set function with an error not greater than 3η.

References

1. Myśliński, A.: Radial Basis Function Level Set Method for Structural Optimization (to be published)
2. Nowakowski, A.: ε -Value Function and Dynamic Programming. Journal of Optimization Theory and Applications 138(1), 85–93 (2008)
3. Pustelnik, J.: Approximation of optimal value for Bolza problem. Ph.D. thesis (2009) (in Polish)
4. Sethian, J.A.: Level Set Methods. Cambridge University Press (1996)

Fixed Domain Algorithms in Shape Optimization for Stationary Navier-Stokes Equations

Andrei Halanay[1] and Cornel Marius Murea[2]

[1] Department of Mathematics and Informatics,
University Politehnica of Bucharest,
313 Splaiul Independenţei RO-060042 Bucharest, Romania
halanay@mathem.pub.ro
[2] Laboratoire de Mathématiques, Informatique et Applications,
Université de Haute Alsace 4-6, rue des Frères Lumière,
68093 Mulhouse, France
cornel.murea@uha.fr
http://www.edp.lmia.uha.fr/murea/

Abstract. The paper aims to illustrate the algorithm developed in the paper [6] in some specific problems of shape optimization issued from fluid mechanics. Using the fictitious domain method with penalization, the fluid equations will be solved in a fixed domain. The admissible shapes are parametrized by continuous function defined in the fixed domain, then the shape optimization problem becomes an optimal control problem, where the control is the parametrization of the shape. We get the directional derivative of the cost function by solving co-state equation. Numerical results are obtained using a gradient type algorithm.

Keywords: shape optimization, optimal control, penalization, approximate extension, gradient method.

1 Introduction

The paper presents some applications of an algorithm developed in [6]. This algorithm is based on a method that uses a penalization of the stationary Navier-Stokes equation that approximates its solution by functions defined on a larger fixed domain. The unknown domains are parametrized by functions in a certain subspace of the space of continuous functions on the larger fixed domain.

The approximating extension technique makes possible the approximation of the solution to the shape optimization problem by a solution of an optimal control problem. The basic reference will be [6]. For shape optimization, the general references are [13], [3] and for optimal control [7], [10]. In particular, for shape optimization for fluids a standard work is [9]. In optimal design related to optimal control, relevant contributions to the topic of this paper are [1], [4], [15].

In Section 2, the shape optimization problem for steady Navier-Stokes is presented. The directional derivative of the cost function is given in Section 3. A

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 378–386, 2013.

gradient type algorithm is also introduced. In Section 4, numerical results are presented in order to design a nozzle.

2 Formulation of the Shape Optimization Problem and Approximating Extensions

Let d be a natural number, $d \leq 4$, let $D \subset \mathbf{R}^d$ be a bounded fixed domain and suppose a family $\mathcal{O}$ of admissible subdomains $\Omega \subset D$ is given, satisfying an uniform Lipschitz condition on the boundary $\partial\Omega$.

With the standard notations from [16], $\mathcal{V}(\Omega) = \{y \in \mathcal{D}(\Omega)^d | \text{div } y = 0\}$, $V(\Omega)$ = closure of $\mathcal{V}(\Omega)$ in $H_0^1(\Omega)^d = \{y \in H_0^1(\Omega)^d | \text{div } y = 0\}$ (since $\partial\Omega$ is lipschitzian), we have the following weak formulation of the stationary Navier-Stokes equation with Dirichlet boundary (non-slip) conditions:

$$\int_\Omega \left(\nu \sum_{i,j=1}^{d} \frac{\partial y_j}{\partial x_i} \frac{\partial v_j}{\partial x_i} + \sum_{i,j=1}^{d} y_i \frac{\partial y_j}{\partial x_i} v_j \right) dx = \int_\Omega \left(\sum_{j=1}^{d} f_j v_j \right) dx, \forall v \in V(\Omega) \tag{2.1}$$

or (see [16]), $\nu((y,v))_\Omega + b_\Omega(y,y,v) = \int_\Omega f \cdot v \, dx$. Here $f = (f_1, \ldots, f_d) \in H^{-1}(D)^d$, and $\nu > 0$ is the viscosity.

To this equation we associate the minimization problem

$$\min\{J(\Omega) = \int_E ||y - y_0||_e^2 dx, \quad E \subset \Omega \in \mathcal{O}; y \text{ verifies (2.1)}\} \tag{2.2}$$

$E \subset \Omega$ is a fixed set and $y_0 \in L^2(E)^d$ is given.

This functional is a particular case of a larger class,

$$J(\Omega) = \int_\wedge j(x, y(x), \nabla y(x)) dx, \quad \wedge = E \text{ or } \wedge = \Omega,$$

that are studied in [6].

The uniform Lipschitz assumption turns $\mathcal{O}$ into a compact with respect to the Hausdorff-Pompeiu complementary metric (see [11], p. 466). Based on this, it is inferred in [6] that if there exists an admissible $\hat{\Omega}$ and a corresponding solution of (2.1) for which $J(\hat{\Omega})$ is finite then there exists at least an optimal pair $[\Omega^*, y^*] \in \mathcal{O} \times V(\Omega^*)$. So, the optimization problem is well defined but its solution is generally nonunique.

If $X(D) \subset C(\bar{D})$ is a functional space, define, for $g \in X(D)$, $\Omega = \Omega_g = int\{x \in D | g(x) \geq 0\}$. If $E \subset \Omega$ is to be fulfiled one must require $g(x) \geq 0$ $\forall x \in E$. g is called a parametrization of Ω_g and Ω_g an admissible domain. The solutions of (2.1) in Ω_g will be denoted as y_g.

If $H : \mathbf{R} \to \{0,1\}$ is the Heaviside function, $H \circ g = \chi_{\bar{\Omega}_g}$, the characteristis function of $\bar{\Omega}_g$. For $\varepsilon > 0$ the following smoothing of the Yosida approximation of the maximal monotone extension of H will be used:

$$H^\varepsilon(r) = \begin{cases} 1, & r \geq 0 \\ \dfrac{(\varepsilon - 2r)(r+\varepsilon)^2}{\varepsilon^2}, & -\varepsilon < r < 0 \\ 0, & r \leq -\varepsilon \end{cases}$$

(see also [8], [12]). It is easy to see that $H^\varepsilon \in C^1(\mathbf{R})$ and is lipschitzian. The boundary value problem (2.1) has an approximate extension

$$\nu((y_\varepsilon, v))_D + b_D(y_\varepsilon, y_\varepsilon, v) + \frac{1}{\varepsilon}\int_D [1 - H^\varepsilon(g)] y_\varepsilon \cdot v \, dx = \int_D f \cdot v \, dx \tag{2.3}$$

Suppose now that $d = 3$. It is proved in [6] that, for $C_1 = 9m(D)^{1/6}$, if

$$\nu^2 > C_1 ||f||_{V^*} \tag{2.4}$$

then (2.3) has an unique solution $y_\varepsilon(g)$ that depends continuously on g as a function from $(C(D), ||\ ||_\infty)$ to $L^2(D)^3$. The following theorem that is proved in [6], §3, allows the approximation of the shape optimization problem (2.1), (2.2) by the optimal control problem (2.2), (2.3).

Theorem 2.1. *If* (2.4) *holds then there exists a sequence* $\varepsilon_n \to 0$ *such that* $y_{\varepsilon_n}(g)|_{\Omega_g} \to y_g$ *weakly in* $H^1(\Omega_g)^3$ *and strongly in* $L^2(\Omega_g)^3$.

3 The Directional Derivative and a Gradient Type Algorithm

In order to solve the optimal control problem (2.2), (2.3) through a gradient type algorithm an important step is the calculation of the directional derivative of the mapping $g \mapsto J[y_\varepsilon(g)]$ in the direction $w \in X(D)$. It is proved in [6], §4, that, under the uniqueness condition (2.4), this derivative in direction w, $\dfrac{\partial y_\varepsilon}{\partial w}(g) = (z_1, z_2, z_3) \in V(D)$ is the solution of the equation in variations

$$\begin{aligned} &\int_D \left(\nu \sum_{i,j=1}^3 \frac{\partial z_j}{\partial x_i}\frac{\partial v_j}{\partial x_i} + \sum_{i,j=1}^3 y_{\varepsilon,i}\frac{\partial z_j}{\partial x_i} v_j + \sum_{i,j=1}^3 z_i \frac{\partial y_{\varepsilon,j}}{\partial x_i} v_j \right) dx + \\ &+ \frac{1}{\varepsilon}\int_D [1 - H^\varepsilon(g)] z \cdot v \, dx = \frac{1}{\varepsilon}\int_D ((H^\varepsilon)'(g)w) y_\varepsilon \cdot r \, dx \end{aligned} \tag{3.1}$$

It is also proved in [6], §4, that under condition (2.4), equation (3.1) has an unique solution.

For the optimal control problem (2.2), (2.3) the co-state equation (see [2], [5], [14]) is

$$\int_D \left(\nu \sum_{i,j=1}^{3} \frac{\partial p_{\varepsilon,j}}{\partial x_i} \frac{\partial v_j}{\partial x_i} - \sum_{i,j=1}^{3} y_{\varepsilon,i} \frac{\partial p_{\varepsilon,j}}{\partial x_i} v_j + \sum_{i,j=1}^{3} \frac{\partial y_{\varepsilon,j}}{\partial x_i} p_{\varepsilon,j} v_j \right) dx + \\ + \frac{1}{\varepsilon} \int_D [1 - H^\varepsilon(g)] p_\varepsilon \cdot v \, dx = \int_E (y_\varepsilon - y_0) \cdot v \, dx. \tag{3.2}$$

Under condition (2.4) the equation (3.2) has a unique solution $p_\varepsilon \in V(D)$. The algorithm will result from the following theorem

Theorem 3.1 ([6], §4, Th.5). *The direction derivative in $g \in X(D)$ of $J[y_\varepsilon(g)]$ in the direction $w \in X(D)$ is given by*

$$\frac{\partial J}{\partial w}[y_\varepsilon(g)]w = \frac{1}{\varepsilon} \int_D ((H^\varepsilon)'(g)w) y_\varepsilon \cdot p_\varepsilon dx \tag{3.3}$$

(p_ε is the unique solution of (3.2)).

Algorithm

Step 0. Choose a starting parametrization g_0 and a positive scalar ϵ. Set $k = 0$.

Step 1. Find y_ε the solution of (2.3).

Step 2. Find p_ε the solution of (3.2).

Step 3. Set the descent direction $w_k = -y_\varepsilon \cdot p_\varepsilon$. If $\|w_k\| < tol$ stop.

Step 4. Determine $g_{k+1} = g_k + \theta_k w_k$, $\theta_k > 0$ by means of an approximate minimization

$$J(g_{k+1}) \approx \min_{\theta \geq 0} J(g_k + \theta w_k).$$

Step 5. Update $k = k + 1$ and go to the **Step 1**.

For the inaccurate line search at the **Step 4**, the methods of Goldstein and Armijo were used. If we denote by $j : [0, \infty) \to \mathbb{R}$ the function $j(\theta) = J(g_k + \theta w_k)$, we determine $\theta_k > 0$ such that

$$j(0) + (1 - \lambda)\, \theta_k j'(0) \leq j(\theta_k) \leq j(0) + \lambda \theta_k j'(0) \tag{1}$$

where $\lambda \in (0, 1/2)$.

4 Numerical Results. Shape Optimization of a Nozzle

Problem Setting

We have adapted the nozzle problem from [13]. We assume that the flow in a nozzle is governed by the steady Navier-Stokes equation with prescribed traction at the inflow and outflow. The problem is to design a nozzle that gives a prescribed velocity near the exit. This kind of problem arises in rocket engine

industries, in the design of a spray for applying a coating or in the manufacture of high-resolution inkjet printer.

We assume that the polyhedron $[A_1\,A_2\,A_3\,A_4\,A_5\,A_6\,A_7]$ is the fixed computational domain D. The coordinates of its vertices are: $A_1(H,0)$, $A_2(0,0)$, $A_3(L/2,0)$, $A_4(L,H-h)$, $A_5(L+\ell,H-h)$, $A_6(L+\ell,H)$, $A_7(L,H)$, where $L=6$, $\ell=1$, $H=3$, $h=1$. We denote by E the observation zone which is the rectangle $[A_4\,A_5\,A_6\,A_7]$.

We denote by Σ_{in} the boundary $[A_1\,A_2]$ representing the inflow section and by Σ_{out} the boundary $[A_5\,A_6]$ representing the outflow section. The desired fluid velocity in the observation zone E is

$$y_0=\left(v_{out}\frac{4(H-x_2)(x_2-H+h)}{h^2},\ 0\right),\ \text{where } v_{out}=4.$$

The fluid viscosity is $\mu=1$ and its density is $\rho=1$.

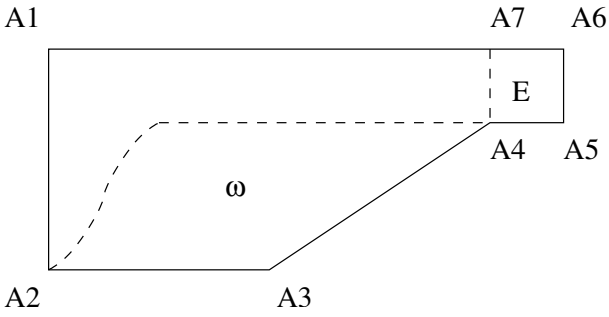


Fig. 1. Computing domain

Let $\omega\subset D$ such that $\omega\cap E=\emptyset$. We look for a connected domain Ω verifying $D\setminus\overline{\omega}\subset\Omega\subset D$ and minimizing the cost function

$$J=\frac{1}{2}\int_E(y_\epsilon-y_0)\cdot(y_\epsilon-y_0)\,dx.$$

The traction imposed on the inflow is $(100,0)$ and on the outflow it is $(0,0)$. We impose no-slip condition on the other boundaries, including the free boundary.

Descent Direction

In Figure 2, we show $1-H^\epsilon(g)$ which is an approximation of the characteristic function of the domain $D\setminus\Omega$, for a typical admissible parametrization g.

We remark that the $(H^\epsilon)'(r)$ vanishes on $\mathbb{R}$, excepting for $r\in(-\epsilon,0)$. Consequently, the zone in D, where $(H^\epsilon)'(g)\neq 0$ is very narrow, see Figure 3. When ϵ is very small, this zone could be empty. For this season, we have taken as descent direction not $-(H^\epsilon)'(g)y_\epsilon\cdot p_\epsilon$ which is given by (3.3), but $w_d=-y_\epsilon\cdot p_\epsilon$. We

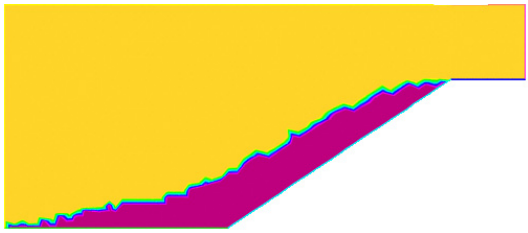


Fig. 2. The value of $1 - H^{\epsilon}(g)$ on D for $\epsilon = 10^{-4}$

Table 1. Mesh parameters used in Figure 3

mesh	no. triangles	no. vertices	mesh size
1	828	460	0.340972
2	3316	1749	0.171328
3	7412	3842	0.146829
4	13168	6765	0.095475

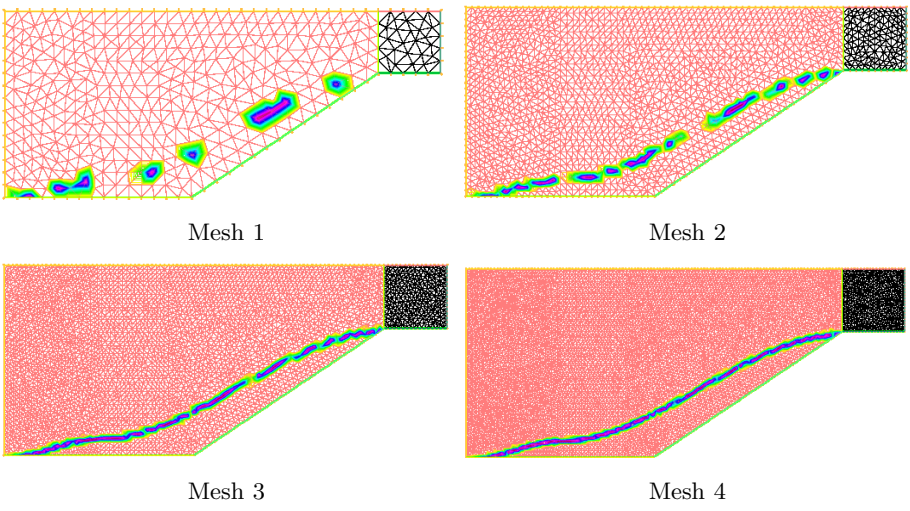


Fig. 3. The zone where the derivative of the Yosida approximation of the Heaviside function is not vanishes for $\epsilon = 10^{-1}$. The mesh parameters are prezented in Table 1.

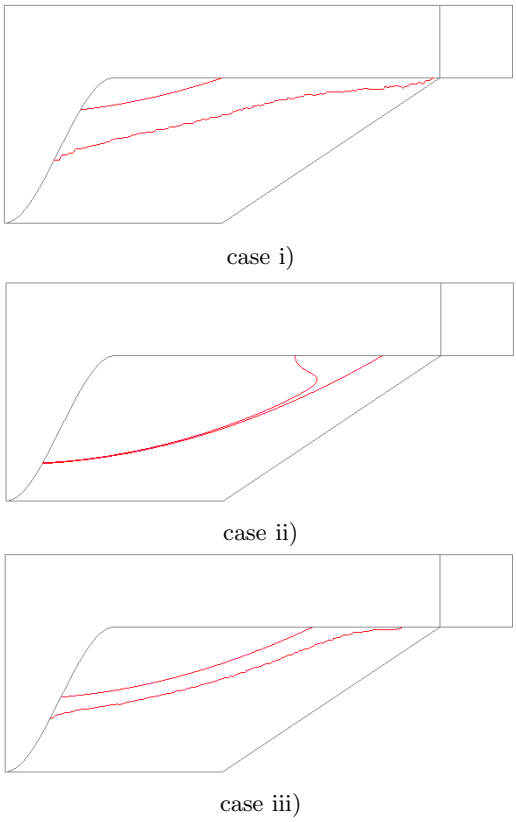


Fig. 4. Case i): final shape (bottom) obtained from the initial shape (top). Case ii): final shape (top) obtained from the initial shape (bottom). Case iii): final shape (bottom) obtained from the initial shape (top).

recall that $(H^\epsilon)'(g) \geq 0$, consequently $w_d = -y_\epsilon \cdot p_\epsilon$ is a descent direction in view of (3.3).

Numerical Parameters

The mesh of D has 15032 triangles and 7697 vertices. We have used the following finite elements: $\mathbb{P}_1 + bubble$ for the velocity, $\mathbb{P}_1$ for the pressure and for the g.

We set the penalization parameter to be $\epsilon = 0.0001$, the number of iterations for the descent algorithm to be 10 and number of iterations for the line search to be 10.

We have tested our algorithm for three initial values of g:

i) for $g(x_1, x_2) = 10^{-4}(x_2 - \frac{x_1^2}{18} - 1.5)$, the initial value of J is 0.518529 and its final value of is 0.00697268;
ii) for $g(x_1, x_2) = 10^{-4}(x_2 - \frac{x_1^2}{18} - 0.5)$, the initial value of J is 0.185802 and its final value of is 0.00466894;

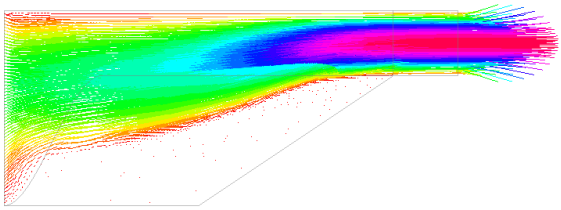


Fig. 5. Fluid velocity for the optimal shape in the case iii)

iii) for $g(x_1, x_2) = 10^{-4}(x_2 - \frac{x_1^2}{18} - 1)$, the initial value of J is 0.0735282 and its final value is 0.00448260.

Numerical Results

We have obtained three different optimal shapes, see Figure 4, that means the algorithm find only local optimum. The minimum final value of the cost function among the three tests is obtained in the case iii).

We remark in Figure 4 case ii), that the zero level set of the initial g partially coincides with the zero level set of the final g. In fact, the initial g vanishes on the free boundary. Since, we impose non-slip boundary condition for y_ϵ and p_ϵ on the free boundary, the descent direction $w_d = -y_\epsilon \cdot p_\epsilon$ vanishes on the free boundary, also. Consequently, g_{k+1} could have the same zero level set as g_k.

The fluid velocity is ploted in Figure 5. We observe that the fluid velocity is very small in the exterior of the optimal shape, more precisely we have

$$Error(g) = \int_\omega (1 - H^\epsilon(g))\, y_\epsilon \cdot y_\epsilon \, dx = 0.000446$$

Acknowledgement. This work was supported by ANCS Grant 480/17.03.2011 in the framework of the "Brâncuşi" programm of cooperation between Romania and France.

References

1. Borrvall, T., Petersson, J.: Topology optimization of fluids in Stokes flow. International Journal for Numerical Methods in Fluids 41, 77–107 (2003)
2. Dede, L.: Optimal flow control for Navier-Stokes equations, drag minimization. J. Numer. Meth. Fluids 44(4), 347–366 (2007)
3. Delfour, M.C., Zolesio, J.P.: Shapes and Geometrics, Analysis, Differential Calculus and Optimization. SIAM, Philadelphia (2001)
4. Gao, Z., Ma, Y.: Optimal shape design for viscous incompressible flow, arxiv: math.oc/0701470v1 (2007)
5. Gunzburger, M.: Adjoint equation-based methods for control problems in viscous, incompressible flows. Flow, Turbul. Comb. 65, 249–272 (2000)
6. Halanay, A., Tiba, D.: Shape optimization for stationary Navier-Stokes equations. Control and Cybernetics 38(4), 1359–1375 (2009)

7. Lions, J.L.: Optimal control of systems governed by partial differential equations. Springer, Berlin (1971)
8. Makinen, R., Neittaanmaki, P., Tiba, D.: On a fixed domain approach for shape optimization problem. In: Ames, W.F., van der Houwer, P.J. (eds.) Computational and Applied Mathematics II: Differential Equations, pp. 317–326. North-Holland, Amsterdam (1992)
9. Mohammadi, B., Pironneau, O.: Applied Shape Optimization for Fluids. Oxford University Press, New York (2001)
10. Neittaanmaki, P., Tiba, D.: Optimal control of nonlinear parabolic systems, Theory, algorithms and applications. Monographs and Textbooks in Pure and Applied Mathematics, vol. 179. Marcel Dekker, New York (1994)
11. Neittaanmaki, P., Sprekels, J., Tiba, D.: Optimization of elliptic systems. Theory and applications. Springer, New York (2006)
12. Neittaanmaki, P., Pennanen, A., Tiba, D.: Fixed domain approaches in shape optimization problems with Dirichlet boundary conditions. J. of Inverse Problems 25, 1–18 (2009)
13. Pironneau, O.: Optimal shape design for elliptic systems. Springer, Berlin (1984)
14. Posta, M., Roubicek, T.: Optimal control of Navier-Stokes equations by Oseen approximations, Necas Center, Prague (2007) (preprint 2007–013)
15. Roubicek, T., Troltzsch, F.: Lipschitz stability of optimal control for steady-state Navier-Stokes equations. Control and Cybernetics 32, 683–705 (2003)
16. Temam, R.: Navier-Stokes equations. Theory and numerical analysis. North-Holland, Amsterdam (1979)

An Electrohydrodynamic Equilibrium Shape Problem for Polymer Electrolyte Membranes in Fuel Cells

Sven-Joachim Kimmerle[1,⋆], Peter Berg[2], and Arian Novruzi[3]

[1] Universität der Bundeswehr München, Institut für Mathematik und Rechneranwendung, Werner-Heisenberg-Weg 39, 85577 Neubiberg, Germany
sven-joachim.kimmerle@unibw.de
[2] Faculty of Science, University of Ontario Institute of Technology, 2000 Simcoe Street N, Oshawa, ON, L1H 7K4, Canada
peter.berg@uoit.ca
[3] Department of Mathematics and Statistics, University of Ottawa, 585 King Edward Avenue, Ottawa, ON, K1N 6N5, Canada
novruzi@uottawa.ca

Abstract. We present a novel, thermodynamically consistent, model for the charged-fluid flow and the deformation of the morphology of polymer electrolyte membranes (PEM) in hydrogen fuel cells. The solid membrane is assumed to obey linear elasticity, while the pore is completely filled with protonated water, considered as a Stokes flow. The model comprises a system of partial differential equations and boundary conditions including a free boundary between liquid and solid. Our problem generalizes the well-known Nernst-Planck-Poisson-Stokes system by including mechanics. We solve the coupled non-linear equations numerically and examine the equilibrium pore shape. This computationally challenging problem is important in order to better understand material properties of PEM and, hence, the design of hydrogen fuel cells.

Keywords: Nernst-Planck-Poisson-Stokes system, Free boundary problem, Equilibrium shape, Fluid-structure interaction, Polymer electrolyte membrane, Proton exchange membrane fuel cell, Nafion, Mechanical deformation of pores, Ohmic interface resistance.

1 Introduction

Fuel cells running at low temperature provide a possibility for the electrification of the power train in automotive devices. Proton exchange membrane fuel cells (PEMFC) are based on hydrogen as fuel and do not rely on the use of fossil combustible material, which is likely getting increasingly scarce and expensive in the future. PEMFC allow to produce electric current by only emitting water as a byproduct and no carbon dioxide. In a hydrogen fuel cell, hydrogen enters

⋆ Corresponding author.

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 387–396, 2013.

into the fuel cell at the anode (negative), while oxygen flows in at the cathode (positive). The design of the PEM allows for control of the potentially explosive chemical reaction between hydrogen and oxygen. The reaction product, just water, leaves the fuel cell predominantly through the cathode outlet. The electric load is applied between anode and cathode, which closes the electric circuit. Within the PEM, the protons migrate from anode to cathode.

The PEM consists of a polymer with a nanopore structure. Negatively charged sulfonic acid groups (SO_3^-) at the walls of the nanoscale pores allow the dissociation of protons (H_3O^+) in the presence of water. A typical material for a PEM is Nafion, a perfluorosulfonic acid ionomer. A hydrated Nafion membrane exhibits a hydrophobic elastic backbone and hydrophilic pores, that are filled with protons and water molecules [1]. The precise morphology of Nafion on nanoscales remains a controversially discussed issue. We follow the widely accepted approach by Schmidt-Rohr and Chen [2], that a Nafion membrane consists mainly of parallel cylindrical channels surrounded by hydrophilic side chains.

Within the production of PEMs, a possibility is to press together several layers of thinner membranes. An increased ohmic resistance between interfaces of two joint PEM is observed in experiments. Our objective is to establish a mathematical model, describing the charged fluid flow and the change of the morphology of the pore due to mechanical deformations, that allows to understand better material properties of PEM. In particular we are interested in finding an equilibrium pore shape and, finally, in explaining this ohmic resistance.

The behaviour of water in a nanochannel may be very different compared to that of bulk water. The physics of water in a confined small channel depends significantly on the type of surface, i.e. whether the interface is hydrophilic or hydrophobic, and on the presence of surface charges [3]. There are several approaches, adapted to different scales, in order to study the charged fluid within PEM pores. On a microscopic level these are mainly molecular dynamics and Brownian motion, on a mesoscopic level there are continuum models, e.g. the Nernst-Planck-Poisson-Stokes (NPPS) system. We follow a continuum approach that is applicable in our situation since the Debye length is small, see [4].

We generalize the well-known NPPS model [5] by fluid-structure interaction between the charged fluid flow and the elastic wall of the channel. Previous electrohydrodynamic models, e.g. [4,5,6], do not incorporate the coupling to the mechanical displacement field. The equations for the flow stated by Castellanos [5] and analysed by Schmuck [7] are more general. They deal with the full nonstationary Navier-Stokes flow and allow also negative charge carriers, but the liquid domain is fixed. The equations in [4,6] represent the first-order approximation for a stationary version of our model without mechanical deformations. The significant effects of the radial variation of system parameters in this situation has been underlined in [4]. However, we work with a varying viscosity instead of a no-slip surface as in [4], modelling essentially the Stern layer in the pore. Furthermore, we consider slightly different boundary conditions (b.c.), e.g. we work with homogeneous Neumann b.c. (13), (14) for the chemical potential on in-/outlet and on the interface instead of Neumann b.c. for the proton

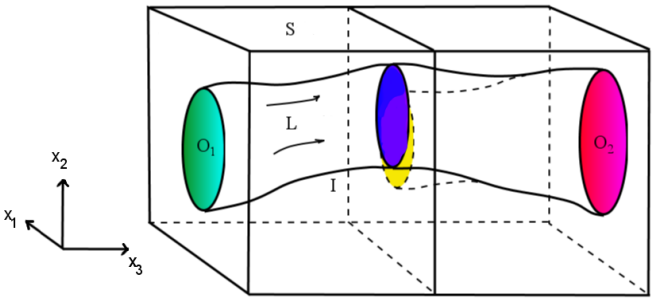


Fig. 1. Domains: L, the liquid, S the solid, I the interface, O_1 inlet (anode), O_2 outlet (cathode). Unknowns: in L: $\boldsymbol{u}$, p, ϕ, c; in S: $\boldsymbol{U}$. Dashed part of the interface: initial pore shape, part of the interface with continuous lines: equilibrium pore shape.

concentration, which are only an approximation of our thermodynamically consistent b.c. A free boundary problem for an interface between elastic solids and fluids is described in [8,9]. These models are similar to our model, when any coupling of the flow to electric field and concentration is neglected. A similar setting as ours has been considered by [10], but the authors focus on possible equilibrium configurations depending on the hydration of the pore and they make the assumption, arguably too strong in nature, of a constant proton concentration. Secondly, we study a stationary fluid flow in a fully saturated PEM.

To the best knowledge of the authors, a full problem involving both charged fluid flow and fluid-structure interaction has not been considered yet. We derive partial differential equations that we state in Sect. 2. We focus on a single cylindrical channel that joins another cylindrical channel from the adjacent PEM layer. Both channels are assumed to be completely filled with protonated water. In §2.1 we define our geometry and introduce the crucial physical quantities. The model is solved numerically by a commercial finite element software in Sect. 3. We discuss the impact of our results in the last part, Sect. 4, of this short paper.

2 An Elasto-Electrohydrodynamical Model for Polymer Electrolyte Membranes

2.1 Geometry and Relevant Physical Quantities

We assume that a single Nafion pore consists of several nanochannels and that the length of nanochannels is larger than their diameter d, typically a few nanometre. Instead of solving the full problem for many channels, we examine the situation around the region where one channel from one PEM layer meets one other channel from another PEM layer. We consider a nanochannel segment of length l including the interface (see Fig. 1). The joint channel extends further on in both directions and is assumed to be connected to other pores at both ends. Let $C = L \cup S \cup I$ denote the whole domain with the open domains L and

S, modelling the liquid channel and the solid Nafion backbone (of both PEM). $I = \partial L \cap \partial S$ denotes the interior free boundaries (or interface) between S and L and ∂C are all outer boundaries of C. We denote the inlet by O_1 and the outlet by O_2. Here the outer boundaries are considered as fixed boundaries, since in our model the channel extends further out of the considered domain C. Free boundaries may move with a normal speed ω. We make the following convention: $\boldsymbol{\nu}$ denotes the outer normal on $I \cup \partial C$, pointing on ∂C outside C and pointing on $I \setminus \partial C$ always from the liquid into the solid. Consequently $\partial_\nu f = \boldsymbol{\nabla} f \cdot \boldsymbol{\nu}$ denotes the derivative in direction of the outer normal for a differentiable function f. The mean curvature (times a factor 2) of a surface is the tangential divergence of the outer normal, i.e. $\kappa = \boldsymbol{\nabla}_{\boldsymbol{\tau}} \cdot \boldsymbol{\nu}$, being positive on I, if I is convex (seen from L). τ_1 and τ_2 denote the two orthonormal tangential vectors on I.

In this study, we suppose that the channel is completely filled with protonated water. Counter-ions, i.e. protons, are considered to be the only charge carriers in our model. On the interface I between liquid and solid we have negatively charged sulfonic acid groups, that are modelled by the negative surface charge density σ_C, see [6]. We like to solve for the velocity field $\boldsymbol{u}$, the pressure p, the proton concentration c, and the electric potential ϕ in the liquid L, and we are looking for the mechanical displacement field $\boldsymbol{U}$ in the solid S.

2.2 Modelling of Strains and Stresses

Mechanical strains and stresses are defined w.r.t. a reference configuration, where the system is free of strains and stresses or the stresses are at least known *a-priori*, either from experiments or from calculations (as it is possible e.g. in the case of a symmetrical geometry). In order to define a mechanical reference configuration, we consider at first another case where we have a straight circular channel with a fixed diameter cut out of a box of solid Nafion. An outer pressure p^* is exerted on the liquid channel without flow. At the interface the well-known Young-Laplace law $p^* + \gamma\kappa^R = p^R$ holds. Here γ is the surface tension, $\kappa^R = 2/d$ is the mean curvature of the straight channel, and p^R is the reference pressure in the solid for a straight cylinder.

As a reference configuration we consider the case of two joint straight cylindrical channel segments with the same radius and parallel axes. These channels are shifted by a fixed offset s between the axes. The offset is the distance between the circle centres at the interface plane. The reference pressure $\overline{p}$ in the solid, corresponding to the situation of two channels with an offset, is varying in space. Namely we have $\overline{p} = p^R = p^* - 2\gamma/d$ on the part of I that belongs to the cylinder barrels, while $\overline{p} = p^*$ on the part of I near the offset that belongs to the cylinder covers. Stresses and strains are to be formulated w.r.t. this reference configuration. The geometry within the reference configuration will be denoted by S_0, L_0 and I_0, while in the current (actual or deformed) configuration we write S, L and I. We assume that we may neglect here inelastic deformations that are due to changes of the chemical composition.

We consider mechanical deformations and displacements $\boldsymbol{U_0}$ in the reference configuration on S_0 and then as $\boldsymbol{U}$ in the current (actual) configuration on S w.r.t. the reference configuration. We consider a material point $\boldsymbol{X}$ in the reference configuration, whose location at time t is given by $\boldsymbol{x}$ in the current configuration. $\boldsymbol{\chi}(t, \boldsymbol{X}) = \boldsymbol{x}$ is called the deformation of material points. The displacement of a material point is defined by $\boldsymbol{U_0}(t, \boldsymbol{X}) := \boldsymbol{\chi}(t, \boldsymbol{X}) - \boldsymbol{X}$. We assume that we may invert the deformation $\boldsymbol{\chi}$ at any time w.r.t. material points $\boldsymbol{X}$ and, thus, we express the displacement w.r.t. the current configuration, $\boldsymbol{U}(t, \boldsymbol{x}) := \boldsymbol{U_0}(t, \boldsymbol{\chi^{-1}}(t, \boldsymbol{x}))$. It is convenient to consider fluid flow in the current configuration, i.e. in Eulerian coordinates, while mechanical displacements in a solid are considered in the reference configuration, i.e. in Lagrangian coordinates. We may work in good approximation with linear elasticity instead of nonlinear elasticity, which would be needed for a full description of large deformations of polymers, see e.g. [11]. Therefore solid stresses are represented by the Cauchy stress tensor $\sigma_S(\nabla U) = (-\overline{p} + \lambda_S \mathrm{tr}(\boldsymbol{\nabla U}))\mathbf{1} + \mu_S\, e(\boldsymbol{\nabla U})$, where $\mathbf{1}$ is the unit tensor, tr denotes the trace, and $e(t) = t + t^\top$ is the symmetrization of a tensor t times 2. The Lamé constants λ_S, μ_S are given material parameters. In the liquid we deal with a Newtonian fluid and the stress tensor reads $\sigma_L = -(p + \frac{2}{3}\mu_d \mathrm{tr}(\boldsymbol{\nabla u}))\mathbf{1} + \mu_d e(\boldsymbol{\nabla u})$, with μ_d being the dynamic viscosity.

2.3 Governing Equations

Now we may state our mathematical problem. For details of its derivation from first principles and the choice of thermodynamically consistent constitutive relations, see [12]. We remark that we assume that the dynamic viscosity μ_d, the electric permittivity ε_r, and the diffusion coefficient of protons, D, may vary in space [4]. Furthermore, we have as given data the outer pressure at in-/outlet $p_0^{(1)}/p_0^{(2)}$, a given displacement $\boldsymbol{g_0}$ on the boundary $\partial S_0 \setminus I_0$, and a constant external electric field E_{ext}.

Our problem consists of Stokes equations for $\boldsymbol{u}$ and p,

$$\text{[Momentum balance]} \qquad -\boldsymbol{\nabla} \cdot (\mu_d\, e(\boldsymbol{\nabla u})) + \boldsymbol{\nabla} p = -Fc\boldsymbol{\nabla}\phi \qquad \text{in } L, \tag{1}$$

$$\text{[Incompressibility]} \qquad \boldsymbol{\nabla} \cdot \boldsymbol{u} = \mathbf{0} \qquad \text{in } L, \tag{2}$$

where $Fc\nabla\phi$ accounts for electro-osmotic pressure, F being Faraday's constant. This is complemented by the boundary conditions

$$\text{[Normal pressure bal.]} \qquad \mu_d\, e(\boldsymbol{\nabla u})\boldsymbol{\nu} - p\boldsymbol{\nu} = -p_0^{(i)}\boldsymbol{\nu} \qquad \text{on } O_i,\ i = 1, 2, \tag{3}$$

$$\text{[Tangential moment. bal.]} \qquad \boldsymbol{u} \cdot \boldsymbol{\tau_j} = 0 \qquad \text{on } O_i,\ i, j = 1, 2, \tag{4}$$

$$\text{[Momentum balance]} \qquad \boldsymbol{u} = \partial_t \boldsymbol{U} \qquad \text{on } I. \tag{5}$$

The mechanical displacement field $\boldsymbol{U_0}$ is determined by the following problem of linear elasticity, formulated in the reference configuration,

$$-\boldsymbol{\nabla} \cdot \sigma_{\mathsf{S}}(\boldsymbol{\nabla U_0}) = \mathbf{0} \qquad \text{in } S_0, \tag{6}$$

$$\boldsymbol{U_0} = \boldsymbol{g_0} \qquad \text{on } \partial S_0 \setminus I_0, \tag{7}$$

$$-\sigma_{\mathsf{S}}(\boldsymbol{\nabla U_0})\boldsymbol{\nu_0} = -\mu_d\, e(\boldsymbol{\nabla u_0})\boldsymbol{\nu_0} + (p_0 - \gamma\kappa_0)\boldsymbol{\nu_0} \qquad \text{on } I_0, \tag{8}$$

where the last line is the Young-Laplace equation, a pressure balance. $\boldsymbol{u_0}$, p_0 and κ_0 denote u, p and κ expressed in the reference configuration. For the electric potential ϕ, we solve the Poisson equation with Neumann b.c.

$$\text{[Electrostatics]} \qquad -\varepsilon_0 \boldsymbol{\nabla} \cdot (\varepsilon_r \boldsymbol{\nabla} \phi) = Fc \qquad \text{in } L, \tag{9}$$

$$\text{[External electric field]} \qquad -\partial_\nu \phi = (-1)^i E_{ext} \quad \text{on } O_i,\ i = 1, 2, \tag{10}$$

$$\text{[Interface charges]} \qquad -\varepsilon_0 \varepsilon_r \partial_\nu \phi = -\sigma_c \qquad \text{on } I, \tag{11}$$

where ε_0 is the vacuum permittivity. For the proton concentration c, we solve the Nernst-Planck equation

$$-\boldsymbol{\nabla} \cdot \left(D \boldsymbol{\nabla} c + \frac{F}{RT} D c \boldsymbol{\nabla} \phi \right) + \boldsymbol{u} \cdot \boldsymbol{\nabla} c = 0 \qquad \text{in } L, \tag{12}$$

$$\partial_\nu c + (-1)^{i+1} \frac{F}{RT} E_{ext}\, c = 0 \qquad \text{on } O_i,\ i = 1, 2, \tag{13}$$

$$\partial_\nu c + \frac{F}{RT \varepsilon_0 \varepsilon_r} \sigma_c\, c = 0 \qquad \text{on } I, \tag{14}$$

the last two lines representing homogeneous Neumann b.c. for the chemical potential $RT \ln(c/\overline{c}) + F\phi$, R denoting the universal gas constant and T the temperature. In (12), the first term represents the diffusion of protons, the second term the migration of protons within the electric field and the third term is the advection due to the moving fluid. The normal velocity ω of the free boundary I is determined by

$$\text{[Normal momentum balance]} \qquad \omega = \partial_t \boldsymbol{U} \cdot \boldsymbol{\nu} \quad \text{on } I. \tag{15}$$

In this study, we are looking for stationary solutions of (1) – (15). Consequently, we neglect the time-derivatives in (5) and (15). Otherwise, we would have to prescribe an initial condition for I, and consider time-dependent domains and boundaries. The equilibrium pore shape corresponds to a displacement on the interface, $\boldsymbol{U_0}|_I$, that is constant in time. It is designated by (8).

We determine uniquely the electric potential ϕ by imposing $\phi = \overline{\phi}$ at the centre point P of the inlet O_1. By solving a nonlinear eigenvalue problem, a typical value $\overline{c}$ for the proton concentration is derived in [6] that corresponds to setting $\phi = \overline{\phi}$ at the centre of a straight cylindrical channel in the case of constant permittivity. The arbitrariness of $\overline{\phi}$ or $\overline{c}$ corresponds to the gauge invariance of the electric potential.

We emphasize that the b.c. (10) is consistent with global electroneutrality, meaning that we rule out that the electric field extends into the solid and, hence, ions would cross the interface. This requires that the hydronium charges (H_30^+) in the liquid and the negative ions (SO_3^-) on the boundary balance, $\int_L Fc + \int_I \sigma_c = 0$. Together with (9), (11), and Gauss' theorem we find $\int_{O_1 \cup O_2} \varepsilon_0 \varepsilon_r \partial_\nu \phi = 0$. The last equation is guaranteed by (10), a choice among many others.

3 Numerical Solution

For any solution approach, numerical or analytical, it is crucial to non-dimensionalize the system, and to split up the equations and, consequently, the quantities according to their different scales. We remark that by non-dimensionalization of the system we find the relevant dimensionless coupling parameters, e.g. a parameter resulting from non-dimensionalization of (9) is the Debye length of the nanochannel. The dimensionless parameter Pe arising in (12) is the Peclet number describing the relation between advection and diffusion for mass transfer. Since $Pe \ll 1$ we could be tempted to neglect the term $\boldsymbol{u} \cdot \boldsymbol{\nabla} c$ in the Nernst-Planck equation. This observation motivates the first-order approximation that is considered and compared with the full model in [12]. However, for the flux of protons, $\boldsymbol{j_+} = -D\left(\boldsymbol{\nabla} c + \frac{F}{RT} c \boldsymbol{\nabla}\phi\right) + c\boldsymbol{u}$, that enters significantly into the ohmic resistance within the nanochannel, we would obtain zero within this first-order approximation. Hence, it is crucial to consider the full equation (12). Sorting the r.h.s. of the PDE by their scales suggests to introduce an internal electric potential ϕ_0 (with a non-zero b.c. only on the interface), an external electric potential ϕ_1 (with a non-zero b.c. only at in-/outlet) and a remainder potential ϕ_2. Corresponding to ϕ_0 we introduce a concentration $c_0 = \overline{c}\exp\left(-F/(RT)\phi_0\right)$, yielding a remainder concentration c_1 with a well scaled PDE. Finally for the pressure it is convenient to replace p by $q := p - RTc_0$, since $q = 0$ implies $u = 0$.

Since the analytical solution of our non-linear coupled system is quite ambitious we focus on a numerical solution. Schmuck's analytical results [7], for a model without coupling to linear elasticity, rely on the parabolic structure of the non-stationary equations for $\boldsymbol{u}$ and c and cannot be transferred to our stationary case directly. We remark that we have a two-sided fluid-structure interaction. The charged fluid flow influences by means of (8) the deformation of the elastomer membrane, while a narrowing (e.g. a complete closing of the channel) or widening of L has large influence on velocity or pressure.

3.1 Description of the Numerical Algorithm

Our numerical algorithm solves for the state variables as well as for the free interface. It should be emphasized that it is a non-trivial matter in which order the coupled equations are solved so as to obtain fast convergence and numerically stable results. We make use of the ALE (Arbitrary Lagrangian Eulerian) method, i.e. we discretize the original geometry in Lagrangian coordinates. Our numerical algorithm is built up in the following way:

a) Considering the reference configuration, we initialize the mesh for L_0, S_0 and I_0. Formally, we set $\boldsymbol{U_0} = \boldsymbol{0}$, $\boldsymbol{u_0} = \boldsymbol{0}$, and $q_0 = 0$, where q_0 is q transformed into the reference configuration.
b) We compute smoothed normal vectors on the boundaries by extending the normal vector field $\boldsymbol{\nu_0}$ by means of $\Delta\boldsymbol{\nu_0} = \boldsymbol{0}$ into $L_0 \cup S_0$, using the definition of $\boldsymbol{\nu_0}$ on $\partial C_0 \cup I_0$ as boundary condition. This enables us to compute the mean curvature κ_0, even at corners that would become smooth instantly due to surface tension anyways.

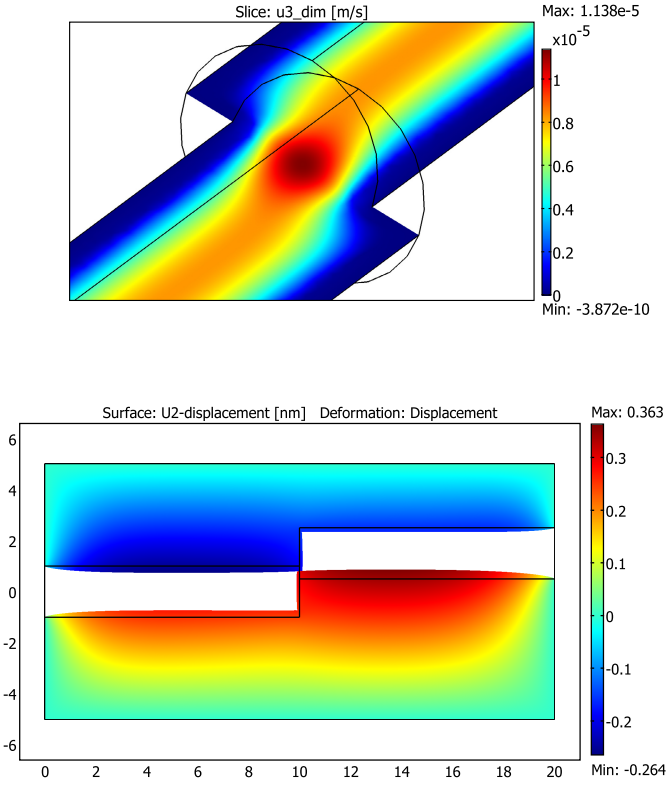


Fig. 2. Top: u_3, velocity in x_3 direction, zoom towards the channel intersection, plotted on the cross section $L \cap \{x_1 = 0\}$ (s.t. the channels are sliced in half). Bottom: Cross section $S \cap \{x_1 = 0\}$. $\boldsymbol{U}_2$, mechanical displacement field in x_2 direction, deformed configuration. A channel may close completely. In both figures the cylinders have diameter $d = 2$ nm in the reference configuration, length 10 nm and offset $s = 0.5$ nm; $\boldsymbol{g_0} = \boldsymbol{0}$.

c) We store $\boldsymbol{U_0^{(old)}} = \boldsymbol{U_0}$, then we solve the linear elasticity problem (6) – (8) for $\boldsymbol{U_0}$ in S_0, using the present values for $\boldsymbol{u}_0$ and q_0.
d) We update the geometry. The free boundary is moved by the mechanical displacement $\boldsymbol{U_0} - \boldsymbol{U_0^{(old)}}$ on I_0, yielding the update of I, and hence of S and L.
e) We solve the electrohydrodynamical system (1) – (5), (9) –(14) as follows.
   (i) First we solve for ϕ_1, ϕ_0, and c_0, that do not depend on other variables.
   (ii) We solve the remaining equations iteratively. We start solving for $\boldsymbol{u}$ and q simultaneously, and then for ϕ_2 and c_1 simultaneously.
   (iii) Then we reiterate e)(i) and e)(ii) until we have a suitable residual error.

f) If $\max_{I_0} \left\| \boldsymbol{U_0} - \boldsymbol{U_0^{(old)}} \right\| < err$, err a prescribed error tolerance, or if a topological event has occured (e.g. closing of the pore), or a specified maximal number of iterations has been reached, we terminate our algorithm with the obtained numerical solution. Else we take the updated geometry as new reference configuration. Should the situation of a geometrically unsuitable mesh arise after the deformation, we remesh. We restart with step b).

For more details of the algorithm see [13]. Our algorithm has been implemented in the commercial finite element software COMSOL 3.4. The use of a parallel solver, like PARDISO, is crucial for efficient calculations.

3.2 Numerical Results

We show two plots, corresponding to two different choices for external parameters, namely, (i) a situation where the initial interface is close to an equilibrium (due to large $p_0^{(1)} - p_0^{(2)}$ and p^* the surface tension term is negligible) (see Fig. 2, top), and (ii) a situation where the charged fluid flow is close to zero ($p_0^{(1)} - p_0^{(2)}$ small, E_{ext} negligible) (see Fig. 2, bottom). All remaining data for our simulations is discussed and summarized in [12]. We emphasize that the database, e.g. for surface tension of protonated water and typical pressures, is thin.

4 Conclusions and Open Questions

We have stated a continuum model describing the charged fluid flow within nanochannels of PEM. The numerically accurate simulation of the differential equations allows further investigation of system characteristics. In our study [12], we focus on the ohmic interface resistance between two circular cylindrical pores, and analyze its dependence on system parameters. By splitting our model into two models, we consider the main effects separately. In this context, our model suggests that this interface resistance depends mainly on two factors: (a) the offset value (the distance between the pore centres at the intersection plane) and its stochastic distribution and (b) on the deformed pore shape due to the balance of elastic and electrohydrodynamic forces. Furthermore, for a straight channel the electro-osmotic drag and the specific pore conductivity match results from experiments, see the discussion in [12], supporting the validity of our model. The effect (b) raises the question whether it is possible to find an equilibrium shape for the nanochannel pore, balancing the solid pressure with the liquid pressure and the *interfacial* pressure, due to surface tension and mean curvature of the interface. This question yields a mathematically challenging problem and requires refined numerical techniques. For suitable outer pressures and external electric field, the above described algorithm suggests fast convergence, but a mathematical justification thereof is missing. However, for a simplified version of our model we can prove existence and uniqueness of a solution and the shape differentiability of the solution. Using a variational energy formulation, we show that this numerically determined optimal shape minimizes the free energy. The latter results are the subject of an upcoming paper [13].

In summary, our full model, being derived from thermodynamical first principles, presents a generalisation of previous models [4,5,6] by including surface tension and bulk stresses in the solid Nafion, or compared to [8,9,10], by including charged fluid flow. We emphasize that, contrary to our last contributions [4,6], we have solved additionally, in the electrohydrodynamical part of our model, the full Nernst-Planck-Poisson-Stokes system without neglecting higher order terms. In particular, our novel approach, i.e. a continuum model of the electrochemical fluid flow within a nanochannel of a PEM including surface tension and bulk stresses, allows to explain ohmic resistance and electro-osmotic drag. Our mathematical results might turn out to be important for further advances in the design of hydrogen fuel cells.

Acknowledgements. The authors would like to thank Toyota Motor Engineering and Manufacturing North America (TEMA) for financial support of this research. S.-J. K. thanks the University of Ottawa for its hospitality.

References

1. Promislow, K., Wetton, B.: PEM Fuel cells: A mathematical Overview. SIAM J. Appl. Math. 70, 369–409 (2009)
2. Schmidt-Rohr, K., Chen, Q.: Parallel cylindrical water nanochannels in Nafion fuel-cell membranes. Nat. Mater. 7, 75–83 (2008)
3. Karniadakis, G., Beskok, A., Aluru, N.: Microflows and nanoflows. Fundamentals and simulation. Springer, New York (2005)
4. Ladipo, K., Berg, P., Kimmerle, S.-J., Novruzi, A.: Effects of radially dependent parameters on proton transport in polymer electrolyte membrane nanopores. J. Chem. Phys. 134, 074103-1–074103-12 (2011)
5. Castellanos, A.: Electrohydrodynamics. Springer, Wien (1998)
6. Berg, P., Ladipo, K.: Exact solution of an electro-osmotic flow problem in a cylindrical channel of polymer electrolyte membranes. Proc. R. Soc. A 465, 2663–2679 (2009)
7. Schmuck, M.: Analysis of the Navier-Stokes-Nernst-Planck-Poisson system. M3AS 9, 993–1014 (2009)
8. Fortin, M.: Problèmes de surfaces libres en mécanique des fluides. In: Shape Optimization and free Boundaries, pp. 143–172. Kluwer Academic Publishers, Dordrecht (1992)
9. Discacciati, M., Fourestey, G., Quarteroni, A., Deparis, S.: Fluid-structure algorithms based on Steklov-Poincaré operators. Comput. Methods Appl. Mech. Engrg. 195, 5797–5812 (2006)
10. Elfring, G.J., Struchtrup, H.: Thermodynamics of pore wetting and swelling in Nafion. J. Membr. Sci. 315, 125–132 (2008)
11. Müller, I., Strehlow, P.: Rubber and rubber balloons. Paradigms of thermodynamics. Springer, Heidelberg (2004)
12. Kimmerle, S.-J., Novruzi, A., Berg, P., Ladipo, K.: Ohmic resistance of charged fluid flow in deformable nanochannels connecting polymer electrolyte membranes (preprint submitted)
13. Berg, P., Kimmerle, S.-J., Novruzi, A.: Modeling, shape analysis and computation of equilibrium pore shape near a PEM-PEM intersection (preprint)

Reduction Strategies for Shape Dependent Inverse Problems in Haemodynamics*

Toni Lassila[1], Andrea Manzoni[2], and Gianluigi Rozza[2]

[1] EPFL, École Polytechnique Fédérale de Lausanne
CMCS, Chair of Modelling and Scientific Computing
Station 8, CH-1015, Lausanne, Switzerland
[2] SISSA MathLab, International School for Advanced Studies, Trieste, Italy
{toni.lassila,andrea.manzoni,gianluigi.rozza}@epfl.ch
http://cmcs.epfl.ch

Abstract. This work deals with the development and application of reduction strategies for *real-time* and *many query* problems arising in fluid dynamics, such as shape optimization, shape registration (reconstruction), and shape parametrization. The proposed strategy is based on the coupling between reduced basis methods for the reduction of computational complexity and suitable shape parametrizations – such as free-form deformations or radial basis functions – for low-dimensional geometrical description. Our focus is on problems arising in haemodynamics: efficient shape parametrization of cardiovascular geometries (e.g. bypass grafts, carotid artery bifurcation, stenosed artery sections) for the rapid blood flow simulation – and related output evaluation – in domains of variable shape (e.g. vessels in presence of growing stenosis) provide an example of a class of problems which can be recast in the *real-time* or in the *many-query* context.

Keywords: Model order reduction, reduced basis methods, free-form deformation, radial basis functions, computational fluid dynamics, shape parametrization, blood flows.

1 Introduction and Motivation

In last decades more and more powerful computers have allowed to solve numerical problems of very large dimensions and describing very complex phenomena. Nevertheless, a computational reduction is still crucial whenever interested to high performances in rapid – even *real-time* – simulations and/or repeated *output* evaluations – seen as *many queries* evaluations– for different values of some *inputs* of interest.

* Work supported by Swiss National Science Foundation (SNSF) under grant 200021-122136 and European Research Council under Project Mathcard ERC-2008-AdG-227058.

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 397–406, 2013.

1.1 A General Strategy for Reduction in Shape Dependent Flows

Flow control and optimization problems can be formulated as the minimization of a given cost functional (or *output*) controlling some *input* parameters which can be physical quantities (e.g. source terms or boundary values) or, alternatively, geometrical quantities; we refer to the latter case as flow control by shape variation, and the optimization of the corresponding flow geometries is thus one possibility to reach that goal; we refer to this case – the most difficult one among flow control problems – such as shape optimization or shape registration/reconstruction problems [4]. Concerning applications arising in fluid mechanics, cost functionals are expressed as functions of flow variables (such as velocity, pressure, temperature), while constraints are usually given in form of PDE systems (Stokes, Navier-Stokes equations, with or without coupling with a structural equation to account for fluid-structure interaction effects) describing the flow, besides topological constraints on the shape of the domain, if necessary. Since *(i)* optimization procedures require repetitive evaluations of outputs, *(ii)* PDEs can be hard to solve and *(iii)* discretization is expensive when geometry keeps changing, computational costs are usually very high; we thus want to address suitable strategies to reduce numerical efforts in *many-query* problems.

Substantial computational saving becomes possible thanks to a *reduced order model* which relies on two reduction steps: *(i)* parameterization of the admissible shapes and *(ii)* substitution of the full-order finite element (FE) solution of flow problems with a reduced solution obtained by the reduced basis (RB) method [17]. In fact, once an equivalent parametrized formulation of the flow problem – now embedding the shape as a parametric quantity – can be derived, reduced basis method for parametrized PDEs, enables to evaluate the *output* very rapidly. In the end, at the outer level a suitable iterative procedure for the optimization is performed. A brief presentation of the whole framework can be found in [8], while a more detailed analysis has been recently addressed in [5,6].

1.2 Abstract Setting

From an abstract point of view, a shape optimization/identification can be seen as an optimal control problem for which the control variable is the shape of the domain Ω itself. This entails the minimization of a cost functional $\mathcal{J}(\cdot)$ over a set of admissible shapes $\mathcal{O}_{ad}$, by finding the optimal shape of the domain where the PDE is defined:

$$\text{find} \quad \hat{\Omega} = \arg \min_{\Omega \in \mathcal{O}_{ad}} \mathcal{J}(Y(\Omega)) \tag{1}$$

where $\mathcal{J}(Y(\Omega))$ depends on the solution $Y = Y(\Omega)$ of a PDE state problem – defined on Ω – which can be written in an abstract form as

$$Y \in \mathcal{Y}(\Omega): \quad \mathcal{A}(Y, W; \Omega) = \mathcal{F}(W; \Omega), \quad \forall\, W \in \mathcal{Y}(\Omega). \tag{2}$$

Here $\mathcal{A}(\cdot, \cdot; \Omega)$ is a continuous, uniformly inf-sup stable bilinear form and $\mathcal{F}(\cdot; \Omega)$ is a bounded linear form, both defined on the original domain Ω; $\mathcal{Y}(\Omega)$ denotes a

suitable functional space defined over Ω. Let us assume that the shape $\Omega = \Omega(\boldsymbol{\mu})$ depends on a set of *input* parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p) \in \mathcal{D} \subset \mathbb{R}^p$; in this way, problem (1)-(2) can be reduced to the following *parametric optimization* inverse problem:

$$\text{find} \quad \hat{\boldsymbol{\mu}} = \arg \min_{\boldsymbol{\mu} \in \mathcal{D}_{ad}} \mathcal{J}(Y(\boldsymbol{\mu})) \tag{3}$$

where $\mathcal{D}_{ad} \subseteq \mathcal{D}$ and $Y(\boldsymbol{\mu})$ solves

$$Y(\boldsymbol{\mu}) \in \mathcal{Y}(\Omega(\boldsymbol{\mu})): \quad \mathcal{A}(Y(\boldsymbol{\mu}), W; \boldsymbol{\mu}) = \mathcal{F}(W; \boldsymbol{\mu}), \quad \forall\, W \in \mathcal{Y}(\Omega(\boldsymbol{\mu})). \tag{4}$$

For a more general setting and overview, see e.g. [5].

2 Reduced Basis Method for Computational Reduction

Our approach to shape dependent flow problems takes advantage of *reduced basis* (RB) methods for rapid and reliable prediction of engineering outputs associated with parametric PDEs [17,12,14]; see e.g. [15,19,16] for applications to the Stokes problem and [13,20,3] for the Navier-Stokes case. The method is built upon a classical finite element (FE) "truth" approximation space $\mathcal{Y}^{\mathcal{N}}$ of (typically very large) dimension $\mathcal{N}$ and is based on the use of "snapshot" FE solutions of the PDEs, corresponding to certain parameter values, as global approximation basis functions previously computed and stored. The RB framework requires a reference ($\boldsymbol{\mu}$-independent) domain $\tilde{\Omega}$ in order to compare, and combine, FE solutions that would be otherwise computed on different domains and grids; moreover, this procedure enables to avoid shape deformation and remeshing that normally occur at each step of an iterative optimization procedure [18]. In Sect. 3 two possible techniques for the construction of such a mapping will be briefly recalled.

We thus consider $\tilde{\Omega}$ as reference domain related to the parameter-dependent "original" domain of interest $\Omega(\boldsymbol{\mu})$ through a parametric mapping $T(\cdot; \boldsymbol{\mu})$, s.t. $\Omega(\boldsymbol{\mu}) = T(\tilde{\Omega}; \boldsymbol{\mu})$. By mapping the problem (3) back to the reference domain $\tilde{\Omega}$, we obtain the following problem in its abstract form:

$$\begin{aligned} &\text{find} \quad \hat{\boldsymbol{\mu}} = \arg \min_{\boldsymbol{\mu} \in \mathcal{D}_{ad}} s(\boldsymbol{\mu}) = \tilde{\mathcal{J}}(Y(\boldsymbol{\mu})) \quad \text{s.t.} \\ &Y(\boldsymbol{\mu}) \in \mathcal{Y}(\tilde{\Omega}): \ \tilde{\mathcal{A}}(Y(\boldsymbol{\mu}), W; \boldsymbol{\mu}) = \tilde{\mathcal{F}}(W; \boldsymbol{\mu}), \ \forall\, W \in \mathcal{Y}(\tilde{\Omega}). \end{aligned} \tag{5}$$

Focusing on shape optimization and/or registration problems, and following the so-called *discretize than optimize* approach, the standard Galerkin FE approximation of (5) reads as follows:

$$\begin{aligned} &\text{find } \hat{\boldsymbol{\mu}} = \arg \min_{\boldsymbol{\mu} \in \mathcal{D}_{ad}} s^{\mathcal{N}}(\boldsymbol{\mu}) = \tilde{\mathcal{J}}(Y^{\mathcal{N}}(\boldsymbol{\mu})) \ \text{ s.t.} \\ &Y^{\mathcal{N}}(\boldsymbol{\mu}) \in \mathcal{Y}^{\mathcal{N}}: \ \tilde{\mathcal{A}}(Y^{\mathcal{N}}(\boldsymbol{\mu}), W; \boldsymbol{\mu}) = \tilde{\mathcal{F}}(W; \boldsymbol{\mu}), \ \forall\, W \in \mathcal{Y}^{\mathcal{N}}. \end{aligned}$$

The reduced basis method provides an efficient way to compute an approximation $Y_N(\boldsymbol{\mu})$ of $Y^{\mathcal{N}}(\boldsymbol{\mu})$ (and related output) by using a Galerkin projection on

a reduced subspace made up of well-chosen FE solutions, corresponding to a specific choice $S_N = \{\boldsymbol{\mu}^1, \ldots, \boldsymbol{\mu}^N\}$ of parameter values. Denoting

$$\mathcal{Y}_N^{\mathcal{N}} = \text{span}\{Y^{\mathcal{N}}(\boldsymbol{\mu}^n), n = 1, \ldots, N\}, \tag{6}$$

the RB space, the RB formulation of (5) is as follows:

$$\text{find } \hat{\boldsymbol{\mu}} = \arg \min_{\boldsymbol{\mu} \in \mathcal{D}_{ad}} s_N(\boldsymbol{\mu}) = \tilde{\mathcal{J}}(Y_N(\boldsymbol{\mu})) \text{ s.t.}$$
$$Y_N(\boldsymbol{\mu}) \in \mathcal{Y}_N^{\mathcal{N}} : \ \tilde{\mathcal{A}}(Y_N(\boldsymbol{\mu}), W; \boldsymbol{\mu}) = \tilde{\mathcal{F}}(W; \boldsymbol{\mu}), \ \forall W \in \mathcal{Y}_N^{\mathcal{N}}.$$

Thanks to the (considerably) reduced dimension $O(N) \ll O(\mathcal{N})$ of the systems obtained from RB approximation, we can provide both reliable results and rapid response in the real-time and multi-query contexts. In particular:

- *Reliability* is ensured by rigorous a posteriori estimations for the error in the RB approximation w.r.t. truth FE discretization (see e.g. [17,16]);
- *Rapid response* is achieved by an Offline–Online computational strategy and a rapidly convergent RB space assembling, based on a *greedy algorithm*. To achieve this goal, RB methods rely on the assumption of affine parametric dependence[1] in $\mathcal{A}(\cdot, \cdot; \boldsymbol{\mu})$ and $\mathcal{F}(\cdot; \boldsymbol{\mu})$.

Hence, in an expensive Offline stage we prepare a very small RB "database", while in the Online stage, for each new $\boldsymbol{\mu} \in \mathcal{D}$, we rapidly evaluate both the field and the output (with error bounds) whose computational complexity is independent of FE dimension $\mathcal{N}$.

3 Efficient Shape Parametrization Techniques for Geometrical Complexity Reduction

In general, shape optimization problems feature more difficulties than optimal control problems, such as shape deformation, shape derivatives and the evaluation of shape-dependent quantities: a crucial aspect of optimal shape design is thus the geometrical treatment of the shapes during the optimization process. Common strategies for shape deformation involve the use of *(i)* the coordinates of the boundary points as design variables (*local boundary variation*) or *(ii)* some families of basis shapes combined by means of a set of control point (*polynomial boundary parametrizations*).

These techniques are not well suited within the RB framework, since a global transformation $T(\cdot; \boldsymbol{\mu})$ is needed, rather than a boundary representation [18]. A more versatile parametrization can be introduced by exploiting the *free-form deformation* (FFD) techniques, in which the deformations of an initial design, rather than the geometry itself, are parametrized [7]. In this case, the shape

[1] If this assumption does not hold, it could be recovered in through an intermediate *empirical interpolation* process.

parametrization is constructed on a regular lattice of control points, by combining the deformations acting on a subset of active control points through a basis of (tensor products of) Berstein polynomials. Input parameters are given by the deformations of the active control points, which have to be properly chosen, following some problem-dependent criteria [10].

Despite its flexibility, the FFD techniques do not satisfy any interpolation property and control points must reside on a regular lattice. In order to overcome these possible limitations, other different techniques based on interpolation properties may be recovered. In particular, we have been focusing on the *radial basis functions* (RBF) techniques [9], which are traditionally used for nonlinear multidimensional interpolation on scattered data (for example in image registration). With respect to FFD techniques, RBF techniques allow a better local boundary control and a free choice of the position of the control points (also on the boundary of the shape domain).

4 Application in Haemodynamics: Real-Time Blood Flow Simulations in Parametrized Cardiovascular Geometries

The framework based on the coupling between FFD or RBF techniques (or other low-dimensional shape parametrizations) and RB methods has turned out to be useful also for a real-time simulation of blood flows in arterial vessels which might show a deep variation in geometrical configuration, as for example carotid artery bifurcations. Our goal is twofold:

- spanning a variety of carotid configurations through low-dimensional shape parametrizations [1], and shape registration of parametrized carotid shapes from patient data measured in the form of flow velocities;
- real-time simulation of blood flows in reconstructed geometries and computing indices related to arterial occlusion risk and highly dependent on geometrical configurations, possibly for predictive surgery applications.

In the first approach we might minimize some discrepancy functional between the simulated velocity and the observed velocity in an atlas-based variational data assimilation method (see e.g. [11]); in the latter we minimize a cost function such as the viscous energy dissipation

$$\mathcal{J}(Y(\boldsymbol{\mu})) = \frac{\nu}{2} \int_{\Omega} |\nabla \mathbf{u}(\boldsymbol{\mu})|^2 \, d\Omega.$$

to obtain carotid shapes exhibiting the least disturbance to the blood flow, being $Y(\boldsymbol{\mu}) = (\mathbf{u}(\boldsymbol{\mu}), p(\boldsymbol{\mu}))$ the velocity and the pressure of the fluid, respectively.

4.1 Validation of the Reduced Basis Methodology

A first numerical test has been performed exploiting a coupled FFD+RB framework on a simple geometrical configuration (see Fig. 1), given by a stenosed

carotid artery parametrized with respect to the displacement of two control points ($p = 2$) located close to the bifurcation (see e.g. [9] for further details about representation of carotid bifurcations). Flow simulations through a steady Stokes model show a remarkable dependence of the flow even on small variation of the shape configuration. In particular, our interest has been focused on the evaluation of an output related both with the flow and the shape, given by the viscous energy dissipation.

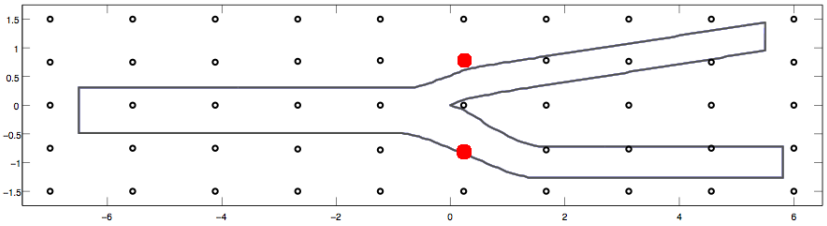


Fig. 1. Schematic diagram of the FFD setting; bold (red) control points can be freely moved in vertical direction and used as parameters representing small deformations

Some details concerning the reduced basis spaces are listed in Tab. 1; we remark the strong reduction in the system dimensions and a large computational speedup, concerning performances for each new geometrical configuration, of about two orders of magnitude. We provide a certification of the accuracy of the methodology: in Fig. 2 the true errors between the FE and the RB approximation are reported, the related error bounds (see [16] for error bound expression and derivation), as well as the error between the FE and the RB output. We observe fast, nearly exponential convergence in N. Furthermore, the a posteriori error bounds are both reliable and reasonably effective.

Table 1. RB + FFD for the carotid artery bifurcation: numerical details

Number of FE dof $\mathcal{N}_{\mathbf{v}} + \mathcal{N}_p$	24046
Number of RB functions N	16
Number of design variables P	2
Linear system dimension reduction	500:1
FE evaluation t_{FE}^{online} (s)	2.8039
RB evaluation t_{RB}^{online} (s)	0.0231

4.2 A Comparison between FFD and RBF Parametrizations

Next, we report here some preliminary results on the comparison between a FFD and a RBF setting defined on the carotid configuration already introduced. Also in this case we are interested in the evaluation of the viscous energy dissipation;

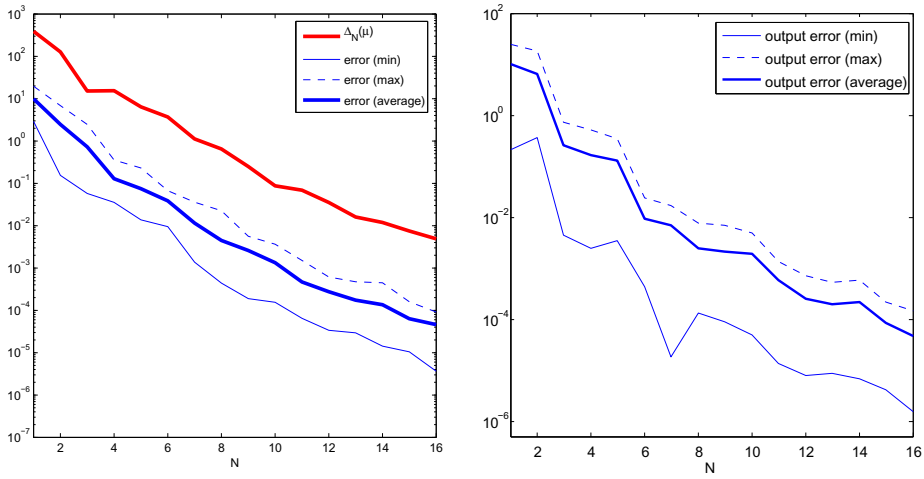


Fig. 2. Left: • error estimation (natural norm) and • true error between RB and FE approximation; right: true error between FE and RB output (vorticity)

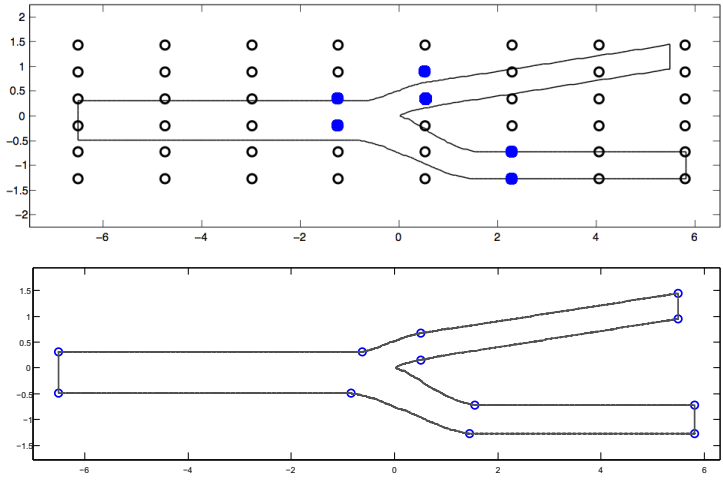


Fig. 3. FFD setting (top) and RBF setting (bottom); for each case, parameters are given by the displacements of the selected (blue) control points

we just compare the two parametrizations techniques by considering a Navier-Stokes model for the fluid flow. Both the settings deal with $p = 6$ parameters, given by the vertical displacements of some selected control points; in the FFD case we introduce a 6×8 lattice of control points, while in the RBF case we introduce in total 12 control points close to the bifurcation and at the extrema (see Fig. 3), using the *thin-plate spline* (TPS) and the *Gaussian* shape functions

Table 2. Results for the minimization of the viscous energy dissipation obtained by using the FFD and the RBF settings introduced above

	FFD	RBF (thin-plate)	RBF (Gaussian)
output reduction	$39,1\%$	$45,9\%$	$36,7\%$
iterations	84	117	91
parameters	6 (48)	6 (12)	6 (12)

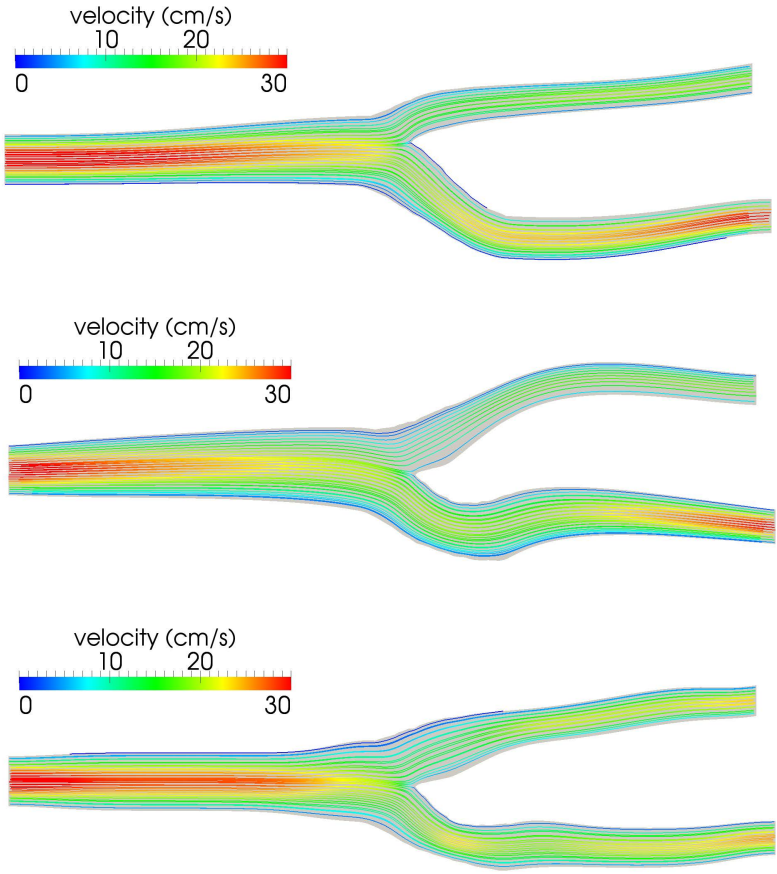


Fig. 4. Optimal configuration obtained by minimizing the viscous energy dissipation for the FFD case (top), the RBF case with the thin-plate spline option (middle) and the RBF case with the Gaussian option (bottom)

[2]. In this last case, we deal with the displacement of the six control points located at the center of the configuration.

We compare the shapes obtained by minimizing the viscous energy dissipation: in Tab. 2 are reported the results for the two cases, while the configurations

corresponding to the minimum values of the viscous energy dissipation are represented in Fig. 4. We can remark that the three shapes are quite similar, as well as the output reduction. The number of iterations taken by the optimization procedure is comparable among the three options. Regarding the main qualities of these two shape parametrization tools, the RBF technique proves to be more versatile and accurate for this kind of applications – it enables to choose freely the location of control points rather than selecting the most relevant control points on the regular FFD lattice, as well as to impose interpolation constraints – even if its construction and the computation of the related parametrized tensors is much more difficult. Not only, by considering the same amount, location and available displacements of control points, the Gaussian RBF is found to be more suitable for describing local deformations; the TPS option allows to get a more global and regular deformation, where an enhanced shape smoothness is ensured by a minimization of the bending deformation energy property, fullfilled by this kind of RBF.

5 Conclusion and Perspectives

The capability of the reduced basis method to solve shape registration and optimization problems involving incompressible flows in real-time looks promising if coupled with an efficient and versatile geometrical parametrization. The integration of the RBF parametrization technique within the reduced basis framework, as well as its application to blood flow simulation on geometries reconstructed from patient data, looks promising in its flexibility and ability to express a variety of shape deformations. Further elements that may be explored deal with the uncertainty quantification [5] and/or robust optimization and control problems [6] for patient-specific scenarios.

References

1. Bressloff, N.W.: Parametric geometry exploration of the human carotid artery bifurcation. J. Biomech. 40, 2483–2491 (2007)
2. Buhmann, M.D.: Radial Basis Functions. Cambridge University Press, UK (2003)
3. Deparis, S., Rozza, G.: Reduced basis method for multi-parameter-dependent steady Navier-Stokes equations: Applications to natural convection in a cavity. J. Comp. Phys. 228(12), 4359–4378 (2009)
4. Haslinger, J., Mäkinen, R.A.E.: Introduction to shape optimization: theory, approximation, and computation. SIAM (2003)
5. Lassila, T., Manzoni, A., Quarteroni, A., Rozza, G.: A reduced computational and geometrical framework for inverse problems in haemodynamics (2011) (submitted)
6. Lassila, T., Manzoni, A., Quarteroni, A., Rozza, G.: Boundary control and shape optimization for the robust design of bypass anastomoses under uncertainty. Accepted for publication in ESAIM Math. Model. Numer. Anal. (2012), doi: 10.1051/m2an/2012059
7. Manzoni, A., Quarteroni, A., Rozza, G.: Shape optimization for viscous flows by reduced basis methods and free-form deformation. Int. J. Numer. Meth. Fluids 70(5), 646–670 (2012)

8. Manzoni, A.: Model order reduction by reduced basis for optimal control and shape optimization. Paper Awarded with the 3rd BGCE Student Paper Prize at the 2011 SIAM Computational Science and Engineering Conference, Reno, NV, USA (2011)
9. Manzoni, A., Quarteroni, A., Rozza, G.: Model reduction techniques for fast blood flow simulation in parametrized geometries. Int. J. Numer. Methods Biomed. Engng. (2011), doi:10.1002/cnm.1465) (in press)
10. Manzoni, A., Quarteroni, A., Rozza, G.: Shape optimization of cardiovascular geometries by reduced basis methods and free-form deformation techniques. Int. J. Numer. Methods Fluids (2011), doi:10.1002/fld.2712) (in press)
11. McLeod, K., Caiazzo, A., Fernández, M., Mansi, T., Vignon-Clementel, I., Sermesant, M., Pennec, X., Boudjemline, Y., Gerbeau, J.F.: Atlas-based reduced models of blood flows for fast patient-specific simulations. Statistical Atlases and Computational Models of the Heart, 95–104 (2010)
12. Patera, A.T., Rozza, G.: Reduced Basis Approximation and A Posteriori Error Estimation for Parametrized Partial Differential Equations. Version 1.0, Copyright MIT, to appear in (tentative rubric) MIT Pappalardo Graduate Monographs in Mechanical Engineering (2006), http://augustine.mit.edu
13. Quarteroni, A., Rozza, G.: Numerical solution of parametrized Navier-Stokes equations by reduced basis methods. Numer. Methods Partial Differential Equations 23(4), 923–948 (2007)
14. Quarteroni, A., Rozza, G., Manzoni, A.: Certified reduced basis approximation for parametrized partial differential equations in industrial applications. J. Math. Ind. 1(3) (2011)
15. Rozza, G.: Reduced basis methods for Stokes equations in domains with non-affine parameter dependence. Comput. Vis. Sci. 12(1), 23–35 (2009)
16. Rozza, G., Huynh, D.B.P., Manzoni, A.: Reduced basis approximation and error bounds for Stokes flows in parametrized geometries: roles of the inf-sup stability constants. Numer. Math (in press, 2013)
17. Rozza, G., Huynh, D.B.P., Patera, A.T.: Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations. Arch. Comput. Methods Engrg. 15, 229–275 (2008)
18. Rozza, G., Manzoni, A.: Model order reduction by geometrical parametrization for shape optimization in computational fluid dynamics. In: Pereira, J.C.F., Sequeira, A. (eds.) Proceedings of ECCOMAS CFD 2010, V European Conference on Computational Fluid Dynamics, Lisbon, Portugal (2010)
19. Rozza, G., Veroy, K.: On the stability of the reduced basis method for Stokes equations in parametrized domains. Comput. Methods Appl. Mech. Engr. 196(7), 1244–1260 (2007)
20. Veroy, K., Patera, A.T.: Certified real-time solution of the parametrized steady incompressible Navier-Stokes equations: rigorous reduced-basis a posteriori error bounds. Int. J. Numer. Methods Fluids 47, 773–788 (2005)

Structural Optimization of Variational Inequalities Using Piecewise Constant Level Set Method

Andrzej Myśliński

Systems Research Institute, ul. Newelska 6, 01-447 Warsaw, Poland
myslinsk@ibspan.waw.pl

Abstract. The paper deals with the shape and topology optimization of the elliptic variational inequalities using the level set approach. The standard level set method is based on the description of the domain boundary as an isocountour of a scalar function of a higher dimensionality. The evolution of this boundary is governed by Hamilton-Jacobi equation. In the paper a piecewise constant level set method is used to represent interfaces rather than the standard method. The piecewise constant level set function takes distinct constant values in each subdomain of a whole design domain. Using a two-phase approximation and a piecewise constant level set approach the original structural optimization problem is reformulated as an equivalent constrained optimization problem in terms of the level set function. Necessary optimality condition is formulated. Numerical examples are provided and discussed.

Keywords: shape and topology optimization, unilateral problems, piecewise constant level set method, Uzawa method.

1 Introduction

The paper deals with the solution of a structural optimization problem for an elliptic variational inequality. This inequality governs unilateral contact between an elastic body and a rigid foundation. The structural optimization problem for the elastic body in unilateral contact consists in finding such topology of the domain occupied by the body and the shape of its boundary that the normal contact stress along the boundary of the body is minimized. The volume of the body is bounded.

In structural optimization the standard level set method [1,15] is employed in the numerical algorithms for tracking the evolution of the domain boundary on a fixed mesh and finding an optimal domain. This method is based on an implicit representation of the boundaries of the optimized structure, i.e., the position of the boundary of the body is described as an isocountour of a scalar function of a higher dimensionality. While the shape of the structure may undergo major changes the level set function remains to be simple in its topology. The evolution of the domain boundary is governed by Hamilton - Jacobi equation. The speed

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 407–416, 2013.

vector field driving the propagation of the level set function is given by the Eulerian derivative of the cost functional with respect to the variations of the free boundary. The solution of this equation requires reinitialization procedure to ensure that it is as close as possible to the signed distance function to the interface. Moreover this approach requires regularization of non-differentiable Heaviside and Dirac functions. Applications of the level set methods in structural optimization can be found, among others, in [1,6,7,8,11,12,14,17].

Recently, a piecewise constant level set method as a variant of traditional level set method has been proposed for the image segmentation [10], shape recovery [4] or elliptic inverse problems. For a domain divided into 2^N subdomains in standard level set approach is required 2^N level set functions to represent them. Piecewise constant level set method can identify an arbitrary number of subdomains using only one discontinuous piecewise constant level set function. This function takes distinct constant values on each subdomain. The interfaces between subdomains are represented implicitly by the discontinuity of a set of characteristic functions of the subdomains [10]. Comparing to the classical level set method, this method is free of the Hamilton-Jacobi equation and do not require the use of the signed distance function as the initial one. Piecewise constant level set method has been used in [18] to solve numerically topological optimization problem in plane elasticity. Moreover in [19] this method has been used to solve structural optimization problem for the Laplace equation in 2D domain.

In the paper the original structural optimization problem is approximated by a two-phase material optimization problem. Using the piecewise constant level set method this approximated problem is reformulated as an equivalent constrained optimization problem in terms of the piecewise constant level set function only. Therefore neither shape nor topological sensitivity analysis is required. During the evolution of the piecewise constant level set function small holes can be created without use of the topological derivatives. Necessary optimality condition is formulated. This optimization problem is solved numerically using the augmented Lagrangian method. Numerical examples are provided and discussed.

2 Problem Formulation

Consider deformations of an elastic body occupying two-dimensional domain Ω with the smooth boundary Γ (see Fig.1). Assume $\Omega \subset D$ where D is a bounded smooth hold-all subset of R^2. Let $E \subset R^2$ and $D \subset R^2$ denote given bounded domains. So-called hold-all domain D is assumed to possess a piecewise smooth boundary. Domain Ω is assumed to belong to the set O_l defined as follows:

$$O_l = \{\Omega \subset R^2 : \Omega \text{ is open}, \;\; E \subset \Omega \subset D, \;\; {}^{\#}\Omega^c \leq l\}, \tag{1}$$

where ${}^{\#}\Omega^c$ denotes the number of connected components of the complement Ω^c of Ω with respect to D and $l \geq 1$ is a given integer. Moreover all perturbations $\delta\Omega$ of Ω are assumed to satisfy $\delta\Omega \in O_l$. The body is subject to body forces

$f(x) = (f_1(x),\ f_2(x))$, $x \in \Omega$. Moreover, surface tractions $p(x) = (p_1(x), p_2(x))$, $x \in \Gamma$, are applied to a portion Γ_1 of the boundary Γ. We assume, that the body is clamped along the portion Γ_0 of the boundary Γ, and that the contact conditions are prescribed on the portion Γ_2, where $\Gamma_i \cap \Gamma_j = \emptyset$, $i \neq j,\ i, j = 0, 1, 2$, $\Gamma = \bar{\Gamma}_0 \cup \bar{\Gamma}_1 \cup \bar{\Gamma}_2$. We denote by $u = (u_1, u_2)$, $u = u(x)$, $x \in \Omega$, the displacement

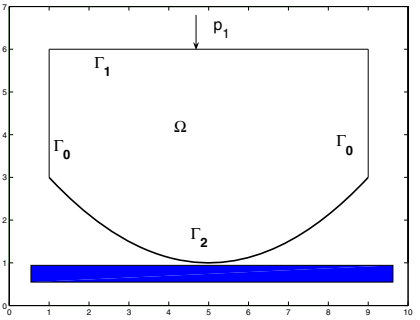


Fig. 1. Initial Domain Ω

of the body and by $\sigma(x) = \{\sigma_{ij}(u(x))\}$, $i, j = 1, 2$, the stress field in the body. Consider elastic bodies obeying Hooke's law, i.e., for $x \in \Omega$ and $i, j, k, l = 1, 2$

$$\sigma_{ij}(u(x)) = a_{ijkl}(x) e_{kl}(u(x)). \tag{2}$$

We use here and throughout the paper the summation convention over repeated indices [9]. The strain $e_{kl}(u(x))$, $k, l = 1, 2$, is defined by:

$$e_{kl}(u(x)) = \frac{1}{2}(u_{k,l}(x) + u_{l,k}(x)), \tag{3}$$

where $u_{k,l}(x) = \frac{\partial u_k(x)}{\partial x_l}$. The stress field σ satisfies the system of equations [9]

$$-\sigma_{ij}(x)_{,j} = f_i(x) \quad x \in \Omega, i, j = 1, 2, \tag{4}$$

where $\sigma_{ij}(x)_{,j} = \frac{\partial \sigma_{ij}(x)}{\partial x_j}$, $i, j = 1, 2$. The following boundary conditions are imposed

$$u_i(x) = 0 \quad \text{on} \quad \Gamma_0, \quad i = 1, 2, \tag{5}$$

$$\sigma_{ij}(x) n_j = p_i \quad \text{on} \quad \Gamma_1, \quad i, j = 1, 2, \tag{6}$$

$$u_N \leq 0, \quad \sigma_N \leq 0, \quad u_N \sigma_N = 0 \quad \text{on } \Gamma_2, \tag{7}$$

$$|\ \sigma_T\ | \leq 1, \quad u_T \sigma_T + |\ u_T\ | = 0 \quad \text{on } \Gamma_2, \tag{8}$$

where $n = (n_1, n_2)$ is the unit outward versor to the boundary Γ. Here $u_N = u_i n_i$ and $\sigma_N = \sigma_{ij} n_i n_j$, $i, j = 1, 2$, represent the normal components of the displacement u and the stress σ, respectively. The tangential components of displacement u and stress σ are given by $(u_T)_i = u_i - u_N n_i$ and $(\sigma_T)_i = \sigma_{ij} n_j - \sigma_N n_i$,

$i,j = 1,2$, respectively. $\mid u_T \mid$ denotes the Euclidean norm in R^2 of the tangent vector u_T. The results concerning the existence and uniqueness of solutions to (2)-(8) can be found in [9,16].

2.1 Variational Formulation of Contact Problem

Let us formulate contact problem (4)-(8) in variational form. Denote by V_{sp} and K the space and set of kinematically admissible displacements:

$$V_{sp} = \{z \in [H^1(\Omega)]^2 = H^1(\Omega) \times H^1(\Omega) \ : \ z_i = 0 \text{ on } \Gamma_0, \ i = 1,2\}, \tag{9}$$

$$K = \{z \in V_{sp} \ : \ z_N \leq 0 \ \text{ on } \ \Gamma_2\}. \tag{10}$$

Denote also by Λ the set

$$\Lambda = \{\zeta \in L^2(\Gamma_2) \ : \mid \zeta \mid \leq 1\}. \tag{11}$$

Variational formulation of problem (4)-(8) has the form: *find a pair* $(u,\lambda) \in K \times \Lambda$ *satisfying*

$$\int_\Omega a_{ijkl} e_{ij}(u) e_{kl}(\varphi - u) dx - \int_\Omega f_i(\varphi_i - u_i) dx - \int_{\Gamma_1} p_i(\varphi_i - u_i) ds + \int_{\Gamma_2} \lambda(\varphi_T - u_T) ds \geq 0 \ \ \forall \varphi \in K, \tag{12}$$

$$\int_{\Gamma_2} (\zeta - \lambda) u_T ds \leq 0 \quad \forall \zeta \in \Lambda, \tag{13}$$

$i,j,k,l = 1,2$. Function λ is interpreted as a Lagrange multiplier corresponding to term $\mid u_T \mid$ in equality constraint in (8) [9,16]. In general, function λ belongs to the space $H^{-1/2}(\Gamma_2)$. Here following [9] function λ is assumed to be more regular. The results concerning the existence and uniqueness of solutions to system (12)-(13) can be found, among others, in [9].

2.2 Structural Optimization Problem

Before formulating a structural optimization problem for the state system (12)-(13) let us introduce first the set U_{ad} of admissible domains. Domain Ω is assumed to satisfy the volume constraint of the form

$$Vol(\Omega) - Vol^{giv} \leq 0, \ \ Vol(\Omega) \stackrel{def}{=} \int_\Omega dx, \tag{14}$$

where the constant $Vol^{giv} = const_0 > 0$ is given. Moreover this domain is assumed to satisfy the perimeter constraint [6], [16, p. 126]

$$Per(\Omega) \leq const_1, \ \ Per(\Omega) \stackrel{def}{=} \int_\Gamma dx. \tag{15}$$

The constant $const_1 > 0$ is given. The set U_{ad} has the following form

$$U_{ad} = \{\Omega \in O_l \ : \ \Omega \text{ is Lipschitz continuous,} \\ \Omega \text{ satisfies conditions (14) and (15) } \}. \tag{16}$$

The set U_{ad} is assumed to be nonempty. In order to define a cost functional we shall also need the following set M^{st} of auxiliary functions

$$M^{st} = \{\eta = (\eta_1, \eta_2) \in [H^1(D)]^2 : \eta_i \leq 0 \ \text{ on } D, \ i = 1, 2, \\ \| \eta \|_{[H^1(D)]^2} \leq \ 1\}, \tag{17}$$

where the norm $\| \eta \|_{[H^1(D)]^2} = (\sum_{i=1}^{2} \| \eta_i \|^2_{H^1(D)})^{1/2}$. Recall from [12] the cost functional approximating the normal contact stress on the contact boundary Γ_2

$$J_\eta(u(\Omega)) = \int_{\Gamma_2} \sigma_N(u)\eta_N(x)ds, \tag{18}$$

depending on the auxiliary given bounded function $\eta(x) \in M^{st}$. σ_N and ϕ_N are the normal components of the stress field σ corresponding to a solution u satisfying system (12)-(13) and the function η, respectively.

Consider the following structural optimization problem: *for a given function* $\eta \in M^{st}$*, find a domain* $\Omega^\star \in U_{ad}$ *such that*

$$J_\eta(u(\Omega^\star)) = \min_{\Omega \in U_{ad}} J_\eta(u(\Omega)) \tag{19}$$

Lemma 1. *There exists an optimal domain* $\Omega^\star \in U_{ad}$ *to the problem (19).*

The proof follows from Šverák theorem and arguments provided in [3, Theorem 2]. Recall from [3] the class of domains O_l determined by (1) is endowed with the complementary Hausdorff topology that guarantees the class itself to be compact. The admissibility condition $^{\#}\Omega^c \leq l$ is crucial to provide the necessary compactness property of U_{ad} [3].

3 Level Set Approach

In [11,12] the standard level set method [16] is employed to solve numerically problem (19). Let $t > 0$ denote the time variable. Consider the evolution of a domain Ω under a velocity field $V = V(x, t)$. Under the mapping $T(t, V)$ we have

$$\Omega_t = T(t, V)(\Omega) = (I + tV)(\Omega), \ \ t > 0.$$

By Ω_t^- and Ω_t^+ we denote the interior and the outside of the domain Ω_t, respectively. This domain and its boundary $\partial\Omega_t$ are defined by a function $\phi = \phi(x, t) \ : R^2 \times [0, t_0) \to R$ satisfying the conditions:

$$\phi(x,t) \ = \ 0, \text{ if } \ x \in \partial\Omega_t, \ \ \phi(x,t) \ < \ 0, \text{ if } \ x \in \Omega_t^-, \\ \phi(x,t) \ > \ 0, \text{ if } \ x \in \Omega_t^+. \tag{20}$$

In the standard level set approach Heaviside function and Dirac function are used to transform integrals from domain Ω into domain D. Assume that velocity field V is known for every point x lying on the boundary $\partial\Omega_t$, i.e., such that $\phi(x,t) = 0$. Therefore the equation governing the evolution of the interface in $D \times [0, t_0]$, known as Hamilton-Jacobi equation, has the form [15]

$$\frac{\partial \phi(x,t)}{\partial t} + V(x,t) \cdot \nabla_x \phi(x,t) = 0. \tag{21}$$

Moreover $\phi(x,0) = \phi_0$ where $\phi_0(x)$ is a given function close to the signed distance function [15].

3.1 Piecewise Constant Level Set Formulation

Define a piecewise constant level set function [4,10,18,19]. Recall D is an open bounded domain in R^2. Let us assume D is partitioned into N subdomains $\{\Omega_i\}_{i=1}^N$ such that

$$D = \bigcup_{i=1}^{N} (\Omega_i \cup \partial\Omega_i), \tag{22}$$

where N is a given integer and $\partial\Omega_i$ denotes the boundary of the subdomain Ω_i. Define function $\phi = \phi(x) : D \to R$ such that [10,18,19]

$$\phi = i \ \text{ in } \Omega_i, \quad i = 1, 2, ..., N. \tag{23}$$

This function is used to identify all the phases in D. In order to ensure that there is no vacuum or overlap between different subdomains Ω_i assume function ϕ satisfies the following constraint:

$$W(\phi) = 0, \tag{24}$$

$$W(\phi) \stackrel{def}{=} (\phi - 1)(\phi - 2)...(\phi - N) = \prod_{i=1}^{N} (\phi - i). \tag{25}$$

The constraint (25) means that for every $x \in D$ there exists a unique $i \in \{1, 2..., N\}$ such that $\phi(x) = i$. Using this approach the characteristic function χ_i, $i = 1, 2, ..., N$, of the subdomain Ω_i is represented as [10,18,19]

$$\chi_i = \frac{1}{\alpha_i} \prod_{j=1, j\neq i}^{N} (\phi - j) \ \text{ and } \ \alpha_i = \prod_{k=1, k\neq i}^{N} (i - k), \tag{26}$$

i.e., it is constructed using one level set function ϕ only. Each characteristic function χ_i is expressed as a product of linear factors of the form $(\phi - j)$ with the ith factor omitted. Therefore as long as (23) holds, $\chi_i(x) = 1$ for $x \in \Omega_i$ and equals zero elsewhere. Any piecewise constant density function $\rho = \rho(x) : D \to R^2$ defined as

$$\rho(x) = \begin{cases} \epsilon \text{ if } x \in D \setminus \bar{\Omega} \ , \\ 1 \text{ if } x \in \Omega, \end{cases} \tag{27}$$

where $\epsilon > 0$ is a small constant, can be constructed as a weighted sum of the characteristic functions χ_i. Denoting by $\{\rho_i\}_{i=1}^N$ a set of real scalars, we can represent a piecewise constant function ρ taking these N distinct constant values by

$$\rho(x) = \sum_{i=1}^{N} \rho_i \chi_i(\phi(x)). \tag{28}$$

We confine to consider a two-phase problem in the domain D, i.e., we set $N = 2$. Therefore

$$\chi_1(x) = 2 - \phi(x) \ \text{ and } \ \chi_2(x) = \phi(x) - 1, \tag{29}$$

$$\rho(x) = \rho_1\chi_1(x) + \rho_2\chi_2(x) = (1-\epsilon)\phi(x) + 2\epsilon - 1. \tag{30}$$

Moreover function (25) takes the form

$$W(\phi) = (\phi - 1)(\phi - 2). \tag{31}$$

Using (23) as well as (30) the structural optimization problem (19) can be transformed into the following one: *find* $\phi \in U_{ad}^{\phi}$ *such that*

$$\min_{\phi \in U_{ad}^{\phi}} J_\eta(\phi) = \int_{\Gamma_2} \rho(\phi)\sigma_N(u_\epsilon)\eta_N ds, \tag{32}$$

where the set U_{ad}^{ϕ} of the admissible functions is given as

$$U_{ad}^{\phi} = \{\phi \in H^1(D) : Vol(\phi) - Vol^{giv} \le 0,\ W(\phi) = 0,\ Per(\phi) \le const_1\}, \tag{33}$$

$$Vol(\phi) \stackrel{def}{=} \int_\Omega \rho(\phi)dx,\ W(\phi) \stackrel{def}{=} (\phi-1)(\phi-2),\ Per(\phi) \stackrel{def}{=} \int_\Omega \mid \nabla\phi \mid dx. \tag{34}$$

The element $(u_\epsilon, \lambda_\epsilon) \in K \times \Lambda$ satisfies the state system (12)-(13) in the domain D rather than Ω:

$$\int_D \rho(\phi)a_{ijkl}e_{ij}(u_\epsilon)e_{kl}(\varphi - u_\epsilon)dx - \int_D \rho(\phi)f_i(\varphi_i - u_{\epsilon i})dx - \int_{\Gamma_1} p_i(\varphi_i - u_{\epsilon i})ds + \int_{\Gamma_2} \lambda_\epsilon(\varphi_T - u_{\epsilon T})ds \ge 0 \ \ \forall\varphi \in K, \tag{35}$$

$$\int_{\Gamma_2} (\zeta - \lambda_\epsilon)u_{\epsilon T}ds \le 0 \quad \forall\zeta \in \Lambda. \tag{36}$$

Lemma 2. *There exists an optimal solution* $\phi \in H^1(D)$ *to the optimization problem (32)-(36).*

The proof follows from the presence of the regularization term in (33) and its lower semicontinuity in $L^1(D)$ (see [2, Theorem 3.2.1, p. 75]). For the similar approach see [4].

3.2 Necessary Optimality Conditions

In order to formulate the necessary optimality condition for the optimization problem (32)-(36) we introduce the Lagrangian $L(\phi, \tilde{\lambda}) = L(\phi, u_\epsilon, \lambda_\epsilon, p^a, q^a, \tilde{\lambda})$:

$$L(\phi,\tilde{\lambda}) = J_\eta(\phi) + \int_D \rho(\phi) a_{ijkl} e_{ij}(u_\epsilon) e_{kl}(p^a) dx - \int_D \rho(\phi) f_i p_i^a dx - \int_{\Gamma_1} p_i p_i^a ds + \int_{\Gamma_2} \lambda_\epsilon p_T^a ds + \int_{\Gamma_2} q^a u_{\epsilon T} ds + \tilde{\lambda} c(\phi) + \sum_{i=1}^{3} \frac{1}{2\mu_i} c_i^2(\phi), \quad (37)$$

where $i,j,k,l = 1,2$, $\tilde{\lambda} = \{\tilde{\lambda}_i\}_{i=1}^3$, $c(\phi) = \{c_i(\phi)\}_{i=1}^3 = [Vol(\phi), W(\phi), Per(\phi)]^T$, $c^T(\phi)$ denotes a transpose of $c(\phi)$, $\mu_m > 0$, $m = 1,2,3$, is a given real. Element $(p^a, q^a) \in K_1 \times \Lambda_1$ denotes an adjoint state defined as follows:

$$\int_D \rho(\phi) a_{ijkl} e_{ij}(\eta + p^a) e_{kl}(\varphi) dx + \int_{\Gamma_2} q^a \varphi_T ds = 0 \quad \forall \varphi \in K_1, \quad (38)$$

$$\int_{\Gamma_2} \zeta(p_T^a + \eta_T) ds = 0 \ \ \forall \zeta \in \Lambda_1. \quad (39)$$

The sets K_1 and Λ_1 are given by

$$K_1 = \{\xi \in V_{sp} \ : \ \xi_N = 0 \ \text{ on } A^{st} \ \}, \quad (40)$$

$$\Lambda_1 = \{\zeta \in \Lambda \ : \ \zeta(x) = 0 \text{ on } B_1 \cup B_2 \cup B_1^+ \cup B_2^+\}, \quad (41)$$

while the coincidence set $A^{st} = \{x \in \Gamma_2 \ : \ u_N + v = 0\}$. Moreover $B_1 = \{x \in \Gamma_2 : \lambda(x) = -1\}$, $B_2 = \{x \in \Gamma_2 : \lambda(x) = +1\}$, $\tilde{B}_i = \{x \in B_i : u_N(x) + v = 0\}$, $i = 1,2$, $B_i^+ = B_i \setminus \tilde{B}_i$, $i = 1,2$. The derivative of the Lagrangian L with respect to ϕ has the form:

$$\frac{\partial L}{\partial \phi}(\phi,\tilde{\lambda}) = \int_D \rho'(\phi)[a_{ijkl} e_{ij}(u_\epsilon) e_{kl}(p^a + \eta) - f(p^a + \eta)] dx + \tilde{\lambda} c'(\phi) + \sum_{i=1}^{3} \frac{1}{\mu_i} c(\phi) c'(\phi), \quad (42)$$

where $\rho'(\phi) = 1 - \epsilon$, $c'(\phi) = [Vol'(\phi), W'(\phi), Per'(\phi)]$ and

$$Vol'(\phi) = 1, \ \ W'(\phi) = 2\phi - 3, \ \ Per'(\phi) = \quad (43)$$

$$\chi_{\{\partial\Omega = const_0\}} \max\{0, -\nabla \cdot (\frac{\nabla\phi}{\mid \nabla\phi \mid})\} - \chi_{\{\partial\Omega > const_0\}} \nabla \cdot (\frac{\nabla\phi}{\mid \nabla\phi \mid}). \quad (44)$$

Using (38)-(44) we can formulate the necessary optimality condition:

Lemma 3. *If $\hat{\phi} \in U_{ad}^{\phi}$ is an optimal solution to the problem (32)-(36) than there exists Lagrange multiplier $\tilde{\lambda}^\star = (\tilde{\lambda}_1^\star, \tilde{\lambda}_2^\star, \tilde{\lambda}_3^\star) \in R^3$ such that $\tilde{\lambda}_1^\star, \tilde{\lambda}_3^\star \geq 0$ satisfying*

$$L(\hat{\phi},\tilde{\lambda}) \leq L(\hat{\phi},\tilde{\lambda}^\star) \leq L(\phi,\tilde{\lambda}^\star) \ \ \forall (\phi,\tilde{\lambda}) \in U_{ad}^{\phi} \times R^3. \quad (45)$$

Proof follows from standard arguments [5,16]. Recall [9,16] condition (45) implies that for all $\phi \in U_{ad}^{\phi}$ and $\tilde{\lambda} \in R^3$

$$\frac{\partial L(\hat{\phi}, \tilde{\lambda})}{\partial \phi} \geq 0 \ \text{ and } \ \frac{\partial L(\phi, \tilde{\lambda}^{\star})}{\partial \tilde{\lambda}} \leq 0. \tag{46}$$

4 Numerical Experiments

The optimization problem (32)-(36) is discretized using the finite difference approximation [9,15,19]. The discretized structural optimization problem (32)-(36) is solved numerically. We employ Uzawa type algorithm to solve numerically optimization problem (32)-(36). The algorithm is programmed in Matlab environment. For details of numerical implementation see [13]. As an example a body occupying 2D domain

$$\Omega = \{(x_1, x_2) \in R^2 : 0 \leq x_1 \leq 8 \ \wedge \ 0 < v(x_1) \leq x_2 \leq 4\}, \tag{47}$$

is considered. The boundary Γ of the domain Ω is divided into three pieces

$$\begin{aligned} \Gamma_0 &= \{(x_1, x_2) \in R^2 : x_1 = 0, 8 \ \wedge \ 0 < v(x_1) \leq x_2 \leq 4\}, \\ \Gamma_1 &= \{(x_1, x_2) \in R^2 : 0 \leq x_1 \leq 8 \ \wedge \ x_2 = 4\}, \\ \Gamma_2 &= \{(x_1, x_2) \in R^2 : 0 \leq x_1 \leq 8 \ \wedge \ v(x_1) = x_2\}. \end{aligned} \tag{48}$$

The domain Ω and the boundary Γ_2 depend on the function v given as in [13]. Fig. 2 presents the obtained optimal domain. The areas with low values of density function appear in the central part of the body and near the fixed edges. The obtained normal contact stress is almost constant along the optimal shape boundary and has been significantly reduced comparing to the initial one.

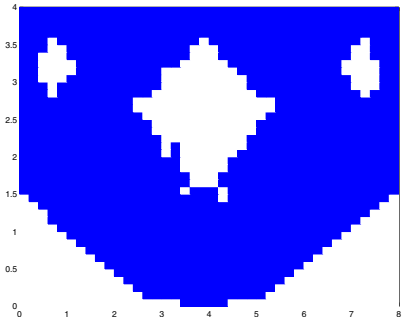


Fig. 2. Optimal domain $\Omega^{\star}$

References

1. Allaire, G., Jouve, F., Toader, A.: Structural Optimization Using Sensitivity Analysis and a Level Set Method. Journal of Computational Physics 194, 363–393 (2004)
2. Aubert, G., Kornprobst, P.: Mathematical Problems in Image Processing. Springer (2006)
3. Chambolle, A.: A density result in two-dimensional linearized elasticity and applications. Arch. Ration. Mech. Anal. 167, 211–233 (2003)
4. De Cezaro, A., Leitão, A.: Level-set approaches of L^2-type for recovering shape and contrast in ill-posed problems. Inverse Problems in Science and Engineering 20(4), 571–587 (2011)
5. Delfour, M.C., Zolesio, J.P.: Shapes and Geometries: Analysis, Differential Calculus and Optimization. SIAM Publications, Philadelphia (2001)
6. Dijk van, N.P., Yoon, G.H., Keulen van, F., Langelaar, M.: A level-set based topology optimization using the element connectivity parameterization method. Struct. Multidisc. Optim. 42, 269–282 (2010)
7. Fulmański, P., Laurain, A., Scheid, J.F., Sokołowski, J.: A Level Set Method in Shape and Topology Optimization for Variational Inequalities. Int. J. Comp. Math. 85, 1491–1514 (2008)
8. He, L., Kao, Ch.Y., Osher, S.: Incorporating topological derivatives into shape derivatives based level set methods. Journal of Computational Physics 225, 891–909 (2007)
9. Haslinger, J., Mäkinen, R.: Introduction to Shape Optimization. Theory, Approximation, and Computation. SIAM Publications, Philadelphia (2003)
10. Lie, J., Lysaker, M., Tai, X.C.: A piecewiase constant level set framework. International Journal of Numerical Analysis and Modeling 2(4), 422–438 (2005)
11. Myśliński, A.: Level Set Method for Optimization of Contact Problems. Engineering Analysis with Boundary Elements 32, 986–994 (2008)
12. Myśliński, A.: Radial Basis Function Level Set Method for Structural Optimization. Control and Cybernetics 39(3), 627–645 (2010)
13. Myśliński, A.: Structural Optimization of Elastic Contact Problems using Piecewise Constant Level Set Method. In: Yamakawa, H. (ed.) CD-ROM Proceedings of the 9th World Congress on Structural and Multidisciplinary Optimization, ISSMO, Shizuoka, Japan, June 13 - 17 (2011)
14. Yamada, T., Izui, K., Nishiwaki, S., Takezawa, A.: A topology optimization method based on the level set method incorporating a fictitious interface energy. Comput. Methods Appl. Mech. Engrg. 199(45-48), 2876–2891 (2010)
15. Osher, S., Fedkiw, R.: Level Set Methods and Dynamic Implicit Surfaces. Springer, New York (2003)
16. Sokołowski, J., Zolesio, J.P.: Introduction to Shape Optimization. Shape Sensitivity Analysis. Springer, Berlin (1992)
17. Wang, S.Y., Lim, K.M., Khao, B.C., Wang, M.Y.: An extended level set method for shape and topology optimization. Journal of Computational Physics 221, 395–421 (2007)
18. Wei, P., Wang, M.Y.: Piecewise constant level set method for structural topology optimization. International Journal for Numerical Methods in Engineering 78, 379–402 (2009)
19. Zhu, S., Wu, Q., Liu, C.: Shape and topology optimization for elliptic boundary value problems using a piecewise constant level set method. Applied Numerical Mathematics 61, 752–767 (2011)

Numerical Shape Optimization via Dynamic Programming

Jan Pustelnik

University of Lodz, Fac. of Math. & Computer Science,
Banacha 22, 90-128 Lodz, Poland

Abstract. In this paper we describe a novel framework for finding numerical solutions to a wide range of shape optimization problems. It is based on classical dynamic programming approach augmented with discretization of the space of trajectories and controls. This allows for straightforward algorithmic implementation. This method has been used to solve a well known problem called the "dividing tube problem", a state problem related to fluid mechanics, that requires simultaneous topology and shape optimization in case of elastic contact problems and involves solving the Navier-Stokes equations for viscous incompressible fluids.

Keywords: dynamic programming, numerical approximation, contact problem, shape optimization, sufficiency optimality condition, structural optimization, topological derivative, stationary Navier-Stokes equations.

1 Introduction

In the paper, as a model problem, we consider state problems related to fluid mechanics, namely the Navier-Stokes equations for viscous incompressible fluids. The main problem is to search for optimal shape of a given objective. For an incompressible fluid, conservation laws for momentum and mass are assumed to be in force. The displacement field of the body is governed by the Reynolds-averaged Navier-Stokes (RANS) equations with an algebraic mixing length turbulence. The volume of the body is assumed to be bounded.

The results pertaining the existence, regularity and finite-dimensional approximation of solutions to mentioned problems are given in [4], [5]. The primal–dual algorithms for numerical solving of contact problems were developed in [6], [8]. In the course of solution the necessary optimality condition for simultaneous shape and topology optimization is formulated, while the shape and topological derivatives are employed, what stays close to classical optimization problems and gives sufficient optimality conditions. It is a different approach than the one applied e.g. in [2] where the notion of topological derivative and results concerning its application in optimization of elastic structures are reported.

We describe a new numerical algorithm for that optimization problem.

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 417–426, 2013.

2 General Shape Optimization Problem

We consider the following shape optimization problem, which is being analyzed and subsequently solved in [3]:

$$\text{minimize} \quad J(\Omega) = \int_{\Omega} L(x, u(x), \nabla u(x))dx \tag{1}$$

subject to

$$\Omega \in \Theta, \quad \mathbb{A}u(x) = f(x, u(x)), \quad \mathbb{B}u(x) = \phi(x) \quad \text{on} \quad \partial\Omega \tag{2}$$

where Θ is a certain family of bounded with $C^{0,1}$ boundary domains of $D \subset R^n$ which will be defined precisely in subsection 2.1 and $\mathbb{A}$ is a differential operator e.g. defining Navier-Stokes equations and $\mathbb{B}$ an operator acting on the boundary. We assume that $L : R^n \times R \times R^n \to R$ is Lipschitz continuous with respect to all variables, $f : R^n \times R^m \to R^m$ is continuous and Lipschitz continuous with respect to last variable, $\phi(\cdot)$ is continuous on $\partial\Omega$.

2.1 Reduction of Shape Optimization Problem to Classical Control Problem

We will summarize in this subchapter the results of research published in [3] in order to introduce relevant objects on which the presented numerical method operates.

Let U be a given nonempty, compact set in $C^{0,1}$ of surfaces defined on $E \subset R^{n-1}$. We assume that each supremum of each subfamily of U also belongs to U as well as finite concatenation of element of U belongs to U. Let $U \ni v \to \Omega(v)$ be a given family of simply connected domains in $D \subset R^n$ with $C^{0,1}$ boundary such that some fixed part of their boundary is changing and is a surface v from U. We assume that $\Omega(v)$ depends in a smooth way on v and that there exists a $v_{\max} \in U$ such that $\Omega(v) \subset \Omega(v_{\max})$, for all $v \in U$ and there exists a $v_{\min} \in U$ such that $\Omega(v_{\min}) \subset \Omega(v)$, for all $v \in U$. Let us denote the part of the boundary $\partial\Omega(v_{\min})$ corresponding to the surface $v_{\min}$ as Γ_0 while that corresponding the surface v as $\Gamma(v)$. We have the following BVP:

Find $z_{\max} \in C^{1,k}(\Omega(v_{\max}))$ such that $\Delta z(x) = 0$ in $\Omega(v_{\max})\backslash\bar{\Omega}(v_{\min})$, $z(x) = 0$ on Γ_0, $z(x) = 1$ on $\Gamma(v_{\max})$. Next for $v \in U$, $v \neq v_{\min}$, find $z \in C^{1,k}(\Omega(v)\backslash\bar{\Omega}(v_{\min}))$ such that: $\Delta z(x) = 0$ in $\Omega(v_{\max})\backslash\bar{\Omega}(v_{\min})$, $z(x) = 0$ on Γ_0, $z(x) = z_{\max}(x)$ on $\Gamma(v)$. Solutions to the BVP above belong to $C^{1,k}(\Omega(v)\backslash\bar{\Omega}(v_{\min}))$and in fact they are restrictions of $z_{\max}$ to $\Omega(v)\backslash\bar{\Omega}(v_{\min})$. Because $z(x)$ depends (in a continuous way) on v, we will use the notation $z(x, v)$.We define the family Θ of sets over which the problem (1)-(2) is considered as: $\Theta = \{\Omega(v) : v \in U\}$. The sets from Θ are called *admissible sets* for problem (1)-(2). For a given $\Omega(v)\backslash\bar{\Omega}(v_{\min})$, we introduce the field and the deformation: $V(x, v) = ||\nabla z(x, v)||^{-2}\nabla z(x, v), T(w, v) = x(s, w, v)$, $s \in [0, 1]$, where $x(\cdot, w, v)$ is a solution to $\frac{d}{ds}x(s, w, v) = V(x(s, w, v), v)$, $s \in [0, 1]$, $x(0, w, v) = w$,

$w \in \Gamma_0$. Notice, that for a given fixed $w \in \Gamma_0$, the point $x(1, w, v)$ belongs to $\Gamma(v)$.

For a given control $v \in U$ we can write:

$$\frac{d}{ds}x(s, w, v) = V(x(s, w, v), v), \quad s \in [0, 1],\ x(0, w, v) = w. \tag{3}$$

Then the boundary $\Gamma(v)$ is the image of Γ_0 by the map $x(1, \cdot, v)$. Thus, for a given $v \neq v_{\min}$, we have an alternative definition of the $\Omega(v)\backslash\bar{\Omega}(v_{\min})$: $\Omega(v)\backslash\bar{\Omega}(v_{\min}) = \{x : x = x(s, w, v), 0 < s < 1,\ w \in \Gamma_0\}$. This means that we can construct and study some objects over the set $\Omega(v)$ with the help of the family $F(v)$: $F(v) = \{x(s, w, v) : 0 < s < 1,\ w \in \Gamma_0\}$. The original functional $J(\Omega)$ in terms of the family $F(v)$ can be rewritten as $J(F(v)) = \mathbf{I}(v)$,

$$\begin{aligned}\mathbf{I}(v) &= \int_{\Omega(v_{\min})} L(y, u(y), \nabla u(y))dy + \int_{\Omega(v)\backslash\bar{\Omega}(v_{\min})} L(x, u(x), \nabla u(x))dx \\ &= \int_{\Omega(v_{\min})} L(y, u(y), \nabla u(y))dy + \int_0^1 \int_{\Gamma_0} \hat{L}(x(s, w, v))dwds,\end{aligned}$$

where $\hat{L}(x(s, w, v)) = L(x(s, w, v), u(x(s, w, v)), \nabla u(x(s, w, v)))\left|\frac{\partial}{\partial s}x\ \frac{\partial}{\partial w}x\right|$.

Therefore we are able to reduce the original shape optimal control problem to classical optimal control problem (P): minimize $\mathbf{I}(v)$ subject to $\frac{d}{ds}x(s, w, v) = V(x(s, w, v), v),\ s \in [0, 1],\ x(0, w, v) = w,\ w \in \Gamma_0,\ v \in U$,

$$\Omega(v) \in \Theta, \mathbb{A}u(x) = f(x, u(x)),\ x \in \Omega(v). \tag{4}$$

In order to formulate any sufficient optimality conditions for this problem we apply classical dynamic programming scheme.

2.2 Dynamic Programming Approach as a Method to Solution of (P)

Let us take any $x \in \Omega(v_{\max})\backslash\bar{\Omega}(v_{\min})$ and denote by U_x a subfamily of U such that $x \in v$ for each $v \in U_x$. Next denote by $v_x = \max U_x$, where the maximum over U_x means that $\Omega(v) \subset \Omega(v_x)$ for all $v \in U_x$. By our assumption on U, v_x exists and $v_x \in U$. Put $\bar{U}_x = \{v \in U :\ \Omega(v) \subset \Omega(v_x)\}$. By (3) for each $v \in U_x$ there is a trajectory $x(\cdot, w, v)$ such that $x = x(1, w_v, v)$, for some $w_v \in \Gamma_0$. The problem (P) falls into the category of Lagrange control problems treated in many books (e.g. [1]). Following Chapter IV of this book we define a value function for (P), for $x \in \Omega(v_{\max})$:

$$S(x) = \inf\left\{\int_{\Omega(v_{\min})} L(y, u(y), \nabla u(y))dy + \int_0^1 \int_{\Gamma_0} \hat{L}(x(s, w, v))dwds\right\}, \tag{5}$$

where infimum in (5) is taken over all pairs $(x(\cdot, w, v), v)$ satisfying $\frac{d}{ds}x(s, w, v) = V(x(s, w, v), v),\ s \in [0, 1],\ v \in \bar{U}_x,\ w \in \Gamma_0$ and for $v \in U_x$, $x(1, w_v, v) = x$, for

some $w_v \in \Gamma_0$. Each pair $(x(\cdot, w, v), v)$ satisfying these equations will be called admissible for the point $x \in \Omega(v_{\max})\backslash\bar{\Omega}(v_{\min})$. However, in practice, we cannot expect that $S(\cdot)$ is of C^1 in $\Omega(v_{\max})\backslash\bar{\Omega}(v_{\min})$, this is why we are interested in numerical approximation of $S(\cdot)$. Therefore, we shall look for ε-value function: $S_\varepsilon(\cdot)$. For given $\varepsilon > 0$ we call $S_\varepsilon : \Omega(v_{\max})\backslash\bar{\Omega}(v_{\min}) \to R$, ε-value function if

$$S(x) \leq S_\varepsilon(x) \leq S(x) + \varepsilon,\ x \in \Omega(v_{\max})\backslash\bar{\Omega}(v_{\min}). \tag{6}$$

It is clear that there exists infinitely many ε-value functions $S_\varepsilon(\cdot)$.

3 Numerical Approximation of the Value Function

This section is an adaptation of the method developed by Pustelnik in his Ph.D. thesis [7] for numerical approximation of value function for Bolza problem from optimal control theory.

Let us define the following set $T = \left\{x :\ x \in \Omega(v_{\max})\backslash\bar{\Omega}(v_{\min})\right\}$. Since $\Omega(v_{\max})\backslash\bar{\Omega}(v_{\min})$ is bounded, the set $\bar{T}$ is compact. Let $T \ni x \to g(x)$ be an arbitrary function of class C^1 in $\bar{T}$ such that $g(x) = c$, $x \in \Gamma_0$, where c is some constant which will be determined later. For a given function g, we define $(x, v) \to G_g(x, v)$ as

$$G_g(x, v) = g_x(x)V(x, v) - \int_{\Gamma_0} \hat{L}(x(1, w, v))dw, \tag{7}$$

$v \in \bar{U}_x$, where $x(\cdot, w, v)$, u are defined as previously. Next, we define the function $x \to F_g(x)$ as

$$F_g(x) = \max\left\{G_g(x, v) :\ v \in \bar{U}_x\right\}. \tag{8}$$

Note that by the assumptions on L and V, the function F_g is continuous in T. By the continuity of F_g and compactness of $\bar{T}$, there exist k_d and k_g such that $k_d \leq F_g(x) \leq k_g$ for $x \in \Omega(v_{\max})\backslash\bar{\Omega}(v_{\min})$.

3.1 Definition of Covering of T

Let $\eta > 0$ be fixed and $\{q_j^\eta\}_{j\in\mathbb{Z}}$ be a sequence of real numbers such that $q_j^\eta = j\eta$, $j \in \mathbb{Z}$ ($\mathbb{Z}$ - set of integers). Denote

$$J = \{j \in \mathbb{Z} :\ \text{there is}\ \ x \in T,\ j\eta < F_g(x) \leq (j+1)\eta\},$$

Next, let us divide the set T into the sets $P_j^{\eta,g}$, $j \in J$, as follows

$$P_j^{\eta,g} := \left\{x \in T : q_j^\eta < F_g(x) \leq q_{j+1}^\eta\right\},\ j \in J.$$

3.2 Discretization of F_g

Define in T a function

$$h^{\eta,g}(x) = -q^{\eta}_{j+1},\ x \in P^{\eta,g}_j, \quad j \in J. \tag{9}$$

Then, by the construction of the covering of T, we have

$$0 \le F_g(x) + h^{\eta,g}(x) \le \eta,\ x \in T. \tag{10}$$

Let $(x(\cdot, w, v), v)$ be any admissible pair with the trajectory defined in $[0, 1]$, starting at the point $x(0, w, v))$, $w \in \Gamma_0$ fixed. We show that there exists an increasing sequence of m points $\{\tau_i\}_{i=1,\dots,m}$, $\tau_1 = 0$, $\tau_m = 1$, such that for $\tau \in [\tau_i, \tau_{i+1}]$

$$|F_g(x(\tau_i, w, v)) - F_g(x(\tau, w, v))| \le \frac{\eta}{2},\ i = 2, \dots, m-2, \tag{11}$$

$$|F_g(x(\tau_2, w, v)) - F_g(x(\tau, w, v))| \le \frac{\eta}{2},\ \tau \in (\tau_1, \tau_2],$$

$$|F_g(x(\tau_{m-1}, w, v)) - F_g(x(\tau, w, v))| \le \frac{\eta}{2},\ \tau \in [\tau_{m-1}, \tau_m).$$

Indeed, it is a direct consequence of two facts: Lipschitz continuity of $x(\cdot, w, v)$ with a common Lipschitz constant and continuity of F_g. From (11) we infer that for each $i \in \{1, \dots, m-1\}$ if $x(\tau_i, w, v) \in P^{\eta,g}_j$ for a certain $j \in J$, then we have for $\tau \in [\tau_i, \tau_{i+1})$

$$x(\tau, w, v) \in P^{\eta,g}_{j-1} \cup P^{\eta,g}_j \cup P^{\eta,g}_{j+1}.$$

Define

$$h^{\eta,g}(x(\tau_1, w, v)) = h^{\eta,g}(x(\tau, w, v)) \text{ for some } \tau \text{ near } \tau_1,$$
$$h^{\eta,g}(x(\tau_m, w, v)) = h^{\eta,g}(x(\tau, w, v)) \text{ for some } \tau \text{ near } \tau_m.$$

Thus for $\tau \in [\tau_i, \tau_{i+1}]$

$$h^{\eta,g}(x(\tau_i, w, v)) - \eta \le h^{\eta,g}(x(\tau, w, v)) \le h^{\eta,g}(x(\tau_i, w, v)) + \eta, \tag{12}$$

and so, for $i \in \{2, \dots, m-1\}$

$$h^{\eta,g}(x(\tau_i, w, v)) - h^{\eta,g}(x(\tau_{i-1}, w, v)) = \eta^i_{x(\cdot,w,v)}, \tag{13}$$

where $\eta^i_{x(\cdot,w,v)}$ is equal to $-\eta$ or 0 or η. Integrating (12) we get

$$\Big|\int_0^1 h^{\eta,g}(x(\tau, w, v))d\tau - \sum_{i\in\{1,\dots,m-1\}} [h^{\eta,g}(x(\tau_i, w, v))(\tau_{i+1} - \tau_i)]\Big| \le \eta.$$

By using the formula above and the following simple arithmetic transformations

$$\sum_{i\in\{2,\dots,m-1\}} [h^{\eta,g}(x(\tau_i, w, v)) - h^{\eta,g}(x(\tau_{i-1}, w, v))](\tau_m - \tau_i)$$
$$= \sum_{i\in\{1,\dots,m-1\}} [h^{\eta,g}(x(\tau_i, w, v))(\tau_{i+1} - \tau_i)] - h^{\eta,g}(x(\tau_1, w, v))(\tau_1 - \tau_m),$$

we obtain

$$\begin{aligned}
&\sum_{i\in\{2,\dots,m-1\}} [h^{\eta,g}(x(\tau_i,w,v)) - h^{\eta,g}(x(\tau_{i-1},w,v))](\tau_m-\tau_i)\\
&+ h^{\eta,g}(x(\tau_1,w,v))(\tau_m-\tau_1) - \eta(\tau_m-\tau_1)\\
&\le \int_{\tau_1}^{\tau_m} h^{\eta,g}(x(\tau,w,v))d\tau\\
&\le \sum_{i\in\{2,\dots,m-1\}} [h^{\eta,g}(x(\tau_i,w,v)) - h^{\eta,g}(x(\tau_{i-1},w,v))](\tau_m-\tau_i)\\
&+ h^{\eta,g}(x(\tau_1,w,v))(\tau_m-\tau_1) + \eta(\tau_m-\tau_1).
\end{aligned}$$

and, taking into account (13), we infer that

$$\begin{aligned}
&\sum_{i\in\{2,\dots,m-1\}} \eta^i_{x(\cdot,w,v)}(\tau_m-\tau_i) + h^{\eta,g}(x(\tau_1,w,v))(\tau_m-\tau_1) - \eta(\tau_m-\tau_1)\\
&\le \int_{\tau_1}^{\tau_m} h^{\eta,g}(x(\tau,w,v))d\tau \qquad (14)\\
&\le \sum_{i\in\{2,\dots,m-1\}} \eta^i_{x(\cdot,w,v)}(\tau_m-\tau_i) + h^{\eta,g}(x(\tau_1,w,v))(\tau_m-\tau_1) + \eta(\tau_m-\tau_1).
\end{aligned}$$

We would like to stress that (14) is very useful from numerical point of view: we can estimate the integral $h^{\eta,g}(\cdot,\cdot)$ along any trajectory $x(\cdot,w,v)$ as a sum of finite number of values, where each value consists of a number from the set $\{-\eta,0,\eta\}$ multiplied by $\tau_m-\tau_i$. Moreover, for two different trajectories: $x(\cdot,w^1,v^1)$, $x(\cdot,w^2,v^2)$, the expressions

$$\sum_{i\in\{2,\dots,m-1\}} \eta^i_{x(\cdot,w^1,v^1)}(\tau_m-\tau_i) + h^{\eta,g}(x(\tau_1,w^1,v^1))(\tau_m-\tau_1)$$

and

$$\sum_{i\in\{2,\dots,m-1\}} \eta^i_{x(\cdot,w^2,v^2)}(\tau_m-\tau_i) + h^{\eta,g}(x(\tau_1,w^2,v^2))(\tau_m-\tau_1)$$

are identical if

$$h^{\eta,g}(x(\tau_1,w^1,v^1)) = h^{\eta,g}(x(\tau_1,w^2,v^2)) \qquad (15)$$

and

$$\eta^i_{x(\cdot,w^1,v^1)} = \eta^i_{x(\cdot,w^2,v^2)} \text{ for all } i\in\{2,\dots,m-1\}. \qquad (16)$$

The last one means that in the set B of all trajectories $x(\cdot,w,v)$, $w\in\Gamma_0$, $v\in U$, we can introduce an equivalence relation r: we say that two trajectories $x(\cdot,w^1,v^1)$ and $x(\cdot,w^2,v^2)$, $w^1,w^2\in\Gamma_0$, $v^1,v^2\in U$ are equivalent if they satisfy (15) and (16). We denote the set of all disjoint equivalence classes by B_r. The cardinality of B_r, denoted by $||B_r||$, is finite and bounded from above by 3^{m+1}.

Define

$$\begin{aligned}
X = \big\{ &x=(x_1,\dots,x_{m-1}) : x_1=0,\ x_i=\eta^i_{x^j},\\
&i=2,\dots,m-1,\ x^j\in B_r,\ j=1,\dots,\ ||B_r||\big\}.
\end{aligned}$$

It is easy to see that the cardinality of X is finite.

The considerations above allow us to estimate the approximation of the value function.

Theorem 1. *We have the following estimation*

$$\begin{aligned}
&\min_{x\in B_r, w_0\in\Gamma_0}\left(-\int_{\tau_1}^{\tau_m} h^{\eta,g}(x(\tau,w_0,v))d\tau - g(x(\tau_m,w_0,v))\right)\\
&\le \max_{x\in B_r}\left\{\int_{\tau_1}^{\tau_m}\left(-\int_{\Gamma_0}\hat{L}(x(s,w,v))dw\right)ds - g(x(\tau_1,w_0,v))\right\}\\
&\le \max_{x\in B_r, w_0\in\Gamma_0}\left(-\int_{\tau_1}^{\tau_m} h^{\eta,g}(x(\tau,w_0,v))d\tau - g(x(\tau_m,w_0,v))\right)+\eta(\tau_m-\tau_1),
\end{aligned}$$

where u is a solution to (4) for $\Omega(v)$.

Proof. By inequality (10) $0 \le F_g(x) + h^{\eta,g}(x) \le \eta$ we have $-h^{\eta,g}(x) \le F_g(x) \le -h^{\eta,g}(x)+\eta$. Integrating the last inequality along any $x(\cdot, w_0, \tilde{v})$ in the interval $[\tau_1, \tau_m]$ we get

$$\begin{aligned}
&-\int_{\tau_1}^{\tau_m} h^{\eta,g}(x(\tau,w_0,\tilde{v}))d\tau\\
&\le \max_{v\in U}\int_{\tau_1}^{\tau_m}\left(g_x(x(\tau,w_0,\tilde{v}))V(x(\tau,w_0,\tilde{v}),v) - \int_{\Gamma_0}\hat{L}(x(\tau,w,v))dw\right)d\tau\\
&\le -\int_{\tau_1}^{\tau_m} h^{\eta,g}(x(\tau,w_0,\tilde{v}))d\tau+\eta(\tau_m-\tau_1).
\end{aligned}$$

Hence, we get two inequalities

$$\begin{aligned}
&\min_{x\in B_r, w_0\in\Gamma_0}\left(-\int_{\tau_1}^{\tau_m} h^{\eta,g}(x(\tau,w_0,\tilde{v}))d\tau - g(x(\tau_m,w_0,\tilde{v}))\right)\\
&\le \min_{x\in B_r, w_0\in\Gamma_0}\max_{v\in U}\int_{\tau_1}^{\tau_m}(-g(x(\tau_m,w_0,\tilde{v}))\\
&\quad +g_x(x(\tau,w_0,\tilde{v}))V(x(\tau,w_0,\tilde{v}),v) - \int_{\Gamma_0}\hat{L}(x(\tau,w,v))dw\Big)d\tau
\end{aligned}$$

and

$$\begin{aligned}
&\max_{x\in B_r, w_0\in\Gamma_0}\max_{v\in U}\int_{\tau_1}^{\tau_m}(-g(x(\tau_m,w_0,\tilde{v}))\\
&+g_x(x(\tau,w_0,\tilde{v}))V(x(\tau,w_0,\tilde{v}),v) - \int_{\Gamma_0}\hat{L}(x(\tau,w,v))dw\Big)d\tau\\
&\le \max_{x\in B_r, w_0\in\Gamma_0}\left(-\int_{\tau_1}^{\tau_m} h^{\eta,g}(x(\tau,w_0,\tilde{v}))d\tau - g(x(\tau_m,w_0,\tilde{v}))\right)+\eta(\tau_m-\tau_1).
\end{aligned}$$

As a consequence of the above we get

$$\begin{aligned}
&\min_{x\in B_r, w_0\in\Gamma_0}\left(-\int_{\tau_1}^{\tau_m} h^{\eta,g}(x(\tau,w_0,\tilde{v}))d\tau - g(x(\tau_m,w_0,\tilde{v}))\right)\\
&\leq \max_{x\in B_r}\left\{\int_{\tau_1}^{\tau_m}\left(-\int_{\Gamma_0}\hat{L}(x(\tau,w,v))dw\right)d\tau - g(x(\tau_1,w_0,v))\right\}\\
&\leq \max_{x\in B_r, w_0\in\Gamma_0}\left(-\int_{\tau_1}^{\tau_m} h^{\eta,g}(x(\tau,w_0,v))d\tau - g(x(\tau_m,w_0,\tilde{v}))\right) + \eta(\tau_m-\tau_1)
\end{aligned}$$

and thus the assertion of the theorem follows.

Let us now define four following symbols: $\mathcal{F}_{(\eta,1)}(x) := -\sum_{i=2,\dots,m-1}\eta_x^i(\tau_m-\tau_1)$, $\mathcal{F}_1(x) := -\sum_{i\in\{1,\dots,m-1\}} x^i(\tau_m-\tau_1)$, $\mathcal{F}_{(\eta,i)}(x) := -\sum_{i=2,\dots,m-1}\eta_x^i(\tau_m-\tau_i)$, $\mathcal{F}_i(x) := -\sum_{i\in\{1,\dots,m-1\}} x^i(\tau_m-\tau_i)$. Now, we use the definition of equivalence class to reformulate the theorem above in a way that is more useful in practice. To this effect let us note that, by definition of equivalence relation r, we have

$$\min_{x\in B_r}\left\{\mathcal{F}_{(\eta,1)}(x)\right\} = \min_{x\in X}\left\{\mathcal{F}_1(x)\right\}, \max_{x\in B_r}\left\{\mathcal{F}_{(\eta,1)}(x)\right\} = \max_{x\in X}\left\{\mathcal{F}_1(x)\right\}.$$

Let us now also define the following auxiliary symbol $\mathcal{H}(x,w_0) := -h^{\eta,g}(x(\tau_1,w_0,v))(\tau_m-\tau_1) - g(x(\tau_m,w_0,v))$. Taking into account (14) we get

$$\begin{aligned}
&\min_{x\in X}\left\{\mathcal{F}_i(x)\right\} + \min_{x\in B_r, w_0\in\Gamma_0}\left\{\mathcal{H}(x,w_0)\right\} - \eta(\tau_m-\tau_1)\\
&\leq \min_{x\in B_r}\left\{-\int_{\tau_1}^{\tau_m} h^{\eta,g}(x(\tau,w_0,v))d\tau - g(x(\tau_m,w_0,v))\right\}\\
&\leq \min_{x\in X}\left\{\mathcal{F}_i(x)\right\} + \max_{x\in B_r, w_0\in\Gamma_0}\left\{\mathcal{H}(x,w_0)\right\} + \eta(\tau_m-\tau_1)
\end{aligned}$$

and a similar formula for supremum. Applying that to the result of the theorem above, we obtain the following estimation

$$\begin{aligned}
&\min_{x\in X}\left\{\mathcal{F}_i(x)\right\} + \min_{x\in B_r, w_0\in\Gamma_0}\left\{\mathcal{H}(x,w_0)\right\} - 2\eta(\tau_m-\tau_1)\\
&\leq \max_{x\in B_r}\left\{\int_{\tau_1}^{\tau_m}\left(-\int_{\Gamma_0}\hat{L}(x(\tau,w,v))dw\right)d\tau - g(x(\tau_1,w_0,v))\right\}\\
&\leq \max_{x\in X}\left\{\mathcal{F}_i(x)\right\} + \max_{x\in B_r, w_0\in\Gamma_0}\left\{\mathcal{H}(x,w_0)\right\} + \eta(\tau_m-\tau_1).
\end{aligned} \tag{17}$$

Thus, we come to the main theorem of this section, which allows us to reduce an infinite dimensional problem to the finite dimensional one.

Theorem 2. *Let $\eta > 0$. Assume that there is $\theta > 0$ and $\bar{v}$ such that*

$$\begin{aligned}
&\max_{x\in X}\left\{\mathcal{F}_i(x)\right\}\\
&\quad + \max_{x\in B_r, w_0\in\Gamma_0}\left\{\mathcal{H}(x,w_0)\right\}\\
&\leq \min_{x\in X}\left\{\mathcal{F}_i(x)\right\}\\
&\quad + \min_{x\in B_r, w_0\in\Gamma_0}\left\{\mathcal{H}(x,w_0)\right\} + \theta(\tau_m-\tau_1),
\end{aligned} \tag{18}$$

$$\min_{x\in B_r, w_0\in\Gamma_0}\{\mathcal{H}(x,w_0)\} = \min_{w_0\in\Gamma_0}\{-h^{\eta,g}(x(\tau_1,w_0,\bar{v}))(\tau_m-\tau_1) - g(x(\tau_m,w_0,\bar{v}))\}$$

Then

$$(\eta+\theta)(\tau_m-\tau_1) + \min_{x\in B_r, w_0\in\Gamma_0}\{\mathcal{H}(x,w_0)\} + \min_{x\in X}\{\mathcal{F}_i(x)\} \tag{19}$$

is ε-optimal value at (τ_1, w_0) for $\varepsilon = 2\eta+\theta$ with

$$g(w) = \int_{\Omega(v_{\min})} L(y,\bar{u}(y),\nabla\bar{u}(y))dy,\ \ w\in\Gamma_0,$$

where $\bar{u}$ is a solution to (4) for $\Omega(\bar{v})$.

Proof. From the formulae (17), (18) we infer

$$\begin{aligned}
&\min_{x\in X}\{\mathcal{F}_i(x)\} + \min_{x\in B_r, w_0\in\Gamma_0}\{\mathcal{H}(x,w_0)\} - 2\eta(\tau_m-\tau_1)\\
&\le \max_{x\in B_r}\left\{\int_{\tau_1}^{\tau_m}\left(-\int_{\Gamma_0}\hat{L}(x(\tau,w,v))dw\right)d\tau - \int_{\Omega(v_{\min})} L(y,\bar{u}(y),\nabla\bar{u}(y))dy\right\}\\
&\le \max_{x\in X}\{\mathcal{F}_i(x)\} + \max_{x\in B_r, w_0\in\Gamma_0}\{\mathcal{H}(x,w_0)\} + \eta(\tau_m-\tau_1) + \theta(\tau_m-\tau_1).
\end{aligned}$$

Next, using the definition of value function (5), we get (19).

3.3 The Algorithm for Numerical Solution of (P)

In the previous section the last theorem allows us to estimate an ε-optimal value of function (see (6)) for problem (P). As can be seen from the formulas (18) and (19) the essence of the approximation is to be able to calculate the value of the following expressions:

$$\sup_{x\in X}\left\{-\sum_{i\in\{1,\dots,m-1\}} x^i(\tau_m-\tau_i)\right\},\quad \inf_{x\in X}\left\{-\sum_{i\in\{1,\dots,m-1\}} x^i(\tau_m-\tau_i)\right\}.$$

To achieve this aim we construct a particular directed weighted graph G, in which the weight of every edge is the value of the expression $x^i(\tau_m-\tau_i)$. This graph has following properties

1. Every path has length of $m-1$ edges.
2. Every two vertices connected by an edge correspond to points (τ, x_1) and $(\tau+\Delta\tau, x_2)$ such, that the point x_2 is reachable from the point x_1 in the next unit of time $\tau+\Delta\tau$.

Therefore by identifying in the graph G the path with lowest (greatest) cost we find the value of the expression $\inf_{x\in X}\{\cdot\}$ ($\sup_{x\in X}\{\cdot\}$).

The Algorithm for Generation of the Graph G.

1. Let B – a set of trajectories – be a finite set Bezier curves.
2. Let P be a set of points. At the beginning, P contains only one point: $p = w_0$, where w_0 is any but fixed point such that $w_0 \in \Omega_0$
3. Create in graph G node which corresponds to point p.
4. Calculate $F_g(x_0)$ (equation (8)) where $x_0 = (0, w_0)$, i.e. find a Bezier curve $\alpha \in B$ which minimize value of $G_g(x_0, \cdot)$.
5. For $t = d\tau, \ldots, 1$ repeat
   (a) Let P' be an empty set of points.
   (b) For each point p from P repeat
      i. For each Bezier curve β from B repeat
         A. Find point p' reachable from p under ,,control" β in time t.
         B. Calculate $F_g(x')$ (equation (8)) where $x' = (t, p')$, i.e. find a Bezier curve $\alpha \in B$ which minimize value of $G_g(x', \cdot)$.
         C. Create in graph G node which correspond to point p'.
         D. Create in graph G edge $e_{(p,p')}$ from point p to p'.
         E. Label edge $e_{(p,p')}$ with a weight $x^t(1-t)$ calculated basing on the indexes j of sets $P_j^{\eta,g}$ which contain points $(t - dt, p)$ and (t, p') generated in the t-th step (depending on the difference in those indexes, x^t itself is equal to $-\eta$, 0 or η).
         F. Save p' in P'.
   (c) Replace set P by P'.
6. Generation of the graph G is complete.

References

1. Fleming, W.H., Rishel, R.W.: Deterministic and Stochastic Optimal Control. Springer, New York (1975)
2. Fulmański, P., Laurin, A., Scheid, J.F., Sokołowski, J.: A Level Set Method in Shape and Topology Optimization for Variational Inequalities. International Journal of Applied Mathematocs and Computer Science 17, 413–430 (2007)
3. Fulmański, P., Nowakowski, A., Pustelnik, J.: Dynamic programming approach to structural optimization problem – numerical algorithm (submitted for publication)
4. Haslinger, J., Mäkinen, R.: Introduction to Shape Optimization. Theory, Approximation and Computation. SIAM Publications, Philadelphia (2003)
5. Hlavaček, I., Haslinger, J., Nečas, J., Lovišek, J.: Solving of variational Inequalities in Mechancs. Mir, Moscow (1996) (in Russian)
6. Hüber, S., Stadler, G., Wohlmuth, B.: A Primal-Dual Active Set Algorithm for Three Dimensional Contact Problems with Coulomb Fristion. SIAM J. Sci. Comput. 30(2), 572–596 (2008)
7. Pustelnik, J.: Approximation of optimal value for Bolza problem, Ph.D. thesis (2009) (in Polish)
8. Stadler, G.: Semismooth Newton and Augmented Lagrangian methods for a Simplified Friction Problem. SIAM Journal on Optimization 15(1), 39–62 (2004)

Shape Sensitivity Analysis of Incompressible Non-Newtonian Fluids

Jan Sokołowski[1] and Jan Stebel[2]

[1] Institut Elie Cartan de Nancy, Université de Lorraine, Campus des Aiguillettes, B.P. 70239, 54506 Vandœuvre-lès-Nancy Cedex, France
Jan.Sokolowski@univ-lorraine.fr
[2] Institute of Mathematics of the Academy of Sciences of the Czech Republic, Žitná 25, 115 67 Praha 1, Czech Republic
stebel@math.cas.cz

Abstract. We study the shape differentiability of a cost function for the steady flow of an incompressible viscous fluid of power-law type. The fluid is confined to a bounded planar domain surrounding an obstacle. For smooth perturbations of the shape of the obstacle we express the shape gradient of the cost function which can be subsequently used to improve the initial design.

Keywords: shape optimization, shape gradient, incompressible fluid, non-Newtonian fluid, Navier-Stokes equations.

1 Introduction

Shape optimization for nonlinear partial differential equations is a growing field in the contemporary optimum design of structures. In this field systems of the solid and fluid mechanics as well as e.g., the coupled models of fluid-structure interaction are included for real life problems. The main difficulty associated with the mathematical analysis of nonlinear state equations is the lack of existence of global strong solutions for mathematical models in three spatial dimensions.

In numerical methods of shape optimization the common approach is the discretization of continuous shape gradient. Therefore, the proper derivation and analysis of the regularity properties of the shape gradient is crucial for numerical solution of the shape optimization problem. The shape sensitivity analysis requires, in particular, the proof of the Lipschitz continuity of solutions the the state equations with respect to the boundary variations. This property of solutions can be obtained e.g. by analysis of the state equation transported to the fixed reference domain which is explained in the case of linear elliptic boundary value problems in monograph [11]. For the nonlinear problems the Lipschitz continuity is not obvious and it requires the additional regularity of solutions to the state equation. In addition, for the applications of levelset method of shape optimization it is required that the obtained shape gradient of the cost functional is given by a function while the general theory gives only the existence of a distribution. In conlusion, it seems that the shape sensitivity analysis in the

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 427–436, 2013.

case of a nonlinear state equation is the main step towards the numerical solution of the shape optimization problems.

In gas dynamics described by the compressible Navier-Stokes there is the existence of weak global solutions. However, the shape sensitivity analysis can be performed only for specific local solutions. The state of art in shape optimization for compressible Navier-Stokes equations is presented in the monograph [8], see also [7]. For incompressible Navier-Stokes equations, the sensitivity analysis of shape functionals is performed e.g. in [2] and [6]. In this paper we are concerned with the non-Newtonian model where the stress is a (nonlinear) function of the velocity gradient. Optimal control problem for this model was studied in [9, 13]. Numerical shape optimization was done in [1], see also [3]. We present new results on the existence of the shape gradient.

We consider the steady flow of an incompressible fluid in a bounded domain $\Omega := B \setminus S$ in $\mathbb{R}^2$, where B is a container and S is an obstacle. Motion of the fluid is described by the system of equations

$$\begin{aligned} \operatorname{div}(\mathbf{v}\otimes\mathbf{v}) - \operatorname{div}\mathbb{S}(\mathbb{D}\mathbf{v}) + \nabla p + \mathbb{C}\mathbf{v} &= \mathbf{f} && \text{in } \Omega, \\ \operatorname{div}\mathbf{v} &= 0 && \text{in } \Omega, \\ \mathbf{v} &= \mathbf{g} && \text{on } \partial\Omega. \end{aligned} \tag{$P(\Omega)$}$$

Here $\mathbf{v}$, p, $\mathbb{C}$, $\mathbf{f}$ stand for the velocity, the pressure, the constant skew-symmetric Coriolis tensor and the body force, respectively. The traceless part $\mathbb{S}$ of the Cauchy stress can depend on the symmetric part $\mathbb{D}\mathbf{v}$ of the velocity gradient in the following way:

$$\mathbb{S}(\mathbb{D}\mathbf{v}) = \nu(|\mathbb{D}\mathbf{v}|^2)\mathbb{D}\mathbf{v}, \tag{1}$$

where ν, $|\mathbb{D}\mathbf{v}|^2$ is the viscosity and the shear rate, respectively. In particular, we assume that ν has a polynomial growth (see Section 2.1 below), which includes e.g. the Carreau and the power-law models.

In the model the term of Coriolis type is present. This term appears e.g. when the change of variables is performed in order to take into account the flight scenario of the obstacle in the fluid.

The aim of this paper is to investigate differentiability of a shape functional depending on the solution to ($P(\Omega)$) with respect to the variations of the shape of the obstacle. We consider a model problem with the drag functional

$$J(\Omega) := \int_{\partial S} (\mathbb{S}(\mathbb{D}\mathbf{v}) - p\mathbb{I})\mathbf{n}\cdot\mathbf{d}, \tag{2}$$

with a given constant unit vector $\mathbf{d}$. Instead of J one could take other type of functional, since our method does not rely on its specific form.

Our main interest is the rigorous analysis of the shape differentiability for ($P(\Omega)$) and (2). We follow the general framework developed in [11] using the speed method and the notion of the material derivative. Let us point out that due to (1) the state problem is nonlinear in its nature. We refer the reader to [12] for an introduction to optimization problems for nonlinear partial differential equations.

1.1 Shape Derivatives

We start by the description of the framework for the shape sensitivity analysis. For this reason, we introduce a vector field $\mathbf{T} \in \mathcal{C}^2(\mathbb{R}^2, \mathbb{R}^2)$ vanishing in the vicinity of ∂B and define the mapping

$$\mathbf{y}(\mathbf{x}) = \mathbf{x} + \varepsilon \mathbf{T}(\mathbf{x}).$$

For small $\varepsilon > 0$ the mapping $\mathbf{x} \mapsto \mathbf{y}(\mathbf{x})$ takes diffeomorphically the region Ω onto $\Omega_\varepsilon = B \setminus S_\varepsilon$ where $S_\varepsilon = \mathbf{y}(S)$. We consider the counterpart of problem ($P(\Omega)$) in Ω_ε, with the data $\mathbf{f}_{|\Omega_\varepsilon}$ and $\mathbf{g}_{|\Omega_\varepsilon}$. The new problem will be denoted by $(P(\Omega_\varepsilon))$ and its solution by $(\bar{\mathbf{v}}_\varepsilon, \bar{p}_\varepsilon)$.

For the nonlinear system ($P(\Omega)$) we introduce the shape derivatives of solutions. To this end we need the linearized system of the form:

Find the couple $(\mathbf{u}, \pi)$ *such that*

$$\begin{aligned} \operatorname{div}\left[\mathbf{u} \otimes \mathbf{v} + \mathbf{v} \otimes \mathbf{u} - \mathbb{S}'(\mathbb{D}\mathbf{v})\mathbb{D}\mathbf{u}\right] + \nabla\pi + \mathbb{C}\mathbf{u} &= \mathbf{F} && \text{in } \Omega, \\ \operatorname{div}\mathbf{u} &= 0 && \text{in } \Omega, \\ \mathbf{u} &= \mathbf{h} && \text{on } \partial\Omega, \end{aligned} \qquad (P_{\mathrm{lin}}(\Omega))$$

where $\mathbf{F}$ *and* $\mathbf{h}$ *are given elements.*

The shape derivative $\mathbf{v}'$ and the material derivative $\dot{\mathbf{v}}$ of solutions are formally introduced by

$$\mathbf{v}' := \lim_{\varepsilon \to 0} \frac{\bar{\mathbf{v}}_\varepsilon - \mathbf{v}}{\varepsilon}, \quad \dot{\mathbf{v}} := \lim_{\varepsilon \to 0} \frac{\bar{\mathbf{v}}_\varepsilon \circ \mathbf{y} - \mathbf{v}}{\varepsilon},$$

where $\bar{\mathbf{v}}_\varepsilon \circ \mathbf{y}(\mathbf{x}) := \bar{\mathbf{v}}_\varepsilon(\mathbf{y}(\mathbf{x}))$, and are related to each other as follows:

$$\dot{\mathbf{v}} = \mathbf{v}' + (\nabla \mathbf{v})\mathbf{T}.$$

The standard calculus for differentiating with respect to shape yields that $\mathbf{v}'$ is the solution of ($P_{\mathrm{lin}}(\Omega)$) with the data $\mathbf{F} = \mathbf{0}$ and $\mathbf{h} = -\partial\mathbf{v}/\partial\mathbf{n}(\mathbf{T} \cdot \mathbf{n})$. Using (7) as the definition of J we obtain the expression for the shape gradient:

$$\begin{aligned} dJ(\Omega; \mathbf{T}) &:= \lim_{\varepsilon \to 0} \frac{J(\Omega_\varepsilon) - J(\Omega)}{\varepsilon} \\ &= \int_\Omega \left[(\mathbb{C}\mathbf{v}') \cdot \boldsymbol{\xi} + (\mathbb{S}'(\mathbb{D}\mathbf{v})\mathbb{D}\mathbf{v}' - \mathbf{v}' \otimes \mathbf{v} - \mathbf{v} \otimes \mathbf{v}') : \nabla\boldsymbol{\xi}\right] - \int_{\partial S} (\mathbf{f} \cdot \mathbf{d})\mathbf{T} \cdot \mathbf{n}. \end{aligned} \qquad (3)$$

In the above formula, the part containing $\mathbf{v}'$ depends implicitly on the direction $\mathbf{T}$. This is not convenient for practical use, hence we introduce the adjoint problem for further simplification of (3):

Find the couple $(\mathbf{w}, s)$ *such that*

$$\begin{aligned} -2(\mathbb{D}\mathbf{w})\mathbf{v} - \operatorname{div}\left[\mathbb{S}'(\mathbb{D}\mathbf{v})^\top \mathbb{D}\mathbf{w}\right] + \nabla s - \mathbb{C}\mathbf{w} &= \mathbf{0} && \text{in } \Omega, \\ \operatorname{div}\mathbf{w} &= 0 && \text{in } \Omega, \\ \mathbf{w} &= \mathbf{d} && \text{on } \partial\Omega. \end{aligned} \qquad (P_{\mathrm{adj}}(\Omega))$$

Consequently, the expression for dJ reduces to

$$dJ(\Omega;\mathbf{T}) = -\int_{\partial S}\left[\left(\mathbb{S}'(\mathbb{D}\mathbf{v})^{\top}\mathbb{D}\mathbf{w} - s\mathbb{I}\right) : \frac{\partial \mathbf{v}}{\partial \mathbf{n}} \otimes \mathbf{n} + \mathbf{f}\cdot\mathbf{d}\right]\mathbf{T}\cdot\mathbf{n}. \tag{4}$$

In order to prove the result given by (3) and (4) we need the material derivatives. In particular, it is sufficient to show that the linear mapping

$$\mathbf{T} \mapsto dJ(\Omega;\mathbf{T})$$

is continuous in an appropriate topology, see the structure Theorem in the book [11] for details.

2 Preliminaries

We impose the structural assumptions on the data, state the known results on well-posedness of $(P(\Omega))$ and introduce the elementary notation for shape sensitivity analysis.

2.1 Structural Assumptions

We require that $\mathbb{S}$ has a potential $\Phi : [0,\infty) \to [0,\infty)$, i.e. $\mathbb{S}_{ij}(\mathbb{D}) = \partial\Phi(|\mathbb{D}|^2)/\partial\mathbb{D}_{ij}$. Further we assume that Φ is a $\mathcal{C}^3$ function with $\Phi(0) = 0$ and that there exist constants $C_1, C_2, C_3 > 0$ and $r \geq 2$ such that

$$C_1(1+|\mathbb{A}|^{r-2})|\mathbb{B}|^2 \leq \mathbb{S}'(\mathbb{A}) :: (\mathbb{B}\otimes\mathbb{B}) \leq C_2(1+|\mathbb{A}|^{r-2})|\mathbb{B}|^2, \tag{5a}$$

$$|\mathbb{S}''(\mathbb{A})| \leq C_3(1+|\mathbb{A}|^{r-3}) \tag{5b}$$

for any $0 \neq \mathbb{A}, \mathbb{B} \in \mathbb{R}^{2\times 2}_{sym}$. Here the symbol :: stands for the usual scalar product in $\mathbb{R}^{2^4}$. The above inequalities imply the monotone structure of $\mathbb{S}$, see e.g. [5].

2.2 Weak Formulation

For the definition of the weak solution we will use the space

$$\mathbf{W}^{1,r}_{0,\mathrm{div}}(\Omega) := \{\boldsymbol{\phi} \in \mathbf{W}^{1,r}_0(\Omega);\ \mathrm{div}\,\boldsymbol{\phi} = 0\}.$$

Let $\mathbf{f} \in (\mathbf{W}^{1,2}_{0,\mathrm{div}}(\Omega))^*$ and $\mathbf{g} \in \mathbf{W}^{1,r}(\Omega)$ with $\mathrm{div}\,\mathbf{g} = 0$. Then a function $\mathbf{v} \in \mathbf{g} + \mathbf{W}^{1,r}_{0,\mathrm{div}}(\Omega)$ is said to be a weak solution to the problem $(P(\Omega))$ if

$$\int_\Omega \left[\mathbb{S}(\mathbb{D}\mathbf{v}) : \mathbb{D}\boldsymbol{\phi} - \mathbf{v}\otimes\mathbf{v} : \nabla\boldsymbol{\phi} + \mathbb{C}\mathbf{v}\cdot\boldsymbol{\phi}\right] = \int_\Omega \mathbf{f}\cdot\boldsymbol{\phi} \tag{6}$$

for every $\boldsymbol{\phi} \in \mathbf{W}^{1,r}_{0,\mathrm{div}}(\Omega)$. Note that the pressure is eliminated since test functions are divergence free.

The following result was shown in [4].

Theorem 1 (Kaplický et al. [4]). *Let $\Omega \in \mathcal{C}^2$, $\mathbf{f} \in \mathbf{L}^{2+\epsilon_0}(\Omega)$, $\epsilon_0 > 0$ and (5a)–(5b) hold with $r > \frac{3}{2}$. Then there exists a constant $\delta > 0$ such that for every $\mathbf{g}$ satisfying*

$$\|\mathbf{g}\|_{3,q} \leq \delta, \ (q > 2),$$

problem $(P(\Omega))$ has a weak solution satisfying $\mathbf{v} \in \mathbf{W}^{2,2+\epsilon}(\Omega)$, $p \in W^{1,2+\epsilon}(\Omega)$, $\epsilon > 0$.

Note that the above result applies only to the unperturbed domain, i.e. $\varepsilon = 0$. Assuming smallness of $\|\mathbf{f}\|_{2,B}$ and $\|\mathbf{g}\|_{3,q,B}$, one can prove that $(P(\Omega))$, $(P(\Omega_\varepsilon))$ has a unique weak solution satisfying

$$\|\mathbf{v}\| \leq C_E(\|\mathbf{f}\|_{2,B}, \|\mathbf{g}\|_{3,q,B}) \quad \text{and} \quad \|\bar{\mathbf{v}}_\varepsilon\| \leq C_E(\|\mathbf{f}\|_{2,B}, \|\mathbf{g}\|_{3,q,B}),$$

respectively, where C_E is independent of ε. At this point we summarize the main hypotheses.

Assumption 1. *In what follows, $\Omega \in \mathcal{C}^2$ is a bounded planar domain of the form $\Omega = B \setminus S$, $\mathbf{f} \in \mathbf{L}^{2+\epsilon_0}(B)$, $\epsilon_0 > 0$, $\mathbf{g} \in \mathbf{W}^{3,q}(B)$ $(q > 2)$ is supported in the vicinity of ∂B, (5a)–(5b) hold with $r \in [2, 4)$ and $\|\mathbf{f}\|_{2,B}$, $\|\mathbf{g}\|_{3,q,B}$ are small enough.*

Let us point out that equation (2) which defines J is not suitable for weak solutions in general, since the energy inequality does not provide enough information about the trace of p and $\mathbb{D}\mathbf{v}$. We therefore introduce an alternative definition that requires less regularity. Let us fix an arbitrary divergence free function $\boldsymbol{\xi} \in \mathcal{C}_c^\infty(B, \mathbb{R}^2)$ such that $\boldsymbol{\xi} = \mathbf{d}$ in a vicinity of S. Then, integrating (2) by parts and using $(P(\Omega))$ yields:

$$J(\Omega) = \int_\Omega \left[(\mathbb{C}\mathbf{v} - \mathbf{f}) \cdot \boldsymbol{\xi} + (\mathbb{S}(\mathbb{D}\mathbf{v}) - \mathbf{v} \otimes \mathbf{v}) : \nabla \boldsymbol{\xi}\right]. \tag{7}$$

Note that this volume integral is finite for any $\mathbf{v} \in \mathbf{W}^{1,2}(\Omega)$.

2.3 Deformation of the Shape

Let us introduce the following notation: We will denote by $\mathrm{D}\,\mathbf{T}$ the Jacobian matrix whose components are $(\mathrm{D}\,\mathbf{T})_{ij} = (\nabla \mathbf{T})_{ji} = \partial_i T_j$. Further,

$$\mathbb{N}(\mathbf{x}) := \mathfrak{g}(\mathbf{x})\mathbb{M}^{-1}(\mathbf{x}), \quad \mathbb{M}(\mathbf{x}) := \mathbb{I} + \varepsilon\,\mathrm{D}\,\mathbf{T}(\mathbf{x}), \quad \mathfrak{g}(\mathbf{x}) := \det \mathbb{M}(\mathbf{x}).$$

One can easily check that the matrix $\mathbb{N}$ and the determinant $\mathfrak{g}$ admit the expansions:

$$\mathfrak{g} = 1 + \varepsilon \mathrm{div}\,\mathbf{T} + O(\varepsilon^2), \quad \mathbb{N} = \mathbb{I} + \varepsilon \mathbb{N}' + O(\varepsilon^2), \quad \mathbb{N}' = (\mathrm{div}\,\mathbf{T})\mathbb{I} - \mathrm{D}\,\mathbf{T}, \tag{8}$$

where the symbol $O(\varepsilon^2)$ denotes a function whose norm in $\mathcal{C}^1(\overline{\Omega})$ is bounded by $C\varepsilon^2$, see [11].

The value of the shape functional for Ω_ε is given by

$$J(\Omega_\varepsilon) := \int_{\Omega_\varepsilon} [(\mathbb{C}\bar{\mathbf{v}}_\varepsilon - \mathbf{f}) \cdot \boldsymbol{\xi}_\varepsilon + (\mathbb{S}(\mathbb{D}\bar{\mathbf{v}}_\varepsilon) - \bar{\mathbf{v}}_\varepsilon \otimes \bar{\mathbf{v}}_\varepsilon) : \nabla \boldsymbol{\xi}_\varepsilon],$$

where $\boldsymbol{\xi}_\varepsilon := (\mathbb{N}^{-\top}\boldsymbol{\xi}) \circ \mathbf{y}^{-1}$. Using the properties of the Piola transform one can check that $\operatorname{div} \boldsymbol{\xi}_\varepsilon = 0$. If $\bar{\mathbf{v}}_\varepsilon$ and $\bar{p}_\varepsilon$ were sufficiently smooth, it would hold that

$$J(\Omega_\varepsilon) = \int_{\partial S_\varepsilon} (\mathbb{S}(\mathbb{D}\bar{\mathbf{v}}_\varepsilon) - \bar{p}_\varepsilon \mathbb{I})\mathbf{n}_\varepsilon \cdot \mathbf{d}. \tag{9}$$

Nevertheless, as opposed to ($P(\Omega)$), we do not require any additional regularity of the solution to the perturbed problem $(P(\Omega_\varepsilon))$ and hence the expression in (9) need not be well defined.

We introduce the auxiliary function $\tilde{\mathbf{v}}$:

$$\tilde{\mathbf{v}} := \lim_{\varepsilon \to 0} \frac{\mathbb{N}^\top \bar{\mathbf{v}}_\varepsilon \circ \mathbf{y} - \mathbf{v}}{\varepsilon},$$

which is related to the material derivative $\dot{\mathbf{v}}$ by the identity

$$\tilde{\mathbf{v}} = \mathbb{N}'^\top \mathbf{v} + \dot{\mathbf{v}}.$$

For the justification of the results of the paper we will use $\tilde{\mathbf{v}}$ since, unlike the material derivative, it preserves the divergence free condition.

3 Main Results

The first result is the existence of $\tilde{\mathbf{v}}$ and hence also of the material derivative.

Theorem 2. *Let Assumption 1 be satisfied. Then the function $\tilde{\mathbf{v}}$ exists and is the unique weak solution of ($P_{\mathrm{lin}}(\Omega)$) with the data*

$$\begin{aligned}\mathbf{F} = \mathbf{A}_0' &:= \operatorname{div}(\mathbf{v} \otimes \mathbb{N}'^\top \mathbf{v}) + \mathbb{N}' \operatorname{div}(\mathbf{v} \otimes \mathbf{v}) \\ &\quad + \operatorname{div}\left[\mathbb{S}'(\mathbb{D}\mathbf{v})\left(((\mathbb{N}' - \mathbb{I}\operatorname{tr}\mathbb{N}')\nabla\mathbf{v})_{sym} - \mathbb{D}(\mathbb{N}'^\top \mathbf{v})\right) + \mathbb{N}'^\top \mathbb{S}(\mathbb{D}\mathbf{v})\right] \\ &\quad - \mathbb{N}' \operatorname{div} \mathbb{S}(\mathbb{D}\mathbf{v}) + \left((\mathbb{N}' - \mathbb{I}\operatorname{tr}\mathbb{N}')\mathbb{C} + \mathbb{C}\mathbb{N}'^\top\right)\mathbf{v} + (\mathbb{I}\operatorname{tr}\mathbb{N}' - \mathbb{N}')\mathbf{f} + (\nabla \mathbf{f})\mathbf{T}, \end{aligned} \tag{10a}$$

$$\mathbf{h} = \mathbf{0}. \tag{10b}$$

The following estimate holds:

$$\|\tilde{\mathbf{v}}\|_{1,2,\Omega} \le C\|\mathbf{A}_0'\|_{\mathbf{W}^{1,2}_{0,\mathrm{div}}(\Omega)^*} \le C\|\mathbf{T}\|_{\mathcal{C}^2(\overline{\Omega})}. \tag{11}$$

Next we establish the existence of the shape gradient of J.

Theorem 3. *Let Assumption 7 be satisfied and* $\mathbf{f} \in \mathbf{W}^{1,2}(\Omega)$. *Then the shape gradient of* J *reads*

$$dJ(\Omega, \mathbf{T}) = J_{\mathbf{v}}(\tilde{\mathbf{v}}) + J_e(\mathbf{T}),$$

where the dynamical part $J_{\mathbf{v}}$ *and the geometrical part* J_e *is given by*

$$J_{\mathbf{v}}(\tilde{\mathbf{v}}) = \int_\Omega \left[(\mathbb{C}\tilde{\mathbf{v}}) \cdot \boldsymbol{\xi} + (\mathbb{S}'(\mathbb{D}\mathbf{v})\mathbb{D}\tilde{\mathbf{v}} - \tilde{\mathbf{v}} \otimes \mathbf{v} - \mathbf{v} \otimes \tilde{\mathbf{v}}) : \nabla\boldsymbol{\xi} \right],$$

$$\begin{aligned} J_e(\mathbf{T}) = \int_\Omega \Big\{ & \left[(\mathbb{I}\,\mathrm{tr}\,\mathbb{N}' - \mathbb{N}')\,\mathbb{C}\mathbf{v} - \mathbb{C}\mathbb{N}'^{\top}\mathbf{v} - (\mathbb{I}\,\mathrm{tr}\,\mathbb{N}' - \mathbb{N}')\,\mathbf{f} - (\nabla\mathbf{f})\mathbf{T} \right] \cdot \boldsymbol{\xi} \\ & + \left[\mathbf{v} \otimes \mathbb{N}'^{\top}\mathbf{v} + \mathbb{S}'(\mathbb{D}\mathbf{v}) \left((\mathbb{N}'\nabla\mathbf{v} - \nabla(\mathbb{N}'^{\top}\mathbf{v}))_{sym} - (\mathrm{tr}\,\mathbb{N}')\mathbb{D}\mathbf{v} \right) + \mathbb{N}'^{\top}\mathbb{S}(\mathbb{D}\mathbf{v}) \right] : \nabla\boldsymbol{\xi} \\ & + \left[\mathbf{v} \otimes \mathbf{v} - \mathbb{S}(\mathbb{D}\mathbf{v}) \right] : \nabla(\mathbb{N}'^{\top}\boldsymbol{\xi}) \Big\}, \end{aligned}$$

respectively. In particular, as $\tilde{\mathbf{v}}$ *depends continuously on* $\mathbf{T}$, *the mapping*

$$\mathbf{T} \mapsto dJ(\Omega, \mathbf{T})$$

is a bounded linear functional on $\mathcal{C}^2(\mathbb{R}^2, \mathbb{R}^2)$.

Based on the previous result we can deduce that the shape gradient has the form of a distribution supported on the boundary of the obstacle. Since this representation is unique, the formal results derived in Section 1.1 are justified provided that the shape derivatives and adjoints exist and are sufficiently regular.

Corollary 1. *Let Assumption 7 be satisfied. Then*

(i) the shape derivative $\mathbf{v}'$ *exists and is the unique weak solution to* ($P_{\mathrm{lin}}(\Omega)$) *with* $\mathbf{F} = \mathbf{0}$, $\mathbf{h} = -\dfrac{\partial \mathbf{v}}{\partial \mathbf{n}}(\mathbf{T} \cdot \mathbf{n})$;

(ii) the adjoint problem ($P_{\mathrm{adj}}(\Omega)$) *has a unique weak solution that satisfies:* $\mathbf{w} \in \mathbf{W}^{2,2}(\Omega)$ *and* $s \in W^{1,2}(\Omega)$.

If in addition $\mathbf{f} \in \mathbf{W}^{1,2}(\Omega)$, *then*

(iii) the shape gradient of J *satisfies* (3);

(iv) the representation (4) *is satisfied in the following sense:*

$$dJ(\Omega; \mathbf{T}) = -\int_{\partial S} \left[\left(\mathbb{S}'(\mathbb{D}\mathbf{v})^{\top}\mathbb{D}\mathbf{w} - s\mathbb{I} \right) : \frac{\partial \mathbf{v}}{\partial \mathbf{n}} \otimes \mathbf{n} + \mathbf{f} \cdot \mathbf{d} \right] \mathbf{T} \cdot \mathbf{n}. \tag{12}$$

In the remaining part we show the main steps of the proof of Theorem 3. Details can be found in [10], where the time-dependent problem is treated.

4 Formulation in the Fixed Domain

In this section we transform the problem $(P(\Omega_\varepsilon))$ to the fixed domain Ω. Let us introduce the following notation:

$$\mathbf{v}_\varepsilon(\mathbf{x}) := \mathbb{N}^\top(\mathbf{x})\bar{\mathbf{v}}_\varepsilon(\mathbf{y}(\mathbf{x})),\ \mathbf{x}\in\Omega.$$

Note that the definition of $\mathbf{v}_\varepsilon$ implies that $\operatorname{div}\mathbf{v}_\varepsilon = 0$. The new function $\mathbf{v}_\varepsilon \in \mathbf{g} + \mathbf{W}^{1,r}_{0,\operatorname{div}}(\Omega)$ satisfies the equality

$$\int_\Omega \Big[\mathfrak{g}\mathbb{S}(\mathbb{D}_\varepsilon\mathbf{v}_\varepsilon) : \mathbb{D}_\varepsilon\boldsymbol{\phi} - \mathbf{v}_\varepsilon\otimes\mathbf{v}_\varepsilon : \nabla\boldsymbol{\phi} + \mathbb{C}\mathbf{v}_\varepsilon\cdot\boldsymbol{\phi}\Big] = \int_\Omega \mathbf{f}\cdot\boldsymbol{\phi} + \langle\mathbf{A}^1_\varepsilon,\boldsymbol{\phi}\rangle_{\mathbf{W}^{1,2}_{0,\operatorname{div}}(\Omega)} \quad \text{for all } \boldsymbol{\phi}\in\mathbf{W}^{1,r}_{0,\operatorname{div}}(\Omega), \quad (13)$$

where the term $\mathbf{A}^1_\varepsilon$ on the right hand side is defined for $\boldsymbol{\phi}\in\mathbf{W}^{1,2}_{0,\operatorname{div}}(\Omega)$ by

$$\langle\mathbf{A}^1_\varepsilon,\boldsymbol{\phi}\rangle_{\mathbf{W}^{1,2}_{0,\operatorname{div}}(\Omega)} = \int_\Omega \Big[\mathbf{v}_\varepsilon\otimes\mathbb{N}^{-\top}\mathbf{v}_\varepsilon : \nabla(\mathbb{N}^{-\top}\boldsymbol{\phi}) - \mathbf{v}_\varepsilon\otimes\mathbf{v}_\varepsilon : \nabla\boldsymbol{\phi} + (\mathbb{C}-\mathfrak{g}\mathbb{N}^{-1}\mathbb{C}\mathbb{N}^{-\top})\mathbf{v}_\varepsilon\cdot\boldsymbol{\phi} + (\mathfrak{g}\mathbb{N}^{-1}\mathbf{f}\circ\mathbf{y}-\mathbf{f})\cdot\boldsymbol{\phi}\Big]. \quad (14)$$

Here $\mathbb{D}_\varepsilon\mathbf{v}_\varepsilon := \mathfrak{g}^{-1}(\mathbb{N}\nabla(\mathbb{N}^{-\top}\mathbf{v}_\varepsilon))_{sym}$.

Applying change of coordinates we further get:

$$J(\Omega_\varepsilon) = \int_\Omega \Big[\mathfrak{g}\left(\mathbb{N}^{-1}\mathbb{C}\mathbb{N}^{-\top}\mathbf{v}_\varepsilon - \mathbb{N}^{-1}\mathbf{f}\circ\mathbf{y}\right)\cdot\boldsymbol{\xi} + \left(\mathbb{N}^\top\mathbb{S}(\mathbb{D}_\varepsilon\mathbf{v}_\varepsilon) - \mathbf{v}_\varepsilon\otimes(\mathbb{N}^{-\top}\mathbf{v}_\varepsilon)\right) : \nabla(\mathbb{N}^{-\top}\boldsymbol{\xi})\Big]. \quad (15)$$

Now after all quantities and equations have been transformed to the fixed domain Ω, we can analyze the limit $\varepsilon\to 0$.

Lemma 1. *The sequence $\{\mathbf{v}_\varepsilon\}_{\varepsilon>0}$ is bounded in $\mathbf{W}^{1,r}_{0,\operatorname{div}}(\Omega)$ and satisfies:*

$$\begin{aligned} \mathbf{v}_\varepsilon &\rightharpoonup \mathbf{v} && \text{weakly in } \mathbf{W}^{1,r}_{0,\operatorname{div}}(\Omega),\\ \mathbb{N}^\top\mathbb{S}(\mathbb{D}_\varepsilon\mathbf{v}_\varepsilon) &\rightharpoonup \mathbb{S}(\mathbb{D}\mathbf{v}) && \text{weakly in } L^{r'}(\Omega,\mathbb{R}^{2\times 2}),\\ \mathbf{A}^1_\varepsilon &\rightharpoonup \mathbf{0} && \text{weakly in } \mathbf{W}^{1,r}_{0,\operatorname{div}}(\Omega)^*. \end{aligned}$$

In particular, $\mathbf{v}$ is the unique weak solution to $(P(\Omega))$.

5 Existence of Material Derivative

Our next task is to identify $\tilde{\mathbf{v}}$ as the limit of the sequence $\{\mathbf{u}_\varepsilon\}$, where

$$\mathbf{u}_\varepsilon := \frac{\mathbf{v}_\varepsilon - \mathbf{v}}{\varepsilon}.$$

First we write down the system for the differences $\mathbf{u}_\varepsilon$. Subtracting (13) and (6) we find that $\mathbf{u}_\varepsilon \in \mathbf{W}^{1,r}_{0,\mathrm{div}}(\Omega)$ satisfies the equality

$$\int_\Omega \Big[\frac{1}{\varepsilon}\mathfrak{g}(\mathbb{S}(\mathbb{D}_\varepsilon \mathbf{v}_\varepsilon) - \mathbb{S}(\mathbb{D}_\varepsilon \mathbf{v})) : \mathbb{D}_\varepsilon \boldsymbol{\phi} + \mathbb{C}\mathbf{u}_\varepsilon \cdot \boldsymbol{\phi} - (\mathbf{v}_\varepsilon \otimes \mathbf{u}_\varepsilon + \mathbf{u}_\varepsilon \otimes \mathbf{v}) : \nabla \boldsymbol{\phi}\Big] = \frac{1}{\varepsilon}\langle \mathbf{A}_\varepsilon, \boldsymbol{\phi}\rangle_{\mathbf{W}^{1,2}_{0,\mathrm{div}}(\Omega)} \quad (16)$$

for all $\boldsymbol{\phi} \in \mathbf{W}^{1,r}_{0,\mathrm{div}}(\Omega)$. The term $\mathbf{A}_\varepsilon \in \mathbf{W}^{1,2}_{0,\mathrm{div}}(\Omega)^*$ on the right hand side is defined as follows:

$$\mathbf{A}_\varepsilon := \mathbf{A}^1_\varepsilon + \mathbf{A}^2_\varepsilon,$$

$$\mathbf{A}^1_\varepsilon \text{ is given by (14)},$$

$$\langle \mathbf{A}^2_\varepsilon, \boldsymbol{\phi}\rangle_{\mathbf{W}^{1,2}_{0,\mathrm{div}}(\Omega)} := \int_\Omega \Big[\mathbb{N}^\top \mathbb{S}(\mathbb{D}_\varepsilon \mathbf{v}) : \nabla(\mathbb{N}^{-\top}\boldsymbol{\phi}) - \mathbb{S}(\mathbb{D}\mathbf{v}) : \mathbb{D}\boldsymbol{\phi}\Big].$$

Next we state the properties of the sequence $\{\mathbf{u}_\varepsilon\}_{\varepsilon>0}$.

Lemma 2. *The sequence $\{\mathbf{u}_\varepsilon\}_{\varepsilon>0}$ is bounded in $\mathbf{W}^{1,2}_{0,\mathrm{div}}(\Omega)$. Further it holds:*

$$\begin{aligned}
\frac{\mathbf{A}_\varepsilon}{\varepsilon} &\rightharpoonup \mathbf{A}'_0 && \text{weakly in } \mathbf{W}^{1,2}_{0,\mathrm{div}}(\Omega)^*,\\
\mathbf{u}_\varepsilon &\rightharpoonup \tilde{\mathbf{v}} && \text{weakly in } \mathbf{W}^{1,2}_{0,\mathrm{div}}(\Omega),\\
\frac{1}{\varepsilon}(\mathfrak{g}(\mathbb{S}(\mathbb{D}_\varepsilon \mathbf{v}_\varepsilon) - \mathbb{S}(\mathbb{D}_\varepsilon \mathbf{v})), \mathbb{D}_\varepsilon \boldsymbol{\phi}) &\to (\mathbb{S}'(\mathbb{D}\mathbf{v})\mathbb{D}\tilde{\mathbf{v}}, \mathbb{D}\boldsymbol{\phi}) && \text{for all } \boldsymbol{\phi} \in \mathbf{W}^{1,\frac{2r}{4-r}}(\Omega),
\end{aligned}$$

where $\mathbf{A}'_0$ is defined in (10a) and $\tilde{\mathbf{v}}$ is the solution of ($P_{\mathrm{lin}}(\Omega)$) with $\mathbf{F} := \mathbf{A}'_0$ and $\mathbf{h} = \mathbf{0}$.

This completes the proof of Theorem 2.

6 Shape Gradient of J

To prove Theorem 3, we decompose the fraction

$$\frac{J(\Omega_\varepsilon) - J(\Omega)}{\varepsilon} = J^\varepsilon_1 + J^\varepsilon_2$$

in a suitable way. Using Lemma 1 and Lemma 2 and the properties of $\mathfrak{g}$ and $\mathbb{N}'$, it is then possible to show that

$$J^\varepsilon_1 \to J_{\mathbf{v}}(\tilde{\mathbf{v}}) \quad \text{and} \quad J^\varepsilon_2 \to J_e(\mathbf{T}).$$

The continuity of the map $\mathbf{T} \mapsto dJ(\Omega; \mathbf{T})$ follows from the estimate (11).

Acknowledgement. The work of J. Stebel was supported by the Czech Science Foundation (GAČR) grant No. 201/09/0917 and by the ESF grant Optimization with PDE Constraints. The research of J. Sokołowski is performed in the framework of the ANR project GAOS.

References

[1] Abraham, F., Behr, M., Heinkenschloss, M.: Shape optimization in steady blood flow: A numerical study of non-newtonian effects. Computer Methods in Biomechanics and Biomedical Engineering 8(2), 127–137 (2005)
[2] Consiglieri, L., Nečasová, Š., Sokolowski, J.: New approach to the incompressible Maxwell-Boussinesq approximation: existence, uniquness and shape sensitivity. J. Differential Equations 249(12), 3052–3080 (2010) ISSN 0022-0396
[3] Haslinger, J., Stebel, J.: Shape optimization for navier–stokes equations with algebraic turbulence model: Numerical analysis and computation. Applied Mathematics and Optimization 63(2), 277–308 (2011)
[4] Kaplický, P., Málek, J., Stará, J.: On global existence of smooth two-dimensional steady flows for a class of non-newtonian fluids under various boundary conditions. Applied Nonlinear Analysis, 213–229 (2002)
[5] Málek, J., Rajagopal, K.R.: Mathematical issues concerning the Navier-Stokes equations and some of its generalizations. In: Evolutionary Equations, Handb. Differ. Equ., vol. II, pp. 371–459. Elsevier/North-Holland, Amsterdam (2005)
[6] Moubachir, M., Zolésio, J.-P.: Moving shape analysis and control. Pure and Applied Mathematics, vol. 277. Chapman & Hall/CRC, Boca Raton, FL (2006) ISBN 978-1-58488-611-2; 1-58488-611-0; Applications to fluid structure interactions
[7] Plotnikov, P.I., Sokolowski, J.: Shape derivative of drag functional. SIAM J. Control Optim. 48(7), 4680–4706 (2010) ISSN 0363-0129
[8] Plotnikov, P.I., Sokolowski, J.: Compressible Navier-Stokes equations. Theory and shape optimization. Monografie Matematyczne, vol. 73. Birkhäuser, Basel (2012) ISBN 978-3-0348-0366-3
[9] Slawig, T.: Distributed control for a class of non-newtonian fluids. Journal of Differential Equations 219(1), 116–143 (2005)
[10] Sokołowski, J., Stebel., J.: Shape sensitivity analysis of time-dependent flows of incompressible non-Newtonian fluids. Control and Cybernetics 40(4), 1077–1097 (2011)
[11] Sokołowski, J., Zolésio, J.-P.: Introduction to shape optimization. Shape sensitivity analysis. Springer Series in Computational Mathematics, vol. 16. Springer, Berlin (1992)
[12] Tröltzsch, F.: Optimal control of partial differential equations. Graduate Studies in Mathematics, vol. 112. American Mathematical Society, Providence (2010)
[13] Wachsmuth, D., Roubíček, T.: Optimal control of planar flow of incompressible non-Newtonian fluids. Z. Anal. Anwend. 29(3), 351–376 (2010)

Finite Element Discretization in Shape Optimization Problems for the Stationary Navier-Stokes Equation

Dan Tiba

Institute of Mathematics (Romanian Academy)
and Academy of Romanian Scientists Bucharest
dan.tiba@imar.ro

Abstract. For shape optimization problems associated to stationary Navier-Stokes equations, we introduce the corresponding finite element approximation and we prove convergence results.

Keywords: shape optimization, full discretization, finite elements, convergence.

1 Introduction

Optimal design and optimal control problems for partial differential equations are extensively studied in the recent mathematical literature. In the case of stationary Navier-Stokes equations, we quote the works Casas, Mateos and Raymond [2007], Rösch and Vexler [2006], Los Reyes and Tröltzsch [2007] devoted to optimal control problems or to approximation procedures. Shape optimization problems related to fluid mechanics have been discussed in Borrvall and Petersson [2003], Mohammadi and Pironneau [2001], Posta and Roubicek [2007], Roubicek and Tröltzsch [2003], Halanay and Tiba [2009]. See as well [7], [8] for related problems and arguments.

This work is concerned with the discretization and the associated convergence analysis, in the spirit of general shape optimization problems for linear elliptic systems, as discussed in Chenais and Zuazua [2006] and in Tiba [2011]. Another approximation procedure for such problems is due to Neittaamäki, Pennanen and Tiba [2009].

In the next section we formulate the problem and review briefly some preliminaries, necessary in the subsequent parts. Section 3 investigates some approximation properties of the stationary Navier-Stokes equation under our discretization approach. The last section introduces the fully discretized optimization problem and studies its convergence.

2 Problem Formulation and Preliminaries

Let $\Omega \subset R^d$ be an (unknown) lipschitzian domain, such that $E \subset \Omega \subset D \subset R^d$ with $E \subset D$ some given bounded domains and d an arbitrary natural number. We recall from Temam [1979] the definition of the following spaces :

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 437–444, 2013.

$$\mathcal{V}(\Omega) = \{y \in \mathcal{D}(\Omega)^d;\ \ div\ y = 0\}, \tag{1}$$

$$V(\Omega) = \ closure\ of\ \mathcal{V}(\Omega)\ in\ H_0^1(\Omega)^d. \tag{2}$$

Then, it is known that $V(\Omega) = \{y \in H_0^1(\Omega)^d, div\ y = 0\}$, as Ω is assumed lipschitzian. For any $y \in V(\Omega)$, if $\widetilde{y}$ is its extension by 0 to D, then $\widetilde{y} \in V(D)$ and conversely, if $\widetilde{z} \in V(D)$ and $\widetilde{z} = 0$ a.e. in $D\backslash\Omega$; then $z = \widetilde{z}|_\Omega \in V(\Omega)$. Such properties may be partially extended to domains with the segment property, Wang and Yang [2008].

The weak formulation of the stationary Navier-Stokes equation with Dirichlet (no-slip) boundary conditions is

$$\int_\Omega (\nu \sum_{i,j=1}^d \frac{\partial y_j}{\partial x_i}\frac{\partial v_j}{\partial x_i} + \sum_{i,j=1}^d y_i \frac{\partial y_j}{\partial x_i} v_j)dx = \int_\Omega \sum_{j=1}^d f_j v_j dx, \forall v \in V(\Omega) \tag{3}$$

where $f = (f_1, \cdots, f_d) \in H^{-1}(D)^d$ and $\nu > 0$ is the viscosity.

By Theorem 1.2 from Temam [1979], the equation (3) has at least one solution $y \in V(\Omega)$. If $d > 4$, the supplementary condition $y \in [L^d(\Omega)]^d$ should be included in the definition (1), (2) of $V(\Omega)$.

We associate to (3) an integral cost functional of the form

$$\int_\Lambda j(x, y(x))dx \tag{4}$$

where Λ is either $E \subset \Omega$ or Ω and y is one of the weak solutions of (3). The integrand $j : D \times R^d \to R$ satisfies measurability and continuity properties to be precised later.

The shape optimization problem considered in this paper consists in the minimization of the performance index (4) subject to the state system (3) and to the constraints

$$E \subset \Omega \subset D, \tag{5}$$

for any $\Omega \in \mathcal{O}$, where $\mathcal{O}$ is a prescribed family of domains. If the Lipschitz assumption is valid for any $\Omega \in \mathcal{O}$ with a uniform constant, then $\mathcal{O}$ is compact with respect to the Hausdorff-Pompeiu complementary metric. A similar compactness result holds for domains with the uniform segment property according to Theorem A3.9, Neittaanmäki, Sprekels and Tiba [2006]. The following existence result is a simplified version of Theorem 1 in Halanay and Tiba [2009].

Theorem 1. *Assume that $j(x, y_n(x)) \to j(x, y(x))$ weakly in $L^2(E)$ if $y_n \to y$ strongly in $L^2(E)$ and $\mathcal{O}$ is compact. Then, the shape optimization problem (3)-(5), with $\Lambda = E$ has at least one optimal pair $[\Omega^*, y^*] \in \mathcal{O} \times V(\Omega^*)$ if it has an admissible pair.*

Remark. This theorem should be understood in the sense of singular control problems Lions [1983], Neittaanmäki, Sprekels and Tiba [2006, 3.1.3.1]. The state system is ill-posed (nonuniqueness), but the optimization problem (3)-(5) is well defined as minimization over admissible pairs $[\Omega, y]$, $\Omega \in \mathcal{O}$ satisfying (5) and $y \in V(\Omega)$ being one of the weak solutions of (3).

3 Discretization of the State Equation

We assume now that D is a smooth bounded subdomain of R^2 and we consider a family of uniformly regular finite element meshes $\{\mathcal{T}_h\}_{h>0}$ in D with $h = \max_{T_h \in \mathcal{T}_h} diam(T_h)$.

For any admissible $\Omega \in \mathcal{O}$, we define its discrete approximation as follows (Chenais and Zuazua [2006] or Tiba [2011] where other variants are also discussed) :

$$\Omega_h = int \cup \{\overline{T}_h; T_h \in \mathcal{T}_h, T_h \subset \Omega\} \tag{6}$$

According, for instance, to Temam [1979], there are many possibilities to introduce a finite element space V_h in Ω_h approximating (2), that is approximating $H_0^1(\Omega)$ and the divergence free condition. In particular, the piecewise linear finite elements are not possible to be used in this setting. One also has to impose null values on $\partial\Omega_h$ in order to take account the Dirichlet boundary condition and any $y_h \in V_h$ may be extended by 0 to Ω, respectively to D. We shall also write $V_h(\Omega)$ or $V_h(D)$ in order to avoid possible confusions.

One example of space V_h (in dimension 2 as assumed here) is the space of continuous functions, vanishing outside Ω_h, that are polynomials of degree less or equal two on any simplex $T \in \mathcal{T}_h$ and satisfy :

$$\int_T div y_h dx = 0, \forall T \in \mathcal{T}_h, \forall y_h \in V_h \tag{7}$$

On V_h we take the scalar product $(\cdot,\cdot)_h$ induced by $H_0^1(\Omega)$. Note that V_h is an external approximation of V due to (7). The discrete approximation of (3) is

$$\nu(y_h, v_h)_h + b_h(y_h, y_h, v_h) = \int_\Omega f \cdot v_h dx, \forall v_h \in V_h \tag{8}$$

Notice that the last integral in (8) is over Ω_h in fact, as v_h vanishes outside Ω_h. We have denoted by "$\cdot$" the scalar product in R^2 and $b_h(\cdot,\cdot,\cdot)$ is the trilinear form approximating

$$b(y, v, w) = \sum_{i,j=1}^{2} \int_\Omega y_i D_i v_j w_j dx, \forall y, v, w \in H_0^1(\Omega).$$

A detailed construction of $b_h(\cdot,\cdot,\cdot)$ and the proof of

$$b_h(u_h, u_h, r_h v) \to b(u, u, v), \forall v \in \mathcal{V}(\Omega) \tag{9}$$

if $u_h \to u$ weakly in $H_0^1(\Omega)$ can be found in Teman [1979], Ch. II.3.

Here $r_h v \in V_h$ is given by a term that takes the same values as $v \in \mathcal{V}(\Omega)$ in the interior nodes and edge midpoints of Ω_h plus a correction term defined in Temam [1979, p.81]. On $\partial\Omega_h$, $r_h v$ should be zero.

Then, the following convergence property is also valid.

Proposition 3.1. Under the above conditions, there exists at least one $u_h \in V_h$, solution of (8) *, for each $h > 0$.*

The Family. $\{u_h\}$ in $H_0^1(\Omega)$ has strong accumulation points, denoted $\bar{u}$, which are solutions of (3)

Remark. If the uniqueness property is valid for (3), the convergence is valid without taking subsequences. In Casas, Mateos and Raymond [2007] and in Girault and Raviart [1989] Ch. II 4, finite element approximations with uniform convergence properties are indicated, including error estimates.

4 Approximation of the Shape Optimization Problem

We also discretize the cost functional (4) and the constraint (5) :

$$J_h(y_h) = \int\limits_{E_h} j(x, y_h(x))dx \tag{10}$$

where y_h is any of the solutions of (8), associated to Ω_h and E_h is obtained as in (6), starting from E ;

$$E_h \subset \Omega_h \subset D. \tag{11}$$

Notice that for any admissible $\Omega \in \mathcal{O}$, restriction (11) is automatically fulfilled by our discretization construction. The collection of all admissible discretized open sets is denoted by $\mathcal{O}_h$. The discrete shape optimization problem is defined by (8), (10), (11). By (6), the family $\mathcal{O}_h$ is always finite, for any given $h > 0$. Then, the discrete minimization problem has at least one discrete optimal solution denoted by $\Omega_h^* \in \mathcal{O}_h$. Since (8) may have, in principle, an infinity of solutions y_h^n, we remark that in each $T \in \mathcal{T}_h$, $T \subset \Omega_h$, the corresponding coefficients of y_h^n are bounded, by the construction of the finite elements. This is a consequence of $|y_h^n|_{V_h}$ bounded and it is enough to pass to the limit in (8), (10) on a minimizing sequence (with respect to n) of admissible states (h and Ω_h are fixed here). The minimization in (10) should be understood as minimization over pairs $[\Omega_h, y_h] \in \mathcal{O}_h \times V_h(\Omega_h)$, similar to the situation in *Theorem 1.*

We recall first some convergence properties of the admissible pairs $[\Omega_h, y_h] \in \mathcal{O}_h \times V_h(\Omega_h)$, when $h \to 0$.

Proposition 4.1 i) If $\Omega \in \mathcal{O}$, then $\Omega_h \in \mathcal{O}_h$ and $\Omega_h \to \Omega$ in the Hausdorff-Pompeiu complementary topology.

ii) If $\Omega_h \in \mathcal{O}_h$ and $\Omega_h \to \hat{\Omega}$ in the Hausdorff-Pompeiu complementary topology, then $\hat{\Omega} \in \mathcal{O}$.

Remark. At point ii), the discrete sets Ω_h are not necessarily constructed via (6) starting from $\hat{\Omega}$. Point i) also applies to the discretization of E and $E_h \to E$ in the Hausdorff-Pompeiu complementary topology. The proof of this proposition and other related properties may be found in Chenais and Zuazua [2006] and in Tiba [2011].

In the sequel, a crucial role is played by the following result which is an extension of *Proposition 3.1.*

Theorem 2. *If $\Omega_h \in \mathcal{O}_h$ and $y_h \in V_h$ is any solution of (8) and if $\Omega_h \to \hat{\Omega}$ in the Hausdorff-Pompeiu complementary topology, then for any subdomain $\mathcal{K}$, compactly included in $\hat{\Omega}$ there is $h_0 > 0$ such that $\mathcal{K} \subset \Omega_h$, $h < h_0$ and*

$$y_h|_{\mathcal{K}} \to \hat{y}|_{\mathcal{K}} \tag{12}$$

weakly in $H^1(\mathcal{K})$, on a subsequence, where $\hat{y} \in V(\hat{\Omega})$ is a solution of (3) in $\hat{\Omega} \in \mathcal{O}$.

Proof
The fact that $\hat{\Omega} \in \mathcal{O}$ is a consequence of *P4.1.* The inclusion $\mathcal{K} \subset \Omega_h$ for $h < h_0$ is known as the Γ-property of the Hausdorff-Pompeiu complementary convergence, Neittaanmäki, Sprekels and Tiba [2006], p. 63.

Extend y_h by 0 to D and denote it by $\tilde{y}_h \in H_0^1(D)$. By Temam [1979], p. 209, we have

$$b_h(u_h, v_h, v_h) = 0, |b_h(u_h, v_h, w_h) \ \leq \ c|u_h|_{V_h}|v_h|_{V_h}|w_h|_{V_h} \tag{13}$$

for any u_h, v_h, w_h in V_h, where $c > 0$ is an absolute constant.

Fixing $v_h = y_h \in V_h$ in (8) we get that $\{|y_h|_{V_h}\}$ is bounded, due to (13), and $\{\tilde{y}_h\}$ is bounded in $H_0^1(D)$. On a subsequence, we have $\tilde{y_h} \to \tilde{y} \in H_0^1(D)$. A simple distributions argument gives that $\tilde{y}|_{D\setminus\hat{\Omega}} = 0$ almost everywhere. Then $\tilde{y}|_{\hat{\Omega}} \in H_0^1(\hat{\Omega})$ as we have assumed that any admissible domain $\hat{\Omega} \in \mathcal{O}$ is lipschitzian and the trace theorem may be applied.We also get $\tilde{y} \in V(\hat{\Omega})$ by an adaptation of Proposition 4.3, Temam[1979], p.83. In particular $y_h|_{\mathcal{K}} \to \tilde{y}|_{\mathcal{K}}$ weakly in $H^1(\mathcal{K})$, on a subsequence.

We have to show that $\tilde{y}|_{\hat{\Omega}}$ is a solution of (3). We fix in (8) $v_h = r_h v$ for any $v \in \mathcal{V}(\hat{\Omega})$. In particular *supp* $v \subset \hat{\Omega}$ is a compact subset and the Γ-property gives that *supp* $v \subset \Omega_h$ for $h < h_0$. Consequently $r_h v \in V_h$ for $h < h_0$ and may be used in (8). Moreover, by (9) we have

$$b_h(y_h, y_h, r_h v) \to b(\tilde{y}, \tilde{y}, v), \forall v \in \mathcal{V}(\hat{\Omega}). \tag{14}$$

Relation (14) is obtained by applying (9) in D as $\tilde{y} \in H_0^1(D)$, $v \in \mathcal{V}(D)$ by extending it with 0 outside $\hat{\Omega}$ and since $\tilde{y}_h \to \tilde{y}$ weakly in $H_0^1(D)$. The formulas for $b(\cdot,\cdot,\cdot)$ and $b_h(\cdot,\cdot,\cdot)$ are not affected by these extensions.

One can pass to the limit in (8) by (14) and the strong convergence $\widetilde{r_h v} \to \tilde{v}$ in $H_0^1(D)$ due to the regularity of $v \in \mathcal{V}(\hat{\Omega})$. This ends the proof since $\mathcal{V}(\hat{\Omega})$ is dense in $V(\hat{\Omega})$ and (3) may be obtained.

Remark. In fact, we have shown that the extensions

$$\tilde{y}_h \to \tilde{y}$$

weakly in $H_0^1(D)$, on a subsequence. If the solution of (3) is unique, the convergence is valid on the whole sequence.

Theorem 3. *i) Any accumulation point of any sequence* $\{\Omega_h^*\}_{h\to 0}$ *of discrete minimizers of (10) is a continuous minimizer* Ω^* *of (4).*

ii) $J_h(\Omega_h^*) \to J(\Omega^*)$ *for* $h \to 0$*, on the initial sequence.*

Proof i) Clearly $\{\Omega_h^*\}$, $h > 0$ is relatively compact in the Hausdorff-Pompeiu complementary metric and we may assume that $\Omega_h^* \to \hat{\Omega}$ on a subsequence; where $\hat{\Omega} \in \mathcal{O}$ by *Proposition 4.1.*

By *Theorem 2*, we get $\tilde{y}_h|_E \to \hat{y}|_E$ strongly in $L^2(E)$, where $\tilde{y}_h$ is the extension by 0 of y_h and $\hat{y}$ is a solution of (3) in $\hat{\Omega}$. The convergence is valid on a subsequence.

We have $J_h(\Omega_h^*) \to J(\hat{\Omega})$. This is a consequence of $j(x, \tilde{y}_h) \to j(x, \hat{y})$ weakly in $L^2(E)$ (see the assumption on $j(\cdot,\cdot)$ in *Theorem 1*) and of

$$J_h(\Omega_h^*) = \int\limits_{E_h} j(x, y_h)dx = \int\limits_{E} j(x, \tilde{y}_h)dx - \int\limits_{E\setminus E_h} j(x, \tilde{y}_h)dx \tag{15}$$

The last integral in (15) converges to 0 as $meas(E\setminus E_h) \to 0$, Tiba [2011], and $j(x, \tilde{y}_h)$ is bounded in $L^2(E)$, which is argued above.

For any $\Omega \in \mathcal{O}$, we can construct Ω_h as in (6) and again by *Theorem 2* and *Proposition 4.1* we obtain that $J_h(\Omega_h) \to J(\Omega)$. Taking into account that

$$J_h(\Omega_h^*) \leq J_h(\Omega_h)$$

we infer that $J(\hat{\Omega}) \leq J(\Omega)$ for any $\Omega \in \mathcal{O}$, i.e. $\hat{\Omega}$ is optimal for the problem (3)-(5) and we redenote it by Ω^*.

ii) This is a consequence of i) as the minimal value $J(\Omega^*)$ is uniquely associated to $\mathcal{O}$.

Remark. The results of this section may be extended to the cost functional corresponding to the choice $\Lambda = \Omega$ by using supplementary arguments as in Neittaanmäki, Sprekels and Tiba [2006], p. 472.

Remark. The approach of this paper is based on a fixed grid given in the whole domain D, i.e. it is a fixed domain method. It should be noticed that the finite dimensional optimization problem is nonconvex and it is not easy to find a global minimum Ω_h^* , $h > 0$.

Starting with some initial guess $\tilde{\Omega} \in \mathcal{O}$, one can define $\tilde{\Omega}_h \in \mathcal{O}_h$ by (6) and use it as initial iteration in some descent algorithm for the finite dimensional problem. Denote by $\dot{\Omega}_h$ the obtained finite dimensional "solution" (which is not necessarily a global minimum of J_h). Then, reading (6) in the converse sense, we get at least one $\dot{\Omega} \in \mathcal{O}$, corresponding to $\dot{\Omega}_h$. If the descent property for J_h "dominates" the approximation error between (3) and (8), then $J(\dot{\Omega}) < J(\tilde{\Omega})$, i.e. the method may find a better admissible domain from the point of view of the cost J.

Acknowledgement. This work was supported by Grant 145/2011 of CNCS, Romania.

References

1. Borrvall, T., Petersson, J.: Topology optimization of fluids in Stokes flow. International Journal for Numerical Methods in Fluids 41, 77–107 (2003)
2. Casas, E., Mateos, M., Raymond, J.-P.: Error estimates for the numerical approximation of a distributed control problem for the steady-state Navier-Stokes equations. SIAM J. Control Optim. 46(3), 952–982 (2007)
3. Chenais, D., Zuazua, E.: Finite-element approximation of 2 D elliptic optimal design. J. Math. Pures Appl. 85, 225–249 (2006)
4. Girault, P., Raviart, P.-A.: Finite element methods for Navier-Stokes equations. Theory and algorithms. Springer, Berlin (1986)
5. Halanay, A., Tiba, D.: Shape optimization for stationary Navier-Stokes equations. Control and Cybernetics 38(4B), 1359–1374 (2009)
6. Lions, J.L.: Contrôle des systèmes distribuiés singuliers. Gauthier-Villars, Paris (1983)
7. Liu, W.B., Neittaanmäki, P., Tiba, D.: On the structural optimization problems. C.R.A.S. Paris, Serie I - Mathematique 331(1), 101–106 (2000)
8. Liu, W.B., Tiba, D.: Error estimates in the approximation of optimization problems governed by nonlinear operators. Numer. Funct. Anal. Optim. 22(7-8), 953–972 (2001)
9. De Los Reyes, J.C., Tröltzsch, F.: Optimal control of the stationary Navier-Stokes equations with mixed control-state constraints. SIAM J. Control Optimiz. 46(2), 604–629 (2007)
10. Mohhamadi, B., Pironneau, O.: Applied shape optimization for fluids. Oxford University Press, New York (2001)
11. Neittaanmäki, P., Sprekels, J., Tiba, D.: Optimization of elliptic systems. Theory and applications. Springer, New York (2006)
12. Neittaanmäki, P., Pennanen, A., Tiba, D.: Fixed domain approaches in shape optimization problems with Dirichlet boundary conditions. J. of Inverse Problems 25, 1–18 (2009)
13. Posta, M., Roubicek, T.: Optimal control of Navier-Stokes equations by Oseen aproximations. Comput. Math. Appl. 53(3-4), 569–581 (2007)
14. Roubicek, T., Tröltzsch, F.: Lipschitz stability of optimal controls for steady-state Navier-Stokes equation. Control and Cybernetics 32, 683–705 (2003)
15. Rösch, A., Vexler, B.: Optimal control of the Stokes equations: a priori error analyses for finite element discretization with postprocessing. SIAM J. Numer. Anal. 44(5), 1903–1920 (2006)

16. Temam, R.: Navier-Stokes equations. Theory and numerical analysis. North-Holland, Amsterdam (1979)
17. Tiba, D.: Finite element approximation for shape optimization problems with Neumann and mixed boundary conditions. Submitted for SIAM J. Control Optim. 49(3), 1064–1077 (2011)
18. Wang, G., Young, D.: Decomposition of vector-valued divergence free Sobolev functions and shape optimization for stationary Navier-Stokes equations. Comm. Part. Diff. Eq. 33, 429–449 (2008)

Strong Shape Derivative for the Wave Equation with Neumann Boundary Condition

Jean-Paul Zolésio[1] and Lorena Bociu[2]

[1] CNRS-INLN, 1136 route des Lucioles, 06902 Sophia Antipolis France and CRM (Applied Math lab.) Montréal, Canada
Jean-Paul.Zolesio@inln.cnrs.fr
[2] NC State University, Department of Mathematics, Raleigh, NC 27695, USA

Abstract. The paper provides shape derivative analysis for the wave equation with mixed boundary conditions on a moving domain Ω_s in the case of non smooth neumann boundary datum. The key ideas in the paper are (i) bypassing the classical sensitivity analysis of the state by using parameter differentiability of a functional expressed in the form of Min-Max of a convex-concave Lagrangian with saddle point, and (ii) using a new regularity result on the solution of the wave problem (where the Dirichlet condition on the fixed part of the boundary is essential) to analyze the strong derivative.

1 Introduction

The aim of this paper is to give a full analysis of the shape differentiability for the solution to the wave equation with mixed boundary conditions on a moving domain Ω_s. The shape derivative investigation has been solved for the wave equation with homogeneous and non-homogeneous Dirichlet boundary conditions [1, 9, 13]. The novelty and difficulty of the paper are represented by the fact that the wave equation has non-homogeneous Neumann boundary condition on part of the boundary of its geometrical domain. Moreover, the Neumann datum g is non-smooth, i.e. $g \in H^{-1/2}(\partial\Omega_s)$.

First, we prove existence of weak shape derivative by using parameter differentiability of a functional expressed in the form of Min-Max of a convex-concave Lagrangian with saddle point [3, 13]. This completely bypasses the classical sensitivity analysis of the state (solution) [9]. A lot of problems in shape sensitivity analysis can be expressed as a Min Max of some Lagrangian dependent on the domain Ω. Using a velocity field of deformations V over Ω one can build a family of perturbations Ω_s, $s \geq 0$, and then the tunics extends to situation when after some change of variable the sensitivity analysis reduces to the study of the differentiability of a Min Max Lagrangian functional with respect to the parameter s for fixed velocity fields V and domains Ω.

Then we use a new regularity result for the solution to the wave problem to analyze the strong derivative via a brute force estimate on the differential quotient. In particular, we study the variational solution y_s in the space

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 445–460, 2013.

$W^{1,\infty}(0,\tau,L^2(\Omega_s)) \cap L^\infty(0,\tau,H^1_*(\Omega_s) = \{\phi \in H^1(\Omega_s),\ \phi = 0 \text{ on } S\})$ by taking advantage of the Dirichlet condition on the fixed part of the boundary S and we derive sharp estimate for y_s at the boundary Γ_s in terms of geometrical constants, in view of controlling the differential quotient's regularity that is necessary in the proof of our main result.

The new results obtained in this paper are: (i) existence of strong material and shape derivatives for the solution to the wave problem with mixed boundary conditions, and (ii) the new wave equation that the shape derivative solves.

The rest of the paper is organized as follows. In Section 2, we provide a preliminary result on existence and uniqueness of a Galerkin solution for the classical wave equation with variable coefficients. This result will be needed in the proof of our main theorem. In Section 3, we introduce the PDE model for the wave equation on a moving domain Ω_s and briefly recall the velocity method from shape optimization. In Section 4, we prove existence for the weak material derivatives. Finally, in Section 5 we complete the analysis by proving existence of strong material and shape derivatives, using the "extractor strategy" introduced in [5, 6] and Fourier transform techniques.

2 Galerkin Solution for the Wave Equation

Let $D \subset \mathbb{R}^N$ be fixed (potentially included in a C^2 manifold). Let $\Omega \subset D$ be an open bounded domain, with smooth boundary $\partial\Omega = \Gamma \cup S$, where $\bar{\Gamma} \cap \bar{S} = \emptyset$. Let $A(x) = \{a_{ij}(x)\}$ be a matrix of functions defined on Ω with the following properties:

$$\begin{cases} a_{ij} \in L^\infty(\Omega), \quad \text{and} \\ \exists \alpha > 0 \text{ such that } \forall x \in \Omega,\ \forall \zeta = \{\zeta_i\} \in \mathbb{R}^N,\ a_{ij}\,\zeta_i\zeta_j \geq \alpha|\zeta|^2. \end{cases} \tag{2.1}$$

We associate with the matrix $A(x)$ the following operator

$$Ay \stackrel{\text{def}}{=} -\operatorname{div}(A(x)\nabla y).$$

Given $\tau > 0$ and the interval $I = [0, 2\tau]$, consider the following wave equation problem:

$$\begin{cases} y_{tt} + A.y = f & \Omega \\ \frac{\partial y}{\partial n_A} = g & \Gamma \\ y = 0 & S \\ y(0) = y_0, \quad y_t(0) = y_1 \end{cases} \tag{2.2}$$

Galerkin solution for (2.2) in $L^2(I, H^1(\Omega)) \cap H^1(I, L^2(\Omega))$.

Let $H^1_*(\Omega) = \{\phi \in H^1(\Omega),\ \phi = 0 \text{ on } S\ \}$ with norm $||\phi||^2 = \displaystyle\int_\Omega |\nabla\phi|^2\,dx$.

Proposition 2.1. *Let* $(y_0, y_1) \in H^1_*(\Omega) \times L^2(\Omega)$ *and* $f \in L^2(I \times \Omega)$ *or* $f \in H(\Omega) := W^{1,\infty}(I, H^{-1}(\Omega))$, *and* $g \in G(\Gamma) := W^{1,1}(I, H^{-1/2}(\Gamma))$. *Then the solution* y *to (2.2) verifies* $y \in E(\Omega) := L^\infty(I, H^1_*(\Omega)) \cap W^{1,\infty}(I, L^2(\Omega))$. *Moreover when* $f \in L^2(I \times \Omega))$, *the trace* y_Γ *of* y *at the boundary* Γ *verifies :*

$$y_\Gamma \in F(\Gamma) := H^{1/2}(0, \tau, L^2(\Gamma)) \cap H^{-1/2}(0, \tau, H^1(\Gamma)),$$

and we have the following estimates: There exists constants c, k *such that*

$$||y||^2_{E(\Omega)} = ||y_t||^2_{L^\infty(I,L^2(\Omega))} + ||y||^2_{L^\infty(I,H^1_*(\Omega))} \leq c\ |||(f,g)|||^2, \tag{2.3}$$

and

$$||y||^2_{F(\Gamma)} = ||y||^2_{H^{1/2}(0,\tau,L^2(\Gamma))} + ||y||^2_{H^{-1/2}(0,\tau,H^1(\Gamma))} \leq k|||(f,g)|||^2, \tag{2.4}$$

where

$$\begin{aligned} |||(f,g)|||^2 &= ||g||^2_{W^{1,1}(I,H^{-1/2}(\Gamma))} + ||f||^2 \\ +|y_0|^2_{H^1_*(\Omega)} &+ |y_1|^2_{L^2(\Omega)} + ||f(0)||^2_{H^{-1}(\Omega)} + ||g(0)||^2_{H^{-1/2}(\Gamma)}, \end{aligned} \tag{2.5}$$

and

$$||f||^2 = ||f||^2_{W^{1,1}(I,H^{-1}(\Omega))} \text{ or } ||f||^2_{L^2(I,L^2(\Omega))}.$$

3 The Wave Equation on Ω_s

Let $D \subset \mathbb{R}^N$ be fixed (potentially included in a C^2 manifold). Let $\Omega \subset D$ be an open bounded domain, with smooth boundary $\partial\Omega = \Gamma \cup S$, where $\bar{\Gamma} \cap \bar{S} = \emptyset$.

The moving domain: For $s \in [0, s^*]$, let V be a smooth vector field,

$$V \in C^0([0, s^*[, C^1(D, \mathbb{R}^N)) \text{ with } V \cdot n = 0 \text{ on } \partial D, \text{ and } V = 0 \text{ on } S.$$

The flow transformation associated to V is given by:

$$T_s(V) : \bar{D} \to \bar{D}, \text{ such that } T_s(V)(S) = S.$$

Using $T_s(V)$, we build the family of perturbed domains $\{\Omega_s\}_s$ as follows: $\Omega_s = T_s(V)(\Omega)$ and $\partial\Omega_s = S \cup \Gamma_s$, where $\Gamma_s = T_s(V)(\Gamma)$. The normal component of the vector field $V(s)$ on the boundary Γ_s (called the "normal speed") is denoted by $v(s)$, i.e. $v(s) = \langle V(s), n_s \rangle$, where $n_s = \nabla b_{\Omega_s}$ where $b(s) = b_{\Omega_s}$ stands for the oriented distance function to Ω_s. From [4], we know that its shape derivative verifies

$$b'(s) = -V(s) \circ p_s \text{ in a neighborhood of } \Gamma_s,$$

where p_s is the projection map onto Γ_s. Most of the time, the normal speed appears in the calculus evaluated at 0, hence we will use the following notation $v = v(0)$ for it.

The PDE model for the wave problem. Let $\Omega_s \subset D$ be a moving domain with boundary $\partial\Omega_s = \Gamma_s \cup S$, where $\bar{\Gamma}_s \cap \bar{S} = \emptyset$, and S is fixed with respect to the parameter s.

We consider the solution y^s to the wave equation in the cylinder $]0,\tau[\times\Omega_s$, with homogeneous Dirichlet condition on the fixed part S of the boundary, and verifying a non homogeneous Neumann condition on the moving part Γ_s:

$$\begin{cases} y^s_{tt} - \Delta y^s = 0 & \Omega_s \\ y^s = 0 & S \\ \frac{\partial}{\partial n_s} y_s = g(s) & \Gamma_s \end{cases}$$

For each s let the boundary datum g have the following regularity $g_s \in W^{1,1}(I, H^{-1/2}(\Gamma_s))$. Then we consider the element $y_s \in H_s = L^2(I, H^1_*(\Omega_s)) \cap H^1(I, L^2(\Omega_s))$ solution to

$$\forall \phi_s \in H_s,\ \int_0^{2\tau}\int_{\Omega_s} (-\frac{\partial}{\partial t} y_s \frac{\partial}{\partial t}\phi_s + \nabla y_s \cdot \nabla \phi_s) dx dt = \int_0^{2\tau}\int_{\Gamma_s} g_s \phi_s d\Gamma_s\, dt$$

As mentioned before, the goal of this paper is threefold. We want to prove existence of the material derivative $\dot{y}(\Omega; V) = \frac{\partial}{\partial s}[y_s \circ T_s]\Big|_{\{s=0\}}$ and of the shape derivative $y'(\Omega; V) = \dot{y}(\Omega; V) - \nabla y \cdot V(0)$, and to render the new wave problem whose solution is the shape derivative $y'(\Omega; V)$. The first step consists in proving the existence of weak material derivative.

4 Weak Material Derivatives

To prove existence of material derivatives, we will take advantage of the regularity of solution for the linear wave equation and use the parameter differentiability for any functional expressed in form of a Min Max of a convex-concave Lagrangian with saddle points. The complete prove of MinMax parameter differentiability under saddle point was given in [2]. The result in case of single unique saddle point (which is easier) is given in [3] with application to PDE problem. Of course on formal view point such results was known by ingeeniers as a "necessary expression" (assume it is differentiable then such is the expression. The difficult part being to prove the differentiability it self)

Let $R \in L^2(I, H^{-1}(\Omega))$ and $q \in H^{1/2}(0,\tau, H^{-1}(\Gamma))$. We consider the transported solution y^s in the non perturbed geometry, that is $y^s = y_s \circ T_s(V)$, and we set

$$j(s) = \int_0^\tau (\int_\Omega y^s R dx + \int_\Gamma y^s\, q\, d\Gamma\,)\, dt.$$

Then we obtain that

$$j(s) = MinMax_{\{(\phi;\psi)\in\times\}} \quad \mathcal{L}(s, \phi, \psi),$$

where the Lagrangian $\mathcal{L}$ is given by

$$\mathcal{L}(s,\phi;\,\psi) = \int_0^\tau (\int_\Omega \phi R dx + \int_\Gamma \phi\, q\, d\Gamma\,)\, dt$$

$$+ \int_0^\tau \int_\Omega (-J(s)\,\frac{\partial\phi}{\partial t}\frac{\partial\psi}{\partial t} + \langle A(s)\nabla\phi, \nabla\psi\rangle) dx dt - \int_0^\tau \int_\Gamma \omega(s)\, g_s \circ T_s(V)\, \psi\, d\Gamma dt,$$

where $J(s) = \det(DT_s)$ is the Jacobian of the transformation $T_s(V)$,

$$A(s) = J(s)(DT_s)^{-1}(DT_s)^{-*},$$

and the density $\omega(s)$ is given as

$$\omega(s) = \det(DT_s)|(DT_s(V))^{-*} n_s|. \tag{4.1}$$

At $s = 0$ the unique saddle point of $\mathcal{L}$ is (y, p), where the co-state p is the solution to the following problem:

$$\frac{\partial^2}{\partial t^2} p - \Delta p = R, \quad \frac{\partial}{\partial n} p = q.$$

We derive that the functional $\mathcal{L}$ is differentiable and we get the explicit expression for its derivative w.r.t s at $s = 0$:

$$\begin{gathered} j'(0) = \tfrac{\partial}{\partial s}\mathcal{L}(0, y, p) \\ = \int_0^\tau \int_\Omega (-\mathrm{div}V(0)\, \tfrac{\partial y}{\partial t}\, \tfrac{\partial p}{\partial t} + \langle [\mathrm{div}V(0) - 2\epsilon(V(0))\,]\nabla y, \nabla p\rangle)\, dx dt \\ - \int_0^\tau \int_\Gamma\ p\ \ (\dot{g}(V)\ +\ H\ g\ v\ ) d\Gamma dt \end{gathered}$$

where $2\epsilon(V) = DV + (DV)^*$, $\dot{g}(V) = [\frac{d}{ds} g_s \circ T_s(V)]_{s=0}$, $v = \langle V(0), n\rangle$, and H is the mean curvature of the boundary.

As a conclusion we get the existence of the weak derivative of the map $s \to y^s$ in $L^2(0, \tau, H^1(\Omega))$, and the weak differentiability of the trace mapping: $s \to y^s|_\Gamma$ in $H^{-1/2}(0, \tau, H^1(\Gamma))$.

5 Strong Material Derivative

Theorem 5.1. *Assume $g_s \in W^{1,1}(I, H^{-1/2}(\Gamma_s))$ such that there exists $\dot{g} \in L^1(I, H^{-1/2}(\Gamma))$ verifying*

$$\frac{g_s \circ T_s(V)}{s} - \dot{g}(V) \to 0 \ \textit{strongly in}\ L^1(I, H^{-1/2}(\Gamma)). \tag{5.1}$$

Then the solution y_s has a ***strong material derivative*** *in the following topology:*

$$\frac{y_s \circ T_s(V) - y}{s} - (\dot{Y} - div V(0)\ y) \to 0\ in\ L^\infty(0, \tau, L^2(\Omega)) \cap W^{-1,\infty}(0, \tau, H^1_*(\Omega)), \tag{5.2}$$

where $\dot{Y}$ is the solution to problem (5.10).

Corollary 5.1.1. *Let* $g_s \in W^{2,1}(I, H^{-1/2}(\Gamma_s))$ *such that there exists* $\dot{g} \in W^{1,1}(I, H^{-1/2}(\Gamma))$ *verifying*

$$\frac{g_s \circ T_s(V)}{s} - \dot{g}(V) \to 0 \text{ strongly in } W^{1,1}(I, H^{-1/2}(\Gamma)). \tag{5.3}$$

Then the solution y_s *has a* ***strong boundary material derivative*** *in the following topology:*

$$\frac{y_s \circ T_s(V) - y}{s} - (\dot{Y} - divV(0)\ y) \to 0 \text{ in } L^\infty(0, \tau, H^{1/2}(\Gamma)), \tag{5.4}$$

where $\dot{Y}$ *is the solution to problem (5.10).*

5.1 Shape Derivative

We know that

$$\exists\, \mathcal{Y} \in C^0\{[0, s*[,\ W^{1,\infty}(I, L^2(D)) \cap L^\infty(I, H^1_*(D))\} \ \cap\ C^1\{[0, s_*[,\ W^{1,\infty}(I, H^{-1}(D)) \cap L^\infty(I, L^2(D))\}$$

such that

$$\forall s, \quad \mathcal{Y}(s, .) = y_s(.) \text{ on } \Omega_s.$$

Now the term

$$y'(\Omega; V) := [\frac{\partial}{\partial s}\mathcal{Y}(0, x)]_{\{x \in \Omega\}} \in W^{1,\infty}(I, H^{-1}(\Omega)) \cap L^\infty(I, L^2(\Omega))$$

is independent of the choice of the function $\mathcal{Y}$ and is given by

$$y'(\Omega; V) = \dot{y}(\Omega; V) - \nabla y \cdot V(0).$$

For simplicity we write $\dot{y}$ and y' for the material and shape derivatives.

5.1.1 Characterization of $y'(\Omega; V)$

Proposition 5.1. *The element* y' *is solution to the wave problem:*

$$\begin{cases} \frac{\partial^2}{\partial t^2} y' - \Delta y' = 0 \text{ in } \Omega \\ \frac{\partial}{\partial n} y' = div_\Gamma(y \nabla_\Gamma v) - v \frac{\partial^2 y}{\partial t^2} + H g v + g'_\Gamma \text{ on } \Gamma, \end{cases} \tag{5.5}$$

where g'_Γ *stands for the* ***boundary shape derivative*** *of* g *given by*

$$g'_\Gamma = \dot{g} - \nabla_\Gamma g . V_\Gamma(0).$$

The proof is done in several steps. Let $Y_s = \int_0^t y_s(\sigma)\, d\sigma$ and $G_s = \int_0^t g_s(\sigma)\, d\sigma$ be the solution to

$$\forall \phi_s \in H(D),\ \int_0^{2\tau} \int_{\Omega_s} (-\frac{\partial}{\partial t} Y_s \frac{\partial}{\partial t}\phi_s + \nabla Y_s . \nabla \phi_s) dx dt = \int_0^{2\tau} \int_{\Gamma_s} G_s \phi_s d\Gamma_s\, dt.$$

From 5.9 we get

$$||\frac{\partial}{\partial t}Y_s||^2_{L^\infty(I,L^2(\Omega_s))} + ||Y_s||^2_{L^\infty(I,H^1_*(\Omega_s))} \leq c\,(||G_s||_{L^\infty(I,H^{-1/2}(\Gamma_s))} + ||g_s||_{L^1(I,H^{-1/2}(\Gamma_s))}) \quad (5.6)$$

Consider the symmetrical matrix $A(s) = J(s)\ DT_s(V)^{-1}.DT_s(V)^{-*} \in L^\infty(D, R^{N^2})$. Setting $Y^s = J_s\ \ Y_s \circ T_s(V)$ we get (with $\phi_s = \psi \circ T_s(V)^{-1}$)

$$\forall \psi \in H = H_0,\ \int_0^{2\tau}\int_\Omega (-\frac{\partial}{\partial t}Y^s\frac{\partial}{\partial t}\psi + A(s).\nabla(J_s^{-1}\ Y^s).\nabla\psi dxdt = \int_0^{2\tau}\int_\Gamma G^s w(s)\psi d\Gamma\, dt$$

Concerning the continuity of $s \to Y^s$ let $Z^s = Y^s - Y \in H^1(I, L^2(\Omega)) \cap L^2(I, H^1_*(\Omega))$, setting $\boldsymbol{m}(s) = (DT_s^{-1}.DT_s^{-*} - I).\nabla_\Gamma Y^s + Y^s A.\nabla(J_s^{-1}) + G^s (DT_s^{-1}.DT_s^{-*} - I).n$

Lemma 5.1. *the term* $\frac{1}{s}\ ||m(s)||_{L^\infty(0,2\tau,L^2(\Omega))}$ *remains bounded when* $s \to 0$

Indeed from classical estimates (see [9]) we have

$$||DT_s^{-1}.DT_s^{-*} - I||_{L^\infty(D)^{N^2}} + ||A(s).\nabla J_s^{-1}||_{L^\infty(D)} \leq C_V\ s$$

We get with 5.6

$$\frac{1}{s}\ ||m(s)||_{L^\infty(0,2\tau,L^2(\Omega))} \leq C_V\ c\,(||G_s||_{L^\infty(I,H^{-1/2}(\Gamma_s))} + ||g_s||_{L^1(I,H^{-1/2}(\Gamma_s))}) \quad (5.7)$$

and rom the sharp regularity at the boundary 2.4 we get

Lemma 5.2.

$$m(s) \in H^{1/2}(0,\tau,L^2(\Gamma))$$

with the following estimate

$$\frac{1}{s}\ ||m(s)||^2_{H^{1/2}(0,\tau,L^2(\Gamma))} \leq c\,C_V \quad (5.8)$$

$$||G^s w(s) - G||_{L^\infty(\Gamma)} \leq C_V\ s$$

The element Z^s is solution of, $\forall \psi \in H = H_0$

$$\int_0^{2\tau}\int_\Omega (-\frac{\partial}{\partial t}Z^s\frac{\partial}{\partial t}\psi + <\nabla Z^s, \nabla\psi> dxdt = -\int_0^{2\tau}\int_\Omega <m(s), \nabla\psi> dxdt$$

$$+\int_0^{2\tau}\int_\Gamma (G^s w(s) - G)\psi d\Gamma\, dt$$

That is

$$\int_0^{2\tau}\int_\Omega(-\frac{\partial}{\partial t}Z^s\frac{\partial}{\partial t}\psi+<\nabla Z^s,\nabla\psi>dxdt=-\int_0^{2\tau}\int_\Omega\bar{f}(s)\psi dxdt$$

$$+\int_0^{2\tau}\int_\Gamma\bar{g}(s)\psi d\Gamma\,dt$$

Where

$$\bar{f}(s)=-div(\boldsymbol{m}(s))\to 0\ \ in\ \ W^{1,1}(I,H^{-1}(\Omega))$$

$$\bar{g}(s)=G^s\omega(s)-G+<\boldsymbol{m}(s),\boldsymbol{n}>\to 0\ \ in\ \ W^{1,1}(I,H^{-1/2}(\Gamma))$$

From 2.5 we get:

$$||Z_t||^2_{L^\infty(I,L^2(\Omega))}+||Z||^2_{L^\infty(I,H^1_*(\Omega))}\le c\,(||G^s\omega-G+<m(s).n>||_{L^\infty(I,H^{-1/2}(\Gamma))}$$

$$+||g^s\omega-g+<\frac{\partial}{\partial t}m(s),n>||_{L^1(I,H^{-1/2}(\Gamma))})$$

$$+||\bar{f}||_{L^\infty(I,H^{-1}(\Omega))}+||\frac{\partial}{\partial t}\bar{f}||_{L^1(I,H^{-1}(\Omega))}+[c\,||||^2+||||\qquad(5.9)$$

$$2+||\bar{f}(0)||_{H^{-1}(\Omega)}+||\bar{G}(0)||_{H^{-1/2}(\Gamma)}]^{1/2}$$

Proposition 5.2. *Assume that $s\to g^s$ is continuous in $L^1(I,H^{-1/2}(\Gamma))$.*

We consider the element $\dot{Y}\in H_0$ solution to the problem

$$\forall\psi\in H=H_0,\ \int_0^{2\tau}\int_\Omega(-\frac{\partial}{\partial t}\dot{Y}\frac{\partial}{\partial t}\psi+\nabla\dot{Y}.\nabla\psi)dxdt$$

$$=\int_0^{2\tau}\int_\Omega<\nabla(Y\,divV(0))-\dot{A}.\nabla Y,\nabla\psi>dxdt$$

$$+\int_0^{2\tau}\int_\Gamma(\dot{G}+GH\,v)\,\psi d\Gamma\,dt$$

$$=\int_0^{2\tau}<\ div([\dot{A}-divV(0)\,I].\nabla Y),\psi>_{H^{-1}_{\bar{\Omega}}(D)\times H^1(\Omega)}dt$$

$$+\int_0^{2\tau}\int_\Gamma(\dot{G}+GH-<\dot{A}.n,\nabla Y>+<\nabla(Y\,divV(0)),n>)\,\psi d\Gamma\,dt\qquad(5.10)$$

This problem is relevant from the previous variational approach as the right hand side

$$f=div(\dot{A}\nabla Y-\nabla Y\ divV(0))=-2div(\epsilon(V(0)).\nabla Y)\in H^1(I,H^{-1}(\Omega))$$

$$\bar{g}=-<\dot{A}.n,\nabla Y>+<\nabla(Y\,divV(0)),n>$$

$$=2<\epsilon(V).\nabla Y,n>+Y\,\frac{\partial}{\partial n}divV(0)\in H^{1/2}(I,L^2(\Gamma))$$

5.2 Differential Quotient

We consider the elements

$$d(s) = \frac{y^s - y}{s} - \dot{y}, \quad \delta(s) = \frac{Y^s - Y}{s} - \dot{Y}$$

In order to characterise $\delta(s)$ we introduce the following vectors functions:

$$\mathcal{M}_1(s) = [\ \frac{DT_s^{-1}.DT_s^{-*}}{s} - 2\epsilon V(0)\ ].\nabla Y^s$$

$$\mathcal{M}_2(s) = 2\,\epsilon(V).\nabla(Y^s - Y)$$

$$\mathcal{M}_3(s) = [\,Y^s A.(\frac{\nabla(J_s^{-1})}{s} - \nabla div V(0)\ )\ +\ (YI - Y^s A).\nabla div V(0)\ ]$$

$$\mathcal{M}(s) = \mathcal{M}_1(s) + \mathcal{M}_2(s) + \mathcal{M}_3(s)$$

And the function on the boundary:

$$G_\delta(s) = \frac{G^s\omega(s) - G}{s} - (\dot{G} + G\,H\,v\,)$$

the element $\delta(s)$ is then solution to

$$\forall\psi, \int_0^{2\tau}\int_\Omega(-\frac{\partial}{\partial t}\delta\frac{\partial}{\partial t}\psi\ + \nabla\delta.\nabla\psi)dxdt = -\int_0^{2\tau}\int_\Omega <\ \mathcal{M}\ ,\ \nabla\psi\ > dt + \int_0^{2\tau}\int_\Gamma\ G_\delta(s)\ \ \psi d\Gamma\, dt$$

That is

$$\int_0^{2\tau}\int_\Omega(-\frac{\partial}{\partial t}\delta\frac{\partial}{\partial t}\psi\ + \nabla\delta.\nabla\psi)dxdt =$$

$$\int_0^{2\tau} <\ f_\delta(s), \psi\ >_{H^{-1}_{\tilde{\Omega}}(D)\times H^1(\Omega)}\ dt + \int_0^{2\tau}\int_\Gamma\ g_\delta(s)\ \psi d\Gamma\, dt \qquad (5.11)$$

Where

$$f_\delta(s) = div(\mathcal{M}\,) \to 0\ \ in\ H^1(I, H^{-1}(\Omega)),\ \ s \to 0,$$

$$g_\delta(s) =\ \ G_\delta(s)\ -\ \mathcal{M}.n\ \ \to 0\ \ in\ \ H^1(I, H^{-1/2}(\Gamma)),\ \ s \to 0.$$

5.3 Fourier Transform

From now on we assume the data, $y_0 = y_1 = 0$ and the Neumann data $g \in W^{1,1}(I, H^{-1/2}(\Gamma))$ being approached in this space by a smooth element g^m. Then the associated solution y^m is smoother. We consider a smooth cuting function $0 \le \rho(t) \le 1$ such that $\rho(t) = 0$ for $|t| \ge 2\tau$ while $\rho(t) = 1$ when $|t| \le \tau$. We consider $(y^m)^0$, $(g^m)^0$ the extension by zero out of I and we set

$$\tilde{y}^m(t,x) =\ \rho(t)\ (y^m)^0(t,x),\ \ \tilde{g}^m(t,x) =\ \rho(t)\ (g^m)^0(t,x)$$

With

$$F^m = \rho f + 2\rho_t\, y_t^m + \rho_{tt}\, y^m \to F = \rho f + 2\rho_t\, y_t + \rho_{tt}\, y \ \ in\ L^2(I, L^2(\Omega))\ as\ m \to \infty. \tag{5.12}$$

We get

$$\tilde{y}_{tt}^m + A.\tilde{y}^m = F^m, \frac{\partial \tilde{y}^m}{\partial n_A} = \tilde{g}^m,\ on\ \Gamma,\ \ \tilde{y}^m = 0\ on\ S. \tag{5.13}$$

For each m we consider the Fourier transform

$$\mathcal{F}^m(\zeta, x) = \int_{-\infty}^{+\infty} \tilde{F}^m(t,x)\, e^{-i\,\zeta\, t}\, dt,\ \ z^m(\zeta, x) = \int_{-\infty}^{+\infty} \tilde{y}^m(t,x)\, e^{-i\,\zeta\, t}\, dt,$$
$$\mathcal{G}^m(\zeta, x) = \int_{-\infty}^{+\infty} \tilde{g}^m(t,x)\, e^{-i\,\zeta\, t}\, dt$$

$$-\zeta^2\, z^m + A.z^m\ = \mathcal{F}^m, \quad \frac{\partial z^m}{\partial n_A} = \mathcal{G}^m,\ on\ \Gamma,\ \ z^m = 0\ on\ S, \tag{5.14}$$

5.3.1 Extractor. Given $\mu > 0$, consider the velocity field $V \in C^0([0, \mu[\,; W^{2,\infty}(D, \mathbf{R}^N))$ and its associated flow mapping $T_s(V)$. Given $s \geq 0$, denote by $\Omega_s = T_s(V)(\Omega)$ the perturbed domain with boundary $\Gamma_s = T_s(V)(\Gamma)$. Consider the functional

$$\mathcal{E}^m(s,V) \stackrel{\text{def}}{=} \int_{-\infty}^{+\infty} d\zeta \int_{\Omega_s(V)} \Big[|\zeta|\ \left| z^m \circ T_s(V)^{-1} \right|^2 + \frac{1}{1+|\zeta|}\, | < A.\nabla(z^m \circ T_s(V)^{-1}), \nabla(z^m \circ T_s(V)^{-1}) > | \Big] dx$$

and its derivative

$$e^m \stackrel{\text{def}}{=} \frac{d}{ds}\mathcal{E}^m(s,V)\Big|_{s=0}$$

that will be computed in two different ways. For simplicity in the following computations we denote by V the autonomeous vector field $V(0)$. Derivative by moving boundary results : Let

$$e_1^m \stackrel{\text{def}}{=} \int_{-\infty}^{+\infty} d\zeta \int_\Omega |\zeta|\, 2\mathcal{R}e\{z^m\, \nabla z^{\bar{m}} \cdot (-V)\} + \frac{1}{1+|\zeta|}\, 2\mathcal{R}e\{< A.\nabla(z^m),\ \nabla(\,\nabla z^{\bar{m}}(-V)\,) >)\}\, dx$$
$$+ \int_{-\infty}^{+\infty} d\zeta \left(\int_\Gamma \{|\zeta|\, |z^m|^2 + \frac{1}{1+|\zeta|}\ < A.\nabla z^m, \nabla z^m >\}\ < V, n >\ d\Gamma(x) \right).$$

Consider the first integral term over Ω:

$$a \stackrel{\text{def}}{=} \int_{-\infty}^{+\infty} d\zeta \int_{\Omega} \left(|\zeta|\, 2\,\mathcal{R}e\,\{z^m \nabla z^{\bar{m}} \cdot (-V)\} - \frac{1}{1+|\zeta|}\, 2\,\mathcal{R}e\,\{< A.\nabla z^m, \nabla(\nabla z^{\bar{m}}.V) >\} \right) dx$$

By Stokes theorem using the fact that $\partial z^m / \partial n_A = \mathcal{G}^m$, we get the following expression:

$$a = \int_{-\infty}^{+\infty} d\zeta \int_{\Omega} \left(|\zeta|\, 2\,\mathcal{R}e\,\{-z^m\, \nabla z^{\bar{m}} \cdot V\} + \frac{1}{1+|\zeta|}\, 2\,\mathcal{R}e\{\mathrm{div}(A.\nabla z^m)\, \nabla z^{\bar{m}}.V\} \right) dx$$
$$- \int_{-\infty}^{+\infty} d\zeta \int_{\Gamma} \frac{1}{1+|\zeta|}\, 2\mathcal{R}e\{\mathcal{G}^m\, \nabla z^{\bar{m}}.V\}\, d\Gamma$$

As we have $-\,\mathrm{div}(A.\nabla z^m) = A.z^m = (\zeta^2\, z^m\, + \mathcal{F}^m(\zeta, x))$ we get

$$a = 2\,\mathcal{R}e \left\{ \int_{-\infty}^{+\infty} d\zeta \int_{\Omega} \left(-z^m\, (|\zeta| + \frac{\zeta^2}{1+|\zeta|}\,) \right) \nabla z^{\bar{m}} \cdot V dx \right\}$$
$$+ 2\mathcal{R}e \left\{ \int_{-\infty}^{+\infty} d\zeta \int_{\Omega} \frac{1}{1+|\zeta|}\, \mathcal{F}^m \nabla z^{\bar{m}} \cdot V\, dx \right\}$$
$$- \int_{-\infty}^{+\infty} d\zeta \int_{\Gamma} \frac{1}{1+|\zeta|}\, 2\,\mathcal{R}e\{\mathcal{G}^m\, \nabla z^{\bar{m}}.V\}\, d\Gamma.$$

Lemma 5.3. *For each $m > 0$ we get*

$$e_1^m = \int_{-\infty}^{+\infty} d\zeta \left(\int_{\Gamma} \{|\zeta|\, |z|^2 + \frac{1}{1+|\zeta|}\, < A.\nabla z^m, \nabla z^m >\}\, < V, n >\, d\Gamma(x) \right)$$
$$- \int_{-\infty}^{+\infty} d\zeta \int_{\Gamma} \frac{1}{1+|\zeta|}\, 2\,\mathcal{R}e\{\mathcal{G}^m\, \nabla z^{\bar{m}}.V\}\, d\Gamma$$
$$+ 2\,\mathcal{R}e \left\{ \int_{-\infty}^{+\infty} d\zeta \int_{\Omega} \left(-z^m\, (|\zeta| + \frac{\zeta^2}{1+|\zeta|} \right) \nabla z^{\bar{m}} \cdot V\, dx \right\}$$
$$+ 2\,\mathcal{R}e \left\{ \int_{-\infty}^{+\infty} d\zeta \int_{\Omega} \frac{1}{1+|\zeta|}\, \mathcal{F}^m \nabla z^{\bar{m}} \cdot V\, dx \right\}.$$

5.3.2 Derivative by change of variable $T_s(V)$ Consider now the expression

$$\mathcal{E}^m(s, V) \stackrel{\text{def}}{=} \int_{-\infty}^{+\infty} d\zeta \int_{\Omega} \left[|\zeta|\, |z^m|^2 + \frac{1}{1+|\zeta|}\, < B(s).\nabla z^m, \nabla z^m > \right] j(s)\, dx$$

where

$$\boxed{B(s) \stackrel{\text{def}}{=} DT_s(V)^{-1}.A \circ T_s(V).(DT_s(V)^{-1})^*, \quad j(s) \stackrel{\text{def}}{=} \det DT_s(V)}$$

and

$$\boxed{(B'.V) = -DV.A - A.DV^* + \nabla A.V} \tag{5.15}$$

where $\nabla A.V$ is the matrix

$$\boxed{(\nabla A.V)_{i,j} \stackrel{\text{def}}{=} \nabla a_{i,j}.V.}$$

Then we get

Lemma 5.4.

$$e_2^m = \int_{-\infty}^{+\infty} d\zeta \int_\Omega \left[|\zeta|\,|z^m|^2 + \frac{1}{1+|\zeta|} < A.\nabla z^m, \nabla z^m > \right] \operatorname{div} V\, dx$$
$$+ \int_{-\infty}^{+\infty} d\zeta \int_\Omega \frac{1}{1+|\zeta|} < (B'.V).\nabla z^m, \nabla z^m > dx.$$

5.4 Extractor Identity

We now equate the two expressions, $e_1^m = e_2^m$, to get

Lemma 5.5.

$$\int_{-\infty}^{+\infty} d\zeta \left(\int_\Gamma \{|\zeta|\,|z^m|^2 + \frac{1}{1+|\zeta|} < A.\nabla z^m, \nabla z^m >\} < V, n > d\Gamma(x) \right)$$
$$- \int_{-\infty}^{+\infty} d\zeta \int_\Gamma \frac{1}{1+|\zeta|}\, 2\,\mathcal{R}e\{\mathcal{G}^m\, \nabla z^{\bar{m}}.V\}\, d\Gamma$$
$$= -2\,\mathcal{R}e \left\{ \int_{-\infty}^{+\infty} d\zeta \int_\Omega \left(-z^m\,(|\zeta| + \frac{\zeta^2}{1+|\zeta|}) \right) \nabla z^{\bar{m}} \cdot V\, dx \right\}.$$
$$- 2\,\mathcal{R}e \left\{ \int_{-\infty}^{+\infty} d\zeta \int_\Omega \frac{1}{1+|\zeta|} \mathcal{F}^m \nabla z^{\bar{m}} \cdot V\, dx \right\}$$
$$+ \int_{-\infty}^{+\infty} d\zeta \int_\Omega \left[|\zeta|\,|z^m|^2 + \frac{1}{1+|\zeta|} < A.\nabla z^m, \nabla z^m > \right] \operatorname{div} V\, dx$$
$$+ \int_{-\infty}^{+\infty} d\zeta \int_\Omega \frac{1}{1+|\zeta|} < (B'.V).\nabla z^m, \nabla z^m > dx.$$

The velocity vector V will be choosen in terms of the geometry of the boundary which is best handled by using the oriented distance function, (cf. [7]) $b_\Omega = d_\Omega - d_{\Omega^c}$, $\Omega^c = \mathbf{R}^N \setminus \Omega$. When Ω is of class $C^{1,1}$, the unit outward normal n to the boundary $\Gamma = \partial\Omega$ is equal to ∇b_Ω. Moreover when Γ is compact, there exist $h > 0$ and an h-tubular neighboorhood $\mathcal{U}_h = \{x \in \mathbf{R}^N : |b_\Omega(x)| < h\}$ of Γ such that $b_\Omega \in C^{1,1}(\mathcal{U}_h)$. In order to work in $\mathbf{R}^N$, localize the oriented distance function to this neighboorhood and work with a global $C^{1,1}$ function on $\mathbf{R}^N$, define $b_\Omega^h \stackrel{\text{def}}{=} \theta^h \circ b_\Omega$ for some function $\theta^h \in C^{1,1}(\mathbf{R}; [0,1])$ such that

$$\theta^h(t) \stackrel{\text{def}}{=} \begin{cases} 1, & , \quad |t| < h/3 \\ 0, & \quad |t| > 2h/3. \end{cases}$$

Obviously b^h_Ω is equal to b_Ω in $\mathcal{U}_{h/3}$ and supp $b^h_\Omega \subset \mathcal{U}_h$, with $b^h_\Omega \in C^{1,1}(\overline{\mathbf{R}})$ with the unit outward normal $n = \nabla b_\Omega = \nabla b^h_\Omega$ on Γ. In order to take care of the boundary condition term $\mathcal{G}$ in the extractor identity we have to make a specific choice for the vector field V. Choose V in the following form (here we assume the matrix A to be defined in the neighbourhood of $\bar{\Omega}$): $V \stackrel{\text{def}}{=} A.\nabla b^h_\Omega \quad \Rightarrow \quad V = A.n$ on Γ, so that $\nabla z^m.V =< \nabla z^m, A.n >= \mathcal{G}^m$, while $v =< V, n >=< A.n, n > \geq \alpha > 0$ on Γ. So the first term yields

$$
\begin{aligned}
&\alpha \int_{-\infty}^{+\infty} d\zeta \left(\int_\Gamma \{|\zeta|\,|z^m|^2 + \frac{1}{1+|\zeta|}\,|\nabla z^m|^2\}\, d\Gamma(x) \right) \\
&\leq \int_{-\infty}^{+\infty} d\zeta \int_\Gamma \frac{1}{1+|\zeta|}\, 2\,\mathcal{R}e\{\mathcal{G}^m \bar{\mathcal{G}}^m\}\, d\Gamma \ldots \\
&\quad \cdot \Bigg| - 2\,\mathcal{R}e \left\{ \int_{-\infty}^{+\infty} d\zeta \int_\Omega \left(-z^m(|\zeta| + \frac{\zeta^2}{1+|\zeta|}) \right) \nabla z^{\bar{m}} \cdot V dx \right\}. \\
&\qquad - 2\,\mathcal{R}e \left\{ \int_{-\infty}^{+\infty} d\zeta \int_\Omega \frac{1}{1+|\zeta|}\, \mathcal{F}^m \nabla z^{\bar{m}} \cdot V\, dx \right\} \\
&\qquad + \int_{-\infty}^{+\infty} d\zeta \int_\Omega \left[|\zeta|\,|z^m|^2 + \frac{1}{1+|\zeta|} < A.\nabla z^m, \nabla z^m > \right] \operatorname{div} V\, dx \\
&\qquad + \int_{-\infty}^{+\infty} d\zeta \int_\Omega \frac{1}{1+|\zeta|} < (B'.V).\nabla z^m, \nabla z^m > dx \Bigg|.
\end{aligned}
$$

and we obtain the following estimate.

Proposition 5.3. *There exists a constant $M > 0$ such that for all $m > 0$*

$$
\begin{aligned}
&\int_{-\infty}^{+\infty} d\zeta \left(\int_\Gamma \{|\zeta|\,|z^m|^2 + \frac{1}{1+|\zeta|}\,|\nabla z^m|^2\}\, d\Gamma(x) \right) \\
&\leq \frac{1}{\alpha} M \Bigg\{ \left\| \sqrt{\frac{1}{1+|\zeta|}}\, \mathcal{G}^m \right\|^2_{L^2(R_\zeta, L^2(\Gamma))} \\
&\qquad + \|V(0)\|_{L^\infty(D,\mathbf{R}^N)} \|z^m\|_{L^2(\mathbf{R}_\zeta, L^2(\Omega))}\, \|\nabla z^m\|_{L^2(\mathbf{R}_\zeta, L^2(\Omega, \mathbf{R}^N))} \\
&\qquad + \|V(0)\|_{L^\infty(D,\mathbf{R}^N)} \left\| \sqrt{\frac{1}{1+|\zeta|}} \mathcal{F}^m \right\|_{L^2(\mathbf{R}_\zeta, L^2(\Omega))} \|\nabla z^m\|_{L^2(\mathbf{R}_\zeta, L^2(\Omega, \mathbf{R}^N))} \\
&\qquad + \| \sqrt{\zeta}\, z^m \|_{L^2(\mathbf{R}_\zeta, L^2(\Omega))} \| \operatorname{div} V(0)\|_{L^\infty(D)} \\
&\qquad + \left\| \sqrt{\frac{1}{1+|\zeta|}} \nabla z^m \right\|_{L^2(\mathbf{R}_\zeta, L^2(\Omega, \mathbf{R}^N))} \|(B'.V) + A\|_{L^\infty(D, \mathbf{R}^{N^2})} \Bigg\}
\end{aligned}
\tag{5.16}
$$

As $V = A.\nabla b^h_\Omega$, we get

$$
\| \operatorname{div} V(0)\|_{L^\infty(D)} \leq M_3(\|A\|_{W^{1,\infty}(\Omega)} + \|\Delta b^h_\Omega\|_{L^\infty(D)})
$$

while

$$\|B'.V\|_{L^\infty(D)} \leq M_4(\ \|A\|_{W^{1,\infty}(\Omega)} + \|D^2 b^h_\Omega\|_{L^\infty(D)}\ \|A\|_{L^\infty(D,\mathbf{R}N^2)})$$

Each terms depending on m in the right-hand side of (5.16) converges in the respective norms:

$$\mathcal{G}^m \to \mathcal{G}, \quad \mathcal{F}^m \to \mathcal{F}, \quad z^m \to z.$$

So that there exists a constant $M_2 > 0$ such that, for all $m > 0$ we have:

$$\forall m > 0, \quad \int_{-\infty}^{+\infty} d\zeta \left(\int_\Gamma \{|\zeta|\,|z^m|^2 + \frac{1}{1+|\zeta|}\,|\nabla z^m|^2\}\, d\Gamma(x) \right) \leq M_2.$$

Consider the following two measures on $\mathbf{R}_\zeta$:

$$\mu_1(\zeta) \stackrel{\text{def}}{=} \sqrt{|\zeta|}\, d\zeta \quad \text{and} \quad \mu_2(\zeta) \stackrel{\text{def}}{=} \sqrt{\frac{1}{1+|\zeta|}}\, d\zeta.$$

Then we get

$$\|z^m\|_{L^2_{\mu_1}(\mathbf{R}_\zeta, L^2(\Gamma))} \leq M_2 \quad \text{and} \quad \|\nabla z^m\|_{L^2_{\mu_2}(\mathbf{R}_\zeta, L^2(\Gamma,\mathbf{R}^N))} \leq M_2$$

and there exist elements $\phi, \boldsymbol{\Phi}$ such that for the two associated weighted topologies we get the weak convergences:

$$z^m_\Sigma \rightharpoonup \phi \text{ weakly in } L^2_{\mu_1}(\mathbf{R}_\zeta, L^2(\Gamma))$$
$$\nabla z^m_\Sigma \rightharpoonup \Phi \text{ weakly in } L^2_{\mu_2}(\mathbf{R}_\zeta, L^2(\Gamma, \mathbf{R}^N))$$

Obviously $\phi = z|_\sigma$ and $\Phi = \nabla\phi|_\Sigma$ then $\Phi = \nabla z|_\Sigma$ and the norms being weakly l.s.c. in the limit we get the estimate:

$$\int_{-\infty}^{+\infty} d\zeta \left(\int_\Gamma \left\{ |\zeta|\,|z|^2 + \frac{1}{1+|\zeta|}|\nabla z|^2 \right\} d\Gamma(x) \right) \leq M_2$$

By Plancherel isomorphism we get in the real line, we get

Corollary 5.1.2.

$$|\zeta|^{1/2} z \in L^2(\mathbf{R}, L^2(\Gamma)) \quad \Longleftrightarrow \quad \tilde{y} \in H^{1/2}(\mathbf{R}, L^2(\Gamma)),$$
$$((1+|\zeta|)^{-1/2})\,\nabla z \in L^2(\mathbf{R}, L^2(\Gamma, \mathbf{R}^N)). \quad \Longleftrightarrow \quad \nabla\tilde{y} \in H^{-1/2}(\mathbf{R}, L^2(\Gamma, \mathbf{R}^N)).$$

Obviously we have:

$$\|y\|_{H^{1/2}(I, L^2(\Gamma))} \leq \|\tilde{y}\|_{H^{1/2}(\mathbf{R}, L^2(\Gamma))}$$

Also:

$$\nabla y|_\Sigma = \nabla_\Gamma y + \frac{\partial y}{\partial n}\, n$$

But

$$\frac{\partial y}{\partial n} = g \in L^2(\Sigma)$$

So that

$$\nabla_\Gamma y \in H^{-1/2}(\mathbf{R}, L^2(\Gamma, \mathbf{R}^N))$$

Which implies

$$y|_\Sigma \in H^{-1/2}(\mathbf{R}, H^1(\Gamma))$$

from which, as $y \in C^0([0,\tau], H^{3/5-\sigma}(\Omega))$*, we get (**??**).*

Proposition 5.4. *Let* Γ *be a* C^2 *submanifold in* $\mathbf{R}^N$ *and* $g \in L^2(I; L^2(\Gamma))$*,* $f \in L^2(I; L^2(\Omega))$*,* $y_0 \in H^2(\Omega)$*,* $y_1 \in L^2(\Omega)$*, and* $A \in W^{1,\infty}(D, \mathbf{R}^{N^2})$*. Then the trace at the boundary* Γ *of* y *verifies the regularity (2.4).*

We have:

$$||y||^2_{H^{1/2}(0,\tau,L^2(\Gamma))} + ||y||^2_{H^{-1/2}(0,\tau,H^1(\Gamma))} \leq ||\tilde{y}||^2_{H^{1/2}(0,2\tau,L^2(\Gamma))} + ||\tilde{y}||^2_{H^{-1/2}(0,2\tau,H^1(\Gamma))}$$

from the Plancherel isometry in the estimation 5.16, we get:

$$\leq C\{||\tilde{g}||^2_{H^{-1/2}(0,2\tau,L^2(\Gamma))} + ||\tilde{y}||^2_{H^{-1/2}(0,2\tau,H^1(\Omega))} + ||\tilde{F}||^2_{H^{-1/2}(0,2\tau,L^2(\Omega))}$$
$$+||\tilde{y}||^2_{H^{1/2}(0,2\tau,L^2(\Omega))} + ||\nabla\tilde{y}||^2_{H^{-1/2}(0,2\tau,L^2(\Omega))}$$

and obviously

$$\leq C_2\,\{||g||^2_{H^{1/2}(0,2\tau,L^2(\Gamma))} + ||y||^2_{H^{-1/2}(0,2\tau,H^1(\Omega))} + ||F||^2_{H^{-1/2}(0,2\tau,L^2(\Omega))}$$
$$+||y||^2_{H^{1/2}(0,2\tau,L^2(\Omega))} + ||\nabla y||^2_{H^{-1/2}(0,2\tau,L^2(\Omega)^N)}$$

5.4.1 Boundedness of *F*. From 5.12 there exists a constant $c_\rho > 0$ such that

$$||F||_{H^{-1/2}(0,2\tau,L^2(\Omega))} \leq ||F||_{L^2(0,2\tau,L^2(\Omega))} \leq c_\rho\,\{\,||y_t||^2_{L^2(I\times\Omega)} + ||y||^2_{L^2(I\times\Omega)} + ||f||^2_{L^2(I\times\Omega)}\,\}$$

and from 5.9 we get

$$||F||_{H^{-1/2}(0,2\tau,L^2(\Omega))} \leq ||F||_{L^2(0,2\tau,L^2(\Omega))} \leq c\,(||f|_{L^2(I\times\Omega)} + ||g(t)||_{L^\infty(I,H^{-1/2}(\Gamma))} + ||\frac{\partial}{\partial t}g(t)||_{L^1(I,H^{-1/2}(\Gamma))}\,)$$

$$+\,||f||_{L^\infty(I,H^{-1}(\Omega))} + ||\frac{\partial}{\partial t}f||_{L^1(I,H^{-1}(\Omega))} + [c\,E(0) + ||f(0)||_{H^{-1}(\Omega)} + ||g(0)||_{H^{-1/2}(\Gamma)}]^{1/2} \qquad (5.17)$$

Theorem 5.2. *Let $g \in W^{1,1}(0,2\tau,H^{-1/2}(\Gamma)) \cap H^{1/2}(0,2\tau,L^2(\Gamma))$, $f \in L^2(I \times \Omega) \cap W^{1,1}(I,H^{-1}(\Omega))$. Then there exist a constant k depending on the domain Ω, on the $L^\infty(I,W^{1,\infty}(\Omega))$-norm of the coefficients matrix A, on the cutting function ρ and on the trace of the linear trace operator (restriction to Γ in the norm of $\mathcal{L}(H^1_*(\Omega),H^{1/2}(\Gamma))$ such that:*

$$
\begin{aligned}
||y||^2_{H^{1/2}(0,\tau,L^2(\Gamma))} + ||y||^2_{H^{-1/2}(0,\tau,H^1(\Gamma))} &\leq k\,\{(||f|_{L^2(I\times\Omega)} \\
&+||g(t)||_{L^\infty(I,H^{-1/2}(\Gamma))} + ||\frac{\partial}{\partial t}g(t)||_{L^1(I,H^{-1/2}(\Gamma))}) \\
&+ \ ||f||_{L^\infty(I,H^{-1}(\Omega))} + ||\frac{\partial}{\partial t}f||_{L^1(I,H^{-1}(\Omega))} \\
&+ \ [\ E(0) + ||f(0)||_{H^{-1}(\Omega)} + ||g(0)||_{H^{-1/2}(\Gamma)}]^{1/2}\} \qquad (5.18)
\end{aligned}
$$

References

[1] Cagnol, J., Zolésio, J.-P.: Shape derivative in the wave equation with Dirichlet boundary conditions. J. Differential Equations 158(2), 175–210 (1999)

[2] Correa, R., Seeger, A.: Directional derivative of an minmax function. Nonlinear Anal. 9, 13–22 (1985)

[3] Cuer, M., Zolésio, J.-P.: Control of singular problem via differentiation of a min-max. Systems & Control Letters 11(2), 151–158 (1988)

[4] Delfour, M., Zolésio, J.-P.: Shapes and Geometries. Analysis, Differential Calculus, and Optimization. SIAM Advances in Design and Control (2001)

[5] Delfour, M.C., Zolésio, J.-P.: Hidden boundary smoothness for some classes of differential equations on submanifolds, Optimization methods in partial differential equations. Contemp. Math., vol. 209, pp. 59–73. Amer. Math. Soc. Providence, RI (1997)

[6] Delfour, M.C., Zolésio, J.-P.: Curvatures and skeletons in shape optimization. Z. Angew. Math. Mech. 76(3), 198–203 (1996)

[7] Desaint, F.R., Zolésio, J.-P.: Manifold derivative in the Laplace-Beltrami equation. J. Funct. Anal. 151(1), 234–269 (1997)

[8] Lasiecka, I., Triggiani, R.: Control theory for partial differential equation Encyclopedia of mathematics. Cambridge University Press (2000)

[9] Sokolowski, J., Zolésio, J.-P.: Introduction to shape optimization. Shape sensitivity analysis. Springer Ser. Comput. Math., vol. 16. Springer, Berlin (1992)

[10] Zolésio, J.-P.: The material derivative (or speed) method for shape optimization. In: Haug, E.J., Céa, J. (eds.) Optimization of Distributed Parameter Structures (Iowa City, Iowa, 1980). NATO Adv. Sci. Inst. Ser. E: Appl. Sci., vol. II, 50, pp. 1089–1151. Sijhofff and Nordhoff, Alphen aan den Rijn, Nijhoff, The Hague (1981)

[11] Zolésio, J.-P., Goatin, P.: N Dimensional Crowd Motion (in this book)

[12] Zolésio, J.-P.: Identification de Domaine, These de doctorat d'état, Nice, France (1979)

[13] Zolésio, J.-P.: Hidden boundary shape derivative for the solution to Maxwell equations and non cylindrical wave equations. In: Optimal Control of Coupled Systems of Partial Differential Equations. Internat. Ser. Numer. Math., vol. 158, pp. 319–345. Birkhuser Verlag, Basel (2009)

The Exact l_1 Penalty Function Method for Constrained Nonsmooth Invex Optimization Problems

Tadeusz Antczak

Faculty of Mathematics and Computer Science, University of Łódź
Banacha 22, 90-238 Łódź, Poland

Abstract. The exactness of the penalization for the exact l_1 penalty function method used for solving nonsmooth constrained optimization problems with both inequality and equality constraints is considered. Thus, the equivalence between the sets of optimal solutions in the nonsmooth constrained optimization problem and its associated penalized optimization problem with the exact l_1 penalty function is established under locally Lipschitz invexity assumptions imposed on the involved functions.

Keywords: exact l_1 penalty function method, absolute value penalty function, penalized optimization problem, locally Lipschitz invex function, Generalized Karush-Kuhn-Tucker optimality conditions.

1 Introduction

Considerable attention has been given in recent years to devising methods for solving nonlinear programming problems via unconstrained minimization techniques. One class of methods which has emerged as very promising is the class of exact penalty function methods. Methods using exact penalty function transform a constrained extremum problem into a single unconstrained optimization problem. The constraints are placed into the objective function via a penalty parameter c in a way that penalizes any violation of the constraints.

One important property that distinguishes exact penalty functions is the exactness of the penalization. The concept of exact penalization is sometimes ambiguous, or at least varies from author to author. One of the definitions of the exactness of the penalization is the following: there is an appropriate penalty parameter choice such that a single unconstrained minimization of the penalty function yields a solution of the constrained optimization problem.

Nondifferentiable exact penalty functions were introduced for the first time by Eremin [6] and Zangwill [17]. In almost all of the introduced penalized approaches the notion of convexity plays a dominant role. In 1970, Luenberger [13] showed that, under convex assumptions, there is a lower bound for a penalty parameter c, equal to the largest Lagrange multiplier in absolute value, associated to one of the constraints of the nonlinear constrained optimization problem.

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 461–470, 2013.

Later, Charalambous [3] generalized the result of Luenberger for the absolute value penalty function, assuming the second-order sufficient conditions. Under the assumptions that the minimization problem is solvable and that it satisfies the relaxed Slater constraint qualification, Mangasarian [14] characterized solutions of the convex optimization problem in terms of minimizers of the exact penalty function for a single value of the penalty parameter exceeding some treshold. Bazaraa et al. [2] also used the exact l_1 penalty function method to solve nonlinear convex optimization problems with both inequality and equality constraints. They assumed that the objective function and the inequality constraints are convex and the equality constraints are affine functions to prove that a Karush-Kuhn-Tucker point in the original optimization problem is a minimizer of the exact l_1 penalty function in the associated penalized optimization problem with sufficiently large value of a penalty parameter. In the mentioned above works, the lower bound of the penalty parameter above which, for all penalty parameters, any optimal solution of the original nonlinear optimization problem is also a minimizer of the penalized problem has been given for differentiable optimization problems involving convex functions. However, from the practical point of view, the converse result is also important.

In recent years, some numerous generalizations of convex functions have been derived which proved to be useful for extending optimality conditions and some classical duality results, previously restricted to convex programs, to larger classes of nonconvex optimization problems. One of them is the invexity notion introduced by Hanson [10] for differentiable scalar functions and later generalized from different points of view, also in the case of nondifferentiable functions (see [1], [5], [8], [11], [12], [16], [18], and others).

Now, we show that there is the equivalence between the set of optimal solutions in a nondifferentiable nonconvex optimization problem and the set of minimizers in its associated exact penalized problem with the absolute value penalty function. It turns out that this property is not true only for (differentiable) convex optimization problems, but it still holds for nonlinear optimization problems involving locally Lipschitz invex functions with respect to the same function η (with the exception of those equality constraint functions for which the associated Lagrange multipliers are negative – these functions should be assumed to be incave with respect to the same function η). The result established here shows that there does exist a lower bound for a penalty parameter c, equal to the largest Lagrange multiplier in absolute value, associated to a Karush-Kuhn-Tucker point in the original nonlinear optimization problem, above which this equivalence holds. Further, in the case when at least one of the functions constituting the nondifferentiable constrained optimization problem is not locally Lipschitz invex and in the case when the objective function is coercive but not invex, then the equivalence in the sense discussed here might not hold between these optimization problems.

2 Preliminaries and Problem Formulation

Throughout this section, X is a nonempty subset of R^n. A real-valued function $f : X \to R$ is said to be locally Lipschitz on X if, for any $x \in X$, there exist a neighborhood U of x and a positive constant $K_x > 0$ such that, for every $y, z \in U$, it holds $|f(y) - f(z)| \leqq K_x \|y - z\|$. The Clarke generalized directional derivative [4] of a locally Lipschitz function $f : X \to R$ at $x \in X$ in the direction $v \in R^n$, denoted $f^0(x; v)$, is given by $f^0(x; v) = \limsup_{\substack{y \to x \\ \lambda \downarrow 0}} \frac{f(y+\lambda v) - f(y)}{\lambda}$.

Definition 1. *The Clarke generalized subgradient [4] of f at $x \in X$, denoted $\partial f(x)$, is defined by $\partial f(x) = \left\{\xi \in R^n : f^0(x; v) \geq \xi^T v \text{ for all } v \in R^n\right\}$.*

The following definition is a generalization of the definition of a class of differentiable convex functions to the case of a class of locally Lipschitz invex functions (see [10]).

Definition 2. *[10] Let a function $f : X \to R$ be a locally Lipschitz function on X and $u \in X$. If there exists a vector-valued function $\eta : X \times X \to R^n$ such that, for each $x \in X$, the inequality $f(x) - f(u) \geq \xi^T \eta(x, u)$ holds for any $\xi \in \partial f(u)$, then f is said to be a locally Lipschitz invex function at u on X with respect to η. If the inequality above is satisfied at any point u, then f is said to be a locally Lipschitz invex function on X with respect to η.*

In order to define an analogous class of Lipschitz incave functions with respect to η, the direction of the inequality in the definition of invex functions should be changed to the opposite one.

Definition 3. *[15] A continuous function $f : R^n \to R$ is said to be coercive if* $\lim_{\|x\| \to \infty} f(x) = \infty$.

Consider the following constrained optimization problem:

$$\begin{array}{c} \text{minimize } f(x) \\ \text{subject to } \quad g_i(x) \leq 0, \;\; i \in I = \{1, \ldots m\}, \\ h_j(x) = 0, \;\; j \in J = \{1, \ldots, s\}, \\ x \in X, \end{array} \qquad \text{(P)}$$

where $f : X \to R$ and $g_i : X \to R$, $i \in I$, $h_j : X \to R$, $j \in J$, are locally Lipschitz functions on a nonempty set $X \subset R^n$.

Let $D := \{x \in X : g_i(x) \leq 0, \;\; i \in I, \; h_j(x) = 0, \;\; j \in J\}$ be the set of all feasible solutions of problem (P). Further, we denote a set of active inequality constraints at point $\overline{x} \in X$ by $I(\overline{x}) = \{i \in I : g_i(\overline{x}) = 0\}$.

Theorem 4. *[4], [18] (Generalized Karush-Kuhn-Tucker necessary optimality conditions). Let $\overline{x} \in D$ be an optimal solution in problem (P) and some suitable constraint qualification be satisfied at $\overline{x}$. Then, there exist $\overline{\lambda} \in R^m$, $\overline{\mu} \in R^s$ such that*

$$0 \in \partial f(\overline{x}) + \sum_{i=1}^{m} \overline{\lambda}_i \partial g_i(\overline{x}) + \sum_{j=1}^{s} \overline{\mu}_j \partial h_j(\overline{x}), \qquad (1)$$

$$\overline{\lambda}_i g_i(\overline{x}) = 0,\ i \in J, \tag{2}$$

$$\overline{\lambda}_i \in R_+, \quad i \in J. \tag{3}$$

We will assume that a suitable constraint qualification is satisfied at any optimal point in problem (P).

Definition 5. *The point $\overline{x} \in D$ is said to be Karush-Kuhn-Tucker point (a KKT point, for short) if there exist the Lagrange multipliers $\overline{\lambda} \in R^m$, $\overline{\mu} \in R^s$ such that the conditions (1)-(3) are satisfied at $\overline{x}$.*

3 The Exactness of the Exact l_1 Penalty Function Method

The most popular nondifferentiable exact penalty function is the absolute value penalty function also called the exact l_1 penalty function. Its definition, for the considered optimization problem (P), is the following

$$\text{minimize } P(x,c) = f(x) + c\left[\sum_{i\in I} g_i^+(x) + \sum_{j\in J} |h_j(x)|\right], \quad (\mathrm{P}(c)) \tag{4}$$

where, for a given constraint $g_i(x) \leq 0$, the function g_i^+ is defined by

$$g_i^+(x) = \begin{cases} 0 & \text{if } g_i(x) \leq 0, \\ g_i(x) & \text{if } g_i(x) > 0. \end{cases} \tag{5}$$

The unconstrained optimization problem defined above, we call the penalized optimization problem with the absolute value penalty function.

It is known (see, for example, [2]) that under suitable convexity assumptions and a constraint qualification, there exists a finite value c that will recover an optimal solution in the constrained optimization problem (P) via the minimization of the exact penalty function being the objective function in the exact penalized optimization problem (P(c)). Now, we generalize this result by weakening the convexity assumption imposed on the functions constituting the considered nonsmooth optimization problem (P).

Theorem 6. *Let $\overline{x} \in D$ be a Karush-Kuhn-Tucker point in the constrained optimization problem (P), at which the Generalized Karush-Kuhn-Tucker conditions (1)-(3) are satisfied with the Lagrange multipliers $\overline{\lambda} \in R^m$ and $\overline{\mu} \in R^s$. Let $J^+(\overline{x}) = \{j \in J : \overline{\mu}_j > 0\}$ and $J^-(\overline{x}) = \{j \in J : \overline{\mu}_j < 0\}$. Furthermore, assume that the functions f, g_i, $i \in I$, h_j, $j \in J^+(\overline{x})$, are locally Lipschitz invex at $\overline{x}$ on X with respect to the same function η and the functions h_j, $j \in J^-(\overline{x})$, are locally Lipschitz incave at $\overline{x}$ on X with respect to the same function η. If c is assumed to be sufficiently large (it is sufficient to set $c \geq \max\{\overline{\lambda}_i,\ i \in I,\ |\overline{\mu}_j|,\ j \in J\}$, where $\overline{\lambda}_i$, $i = 1, ..., m$, $\overline{\mu}_j$, $j = 1, ..., s$, are the Lagrange multipliers associated to the constraints g_i and h_j, respectively), then $\overline{x}$ is also a minimizer of its penalized optimization problem (P(c)) with the absolute value penalty function.*

Proof. By assumption, $\overline{x}$ is a Karush-Kuhn-Tucker point in the constrained optimization problem (P), at which the Generalized Karush-Kuhn-Tucker conditions (1)-(3) are satisfied with the Lagrange multipliers $\overline{\lambda} \in R^m$ and $\overline{\mu} \in R^s$. Since $c \geq \max\left\{\overline{\lambda}_i,\ i \in I,\ \left|\overline{\mu}_j\right|,\ j \in J\right\}$, then, by definition of the objective function in the penalized optimization problem (P(c)), it follows that

$$P(x,c) = f(x) + c\sum_{i=1}^{m} g_i^+(x) + c\sum_{j=1}^{s} |h_j(x)| \geq f(x) + \sum_{i=1}^{m} \overline{\lambda}_i g_i^+(x) + \sum_{j=1}^{s} \left|\overline{\mu}_j h_j(x)\right|. \tag{6}$$

Thus, (5) gives

$$f(x) + \sum_{i=1}^{m} \overline{\lambda}_i g_i^+(x) + \sum_{j=1}^{s} \left|\overline{\mu}_j h_j(x)\right| \geq f(x) + \sum_{i=1}^{m} \overline{\lambda}_i g_i(x) + \sum_{j=1}^{s} \overline{\mu}_j h_j(x). \tag{7}$$

By assumption, the inequality constraints g_i, $i \in I$, and the equality constraints h_j, $j \in J^+(\overline{x})$, are locally Lipschitz invex at $\overline{x}$ on X and the equality constraints h_j, $j \in J^-(\overline{x})$, are locally Lipschitz incave at $\overline{x}$ on X. Hence, by the Generalized Karush-Kuhn-Tucker conditions (2) and (3) together with the feasibility of $\overline{x}$ in problem (P), it follows that the inequality

$$f(x) + \sum_{i=1}^{m} \overline{\lambda}_i g_i(x) + \sum_{j=1}^{s} \overline{\mu}_j h_j(x) \geq f(x) + \sum_{i=1}^{m} \overline{\lambda}_i \zeta_i^T \eta(x,\overline{x}) + \sum_{j=1}^{s} \overline{\mu}_j \gamma_j^T \eta(x,\overline{x}) \tag{8}$$

holds for any $\zeta_i \in \partial g_i(\overline{x})$, $i = 1,...,m$, and for any $\gamma_j \in \partial h_j(\overline{x})$, $j = 1,...,s$. Then, using the Generalized Karush-Kuhn-Tucker condition (2) , we get

$$\begin{aligned} f(x) + \textstyle\sum_{i=1}^{m} \overline{\lambda}_i \left[g_i(\overline{x}) + \zeta_i^T \eta(x,\overline{x})\right] + \sum_{j=1}^{s} \overline{\mu}_j \left[h_j(\overline{x}) + \gamma_j^T \eta(x,\overline{x})\right] \\ = f(x) + \textstyle\sum_{i=1}^{m} \overline{\lambda}_i \zeta_i^T \eta(x,\overline{x}) + \sum_{j=1}^{s} \overline{\mu}_j \gamma_j^T \eta(x,\overline{x}). \end{aligned} \tag{9}$$

Thus, by the Generalized Karush-Kuhn-Tucker necessary optimality condition (1), it follows that

$$f(x) + \sum_{i=1}^{m} \overline{\lambda}_i \zeta_i^T \eta(x,\overline{x}) + \sum_{j=1}^{s} \overline{\mu}_j \gamma_j^T \eta(x,\overline{x}) = f(x) - \xi^T \eta(x,\overline{x}), \tag{10}$$

where $\xi \in \partial f(\overline{x})$. By assumption, f is locally Lipschitz invex at $\overline{x}$ on X also with respect to the function η. Using Definition 2 together with the feasibility of $\overline{x}$ in problem (P), we get

$$f(x) - \xi^T \eta(x,\overline{x}) \geq f(\overline{x}) = f(\overline{x}) + c\sum_{i=1}^{m} g_i^+(\overline{x}) + c\sum_{j=1}^{s} |h_j(\overline{x})| = P(\overline{x},c). \tag{11}$$

Then, by (6)-(11), we conclude that the inequality $P(x,c) \geq P(\overline{x},c)$ holds for all $x \in X$. This means that $\overline{x}$ is a minimizer of the penalized optimization problem (P(c)) with the absolute value penalty function and the proof of theorem is complete. ■

Corollary 7. *Let $\overline{x}$ be an optimal point in the considered optimization problem (P). Furthermore, assume that all hypotheses of Theorem 6 are fulfilled. Then $\overline{x}$ is also a minimizer in the penalized optimization problem (P(c)) with the absolute value penalty function.*

Theorem 8. *Let the point $\overline{x}$ be a minimizer of the penalized optimization problem (P(c)) with the absolute value penalty function. Furthermore, assume that the functions f, g_i, $i \in I$, h_j, $j \in J^+(\widetilde{x})$, are locally Lipschitz invex at $\widetilde{x}$ on X with respect to the same function η, and the functions h_j, $j \in J^-(\widetilde{x})$, are locally Lipschitz incave at $\widetilde{x}$ on X with respect to the same function η, where $\widetilde{x}$ is any Karush-Kuhn-Tucker point in problem (P), at which the Karush-Kuhn-Tucker necessary optimality conditions (1)-(3) are satisfied with the Lagrange multipliers $\widetilde{\lambda} \in R^m$ and $\widetilde{\mu} \in R^s$. If the set of all feasible solutions in the constrained optimization problem (P) is compact and the penalty parameter c is sufficiently large (it is sufficient if c satisfies the following condition $c > \max\left\{\widetilde{\lambda}_i,\ i \in I,\ |\widetilde{\mu}_j|,\ j \in J\right\}$), then $\overline{x}$ is also optimal in problem (P).*

Proof. We assume that $\overline{x}$ is a minimizer in the penalized optimization problem (P(c)) with the absolute value penalty function. Then, by the definition of the penalized optimization problem (P(c)) and (5), the following inequalities $f(x) + c\left(\sum_{i=1}^m g_i^+(x) + \sum_{j=1}^s |h_j(x)|\right) \geq f(\overline{x}) + c\left(\sum_{i=1}^m g_i^+(\overline{x}) + \sum_{j=1}^s |h_j(\overline{x})|\right) \geq f(\overline{x})$ hold for all $x \in X$. Thus, for all $x \in D$, the following inequality

$$f(x) \geq f(\overline{x}) \tag{12}$$

holds. The inequality above means that values of the function f are bounded below on the set D of all feasible solutions in the constrained optimization problem (P). Since f is a continuous function bounded below on the compact set D, therefore, by Weierstrass' theorem, f admits its minimum $\widetilde{x}$ on D.

Now, we prove that $\overline{x}$ is also optimal in the considered optimization problem (P). First, we show that $\overline{x}$ is feasible in problem (P). By means of contradiction, suppose that $\overline{x}$ is not feasible in problem (P). As we have established above, the given constrained optimization problem (P) has an optimal solution $\widetilde{x}$. Since a constraint qualification is satisfied at $\widetilde{x}$, then there exist the Lagrange multipliers $\widetilde{\lambda} \in R^m$ and $\widetilde{\mu} \in R^s$ such that the Generalized Karush-Kuhn-Tucker necessary optimality conditions (1)-(3) are satisfied at $\widetilde{x}$. By assumption, the functions f, g_i, $i \in I$, h_j, $j \in J^+(\widetilde{x})$, are invex at $\widetilde{x}$ on X with respect to the same function η and the functions h_j, $j \in J^-(\widetilde{x})$, are incave at $\widetilde{x}$ on X with respect to the same function η. Therefore, by Definition 2, respectively, it follows that the inequalities

$$f(\overline{x}) - f(\widetilde{x}) \geq \xi^T \eta(\overline{x}, \widetilde{x}), \tag{13}$$

$$g_i(\overline{x}) - g_i(\widetilde{x}) \geq \zeta_i^T \eta(\overline{x}, \widetilde{x}), \quad i \in I, \tag{14}$$

$$h_j(\overline{x}) - h_j(\widetilde{x}) \geq \gamma_j^T \eta(\overline{x}, \widetilde{x}), \quad j \in J^+(\widetilde{x}), \tag{15}$$

$$h_j(\overline{x}) - h_j(\widetilde{x}) \leq \gamma_j^T \eta\left(\overline{x}, \widetilde{x}\right), \quad j \in J^-(\widetilde{x}) \tag{16}$$

hold for each $\xi \in \partial f\left(\widetilde{x}\right)$, $\zeta_i \in \partial g_i\left(\widetilde{x}\right)$, $i = 1, ..., m$, and $\gamma_j \in \partial h_j\left(\widetilde{x}\right)$, $j = 1, ..., s$. Multiplying (14), (15) and (16) by the associated Lagrange multiplier and then adding both sides of the obtained inequalities and both sides of (13), we get

$$\begin{aligned} f\left(\overline{x}\right) - f\left(\widetilde{x}\right) + \sum_{i=1}^m \widetilde{\lambda}_i g_i(\overline{x}) - \sum_{i=1}^m \widetilde{\lambda}_i g_i(\widetilde{x}) + \sum_{j=1}^s \widetilde{\mu}_j h_j(\overline{x}) - \sum_{j=1}^s \widetilde{\mu}_j h_j(\widetilde{x}) \\ \geq \left[\xi^T + \sum_{i=1}^m \widetilde{\lambda}_i \zeta_i^T + \sum_{j=1}^s \widetilde{\mu}_j \gamma_j^T\right] \eta\left(\overline{x}, \widetilde{x}\right). \end{aligned}$$

Using (5) with the Karush-Kuhn-Tucker necessary optimality conditions (11), (12) and the feasibility of $\widetilde{x}$ in problem (P), we get

$$f\left(\overline{x}\right) + \sum_{i=1}^m \widetilde{\lambda}_i g_i^+(\overline{x}) + \sum_{j=1}^s \widetilde{\mu}_j \left|h_j(\overline{x})\right| \geq f\left(\widetilde{x}\right). \tag{17}$$

By assumption, the penalty parameter c is sufficiently large (it is sufficient that $c > \max\left\{\widetilde{\lambda}_i,\ i \in I,\ \left|\widetilde{\mu}_j\right|,\ j \in J\right\}$. Since $\overline{x}$ is assumed to be not feasible in the given optimization problem (P), therefore, at least one of $g_i^+(\overline{x})$ and $|h_j(\overline{x})|$ must be nonzero. Therefore, (17) yields

$$f\left(\overline{x}\right) + c\left[\sum_{i=1}^m g_i^+(\overline{x}) + \sum_{j=1}^s \left|h_j(\overline{x})\right|\right] > f\left(\widetilde{x}\right). \tag{18}$$

Then, by $\widetilde{x} \in D$ and (2), we get

$$f\left(\overline{x}\right) + c\left[\sum_{i=1}^m g_i^+(\overline{x}) + \sum_{j=1}^s \left|h_j(\overline{x})\right|\right] > f\left(\widetilde{x}\right) + c\left[\sum_{i=1}^m g_i^+(\widetilde{x}) + \sum_{j=1}^s \left|h_j(\widetilde{x})\right|\right].$$

Then, by the definition of the exact l_1 penalty function (see (4)), it follows that the following inequality $P\left(\overline{x}, c\right) > P\left(\widetilde{x}, c\right)$ holds, which is a contradiction to the assumption that $\overline{x}$ is a minimizer in the penalized optimization problem (P(c)) with the absolute value penalty function. Thus, we have proved that $\overline{x}$ is feasible in the given constrained optimization problem (P). Hence, the optimality of $\overline{x}$ in problem (P) follows directly from (12). ■

Corollary 9. *Let the hypotheses of Corollary 7 and Theorem 8 are fulfilled. Then, the set of optimal solutions in the considered extremum problem (P) and the set of minimizers in its associated exact penalized optimization problem (P(c)) with the absolute value penalty function coincide.*

Example 10. Consider the following nonsmooth optimization problem

$$\begin{aligned} f(x) &= \arctan\left(|x|\right) \to \min \\ g(x) &= \tfrac{1}{2}\left(e^{|x|-x} - 1\right) \leq 0. \end{aligned} \tag{P1}$$

Note that $D = \{x \in R : x \geq 0\}$ and $\overline{x} = 0$ is an optimal solution in the considered nonsmooth optimization problem (P1). Since we use the exact l_1 penalty method for solving problem (P1), then we construct the following unconstrained optimization problem

$$P(x,c) = \arctan(|x|) + c\max\left\{0, \frac{1}{2}\left(e^{|x|-x} - 1\right)\right\} \to \min. \quad (\mathrm{P1}(c))$$

Note that $\overline{x} = 0$ is feasible in problem (P1) and the Generalized Karush-Kuhn-Tucker necessary optimality conditions (1)-(3) are fulfilled at $\overline{x}$ with the Lagrange multiplier $\overline{\lambda}$ satisfying the following condition: $0 \in \partial f(\overline{x}) + \overline{\lambda}\partial g(\overline{x})$, where $\partial f(\overline{x}) = [-1,1]$ and $\partial g(\overline{x}) = [-1,0]$. Further, it can be established by Definition 2 that the objective function f and the constraint function g are locally Lipschitz invex at $\overline{x}$ on R with respect to the same function η defined by $\eta(x,\overline{x}) = \frac{1}{2}(\arctan(|x|) - \arctan(|\overline{x}|))$. Then, by Theorems 6 and 8, it follows that, for any penalty parameter c satisfying $c > \overline{\lambda}$, there is the equivalence between the sets of optimal solutions in optimization problems (P1) and (P1(c)). Further, note that not all functions involved in problem (P1) are differentiable and convex. Therefore, in order to show that the point $\overline{x} = 0$, being optimal in (P1), is also a minimizer in the unconstrained optimization problem (P1(c)), we can not use the conditions for convex smooth optimization problems (see, for instance, Theorem 9.3.1 [2]).

Example 11. *Consider the following nonsmooth constrained optimization problem*

$$\min\ f(x) = \begin{cases} -x+4 & \text{if} \quad x < -4, \\ \frac{1}{2}x + 10 & \text{if} -4 \leq x < 0, \\ -5x + 10 & \text{if}\ \ 0 \leq x < 2, \\ x - 2 & \text{if} \quad x \geq 2, \end{cases} \qquad (P2)$$

$$g(x) = x - \tfrac{1}{4} \leq 0,$$

in which not all functions are locally Lipschitz invex. Note that $D = \left\{x \in R : x \leq \frac{1}{4}\right\}$ *and* $\overline{x} = -4$ *is an optimal solution in the considered optimization problem (P2). Since* $0 \in \partial f(0) = \left[-5, \frac{1}{2}\right]$, *then* $\widetilde{x} = 0$ *is a stationary point of* f. *It is not difficult to show that* $\widetilde{x}$ *is not a global minimizer of* f. *Then the objective function* f *is not locally Lipschitz invex on* R *with respect to any function* η *defined by* $\eta : R \times R \to R$ *(see, for example, [16]). However, we use the exact* l_1 *penalty method to solve the considered optimization problem (P2). Therefore, we construct the following unconstrained optimization problem*

$$P(x,c) = f(x) + c\max\left\{0, x - \frac{1}{4}\right\} \to \min \qquad (P2(c))$$

Note that $\overline{x} = -4$, *being an optimal solution in problem (P2), is not a global minimizer in the associated penalized optimization problem (P2(c)) for all values of the penalty parameter* c *satisfying the condition* $c > \overline{\lambda} = 0$, *(where* $\overline{\lambda}$ *is the Lagrange multiplier associated to the inequality constraint* g *satisfying the*

Karush-Kuhn-Tucker necessary optimality conditions (1)-(3)). However, for every penalty parameter $c \in (0, \frac{32}{7})$, *the point* $\widehat{x} = 2$ *is a global minimizer in the above penalized optimization problem (P2(c)). Therefore, there is no the equivalence between the sets of optimal solutions in problems (P2) and (P2(c)) for any penalty parameter c satisfying the condition* $c > \overline{\lambda}$. *This follows from the fact that not all functions constituting the considered optimization problem (P2) are locally Lipschitz invex on R.*

Remark 12. *Peressini et al. [15] considered differentiable convex optimization problems and solved them by using the exact* l_1 *penalty function method. Under assumption that the objective function in the constrained optimization problem is coercive (see Definition 3), they proved that, for sufficiently large values of the penalty parameter c, the constrained optimal solution in (P) is also a minimizer in its associated penalized optimization problem (P(c)) with the exact* l_1 *penalty function. But the finite value of the penalty parameter c, above which this result holds, was not given in [15]. Note that the objective function in the optimization problem (P2) considered in Example 11 is coercive. However, for not all values of the penalty parameter c satisfying the condition* $c > \overline{\lambda}$, *an optimal solution in the considered optimization problem (P2) yields a minimizer in its associated penalized optimization problem (P(c)) with the exact* l_1 *penalty function. But the result proved in the paper shows that, under invexity assumptions imposed on the functions constituting the constrained nonsmooth optimization problem (P), for every value of the penalty parameter c satisfying the condition* $c > \max\left\{\overline{\lambda}_i,\ i \in I,\ \left|\overline{\mu}_j\right|,\ j \in J\right\}$, *the sets of optimal solutions in problems (P) and (P(c)) coincide. Hence, this example shows that in the case when the objective function is coercive but not invex the result established in the paper might not be true for such optimization problems.*

References

[1] Antczak, T.: Lipschitz r-invex functions and nonsmooth programming. Numerical Functional Analysis and Optimization 23, 265–283 (2002)

[2] Bazaraa, M.S., Sherali, H.D., Shetty, C.M.: Nonlinear programming: theory and algorithms. John Wiley and Sons, New York (1991)

[3] Charalambous, C.: A lower bound for the controlling parameters of the exact penalty functions. Mathematical Programming 15, 278–290 (1978)

[4] Clarke, F.H.: Optimization and nonsmooth analysis. A Wiley-Interscience Publication, John Wiley&Sons, Inc. (1983)

[5] Craven, B.D.: Invex functions and constrained local minima. Bulletin of the Australian Mathematical Society 24, 357–366 (1981)

[6] Eremin, I.I.: The penalty method in convex programming. Doklady Akad. Nauk SSSR 143, 748–751 (1967)

[7] Di Pillo, G., Grippo, L.: Exact penalty functions in constrained optimization. SIAM Journal of Control and Optimization 27, 1333–1360 (1989)

[8] Egudo, R.R., Hanson, M.A.: On sufficiency of Kuhn-Tucker conditions in nonsmooth multiobjective programming, FSU Report No. M-888 (1993)

[9] Han, S.P., Mangasarian, O.L.: Exact penalty functions in nonlinear programming. Mathematical Programming 17, 251–269 (1979)
[10] Hanson, M.A.: On sufficiency of the Kuhn-Tucker conditions. Journal of Mathematical Analysis and Applications 80, 545–550 (1981)
[11] Kaul, R.N., Suneja, S.K., Lalitha, C.S.: Generalized nonsmooth invexity. Journal of Information and Optimization Sciences 15, 1–17 (1994)
[12] Kim, M.H., Lee, G.M.: On duality theorems for nonsmooth Lipschitz optimization problems. Journal of Optimization Theory and Applications 110, 669–675 (2001)
[13] Luenberger, D.: Control problem with kinds. IEEE Transaction on Automatic Control 15, 570–574 (1970)
[14] Mangasarian, O.L.: Sufficiency of exact penalty minimization. SIAM Journal of Control and Optimization 23, 30–37 (1985)
[15] Peressini, A.L., Sullivan, F.E., Uhl, Jr., J.J.: The mathematics of nonlinear programming. Springer-Verlag New York Inc. (1988)
[16] Reiland, T.W.: Nonsmooth invexity. Bulletin of the Australian Mathematical Society 42, 437–446 (1990)
[17] Zangwill, W.I.: Nonlinear programming via penalty functions. Management Science 13, 344–358 (1967)
[18] Zhao, F.: On sufficiency of the Kuhn-Tucker conditions in nondifferentiable programming. Bulletin of the Australian Mathematical Society 46, 385–389 (1992)

The Minimum Energy Building Temperature Control

Marek Długosz

AGH University of Science and Technology
Faculty of Electrical Engineering, Automatics, Computer Science and Electronics
Department of Automatics
Al. A. Mickiewicza 30, 30-059 Krakow, Poland
mdlugosz@agh.edu.pl
http://www.agh.edu.pl

Abstract. One of the most important factors of users comfort inside building is air temperature. From the other side one of the most biggest position in home budget is price for heat. Mutually exclusive indices are the cause that the control of temperature task using the smallest amount of energy as it is possible is very difficult. In this paper is presented simple model of temperature changes inside building base on lumped capacity method. Using this method finally obtains mathematical model of temperature changes which model is equivalent in structure to electrical RC-network. The model is composed of linear differential equations. Based on this mathematical model the simple algorithm controlling of temperature inside room is proposed. In this article are also included numerical simulations of the proposed solutions.

Keywords: building temperature model, control, optimization.

1 Introduction

Problem of optimal use of the heat energy for heating residential building is still a current problem. Some of the main reasons for this are: still rising energy price (electricity, gas, coal), still rising power consumption by household or environmental pollution. The goal of this article is to present a control system which stabilises the temperature inside the building with the using minimum amount of energy.

The air temperature inside the building T_i depends on many factors. Some of them like: solar radiation, wind, heating system, light, people, air ventilation are showed on figure 1. Some of these factors are unpredicted like: people inside, light, air ventilation. Some of them are periodical and can be measured or predicted, for example: solar radiation, temperature outside, wind direction and force. The physical phenomena of thermal conductivity are also very complex and described by partial differential equations which depend on time and spatial variables. For those reasons one and general thermal model of the building does not exist. On the other hand for searching optimal controls the mathematical model of the

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 471–480, 2013.

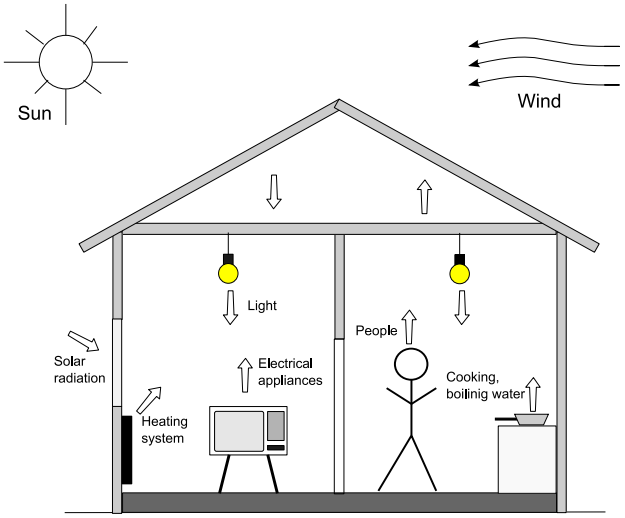


Fig. 1. The temperature inside the building depends on many factors such as: heating system, external air temperature, wind, solar heat, casual heat gains, structure of the building

system is necessary. In the literature can be found three main methods to obtain and identify an approximated thermal model of the building.

The first method: the impulse response factor method [1,8] is based on the response of the model if the excitation is a unit impulse. Making some additional assumptions and using the properties of Laplace transform the response of the wall to this excitation function can be expressed as time series.

The second method is the finite difference method. This is a numerical method for solving partial differential equation of the heat conduction [1,8]. The finite difference method is based on approximation derivates by algebraic equation. The building wall is divided into a finite number of layers and temperature for each layers is computed using set of the algebraic equations.

The third method: the lumped parameter method (or other name the lumped capacitance method) base on assumptions that transfer of the heat flux between two spaces which are divided by partition (wall) can be modeled by the equivalent electrical RC circuit [4,5,8]. The parameters of the electrical RC circuit like resistances are interpreted as thermal resistances, capacities are interpreted as heat capacities of the modeled elements. The physical properties of the construction elements of the building are represented by resistors and capacitors. The lumped parameter method describes changes of the air temperatures or the temperatures of the construction elements in one point. Finally, the mathematical model which is obtained by using the lumped parameter method has the form of linear differential equations. This model can be easily solved by analytical or numerical methods. In this paper, the lumped parameter method was chosen for modeling changes of the indoor air temperature of building.

The plan of the article is as follow, the first section contains short description of the lumped parameter method (LPM), next section contains description of the *LQR* controller. The last section presents some of experimental results. At the end of the article are contained conclusions and plans for the future works.

2 Thermal Modeling Methodologies

The most suitable form of the mathematical model of dynamic system for searching of optimal control solutions, is form of the linear differential equations.

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t) + \boldsymbol{Z}\boldsymbol{z}(t) \tag{1}$$

$$\boldsymbol{y}(t) = \boldsymbol{C}\boldsymbol{x}(t) \tag{2}$$

where $\boldsymbol{A}_{n\times n}$ – state space matrix, $\boldsymbol{B}_{n\times r}$ – control matrix, $\boldsymbol{Z}_{n\times k}$ – noise matrix, $\boldsymbol{C}_{m\times n}$ – output matrix, $\boldsymbol{x}(t) \in X = \mathbb{R}^n$ – state space vector, $\boldsymbol{z}(t) \in Z = \mathbb{R}^k$ – noise vector, $\boldsymbol{u}(t) \in U = \mathbb{R}^r$ – control vector, $\boldsymbol{y}(t) \in Y = \mathbb{R}^m$ – output vector. In all simulations the noise matrix $\boldsymbol{Z}$ was assumed to zero.

2.1 LPM

The assumptions of the lumped parameters method is that the temperature of the solid is spatially uniform at any instant during the transport of the heat process [6]. The result of this assumption is that the heat flow between two spaces which are separated by partition can be replaced by an equivalent RC electrical circuit [4,5,8]. The lumped parameter method describes changes of the temperature in one point so it is only an approximation of the real temperature. These simplifications allow us use the linear differential equations instead of more complicated partial differential equations. The lumped parameters method can be used for materials for which the conductivity in the middle is larger than the conductivity on the material surface [2,6].

As was said, the heat flow between two spaces which are separated by the partition can be replaced by the electrical circuit and figure 2 shows this. The meaning parameters are: the R_{out} and R_{int} thermal resistances of area outer and inner, C_{total} thermal capacity of the partition. The equation of the heat conduction based on the first-order model is:

$$C_{total}\frac{\mathrm{d}T}{\mathrm{d}t} = \frac{(T_o - T)}{R_{out}} + \frac{(T_i - T)}{R_{in}} + q \tag{3}$$

where q represents the other heat sources, T is uniform material temperature, T_o is outer air temperature and T_i is inner air temperature. The wall on figure 2 consists one of the uniform material but in the real wall may be build more than one of layers of the uniform materials. In this case, we can extend the model by adding the next equations for each uniform layer [5,8]. Also, if is needed the more accurate mathematical model we may add the next equations to the model [3,5]. All these operations finally increase total number of the equations and order of the model.

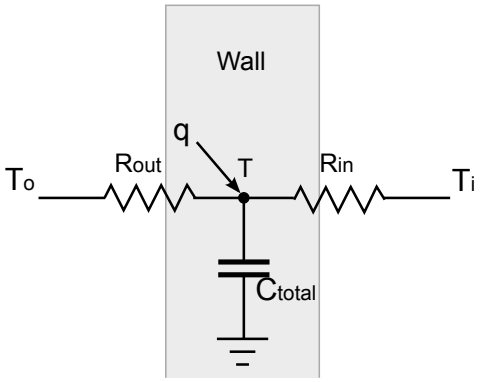


Fig. 2. Representation of the lumped parameter construction element

2.2 The Simple Thermal Model of Building

The state equations from (4) to (9) describe the thermal behaviour of more complex space [4]. The space inside which temperature is modeled contains two external walls , two partitions, floor and ceiling. In this case changes of the temperature of the indoor air depend on much more factors.

$$C_1\dot{T}_1 = U_1(T_i - T_1) + U_2(T_0 - T_1) \tag{4}$$
$$C_2\dot{T}_2 = U_3(T_i - T_2) + U_4(T_0 - T_2) \tag{5}$$
$$C_3\dot{T}_3 = U_5(T_i - T_3) + U_6(T_{z1} - T_3) + Q_s \tag{6}$$
$$C_4\dot{T}_4 = U_6(T_i - T_4) + U_5(T_{z2} - T_4) \tag{7}$$
$$C_5\dot{T}_5 = U_7(T_i - T_5) + U_7(T_{z3} - T_5) \tag{8}$$
$$\begin{aligned} C_6\dot{T}_i = {} & U_1(T_1 - T_i) + U_3(T_2 - T_i) + U_5(T_3 - T_i) \\ & + U_6(T_4 - T_i) + U_7(T_5 - T_i) + U_8(T_o - T_i) \\ & + Q_p + Q_g \end{aligned} \tag{9}$$

The parameters are: T_1 and T_2 – temperature of the building structure, T_3 and T_4 – temperature of the floor and ceiling, T_5 – temperature of the partitions, T_o – outdoor air temperature and T_i – indoor air temperature. The figure 3 shows electrical circuit RC which is equivalent with the building thermal model. The electrical parameters of this circuit correspond with physical parameters of the building. The resistances are equivalent to overall thermal transmittance and capacities are equivalent to thermal capacity. As is shown on the figure 3 electrical circuit has the form of the RC ladder network. The analytical solutions of the model's equations can be easily found and analysed and this is big advantage of this type of the model.

In next simulations values of the parameters of the building like C_i and R_i are the same as those adopted in the article [4].

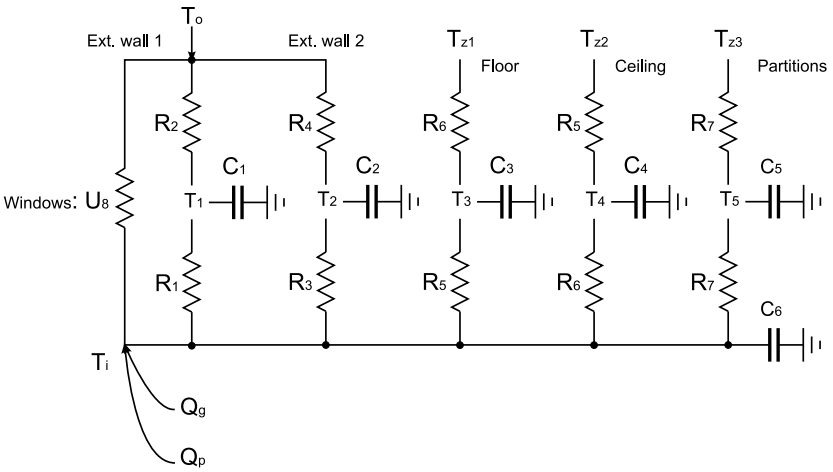


Fig. 3. The equivalent electrical circuit RC for the building thermal model given by equations number from (4) to (9)

3 Temperature Control in the Building

The typical idealised behaviour of indoor air temperature is shown on figure 4. As we can see the three different phases can be highlighted. The first phase, when indoor air temperature should reach the reference value in given time. The second phase when the air temperature should be stabilised on the specified level. The third phase when the air temperature do not need to be stabilised or controlled. This is idealised behaviour of indoor air temperature but generally all more complicated schemas of the temperature changes can be described by using those three phases.

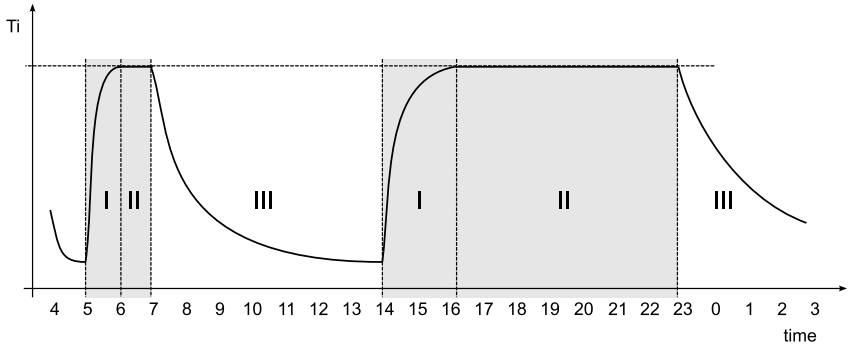


Fig. 4. Typical idealised changes of indoor air temperature

Only in the first and second phases is required active control of indoor air temperature. The main goal of the control system is to control indoor temperature in the first and second phases but in the first phase, time of control is also limited. In this paper is proposed to use two different controllers. The finite-horizon LQR controller which works in the first phase and the infinite-horizon LQR controller which works in the second phase. The next two subsections are describe shortly those controllers and are present their advantages and disadvantages.

3.1 The Finite-Horizon *LQR* Controller

The finite-horizon LQR controller minimizes the cost function (10) [7]:

$$J(\boldsymbol{x}, \boldsymbol{u}) = \frac{1}{2} \int_0^{t_k} \left(\boldsymbol{x}(t)^T \boldsymbol{W} \boldsymbol{x}(t) + \boldsymbol{u}(t)^T \boldsymbol{R} \boldsymbol{u}(t) \right) \mathrm{d}t + \frac{1}{2} \boldsymbol{x}(t_k)^T \boldsymbol{F} \boldsymbol{x}(t_k). \tag{10}$$

The matrices $\boldsymbol{W}$, $\boldsymbol{F}$, $\boldsymbol{R}$ are weight matrices and that matrices must be nonnegative and symmetric and $\boldsymbol{W} = \boldsymbol{W}^T \geq 0$, $\boldsymbol{F} = \boldsymbol{F}^T \geq 0$ $\boldsymbol{R} = \boldsymbol{R}^T > 0$, the pair of matrices $(\boldsymbol{A}, \boldsymbol{B})$ is stabilisabe , the pair of matrices $(\boldsymbol{W}, \boldsymbol{A})$ is detectable, t_k – is the control time. The control law is given by equation (11) [7]:

$$\boldsymbol{u}(t) = -\boldsymbol{R}^{-1} \boldsymbol{B}^T \boldsymbol{K}(t) \boldsymbol{x}(t) \tag{11}$$

where matrix $\boldsymbol{K}$ is unique, symmetric and nonnegative solution of Riccati differential equation (12) [7]:

$$\dot{\boldsymbol{K}}(t) = \boldsymbol{K}(t) \boldsymbol{B} \boldsymbol{R}^{-1} \boldsymbol{B}^T \boldsymbol{K}(t) - \boldsymbol{A}^T \boldsymbol{K}(t) - \boldsymbol{K}(t) \boldsymbol{A} - \boldsymbol{W}. \tag{12}$$

The controller (11) is nonstationary because the values of matrix $\boldsymbol{K}$ depend on time.

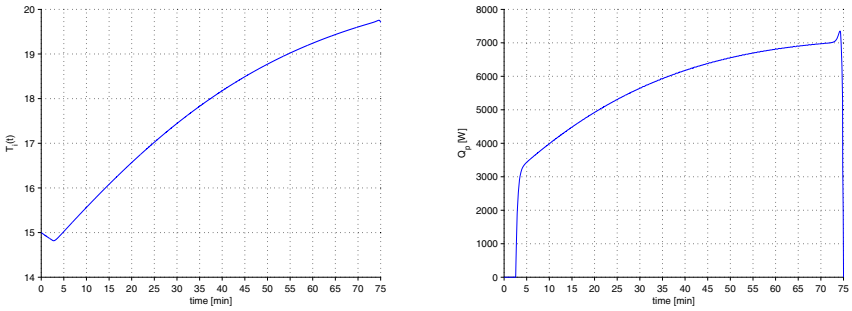
Simulation. The fig. 5 shows the result of simulation of the control system with the finite-horizon LQR controller. The control task was to increase the temperature value from 15 to 20 degrees in a finite time (75 minutes). The first plot shows the change of the indoor air temperature T_i (9), the second graph shows the change of the control signal $\boldsymbol{u}(t)$ (11). ■

The finite-horizon LQR controller is complicated in practical applications. First of all this is the nonstationary controller because the gain matrix $\boldsymbol{K}(t)$ is depends on time. In order to compute matrix $\boldsymbol{K}(t)$ the Riccati differential equation (12) must be solved. The same result, raise the value of indoor temperature from one level to other level in finite time, can be obtained using the infinite-horizon LQR controller with appropriate chose of the weight matrices, $\boldsymbol{W}$ and $\boldsymbol{K}$.

3.2 The Infinite-Horizon *LQR* Controller

The infinite-horizon LQR controller minimizes a cost function (13) [7]:

$$J(\boldsymbol{x}, \boldsymbol{u}) = \int_0^{\infty} \left(\boldsymbol{x}(t)^T \boldsymbol{W} \boldsymbol{x}(t) + \boldsymbol{u}(t)^T \boldsymbol{R} \boldsymbol{u}(t) \right) \mathrm{d}t \tag{13}$$



(a) (b)

Fig. 5. (a): Changes of the indoor air temperature T_i (9) and (b): the control signal $\boldsymbol{u}(t) = Q_p(t)$ (11)

where $\boldsymbol{W} = \boldsymbol{W}^T \geq 0$, $\boldsymbol{R} = \boldsymbol{R}^T > 0$, the pair of matrices $(\boldsymbol{A}, \boldsymbol{B})$ is stabilisable, the pair of matrices $(\boldsymbol{W}, \boldsymbol{A})$ is detectable. The cost function (13) contains two parts: a part which is joined with state space vector $\boldsymbol{x}(t)$ and part which is joined with the control vector $\boldsymbol{u}(t)$. The matrices $\boldsymbol{W}$ and $\boldsymbol{R}$ are called the weight matrices and appropriate selection of their values determines which part of cost function is better stabilised. The infinite-horizon LQR controller is proportional controller and the control law is given by equation (14) [7]:

$$\boldsymbol{u}(t) = -\boldsymbol{R}^{-1}\boldsymbol{B}^T\boldsymbol{K}\boldsymbol{x}(t) \tag{14}$$

where $\boldsymbol{K}$ is unique, symetric, nonnegative solution of algebraic Riccati equation:

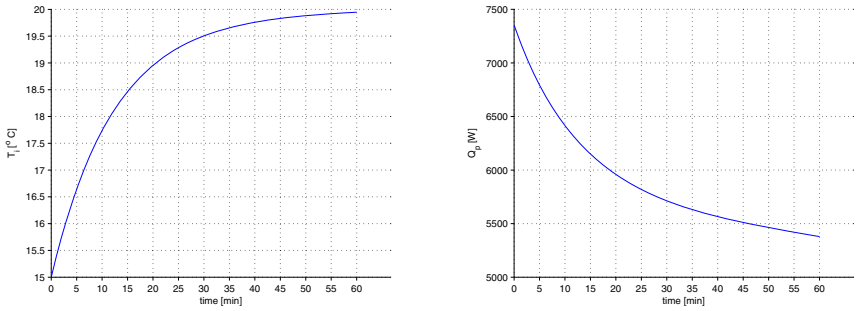
$$\boldsymbol{K}\boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{B}^T\boldsymbol{K} - \boldsymbol{A}^T\boldsymbol{K} - \boldsymbol{K}\boldsymbol{A} - \boldsymbol{W} = 0. \tag{15}$$

Simulation. The fig. 6 shows the result of simulation of the control system with the infinite-horizon LQR controller. The control task was to increase the temperature value from 15 to 20 degrees in a finite time (1 hour). The first plot of figure 6 shows the change of the indoor temperature T_i (9), the second graph shows the change of the feedback control signal $\boldsymbol{u}(t)$ (11). ■

The infinite-horizon LQR controller is easier in practical applications because the values of matrix $\boldsymbol{K}$ are constant and it is a stationary proportional controller. By changing values of the matrices $\boldsymbol{W}$ and $\boldsymbol{R}$ is possible to modify in wide range of how the controller works e.g. approximate time after which the desired value of controlled variable will be achieved.

4 The Complex Control System

The goal is to build a control system which works properly in phase I and phase II, see fig. 4. The control system will be implemented in a computer so this



(a) (b)

Fig. 6. (a): Changes of the inside temperature T_i (9) and (b): the control signal $\boldsymbol{u}(t) = Q_p(t)$ (11)

gives ability to build more complex control system. The infinite-horizon LQR controller will be used because, as was said earlier, this kind of controller is easy to use in practical applications. The figure 7 shows the block diagram of proposed control system. The LQR controller number 1 works during phase I

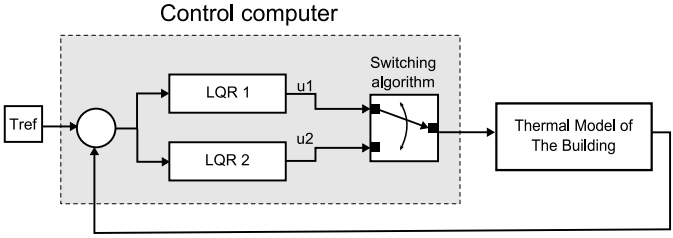
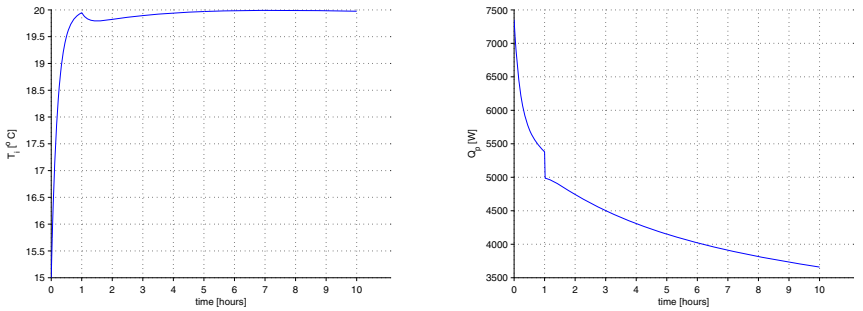


Fig. 7. Scheme of the proposed complex control system

and the LQR controller number 2 works during phase II. The difference between the two LQR controllers is in their values of weight matrices $\boldsymbol{W}$ and $\boldsymbol{R}$. The controller number 1 should achieve desired temperature in finite time of control. The controller number 2 should stabilise temperature on desired level but time of control is unknown. The block with title "Switching algorithm" on figure 4 contains an algorithm which decide which of controllers should work currently.

Simulation. The fig. 8 shows the simulation result of the control system whose block diagram is presented on figure 7. The control task was to increase the temperature value from 15 to 20 degrees in a finite time and next stabilise this temperature on desired level. The figure 8(a) shows the changes of the indoor temperature T_i (9), the figure 8(b) shows the change of the control signal $\boldsymbol{u}(t)$

(11). At the beginning, the LQR controller number 1 is working, after some time when desired the indoor temperature T_i is reached, the LQR controller number 2 starts to work . The moment of switched between the controller number 1 and the controller number 2 can be observed on figure 8(b) as a step change of the value of control signal $\boldsymbol{u}(t)$.



(a) (b)

Fig. 8. Changes of the inside temperature (left plot) and the control signal $\boldsymbol{u}(t) = Q_p(t)$ (right plot)

The fig. 9 shows a comparison of the controls signals for the single LQR controller (dotted line) and the control system which contains the two LQR controllers (solid line). The control system which includes two LQR controllers uses less energy to stabilise the temperature than system with one LQR controller and the area between two curves corresponds to the amount of saved energy. ■

5 Conclusions

This paper is presented the control system which contains two infinite-horizon LQR controllers. The infinite-horizon LQR controllers were chosen because: the control system is closed-loop system with negative feedback, simple structure of the infinite-horizon LQR controller (proportional controller) and modifying the weight matrices $\boldsymbol{W}$ and $\boldsymbol{R}$ can change the nature of the work control system. Also, some disadvantages of the infinite-horizon LQR controller are existing like: the LQR controller is a proportional controller so always is a deviation between desired value and real value of controlled signal, the LQR controller for proper work needs to know values of all coordinates of the state vector $\boldsymbol{x}(t)$ and in some cases reconstruction of the non-measurable coordinates of the state variables is needed. As shown by results of the simulations there is possibility to control indoor air temperature efficiently and using less energy. The LQR controller minimises the cost function which also takes into account the energy consumption of the control signal. Recent times, can be observed the growing popularity

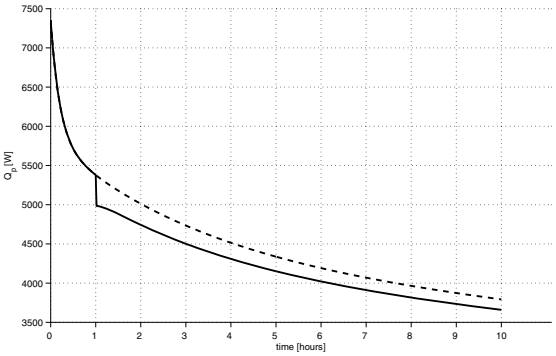


Fig. 9. Comparison of the controls signals $\boldsymbol{u}(t) = Q_p(t)$ for the single LQR controller (dotted line) and the control system which contains the two LQR controllers (solid line)

of wireless home automation devices. The future work will be concentrated on practical implementation of the proposed solutions in devices which work in ZWave and ZigBee standard.

Acknowledgements. Work is financed by NCN-National Science Centre founds for 2011-2013 as a research project. Contract number N N514 644440.

References

1. Clarke, J.A.: Energy simulation in building design. Butterworth-Heinemann (2001)
2. Crabb, J.A., Murdoch, N., Penman, J.M.: A simplified thermal response model. Building Services Engineering Research and Technology 8(1), 13–19 (1987)
3. Deng, K., Barooah, P., Mehta, P.G., Meyn, S.P.: Building thermal model reduction via aggregation of states. In: Proceedings of the 2010 American Control Conference, ACC 2010, pp. 5118–5123 (2010)
4. Gouda, M.M., Danaher, S., Underwood, C.P.: Low-order model for the simulation of a building and its heating system. Building Services Engineering Research and Technology 21(3), 199–208 (2000)
5. Gouda, M.M., Danaher, S., Underwood, C.P.: Building thermal model reduction using nonlinear constrained optimization. Building and Environment 37(12), 1255–1265 (2002)
6. Incropera, F.P., DeWitt, D.P., Bergman, T.L., Lavine, A.S.: Fundamentals of Heat and Mass Transfer. John Wiley & Sons (September 2006)
7. Mitkowski, W.: Stabilizacja Systemów Dynamicznych. WNT Warszawa (1991)
8. Underwood, C., Yik, F.: Modelling methods for energy in buildings. Blackwell Science (2004)

Introducing Periodic Parameters in a Marine Ecosystem Model Using Discrete Linear Quadratic Control

Mustapha El Jarbi, Thomas Slawig, and Andreas Oschlies

Institute for Computer Science, Christian- Albrechts Universitaet zu Kiel,
24098 Kiel, Germany
GEOMAR, Düsternbrooker Weg 20, 24105 Kiel
{mej,ts}@informatik.uni-kiel.de,
{aoschlies}@ifm-geomar.de

Abstract. This paper presents the application of the *Discrete Linear Quadratic Control (DLQC)* method for a parameter optimization problem in a marine ecosystem model. The ecosystem model simulates the distribution of nitrogen, phytoplankton, zooplankton and detritus in a water column with temperature and turbulent diffusivity profiles taken from a three-dimensional ocean circulation model. We present the linearization method which is based on the available observations. The linearization is necessary to apply the DLQC method on the nonlinear system of state equations. We show the form of the linearized time-variant problems and the resulting two algebraic Riccati Equations. By using the DLQC method, we are able to introduce temporally varying periodic model parameters and to significantly improve – compared to the use of constant parameters – the fit of the model output to given observational data.

Keywords: Optimal Control, Non-linear Systems, Parameter Optimization, Biogeochemical Modelling, Discrete Linear Quadratic Regulator Problem, Periodic Parameter, Discrete Riccati Equation.

1 Introduction

We consider nonlinear partial differential diffusion-advection systems of the form

$$\frac{\partial x^i}{\partial t} = -w^i \frac{\partial x^i}{\partial z} + \frac{\partial}{\partial z}\left(\nu_\rho \frac{\partial x^i}{\partial z}\right) + q^i(\mathbf{x}, \mathbf{u}), \quad i = 1, 2, 3, 4 \tag{1}$$

$x^i : [0, T] \times [-H, 0] \longrightarrow \mathbb{R}$.

Here z denotes the vertical spatial coordinate, H the depth in the water column, q^i represents the biogeochemical coupling terms for the four species and $\mathbf{u} = (u_1, \ldots, u_p)$ is the vector of unknown physical and biological parameters. The circulation data are the turbulent mixing coefficient $\nu_\rho = \nu_\rho(z, t)$ and the temperature $\Theta = \Theta(z, t)$, which goes into the non-linear coupling terms q^i, see

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 481–490, 2013.

(3). The vertical sinking velocity w^i is a parameter of the biological model that is nonzero only for x^4, i.e. $w^1 = w^2 = w^3 = 0$, $w^4 = ws > 0$.

The state of the system is denoted by $\mathbf{x} = (x^1, x^2, x^3, x^4)^\top$ and the control by $\mathbf{u}$. A control problem is defined as

$$\min_{\mathbf{u}} \mathcal{F}(\mathbf{x}, \mathbf{u}) \quad \textit{subject to} \quad (1), \tag{2}$$

where $\mathcal{F}$ is a cost functional which will be introduced later.
Our main goals are:

- to minimize a least-squares type cost functional,
- to allow the parameters to vary temporally over the year while remaining periodic over all years of the considered time interval.

The work presented in this paper is motivated by results obtained for a typical marine ecosystem model, namely the NPZD model introduced in [1], [2]. As was reported in several publications with different optimization algorithms, the quality of the model-to-dat fit was not optimal, and in some cases it was difficult to identify the parameters uniquely, see for example [3],[5],[4]. In most cases, the parameters of the marine ecosystem models are assumed to be temporally constant. This reflects the aim to obtain a model that is applicable for arbitrary time intervals. To solve this problems, we discretize and linearize the nonlinear state (1) around a reference trajectories and we interpret it as a Discrete Linear Quadratic Control (DLQC) problem. Therein, we allow the parameters to be time-dependent, apply a well-established method for optimal control, and additionally impose the constraint of annual periodicity. This avoids the process of parametrization in the sense that we do not have to know or assume how the above mentioned periodic functions look like. In contrast, the method itself will generate an optimal periodic function for each parameter. Moreover, it allows to balance the two aims that we have: By introducing weight matrices we can choose if it is more important to obtain a very good fit or nearly perfect periodicity. The method requires a reference trajectory and a reference control, i.e., the vector of model parameters. The former can be taken from the measurement data, and for the latter we use an initial guess for the parameters which can be the output of an optimization with constant parameters. The outline of this paper is as follows. In the next section we briefly described the model Equation and optimization problem (2), the DLQC problem formulation in section 3. A application of the DLQC method on the NPZD model is presented in section 4.3. Afterwards, we present our results with respect to the quality of the fit and the periodicity of the parameters and end the paper with some conclusions.

2 Model Equations and Optimization Problem

In this section we give the formulations of the NPZD model and of the corresponding parameter optimization problem and we formulate the optimization problem for the discrete model.

2.1 Model Equations

This section describes the ecosystem model. The considered system (1) is a spatially one-dimensional marine biogechemical model, that simulates the interaction of dissolved inorganic nitrogen N, phytoplankton P, zooplankton Z and detritus D. It was developed with the aim of simultaneously reproducing observations at three North Atlantic locations by the optimization of free parameters within credible limits, see [4]. The model uses the ocean circulation and temperature field in an off-line modus, i.e. these are used only as forcing, but no feedback on them is modeled. The model simulates one water column at a given horizontal position, which is motivated by the fact that there have been special time series studies at fixed locations, one of which was used here. In the model, the concentrations (in mmol N m^{-3}) of dissolved inorganic nitrogen N, phytoplankton P, zooplankton Z, and detritus D, denoted by $\mathbf{x} = (x^i)_{i=1,\ldots,4} = (N, P, Z, D)$ are described by the PDE system (1).

The biogeochemical source-minus-sink terms $\mathbf{q} = (q^i)_{i=1,\ldots,4}$ are explicit by given in [1]:

$$\left.\begin{aligned} q^1(\mathbf{x},\mathbf{u}) &= -\bar{J}(z,t,N)P + \gamma_2 Z + \mu_D D, \\ q^2(\mathbf{x},\mathbf{u}) &= \bar{J}(z,t,N)P - \mu_X P - G(P)Z, \\ q^3(\mathbf{x},\mathbf{u}) &= \gamma_1 G(P)Z - \gamma_2 Z - \mu_Z Z^2, \\ q^4(\mathbf{x},\mathbf{u}) &= (1-\gamma_1)G(P)Z - \mu_Z Z^2 + \mu_X P - \mu_D D - w_s \tfrac{\partial D}{\partial z} \end{aligned}\right\} \tag{3}$$

where $\bar{J}$ is the daily averaged phytoplankton growth rate as a function of depth z and time t, and G is the grazing function (see below). The remaining parameter in the above equations are defined in [1],

$$G(\epsilon, g) = \frac{g\epsilon P^2}{g + \epsilon P^2} \qquad \bar{J}(z,t,N) = \min\left(L(z,t), J_{max}\frac{N}{K_1 + N}\right), \tag{4}$$

where L denotes the purely light-limited growth rate, and J_{max} is the light-saturated growth. For more details of $\bar{J}$, L and the parameters see [1], [4].

2.2 The Optimization Problem

The aim of the optimization is to fit the aggregated model output $\mathbf{y} = C\mathbf{x}$ (C is called the output matrix) to the given observational data $\mathbf{y}^{obs}$. There are five types of measurement data $\mathbf{y}^{obs} = (y_m^{obs})_{m=1,\ldots,5}$, which correspond to aggregated values $\mathbf{y} := (y_m)_{m=1,\ldots,5}$ of the model output see also [3]. Thus the cost function can be written as:

$$\mathcal{F}(\mathbf{x},\mathbf{u}) := \|C\mathbf{x} - \mathbf{y}^{obs}\|_{2,\sigma}, \tag{5}$$

where $\|\|_{2,\sigma}$ is a Euclidean norm weighted using the vector

$$\sigma = (\sigma_l)_{l=1,\ldots,5} = (0.1, 0.01, 0.01, 0.0357, 0.025)$$

of uncertainties corresponding to the five types of measurement data.

3 DLQC Problem Formulation

We use a discrete *linear time-varying (LTV) system*, i.e. we assume that the dynamical system is already discretized in time, namely at discrete times $t_k, k = 1, \ldots, M$. In the context of the DLQC, one usually considers a discrete-time system of the form:

$$\begin{aligned} \mathbf{x}_{k+1} &= A_k \mathbf{x}_k + B_k \mathbf{u}_k, \quad k = 1, 2, ..., M-1 \\ & x_1 \text{ (the given initial value)}, \end{aligned} \tag{6}$$

where in every time step k

- $\mathbf{x}_k = \mathbf{x}(t_k) \in \mathbb{R}^n$ is called the state vector (here the model output),
- $\mathbf{u}_k = \mathbf{u}(t_k) \in \mathbb{R}^p$ is the control (here the model parameter) vector, with the parameter vector from the model (3).
- The matrix $A_k \in \mathbb{R}^{n \times n}$ and $B_k \in \mathbb{R}^{n \times p}$ are called the system matrix and the input matrix, respectively.

We will use the notations

$$\begin{aligned} \mathbf{x} &= (\mathbf{x}_k)_{k=1,\ldots,M} \in \mathbb{R}^{M \times n} \cong \mathbb{R}^{Mn}, \\ \mathbf{u} &= (\mathbf{u}_k)_{k=1,\ldots,M-1} \in \mathbb{R}^{(M-1) \times p} \cong \mathbb{R}^{(M-1)p} \end{aligned}$$

for the whole discrete trajectories of state and control vector, respectively. The quadratic cost function of this optimal control problem is defined by:

$$\mathcal{J}(\mathbf{u}) = \frac{1}{2} \mathbf{x}_M^\top Q_M \mathbf{x}_M + \frac{1}{2} \sum_{k=1}^{M-1} \mathbf{x}_k^\top Q_k \mathbf{x}_k + \mathbf{u}_k^\top R_k \mathbf{u}_k, \tag{7}$$

where in every time step k

- Q_k is a positive semidefinite diagonal weighting matrix for the state vector for every model time step $k = 1, \ldots, M$,
- R_k is a positive definite diagonal weighting matrix for the control vector for every model time step $k = 1, \ldots, M-1$.

For the solution of a discrete linear quadratic optimal control problem with LTV systems, there exists the following theorem, see [6].

Theorem 1. *If the $Q_k, k = 1, \ldots, M$, are positive semi-definite and the $R_k, k = 1, \ldots, M-1$, are positive definite, then there exists a unique solution of the DLQC (6), (7). The optimal control is given by the* feedback law

$$\begin{aligned} \mathbf{u}_k &= -K_k \mathbf{x}_k, \quad k = 1, \ldots, M-1. \\ K_k &:= (R_k + B_k^\top \mathbf{X}_{k+1} B_k)^{-1} B_k^\top \mathbf{X}_{k+1} A_k, \qquad k = 1, \ldots, M-1 \\ \mathbf{x}_{k+1} &= (A_k - B_k K_k) \mathbf{x}_k, \qquad k = 1, \ldots, M-1. \end{aligned}$$

where the $(\mathbf{X}_k)_{k=1,\ldots,M-1}$, is the unique symmetric solution of the Discrete Riccati Equation (DRE).

$$\left.\begin{aligned} \mathbf{X}_k = Q_k + A_k^\top \mathbf{X}_{k+1} A_k - A_k^\top \mathbf{X}_{k+1} B_k (R_k + B_k^\top \mathbf{X}_{k+1} B_k)^{-1} B_k^\top \mathbf{X}_{k+1} A_k, \\ k = 1, \ldots, M-1. \end{aligned}\right\} \tag{8}$$

4 Application of DLQC to the NPZD Model

In this section we apply the LQOC method to the discretized version of the NPZD model. We present the details of discretization, linearization and the enforcement of the periodicity of the parameters (controls).

4.1 Discretization Scheme

We use a discrete linear quadratic control (DLQC). For this purpose we present the original discretization scheme of the model.

The NPZD model is forced by output from the OCCAM global circulation model, namely the hourly vertical profiles of temperature t and vertical diffusivity ν_ρ. The vertical grid consists of 32 layers with thickness increasing with depth. The time integration of the system (1) is performed by an operator splitting method:

– At first, the nonlinear coupling operators $\mathbf{q}_k = (q_k^1, q_k^2, q_k^3, q_k^4)_{k=1,\ldots,M-1}^\top$ are computed at every spatial grid point and integrated by four explicit Euler steps, each of which is described by the operator:

$$B_k(\mathbf{x}_k, \mathbf{u}_k) := (\mathbf{x}_k + \frac{\tau}{4}\mathbf{q}_k(\mathbf{x}_k, \mathbf{u}_k)). \tag{9}$$

This gives an intermediate iterate

$$\hat{\mathbf{x}}_k := B_k \circ B_k \circ B_k \circ B_k(\mathbf{x}_k, \mathbf{u}_k).$$

– Then, an explicit Euler step with full step-size τ is performed for the sinking term, which is spatially discretized by an upstream scheme. This step is summarized in a matrix S. Since the sinking velocity is temporally constant, this matrix does not depend on the time step k. Thus, at the end of this step, an intermediate tracer vector $\tilde{\mathbf{x}}_k$ is computed as

$$\tilde{\mathbf{x}}_k := S\hat{\mathbf{x}}_k, \tag{10}$$

where $S = (I_k + \tau A^{adv})$.

– Finally, an implicit Euler step is applied for the diffusion operator discretized with second order central differences. The resulting matrix D_k for the diffusion depends on k since the diffusion coefficient depends on time. The matrix is tridiagonal, and the system is solved directly for $\mathbf{x}_{k+1}$

$$\tilde{D}_k\mathbf{x}_{k+1} = \tilde{\mathbf{x}}_k, \tag{11}$$

where $\tilde{D}_k = (I_k - \tau D_k)\mathbf{x}_{k+1}$.

Summarizing, the discrete system can be written as

$$\begin{aligned}\mathbf{x}_{k+1} &= \tilde{D}_k^{-1} S B_k \circ B_k \circ B_k \circ B_k(\mathbf{x}_k, \mathbf{u}_k) \\ &= \tilde{D}_k^{-1} S G(\mathbf{x}_k, \mathbf{u}_k), \qquad k = 1, ..., M-1,\end{aligned} \tag{12}$$

The function G is nonlinear and represents the discretized source minus sink terms.

4.2 Linearization of the Model

The LDQC approach is based on a linearization of (12) to obtain a linear time-varying problem. The linearization is performed around *reference trajectories* $(\mathbf{x}_k^r, \mathbf{u}_k^r)_{k=1,\dots,M-1}$. For the reference state trajectory we take available the observational data, is taken from the Bermura Atlantic Time-series Study (BATS) see also [7], the choice of the reference control trajectory is described in below. The linearized state equation now reads

$$\tilde{\mathbf{x}}_{k+1} = A_k \tilde{\mathbf{x}}_k + B_k \mathbf{v}_k + r_k, \qquad k = 1, \dots, M-1, \tag{13}$$

where

$$\begin{aligned}
A_k &= \tilde{D}_k^{-1} S \frac{\partial G}{\partial x}(\mathbf{x}_k^r, \mathbf{u}_k^r), \quad A_k \in \mathbb{R}^{n \times n} \\
B_k &= \tilde{D}_k^{-1} S \frac{\partial G}{\partial u}(\mathbf{x}_k^r, \mathbf{u}_k^r) \quad B_k \in \mathbb{R}^{n \times p}, \\
r_k &= \tilde{D}_k^{-1} S G(\mathbf{x}_k^r, \mathbf{u}_k^r) - \mathbf{x}_{k+1}^r, \quad r_k \in \mathbb{R}^n \\
\tilde{\mathbf{x}}_k &= \mathbf{x}_k - \mathbf{x}_k^r, \quad \mathbf{v}_k = \mathbf{u}_k - \mathbf{u}_k^r, \quad \tilde{\mathbf{x}}_k \in \mathbb{R}^n, \quad \mathbf{v}_k \in \mathbb{R}^p.
\end{aligned}$$

Now we write the linearized problem in the form of a (LDQC) problem, therefore we set:

$$\hat{\mathbf{x}}_k := \begin{pmatrix} \tilde{\mathbf{x}}_k \\ 1 \end{pmatrix}, \quad \hat{A}_k = \begin{pmatrix} A_k & r_k \\ 0 & 1 \end{pmatrix}, \quad \hat{B}_k = \begin{pmatrix} B_k \\ 0 \end{pmatrix}, \quad \hat{Q}_k = \begin{pmatrix} Q_k & 0 \\ 0 & 0 \end{pmatrix}$$

where $\hat{\mathbf{x}}_k \in \mathbb{R}^{n+1}$, $\hat{A}_k \in \mathbb{R}^{(n+1)\times(n+1)}$, $\hat{B}_k \in \mathbb{R}^{(n+1)\times p}$, $\hat{Q}_k \in \mathbb{R}^{(n+1)\times(n+1)}$. The linearized state equation (13) can be written in a form similar to (6), namely as:

$$\hat{\mathbf{x}}_{k+1} = \hat{A}_k \hat{\mathbf{x}} + \hat{B}_k \mathbf{v}_k \quad k = 1, ..., M-1. \tag{14}$$

Enforcing Periodicity of the Parameters. A main objective of this work is to enforce periodicity of the parameters/controls. For this purpose, let us assume that the length of a time period – measured in time steps – is $T > 0$ and that $M \bmod T = 0$. We now chose the reference trajectory for the control $\mathbf{u}^r = (\mathbf{u}_k^r)_{k=1,\dots,M-1} \in \mathbb{R}^{(M-1)p}$ to be

$$\mathbf{u}_k^r := \begin{cases} \mathbf{u}_0, & \text{if } k \le T \\ \mathbf{u}_{k-T}, & \text{if } k > T. \end{cases} \tag{15}$$

Where $\mathbf{u}_0$ is the parameter vector determined by optimization in [1], that was used as an initial guess here. we will enforce periodicity of $\mathbf{u}_k = \mathbf{u}_k^r + \mathbf{v}_k = \mathbf{u}_{k-T} + \mathbf{v}_k$ for $k \ge T+1$.

4.3 Application to NPZD Model

From Theorem 1 in section 3, the optimal control is given by

$$\mathbf{v}_k = -(R_k + \hat{B}_k^\top \hat{\mathbf{X}}_{k+1} \hat{B}_k)^{-1} \hat{B}_k^\top \hat{\mathbf{X}}_{k+1} \hat{A}_k \hat{\mathbf{x}}_k, \quad k = 1, \ldots, M-1.$$

According to (15), we find

$$\mathbf{u}_k = \begin{cases} \mathbf{u}_0 - (R + \hat{B}_k^\top \hat{\mathbf{X}}_{k+1} \hat{B}_k)^{-1} \hat{B}_k^\top \hat{\mathbf{X}}_{k+1} \hat{A}_k \hat{\mathbf{x}}_k, \text{ if } k \le T, \\ \mathbf{u}_{k-T} - (R_k + \tilde{B}_k^\top \hat{\mathbf{X}}_{k+1} \hat{B}_k)^{-1} \hat{B}_k^\top \hat{\mathbf{X}}_{k+1} \hat{A}_k \hat{\mathbf{x}}_k, \text{ if } k > T. \end{cases}$$

Here the $\hat{\mathbf{X}}_k$ can be computed backwards in discrete time, starting from

$$\hat{\mathbf{X}}_M = \hat{Q}_M, \tag{16}$$

as the unique symmetric solutions of the Discrete Riccati equations (8). We set

$$\hat{\mathbf{X}}_k = \begin{bmatrix} \mathbf{X}_k & h_k \\ h_k^\top & \alpha_k \end{bmatrix} \tag{17}$$

with $h_k \in \mathbb{R}^n$ and $\alpha_k \in \mathbb{R}$ for $k = 1, \ldots, M-1$. we easily get

$$\mathbf{u}_k = \begin{cases} \mathbf{u}_0 + K_k \mathbf{z}_k + S_k, \text{ if } k \le T, \\ \mathbf{u}_{k-T} + K_k \mathbf{z}_k + S_k, \text{ if } k > T, \end{cases}$$

where K_k and S_k are given by

$$\begin{aligned} K_k &= -(R_k + B_k^\top \mathbf{X}_{k+1} B_k)^{-1} B_k^\top \mathbf{X}_{k+1} A_k, \quad k = M-1, \ldots, 1 \\ S_k &= -(R_k + B_k^\top \mathbf{X}_{k+1} B_k)^{-1} B_k^\top (\mathbf{X}_{k+1} r_k + h_{k+1}), \quad k = M-1, \ldots, 1. \end{aligned}$$

Now, the system (16), (17) to compute the $\mathbf{X}_k$ can be separated into

$$\begin{aligned} \mathbf{X}_M &= Q_M, \\ \mathbf{X}_k &= Q_k + A_k^\top \mathbf{X}_k A_k - A_k^\top \mathbf{X}_{k+1} B_k (R_k + B_k^\top \mathbf{X}_{k+1} B_k)^{-1} B_k^\top \mathbf{X}_{k+1} A_k, \\ & \quad k = M-1, \ldots, 1. \end{aligned}$$

To evaluate the $\mathbf{X}_k$ and an additional difference equation for the h_k, namely

$$\begin{aligned} h_M &= 0, \\ h_k &= A_k^\top (\mathbf{X}_{k+1} r_k + h_{k+1}) - A_k^\top \mathbf{X}_{k+1} B_k (R_k + B_k^\top \mathbf{X}_{k+1} B_k)^{-1} B_k^\top (\mathbf{X}_{k+1} r_k + h_{k+1}), \\ & \qquad k = M-1, \ldots, 1. \end{aligned}$$

For the application on the NPZD Model, Q_k is to be constant for all k, this can be written as following

$$Q_k = \operatorname{diag}(\frac{1}{\sigma_l^2})_{l=1,\ldots,5}, \quad k = 1, \ldots, M-1,$$

and the matrix R_k can be written as

$$R_k = diag \begin{cases} \frac{1}{|(\mathbf{u}_0)_i|^2}, & i = 1, \ldots, p, k = 1, \ldots, T \\ \frac{1}{|(\mathbf{u}_{k-T,i}|^2}, & i = 1, \ldots, p, k = T+1, \ldots, M-1 \end{cases}$$

5 Optimization Results

5.1 Fit of Model Output to Observational Data

This section shows a comparison between the optimized model output obtained by the DLQC method with periodic parameters and the observational data. As a reference we also compare the results to those obtained by a direct optimization of the nonlinear model using *constant parameters* with a *Sequential Quadratic Programming (SQP)* method that takes into account parameter bounds. This method was used in [4]. We performed the optimization for the years 1994 to 1998, in contrast to the years 1991 to 1996 that were used in [4]. The reason for this is that no zooplankton data are available at BATS for the years 1991 to 1993, which would be disadvantageous for the linearization procedure in the DLQC method. In [4] a minimum value of the cost function (5) of $\mathcal{F} \approx 70$ was obtained for optimized constant parameters for the five year time interval $[1991, 1996]$. For the time interval $[1994, 1998]$ a comptation with the method used in [4] gave a very similar value. In contrast to these and other (as in [3]) earlier results obtained for constant model parameters, the DLQC method gives a nearly perfect fit of the data. Figure 1 shows the model results $\mathbf{y}$ obtained with the DLQC method together with the observational data $\mathbf{y}^{obs}$ for the years 1994 to 1998. The model-data fit for $\mathbf{y}_2 = P$ (chlorophyll a) is nearly perfect. Even substantial concentration changes that occur between some neighboring measurement points (e.g. for $\mathbf{y}_4 = P + Z + D$ (particulate organic nitrogen), in 1994, 1995 or 1997) can be captured by the optimized trajectory. There are only some parts of the time interval where the trajectories are slightly farther away from the data, for example in 1996 for zooplankton and in the last two years of the simulated time interval for PON.

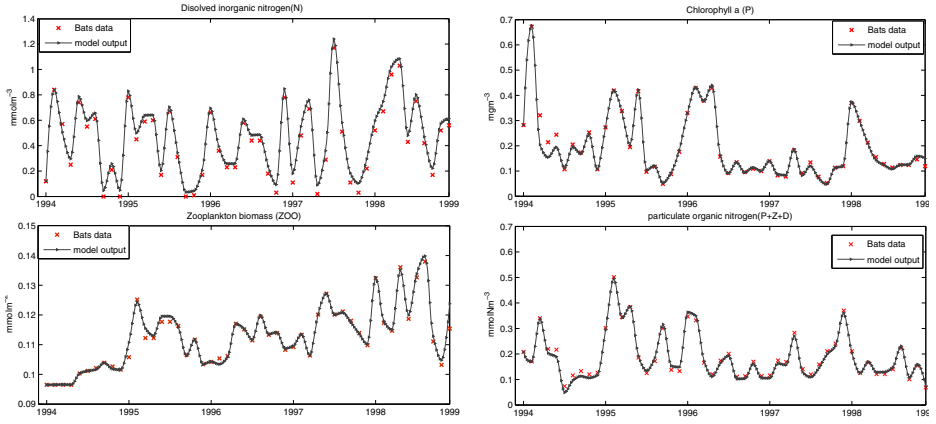


Fig. 1. Observational data $\mathbf{y}_i^{obs}, i = 1, .., 4$ and aggregated model trajectories $\mathbf{y}_i, i = 1, .., 4$, optimized with periodic parameters obtained by the DLQC method. Values are shown for the upper layer (depth less than 5 meters) and years 1994-1998.

5.2 Periodicity of the Parameters

we show here that the above model-to-data fit can be achieved by almost annually periodic parameters. This was possible due to an appropriate adjustment of the matrices Q_k and $R_k, K = 1, \ldots, M-1$, in the cost function (7) used in the DLQC framework, see section 4.3. Thus both the annual periodicity of the parameters. Due to the choice of the reference values for the parameters in the first year, we could also keep the parameters in their desired bounds, although these bounds need not to be imposed explicity. Figures 2 illustrate the temporal behavior of the selected four parameters that were optimized with the DLQC method. In these figure, the temporal changes of the parameters are plotted against the *actual* times of the linearization points which are determined by the available measurements. Obviously, the DLQC method then leads to perfectly periodic parameters.

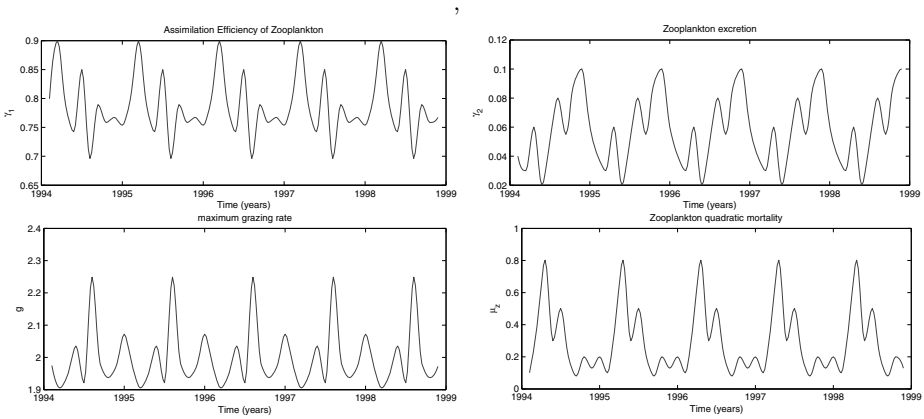


Fig. 2. Periodicity of the selected optimal parameters $u_{n,1} = \gamma_1, u_{n,4} = \gamma_2, u_{n,6} = g, u_{n,8} = \mu_z$, obtained by the DLQC method

6 Conclusion

In this paper, we successfully applied the method linear quadratic optimal control to the optimization of an one-dimensional marine ecosystem model. The model has to be linearized to fit in the LQOC frame work. The method permits perfect periodic evolution of model parameters and additionally notably improves the fit of the data in comparison with the solution with fixed model parameters. We demonstrated that the LQOC optimization is suitable for the considered problem and furthermore have shown that this method provides a very reasonable solution. Even with the available small number of observational data, which is typical to oceanographic time series sites, its quality is very high. Temporal deviations of individual parameters about the annual mean can be analyzed further

to help making inferences about processes that the model cannot describe well when constant parameters are used. This analysis should contribute to a better understanding of model deficiencies and may improve marine ecosystem models. A next step could be to use only a part of the time horizon to estimate the periodic parameters and verifying the model and the parameters on the remaining part of the data.

Acknowledgements. The authors would like to thank Johannes Rückelt, Institute of Computer Science, Christian-Albrechts-Universität zu Kiel, for a direct optimization of the nonlinear model using constant parameters with a Sequential Quadratic Programming (SQP) method over the years 1994 to 1998.

References

1. Oschlies, A., Garon, V.: An eddy-permitting coupled physical-biological model of the North Atlantic 1. Sensitivity to advection numerics and mixed layer physics. Global Biogeochemical Cycles 13, 135–160 (1999)
2. Schartau, M., Oschlies, A.: Simultaneous data-based optimization of a 1d-ecosystem model at three locations in the north Atlantic: Part I - method and parameter estimates. Journal of Marine Research 61, 765–793 (2003)
3. Ward, B.: Marine Ecosystem Model Analysis Using Data Assimilation (2009), http://web.mit.edu/benw/www/Thesis.pdf
4. Rückelt, J., Sauerland, V., Slawig, T., Srivastav, A., Ward, B., Patvardhan, C.: Parameter Optimization and Uncertainty Analysis in a Model of Oceanic CO_2-Uptake using a Hybrid Algorithm and Algorithmic Differentiation. Nonlinear Analysis B Real World Applications 10, 3993–4009 (2010)
5. Ward, B.A., Anderson, M.A.M., Friedrichs, T.R., Oschlies, A.: Parameter optimisation techniques and the problem of underdetermination in marine biogeochemical models. Journal of Marine Systems and Control Letters 81, 34–43 (2010)
6. Rugh, W.J.: Linear System Theory, 2nd edn. Prentice-Hall, Upper Saddle River (1996)
7. Bermuda Atlantic Time-Series Study, http://www.bios.edu/research/bats.html

Avoidance Trajectories Using Reachable Sets and Parametric Sensitivity Analysis*

Matthias Gerdts and Ilaria Xausa

Universität der Bundeswehr, Institut für Mathematik und Rechneranwendung (LRT),
Werner-Heisenberg-Weg 39, 85577 Neubiberg, Germany
{matthias.gerdts,ilaria.xausa}@unibw.de
http://www.unibw.de/lrt1/gerdts

Abstract. The article suggests a conceptual model-based simulation method with the aim to detect collision of cars in all-day road traffic. The benefit of the method within a driver assistance system would be twofold. Firstly, unavoidable accidents could be detected and appropriate actions like full braking maneuvers could be initiated in due course. Secondly, in case of an avoidable accident the algorithm is able to suggest an evasion trajectory that could be tracked by a future active steering driver assistance system. The algorithm exploits numerical optimal control techniques and reachable set analysis. A parametric sensitivity analysis is employed to investigate the influence of inaccurate sensor measurements.

Keywords: driver assistance, collision avoidance, optimal control, reachable sets, parametric sensitivity analysis.

1 Introduction

Over the years many passive and active safety systems have been developed for modern passenger cars with the aim to reduce the number of casualties in traffic accidents. Passive safety systems contain amongst others improvements of the chassis, airbags, seat belts, and seat belt tighteners. These safety systems help to reduce the severeness of accidents once an accident has occured. In contrast, semi-active safety systems and driver assistance systems, for instance anti-blocking system, braking assistant, anti-slip regulation, electronic stability control, adaptive cruise control, lane departure warning, or blind spot intervention, become active in critical situations before an accident occurs and intend to prevent accidents. In future, active driver assistance systems that actively initiate braking maneuvers or even active steering maneuvers to avoid obstacles will become relevant in order to detect potential collisions and reduce severeness of collisions. The availability of high performance sensors will play a central role

* The work is supported by the Initial Training Network Sensitivity Analysis for Deterministic Controller Design (ITN SADCO), which is funded by 7th Framework Programme of the EU.

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 491–500, 2013.

in future collision avoidance systems. But next to the required technical devices, intelligent software systems and algorithms will play a crucial role as well. The main tasks in collision avoidance are to reliably indicate future collisions and – if possible – to provide escape trajectories if such exist. This paper suggests an optimal control based method that has the potential of fulfilling these two tasks.

2 Model Scenarios

We investigate two model scenarios that are likely to occur in all-day traffic. According to [9] wrong velocity, short distance, and overtaking maneuvers are responsible for approximately 29.7 % of accidents with injuries to persons. The following model scenarios address these situations, see Figure 1. The typical time to collision ranges from 0.5 to 3 seconds. For simplicity we assume a straight road throughout. A reference coordinate system is used with the x-axis pointing into the longitudinal direction of the road and the y-axis pointing in the cross-direction of the road.

Scenario 1: A stationary obstacle at a given distance to an approaching car, which drives at a prescribed speed, has to be avoided.

Scenario 2: An overtaking maneuver on a highway is considered. One car (car A) has initiated an overtaking maneuver to overtake car B while another car (car C) is approaching at a prescribed speed.

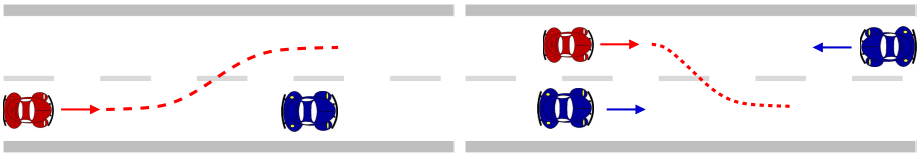


Fig. 1. Collision avoidance model scenarios: stationary obstacle (left) and overtaking maneuver (right)

For these two model scenarios we aim at answering the following questions:

- Can a collision be avoided?
- If a collision can be avoided, how can it be avoided?

3 Model of the Car

In this article the single-track car model is used. It is a simplified car model, which is commonly used in the automobile industry for basic investigations of the dynamical behavior of cars. It is based on the simplifying assumptions that rolling and pitching behavior of the car body can be neglected. The car model

includes two control variables for the driver: the steering angle velocity $|w_\delta| \leq 0.5$ $[rad/s]$ and a function F_B with values in $[F_{Bmin}, F_{Bmax}]$, $F_{Bmin} = -5000$ $[N]$, $F_{Bmax} = 15000$ $[N]$, which models a combined brake (if $F_B > 0$) and acceleration (if $F_B < 0$) assembly. Details on the model can be found in [5,6,7]. The dynamics are given by the following system of differential equations for the car's center of gravity (x, y) in the plane, the yaw angle ψ, the velocities v_x and v_y in x- and y-direction, respectively, the yaw angle rate w_ψ, and the steering angle δ:

$$x' = v_x, \quad y' = v_y, \quad \psi' = w_\psi, \quad \delta' = w_\delta, \tag{1}$$

$$v_x' = \frac{1}{m}\left[F_x \cos\psi - F_y \sin\psi\right], \tag{2}$$

$$v_y' = \frac{1}{m}\left[F_x \sin\psi + F_y \cos\psi\right], \tag{3}$$

$$w_\psi' = \frac{1}{I_{zz}}\left[F_{sf} \cdot l_v \cdot \cos\delta - F_{sr} \cdot l_h + F_{lf} \cdot l_v \cdot \sin\delta\right], \tag{4}$$

The functions F_x, F_y, F_{sf}, F_{sr}, F_{lf} denote forces (in x-, y-direction, as well as lateral and longitudinal tyre forces at front and rear wheels) and are smooth nonlinear functions of the state $(x, y, \psi, v_x, v_y, w_\psi, \delta)$ and m, I_{zz}, l_v, l_h are constants. For further details please refer to [5,6,7]. For the following numerical computations we used realistic data for the various parameters involved in this model. Unfortunately, these parameter values are proprietary and may not be published. For a different parameter set which is quite realistic please refer to [5].

4 Collision Detection and Collision Avoidance

Once an obstacle has been detected by suitable sensors, e.g. radar or lidar, we use the following approaches to decide whether a collision is going to happen or not. As we intend to use optimal control to model the scenarios, as a by-product we obtain evasion trajectories if such exist at all. We investigate three different approaches. Herein, it is assumed for simplicity that the obstacle in scenario 1 is fixed close to the right boundary of a straight road as in the left picture in Figure 1. Moreover, the following approaches assume that the constellation of car and obstacle is such that a collision cannot be avoided by just applying a full braking maneuver.

4.1 Approach 1: Reaching a Safe Target Position for Scenario 1

The first approach aims at reaching a safe target state, which should be defined such that the evading car is able to avoid the obstacle and moreover is able to continue its drive after the obstacle has been passed. This approach is modeled by the following optimal control problem $\mathrm{OCP}(y_0, v_{x,0})$:

Minimize

$$c_1 t_f + c_2 d + c_3 \int_0^{t_f} w_\delta(t)^2 dt$$

subject to the equations of motion (1)-(4) with initial condition

$$(x(0), y(0), \psi(0), v_x(0), v_y(0), w_\psi(0), \delta(0)) = (0, y_0, 0, v_{x,0}, 0, 0, 0),$$

the control constraints $|w_\delta| \le w_{\delta,max}$, $F_B \in [F_{B,min}, F_{B,max}]$, *the pure state constraint*

$$y_{min} \le y(t) \le y_{max},$$

and boundary conditions

$$x(t_f) = d, \quad v_y(t_f) = 0, \quad y(t_f) \ge y_{target}.$$

Herein, the final time t_f is supposed to be free and $c_1, c_2, c_3 \ge 0$ are suitable constants. y_0 is the initial y-position of the evading car on the road, $v_{x,0}$ is the initial velocity in x-direction of the evading car. y_{min} and y_{max} define the boundaries of the road. The terminal constraint $v_y(t_f) = 0$ shall ensure that the evading car can continue its drive beyond t_f without leaving the road immediately. y_{target} defines a y-position sufficiently far away from the obstacle's y-position (safe target position).

d is the initial distance of the evading car to the obstacle. If $c_2 > 0$, then d is assumed to be an additional optimization parameter. The resulting optimal control problem then aims at finding the minimal distance to the obstacle that still allows to avoid a collision. If $c_2 = 0$, then d is supposed to be a fixed distance.

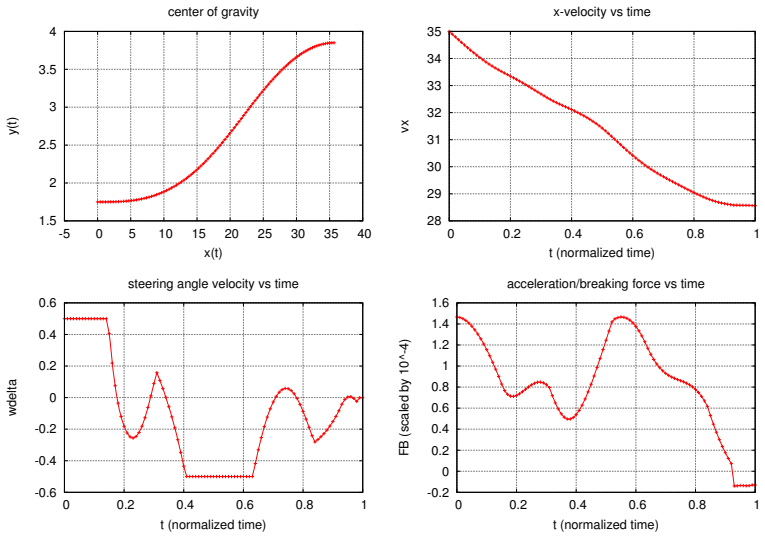


Fig. 2. Avoidance trajectory for $v_{x,0} = 35$, $y_0 = 1.75$, $y_{target} = 3.85$: center of gravity and velocity in x-direction (top), steering angle velocity and braking/acceleration force (bottom). The distance d computes to 35.6798 and the final time t_f to 1.1399.

In this case, it might happen that the problem becomes infeasible owing to the constraint $y(t_f) \geq y_{target}$. A remedy in this case is approach 2 below.

Figure 2 shows the result for $y_{min} = 0.9$, $y_{max} = 6.1$, $y_0 = 1.75$, $v_{x,0} = 35$, $c_1 = c_2 = 0.1$, $c_2 = 0.2$, $w_{\delta,max} = 0.5$, $F_{B,min} = -5000$, $F_{B,max} = 15000$, $y_{target} = 3.85$. The minimal distance computes to $d = 35.6798$ [m] and the final time (time to collision) is $t_f = 1.1399$ [s].

Approach 1 yields one single optimal trajectory provided an avoidance trajectory exists. This avoidance trajectory could be tracked by a real car.

However, the solution depends on the definition of the safe target position y_{target} and it does not exist if a collision is unavoidable. Therefore it would be nicer to have full information about what points on the road can actually be reached by the evading car. This leads to the following reachable set approach.

4.2 Approach 2: Computing the Projected Reachable Set

The second approach aims at providing all points on the road that can be reached by the evading car in finite time t_f from a given initial state with boundary condition $v_y(t_f) = 0$. More precisely, we aim at computing the *projected reachable set*

$$\mathcal{PR} := \bigcup_{d \in [d_{min}, d_{max}]} \bigcup_{y \in \mathcal{PR}(d)} \{(d, y)\},$$

where

$$\begin{aligned} \mathcal{PR}(d) := \big\{ \hat{y} \in \mathbb{R} \mid \exists \text{ final time } t_f > 0, \text{ controls } w_\delta, F_B, \\ \text{and states } x, y, \psi, v_x, v_y, w_\delta, \delta \text{ such that} \\ \text{dynamics and constraints are satisfied} \\ \text{and } \hat{y} = y(t_f), x(t_f) = d, v_y(t_f) = 0 \big\} \end{aligned}$$

denotes the projected reachable set at initial distance d. Note that we are not interested in the reachable set at t_f for the full state vector but only for the components x and y. In order to approximate the projected reachable set we employ the optimal control technique in [1,2], which for a simplified setting allows a first order approximation. The set $\mathcal{PR}$ is approximated as follows. For $N, M \in \mathbb{N}$ and step-sizes $h = (d_{max} - d_{min})/N$ and $k = (y_{max} - y_{min})/M$ let

$$\mathbb{G}_{h,k} = \{(d_i, y_j) \in \mathbb{R}^2 \mid d_i = d_{min} + ih, y_j = y_{min} + jk, i = 0, \ldots, N, j = 0, \ldots, M\}$$

denote a grid covering the road region of interest. Then for each grid point $(d_i, y_j) \in \mathbb{G}_{h,k}$ the following optimal control problem is solved:

Minimize

$$\frac{1}{2}(y(t_f)) - y_j)^2$$

subject to the equations of motion (1)-(4) with initial condition

$$(x(0), y(0), \psi(0), v_x(0), v_y(0), w_\psi(0), \delta(0)) = (0, y_0, 0, v_{x,0}, 0, 0, 0),$$

the control constraints $|w_\delta| \le w_{\delta,max}$, $F_B \in [F_{B,min}, F_{B,max}]$, *the pure state constraint*

$$y_{min} \le y(t) \le y_{max},$$

and boundary conditions $x(t_f) = d_i$, $v_y(t_f) = 0$.

Let $x^*_{i,j}(\cdot)$ and $y^*_{i,j}(\cdot)$ denote the optimal solution components of the state vector. The projected reachable set is approximated by collecting all grid points in $\mathbb{G}_{h,k}$ with distance of order $\mathcal{O}(h+k)$ to end points of trajectories:

$$\mathcal{PR}_{h,k} := \bigcup_{\substack{(d_i,y_j)\in\mathbb{G}_{h,k}:\\ \|(x^*_{i,j}(t_f),y^*_{i,j}(t_f))-(d_i,y_j)\|\le C(h+k)}} \{(d_i, y_j)\}.$$

Herein, $C > 0$ is a constant. In [2] it is shown that the approximation $\mathcal{PR}_{h,k}$ convergences in the Hausdorff distance to $\mathcal{PR}$ of order $\mathcal{O}(h+k)$ as h and k approach zero, if $\mathcal{PR}$ is closed and non-empty. Direct discretization techniques for the numerical solution of the optimal control problem introduce a further approximation to $\mathcal{PR}_{h,k}$ whose convergence properties for a special setting are analyzed in [2] as well.

The projected reachable set approximations $\mathcal{PR}_{h,k}$ are depicted in Figure 3 for different initial velocities and the data $y_{min} = 1.3$, $y_{max} = 5.7$, $d_{min} = 10$, $d_{max} = 200$, $y_0 = 1.75$, $F_{B,min} = -5000$, $F_{B,max} = 15000$. The optimal control problems have been solved by the software OCPID-DAE1 [8].

Note that the obstacle car is not taken into account in the optimal control problems. But once the projected reachable set is known, it can be decided for a given obstacle position whether a collision can be avoided or not by investigating the remaining space in the projected reachable set outside the obstacle at the x-position of the obstacle.

4.3 Approach 3: Feasibility Problem for Scenario 2

For a fixed obstacle it is comparatively simple to define a safe target position or to approximate the projected reachable set, but for moving objects as in the overtaking maneuver in Figure 1 it is not, since a collision with all other moving cars has to be avoided at all times. In the overtaking scenario in Figure 1 let car A denote the car that overtakes a car called car B and car C is the car approaching car A in opposite direction. Cars B and C are supposed to drive at constant velocity in a straight line. Anti-collision constraints lead to the following pure state constraints, where W denotes the maximum width of the cars (for simplicity we use balls to model the anti-collision constraints):

$$\begin{aligned}(x_A(t) - x_B(t))^2 + (y_A(t) - y_B(t))^2 &\ge W^2, \quad \text{(don't hit car B)}\\ (x_A(t) - x_C(t))^2 + (y_A(t) - y_C(t))^2 &\ge W^2, \quad \text{(don't hit car C)}\end{aligned}$$

Unfortunately, these constraints will be infeasible if there is no way to avoid a collision. Of course, in this case the resulting optimal control problems do not

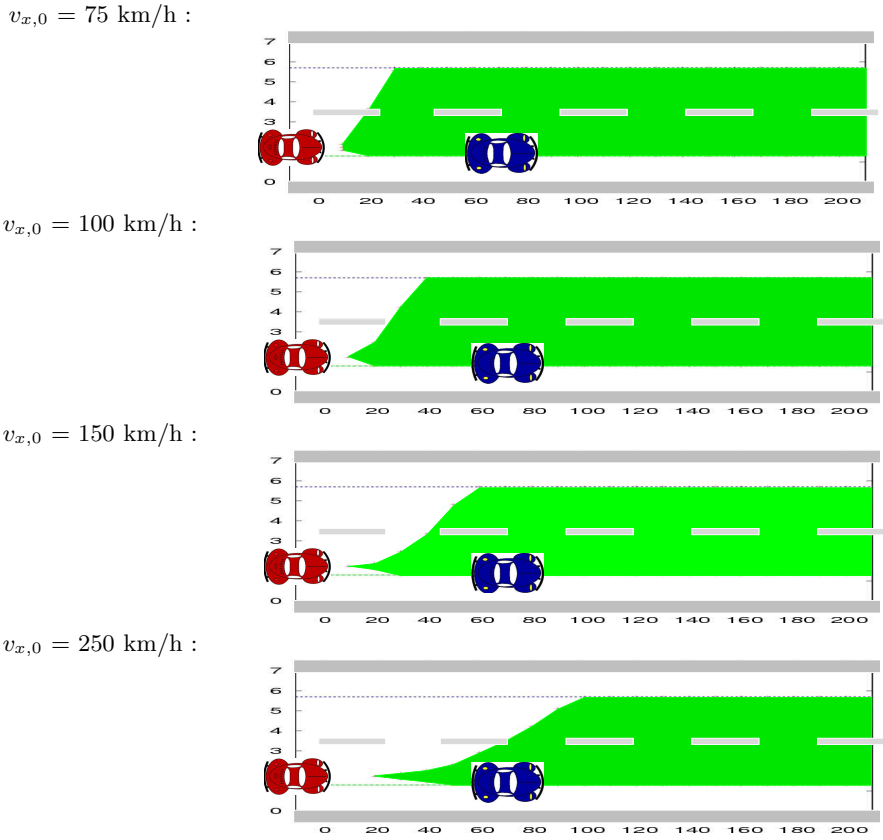


Fig. 3. Projected reachable sets for initial velocities $v_{x,0} = 75, 100, 150, 250$ [km/h]. The approaching car can avoid a collision with the obstacle in the first three settings, but not in the final setting, if the measurements of the obstacle and the approaching car are taken into account.

have a solution and numerical methods will fail. In order to circumvent this problem the relaxed constraints

$$(x_A(t) - x_B(t))^2 + (y_A(t) - y_B(t))^2 + \alpha \geq W^2,$$
$$(x_A(t) - x_C(t))^2 + (y_A(t) - y_C(t))^2 + \alpha \geq W^2$$

are considered, where α denotes the maximal constraint violation. Now, an optimal control problem with the aim to minimize the constraint violation α is solved subject to the above constraints. A collision detection algorithm is then given by considering the minimal constraint violation α^*. If $\alpha^* > 0$, then a collision cannot be avoided (the anti-collision constraints cannot be satisfied). If $\alpha^* \leq 0$, then a collision can be avoided with a trajectory that is produced by the optimal control problem. We illustrate the outcome for the following data:

- car A: 100 [km/h], car B: 75 [km/h], car C: 100 [km/h]
- car width 2.6 [m], road width 7 [m]
- initial y-position of car A : 5.25 [m]
  initial y-position of car B : 1.75 [m]
  initial y-position of car C : 5.25 [m]

Table 1 summarizes the results for different initial distances of cars A and C obtained with OCPID-DAE1 [8]. A movie that visualizes the overtaking maneuver with initial distance of 60 m can be downloaded on the homepage of the first author.

Table 1. Results for the constraint minimization problem for the overtaking maneuver

initial distance [m]	constraint violation α^* [m]	collision
10	0.24780E+01	yes
20	0.22789E+01	yes
30	0.21355E+01	yes
40	0.19351E+01	yes
50	0.94517E-01	yes
60	0.74140E-08	no
⋮	⋮	⋮
190	0.74593E-08	no
200	0.74760E-08	no

5 Sensor Influence

In the collision avoidance scenarios, the initial position, i.e the constellation of evading car and obstacles, is determined by sensor measurements. These sensor measurements are subject to measurement errors and hence it is important to investigate how the optimal solution depends on these sensor measurement errors. We outline this for approach 1 and consider the initial values y_0 and $v_{x,0}$ to be parameters in the optimal control problem $\mathrm{OCP}(y_0, v_{x,0})$. We apply the sensitivity analysis in [3], which exploits the sensitivity results in [4] for finite dimensional optimization problems. To this end, $\mathrm{OCP}(y_0, v_{x,0})$ is discretized using piecewise constant control approximations $w_{\delta,j} \approx w_\delta(t_j)$ and $F_{B,j} \approx F_B(t_j)$, $j = 0, \ldots, K$, and Runge-Kutta approximations for the state on a grid with grid points t_j, $j = 0, \ldots, K$. The discretized problem is a finite dimensional nonlinear optimization problem, which is solved for nominal parameters $\hat{y}_0$ and $\hat{v}_{x,0}$. If the nominal optimal solution satisfies the assumptions of the sensitivity theorem in [4], i.e. second order sufficient conditions, linear independence constraint qualification and strict complementarity, then it was shown that the solution locally depends continuously differentiable on the parameters y_0 and $v_{x,0}$ and the sensitivities of the optimal control discretization with respect to the initial values y_0 and $v_{x,0}$ can be computed, that is we obtain the sensitivities

$$\frac{dw_{\delta,j}}{(y_0, v_{x,0})}(\hat{y}_0, \hat{v}_{x,0}), \quad \frac{dF_{B,j}}{(y_0, v_{x,0})}(\hat{y}_0, \hat{v}_{x,0}), \quad j = 0, \ldots, K.$$

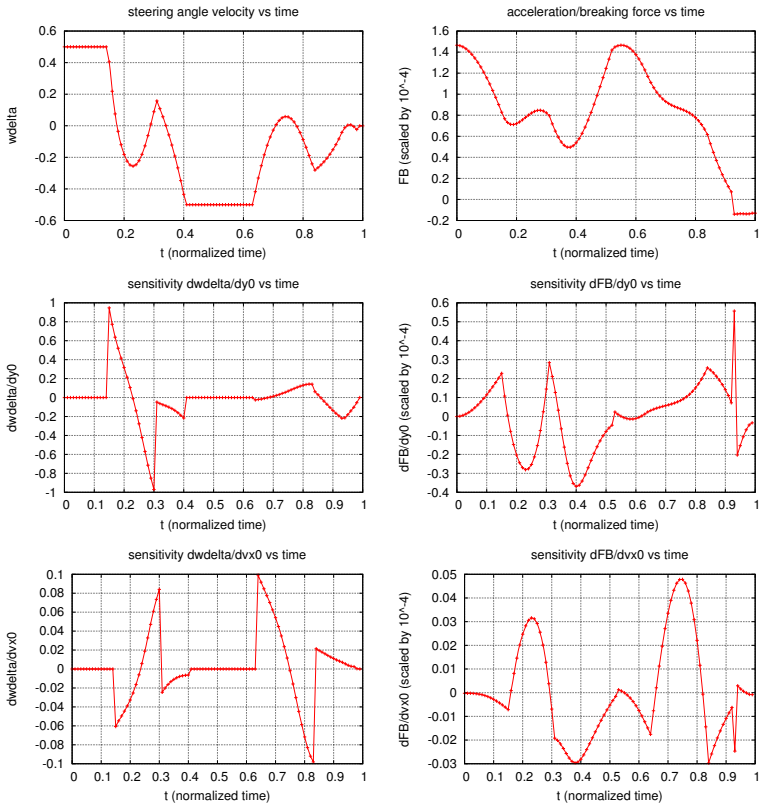


Fig. 4. Sensitivity differentials for the optimal controls of OCP($\hat{y}_0, \hat{v}_{x,0}$) at $\hat{y}_0 = 1.75$ and $\hat{v}_{x,0} = 35$: Control and sensitivities of w_δ w.r.t. y_0 and $v_{x,0}$ (left) and of F_B w.r.t. y_0 and $v_{x,0}$ (right)

The sensitivities indicate how sensitive the solution depends on perturbations in the initial values and hence can help to specify tolerances for sensors. We omit the details here since this sensitivity approach became quite standard in the meanwhile. Details can be found, e.g., in [3].

Figure 4 shows the sensitivity differentials for the controls w_δ and F_B with respect to y_0 and $v_{x,0}$ at the nominal parameters $\hat{y}_0 = 1.75$ and $\hat{v}_{x,0} = 35$. From the pictures it can be concluded that a perturbation in the initial y-position has the highest influence on the controls w_δ and F_B. A perturbation of the initial velocity $v_{x,0}$ has less influence. Hence, the sensor measurement of the y-position (respectively, the offset to the obstacle) should be more accurate than the sensor measurement of the velocity. More elaborate investigations regarding the definition of sensor tolerances that are necessary to achieve a certain performance are currently under investigation.

Please note that the above sensitivity approach does not work for the optimal control problems in approach 2 as those do not satisfy the assumptions of the sensitivity theorem in [4] whenever a grid point is in the projected reachable set. Adding a regularization term in the objective function might help to overcome this difficulty and would allow to investigate the dependence of the projected reachable set on sensor measurements.

6 Outlook

The paper suggests different approaches to an avoidance trajectory system based on optimal control techniques, reachable set computations, and sensitivity analysis. Many extensions are possible, e.g. computation of driver-friendly trajectories for active steering driver assistance systems, more complicated road geometries, real-time approximations, investigation of worst-case scenarios or cooperative control in the presence of many moving objects, and the investigation of parameter dependence of the projected reachable set. These issues are currently under investigation.

References

1. Baier, R., Gerdts, M.: A computational method for non-convex reachable sets using optimal control. In: Proceedings of the European Control Conference (ECC 2009), Budapest, Hungary, August 23-26, pp. 97–102 (2009)
2. Baier, R., Gerdts, M., Xausa, I.: Approximation of Reachable Sets using Optimal Control Algorithms. Technical report (2011) (submitted)
3. Büskens, C., Maurer, H.: Sensitivity Analysis and Real-Time Control of Parametric Optimal Control Problems Using Nonlinear Programming Methods. In: Grötschel, M., Krumke, S.O., Rambau, J. (eds.) Online Optimization of Large Scale Systems, pp. 56–68. Springer (2001)
4. Fiacco, A.V.: Introduction to Sensitivity and Stability Analysis in Nonlinear Programming. Mathematics in Science and Engineering, vol. 165. Academic Press, New York (1983)
5. Gerdts, M.: Solving Mixed-Integer Optimal Control Problems by Branch&Bound: A Case Study from Automobile Test-Driving with Gear Shift. Optimal Control, Applications and Methods 26(1), 1–18 (2005)
6. Gerdts, M.: A variable time transformation method for mixed-integer optimal control problems. Optimal Control, Applications and Methods 27(3), 169–182 (2006)
7. Gerdts, M., Karrenberg, S., Müller-Beßler, B., Stock, G.: Generating Optimal Trajectories for an Automatically Driven Car. Optimization and Engineering 10(4), 439–463 (2009)
8. Gerdts, M.: OCPID-DAE1 – Optimal Control and Parameter Identification with Differential-Algebraic Equations of Index 1: User's Guide. Technical report, Institut für Mathematik und Rechneranwendung, Universität der Bundeswehr München (2011)
9. Statistisches Bundesamt: Unfallentwicklung auf deutschen Strassen 2008, Statistisches Bundesamt, Wiesbaden (2009), http://www.destatis.de

Theoretical Analysis and Optimization of Nonlinear ODE Systems for Marine Ecosystem Models

Anna Heinle and Thomas Slawig*

Institute for Computer Science
Cluster "The Future Ocean", Christian-Albrechts Universität zu Kiel, Germany
{ahe,ts}@informatik.uni-kiel.de

Abstract. We present the investigation of a biogeochemical marine ecosystem model used as part of the climate change research focusing on the enhanced carbon dioxid concentration in the atmosphere. Numerical parameter optimization has been performed to improve represention of observational data using data assimilation techniques. Several local minima were found but no global optimum could be identified. To detect the actual capability of the model in simulating natural systems, a theoretical analysis of the model equations is conducted. Here, basic properties such as continuity and positivity of the model equations are investigated.

Keywords: Climate models, Marine ecosystem models, Parameter optimization, Ordinary differential equations.

1 Introduction

An important part of climate change research is the investigation of biogeochemical processes occurring in the oceans. The Earth's carbon cycle, one of the main climate drivers, is highly dependent on the marine ecosystem and its interactions due to primary carbon producers such as phytoplankton. For this reason, scenarios of the climate's future are commonly created from numerical models, including a submodel to simulate the marine ecosystem. These submodels range from conceptually simple models, like the FDM model by Fasham [1], to highly complex models simulating numerous components of the marine ecosystem such as different types of plankton or multiple nutrients (see e.g. [2,3,4]).

In this work, we consider a model of mid-complexity used at GEOMAR, Kiel studying the CO_2 uptake of the ocean. Four components of the marine ecosystem, namely nitrogen (N), phytoplankton (P), zooplankton (Z) and detritus (D) are simulated in this model. The model, hereafter called NPZD model, depicts the main interactions and feedbacks of the marine ecosystem. However, its feasibility to reproduce real observed data is limited (see e.g. [5,6]). Currently, the reason for this fact is unclear and is the motivation for this study.

* In cooperation with Andreas Oschlies, GEOMAR, Kiel, Germany.

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 501–510, 2013.

We introduce two methods used. Numerical parameter optimization is conducted to assess the influence of different parameter combinations on the model outcome and to detect whether optimal parameters exist reproducing a given observational data set. Further, the theoretical framework of the model is analysed. The knowledge of basic properties of the model as well as the demonstration of particular dependencies of model parameters may yield a better understanding and assessment of the model dynamics.

The paper is structured as follows: In Section 2, we introduce the ecosystem model considered. Biological definitions as well as a mathematical formulation of the model are provided. In Section 3 we show an extract of the numerical experiments conducted focusing on the variability of model outcomes. Theoretical findings regarding the mathematical framework of the NPZD model are subsequently presented in Section 4. A final discussion and an outlook of future work conclude the paper in Section 5.

2 Model

We study the behaviour of a marine ecosystem model of NPZD type developed by Oschlies and Garçon [7]. This model simulates the concentration (in mmol Nm^{-3}) of four geobiochemical components (also named *tracers* in the following) in the atlantic ocean, namely dissolved inorganic nitrogen (N), phytoplankton (P), zooplankton (Z) and detritus (D). The interactions of the four tracers among each other control the dynamics of the model (see below and Fig. 1).

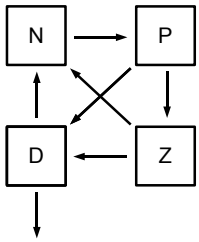


Fig. 1. Scheme of the coupling between the model variables N, P, Z and D. Arrows indicate a nitrogen flux from one to another component.

Aside from light, phytoplankton needs nutrients to photosynthesize and to grow. In the model, these nutrients are represented by the nitrogen component N. Zooplankton graze on phytoplankton and fecal particles of zooplankton as well as other organic, sinking material are summarized in the component detritus. A part of the detritus sinks down to the bottom of the system, another part is recycled by bacteria and comes back to the nitrogen component.

2.1 Mathematical Formulation

The model is directly defined by the interactions of the four tracers N, P, Z and D, in the following denoted by $\mathbf{y} = (y_i)_{i=1,\dots,4} = (N, P, Z, D)$.

A system of coupled, nonlinear ordinary differential equations (ODEs), in detail given by

$$\begin{cases} \frac{\partial y_1}{\partial t} &= -J(\mu, U)y_2 + \Phi_m^Z y_3 + \gamma_m y_4, \\ \frac{\partial y_2}{\partial t} &= (J(\mu, U) - \Phi_m^P)y_2 - G(\epsilon, g)y_3, \\ \frac{\partial y_3}{\partial t} &= (\beta G(\epsilon, g) - \Phi_m^Z - \Phi_Z^* y_3)y_3, \\ \frac{\partial y_4}{\partial t} &= \Phi_m^P y_2 + ((1-\beta)G(\epsilon, g) + \Phi_Z^* y_3)y_3 - (\gamma_m + w_s)y_4 \end{cases}, \tag{1}$$

describes these interactions. J and G are nonlinear functions representing the growth of phytoplankton respectively the grazing of zooplankton on phytoplankton. Parameters appearing in the equations as well as in functions J and G are nonnegative and signify for example growth and mortality rates. Table 1 gives a summary of all parameters. Characteristic, environmental conditions are taken into account by an additional extrinsic physical forcing.

We want to note, that due to our focus on the model specific representation of the interactions of the four tracers this forcing is not further regarded in this work.

Table 1. Model parameters. Units and definitions of the parameters are given together with their value resp. the biological, significant range. Parameters with an index in the first column are included in the optimization processes.

Index	Symbol	Value	Unit	Defintion
	C_{ref}	1.066	1	Growth coefficient of phytoplankton
	c	1	$^\circ$C	Growth coefficient of phytoplankton
	k_{water}	25	m^{-1}	PAR extinction length
	f_{PAR}	0.43	1	PAR fraction of insolation
1	β	[0,1]	1	Assimilation effiency of zooplankton
2	ν_m	$\mathbb{R}_+$	d^{-1}	Phytoplankton growth rate
3	α	$\mathbb{R}_+$	$m^2\ W^{-1} d^{-1}$	Slope of photosynthesis vs light intensity
4	Φ_m^Z	[0,1]	d^{-1}	Zooplankton linear loss rate
5	k_P	[0, 1]	$m^2\ (mmol\ N)^{-1}$	Light attenuation by phytoplankton
6	ϵ	$\mathbb{R}_+$	$m^6\ (mmol\ N)^{-2}\ d^{-1}$	Grazing encounter rate
7	g	$\mathbb{R}_+$	d^{-1}	Maximum grazing rate
8	Φ_m^P	[0,1]	d^{-1}	Phytoplankton linear mortality
9	Φ_Z^*	$\mathbb{R}_+$	$m^3\ (mmol\ N)^{-1}\ d^{-1}$	Zooplankton quadratic mortality
10	γ	[0,1]	d^{-1}	Detritus remineralization rate
11	k_N	$\mathbb{R}_+$	$mmol\ N\ m^{-3}$	Half saturation for NO3 uptake
12	w_s	[0,1]	$m\ d^{-1}$	Detritus sinking

Light and Nutrient Limited Growth Rate of Phytoplankton

Biologically, the growth of phytoplankton, in the model represented by J, is limited by two factors, light and nutrients. Here, this limitation is modeled using

Liebig's law of the minimum (see Eq. (2)). The temperature weighted maximum growth rate $V_p(t) = \mu_m C^{T(t)}$ is multiplied by the minimum of two functions, function μ, describing the light limitation (Eq. (3)) and function U, mirroring the nutrient limitation (Eq. (4)).

$$J(\mu(I, y_2, t), U(y_1, t)) = V_p(t) \min\{\mu(I, y_2, t), U(y_1, t)\} \ . \tag{2}$$

The function μ, given by

$$\mu(I, y_2, t) = \frac{1}{k_{to}(t) z} \ln\left(\frac{I_0(t) + \sqrt{V_p(t)^2 + I_0(t)^2}}{I_z(t) + \sqrt{V_p(t)^2 + I_z(t)^2}}\right) \ , \tag{3}$$

$$I_0(t) = \alpha I_{in}(t) \ ,$$

$$I_z(t) = I_0(t) e^{-k_{to} z(t)} \ ,$$

$$k_{to}(t) = k_w(t) + k_P(t) y_2 \ ,$$

is based on the Smith's curve [8] and describes the relationship between photosynthesis and light for phytoplankton. The motivation for this function is to enable the integration of variable insolation, for example due to seasonal influences or in case a daily cycle is to be simulated.

The nutrient limitation U is described by a so called Holling type II function,

$$U(y_1, t) = \frac{y_1}{k_N + y_1} \ . \tag{4}$$

Holling type functions describe the reproduction of a consumer as a function of food density, here as a function of nutrients y_1. Obviously, U is monotonically increasing and ranges in the interval $[0, 1]$ for $k_N > 0$ and $y_1 \in \mathbb{R}^0_+$.

Zooplankton Growth Rate

The growth of zooplankton is dependent on the phytoplankton availability and is given by a Holling type III function,

$$G(\epsilon, g, y_2) = \frac{g \epsilon y_2^2}{g + \epsilon y_2^2} \ . \tag{5}$$

Such as U, G is monotonically increasing, but ranges in the interval $[0, g]$ for $g, \epsilon > 0$ and $y_2 \in \mathbb{R}^0_+$.

3 Numerical Optimization

We show an extract of the experiments conducted in the course of the parameter optimization of the ecosystem model introduced in Section 2. Two observational

data sets are used to optimize the parameters of the model by data assimilation techniques (see below). In both cases, a cost function of least-squares type,

$$\min_{u \in \mathbb{R}_+^k} J(y(u)) := ||y(u) - y_d||^2 \text{ s.t. } lb \leq u \leq ub \ , \tag{6}$$

is applied. Reference data is denoted by y_d, the model output is given by $y(u)$ and $u \in R_+^k$ is the parameter vector to be optimized, componentwise within the constraints lb and ub.

3.1 Data

The following two data sets are used for parameter optimization of the NPZD model.

D1. The Bermuda Atlantic Time Series (BATS): Vertical profiles obtained during the US JGOFS project. This data is frequently used within optimization approaches of marine ecosystem models, among others the one dimensional version of the NPZD model (see e.g. [9,6,12]). For this study, each profile is vertically averaged from the surface to the respective "mixed layer depth" which is fixed by a temperature decline of more than 0.5 °C (for details see [10]).

D2. Indoor-mesocosm data obtained during the AQUASHIFT project: A 30 day mesocosm experiment performed at GEOMAR, Kiel investigating the impact of ocean warming on the phytoplankton spring bloom in the North Atlantic (for details see [11]).

Since parameter optimization with respect to D1 is subject of [12], we here focus on the optimization using data set D2.

3.2 Results with Respect to D2

In the following we show a cut-out of the numerical optimization experiments. Note, that we show individual examples instead of statistics to highlight specific issues. Table 2 and Figure 2 demonstrate the impact of varying upper bounds of the parameters (E1), Table 3 and Figure 3 provide the results of an initial value experiment (E2).

All results shown are obtained operating an optimization algorithm of SQP (sequential quadratic programming) type. Initial values are taken from literature as well as chosen randomly. The optimization runs include the indexed parameters presented in Table 1.

E1. In the first experiment, optimal parameter values found by Rückelt et al. [6] are used to initialize the model. Table 2 shows the parameters to be optimized, their initial values u^{ini}, optimized values u^* and the bounds lb and ub for two examples. Replacing the upper bounds as given in [6] by a vector of ones[1], the cost function can be decreased less ($J = 3.8091$) in contrast to

[1] Consequently, the initial values of some parameters are modified as well.

Table 2. Variation of the upper bounds. Initial values and upper bounds of the parameters according to [6] *(left)*, resp. modified *(right)*.

	lb	ub	u^{ini}	u^*	ub_1	$\tilde{u}^{ini}$	$\tilde{u}^*$
Φ_P	0.0001	1	0.001	0.0021	1	0.001	0.0205
Φ_Z^*	0.0001	10	0.202	0.1631	1	0.202	0.4179
γ	0.0001	1	0.092	0.0044	1	0.092	0.9051
β	0.0001	1	1.000	0.9484	1	0.8	0.9365
Φ_Z	0.0001	1	0.025	0.0087	1	0.025	0.4074
g	0.0001	100	20.00	20.457	1	0.8	0.9761
ϵ	0.0001	100	5.446	4.6083	1	0.8	0.3319
ν_m	0.0001	10	1.076	1.1482	1	0.8	0.1696
k_N	0.0001	10	1.827	1.1982	1	0.8	0.5867
w_s	0.0001	1	0.230	0.0085	1	0.8	0.3761
α	0.0001	100	0.107	0.0217	1	0.107	0.2462
k_P	0.0001	1	0.026	0.8503	1	0.026	0.0130
Cost				**3.4240**			**3.8091**

final costs using the original setting ($J = 3.4240$). One might assume that narrowing the constraints may yield to worse optimization results. However, the values do not differ so much, and, as visible in Figure 2, both parameter sets result in an unsatisfying model outcome.

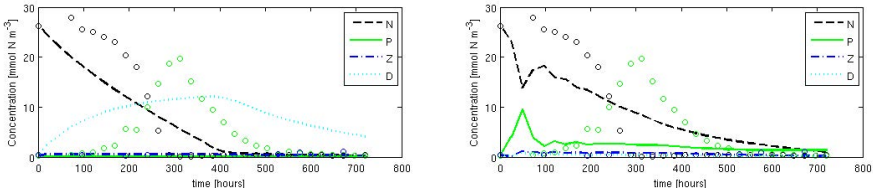


Fig. 2. Simulations of N, P, Z and D for a 30 day period, according to the length of the mesocosm study, using parameter values u^* *(left)* and $\tilde{u}^*$ *(right)* as presented in Table 2. *Circles* depict data points.

E2. The second experiment addresses the influence of initial values on the optimized parameter vector. Table 3 displays the results of four, randomly initialized optimization runs, where $u_{1,2}^{ini} \in [0,1]^{12}$ and $u_{3,4}^{ini}$ are within the constraints (see Tab. 3). While the optimization started at u_1^{ini} reaches costs J near 1, the optimization process initialized with u_2^{ini} gets stuck, just reaching a cost function of $J \simeq 6$. Similar results are obtained for u_3^{ini} and u_4^{ini} indicating the independence of the optimization on the initial values. Further, although the cost function reaches values in the same ranges, the outcomes as well as the final parameter sets $u_i^*, i = 1, \ldots, 4$ differ noticeably and we deduce the existence of multiple distinctive, local minima (see Tab. 3 and Fig. 3).

These two experiments demonstrate the general challenge of finding an optimal parameter set in the context of ecosystem modeling. The experiments performed

Table 3. Impact of different, random initial values u^{ini}. Upper and lower bounds are fix ($lb = (0.0001)^{12}$ and $ub_i = 1.0$ for $i = 1, 3, 4, 5, 12$, $ub_j = 10.0,\ j \neq i$). u^* are the optimized values.

	u_1^{ini}	u_1^*	u_2^{ini}	u_2^*	u_3^{ini}	u_3^*	u_4^{ini}	u_4^*
Φ_P	0.8147	0.0585	0.6787	0.5547	0.1146	0.0036	0.4551	0.0005
Φ_Z^*	0.9058	0.0660	0.7577	1.7077	3.5730	8.8814	3.7551	3.3658
γ	0.1270	0.1883	0.7431	0.9756	0.0947	0.6833	0.0095	0.1377
β	0.9134	0.7945	0.3922	0.4261	0.0942	0.2343	0.2697	0.1882
Φ_Z	0.6324	0.0035	0.6555	0.2604	0.0476	0.5000	0.1297	0.9283
g	0.0975	0.0361	0.1712	3.4484	7.0978	3.4276	6.3543	7.9393
ϵ	0.2785	0.7821	0.7060	1.1090	0.9482	1.3885	2.4897	0.2272
ν_m	0.5469	0.1146	0.0318	0.7970	0.6744	0.0297	4.2283	4.6502
k_N	0.9575	3.0715	0.2769	1.8394	6.6350	6.5590	1.3252	2.2505
w_s	0.9649	0.0458	0.0462	0.3040	9.7396	9.2367	4.8159	5.7531
α	0.1576	9.9534	0.0971	0.1138	5.8376	5.8376	2.1038	2.2880
k_P	0.9706	0.6453	0.8235	0.9280	0.0162	0.0162	0.5233	0.5997
Cost		**1.109**		**5.925**		**1.390**		**5.670**

for the NPZD model indicate the existence of numerous local minima which make the numerical detection of a global minimum - as far as it exists - virtually impossible. Thus, we proceed with a second approach to gain information on the feasibility of the model.

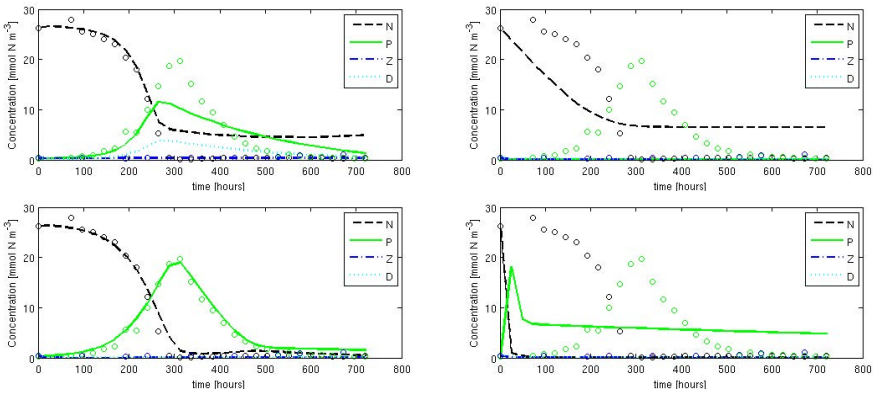


Fig. 3. Simulations created by u_1^* *(top left)*, u_2^* *(top right)*, u_3^* *(bottom left)* and u_4^* *(bottom right)*. Notations are the same as in Figure 1.

4 Theoretical Analysis

In the following, we consider the initial value problem (IVP)

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y}),\ \mathbf{y}(t_0) = \mathbf{y_0} > 0 \text{ and } t \in I = [t_0, t_0 + a], \tag{7}$$

where $\mathbf{f}(t, \mathbf{y})$ is given by the right-hand side of

$$\begin{cases} \frac{\partial y_1}{\partial t} &= -J(U)y_2 + \Phi_m^Z y_3 + \gamma_m y_4, \\ \frac{\partial y_2}{\partial t} &= (J(U) - \Phi_m^P)y_2 - G(\epsilon, g)y_3, \\ \frac{\partial y_3}{\partial t} &= (\beta G(\epsilon, g) - \Phi_m^Z - \Phi_Z^* y_3)y_3, \\ \frac{\partial y_4}{\partial t} &= \Phi_m^P y_2 + ((1-\beta)G(\epsilon, g) + \Phi_Z^* y_3)y_3 - \gamma_m y_4 \end{cases} . \tag{8}$$

In contrast to system (1), we here assume that the ODE system is mass conserving (equivalent to $w_s = 0$ in system (1)) and the limitation of phytoplankton growth due to light is ignored.
To reveal fundamental characteristics of the model dynamics, basic properties of the model equations have to be checked at first. From now on, the differential equations $\frac{\partial y_i}{\partial t}$ are denoted by $f_i(\mathbf{y}),\ i = 1, \ldots, 4$ and the model parameters are summarized in the vector $\mathbf{u} = (u_i)_{i=1,..,12}$.

Remark 1. *The equations $f_i, i = 1, \ldots, 4$ in (8) are continuous in $\mathbf{y}$ and $\mathbf{u}$ for positive $\mathbf{y}$ resp. $\mathbf{u}$.*

This is obvious since the equations are compositions of continuous functions in both, $\mathbf{y}$ and $\mathbf{u}$, that especially holds for the functions J and G.

Next, we consider the range of $\mathbf{f}$. Concentrations, as simulated in our model, are assumed to be nonnegative. To ensure this, we check an important preliminary to investigate this property. The following definition is used.

Definition 1. A function $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ is called *quasipositive* if for $\mathbf{y} \in \mathbb{R}^n$ with $y_i \geq 0,\ i = 1, \ldots, n$, for all $k \in 1, ..., n,\ t \geq t_0$ and $y_k = 0$ hold $f_k(t, \mathbf{y}) \geq 0$.

Proposition 1. *The equations $f_i, i = 1, \ldots, 4$ in (8) are quasipositive given any parameter vector $\mathbf{u} \geq 0$.*

Proof. Let $\mathbf{u} \geq 0$ and $\mathbf{y} = (y_i)_{i=1,...,4} \geq 0$. Then

$$\begin{aligned} f_1(\mathbf{y}) &= \Phi_m^Z y_3 + \gamma_m y_4 \geq 0 \text{ for } y_1 = 0, \\ f_2(\mathbf{y}) &= 0 \text{ for } y_2 = 0, \\ f_3(\mathbf{y}) &= 0 \text{ for } y_3 = 0, \\ f_4(\mathbf{y}) &= \Phi_m^P y_2 + ((1-\beta)G(\epsilon, g) + \Phi_Z^* y_3)y_3 \geq 0 \text{ for } y_4 = 0\ . \end{aligned}$$

□

To investigate the existence and uniqueness of solutions of an IVP, Lipschitz continuity plays a main role. We want to recall the following, well-known Lemma.

Lemma 1. *Let $\mathcal{D}$ be a domain in $\mathbb{R}^n$ and $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$. If f and its partial derivatives $\frac{\partial f_i}{\partial y_j}$ are continuous in $\mathcal{D}$, then f is locally Lipschitz continuous in $\mathcal{D}$.*

Now, for parameters in the range $[0,1]$, the local Lipschitz continuity of functions $f_i, i = 1,\dots,4$ can be shown providing a good base for further analyses regarding the uniqueness of solutions.

Proposition 2. *Let $\mathcal{D} \subset \mathbb{R}^4_+$ be a domain and $\mathbf{u} \in [0,1]^{12}$. Then, the equations $f_i, i = 1,\dots,4$ in (8) are locally Lipschitz continuous in $\mathbf{y}$.*

Proof. As noted before, the functions f_i are continuous for all i. We do not show the calculations in detail, but $\frac{\partial f_i}{\partial y_j} \in \mathcal{C}^1(\mathcal{D})$ for $i,j = 1,\dots,4$. Hence, $f \in \mathcal{C}^1(\mathcal{D})$ and the preliminaries of corollary 1 are fulfilled. The local Lipschitz continuity of f follows. □

Remark 2. The global Lipschitz continuity of $f_i, i = 1,\dots,4$ is not necessarily given. For demonstration, we here give a short counterexample.

Example: Let $L > 0$ and a parameter vector $\mathbf{u} \in [0,1]$. We distinguish 2 cases.

1. $\beta g \le \beta g^2 + 2\Phi_Z^* + \Phi_Z$.
   Choose $\mathbf{y} := (y_1, \sqrt{\frac{1-g}{\epsilon}}, 1, y_4)$ and $\bar{\mathbf{y}} := (y_1, \sqrt{\frac{1-g}{\epsilon}}, 1 + \frac{L+\delta}{\Phi_Z^*}, y_4)$ for a $\delta > 0$. Then

$$||f_3(\mathbf{y}) - f_3(\bar{\mathbf{y}})|| = (L+\delta)||\frac{\Phi_Z}{\Phi_Z^*} + 2 + \frac{L+\delta}{\Phi_Z^*} - \frac{\beta g(1-g)}{\Phi_Z^*}||$$
$$> L\frac{L+\delta}{\Phi_Z^*} = L||\mathbf{y} - \bar{\mathbf{y}}|| \ .$$

2. $\beta g > \beta g^2 + 2\Phi_Z^* + \Phi_Z$.
   Choose $\mathbf{y} := (y_1, \sqrt{\frac{1-g}{\epsilon}}, 1, y_4)$ and $\bar{\mathbf{y}} := (y_1, \sqrt{\frac{1-g}{\epsilon}}, 1 - \frac{L}{\Phi_Z^*}, y_4)$. Then

$$||f_3(\mathbf{y}) - f_3(\bar{\mathbf{y}})|| = L||\frac{\Phi_Z}{\Phi_Z^*} + 2 + \frac{L}{\Phi_Z^*} - \frac{\beta g(1-g)}{\Phi_Z^*}||$$
$$> L\frac{L}{\Phi_Z^*} = L||\mathbf{y} - \bar{\mathbf{y}}||$$

5 Conclusions and Outlook

An extract of the numerical and theoretical investigation of a marine ecosystem model is presented. Parameter optimization and numerical experiments are conducted to get an insight of the model variability. Two examples are shown addressing the influence of different parameter settings. We demonstrate that the model is highly variable and numerous local minima exist, making the determination of a global optimum by numerical methods virtually impossible.

The theoretical part of this study focus on basic properties of the model equations. Continuity and quasipositivity of the equations are established and conditions for local Lipschitz continuity are presented. These first findings facilitate the ongoing investigation of theoretical characteristics of the model making the model and its application in climate research more meaningful.

References

1. Fasham, M.J.R., Ducklow, H.W., McKelvie, S.M.: A nitrogen-based model of plankton dynamics in the oceanic mixed layer. J. Mar. Res. 99, 591–639 (1990)
2. Bissett, W.P., Walsh, J.J., Dieterle, D.A., Carder, K.L.: Carbon cycling in the upper waters of the Sargasso Sea: I. Numerical simulation of differential carbon and nitrogen fluxes. Deep-Sea Res. I 46, 205–269 (1999)
3. Moore, J.K., Doney, S.C., Kleypas, J.A., Glover, D.M., Fung, I.Y.: An intermediate complexity marine ecosystem model for the global domain. Deep-Sea Res. II 49, 403–462 (2002)
4. Lancelot, C., Spitz, Y.H., Gypens, N., Ruddick, K., Becquevort, S., Rousseau, V., Billen, G.: Modelling diatom-Phaeocystis blooms and nutrient cycles in the Southern Bight of the North Sea: the MIRO model. Mar. Ecol. Prog. Ser. 289, 63–78 (2005)
5. Schartau, M., Oschlies, A.: Simultaneous data-based optimization of a 1d-ecosystem model at three locations in the north atlantic: Part I - method and parameter estimates. J. Mar. Res. 61, 765–793 (2003)
6. Rückelt, J., Sauerland, V., Slawig, T., Srivastav, A., Ward, B., Patvardhan, C.: Parameter optimization and uncertainty analysis in a model of oceanic CO2-uptake using a hybrid algorithm and algorithmic differentiation. Nonlinear Anal. Real, Online (2010)
7. Oschlies, A., Garçon, V.: An eddy-permitting coupled physical-biological model of the north atlantic. 1. Sensitivity to advection numerics and mixed layer physics. Global Biogeochem. Cy. 13, 135–160 (1999)
8. Jassby, A.D., Platt, T.: Mathematical formulation of the relationship between photosynthesis and light for phytoplankton. Limnol. Oceanogr. 21, 540–547 (1976)
9. Spitz, Y.H., Moisan, J.R., Abbott, M.R.: Configuring an ecosystem model using data from the Bermuda Atlantic time series (BATS). Deep-Sea Res. II 48, 1733–1768 (2001)
10. Fasham, M.J.R., Evans, G.T.: The use of optimisation techniques to model marine ecosystem dynamics at the JGOFS station at 473N 203W. Philos. T. Roy. Soc. B 348, 206–209 (1995)
11. Sommer, U., Lengfellner, K.: Climate change and the timing, magnitude, and composition of the phytoplankton spring bloom. Glob. Change Biol. 14, 1199–1208 (2008)
12. El Jarbi, M., Slawig, T., Oschlies, A.: Introducing Periodic Parameters in a Marine Ecosystem Model Using Discrete Linear Quadratic Control. In: Hömberg, D., Tröltzsch, F. (eds.) CSMO 2011. IFIP AICT, vol. 391, pp. 485–494. Springer, Heidelberg (2013)

Solving Electric Market Quadratic Problems by Branch and Fix Coordination Methods*

F.-Javier Heredia[1], Cristina Corchero[1], and Eugenio Mijangos[2]

[1] Department of Statistics and Operations Research
Universitat Politècnica de Catalunya UPC
{f.javier.heredia,cristina.corchero}@upc.edu
[2] Department of Applied Mathematics,
Statistics and Operations Research
University of the Basque Country UPV/EHU
eugenio.mijangos@ehu.es

Abstract. The electric market regulation in Spain (MIBEL) establishes the rules for bilateral and futures contracts in the day-ahead optimal bid problem. Our model allows a price-taker generation company to decide the unit commitment of the thermal units, the economic dispatch of the bilateral and futures contracts between the thermal units and the optimal sale bids for the thermal units observing the MIBEL regulation. The uncertainty of the spot prices is represented through scenario sets. We solve this model on the framework of the Branch and Fix Coordination metodology as a quadratic two-stage stochastic problem. In order to gain computational efficiency, we use scenario clusters and propose to use perspective cuts. Numerical results are reported.

Keywords: Liberalized Electricity Market, Optimal Bid, Stochastic Programming, Quadratic Branch-and-Fix Coordination.

1 Introduction

This work is applied to the Iberian Electricity Market (MIBEL) comprising the Spanish and Portuguese electricity systems. The MIBEL market includes in the short-term: the day-ahead market (DAM) and a set of balancing, reserve and adjustment markets (intraday markets); these markets are complemented with the medium- and long-term mechanisms: a derivatives market and different kinds of bilateral contracts. This structure is similar to other European electricity markets and explains why generation companies can no longer optimize their short-term generation planning decisions, i.e. their bidding strategies, without considering the relationship between the short-term bid and the medium-term physical products. The MIBEL's directives dictate specific rules describing how these medium-term mechanisms should be included into the DAM bid. This work deals with the most relevant medium-term mechanisms in the MIBEL:

* This work was partially supported by the Ministry of Science and Technology of Spain through MICINN Project DPI2008-02153.

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 511–520, 2013.

the national bilateral contracts (BC) and the future physical contracts (FC). Stochastic programming techniques are applied to maximize the expected value of the utility's profit coming from the day-ahead market, where the significative random variable is the auction clearing price of the day-ahead electricity market. This random variable is modeled through a set of scenarios of the forecasted prices. The set of scenarios is used to feed a two-stage stochastic optimization model that finds the optimal day-ahead bid of a price-taker GenCo (an electrical utility without influence over the market prices) operating in the MIBEL and holding bilateral and physical futures contracts.

The extensive form of the deterministic equivalent of this stochastic programming problem will be a mixed integer quadratic programming problem (MIQP), which is difficult to solve efficiently, particularly for large-scale instances. Several algorithmic approaches can be adopted to overcome this difficulty. In [2] the quadratic objective function of this problem is approximated by a polyhedral outer approximation by means of *perspective cuts* so that we can exploit the efficiency of general-purpose solvers for mixed integer linear problems (MILP). An alternative to the perspective cuts methodology is the Second-Order Cone Program reformulation (SOCP, [9]), but for quadratic problems the perspective cuts reformulation was reported to be more efficient [6]. Finally, the Branch-and-Fix Coordination (BFC) method has been used successfully to solve two-stage stochastic mixed integer linear problems [3] to solve the day-ahead optimal bid problem. In this work we propose an combination between BFC and PC to efficiently solve the optimal day-ahead bid problem.

2 Day-ahead Electricity Market Bid with Futures and Bilateral Contracts Model (DAMB-FBC)

In this section the model (DAMB-FBC) is formulated as a two-stage stochastic programming problem that allows a price-taker generation company to optimally decide the unit commitment of its thermal units, the economic dispatch of the bilateral and futures contracts between the thermal units, and the optimal generation bid of the committed units to the MIBEL's day-ahead market. The objective function of the model represents the expected profits of the GenCo's participation in the day-Ahead market. The constraints assure that the MIBEL's rules and the operational restrictions of the units are respected. The main decision variables are the ones that model the start-up and shut-down of the units, the quantity that will be bid at instrumental price and the scheduled energy committed to the bilateral and the futures contracts settlement.

2.1 Parameters

The (DAMB-FBC) model considers a price-taker GenCo owning a set of thermal generation units $\mathcal{I}$ that bid to the $t \in \mathcal{T} = \{1, 2, .., 24\}$ hourly auctions of the DAM. The parameters for the i^{th} thermal unit are:

- c_i^b, c_i^l and c_i^q, generation costs with constant, linear and quadratic coefficients (€, €/MWh and €/MWh^2 respectively).
- $\overline{P}_i$ and $\underline{P}_i$, upper and lower bounds on the hourly energy generation (MWh).
- c_i^{on} and c_i^{off}, start-up and shut-down costs (€).
- t_i^{on} and t_i^{off}, minimum operation and minimum idle time (h).

A base load physical futures contract $j \in \mathcal{F}$ is defined by:

- $\mathcal{U}_j$, the set of generation units allowed to cover the FC j.
- L_j^F, the amount of energy (MWh) to be procured each interval of the delivery period by the set $\mathcal{U}_j$ of generation units to cover contract j.
- λ_j^F, the price of contract j (€/MWh).

A base load bilateral contract $k \in \mathcal{B}$ is defined by:

- L_k^B, the amount of energy (MWh) to be procured at each interval of the delivery period by the set of available generation units to cover the BCs.
- λ_k^B, the price of the contract k (€/MWh).

The random variable λ_t^D, the clearing price of the t^{th} hourly auction of the DAM, is represented in the two-stage stochastic model by a set of scenarios $s \in \mathcal{S}$, each one with its associated clearing price for each DAM auction $t \in \mathcal{T}$:

- $\lambda_t^{D,s}$ clearing price for auction t at scenario s (€/MWh).
- P^s probability of scenario s.

2.2 Variables

Those decision variables that doesn't depend on the scenarios are called first stage (or *here-and-now*) variables and in our formulation are, for each $t \in \mathcal{T}$ and $i \in \mathcal{I}$:

- u_{ti}, the unit commitment (binary)
- c_{ti}^u, c_{ti}^d, the start-up/shut-down costs variables.
- q_{ti}, the instrumental price offer bid.
- f_{tij}, the scheduled energy for FC $j \in \mathcal{F}$.
- b_{ti}, the scheduled energy for the pool of BCs .

Decision variables that can adopt different values depending on the scenario are called second stage variables and in our formulation are, for each $t \in \mathcal{T}$, $i \in \mathcal{I}$ and scenario $s \in \mathcal{S}$:

- g_{ti}^s, the total generation.
- p_{ti}^s, the matched energy in the day-ahead market.

2.3 Contraints

Bilateral and Futures Contracts Constraints. The coverage of both the physical futures contracts and the bilateral contracts must be guaranteed. The constraints for each futures contract are:

$$\sum_{i \in \mathcal{U}_j} f_{tij} = L_j^F \qquad t \in \mathcal{T},\ j \in \mathcal{F} \tag{1}$$

$$f_{tij} \geq 0 \qquad t \in \mathcal{T},\ j \in \mathcal{F},\ i \in \mathcal{I} \tag{2}$$

and the bilateral contract constraints are:

$$\sum_{i \in \mathcal{I}} b_{ti} = \sum_{k \in \mathcal{B}} L_k^B \qquad t \in \mathcal{T} \tag{3}$$

$$0 \leq b_{ti} \leq \overline{P}_i u_{ti} \qquad i \in \mathcal{I},\ t \in \mathcal{T} \tag{4}$$

where L_k^B is the energy that has to be settled for contract $k \in \mathcal{B}$

Day-ahead Market and Total Generation Constraints. As we have introduced, we will use the value of the *matched energy* in our formulation. The matched energy is the accepted energy in the clearing process, that is, the energy generated that will be rewarded at the clearing price. This matched energy is uniquely determined by the sale bid and the clearing price and it will play a central role in the presented model [2].

The MIBEL's rules affecting the day-ahead market establishes the relation between the variables representing the matched energy p_{ti}^s, the energy of the bilateral contracts b_{ti}, the energy of the futures contracts f_{tij}, the instrumental price offer bid q_{ti}, and the commitment binary variables u_{ti}. The energies L_j^F and L_k^B must be integrated in the MIBEL's DAM bid observing the two following rules:

1. If generator i contributes with f_{tij} MWh at period t to the coverage of the FC j, then the energy f_{tij} must be offered to the pool for free (*instrumental price bid*).
2. If generator i contributes with b_{ti} MWh at period t to the coverage of any of the BCs, then the remaining production capacity $\overline{P}_i - b_{ti}$ must be bid to the DAM.

These rules can be included in the model by means of the following set of constraints:

$$p_{ti}^s \geq q_{ti} \qquad i \in \mathcal{I},\ t \in \mathcal{T},\ s \in \mathcal{S} \tag{5}$$

$$p_{ti}^s \leq \overline{P}_i u_{ti} - b_{ti} \qquad i \in \mathcal{I},\ t \in \mathcal{T},\ s \in \mathcal{S} \tag{6}$$

$$q_{ti} \geq \underline{P}_i u_{ti} - b_{ti} \qquad i \in \mathcal{I},\ t \in \mathcal{T},\ s \in \mathcal{S} \tag{7}$$

$$q_{ti} \geq \sum_{j \mid i \in \mathcal{U}_j} f_{tij} \qquad i \in \mathcal{I},\ t \in \mathcal{T},\ s \in \mathcal{S} \tag{8}$$

where:

(5) and (6) ensure respectively that the matched energy p^s_{ti} will be greater than the instrumental price bid q_{ti} and less than the total available energy not allocated to BC.
(7) and (8) guarantee respectively that the instrumental price bid will be greater than the minimum generation output of the unit and greater than the contribution of the unit to the FC coverage.

Please note that (2) together with (8) assures that q_{ti} will be always non-negative. The total generation level of a given unit i, g^s_{ti}, is defined as the addition of the allocated energy to the BC plus the matched energy of the DAM:

$$g^s_{ti} = b_{ti} + p^s_{ti} i \in \mathcal{I},\ t \in \mathcal{T},\ s \in \mathcal{S} \tag{9}$$

Contraints (1)-(9) assure that g^s_{ti} will be always either zero or $g^s_{ti} \in [\underline{P}_i, \overline{P}_i]$, that is:

$$\underline{P}_i u_{ti} \leq g^s_{ti} \leq \overline{P}_i u_{ti},\ i \in \mathcal{I},\ t \in \mathcal{T},\ s \in \mathcal{S} \tag{10}$$

Unit Commitment Constraints. This section includes the formulation for the unit commitment of the thermal units [2]. The first two sets of constraints model the start-up and shut-down costs and the next ones control minimum operation and idle time for each unit. First, the start-up and shut-down costs are modeled:

$$c^u_{ti} \geq c^{on}_i [u_{ti} - u_{(t-1)i}] \qquad i \in \mathcal{I},\ t \in \mathcal{T} \setminus \{1\} \tag{11}$$

$$c^d_{ti} \geq c^{off}_i [u_{(t-1)i} - u_{ti}] \qquad i \in \mathcal{I},\ t \in \mathcal{T} \setminus \{1\} \tag{12}$$

$$c^u_{ti}, c^d_{ti} \geq 0 \qquad i \in \mathcal{I},\ t \in \mathcal{T} \tag{13}$$

$$u_{ti} \in \{0, 1\} \qquad i \in \mathcal{I},\ t \in \mathcal{T} \tag{14}$$

The initial state of each thermal unit i can be taken into account through the parameters G_i and H_i that represent, respectively, the number of the initial time periods along which the thermal unit must remain on (G_i) or off (H_i). These conditions are imposed by the following set of constraints:

$$\sum_{j=1}^{G_i} (1 - u_{ji}) = 0 \quad \text{and} \quad \sum_{j=1}^{H_i} u_{ji} = 0, \quad i \in \mathcal{I} \tag{15}$$

Finally, the minimum up and down time, t^{on}_i and t^{off}_i are imposed, up to the periods $|\mathcal{T}| - (t^{on}_i - 1)$ and $|\mathcal{T}| - (t^{off}_i - 1)$, through the following set of constraints:

$$\sum_{n=t}^{t+t^{on}_i-1} u_{in} \geq t^{on}_i [u_{ti} - u_{(t-1)i}] \qquad t = G_i + 1, \ldots, |\mathcal{T}| - t^{on}_i + 1,\ i \in \mathcal{I} \tag{16}$$

$$\sum_{n=t}^{t+t^{off}_i-1} (1 - u_{ni}) \geq t^{off}_i [u_{(t-1)i} - u_{ti}] \quad t = H_i + 1, \ldots, |\mathcal{T}| - t^{off}_i + 1\ i \in \mathcal{I} \tag{17}$$

and for the last $t_i^{on} - 1$ and $t_i^{off} - 1$ time periods:

$$\sum_{n=t}^{|\mathcal{T}|}(u_{ni} - [u_{ti} - u_{(t-1)i}]) \geq 0 \qquad t = |\mathcal{T}| - t_i^{on} + 2, \ldots, |\mathcal{T}|,\ i \in \mathcal{I} \tag{18}$$

$$\sum_{n=t}^{|\mathcal{T}|}(1 - u_{ni} - [u_{(t-1)i} - u_{ti}]) \geq 0 \qquad t = |\mathcal{T}| - t_i^{off} + 2, \ldots, |\mathcal{T}|,\ i \in \mathcal{I} \tag{19}$$

2.4 Objective Function

The quadratic function that gives the long-run expected profits of the GenCo after the participation in the DAM is:

$$\min \mathrm{E}_{\lambda^D}\left[C(u, c^u, c^d, g, p; \lambda^D)\right] = \sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}}\Big(c_{ti}^u + c_{ti}^d + c_i^b u_{ti} + \tag{20}$$

$$+ \sum_{s\in\mathcal{S}} P^s\left[(c_i^l g_{ti}^s + c_i^q (g_{ti}^s)^2) - \lambda_t^{D,s} p_{ti}^s\right]\Big), \tag{21}$$

where the right hand side of (20) is the on/off fixed cost of the unit commitment of the thermal units, deterministic and independent of the realization of the random variable $\lambda_t^{D,s}$ and (21) represents the expected value of the benefits from the DAM. The term between parenthesis corresponds to the expression of the quadratic generation costs associated to the total generation of the unit g_{ti}^s while the last term, $\lambda_t^{D,s} p_{ti}^s$ computes the incomes from the DAM due to a value p_{ti}^s of the matched energy.

Please note that the constant incomes from the BC and FC, i.e. $\sum_{k\in\mathcal{B}} \lambda_k^{BC} L_k^{BC}$ and $\sum_{t\in\mathcal{T},\, j\in\mathcal{J}} (\lambda_j^{FC} - \bar{\lambda}_t^D) L_j^{FC}$, have been dropped from the objective function.

2.5 Model (DAMB-FBC)

The model defined so far can be represented as:

$$(\text{DAMB-FBC}) \begin{cases} \min \mathrm{E}_{\lambda^D}\left[C(u, c^u, c^d, g, p\,; \lambda^D)\right] \\ \text{s.t.} \\ \quad \text{Eq. (1)} - \text{(4)} & \text{BC and FC constraints} \\ \quad \text{Eq. (5)} - \text{(9)} & \text{DAM and total gen. constraints} \\ \quad \text{Eq. (11)} - \text{(19)} & \text{Unit commitment constraints} \end{cases}$$

Model (DAMB-FBC) is the optimization problem associated with the two-stage stochastic programming problem with a set $\mathcal{S}$ of scenarios for the spot price λ_t^D, where $t \in \mathcal{T}$. This optimization problem is a convex MIQP with a well defined global optimal solution.

3 QBFC Method

Model (DAMB-FBC) can be rewritten as the so-called Deterministic Equivalent Model (DEM)

$$\begin{aligned}
\text{minimize } & c^t\delta + \sum_{s\in\mathcal{S}} P^s q^s(x, y^s) \\
\text{subject to: } & l_a \le A\begin{bmatrix}\delta\\ x\end{bmatrix} \le u_a, \\
& l_t^s \le T^s\begin{bmatrix}\delta\\ x\\ y^s\end{bmatrix} \le u_t^s,\ s\in\mathcal{S}, \\
& x \ge 0,\ \underline{y} \le y^s \le \overline{y},\ s\in\mathcal{S}, \\
& \delta \in \{0,1\}^{n_\delta},
\end{aligned}$$

where $\delta = u$, $x = (c^u, c^d)$, $y = (g, p)$, $q(x, y) = b_x^t x + b_y^t y + + y^t Q_{yy} y$, and Q_{yy} being a diagonal matrix.

As is showed by [3] the compact representation (DEM) can be written as a *splitting variable* representation; i.e., δ and x are respectively replaced by δ^s and x^s, for $s \in \mathcal{S}$. So, we have

$$\begin{aligned}
\text{(MIQP)}\quad \text{minimize } & \sum_{s\in\mathcal{S}} P^s(c^t\delta^s + q^s(x^s, y^s)) \\
\text{subject to: } & l_a \le A\begin{bmatrix}\delta^s\\ x^s\end{bmatrix} \le u_a,\ s\in\mathcal{S}, \\
& l_t^s \le T^s\begin{bmatrix}\delta^s\\ x^s\\ y^s\end{bmatrix} \le u_t^s,\ s\in\mathcal{S}, \\
& x^s \ge 0,\ \underline{y} \le y^s \le \overline{y},\quad \delta^s \in \{0,1\}^{n_\delta},\ s\in\mathcal{S}, \\
(\text{NAC}_\delta)\ & \delta^s - \delta^{s'} = 0,\ \forall s, s' \in \mathcal{S}:\ s \ne s', \\
(\text{NAC}_x)\ & x^s - x^{s'} = 0,\ \forall s, s' \in \mathcal{S}:\ s \ne s',
\end{aligned}$$

where NAC_δ and NAC_x are the *nonanticipativity constraints*.

In this method (DEM) is solved by using a Branch-and-Fix-Coordination scheme (BFC) for each scenario $s \in \mathcal{S}$ to fulfill the integrality condition (IC) on the variables δ, so that the NAC_δ are also satisfied when selecting branching nodes and branching variables by the Twin-Node-Families concept (TNF), which was introduced by [1].

A similar approach to that suggested in [3] is used in this work to coordinate the selection of the branching node and branching variable for each scenario-related BF tree, such that the NAC_δ are satisfied when fixing δ^s, for all $s \in \mathcal{S}$, either to 1 or to 0. A *TNF integer set* is a set of integer BF nodes (i.e. they verify IC), one per BF tree, in which the NAC_δ are verified. More details about this metodology can be found in [8].

When the number of scenarios is very hight, in order to gain computational efficiency we can take scenario clusters; i.e., instead a submodel for each scenario $s \in \mathcal{S}$ we can use a submodel (MIQP^p) for each scenario cluster $\mathcal{S}^p \subset \mathcal{S}$ with $p = 1, \ldots, \widehat{p}$, where $\mathcal{S}^p \cap \mathcal{S}^{p'} = \emptyset$, for all $p \neq p'$, and $\cup_{p=1}^{\widehat{p}} \mathcal{S}^p = \mathcal{S}$,

$$(\text{MIQP}^p) \quad \text{minimize} \sum_{s \in \mathcal{S}^p} P^s (c^t \delta^p + q^s(x^p, y^s)), \tag{22}$$

$$\text{subject to :} \; l_a \leq A \begin{bmatrix} \delta^p \\ x^p \end{bmatrix} \leq u_a, \tag{23}$$

$$l_t^s \leq T^s \begin{bmatrix} \delta^p \\ x^p \\ y^s \end{bmatrix} \leq u_t^s, s \in \mathcal{S}^p, \tag{24}$$

$$x^p \geq 0, \; \underline{y} \leq y^s \leq \overline{y}, \; s \in \mathcal{S}^p, \quad \delta^p \in \{0,1\}^{n_\delta}, \tag{25}$$

These submodels are linked by the NACs $\delta^p - \delta^{p'} = 0$ and $x^p - x^{p'} = 0$, for all $p, p' \in \{1, \ldots, \widehat{p}\}$ such that $p \neq p'$.

In order to gain computational efficiency we propose to use perspective cuts (PC) [5,2] to solve the quadratic subproblems in each node of the TNF. Then MIQP^p becomes:

$$\min \sum_{s \in \mathcal{S}^p} P^s \left\{ \left(b_x^t x + x^t Q_{xx} x \right) + \left(\sum_{i=1}^{n} v_i^s \right) \right\}$$

$$\text{s.t.:} \; v_i^s \geq (2q_{ii}^s \underline{y}_i + b_i^s) y_i^s + (c_i - q_{ii}^s \underline{y}_i^2) \delta_i^s, \; i \in \{1, \ldots, n\}, \; s \in \mathcal{S}^p,$$

$$v_i^s \geq (2q_{ii}^s \overline{y}_i + b_i^s) y_i^s + (c_i - q_{ii}^s \overline{y}_i^2) \delta_i^s, \; i \in \{1, \ldots, n\}, \; s \in \mathcal{S}^p,$$

$$\text{Eq. (23)} - \text{(25)}.$$

These methods have been implemented in C++ with the help of Cplex 12.1 to solve only the quadratic subproblems. In this work two algorithmic alternatives have been considered:

- ▷ QBFC: coordination of δ in the TNF of the BF trees for clusters $p \in \{1, \ldots, \widehat{p}\}$ without using PCs.
- ▷ QBFC-PC: coordination of δ in the TNF of the BF trees for clusters $p \in \{1, \ldots, \widehat{p}\}$ using PCs.

For our instances the number of scenarios in each cluster is the same, $|\mathcal{S}^p| = |\mathcal{S}|/\widehat{p}$. Each cluster contains $|\mathcal{S}^p|$ consecutive scenarios, starting from the first one and following in natural order.

4 Numerical Tests

These instances are based on the liberalized electricity market model suggested in [2]. In these problems Q_{xx} is the zero matrix, as a result, when we use perspective cuts the subproblem to solve in each node is linear. The tests have been

performed on HP with Intel(R) Core(TM)2 Quad CPU Q8300 2.50GHz 4 CPU under SUSE Linux Enterprise Desktop 11 (x86_64).

In Table 1 $|\mathcal{S}|$ means the number of scenarios, $|\mathcal{T}|$ the number of periods, "# var" the number of continuous variables, "# var_{PCF}" the number of continuous variables for the PC formulation, "# bin" the number of binary variables, and "# constr" the number of constraints for (DEM).

Table 1. Test problems

Prob.	$\lvert\mathcal{S}\rvert$	$\lvert\mathcal{T}\rvert$	# var	# var_{PCF}	# bin	# constr
P01	10	12	1296	1776	48	1788
P02	20	12	2256	3216	48	3228
P03	30	12	3216	4656	48	4668
P04	40	12	4176	6096	48	6108
P05	50	12	5136	7536	48	7548
P11	10	24	2592	3552	96	3600
P12	20	24	4512	6432	96	6480
P13	30	24	6432	9312	96	9360
P14	40	24	8352	12192	96	12240
P15	50	24	10272	15072	96	15120

For every problem $|\mathcal{F}| = |\mathcal{B}| = 2$ and $|\mathcal{I}| = 4$. If we use the PC formulation, the problem increases the number of variables in $m = |\mathcal{T}| \cdot |\mathcal{I}| \cdot |\mathcal{S}|$ and the number of constraints in $2 \cdot m$.

Table 2. Computational results: CPU-times

Prob.	$\widehat{p}$	QBFC	QBFC-PC	ratio	# PC
P01	2	10.1	3.4	0.34	280
P02	4	18.7	8.7	0.47	825
P03	5	2153.0	39.8	0.02	1685
P04	5	50.0	45.1	0.90	1491
P05	5	113.7	19.5	0.17	1276
P11	2	86.8	27.4	0.32	513
P12	4	469.7	50.3	0.11	1821
P13	5	687.3	176.6	0.26	3454
P14	5	1198.0	276.7	0.23	4239
P15	5	1190.9	246.3	0.21	2592

In Table 2 below the headings QBFC are the times in CPU-seconds used for solving problems with the number of scenario cluster given below the heading $\widehat{p}$ and by solving the quadratic subproblem QP^p for each node using Cplex. Column QBFC-PC gives us the CPU-seconds and indicates that the quadratic subproblems QP^p have been solved by using perspective cuts, which means that instead of solving a quadratic problem QP^p in each node of a TNF for $p \in$

$\{1, 2, \ldots, \widehat{p}\}$, a linear problem is solved. Also, "ratio"$=\frac{\text{QBFC}-\text{PC}}{\text{QBFC}}$ gives us the ratio of CPU-times. Note that the running time with PC is a 30% of the running time without PC (average). The last column, "# PC", means the number of perspective cuts generated in each test.

5 Conclusions

We have presented an Optimal Bidding Model for a price-taker generation company operating both in the MIBEL Derivatives and Day-Ahead Electricity Market (DAMB-FBC). The model developed finds the optimal bid for the spot market, the optimal allocation of the physical futures and bilateral contracts among the thermal units and the unit commitment following in detail the MIBEL rules. The (DAMB-FBC) has been solved both with the standard BFC method and with a PC variation which reduces the running time to a 30% on the average.

References

1. Alonso-Ayuso, A., Escudero, L.F., Ortuño, M.T.: BFC, a branch-and-fix coordination algorithm framework for solving some types of stochastic pure and mixed 0-1 programs. European Journal of Operational Research 151, 503–519 (2003)
2. Corchero, C., Mijangos, E., Heredia, F.J.: A new optimal electricity market bid model solved through perspective cuts. TOP (2011) (published online 2011), doi:10.1007/s11750-011-0240-6
3. Escudero, L.F., Garín, M., Merino, M., Pérez, G.: A general algorithm for solving two-stage stochastic mixed 0-1 first-stage problems. Computers & Operations Research 36, 2590–2600 (2009)
4. Escudero, L.F., Garín, M., Merino, M., Pérez, G.: An algorithmic framework for solving large-scale multistage stochastic mixed 0-1 problems with nonsymmetric scenario trees. Computers & Operations Research 39(5), 1133–1144 (2012)
5. Frangioni, A., Gentile, C.: Perspective cuts for a class of convex 0-1 mixed integer programs. Mathematical Programming 106, 225–236 (2006)
6. Frangioni, A., Gentile, C.: A computational comparison of reformulations of the perspective relaxation: SOCP vs. cutting planes. Operations Research Letters 37, 206–210 (2009)
7. Hull, J.C.: Options, futures and other derivatives, 5th edn. Prentice-Hall International, Englewood Cliffs (2002)
8. Mijangos, E.: An Algorithm for Two-Stage Stochastic Quadratic Problems. In: Hömberg, D., Tröltzsch, F. (eds.) CSMO 2011. IFIP AICT, vol. 391, pp. 181–191. Springer, Heidelberg (2013)
9. Tawarmalani, M., Sahinidis, N.: Semidefinite relaxations of fractional programs via novel convexification techniques. Journal of Global Optimization 20, 137–158 (2001)

Asymptotic Behavior of Nonlinear Transmission Plate Problem

Mykhailo Potomkin

B.Verkin Institute for Low Temperature Physics and Engineering of NASU,
47 Lenin Ave., Kharkov 61103 , Ukraine
mika_potemkin@mail.ru
http://ilt.kharkov.ua

Abstract. We study a nonlinear transmission problem for a plate which consists of thermoelastic and isothermal parts. The problem generates a dynamical system in a suitable Hilbert space. Main result is the proof of asymptotic smoothness of this dynamical system and existence of a compact global attractor in special cases.

Keywords: Transmission problems, long-time behavior, asymptotic smoothness, attractor.

1 Introduction

In this work we deal with a partially thermoelastic plate: one part is of isothermal material, the second one is of material whose structure does not neglect the thermal dissipation. Due to the thermal dissipation, purely thermoelastic plate is exponentially stable in linear case (see, e.g., survey in [1, Chapter 3A]) or possesses a compact global attractor in cases of different kind of nonlinearities (see, e.g., [2]). On the other hand, in the case of purely isothermal plate the energy is constant, thus there could not be any decay to zero point in the linear model and global attractor in the nonlinear model. Here we investigate whether the thermal dissipation on a part of the plate is enough to have any stabilization. Exponential stability of the linear problem of this type was established in [3].

Let Ω_1, Ω_2 and Ω be bounded open sets in $\mathbb{R}^2$, $\Gamma_0 = \overline{\Omega}_1 \cap \overline{\Omega}_2$, $\Gamma_1 = \partial\Omega_1/\Gamma_0$ and $\Gamma_2 = \partial\Omega_2/\Gamma_0$ be smooth surfaces. We also assume that $\Omega = \Omega_1 \cup \Omega_2 \cup \Gamma_0$ and $\overline{\Gamma}_1 \cap \overline{\Gamma}_2 = \emptyset$. In the model under consideration the plate (its middle surface), in equilibrium, occupies the domain Ω which consists of two parts Ω_1 and Ω_2 with common boundary Γ_0. In what follows below ν denotes the outward normal vector on Γ_1 and Γ_2, in cases of common boundary Γ_0 the vector ν is outward to Ω_2.

We consider the following system of equations:

$$\rho_1 u_{tt} + \beta_1 \Delta^2 u + \mu \Delta\theta + F_1(u,v) = 0 \quad \text{in } \Omega_1 \times \mathbb{R}^+, \tag{1}$$

$$\rho_0 \theta_t - \beta_0 \Delta\theta - \mu \Delta u_t = 0 \quad \text{in } \Omega_1 \times \mathbb{R}^+, \tag{2}$$

$$\rho_2 v_{tt} + \beta_2 \Delta^2 v + F_2(u,v) = 0 \quad \text{in } \Omega_2 \times \mathbb{R}^+. \tag{3}$$

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 521–527, 2013.

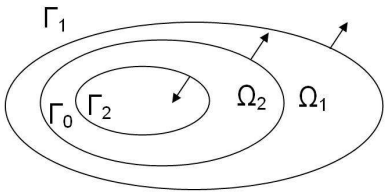


Fig. 1.

We impose the following boundary conditions:

$$u = \frac{\partial u}{\partial \nu} = 0 \ \text{ on } \ \Gamma_1, \quad v = \frac{\partial v}{\partial \nu} = 0 \ \text{ on } \ \Gamma_2, \tag{4}$$

$$u = v, \ \frac{\partial u}{\partial \nu} = \frac{\partial v}{\partial \nu}, \ \beta_1 \Delta u = \beta_2 \Delta v, \ \beta_1 \frac{\partial \Delta u}{\partial \nu} + \mu \frac{\partial \theta}{\partial \nu} = \beta_2 \frac{\partial \Delta v}{\partial \nu} \text{ on } \Gamma_0, \tag{5}$$

$$\theta = 0 \ \text{ on } \ \Gamma_0, \quad \frac{\partial \theta}{\partial \nu} + \lambda \theta = 0 \ \text{ on } \ \Gamma_1. \tag{6}$$

The equations above are equipped with the following initial data:

$$\begin{array}{l} u(\mathbf{x},0) = u_0(\mathbf{x}), \ \ u_t(\mathbf{x},0) = u_1(\mathbf{x}), \ \ \theta(\mathbf{x},0) = \theta_0(\mathbf{x}) \ \text{ in } \ \Omega_1, \\ v(\mathbf{x},0) = v_0(\mathbf{x}), \ \ v_t(\mathbf{x},0) = v_1(\mathbf{x}) \ \text{ in } \ \Omega_2. \end{array} \tag{7}$$

Coefficients ρ_i, β_i and μ are strictly positive, the functions

$$F_i : H^2(\Omega_1) \times H^2(\Omega_2) \longrightarrow \ L^2(\Omega_i), \ i = 1, 2$$

are nonlinear.

Functions $u(\mathbf{x}, t)$ and $v(\mathbf{x}, t)$ describe the vertical displacement of the plate in Ω_1 and Ω_2, respectively, and function $\theta(\mathbf{x}, t)$ describes the temperature regime. Equations (1) and (3) are plate equations, equation (1) is a heat equation. Equalities in (4) mean that the plate is clamped along Γ_1 and Γ_2. Boundary conditions (5) are transmission boundary conditions. Temperature function θ satisfies the Newton law of cooling on Γ_1 with some positive coefficient λ and vanishes on Γ_0 (see equalities (6)).

Let us introduce four problems which are concrete examples of abstract problem (1)-(7).

Problem A corresponds to oscillations of a plate in the Berger approach. In this case

$$F_1(u, v) = -M(u, v)\Delta u, \quad F_2(u, v) = -M(u, v)\Delta v,$$

where

$$M(u, v) = \Gamma + \gamma \left[\int\limits_{\Omega_1} |\nabla u|^2 \mathrm{d}\mathbf{x} + \int\limits_{\Omega_2} |\nabla v|^2 \mathrm{d}\mathbf{x} \right]. \tag{8}$$

Here Γ is a real number, γ is strictly positive.

In **problem B** we consider scalar nonlinearities, namely,

$$F_1(u,v) = f_1(u), \quad F_2(u,v) = f_2(v),$$

where the scalar functions $f_i \in C^2$ satisfy the following conditions: there exist such $p > 1$ and $C > 0$ that

$$|f_i'(s)| \le C(1 + |s|^p),$$
$$\liminf_{|s|\to\infty} \frac{f_i(s)}{s} > 0.$$

In **problem C** we deal with the von Karman nonlinearity. Here we set $\Gamma_2 = \emptyset$ and

$$F_1(u,v) = -[u, \mathcal{F}_1], \quad F_2(u,v) = -[v, \mathcal{F}_2],$$

where $[\psi, \varphi] = \psi_{xx}\varphi_{yy} + \psi_{yy}\varphi_{xx} - 2\psi_{xy}\varphi_{xy}$ is the von Karman brackets; the Airy stress functions $\mathcal{F}_1$ and $\mathcal{F}_2$ solve the following equations (parameters γ_i are strictly positive):

$$\gamma_1 \Delta^2 \mathcal{F}_1 + [u,u] = 0 \text{ in } \Omega_1 \times \mathbb{R}^+ \text{ and } \gamma_2 \Delta^2 \mathcal{F}_2 + [v,v] = 0 \text{ in } \Omega_2 \times \mathbb{R}^+ \qquad (9)$$

with the boundary conditions:

on $\Gamma_1 : \mathcal{F}_1 = \frac{\partial}{\partial \nu}\mathcal{F}_1 = 0,$
on $\Gamma_0 : \mathcal{F}_1 = \mathcal{F}_2, \ \frac{\partial}{\partial \nu}\mathcal{F}_1 = \frac{\partial}{\partial \nu}\mathcal{F}_2, \ \gamma_1 \Delta \mathcal{F}_1 = \gamma_2 \Delta \mathcal{F}_2, \ \gamma_1 \frac{\partial}{\partial \nu}\Delta \mathcal{F}_1 = \gamma_2 \frac{\partial}{\partial \nu}\Delta \mathcal{F}_2.$

Problem D corresponds to the problem of oscillations of the Berger plate on an elastic base. Mathematically, this problem is a a generalization of problems A and B. Nonlinearities F_i are given by the following equalities:

$$F_1(u,v) = -M(||\nabla u||^2_{\Omega_1} + ||\nabla v||^2_{\Omega_2})\Delta u + a_1(\mathbf{x})u|u|^{p-1} + g_1(\mathbf{x},u),$$
$$F_2(u,v) = -M(||\nabla u||^2_{\Omega_1} + ||\nabla v||^2_{\Omega_2})\Delta v + a_2(\mathbf{x})v|v|^{p-1} + g_2(\mathbf{x},v),$$

where $M(s) = s^{1+\alpha}$ with $\alpha > 0$, $a_1(\mathbf{x}) \in L^\infty(\Omega_1)$ and $a_2(\mathbf{x}) \in L^\infty(\Omega_2)$. We assume that the following condition holds:

$$\text{either } a(\mathbf{x}) \ge c_0 \ \forall \mathbf{x} \in \Omega \text{ or } 2(\alpha+2) > p+1, \ p \ge 1. \qquad (10)$$

Here $a = \{a_1, a_2\}$ and $c_0 > 0$ is a small number. The functions $g_1(\mathbf{x},u)$ and $g_2(\mathbf{x},v)$ are scalar and satisfy the growth condition for some $\varepsilon_0 > 0$ and any $\mathbf{x}_i \in \Omega_i$:

$$\left|\frac{\partial}{\partial u} g_1(\mathbf{x}_1,u)\right| + \left|\frac{\partial}{\partial v} g_2(\mathbf{x}_2,v)\right| \le C(1+|u|^{\max\{0,p-1-\varepsilon_0\}} + |v|^{\max\{0,p-1-\varepsilon_0\}}), \quad (11)$$

Our main result is the property of asymptotic smoothness of the dynamical system generated by weak solutions of problem (1)-(7). To achieve it we use method of so-called compensated compactness function first introduced in [7] (see

also [2, Proposition 2.10]). The method for various types of nonlinearities was developed in [2]. We also need to impose the following conditions on parameters:

$$\rho_1 \geq \rho_2 \text{ and } \beta_1 \leq \beta_2 \tag{12}$$

and on geometric structure of Ω_i:

$$(\mathbf{x} - \mathbf{x}_0) \cdot \nu(\mathbf{x}) \geq \delta_0 \text{ on } \Gamma_0, \tag{13}$$

$$(\mathbf{x} - \mathbf{x}_0) \cdot \nu(\mathbf{x}) \leq 0 \text{ on } \Gamma_2 \tag{14}$$

for some $\mathbf{x}_0 \in \mathbb{R}^2$ and $\delta_0 > 0$. Imposing these conditions authors of work [3] proved the exponential stability of linear problem ($F_1 = F_2 = 0$).

Asymptotic smoothness is important property of a dynamical system if one wants to prove the existence of a compact global attractor. In particular, asymptotically smooth dynamical system possesses a compact global attractor, if it possesses an appropriate Lyapunov function (for details we refer to [2, Chapter 2.4]).

Our result on asymptotic smoothness is applicable for each of the concrete problem listed above. For problems A, B and D we proved that the corresponding dynamical system possesses an appropriate Lyapunov function and, thus, there exists a compact global attractor.

Up to our best knowledge asymptotic behavior in transmission problem for a plate of types A, B, C and D was not considered before.

One can find formulations of our results in the next section. For proofs and other details we refer to our works [4] and [5].

2 Formulation of Main Result

2.1 Dynamical System

Below notation $\psi = \{\psi_1, \psi_2\}$ means that $\psi(\mathbf{x})$ defined for $\mathbf{x} \in \Omega$ is equal to $\psi_i(\mathbf{x})$, if $\mathbf{x} \in \Omega_i$, for $i = 1, 2$.

To formulate a well-posedness result we need to impose the following conditions:

$$\int_{\Omega_1} |F_1(w_1^1, w_2^1) - F_1(w_1^2, w_2^2)|^2 \mathrm{d}\mathbf{x} + \int_{\Omega_2} |F_2(w_1^1, w_2^1) - F_2(w_1^2, w_2^2)|^2 \mathrm{d}\mathbf{x} \leq C(r) || \left\{ w_1^1 - w_1^2, w_2^1 - w_2^2 \right\} ||^2_{H_T^2} \tag{15}$$

for all $|| \left\{ w_1^i, w_2^i \right\} ||_{H_0^2(\Omega)} \leq r,\ i = 1, 2$. We also assume that there exists such continuous functional $\Pi : H_0^2(\Omega) \to \mathbb{R}$ that

$$\frac{\mathrm{d}}{\mathrm{d}t} \Pi(w_1, w_2) = \int_{\Omega_1} F_1(w_1, w_2) w_{1,t} \mathrm{d}\mathbf{x} + \int_{\Omega_2} F_2(w_1, w_2) w_{2,t} \mathrm{d}\mathbf{x}, \tag{16}$$

$$\Pi(w_1, w_2) \geq -C, \ \ \exists C > 0, \tag{17}$$

$$\Pi(w_1, w_2) \leq \mathcal{G}\left(|| \{w_1, w_2\} ||_{H_0^2(\Omega)} \right). \tag{18}$$

Condition (16) holds for

$$\{w_1, w_2\} \in L^2(0,T; H_0^2(\Omega)) \text{ and } \{w_{1,t}, w_{2,t}\} \in L^2((0,T) \times \Omega).$$

Conditions (17) and (18) hold for all $\{w_1, w_2\} \in H_0^2(\Omega)$. The scalar function $\mathcal{G} : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ is supposed to be bounded on bounded intervals. The condition (16) also means that feedback forces $\{F_1(u,v), F_2(u,v)\}$ are potential, i.e., $\{F_1(u,v), F_2(u,v)\}$ is a Freshet derivative of $\Pi(u,v)$.

For problem A $\Pi(w_1, w_2) = \frac{1}{4} M^2(w_1, w_2)$, for problem B:

$$\Pi(w_1, w_2) = \int_{\Omega_1} \int_0^{w_1(x)} f_1(s) \mathrm{d}s\mathrm{d}\mathbf{x} + \int_{\Omega_1} \int_0^{w_2(x)} f_2(s) \mathrm{d}s\mathrm{d}\mathbf{x}.$$

For problem C, if calculate $\mathcal{F}_i$ according to (9):

$$\Pi(w_1, w_2) = \frac{\gamma_1}{2} \int_{\Omega_1} |\Delta \mathcal{F}_1|^2 \mathrm{d}\mathbf{x} + \frac{\gamma_2}{2} \int_{\Omega_2} |\Delta \mathcal{F}_2|^2 \mathrm{d}\mathbf{x}. \tag{19}$$

For problem D:

$$\Pi(w) = \frac{1}{2(\alpha+2)} ||\nabla w||^{2(\alpha+2)}_{L^2(\Omega)} + \frac{1}{p+1} \int_\Omega a(\mathbf{x}) |w(\mathbf{x})|^{p+1} d\mathbf{x} + \int_\Omega \int_0^{w(\mathbf{x})} g(\mathbf{x}, s) ds d\mathbf{x},$$

where $a = \{a_1, a_2\}$ and $g = \{g_1, g_2\}$.

We introduce phase space $\mathcal{H} = H_0^2(\Omega) \times L^2(\Omega) \times L^2(\Omega_2)$ and energy function $\mathcal{E} : \mathcal{H} \longrightarrow \mathbb{R}$ which we define for an argument $w = (w_1, w_2, w_3, w_4, w_5)$ (here $\{w_1, w_2\} \in H_0^2(\Omega)$, $\{w_3, w_4\} \in L^2(\Omega)$ and $w_5 \in L^2(\Omega)$) as follows

$$\mathcal{E}(w) = \frac{1}{2} \left[\int_{\Omega_1} \beta_1 |\Delta w_1|^2 + \rho_1 |w_3|^2 + \rho_0 |w_5|^2 d\mathbf{x} + \int_{\Omega_2} \beta_2 |\Delta w_2|^2 + \rho_2 |w_4|^2 d\mathbf{x} + 2\Pi(w_1, w_2) \right]. \tag{20}$$

Now we are in position to formulate the theorem on well-posedness of problem (1)-(7).

Theorem 1. *Let (15), (16), (17) and (18) hold. Then for any initial $w_0 \in \mathcal{H}$ and $T > 0$ there exists a unique weak solution $w(t) \in C([0,T]; \mathcal{H})$. Moreover, it satisfies the energy equality:*

$$\mathcal{E}(w(T)) - \mathcal{E}(w(t)) = -\int_t^T \int_{\Omega_1} \beta_0 |\nabla w_5|^2 d\mathbf{x} d\tau - \int_t^T \int_{\Gamma_1} \beta_0 \lambda |w_5|^2 d\Gamma d\tau \tag{21}$$

for all $0 \le t \le T$. If one sets $S_t w_0 = w(t)$, then $(\mathcal{H}, S_t)$ is a continuous dynamical system.

2.2 Asymptotic Smoothness

In this subsection we define the notion of asymptotic smoothness of a dynamical system, impose additional conditions on nonlinear functions F_i and formulate our result on asymptotic smoothness of the dynamical system $(\mathcal{H}, S_t)$.

Definition 1. *Let (X, S_t) be a dynamical system. Assume that X is a complete metric space and S_t is a semigroup of operators on X. The dynamical system (X, S_t) is said to be asymptotically smooth if for any positively invariant bounded set $D \subset X$ there exists a compact K in the closure $\overline{D}$ of D such that*

$$\lim_{t\to+\infty} \sup_{x\in D} \operatorname{dist}_X (S_t x, K) = 0.$$

For the detailed discussion and applications of asymptotic smoothness we refer to, e.g., [6] and [2].

To obtain asymptotic smoothness we need to impose additional conditions on nonlinear functions F_i.

Assume that there exist such $\delta, \sigma > 0$ that

$$w \longmapsto \Pi(w) \ : \ H^{2-\delta} \longrightarrow \mathbb{R} \ \text{ is a continuous mapping,} \tag{22}$$

$$w \longmapsto \Pi'_{\Phi}(w) : \ H^{2-\delta} \longrightarrow H^{-\sigma} \ \text{ is a continuous mapping.} \tag{23}$$

Our main result is the following theorem.

Theorem 2. *Let (12),(13), (14), (15), (17), (18), (22) and (23) hold. Then the dynamical system $(\mathcal{H}, S_t)$ is asymptotically smooth.*

The method of the proof is based on idea of compensated compactness function (see [7] and [2]). This result is applicable for all concrete problems, A,B,C and D, listed in introduction.

2.3 Compact Global Attractor and Its Properties

The direct application of [2, Corollary 2.29] gives the following theorem.

Theorem 3. *Let conditions of theorem 2 hold. Assume also that for all solution $w(t)$ to problem (1)-(7) the following statement holds:*

$$\text{if } \mathcal{E}(w(t)) \text{ does not depend on } t, \text{ then } w(t) \text{ does not depend on } t. \tag{24}$$

In other words, energy $\mathcal{E}$ is constant only on stationary trajectories.
Then the dynamical system $(\mathcal{H}, S_t)$ possesses a compact global attractor.

Verification of condition (24) could be reduced to the problem which is relative to unique continuation problems. Let $w(t) = (u(t), v(t), u_t(t), v_t(t), \theta(t))$ be a solution of problem (1)-(7) such that $\mathcal{E}(w(t))$ is constant. Let us denote $v^h(t) := v(t+h) - v(t)$, $B(t)v^h := F_2(u, v(t+h)) - F_2(u, v(t))$ for some $h > 0$. If $\mathcal{E}(w(t))$ does not depend on t, then v^h solves the following problem:

$$\rho_2 v^h_{tt} + \beta_2 \Delta^2 v^h = B(t)v^h,$$

$$v^h|_{\Gamma_0} = \frac{\partial v^h}{\partial \nu}|_{\Gamma_0} = \Delta v^h|_{\Gamma_0} = \frac{\partial \Delta v^h}{\partial \nu}|_{\Gamma_0} = 0.$$

Therefore, if we prove that $v^h(t) \equiv 0$, then we will have that $w(t) \equiv w_0$.

Using Pochozhaev multiplicator $(\mathbf{x} - \mathbf{x}_0) \cdot \nabla v^h$, where vector $\mathbf{x}_0$ is the same as in conditions (13) and (14), for concrete problems A and B with $f_2 \equiv 0$ and Carleman estimates obtained in [8] for concrete problems B and D, we manage to verify (24). Thus, we obtain the following corollary.

Corollary 1. *Let (12),(13), (14) hold. The dynamical system* $(\mathcal{H}, S_t)$ *corresponding to one of the concrete problems A, B or D possesses a compact global attractor.*

References

1. Lasiecka, I., Triggiani, R.: Control Theory for PDEs, vol. 1. Cambridge University Press, Cambridge (2000)
2. Chueshov, I.D., Lasiecka, I.: Long-time behavior of second order evolution equations with nonlinear damping. Memoirs of AMS, vol. (912). Americal Mathematical Society, Providence (2008)
3. Rivera, J.E.M., Oquendo, H.P.: A transmission problem for thermoelastic plates. Quarterly of Applied Mathematics 62(2), 273–293 (2004)
4. Potomkin, M.: A nonlinear transmission problem for a compound plate with thermoelastic part (2010), http://arxiv.org/abs/1003.3332
5. Potomkin, M.: On Transmission Problem for Berger Plates on an Elastic Base. Journal of Mathematical Physics, Analysis and Geometry 7(1), 96–102 (2011)
6. Raugel, G.: Global attractors in partial differential equations. In: Fiedler, B. (ed.) Handbook of Dynamical Systems, vol. 2, pp. 885–982. Elsevier, Amsterdam (2002)
7. Khanmamedov, A.: Global attractors for von Karman equations with nonlinear dissipation. J. Math. Anal. Appl. 318, 92–101 (2006)
8. Albano, P.: Carleman estimates for the Euler-Bernoulli plate operator. Electronic Journal of Diff. Eq. 53, 1–13 (2000)

p-th Order Optimality Conditions for Singular Lagrange Problem in Calculus of Variations. Elements of p-Regularity Theory

Agnieszka Prusińska[1], Ewa Szczepanik[1], and Alexey Tret'yakov[1,2]

[1] Siedlce University of Natural Sciences and Humanities, Siedlce, Poland
aprus@uph.edu.pl
[2] System Research Institute of the Polish Academy of Sciences, Warsaw, Poland, Dorodnicyn Computing Center of the Russian Academy of Sciences, Moscow, Russia

Abstract. This paper is devoted to singular calculus of variations problems with constraints which are not regular mappings at the solution point, e.i. its derivatives are not surjective. We pursue an approach based on the constructions of the p-regularity theory. For p-regular calculus of variations problem we present necessary conditions for optimality in singular case and illustrate our results by classical example of calculus of variations problem.

Keywords: singular variational problem, necessary condition of optimality, p-regularity, p-factor operator.

1 Introduction

Let us consider the following Lagrange problem:

$$J_0(x) = \int_{t_1}^{t_2} F(t, x(t), x'(t))dt \to \min \tag{1}$$

subject to the subsidiary conditions

$$H(t, x(t), x'(t)) = 0,\ Ax(t_1) + Bx(t_2) = 0 \tag{2}$$

where $x \in \mathcal{C}_n^2[t_1, t_2]$, $H(t, x(t), x'(t)) = (H_1(t, x(t), x'(t)), \dots, H_m(t, x(t), x'(t)))^T$, $H_i : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, $i = 1, \dots, m$, $F : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, $t \in [t_1, t_2]$, A, B — $n \times n$ matrices, $\mathcal{C}_n^l[t_1, t_2]$ — Banach spaces of n-dimensional l-times continuously differentiable vector functions with usual norms.

Let us introduce a mapping $G(x) = H(\cdot, x(\cdot), x'(\cdot))$ such that $G : X \to Y$, where $X = \{x(\cdot) \in \mathcal{C}_n^2[t_1, t_2] : Ax(t_1)+Bx(t_2) = 0\}$, $Y = \mathcal{C}_m[t_1, t_2]$. It means that G acts as follows $G(x)t = H(t, x(t), x'(t))$. Then the system of equations (2) can be replaced by the following operator equation $G(x) = 0_Y$ (or $G(x(\cdot)) = 0_Y$). We assume that all the functions and their derivatives in (1)–(2) are $p+1$-times continuously differentiable with respect to the corresponding variables t, x, x'.

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 528–537, 2013.

Under these assumptions: $G(x) \in \mathcal{C}^{p+1}(X)$, where by $\mathcal{C}^{p+1}(X)$ we mean a set of $p+1$-times continuously differentiable mappings on X.

Let us denote $\lambda(t) = (\lambda_1(t), \dots, \lambda_m(t))^T$, $\lambda(t)H = \lambda_1(t)H_1 + \dots + \lambda_m(t)H_m$, $\lambda(t)H_x = \lambda_1(t)H_{1x} + \dots + \lambda_m(t)H_{mx}$, $\lambda(t)H_{x'} = \lambda_1(t)H_{1x'} + \dots + \lambda_m(t)H_{mx'}$.

If $\operatorname{Im} G'(\hat{x}) = Y$, where $\hat{x}(t)$ is a solution to (1)–(2), then necessary conditions of Euler-Lagrange $F_x + \lambda(t)H_x - \frac{d}{dt}(F_{x'} + \lambda(t)H_{x'}) = 0$ hold. Here, F_x, H_x, $F_{x'}$, $H_{x'}$ are partial derivatives of the functions $F(t, x(t), x'(t))$ and $H(t, x(t), x'(t))$ with respect to x and x', respectively.

In *singular* (*nonregular* or *degenerate*) case when $\operatorname{Im} G'(\hat{x}) \neq Y$, we can only guarantee that the following equations

$$\lambda_0 F_x + \lambda(t)H_x - \frac{d}{dt}(\lambda_0 F_{x'} + \lambda(t)H_{x'}) = 0 \tag{3}$$

hold, where $\lambda_0^2 + \|\lambda(t)\|^2 = 1$, i.e. λ_0 might be equal to 0, and then we have not constructive information of the functional $F(t, x(t), x'(t))$.

Example 1. Consider the problem

$$J_0(x) = \int_0^{2\pi} (x_1^2(t) + x_2^2(t) + x_3^2(t) + x_4^2(t) + x_5^2(t))dt \to \min \tag{4}$$

subject to

$$\begin{aligned} &H(t, x(t), x'(t)) = \\ &= \begin{pmatrix} x_1'(t) - x_2(t) + x_3^2(t)x_1(t) + x_4^2(t)x_2(t) - x_5^2(t)(x_1(t) + x_2(t)) \\ x_2'(t) + x_1(t) + x_3^2(t)x_2(t) - x_4^2(t)x_1(t) - x_5^2(t)(x_2(t) - x_1(t)) \end{pmatrix} = 0, \end{aligned} \tag{5}$$

$x_i(0) - x_i(2\pi) = 0$, $i = 1, \dots, 5$.

Here $F(t, x(t), x'(t)) = x_1^2(t) + x_2^2(t) + x_3^2(t) + x_4^2(t) + x_5^2(t)$, $A = -B = I_5$, where I_5 is the unit matrix of size 5 and

$$\begin{aligned} &G(x) = \\ &= \begin{pmatrix} x_1'(\cdot) - x_2(\cdot) + x_3^2(\cdot)x_1(\cdot) + x_4^2(\cdot)x_2(\cdot) - x_5^2(\cdot)(x_1(\cdot) + x_2(\cdot)) \\ x_2'(\cdot) + x_1(\cdot) + x_3^2(\cdot)x_2(\cdot) - x_4^2(\cdot)x_1(\cdot) - x_5^2(\cdot)(x_2(\cdot) - x_1(\cdot)) \end{pmatrix} = 0. \end{aligned}$$

The solution of (4)–(5) is $\hat{x}(t) = 0$. At this point $G'(0)$ is singular. Later we explain this in more details.

The corresponding Euler-Lagrange equation (see (3)) in this case is as follows:

$$\begin{aligned} 2\lambda_0 x_1 + \lambda_2 - \lambda_1' + \lambda_1 x_3^2 + \lambda_1 x_5^2 - \lambda_2 x_5^2 - \lambda_2 x_4^2 &= 0 \\ 2\lambda_0 x_2 - \lambda_1 - \lambda_2' + \lambda_1 x_4^2 + \lambda_2 x_3^2 - \lambda_1 x_5^2 - \lambda_2 x_5^2 &= 0 \\ 2\lambda_0 x_3 + 2\lambda_1 x_1 x_3 + 2\lambda_2 x_2 x_3 &= 0 \\ 2\lambda_0 x_4 + 2\lambda_1 x_2 x_4 - 2\lambda_2 x_1 x_4 &= 0 \\ 2\lambda_0 x_5 - 2\lambda_1 x_5 x_1 - 2\lambda_1 x_2 x_5 - 2\lambda_2 x_2 x_5 + 2\lambda_2 x_1 x_5 &= 0 \\ \lambda_i(0) - \lambda_i(2\pi) = 0, \quad i = 1, 2. \end{aligned} \tag{6}$$

(to simplify formulas we omit dependence of t here and further in the paper).

If $\lambda_0 = 0$ we obtain the series of spurious solutions to the system (4)–(5):

$$x_1 = a \sin t, \quad x_2 = a \cos t, \quad x_3 = x_4 = x_5 = 0,$$

$$\lambda_1 = b \sin t, \quad \lambda_2 = b \cos t, \quad a, b \in \mathbb{R}.$$

2 Elements of p-Regularity Theory

Let us recall the p-order necessary and sufficient optimality conditions for degenerate optimization problems (see [1]–[5]):

$$\min \varphi(x) \tag{7}$$

subject to

$$f(x) = 0, \tag{8}$$

where $f : X \to Y$ and X, Y are Banach spaces, $\varphi : X \to \mathbb{R}$, $f \in \mathcal{C}^{p+1}(X)$, $\varphi \in \mathcal{C}^2(X)$ and at the solution point $\hat{x}$ of (7)–(8) we have: $\operatorname{Im} f'(\hat{x}) \neq Y$ i.e. $f'(\hat{x})$ is singular.

Let us recall the basic constructions of p-regularity theory which is used in investigation of singular problems.

Suppose that the space Y is decomposed into a direct sum

$$Y = Y_1 \oplus \ldots \oplus Y_p, \tag{9}$$

where $Y_1 = \overline{\operatorname{Im} f'(\hat{x})}$, $Z_1 = Y$. Let Z_2 be closed complementary subspace to Y_1 (we assume that such closed complement exists), and let $P_{Z_2} : Y \to Z_2$ be the projection operator onto Z_2 along Y_1. By Y_2 we mean the closed linear span of the image of the quadratic map $P_{Z_2} f^{(2)}(\hat{x})[\cdot]^2$. More generally, define inductively,

$$Y_i = \overline{\operatorname{span}\ \operatorname{Im} P_{Z_i} f^{(i)}(\hat{x})[\cdot]^i} \subseteq Z_i, \quad i = 2, \ldots, p-1,$$

where Z_i is a chosen closed complementary subspace for $(Y_1 \oplus \ldots \oplus Y_{i-1})$ with respect to Y, $i = 2, \ldots, p$ and $P_{Z_i} : Y \to Z_i$ is the projection operator onto Z_i along $(Y_1 \oplus \ldots \oplus Y_{i-1})$ with respect to Y, $i = 2, \ldots, p$. Finally, $Y_p = Z_p$. The order p is chosen as the minimum number for which (9) holds. Let us define the following mappings

$$f_i(x) = P_i f(x), \quad f_i : X \to Y_i \ \ i = 1, \ldots, p,$$

where $P_i := P_{Y_i} : Y \to Y_i$ is the projection operator onto Y_i along $(Y_1 \oplus \ldots \oplus Y_{i-1} \oplus Y_{i+1} \oplus \ldots \oplus Y_p)$ with respect to Y, $i = 1, \ldots, p$.

Definition 1. *The linear operator* $\Psi_p(\hat{x}, h) \in \mathcal{L}(X, Y_1 \oplus \ldots \oplus Y_p)$, $h \in X$, $h \neq 0$

$$\Psi_p(\hat{x}, h) = f_1'(\hat{x}) + f_2''(\hat{x})h + \ldots + f_p^{(p)}(\hat{x})[h]^{p-1},$$

is called the p-factor operator.

Definition 2. *We say that the mapping f is p-regular at $\hat{x}$ along an element h, if* $\operatorname{Im} \Psi_p(\hat{x}, h) = Y$.

Remark 1. *The condition of p-regularity of the mapping $f(x)$ at the point $\hat{x}$ along h is equivalent to* $\operatorname{Im} f_p^{(p)}(\hat{x})[h]^{p-1} \circ \operatorname{Ker} \Psi_{p-1}(\hat{x}, h) = Y_p$, *where* $\Psi_{p-1}(\hat{x}, h) = f_1'(\hat{x}) + f_2''(\hat{x})h + \ldots + f_{p-1}^{(p-1)}(\hat{x})[h]^{p-2}$

Definition 3. *We say that the mapping f is p-regular at $\hat{x}$ if it is p-regular along any h from the set*

$$H_p(\hat{x}) = \bigcap_{k=1}^{p} \operatorname{Ker}^k f_k^{(k)}(\hat{x}) \setminus \{\mathbf{0}\},$$

where

$$\operatorname{Ker}^k f_k^{(k)}(\hat{x}) = \{\xi \in X : f_k^{(k)}(\hat{x})[\xi]^k = 0\}.$$

is k-kernel of the k-order mapping $f_k^{(k)}(\hat{x})[\xi]^k$.

For a linear surjective operator $\Psi_p(\hat{x}, h) : X \mapsto Y$ between Banach spaces we denote by $\{\Psi_p(\hat{x}, h)\}^{-1}$ its *right inverse*. Therefore $\{\Psi_p(\hat{x}, h)\}^{-1} : Y \mapsto 2^X$ and we have $\{\Psi_p(\hat{x}, h)\}^{-1}(y) = \{x \in X : \Psi_p(\hat{x}, h)x = y\}$. We define the *norm* of $\{\Psi_p(\hat{x}, h)\}^{-1}$ via the formula

$$\|\{\Psi_p(\hat{x}, h)\}^{-1}\| = \sup_{\|y\|=1} \inf\{\|x\| : x \in \{\Psi_p(\hat{x}, h)\}^{-1}(y)\}.$$

We say that $\{\Psi_p(\hat{x}, h)\}^{-1}$ is *bounded* if $\|\{\Psi_p(\hat{x}, h)\}^{-1}\| < \infty$.

Definition 4. *The mapping f is called strongly p-regular at the point $\hat{x}$ if there exists $\gamma > 0$ such that*

$$\sup_{h \in H_\gamma} \left\| \{\Psi_p(\hat{x}, h)\}^{-1} \right\| < \infty$$

where $H_\gamma(\hat{x}) = \left\{ h \in X : \left\| f_k^{(k)}(\hat{x})[h]^k \right\|_{Y_k} \le \gamma, k = 1, \ldots, p, \|h\| = 1 \right\}$.

3 Optimality Conditions for p-Regular Optimization Problems

We define p-factor Lagrange function

$$\mathcal{L}_p(x, \lambda, h) = \varphi(x) + \left\langle \sum_{k=1}^{p} f_k^{(k-1)}(x)[h]^{k-1}, \lambda \right\rangle,$$

where $\lambda \in Y^*$, $f_1^{(0)}(x) = f(x)$ and

$$\bar{\mathcal{L}}_p(x, \lambda, h) = \varphi(x) + \left\langle \sum_{k=1}^{p} \frac{2}{k(k+1)} f_k^{(k-1)}(x)[h]^{k-1}, \lambda \right\rangle.$$

Let us recall the following basic theorems on optimality conditions in nonregular case.

Theorem 1 (Necessary and sufficient conditions for optimality). *(see [7]) Let X and Y be Banach spaces, $\varphi \in \mathcal{C}^2(X)$, $f \in \mathcal{C}^{p+1}(X)$, $f : X \to Y$, $\varphi : X \to \mathbb{R}$. Suppose that $h \in H_p(\hat{x})$ and f is p-regular along h at the point $\hat{x}$. If $\hat{x}$ is a local solution to the problem (7)–(8) then there exist multipliers, $\hat{\lambda}(h) \in Y^*$ such that*

$$\mathcal{L}'_{px}(\hat{x}, \hat{\lambda}(h), h) = 0 \Leftrightarrow \varphi'(\hat{x}) + \left(f'_1(\hat{x}) + \cdots + f_p^{(p)}(\hat{x})[h]^{p-1}\right)^* \hat{\lambda}(h) = 0. \quad (10)$$

Moreover, if f is strongly p-regular at $\hat{x}$, there exist $\alpha > 0$ and a multipliers $\hat{\lambda}(h)$ such that (10) is fulfilled and $\bar{\mathcal{L}}_{pxx}(\hat{x}, \hat{\lambda}(h), h)[h]^2 \geq \alpha \|h\|^2$ for every $h \in H_p(\hat{x})$, then $\hat{x}$ is a strict local minimizer to the problem (7)–(8).

For our purposes, the following modification of Theorem 1 will be useful (see [3]).

Theorem 2. *Let X and Y be Banach spaces, $\varphi \in \mathcal{C}^2(X)$, $f \in \mathcal{C}^{p+1}(X)$, $f : X \to Y$, $\varphi : X \to \mathbb{R}$, $h \in H_p(\hat{x})$, and f is p-regular along h at the point $\hat{x}$. If $\hat{x}$ is a solution to the problem (7)–(8), then there exist multipliers $\bar{\lambda}_i(h) \in Y_i^*$, $i = 1, \ldots, p$ such that*

$$\varphi'(\hat{x}) + (f'(\hat{x}))^* \bar{\lambda}_1(h) + \ldots + \left(f^{(p)}(\hat{x})[h]^{p-1}\right)^* \bar{\lambda}_p(h) = 0, \quad (11)$$

and

$$\left(f^{(k)}(\hat{x})[h]^{k-1}\right)^* \bar{\lambda}_i(h) = 0,\ k = 1, \ldots, i-1,\ i = 2, \ldots, p. \quad (12)$$

Moreover, if f is strongly p-regular at $\hat{x}$, there exist $\alpha > 0$ and multipliers $\bar{\lambda}_i(h)$, $i = 1, \ldots, p$ such that (11)–(12) hold, and

$$\left(\varphi''(\hat{x}) + \frac{1}{3} f''(\hat{x})\bar{\lambda}_1(h) + \ldots + \frac{2}{p(p+1)} f^{(p+1)}(\hat{x})[h]^{p-1}\bar{\lambda}_p(h)\right)[h]^2 \geq$$

$$\geq \alpha \|h\|^2,$$

for every $h \in H_p(\hat{x})$, then $\hat{x}$ is a strict local minimizer to the problem (7)–(8).

Proof.

We need to prove only the formula (12). From (10) we obtain
$\varphi'(\hat{x}) + \left(P_1 f'(\hat{x}) + \cdots + P_p f^{(p)}(\hat{x})[h]^{p-1}\right)^* \hat{\lambda}(h) = 0.$

This expression can be transformed as follows
$\varphi'(\hat{x}) + f'(\hat{x})^* P_1^* \hat{\lambda}(h) + \cdots + \left(f^{(p)}(\hat{x})[h]^{p-1}\right)^* P_p^* \hat{\lambda}(h) = 0.$

Let $\bar{\lambda}_i(h) := P_i^* \hat{\lambda}(h)$, $i = 1, \ldots, p$. Then, for $k < i$, $i = 1, \ldots, p$,
$\left(f^{(k)}(\hat{x})[h]^{k-1}\right)^* \bar{\lambda}_i(h) = \left(f^{(k)}(\hat{x})[h]^{k-1}\right)^* P_i^* \hat{\lambda}(h) =$
$= \left(P_i f^{(k)}(\hat{x})[h]^{k-1}\right)^* \hat{\lambda}(h) = 0$, which proves (12).

Now we are ready to apply this theorem to singular calculus of variations problems. Let us introduce p-factor Euler-Lagrange function

$$S(x) = F(x) + \left\langle \lambda(t), \left(g_1(x) + g'_2(x)[h] + \ldots + g_p^{(p-1)}(x)[h]^{p-1}\right)\right\rangle =$$
$$= F(x) + \lambda(t) G^{(p-1)}(x)[h]^{p-1},$$

where $G^{(p-1)}(x)[h]^{p-1} = g_1(x) + g_2'(x)[h] + \cdots + g_p^{(p-1)}(x)[h]^{p-1}$, $\lambda(t) = (\lambda_1(t), \ldots, \lambda_m(t))^T$ and $g_k(x)$, for $k = 1, \ldots, p$ are determined for the mapping $G(x)$ similarly like $f_k(x)$, $k = 1, \ldots, p$ for the mapping $f(x)$, i.e. $g_k(x) = P_{Y_k} G(x)$, $k = 1, \ldots, p$. Denote

$$g_k^{(k-1)}(x)[h]^{k-1} = \sum_{i+j=k-1} C_{k-1}^i g_{kx^i(x')^j}^{(k-1)}(x) h^i (h')^j, \ k = 1, \ldots, p,$$

where

$$g_{kx^i(x')^j}^{(k-1)}(x) = g_k^{(k-1)}(\underbrace{x \ldots x}_{i} \underbrace{x' \ldots x'}_{j})(x).$$

Definition 5. *We say that the problem (1)–(2) is p-regular at $\hat{x}$ along $h \in \bigcap_{k=1}^{p} \mathrm{Ker}^k g_k^{(k)}(\hat{x})$, $\|h\| \neq 0$ if*

$$\mathrm{Im}\left(g_1'(\hat{x}) + \ldots + g_p^{(p)}(\hat{x})[h]^{p-1}\right) = \mathcal{C}_m[t_1, t_2].$$

The following theorem holds.

Theorem 3. *Let $\hat{x}(t)$ be a solution of the problem (7)–(8) and assume that the problem is p-regular at $\hat{x}$ along $h \in \bigcap_{k=1}^{p} \mathrm{Ker}^k g_k^{(k)}(\hat{x})$. Then there exists a multiplier $\hat{\lambda}(t) = (\hat{\lambda}_1(t), \ldots, \hat{\lambda}_m(t))^T$ such that the following p-factor Euler-Lagrange equation*

$$\begin{aligned} &S_x(\hat{x}) - \frac{d}{dt} S_{x'}(\hat{x}) = F_x(\hat{x}) + \\ &+ \left\langle \hat{\lambda}(t), \sum_{k=1}^{p} \sum_{i+j=k-1} C_{k-1}^i g_{x^i(x')^j}^{(k-1)}(\hat{x}) h^i (h')^j \right\rangle_x - \\ &- \frac{d}{dt}\left[F_{x'}(\hat{x}) + \left\langle \hat{\lambda}(t), \sum_{k=1}^{p} \sum_{i+j=k-1} C_{k-1}^i g_{x^i(x')^j}^{(k-1)}(\hat{x}) h^i (h')^j \right\rangle_{x'} \right] = 0 \end{aligned} \tag{13}$$

holds.

The proof of this theorem is very similar to the one of analogous result for the singular isoperimetric problem, see in [4], [5].

Consider again the Example 1 and (4)–(5). Here $p = 2$, $\hat{x} = 0$. At the beginning we substantiate that G is singular at the points $\bar{x} = (a \sin t, a \cos t, 0, 0, 0)^T$. Indeed, $G'(\bar{x}) = \begin{pmatrix} (\cdot)_1' - (\cdot)_2 \\ (\cdot)_2' + (\cdot)_1 \end{pmatrix}$, where $G'(\bar{x})x(t) = \begin{pmatrix} x_1'(t) - x_2(t) \\ x_2'(t) + x_1(t) \end{pmatrix}$. Let us denote $\begin{pmatrix} x_1' - x_2 \\ x_2' + x_1 \end{pmatrix}$ by $x' + Lx$, where $L = \begin{pmatrix} 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$.

Then $G'(\bar{x}) = (\cdot)' + L(\cdot)$ and

$$\mathrm{Ker} G'(\bar{x}) = \mathrm{span}\,\{(\Phi_1(t), 0, 0, 0)^T, (\Phi_2(t), 0, 0, 0)^T\} \oplus \{(0, 0, x_3(t), x_4(t), x_5(t))^T,$$

$x_i \in \mathcal{C}^2[0,2\pi],\ i=3,4,5\}$, where $\Phi_1(t)=(\sin t,\cos t)^T$, $\Phi_2(t)=(\cos t,-\sin t)^T$, and moreover $\operatorname{Im} G'(\bar{x})=(\operatorname{Ker}(G'(\bar{x})^*)^{\perp}=\left(\operatorname{Ker}(-\frac{d}{dt}(\cdot)'+L^T(\cdot))\right)^{\perp}=$
$=\left\{\xi\in\mathcal{C}_2[0,2\pi]:\langle\xi,\psi_i\rangle=0,i=1,2,\psi_1(t)=(\sin t,\cos t)^T,\psi_2(t)=\right.$
$\left.=(\cos t,-\sin t)^T\right\}\neq\mathcal{C}_2[0,2\pi]$.

It means that the mapping $G(x)$ is non-regular at the points $\bar{x}$. From the last relation we obtain that $Y_2=(\operatorname{Im} G'(\bar{x}))^{\perp}=\operatorname{span}\{\psi_1,\psi_2\}$ where $\psi_1'=\psi_2$, $\psi_2'=-\psi_1$ and $\langle\Phi_i,\psi_j\rangle=\delta_{ij}$, $\langle\zeta,\eta\rangle=\int_0^{2\pi}\zeta(\tau)\eta(\tau)d\tau$.

The projection operator P_{Y_2} is defined as

$$P_2\begin{pmatrix}y_1\\y_2\end{pmatrix}=P_2y=\bar{y}_1\psi_1+\bar{y}_2\psi_2,$$

where $y=(y_1,y_2)^T$ and

$$\langle y-(\bar{y}_1\psi_1+\bar{y}_2\psi_2),\psi_1\rangle=0,$$

$$\langle y-(\bar{y}_1\psi_1+\bar{y}_2\psi_2),\psi_2\rangle=0,$$

i.e. $\frac{1}{2\pi}\langle y,\psi_1\rangle=\bar{y}_1$, $\frac{1}{2\pi}\langle y,\psi_2\rangle=\bar{y}_2$.
Let us point out that $P_2(x_1,\psi_1+x_2\psi_2)=x_1\psi_1+x_2\psi_2$.

Based on Remark 1 we can verify surjectivity of $P_2G''(\bar{x})h$ only on $\operatorname{Ker}G'(\bar{x})$, for $h\in\operatorname{Ker}G'(\bar{x})\cap\operatorname{Ker}^2P_2G''(\bar{x})$, $h=(a\sin t,a\cos t,1,1,1)^T$. In order to find $P_2G''(\bar{x})h$ let us determine

$$G''(\bar{x})=\begin{pmatrix}\begin{pmatrix}0&0&\sin t&0&0\\0&0&0&\cos t&0\\0&0&0&0&\cos t-\sin t\end{pmatrix}\\\begin{pmatrix}0&0&\cos t&0&0\\0&0&0&-\sin t&0\\0&0&0&0&\sin t-\cos t\end{pmatrix}\end{pmatrix}$$

and

$$G''(\bar{x})h=2a\begin{pmatrix}0&0&h_3\sin t&h_4\cos t&h_5(\cos t-\sin t)\\0&0&h_3\cos t&-h_4\sin t&h_5(\sin t-\cos t)\end{pmatrix}.$$

It is obvious that $h=(a\sin t,a\cos t,1,1,1)^T$ belongs to $\operatorname{Ker}G'(\bar{x})\cap\operatorname{Ker}^2G''(\bar{x})$ and consequently belongs to $\operatorname{Ker}G'(\bar{x})\cap\operatorname{Ker}^2P_2G''(\bar{x})$. We have

$$G''(\bar{x})[h,x]=2a(x_3-x_5)\begin{pmatrix}\sin t\\\cos t\end{pmatrix}+2a(x_4-x_5)\begin{pmatrix}\cos t\\-\sin t\end{pmatrix}.$$

It means that
$P_2G''(\bar{x})[h,x]=G''(\bar{x})[h,x]$ and $G''(\bar{x})[h]\circ\operatorname{Ker}G'(\bar{x})=\operatorname{span}\{\Phi_1,\Phi_2\}=Y_2$. Therefore $G''(\bar{x})[h]$ is surjection. Hence, $G(x)$ is 2-regular along h at the points $\bar{x}=(a\sin t,a\cos t,0,0,0)^T$. Finally, we can apply Theorem 3 with $\lambda_0=1$. We have constructed operator

$$G'(\bar{x})+P_{Y_2}G''(\bar{x})h=$$

$$= \begin{pmatrix} (\cdot)_1' \\ (\cdot)_2' \end{pmatrix} + \begin{pmatrix} 0 & -1 & 2a\sin t & 2a\cos t & 2a(\cos t - \sin t) \\ 1 & 0 & 2a\cos t & -2a\sin t & 2a(\sin t - \cos t) \end{pmatrix}$$

which corresponds to the following system ($F_{x'} = 0$):

$$F_x(\bar{x}) + (G'(\bar{x}) + P_2G''(\bar{x})h)^*\lambda = 0 \Leftrightarrow F_x(\bar{x}) + G'(\bar{x})^T\lambda + (P_2G''(\bar{x})h)^T\lambda = 0, \Leftrightarrow$$

$$\Leftrightarrow \begin{cases} 2\bar{x}_1 - \lambda_1' + \lambda_2 = 0 \\ 2\bar{x}_2 - \lambda_2' + \lambda_1 = 0 \\ 2\bar{x}_3 + 2\lambda_1 a\sin t + 2\lambda_2 a\cos t = 0 \\ 2\bar{x}_4 + 2\lambda_1 a\cos t - 2\lambda_2 a\sin t = 0 \\ 2\bar{x}_5 + 2\lambda_1 a(\cos t - \sin t) + 2\lambda_2 a(\sin t - \cos t) = 0 \end{cases} \tag{14}$$

or

$$\Leftrightarrow \begin{cases} 2a\sin t - \lambda_1' + \lambda_2 = 0 \\ 2a\cos t - \lambda_2' + \lambda_1 = 0 \\ \lambda_1\sin t + \lambda_2\cos t = 0 \\ \lambda_1\cos t - \lambda_2\sin t = 0 \\ \lambda_1(\cos t - \sin t) + \lambda_2(\sin t - \cos t) = 0, \\ \lambda_i(0) - \lambda_i(2\pi) = 0,\ i = 1, 2. \end{cases}$$

One can verify that the false solutions of (6)

$$x_1 = a\sin t,\ x_2 = a\cos t,\ x_3 = x_4 = x_5 = 0$$

do not satisfy the system (14) for $a \neq 0$. It means that $x_1 = a\sin t$, $x_2 = a\cos t$, $x_3 = x_4 = x_5$ do not satisfy 2-factor Euler-Lagrange equation (13)

Let us consider the same problem with higher derivatives $x'(t)$, ..., $x^{(r)}(t)$, $r \geq 2$,

$$J(x) = \int_{t_1}^{t_2} F(t, x(t), x'(t), \ldots, x^{(r)}(t))dt \to \min, \quad x(t) \in C_n^{2r}[t_1, t_2],$$

subject to subsidiary differential relation

$$H(t, x(t), x'(t), \ldots, x^{(r)}(t)) = \begin{pmatrix} H_1(t, x(t), x'(t), \ldots, x^{(r)}(t)) \\ \cdots \\ H_m(t, x(t), x'(t), \ldots, x^{(r)}(t)) \end{pmatrix} = \begin{pmatrix} 0 \\ \ldots \\ 0 \end{pmatrix},$$

$A_kx^{(k)}(t_1) + B_kx^{(k)}(t_2) = 0$, where A_k, B_k are $n \times n$ matrices, $k = 1, \ldots, r$. Let $G(x) = H(\cdot, x(\cdot), \ldots, x^{(r)}(\cdot))$, $G : X \to Y$, where $Y = \mathcal{C}_m([t_1, t_2])$ and $X = \{x(\cdot) \in \mathcal{C}_n^{2r}[t_1, t_2] : A_kx^{(k)}(t_1) + B_kx^{(k)}(t_2) = 0, k = 1, \ldots, r\}$.

Moreover,

$$g_k^{(k-1)}(x)[h]^{k-1} = \sum_{i_1 + \cdots + i_r = k-1} g^{(k-1)}_{x^{i_1} \ldots (x^{(r)})^{i_r}}[h + h' + \cdots + h^{(r)}]^{k-1},\ k = 1, \ldots, p,$$

and introduce the co called p-factor Euler-Poisson function

$$K(x) = F(x) + \left\langle \lambda(t), \left(g_1(x) + g_2'(x)[h] + \ldots + g_p^{(p-1)}(x)[h]^{p-1} \right) \right\rangle$$

Theorem 4. *Let $\hat{x}(t)$ be a solution of the problem (1)–(2) and assume that this problem is p-regular at $\hat{x}$ along $h \in \bigcap_{k=1}^{p} \operatorname{Ker}^k g_k^{(k)}(\hat{x})$. Then there exist a multiplier $\hat{\lambda}(t) = (\hat{\lambda}_1(t), \ldots, \hat{\lambda}_m(t))^T$ such that the following p-factor Euler-Poisson equation*

$$K_x(\hat{x}) - \frac{d}{dt}K_{x'}(\hat{x} + \frac{d^2}{dt^2}K_{x''}(\hat{x}) - \ldots + (-1)^r K_{x^{(r)}}(\hat{x}) =$$
$$= F_x(\hat{x}) + \left\langle \hat{\lambda}(t), \sum_{k=1}^{p} g_k^{(k-1)}(\hat{x})[h]^{k-1} \right\rangle_x -$$
$$-\frac{d}{dt}\left[F_{x'}(\hat{x}) + \left\langle \hat{\lambda}(t), \sum_{k=1}^{p} g_k^{(k-1)}(\hat{x})[h]^{k-1} \right\rangle_{x'} \right] +$$
$$+ \ldots + (-1)^r \frac{d^r}{dt^r}\left[F_{x^{(r)}}(\hat{x}) + \left\langle \hat{\lambda}(t), \sum_{k=1}^{p} g_k^{(k-1)}(\hat{x})[h]^{k-1} \right\rangle_{x^{(r)}} \right] = 0$$

holds.

The proof of Theorem 4 is similar to that one the reader can find in [4] for isoperimetric problem.

Example 2. Consider the following problem

$$J_0(x) = \int_0^{\pi} (x_1^2(t) + x_2^2(t) + x_3^2(t))dt \to \min \tag{15}$$

subject to

$$H(t, x(t), x'(t), x''(t)) = x_1''(t) + x_1(t) + x_2^2(t)x_1(t) - x_3^2(t)x_1(t) = 0, \tag{16}$$

$x_i(0) - x_i(\pi) = 0$, $x_i'(0) + x_i'(\pi) = 0$, $i = 1, 2, 3$. Here $A_1 = -B_1 = I_3$, $A_2 = B_2 = I_3$, where I_3 means the unit matrix of size 3.

The solution of (15)–(16) is $\hat{x}(t) = 0$. The Euler-Poisson equation in this case has the following form

$$\lambda_0 F_x + \lambda(t)H_x - \frac{d}{dt}(\lambda(t)H_{x'}) + \frac{d^2}{dt^2}(\lambda(t)H_{x''}) = 0$$

or

$$2\lambda_0 x_1 + \lambda + \lambda x_2^2 - \lambda x_3^2 + \lambda'' = 0$$
$$2\lambda_0 x_2 + 2\lambda x_2 x_1 = 0$$
$$2\lambda_0 x_3 - 2\lambda x_3 x_1 = 0,$$
$$\lambda(0) - \lambda(\pi) = 0,\ \lambda'(0) + \lambda'(\pi) = 0$$

and gives us the series of spurious solutions $x_1 = a\sin t$, $x_2 = 0$, $x_3 = 0$, $\lambda = b\sin t$, $\lambda_0 = 0$, $a \in \mathbb{R}$. The mapping $G(x)$ is singular at these points $x_1 = a\sin t$, $x_2 = 0$, $x_3 = 0$ and $G'(a\sin t, 0, 0)$ is non surjective.
But $G(x)$ is 2−regular at the points $\bar{x} = (a\sin t, 0, 0)$ along $h = (\sin t, \sin t, -\sin t)$. Indeed, $Y_2 = \text{span}\ \{\sin t\}$,

$$G'(\bar{x})h + P_{Y_2}G''(\bar{x})[h]^2 = h'' + h + 2a\sin t \int_0^{\pi} (\sin^2 t - \sin^2 t)\sin^2 t dt = 0.$$

It means that $h \in \mathrm{Ker}G'(\bar{x}) \cap P_{Y_2}\mathrm{Ker}^2 G''(\bar{x})$ and

$$P_{Y_2}G''(\bar{x})h = 2a\sin t \int_0^\pi (\sin t(\cdot)_2 + \sin t(\cdot)_3)\sin^2 t dt.$$

We have

$$P_{Y_2}G''(\bar{x})h \begin{pmatrix} b\sin t \\ b\sin t \\ b\sin t \end{pmatrix} = 2ab\sin t \int_0^\pi 2\sin^4 t dt = Y_2, b \in \mathbb{R}$$

i.e. G is 2-regular at the points $\bar{x} = (a\sin t, 0, 0)$ along h. At these points $\bar{x}$ we can guarantee $\lambda_0 = 1$ in the $2-$factor Euler-Poisson equation

$$\begin{gathered} 2a\sin t + \lambda'' + \lambda = 0 \\ 2a\sin t \int_0^\pi \sin^3 \tau\lambda(\tau)d\tau = 0 \\ 2a\sin t \int_0^\pi \sin^3 \tau\lambda(\tau)d\tau = 0 \\ \lambda(0) - \lambda(\pi) = 0,\ \lambda'(0) + \lambda'(\pi) = 0 \end{gathered}$$

The first equation has no solutions for $a \neq 0$, which means that the point $\bar{x} = (a\sin t, 0, 0)^T$ is not a local solution of the considered problem.

Acknowledgements. Research of the second author is supported by the Russian Foundation for Basic Research Grant No 11-01-00786a and the Council for the State Support of Leading Scientific Schools Grant 5264.2012.1.

References

1. Brezhneva, O.A., Tret'yakov, A.A.: Optimality conditions for degenerate extremum problems with equality constraints. SIAM J. Contr. Optim. 42(2), 729–745 (2003)
2. Brezhneva, O.A., Tret'yakov, A.A.: Corrigendum: Optimality Conditions for Degenerate Extremum Problems with Equality Constraints. SIAM J. Contr. Optim. 48(5), 3670–3673 (2010)
3. Belash, K.N., Tret'yakov, A.A.: Methods for solving degenerate problems. USSR Comput. Math. and Math. Phys. 28, 90–94 (1988)
4. Korneva, I.T., Tret'yakov, A.A.: Application of the factor-analysis to the calculus of variations. In: Proceedings of Simulation and Analysis in Problems of Decision-Making Theory, pp. 144–162. Computing Center of Russian Academy of Sciences, Moscow (2002) (in Russian)
5. Prusińska, A., Tret'yakov, A.A.: P-order Necessary and Sufficient Conditions for Optimality in Singular Calculus of Variations. Discussiones Mathematicae, Differential Inclusions, Control and Optimization 30, 269–279 (2010)

Mathematical and Implementation Challenges Associated with Testing of the Dynamical Systems

Pawel Skruch

AGH University of Science and Technology,
Faculty of Electrical Engineering, Automatics,
Computer Science and Electronics, Department of Automatics,
al. A. Mickiewicza 30/B1, 30–059 Krakow, Poland
pawel.skruch@agh.edu.pl

Abstract. The paper presents mathematical and implementation challenges associated with testing of embedded software systems with dynamic behavior. These challenges are related to notation of tests, calculation of test coverage, implementation of a test comparator, and automatic generation of test cases. Some author's ideas and solutions are presented with the help of abstract models that describe behavior of the software systems. The models are represented using the state space (or input/state/output) notation. An application example is given to illustrate theoretical analysis and mathematical formulation.

Keywords: software system, model-based testing, dynamical system.

1 Introduction

Designing an embedded control system is a complex and error prone task. Within the last decades embedded systems have become increasingly sophisticated and their software content has grown rapidly. The increasing miniaturization of embedded control systems on the one hand and rising functional demands on the other hand require advanced and automated development and testing methodologies. In this context, model-based development (MBD) and model-based testing (MBT) approaches have the potential to facilitate the development of such systems under pressure of time-to-market constraints, quality assurance, and safety standards.

MBD is a process that provides the ability to graphically represent requirements, specification, and designs using domain-specific notations and simulate the resultant behavior for validation purposes. The code can be then generated from models, ranging from system skeletons to complete, deployable products. MBT is a related part that supports test generation from various kinds of models by application of a number of sophisticated methods.

Testing is the process of trying to discover every conceivable fault or weakness in a work product. The primary goal of the testing process is to found defects; the

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 538–546, 2013.

secondary goal is to show the system's compliance to its requirements. Testing can show that defects are present, but cannot prove that there are no defects [9]. Testing reduces the probability of undiscovered defects remaining in the software but, even if no defects are found, it is not a proof of correctness. Poorly tested systems may cost producers billions of dollars annually especially when defects are found by end users in production environments [7], [8], [13]. Barry Boehm's research analysis [4] indicates that the cost of removing a software defect grows exponentially for each stage of the development life cycle in which it remains undiscovered. Boris Beizer [2] estimates that 30 up to 90 percentage of the effort in put into testing. Another research project conducted by the United States Department of Commerce, National Institute of Standards and Technology [10] estimated that software defects cost the U.S. economy $60 billion per year.

Exhaustive testing is impossible what means that testing everything (all combinations of inputs and preconditions) is not feasible expect for trivial cases. This is valid in particular for software systems with dynamic behavior. The dynamic systems are modeled by difference or differential equations and have usually infinitely many states. Testing dynamic aspects of such systems requires tests that utilize time continuous input signals and time continuous output signals (even when the system is digitally processed). The process of selecting just a few of the many possible scenarios to be tested is a difficult and challenging task and currently is most often based on qualitative best engineering judgment.

In this paper, testing problem as well as test artifacts for software systems with dynamic behavior are formulated using the mathematical formalism. The main results concern the concept of testing with a model as an oracle (section 2), a proposal for test notation (section 4), an implementation of a test comparator (section 5), a calculation of test coverage (section 6), and a selection of tests (sections 7). An example (section 8) is given to present a perspective on the applicability of the approach for industrial projects.

2 Concept of Testing with a Model as an Oracle

The model of a software system shall specify the system's behavior in a clear and unambiguous form. It can be used in computer simulations in an early phase of the development to validate the system concept, calibrate parameters, and optimize the system performance. In the next phase, the physical system is designed (i.e., hardware and software) that shall meet the requirements specified by the model. Testing process shall be considered as the last phase in the development process that allows verifying that the physical system behavior is identical to that observed during computer simulations. When the tests failed then the system needs to be redesigned. The physical system that is being tested for the correct operation is often referred to as the system under test (SUT).

The model fully represents the requirements therefore it can be used an oracle to assess if the algorithm implemented in the electronic control unit (ECU) being tested correctly implements the requirements. The term *test oracle* describes a source to determine expected results to compare with the actual result of the

SUT [1]. The approach of a validated model being used as an oracle (the block *Mathematical Model of the SUT* on figure 1) is very popular in industry and often applied. The execution of a test case consists of exciting the system using actuators to simulate its working conditions and measuring the system's response in terms of electrical signals, motion, force, strain, etc. The signals are physical in case of the SUT and virtual in case of the model. The approach stipulates that the same inputs $\boldsymbol{u}(\cdot)$ are applied to both the SUT and to the model. Next, the responses from the SUT $\boldsymbol{y}_{\mathrm{s}}(\cdot)$ and from the model $\boldsymbol{y}(\cdot)$ are compared by a test comparator to determine whether a test case has passed or failed.

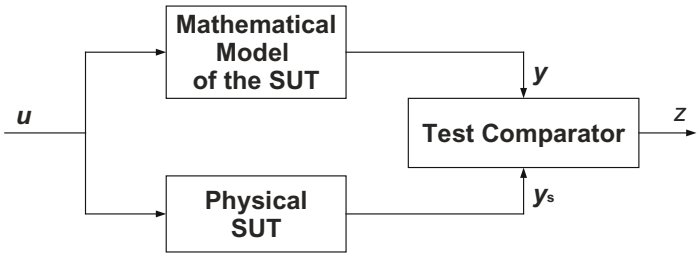


Fig. 1. Testing approach of a validated model being used as an oracle

3 Mathematical Model of the System under Test

The state space (or input/state/output) representation provides a convenient way to model and analyze dynamical systems. The state space model consists of a set of input, output, and internal state variables that are expressed as vectors. The relationship between inputs, outputs, and internal states in a finite-dimensional, time-invariant, nonlinear system with continuous-time parameter can be specified by the following equations:

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}, t), \quad \boldsymbol{x}(0) = \boldsymbol{x}_0 \ , \tag{1}$$

$$\boldsymbol{y}(t) = \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{u}, t) \ , \tag{2}$$

where $\boldsymbol{x}(t) \in X \subset \mathbb{R}^n$ refers to the internal state, $\boldsymbol{u}(t) \in U \subset \mathbb{R}^r$ refers to the input state, $\boldsymbol{y}(t) \in Y \subset \mathbb{R}^m$ refers to the output state, the independent variable $t > 0$ is time, $\boldsymbol{x}_0 \in \mathbb{R}^n$ is the given initial condition, $\boldsymbol{f} : \mathbb{R}^n \times \mathbb{R}^r \times \mathbb{R} \to \mathbb{R}^n$ denotes a mathematical relationship describing the system behavior, $\boldsymbol{g} : \mathbb{R}^n \times \mathbb{R}^r \times \mathbb{R} \to \mathbb{R}^m$ determines the output, X is the internal state space, Y is the output state space, U is called the input state space, $\mathbb{R}^n$, $\mathbb{R}^m$, $\mathbb{R}^r$ are real vector spaces of column vectors, n, m, r are positive integers that determine numbers of internal state, output, and input variables, respectively.

The physical and implementation constraints imposed by computer system resources lead to the assumption that the spaces U, X, and Y shall be bounded. The assumption means that each space is contained in a ball of finite radius.

4 Test Notation

A test case can be considered as a set of inputs, execution preconditions, and expected outcomes developed for a particular objective, such as to exercise a particular program path or to verify compliance with a specific requirement [15]. Adapting this definition to the state space modeling concept of the SUT (1), (2), a single test case $\mathrm{T}_{\mathrm{case}}^{(j)}$ can be defined as

$$\mathrm{T}_{\mathrm{case}}^{(j)} = \left\{ T^{(j)}, \boldsymbol{x}_0^{(j)}, \boldsymbol{u}^{(j)}(\cdot), \boldsymbol{y}^{(j)}(\cdot) \right\} , \tag{3}$$

in case of black-box testing [3], or

$$\mathrm{T}_{\mathrm{case}}^{(j)} = \left\{ T^{(j)}, \boldsymbol{x}_0^{(j)}, \boldsymbol{u}^{(j)}(\cdot), \boldsymbol{x}^{(j)}(\cdot), \boldsymbol{y}^{(j)}(\cdot) \right\} , \tag{4}$$

in case of gray-box testing [12]. Here, $\boldsymbol{u}^{(j)} : [0, T^{(j)}] \to \mathbb{R}^r$ is an input function applied to the SUT, $\boldsymbol{x}^{(j)} : [0, T^{(j)}] \to \mathbb{R}^n$ is an expected state function, and $\boldsymbol{y}^{(j)} : [0, T^{(j)}] \to \mathbb{R}^m$ is an expected output function within the execution time window $[0, T^{(j)}]$ when the system starts from an initial condition $\boldsymbol{x}_0^{(j)}$, $j = 1, 2, \ldots, N$ is a label to indicate different test cases. A collection of one or more test cases forms a test suite $\mathrm{T}_{\mathrm{suite}} = \left\{ \mathrm{T}_{\mathrm{case}}^{(1)}, \mathrm{T}_{\mathrm{case}}^{(2)}, \ldots, \mathrm{T}_{\mathrm{case}}^{(N)} \right\}$.

5 Test Comparator Implementation

The test comparator can be considered as a tool that implements a mechanism for determining whether a test has passed or failed [5]. In the concept, illustrated on figure 1, this tool compares the actual output $\boldsymbol{y}_{\mathrm{s}}(\cdot)$ produced by the SUT with the expected output $\boldsymbol{y}(\cdot)$ produced by the model. If the actual output is within a predefined tolerance range ϵ relative to the expected output, then the test is qualified as *pass* ($z = 0$, *system ok*), otherwise the test is qualified as *fail* ($z = 1$, *system error*). A possible practical realization of the comparison function z for a given test case $\mathrm{T}_{\mathrm{case}}^{(j)}$ is presented below:

$$z(\mathrm{T}_{\mathrm{case}}^{(j)}) = \begin{cases} 0 & \text{if } \forall_{t \in [0, T^{(j)}]} \left\| \boldsymbol{y}^{(j)}(t) - \boldsymbol{y_{\mathrm{s}}}^{(j)}(t) \right\| < \epsilon \left\| \boldsymbol{y}^{(j)}(t) \right\| , \\ 1 & \text{otherwise} . \end{cases} \tag{5}$$

In the formula (5) the standard Euclidean norm $\| \cdot \|$ has been used to measure the distance between two points in the space $\mathbb{R}^m$.

6 Test Coverage Calculation

The degree to which a given test suite $\mathrm{T}_{\mathrm{suite}}$ addresses all specified requirements for a given system is determined by a test coverage measure [15]. The most obvious quantification of the system's behavior exercised by the test suite is computed by dividing the number of the system states explored by the test

suite by the cardinality of the entire state space. However, the formula has limited usefulness for dynamical systems because the state space for such systems contains usually infinite number of states. In such situation, one of the possible ways out is to transform the internal state space X into another one $X_{\boldsymbol{h}}$ that contains countable number of elements.

The test coverage $C_{\boldsymbol{h}}$ of the test suite $\mathrm{T}_{\mathrm{suite}} = \{\mathrm{T}_{\mathrm{case}}^{(1)}, \mathrm{T}_{\mathrm{case}}^{(2)}, \ldots, \mathrm{T}_{\mathrm{case}}^{(N)}\}$ can be defined as follows [14]

$$C_{\boldsymbol{h}}(\mathrm{T}_{\mathrm{suite}}) = \frac{\left| \bigcup_{j=1}^{N} V_{\boldsymbol{h}}(\mathrm{T}_{\mathrm{case}}^{(j)}) \right|}{|X_{\boldsymbol{h}}|}, \tag{6}$$

where

$$X_{\boldsymbol{h}} = \{\boldsymbol{i} \in \mathbb{Z}^n : \exists_{\boldsymbol{x} \in X} : \boldsymbol{x} \in G_{\boldsymbol{h}}(\boldsymbol{i})\} \tag{7}$$

is the transformed internal state space, $\boldsymbol{h} = [h_1\, h_1\, \ldots\, h_n]^{\mathrm{T}}$, $h_k > 0$ for $k = 1, 2, \ldots, n$, $\mathbb{Z}$ stands for the set of integers,

$$G_{\boldsymbol{h}}(\boldsymbol{i}) = \left\{ \boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{x} = [x_1\, x_2\, \ldots\, x_n]^{\mathrm{T}},\ \left\lfloor \frac{x_k}{h_k} \right\rfloor = i_k,\ k = 1, 2, \ldots, n \right\} \tag{8}$$

denotes a partition with the size $\boldsymbol{h}$ in the space $\mathbb{R}^n$, $\left\lfloor \frac{x_k}{h_k} \right\rfloor$ is the largest integer not greater than $\frac{x_k}{h_k}$,

$$V_{\boldsymbol{h}}(\mathrm{T}_{\mathrm{case}}^{(j)}) = \left\{ \boldsymbol{i} \in X_{\boldsymbol{h}} : \exists_{t \in [0, T^{(j)}]} : \boldsymbol{x}^{(j)}(t) \in G_{\boldsymbol{h}}(\boldsymbol{i}) \right\} \tag{9}$$

is a set of states of the transformed internal state space covered by the test case $\mathrm{T}_{\mathrm{case}}^{(j)}$. It should be noticed that the sum

$$V_{\boldsymbol{h}}(\mathrm{T}_{\mathrm{suite}}) = \bigcup_{j=1}^{N} V_{\boldsymbol{h}}(\mathrm{T}_{\mathrm{case}}^{(j)}) \tag{10}$$

will contain the information about the internal states covered by the test suite $\mathrm{T}_{\mathrm{suite}}$.

The proposed test coverage measure is defined using a partition (or discretization) of the system internal state space. The partition forms a rectangular grid and, roughly speaking, the test coverage is defined by the number of the grid boxes visited by the system state during a test.

7 Conformance Test Selection Method

The section presents a proposal of the algorithm for generating test cases. The general principle of the algorithm is to create input functions $\boldsymbol{u}(\cdot)$ for which the system trajectories $\boldsymbol{x}(\cdot)$ cross every element $G_{\boldsymbol{h}}(\boldsymbol{i})$ of the space $X_{\boldsymbol{h}}$. The selection and completeness of test cases is quantified by the coverage metric (6). Test cases are selected to check that the functional specification (here in the form of the mathematical model) is correctly implemented, which is variously referred to in the literature as conformance testing [5], correctness testing [6], or functional testing [15].

Algorithm 1. Conformance test selection method

1: $\boldsymbol{h} = [h_1\, h_2\, \ldots\, h_n]^{\mathrm{T}}$, $h_1, h_2, \ldots, h_n > 0$, $\delta \in (0, 1]$, $T > 0$
2: $\mathrm{T}_{\mathrm{suite}} := \emptyset$, $V_{\boldsymbol{h}}(\mathrm{T}_{\mathrm{suite}}) := \emptyset$, $C_{\boldsymbol{h}}(\mathrm{T}_{\mathrm{suite}}) = 0$, $j := 0$
3: **while** $C_{\boldsymbol{h}}(\mathrm{T}_{\mathrm{suite}}) \leq \delta$ **do**
4: Find $\boldsymbol{x}_a \in G_{\boldsymbol{h}}(\boldsymbol{i}_a)$, $\boldsymbol{x}_b \in G_{\boldsymbol{h}}(\boldsymbol{i}_b)$ where $\boldsymbol{i}_b \in X_{\boldsymbol{h}} \backslash V_{\boldsymbol{h}}(\mathrm{T}_{\mathrm{suite}})$
5: Calculate the control function $\boldsymbol{u}^*(\cdot)$ that steers the system from the initial state $\boldsymbol{x}(0) = \boldsymbol{x}_a$ to the final state $\boldsymbol{x}(T) = \boldsymbol{x}_b$ at finite time T
6: Calculate the trajectory $\boldsymbol{x}^*(\cdot)$ and output function $\boldsymbol{y}^*(\cdot)$
7: $j := j + 1$
8: $\mathrm{T}^{(j)}_{\mathrm{case}} := \{T^{(j)}, \boldsymbol{x}_0^{(j)}, \boldsymbol{u}^{(j)}(\cdot), \boldsymbol{x}^{(j)}(\cdot), \boldsymbol{y}^{(j)}(\cdot)\}$, where $T^{(j)} := T$, $\boldsymbol{x}_0^{(j)} := \boldsymbol{x}_a$, $\boldsymbol{u}^{(j)}(\cdot) := \boldsymbol{u}^*(\cdot)$, $\boldsymbol{x}^{(j)}(\cdot) := \boldsymbol{x}^*(\cdot)$, $\boldsymbol{y}^{(j)}(\cdot) := \boldsymbol{y}^*(\cdot)$
9: $\mathrm{T}_{\mathrm{suite}} := \mathrm{T}_{\mathrm{suite}} \cup \mathrm{T}^{(j)}_{\mathrm{case}}$
10: Calculate $V_{\boldsymbol{h}}(\mathrm{T}_{\mathrm{suite}})$ and $C_{\boldsymbol{h}}(\mathrm{T}_{\mathrm{suite}})$
11: **end while**

8 Embedded PID Controller Example

An embedded PID controller is a system that can be considered as a combination of computer hardware and software designed to perform a dedicated control function. The PID controller works in a closed-loop system (figure 2) and attempts to minimize the error $e(t)$ by adjusting the control input $s(t)$. The error

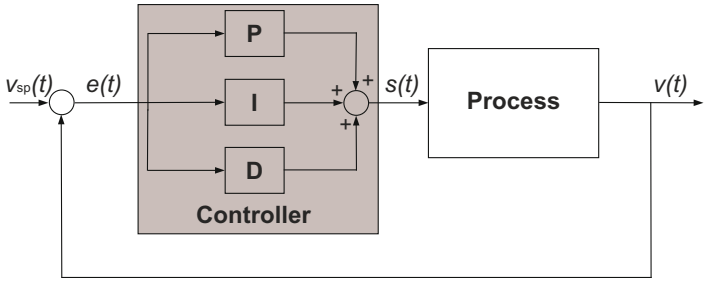


Fig. 2. A block diagram of the closed-loop system with the PID controller

is calculated as the difference between a measured process output $v(t)$ and a desired set point $v_{\mathrm{sp}}(t)$. The control signal is a result of the following calculation

$$s(t) = K\left(e(t) + \frac{1}{T_{\mathrm{i}}}\int_0^t e(\tau)\,\mathrm{d}\tau + T_{\mathrm{d}}\frac{\mathrm{d}e(t)}{\mathrm{d}t}\right)\,, \tag{11}$$

where $K = 3.6$ is proportional gain, $T_{\mathrm{i}} = 1.81\,[\mathrm{s}]$ is integral time, $T_{\mathrm{d}} = 0.45\,[\mathrm{s}]$ is derivative time. The control signal is thus a sum of three terms: the P-term (which is proportional to the error), the I-term (which is proportional to the integral of the error), and D-term (which is proportional to the derivative of the

error). The parameters K, T_i, and T_d can be obtained using the Ziegler-Nichols algorithm [16]. The model of the process to be controlled has been omitted in the example for the purpose of clarity and easy readability.

The algorithm 1 can be used to generate a set of conformance test cases. Before its execution, the equation (11) needs to be rewritten to the form

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}u(t), \quad \boldsymbol{x}(0) = \boldsymbol{0}\ , \tag{12}$$

$$\boldsymbol{y}(t) = \boldsymbol{C}\boldsymbol{x}(t)\ , \tag{13}$$

where $\boldsymbol{x}(t) = [x_1(t)\ x_2(t)]^\mathrm{T}$, $x_1(t) = \int_0^t e(\tau)\,\mathrm{d}\tau$, $x_2(t) = \dot{x}_1(t)$,

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -(T_\mathrm{i}T_\mathrm{d})^{-1} & -T_\mathrm{d}^{-1} \end{bmatrix}\ , \quad \boldsymbol{B} = \begin{bmatrix} 0 \\ (KT_\mathrm{d})^{-1} \end{bmatrix}\ , \quad \boldsymbol{C} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\ . \tag{14}$$

Then, the algorithm has been implemented and executed with the following parameters: $\boldsymbol{h} = [0.3,\ 0.2]^\mathrm{T}$ (size of the partition), $\delta = 0.7$ (acceptable coverage level), $T = 20\,[\mathrm{s}]$ (test execution time), $-1.5 \le x_1(t) < 1.5$, $-1 \le x_2(t) < 1$ (system implementation constraints). The test suite that guarantees the coverage level higher than δ consists of 10 test cases. The elements of these test cases are detailed in table 1, figures 3 and 4.

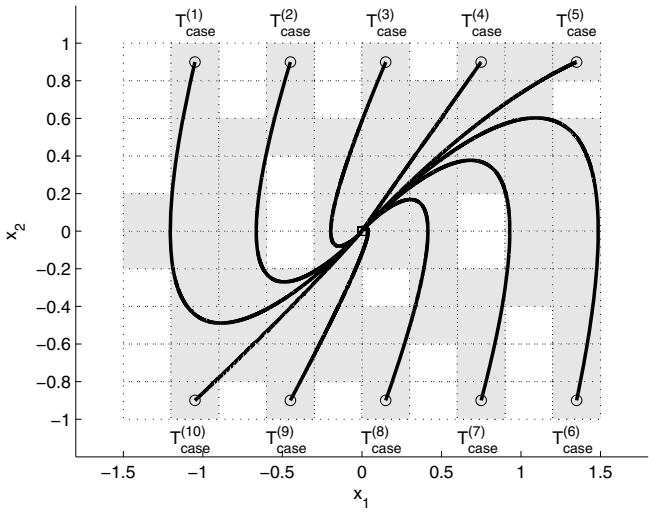


Fig. 3. Trajectories $\boldsymbol{x}^{(j)}$, $j = 1, 2, \ldots, 10$ and elements (*gray rectangles*) of the transformed state space covered by the test cases $T_\mathrm{case}^{(j)}$. The trajectories start in □ and end in ○.

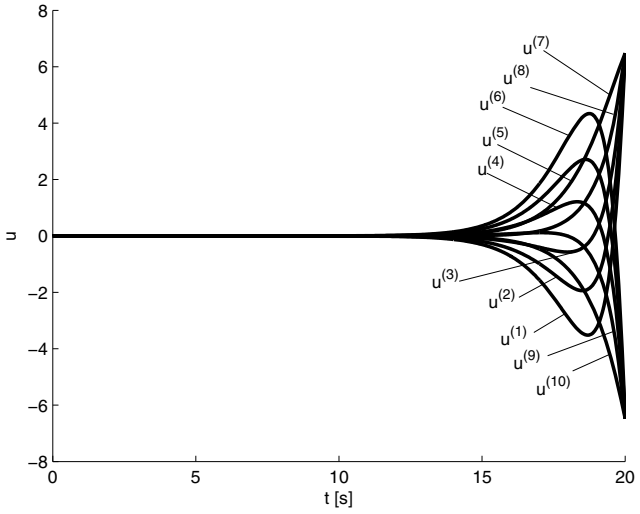


Fig. 4. Illustration of the input functions $u^{(j)}(\cdot)$, $j = 1, 2, \ldots, 10$ belonging to the test cases $\mathrm{T}_{\mathrm{case}}^{(j)}$

Table 1. An example test report for the test suite $\mathrm{T}_{\mathrm{suite}} = \{\mathrm{T}_{\mathrm{case}}^{(1)}, \mathrm{T}_{\mathrm{case}}^{(2)}, \ldots, \mathrm{T}_{\mathrm{case}}^{(10)}\}$ that guarantees the coverage level $C_{\boldsymbol{h}} > \delta$, where $\boldsymbol{h} = [0.3,\ 0.2]^{\mathrm{T}}$, $\delta = 0.7$ (70%). The notation used in the last column means $\mathrm{T}_{\mathrm{suite}}^{(j)} = \{\mathrm{T}_{\mathrm{case}}^{(1)}, \mathrm{T}_{\mathrm{case}}^{(2)}, \ldots, \mathrm{T}_{\mathrm{case}}^{(j)}\}$.

j	$T^{(j)}$[s]	$\boldsymbol{x}_0^{(j)\mathrm{T}}$	$u^{(j)}(\cdot)$	$\boldsymbol{x}^{(j)}(\cdot)$	$\boldsymbol{y}^{(j)}(\cdot)$	$C_{\boldsymbol{h}}(\mathrm{T}_{\mathrm{case}}^{(j)})$	$C_{\boldsymbol{h}}(\mathrm{T}_{\mathrm{suite}}^{(j)})$
1	20	[0, 0]	fig. 4	fig. 3		0.16	0.16
2	20	[0, 0]	fig. 4	fig. 3		0.12	0.24
3	20	[0, 0]	fig. 4	fig. 3		0.08	0.30
4	20	[0, 0]	fig. 4	fig. 3		0.07	0.36
5	20	[0, 0]	fig. 4	fig. 3		0.09	0.40
6	20	[0, 0]	fig. 4	fig. 3		0.14	0.49
7	20	[0, 0]	fig. 4	fig. 3		0.12	0.58
8	20	[0, 0]	fig. 4	fig. 3		0.08	0.65
9	20	[0, 0]	fig. 4	fig. 3		0.08	0.70
10	20	[0, 0]	fig. 4	fig. 3		0.09	0.73

9 Conclusions

In spite of continuing research on test approaches for continuous and mixed discrete-continuous systems, there is still a lack for patterns, processes, methodologies, and tools that effectively support automatic generation and selection of the correct test cases for such systems. The model-based approach presented in the paper looks promising. The functional model of the system under test can be used as an oracle providing the capabilities to assess the results of test cases

in an automatic way and also in test generation algorithms. Additional aspects, such as test notation, implementation of a test comparator, and coverage analysis have been discussed in the paper in order to have complete set of tools and mathematical methods for testing software systems with dynamic behavior. The example has been used to validate the concept and to have a perspective on its applicability for industrial projects.

Acknowledgments. This work was supported by the National Science Centre (Poland) – project No N N514 644440.

References

1. Adrion, W., Brandstad, J., Cherniabsky, J.: Validation, Verification and Testing of Computer Software. Computing Surveys 14, 159–192 (1982)
2. Beizer, B.: Software Testing Techniques, 2nd edn. Van Nostrand Reinhold, Boston (1990)
3. Beizer, B.: Black-Box Testing. Techniques for Functional Testing of Software and Systems. John Willey & Sons, New York (1995)
4. Boehm, B.: Software Engineering Economics. Prentice Hall, Englewood Cliffs (1981)
5. International Software Testing Qualifications Board (ISTQB): Standard Glossary of Terms Used in Software Testing, Version 2.1 (2010), http://www.astqb.org
6. Kaner, C., Faulk, J., Nguyen, H.Q.: Testing Computer Software, 2nd edn. John Willey & Sons, New York (1995)
7. Leveson, N.G., Turner, C.S.: An Investigation of the Therac-25 Accidents. IEEE Computer 27, 18–41 (1993)
8. Lions, J.L.: ARIANE 5. Flight 501 Failure. Ariane 501 Inquiry Board Report. Technical report, Paris, France (1996)
9. Myers, G.: The Art of Software Testing, 2nd edn. John Willey & Sons, New York (2004)
10. National Institute of Standards & Technology, U.S. Department of Commerce: The Physiology of the Grid: The Economic Impacts of Inadequate Infrastructure for Software Testing, Final Report. Technical report, North Carolina, USA (2002)
11. Nelder, J.A., Mead, R.: A Simplex Method for Function Minimization. The Computer Journal 7, 308–313 (1965)
12. Patton, R.: Software Testing, 2nd edn. Sams, Indianapolis (2005)
13. Skeel, R.: Roundoff Error and the Patriot Missile. Society for Industrial and Applied Mathematics (SIAM) News 25, 11 (1992)
14. Skruch, P.: A Coverage Metric to Evaluate Tests for Continuous-Time Dynamic Systems. Central European Journal of Engineering 1, 174–180 (2011)
15. The Institute of Electrical and Electronics Engineers, Inc.: IEEE Standard Glossary of Software Engineering Terminology, IEEE Std 610.12-1990 (1990), http://www.standards.ieee.org
16. Ziegler, J., Nichols, N.: Optimum Settings for Automatic Controllers. Transactions of ASME 64, 759–768 (1942)

Numerical Parameters Estimation in Models of Pollutant Transport with Chemical Reaction

Fabiana Zama*, Roberta Ciavarelli, Dario Frascari, and Davide Pinelli**

Bologna University

Abstract. In this work we present an iterative algorithm for solving a parameter identification problem relative to a system of diffusion, convection and reaction equations. The parameters to estimate are the retardation factors, diffusivity, reaction and transport coefficients relative to a model of pollutant transport with chemical reaction. The proposed method solves the nonlinear least squares problem by means of a sequence of constrained optimization problems. The algorithm does not depend on the type of discretization method used to solve the state equation. The results reported in the numerical tests show the efficiency of the algorithm in terms of performance and solution quality.

1 Introduction

Parameter estimation is a very important topic in applied sciences and chemical engineering: an overview of methods and applications can be found in [3]. The modeling of pollutant transport with (bio)chemical reaction gives raise to partial differential systems which are usually very complex. Therefore there is a need for efficient algorithms for solving parameter estimation problems.

In this work we consider the parameter estimation problem for a system of reaction, diffusion and transport equations:

$$\frac{\partial U}{\partial t} = \nabla \cdot (D \nabla U) - V \nabla U + R(U), \tag{1}$$

where $U \equiv (u_1, u_2, \ldots, u_{Nc})^t$ represents the concentration in the state variable (x,t), $x \in [a,b]$, $t \in [0,T]$. The coefficients D and V represent the diffusion coefficient and the fluid velocity, while the reaction term is represented by the function R which depends on the solution U. Free flow condition is assumed at outlet boundary and pulse functions are given at inlet boundary. Homogenous initial conditions are assumed. A typical model of pollutant transport and biodegradation is illustrated by Frascari et al. [8].

The parameter estimation problem can be formalized as a constrained optimization problem:

$$\min_q J(U,q) \quad s.t. \quad c(U,q) = 0$$

* Department of Mathematics.
** Department of Chemical, Mining and Environmental Engineering.

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 547–556, 2013.

where $c(U,q) = 0$ represents the governing PDEs system (1) or state equation, the objective function $J(U,q)$ is the distance between the measurements $y \in Y$ and the solution U of the state equation $c(U,q) = 0$, corresponding to the parameter $q \in Q$ in the measurement points. Introducing the reduced observation operator $F : Q \longrightarrow Y$, that maps the unique solution of the state equation $U(q)$ into the measurements space Y, we define the equivalent nonlinear least squares problem:

$$\min_q \frac{1}{2}\|F(q) - y\| \tag{2}$$

where $\|\cdot\|$ is the euclidean norm throughout the paper. The ill posedness of the problem is well known [6] and different methods are proposed in the literature to obtain stable solutions in the presence of noisy data.

In this paper we propose an iterative method that solves the nonlinear least squares problem (2) by computing a sequence of constrained optimization problems. The proposed algorithm computes the solution q and the proper smoothing parameters, suitable to overcome the instability problems that arise in the solution of nonlinear least squares problems. The necessary starting values and tolerance parameters are computed using information obtained by the given measurements.

The paper is organized as follows. In section 2 we formulate the parameter estimation problem as an optimization problem and describe the discrete optimization algorithm. In section 3 the described algorithm is tested to evaluate both efficiency and solution quality.

2 The Optimization Algorithm

Aim of this section is to describe the discrete optimization algorithm for parameter estimation in the contest of transport and chemical reaction.

Given a set of measurements $\mathbf{y} \in \mathbb{R}^{Nm}$ relative to the concentration of compound u_i $(i = 1, \ldots, N_c)$ at points $(t_j, x_j) \in [0,T]\times[a,b]$. The problem consists in finding the parameters $\mathbf{q} \in \mathbb{R}^{Np}$ whose image $F(\mathbf{q}) \in \mathbb{R}^{Nm}$ is the least squares approximation of the data $\mathbf{y}$. The discrete nonlinear least squares problem is given by:

$$\min_{\mathbf{q}} J(\mathbf{q}), \quad J(\mathbf{q}) \equiv \frac{1}{2}\|F(\mathbf{q}) - \mathbf{y}\| \tag{3}$$

By applying the first order conditions we obtain the nonlinear system:

$$\mathcal{J}_F^t(\mathbf{q})(F(\mathbf{q}) - \mathbf{y}) = 0, \quad \mathcal{J}_F(\mathbf{q}) \in \mathbb{R}^{Nm\times Np}, \quad (\mathcal{J}_F(\mathbf{q}))_{i,j} = \frac{\partial F_i}{\partial q_j}(\mathbf{q})$$

This problem is solved iteratively by setting an initial guess $\mathbf{q}^{(0)}$ and defining a direction $\mathbf{s}^{(k)}$ s.t. $\mathbf{q}^{(k+1)} = \mathbf{q}^{(k)} + \mathbf{s}^{(k)}$, $k \geq 0$, where $\mathbf{s}^{(k)}$ is obtained as solution of the linear system $H_k\mathbf{s}^{(k)} = -G_k^t\mathbf{r}_k$ where

$$G_k \equiv \mathcal{J}_F(\mathbf{q}^{(k)}), \quad \mathbf{r}_k \equiv F(\mathbf{q}^{(k)}) - \mathbf{y}$$

and H_k is the Hessian of the objective function $J(\mathbf{q}^{(k)})$ in (3). Although the dimensions of the linear system are small ($N_p \times N_p$), the computation of second order information is very expensive. In this work we use first order approximation given by the Gauss Newton method which is equivalent to defining $\mathbf{s}^{(k)}$ as solution of the linearized problem:

$$\min_{\mathbf{s}} \frac{1}{2}\|G_k\mathbf{s} + \mathbf{r}_k\|, \quad k \geq 0. \tag{4}$$

It is well known that instabilities often occur in the solution of this unconstrained linear least squares problem and it is necessary to introduce some smoothing technique to obtain stable solutions in the presence of data noise. A possible strategy is to add a constraint to the problem and compute the direction $\mathbf{s}^{(k)}$ as solution of the following constrained optimization problem:

$$\min_{\mathbf{s}} \frac{1}{2}\|G_k\mathbf{s} + \mathbf{r}_k\|, \quad s.t. \ \|\mathbf{s}\| \leq \Delta_k \tag{5}$$

where Δ_k represents the smoothness level required in the solution $\mathbf{s}$. The algorithm that we propose here allows us to solve problem (5) iteratively by computing the approximate solution of the equivalent dual lagrangian problem. The smoothed solution $\mathbf{s}^{(k)}$ is obtained by applying a few steps of Constrained Least Squares Regularization `CLSRit` algorithm [1] . Furthermore we define a suitable size of the initial trust region Δ_0, by using the problem data and we update it to compute $\Delta_k > 0$ by means of the trust region update method [4].

This algorithm can be viewed as an implementation of the Levemberg Marquardt Trust Region method, widely used both in constrained optimization and in the contest of parameter estimation [5], [4]. The Trust Region Constrained Least Squares Regularization `TRCLSR`, reported in table 1, can be split in the following steps:

- Computation of the initial trust region size Δ_0 (paragraph 2.1).
- Computation of the direction $\mathbf{s}^{(k)}$ (paragraph 2.2).
- Update of the trust region size Δ_k (paragraph 2.1).
- Solution update and stopping rules (paragraph 2.3).

The following input parameters are required: the starting value for the unknown parameters $\mathbf{q}^{(0)}$, the relative tolerance τ_J of the objective function J, the absolute tolerance τ_s of the step size $\|\mathbf{s}^{(k)}\|$, the problem data $\mathbf{y}$ and the function F that maps the parameters into the data space.

2.1 Update of Δ_k

An initial estimate of the size of Trust Region parameter Δ_0 can be obtained by computing a Tikhonov [7] regularized solution of problem (4) with regularization parameter $\alpha = 10^{-6}$ i.e.:

$$\Delta_0 = \|\bar{\mathbf{s}}^{(0)}\|, \quad \bar{\mathbf{s}}^{(0)} \quad s.t. \quad (G_0^t G_0 + \alpha I)\bar{\mathbf{s}}^{(0)} = G_0^t(F(\mathbf{q}^{(0)}) - \mathbf{y})$$

Table 1. Algorithm TRCLSR

Algorithm 1 (TRCLSR(F, $\mathbf{y}$, $\mathbf{q}^{(0)}$, τ_J, τ_s))

Compute $U^{(0)}$ *solving the PDE state equation* $c(U^{(0)}, \mathbf{q}^{(0)}) = 0$;
Compute Jacobian G_0 *as in subsection 2.4*
Compute $\bar{\mathbf{s}}^{(0)}$ *s.t.* $(G_0^t G_0 + 1.e-6I)\bar{\mathbf{s}}^{(0)} = G_0^t(F(\mathbf{q}^{(0)}) - \mathbf{y})$;
Set $\Delta_0 = \|\bar{\mathbf{s}}^{(0)}\|$;
$k = 0$
repeat
  Compute direction $\mathbf{s}^{(k)}$ *as in subsection 2.2*
  Compute Δ_{k+1} *as in subsection 2.1*
  Compute $\mathbf{q}^{(k+1)}$ *as in subsection 2.3*;
  Solve PDE state equation $c(U^{(k+1)}, \mathbf{q}^{(k+1)}) = 0$;
  Compute G_{k+1} *(subsection 2.4)*
  $k = k + 1$
until $\left(|J(\mathbf{q}^{(k+1)}) - J(\mathbf{q}^{(k)})| < \tau_J |J(\mathbf{q}^{(k)})| \quad or \quad \|\mathbf{s}^{(k)}\| < \tau_s\right)$

the value of the parameter α should be small enough to avoid the instability of the linear system without smoothing too much the solution. At each step k the update Δ_{k+1} is performed following the Trust Region algorithm (see algorithm 4.1 in [4]).

2.2 Computation of the Direction $\mathbf{s}^{(k)}$

The direction $\mathbf{s}^{(k)}$ is computed by solving problem (5) in its equivalent lagrangian dual form [2]:

$$\max_{\lambda} \Phi(\lambda), \quad \Phi(\lambda) \equiv \min_{\mathbf{q}} \mathcal{L}(\mathbf{s}, \lambda). \tag{6}$$

where $\mathcal{L}$ is the lagrangian function: $\mathcal{L}(\mathbf{s}, \lambda) \equiv \frac{1}{2}\|G_k\mathbf{s} + \mathbf{r}_k\| + \lambda\,(\|\mathbf{s}\| - \Delta_k)$. Solving the dual problem (6) requires to find $\hat{\lambda}$ s.t. $\|\mathbf{s}(\hat{\lambda})\| = \Delta_k$ where $\mathbf{s}(\hat{\lambda})$ is the solution of the following linear system

$$(G_k^t G_k + \hat{\lambda} I)\mathbf{s}(\hat{\lambda}) = -G_k^t \mathbf{r}_k$$

The nonlinear equation $\|\mathbf{s}(\lambda)\| - \Delta_k = 0$ is solved by the hybrid method proposed in [1]. Given a starting value $\lambda_0 > 0$ s.t. $\mathbf{s}(\lambda_0) \leq \Delta_k$ and a value $k_s > 2$ s. t. $0 < \lambda_0 < \lambda_{k_s} < \hat{\lambda}$, compute $(\mathbf{s}_\ell, \lambda_\ell)$ where $\lambda_0 = \|\mathbf{r}_k\|$ and

$$\lambda_\ell = \lambda_{\ell-1} + \mathcal{S}_{\ell-1}, \quad \ell \geq 1$$

where

$$\mathcal{S}_{\ell-1} = \begin{cases} sign(\|\mathbf{s}(\lambda_{\ell-1})\| - \Delta_k)\dfrac{\lambda_0}{2^{\ell-1}} & \ell \leq k_s + 1 \\ \dfrac{\|\mathbf{s}(\lambda_{\ell-1})\| - \Delta_k}{\|\mathbf{s}(\lambda_{\ell-1})\| - \|\mathbf{s}(\lambda_{\ell-2})\|}\,(\lambda_{\ell-2} - \lambda_{\ell-1}) & \ell > k_s + 1 \end{cases} \tag{7}$$

where $\mathbf{s}(\lambda_\ell)$ satisfies: $(G_k^t G_k + \lambda_\ell I)\mathbf{s}(\lambda_\ell) = -G_k^t \mathbf{r}_k$ Under the given hypotheses it is proven that λ_ℓ converges to the solution $\hat{\lambda}$ of the dual problem (6) and the sequence $\{\mathbf{s}_\ell\}$ converges to $\hat{\mathbf{s}} \equiv \mathbf{s}(\hat{\lambda})$ which is the solution of the problem (5) [1].

2.3 Solution Update $\mathbf{q}^{(k+1)}$ and Stopping Conditions

After the computation of each direction $\mathbf{s}^{(k)}$, the solution $\mathbf{q}^{(k+1)}$ is updated as follows:

$$\mathbf{q}^{(k+1)} = \begin{cases} \mathbf{q}^{(k)} + \mathbf{s}^{(k)} & \rho_k > \eta, \quad 0 \leq \eta \leq 0.25 \\ \mathbf{q}^{(k)} & otherwise \end{cases}$$

where the parameter ρ_k is given by

$$\rho_k = \frac{J(\mathbf{q}^{(k)}) - J(\mathbf{q}^{(k)} + \mathbf{s}^{(k)})}{m_k(0) - m_k(\mathbf{s}^{(k)})}$$

and it represents the ratio between the actual reduction $J(\mathbf{q}^{(k)}) - J(\mathbf{q}^{(k)} + \mathbf{s}^{(k)})$ and the reduction predicted in J by the model function m_k:

$$m_k(\mathbf{s}) \equiv J(\mathbf{q}^{(k)}) + \mathbf{s}^t G_k^t (F(\mathbf{q}^{(k)}) - \mathbf{y}) + \mathbf{s}^t G_k^t G_k \mathbf{s}.$$

The iterations are stopped when the relative reduction of the objective function J is below a given tolerance τ_J or when the increase of the step size $\|\mathbf{s}^{(k)}\|$ is less than a given threshold τ_s.

2.4 Computation of the Jacobian Matrix G_k

In our tests we used central finite difference approximation (FD). The i-th row of the Jacobian matrix $(G_k)_i$ is obtained as:

$$(G_k)_i = \frac{F(U(\mathbf{q}^{(k)} + \varepsilon \mathbf{e}_i)) - F(U(\mathbf{q}^{(k)} - \varepsilon \mathbf{e}_i))}{2\varepsilon}, \quad i = 1, \ldots, N_p$$

where $\mathbf{e}_i$ is the $i-th$ canonical basis vector, and $\varepsilon = 1.e-4$. Each row $(G_k)_i$ requires the solution of two state equations to compute $U(\mathbf{q}^{(k)} + \varepsilon \mathbf{e}_i)$ and $U(\mathbf{q}^{(k)} - \varepsilon \mathbf{e}_i)$. Therefore the number of PDE solutions for each iteration k is $2 \cdot N_p + 1$.

3 Numerical Results

In this section we test the proposed algorithm for the estimation of selected parameters in the time evolution model of Butane (C_B), Oxygen (C_O) and Chloroform (C_{CF}) concentrations in a column bioreactor. Taking advantage of symmetry, the problem is solved along one section of the spatial domain. The model

is given by the following system of diffusion transport and reaction equations representing the concentrations in time $t \in [0, T]$ and space variable $x \in [a, b]$:

$$\begin{cases} \delta_B \dfrac{\partial C_B}{\partial t} = -V \dfrac{\partial C_B}{\partial x} + (D_B + \alpha_L V) \dfrac{\partial^2 C_B}{\partial x^2} \\ \delta_O \dfrac{\partial C_O}{\partial t} = -V \dfrac{\partial C_O}{\partial x} + (D_O + \alpha_L V) \dfrac{\partial^2 C_O}{\partial x^2} + K_O C_O \cdot \\ \delta_{CF} \dfrac{\partial C_{CF}}{\partial t} = -V \dfrac{\partial C_{CF}}{\partial x} + (D_{CF} + \alpha_L V) \dfrac{\partial^2 C_{CF}}{\partial x^2} \end{cases} \tag{8}$$

The parameter α_L represents the longitudinal dispersivity, V is the water velocity, δ_B and δ_{CF} are the Butane and Chloroform retardation factors and K_O is the abiotic oxygen consumption rate. The parameters D_B, D_O and D_{CF} represent the molecular diffusivities in water ($D_B = 1.03e-9\ m^2 s^{-1}$, $D_O = 2.5e-9\ m^2 s^{-1}$, $D_{CF} = 1.e-9\ m^2 s^{-1}$). Inlet boundary conditions are given by $C_B(t, a) = B_1 p(t, 0.625, 1.875)$, $C_O(t, a) = C_1 p(t, 4, 5.625)$, $C_{CF}(t, a) = CF_1 p(t, 7.125, 8.125)$

where $p(t, \tau_1, \tau_2)$ represents the unit smoothed pulse function:

$$p(t, \tau_1, \tau_2) = \begin{cases} 1/(1 + e^{-(t-\tau_1)/\tau}) & t \in [\tau_1 - \Delta_\tau, \tau_1 + \Delta_\tau], \quad \Delta_\tau = 0.321 \\ 1 & t \in (\tau_1 + \Delta_\tau, \tau_2 - \Delta_\tau) \\ 1 - 1/(1 + e^{-(t-\tau_2)/\tau}) & t \in [\tau_2 - \Delta_\tau, \tau_2 + \Delta_\tau] \\ 0 & \text{otherwise} \end{cases}$$

Free flow boundary conditions are assumed at the outlet:

$$\frac{\partial}{\partial x} C_B(t, b) = 0, \frac{\partial}{\partial x} C_O(t, b) = 0, \frac{\partial}{\partial x} C_{CF}(t, b) = 0$$

and homogeneous initial conditions are assumed ($C_B(0, x) = 0, C_O(0, x) = 0, C_{CF}(0, x) = 0$).

The test problem is obtained by solving the state equation (8) in the domain $[a, b] \times [0, T]$ with $a = 0$, $b = 2$ and $T = 15$, using the Crank Nicolson method on a mesh of $M = X_s \times T_s$ uniformly spaced points. The measurements $\mathbf{y}$ are obtained by sampling on a uniform grid, with $N = T_m \times N_m$ points, each component (C_B, C_O, C_{CF}) of the solution of (8), computed with the parameter vector $\mathbf{q} = [V, \alpha_L, \delta_B, \delta_{CF}, K_O, B_1, O_1, CF_1]$, reported in table 2.

Table 2. Value of the parameters used to obtain measurement data $\mathbf{y}$

Parameter	Units	Value	Parameter	Units	Value
V	md^{-1}	0.75	K_O	d^{-1}	0.035
α_L	m	0.12	B_1	$molm^{-3}$	0.26
δ_B	–	1.14	O_1	$molm^{-3}$	0.47
δ_{CF}	–	1.01	CF_1	$molm^{-3}$	3.4e-3

Table 3. Results obtained with 40×12 measurements by changing the starting guess $\mathbf{q}^0$ as in (9). The state equation is solved using a mesh size $M = 257 \times 128$.

δ_0	k	ℓ	$\|\mathbf{r}^k\|$	$\|\mathbf{q}^* - \mathbf{q}^k\|/\|\mathbf{q}^*\|$
8.e-2	5	95	5.607516e-3	4.647818e-4
1.e-1	5	92	5.607516e-3	4.647818e-4
5.e-1	7	107	5.607516e-3	4.647814e-4
6.e-1	8	123	5.607516e-3	4.647789e-4
7.e-1	12	182	5.607516e-3	4.647818e-4
8.e-1	3	71	4.603204e-1	8.000000e-1
9.e-1	3	71	5.018664e-1	9.000000e-1
1.2	4	89	3.544192e-1	1.304530

In the following paragraphs we report the results of experiments to test the algorithm with respect to initial value $\mathbf{q}^0$, mesh size M and data noise. All the experiments were performed using MATLAB (R2010a) on a workstation with 6 Intel(R) Core(TM) i7 processors and 24 GByte ram. In all the tests reported in the following paragraphs, algorithm `TRCLSR` in table 1 has the following tolerance parameters: $\eta = 0$, $\tau_J = 10^{-7}$ and $\tau_s = 10^{-8}$.

3.1 Starting Guess $\mathbf{q}^0$

In this experiment we apply the algorithm `TRCLSR` to estimate the parameters $\mathbf{q}^* = [V, \alpha_L, \delta_B, \delta_{CF}, K_O]$ by changing the starting guess $\mathbf{q}^0$ in order to get an assigned relative error δ_0, i.e. $\|\mathbf{q}^0 - \mathbf{q}^*\|/\|\mathbf{q}^*\| = \delta_0$.

Table 3 shows the relative error (fifth column) and the residual norm (fourth column) obtained with $8\% \leq \delta_0 \leq 120\%$ and mesh size $M = 32896$. The measurements are obtained by uniformly sampling C_B, C_O, C_{CF} on $N = T_m \times N_m$ points with $T_m = 40$ and $N_m = 4$. Computing the starting vector as follows

$$\mathbf{q}^0 = \mathbf{q}^* + \delta_0 \|\mathbf{q}^*\| \eta, \tag{9}$$

where η is a uniform random vector s.t. $\|\eta\| = 1$, we observe that, in order to have an accurate solution, the maximum allowed δ_0 is 70% and this value does not depend on the mesh size used to solve the state equation (8). Figure 1(a) shows the results of the same experiment carried out using meshes of increasing size $8256 \leq M \leq 424571$: it is clear a significant error increase when δ_0 is beyond the percentage allowed.

When $\delta_0 \leq 70\%$ the algorithm converges to the optimal solution with residual norm and relative error independent on $\mathbf{q}^0$.

3.2 Estimate Accuracy

The quality of parameters estimate improves by increasing the accuracy of the solution of the state equation (8). In table 4 are reported the residual norm (fifth column) and relative error (sixth column) obtained by increasing the size of the mesh $T_s \times X_s$ (first and second columns) used to solve (8).

Table 4. Results obtained with 40×12 measurements by changing the mesh size $M = T_s \times X_s$ in the state equation solution

T_s	X_s	k	ℓ	$\|\mathbf{r}^k\|$	$\|\mathbf{q}^* - \mathbf{q}^k\|/\|\mathbf{q}^*\|$
257	128	7	107	5.607516e-3	4.647814e-4
513	256	7	100	9.653799e-4	9.616222e-5
917	463	8	104	1.030716e-4	1.055537e-5
1013	617	7	93	1.394869e-4	8.879562e-6

We can not get the same conclusion by increasing the number of measurements N. As it can be observed in figure 1(b), more than $N = 100$ measurement points do not lead to a sharp decrease of the relative error.

3.3 Noisy Data

In this paragraph we analyze the solution in the presence of noise on the measured data. The noisy data $\mathbf{y}^\delta$ are computed so as to achieve a predetermined level of noise δ: $\mathbf{y}^\delta = \mathbf{y} + \delta\|\mathbf{y}\|\eta$, where η is a random vector with $\|\eta\| = 1$. In table 5 we report the results obtained by solving the state equation with mesh size $M = 424571$ ($X_s = 917, T_s = 463$) and measurements obtained by sampling C_B, C_O, C_{CF} on a uniform grid with $T_m = 40$ and $N_m = 4$ points. The noise δ, reported in column 1, is increased from 0.01% to 10% and we observe the same behavior in the residual norm (column fourth). The quality of the result is still good, as can be observed from the relative error (fourth column table 5) and the graph in figure 3.

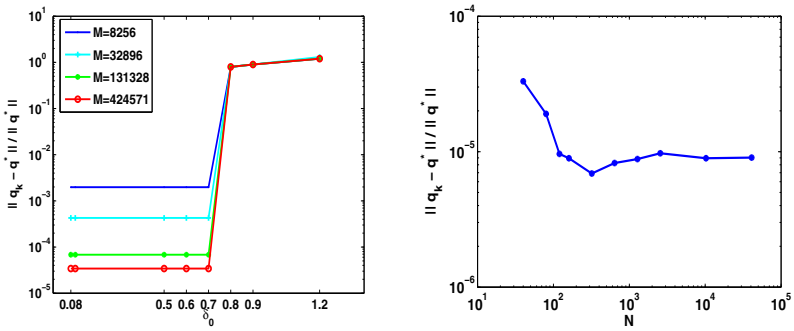
The plots in figure 2 show the relative error and residual convergence history relative to the case $M = 424571$ with noise $\delta = 1.e-3$.

Table 5. Results obtained with noise added to 40×12 measurements and solving (8) with $M = 424571$

δ	k	ℓ	$\|\mathbf{r}^k\|$	$\|\mathbf{q}^* - \mathbf{q}^k\|/\|\mathbf{q}^*\|$
1.e-4	7	98	4.984954e-4	3.761874e-5
1.e-3	7	103	2.404441e-3	1.147139e-4
1.e-2	7	114	2.159983e-2	1.841462e-3
1.e-1	6	111	2.363836e-1	5.460248e-3

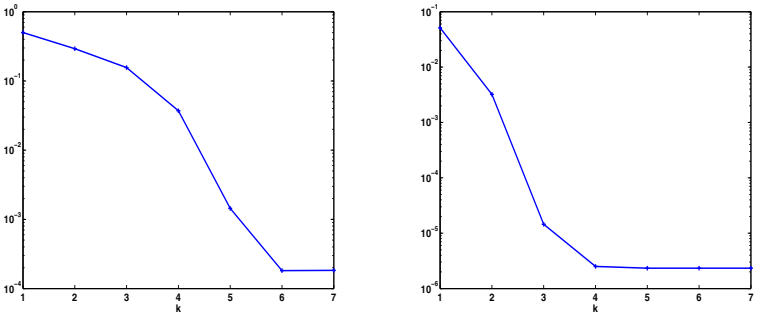
3.4 Algorithm Efficiency

The efficiency of the algorithm can be measured by outer iteration numbers (k) and by the inner iterations ℓ, reported in tables 3, 4 and 5 (columns k and ℓ). We observe a small number of outer iterations (k) with respect to the inner iterations ℓ. The outer iterations (k) are computationally expensive since each step requires $2 \cdot N_p + 1$ solutions of the state equation (8), as shown in paragraph 2.4. Although the number of internal iterations is quite large it is relative to the the solution of a small size linear system $N_p \times N_p$ ($N_p = 5$), so it's generally inexpensive.



(a) (b)

Fig. 1. (a) Relative errors obtained by changing δ_0 and using meshes of incresing size M. (b) Relative Errors obtained by increasing the measurements N (state equation solved with $M = 474571$).



(a) (b)

Fig. 2. Algorithm convergence in the case $M = 424571$ with noise $\delta = 1.e-3$: (a) Relative Error: $\|\mathbf{q}^* - \mathbf{q}^k\|/\|\mathbf{q}^*\|$ (b) Residual norm: $\|\mathbf{r}_k\|$

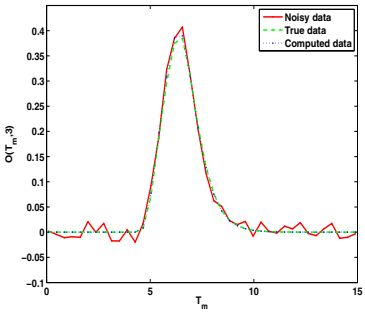


Fig. 3. Results for oxygen concentration at x = 1.1313. The `computed data` are relative to $\mathbf{q}_k$ parameters, the `noisy Data` are obtained by $\mathbf{y}^\delta$ with $\delta = 10\%$ and `true data` are computed using $\mathbf{q}^*$.

4 Conclusions

We can conclude that the algorithm `TRCLSR` computes accurate estimates of the parameters of the model (8), with up to 70% relative error on the initial guess. Furthermore, studies with data affected by noise show that the algorithm can determine accurate solutions with residual norm related to the level of added noise.

Future work will focus on the use of experimental data and more complex nonlinear models. The proposed optimization algorithm is independent of the type of discretization used to solve the state equation, therefore different PDE solutors will be tested.

References

1. Loli Piccolomini, E., Zama, F.: An Iterative algorithm for large size Least-Squares constrained regularization problems. Applied Mathematics and Computation (217) (2011)
2. Nash, J., Sopher, A.: Linear and nonlinear programming. McGraw Hill (1996)
3. Englezos, P., Kalogerakis, N.: Applied Parameter Estimation for Chemical Engineers. Marcel Dekker, New York (2001)
4. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer (2006)
5. Beckrt, R., Braack, M., Vexler, B.: Compust. Theory and Modelling (8) (2004)
6. Chavent, G.: Nonlinear Least Squares for Inverse Problems. Springer (2009)
7. Tikhonov, A.N., Arsenin, V.Y.: Solutions of Ill–Posed Problems. John Wiley & Sons (1977)
8. Frascari, D., Cappelletti, M., Fedi, S., Verboschi, A., Ciavarelli, R., Nocentini, M., Pinelli, D.: Application of the growth substrate pulsed feeding technique to a process of chloroform aerobic cometabolism in a continuous-flow sand-filled reactor. Process Biochemistry (2011), doi:10.1016/j.procbio.2011.08.019

N Dimensional Crowd Motion

Jean-Paul Zolésio[1] and Paola Goatin[2]

[1] CNRS and INRIA, CNRS-INLN, 1136 route des Lucioles,
06902 Sophia Antipolis Cedex, France
Jean-Paul.Zolesio@inln.cnrs.fr
[2] INRIA Sophia Antipolis - Méditerranée, 2004,
route des Lucioles - BP 93, 06902 Sophia Antipolis Cedex, France
paola.goatin@inria.fr

Abstract. We propose a variational formulation of a macroscopic model for crowd motion involving a conservation law describing mass conservation coupled with an eikonal equation giving the flow direction. To get a self contain paper we recall many results concerning flow mapping and convection process associated with non smooth vector field V.

1 The Crowd Motion Problem

We consider a set $\Omega \subset \mathbb{R}^N$ representing a room and we denote by $\Gamma = \Gamma_w \cup \Gamma_o$ its boundary, where Γ_w represents the solid wall, and Γ_o the doors (we also assume $\Gamma_w \cap \Gamma_o = \emptyset$). The natural setting for the problem considered in this paper is $N = 2$, but the 3D problem arises for example for fishes in an aquarium or flock of birds in a portion of sky.

The aim of this paper is to present a variational formulation of a model describing the motion of pedestrians in a finite set Ω. The control parameter of the crowd dynamics is the *speed vector field* V (which is time dependent) and equation (2.1) below expresses the conservation of the pedestrian mass $\int_\Omega \rho(t,x)\, dx$, where $\rho = \rho(t,x)$ denotes the pedestrian density. We study the dynamic system on a time interval $I = (0,\tau)$, the final time τ being arbitrary.

2 Crowd Motion

We denote by $v \in \mathbb{R}^N$ the *velocity*, which is the norm of *speed vector* V, i.e. $v(t,x) = |V(t,x)|$. The conservation of the mass, ρ being the density, is classically expressed by the following equation in conservation form

$$\begin{aligned} &\rho_t + div(\rho\, V) = 0, \text{ in } I \times \Omega, \\ &\rho(0) = \rho_0, \\ &V.n = 0, \qquad\qquad \text{on } \Gamma_w, \\ &V = v\, \boldsymbol{n}, \quad v \geq 0, \quad \text{on } \Gamma_o. \end{aligned} \tag{2.1}$$

D. Hömberg and F. Tröltzsch (Eds.): CSMO 2011, IFIP AICT 391, pp. 557–566, 2013.

The "crowd rheology" modeling is done in two steps. First, we impose that v is a decreasing function of the density ρ, which means that the pedestrian is going faster when there are less people around him. We assume given a decreasing function f (we will consider as an example the function $f(\rho) = 1 - \rho$, so that we have

$$v = f \circ \rho. \tag{2.2}$$

Secondly, we want to take into account the fact that pedestrians try to minimize their travel time. As a consequence, they prefer avoiding high density regions, where they would proceed at low velocity. This behavior can be recovered by means of an eikonal equation whose running cost is given by the reciprocal of the velocity, as proposed by [3]. More precisely, we impose that there exists some potential function Φ, which solves

$$\begin{aligned} &\|\nabla \Phi\| = \frac{1}{f(\rho)}, \text{ in } \Omega, \\ &\Phi = 0, \qquad\qquad \text{on } \Gamma_o, \end{aligned} \tag{2.3}$$

such that

$$V = f^2(\rho)\ \nabla \Phi. \tag{2.4}$$

Since the geometrical domain Ω is assumed to be simply connected, the existence of Φ such that (2.4) holds implies the following curl free condition (assuming that f never reaches zero):

$$curl\left(\frac{V}{f^2(\rho)}\right) = 0 \tag{2.5}$$

We observe that $curl\left(\frac{V}{f^2(\rho)}\right) = f^{-2}\, curl V\ +\ \nabla(\, f^{-2}\,) \times V$, then (2.5) is equivalent to

$$f(\rho)\, curl V - 2 f'(\rho)\ \nabla \rho \times V\ =\ 0, \tag{2.6}$$

which, taking $f = 1 - \rho\ + \kappa$ (for some constant $\kappa > 0$), simplifies to the following *bilinear condition*:

$$\boxed{(1-\rho)\, curl V\ +\ 2\ \nabla \rho \times V\ =\ 0}. \tag{2.7}$$

Concerning the boundary condition for the vector field V, we shall assume that the initial density ρ_0 is compactly supported inside the domain Ω so that during the time τ, the speed of the crowd being bounded, no pedestrian will reach the boundary so that without any loss of generality and for sake of simplicity we shall assume $V.n = 0$ on the wall Γ_w and $||V|| \leq v_{max}$, where $v_{max} > 0$ is the maximum speed for a pedestrian. The "strong" boundary bondition $\Phi = 0$ on Γ_0 is lost in this process but is preserved in weak form as, from $V = v\boldsymbol{n}$ on Γ_0, we get $\Phi = c^{te}$ on Γ_0.

3 Speed Vector V

We assume Ω to be a bounded domain in $\mathbb{R}^N$ with "smooth boundary" Γ. We consider a vector field $V \in L^1(I, L^1(\Omega; \mathbb{R}^N))$ with divergence $divV \in L^1(I, L^1(\Omega))$ and normal component $V.n = 0$ at the boundary (as an element of $W^{-1,1}(\Gamma)$). As a definition we set $E(V)$ the family of such L^1 vector fields V. We denote by $V(t)$ the partial mapping $x \to V(t)(x) = V(t, x)$.

3.1 Regularization

We assume now Ω to be star shaped, and without any loss of generality we assume that $0 \in \Omega$ and the domain to be star shaped with respect to 0. (In fact in all what follows in this section it suffices the domain to be locally star shaped.) We denote by V^e the extension of V to R^N by zero outside of Ω. And we set

$$\bar{V}_n(t, x) = V^e\,(t,\ (1 + 1/n)\, x)\,,$$

which is compactly supported in Ω. Let λ_n be a mollifier and consider

$$V_n(t) = \lambda_n \star \bar{V}_n(t) \in C_c^\infty(\Omega, \mathbb{R}^N))$$

We assume the mollifier suitably chosen, so that $V_n(t)$ is also compactly supported in Ω. We get $div\bar{V}_n(t, x) = (1 + 1/n)\,(divV)^e(t, (1 + 1/n)\, x\,)$

$$divV_n(t) = (1 + 1/n)\, \lambda_n \star div\bar{V}_n$$

So that we have the following strong convergences

$$V_n \to V \ in\ L^1(I, L^1(\Omega, \mathbb{R}^N));\quad divV_n \to divV \ in\ L^1(I, L^1(\Omega))$$

4 Flow Mapping

Consider $V \in L^1(I, C^1(\Omega) \cap H_0^1(\Omega))$. We prove here, following [6], that the mapping mapping $T_t(V)$ defined over the bounded domain Ω .

4.1 Existence

Let $X \in \Omega$, we consider the sequence

$$\begin{aligned} x_0(t) &= X \\ x_1(t) &= X + \int_0^t V(s, x_0(s))ds \\ &\vdots \\ x_{n+1}(t) &= X + \int_0^t V(s, x_n(s))ds \end{aligned} \tag{4.1}$$

As a first result we have $x_n(t) \in \bar{\Omega}$. We apply Ascoli compactness theorem to the family $x_n(.) \in C^0(I, \bar{\Omega})$. We verify the equicontinuity of this family at any $t \in I = [0, \tau]$:

$$x_n(t+\varepsilon) - x_n(t) = \int_t^{t+\varepsilon} V(s, x_n(s))\, ds$$

$$\forall n,\ \ ||x_n(t+\varepsilon) - x_n(t)|| \ \leq \ \int_t^{t+\varepsilon} \|V(s)\|_{L^\infty(\Omega,\mathbb{R}^N)}\, ds$$

Then this sequence converges in $C^0(I, \bar{\Omega})$ to an element $x(t)$, passing to the limit in (4.1) we observe that $x(t)$ is a solution to the flow equation.

4.2 Uniqueness

Assume x^i, $i = 1, 2$ are two solutions. We set $y(t) = x^2(t) - x^1(t)$, we get

$$y(t) = \int_0^t [\int_0^1 DV(s, x^1(s) + \lambda y(s)\,)\, d\lambda\,].y(s)\, ds$$

Then

$$\forall t \in I, \quad \|y(t)\| \leq \int_0^t \|DV(s)\|_{L^\infty}\, \|y(s)\|\ ds \ \leq \ \max_s \|y(s)\| \int_0^t \|DV(s)\|_{L^\infty}\ ds$$

Choose t_V such that $k = \int_0^{t_V} \|DV(s)\|_{L^\infty} ds < 1$, then we get $y = 0$ on $[0, t_V]$. Then the solution is unique on this interval $[0, t_V]$. Now the interval I can be decomposed in a finite number of such intervals, then the solution is unique on I.

Let $X \in \Omega$, we set $T_t(V)(X) = x(t)$, and $T(V)$ denotes the mapping $(t, X) \mapsto T_t(V)(X)$

Proposition 41. *Let $V \in L^1(I, C^1(\Omega) \cap H_0^1(\Omega))$, the flow mapping $T_t(V)$ is defined for any $t \leq \tau$ and $T_t(V) \in C^1(\Omega)$. It is invertible and $T_t(V)^{-1} = T_t(V^t)$ where $V^t(s) = -V(t-s)$, so that $T_t(V)^{-1} \in C^1(\Omega)$.*

4.3 Convection

Let $\zeta_0 \in L^1(\Omega)$ and set $\zeta(t) = \zeta_0 \circ T_t(V)^{-1}$. This function solves the convection problem

$$\zeta_t \ + \nabla\zeta(t).V(t) = 0,\ \zeta(0) = \zeta_0 \tag{4.2}$$

Moreover, as $\zeta_t = -div(\zeta V) + \zeta\, divV(t) \in L^1(I, W^{-1,1}(\Omega))$, we get $\zeta \in C(I, W^{-1,1}(\Omega))$.

5 Solution to Transport Equation (2.1)

5.1 The Homogeneous Equation

Proposition 51. *Let $V \in L^1(I, L^1(\Omega, \mathbb{R}^N))$ with $divV \in L^1(I \times \Omega)$ and $V.n = 0$ on Γ as an element of $W^{-1,1}(\Gamma)$. Let $\rho_0 \in L^\infty(\Omega)$. Then there exists a solution $\rho \in L^\infty(I \times \Omega)$ to the transport equation (4.2). Moreover, we have the following estimates*

$$\|\rho\|_{L^\infty(I\times D)} \leq \|\rho_0\|_{L^\infty(\Omega)} \tag{5.1}$$

Proof. Let V_n a smooth field strongly converging to V. Let ρ_n be the solution of this equation associated to the smooth vector field V_n, that is:

$$\rho_n(t) = \rho_0 \circ T_t(V_n)^{-1} \tag{5.2}$$

then

$$||\rho_n||_{L^\infty(I\times D)} = ||\rho_0||_{L^\infty(\Omega)}$$

There exists a subsequence which is $\sigma*$ weakly converging to some element ρ verifying $\|\rho\|_{L^\infty(I\times D)} \leq \|\rho_0\|_{L^\infty(\Omega)}$ and we can pass to the limit in the weak formulation :

$$\forall \varphi \in C^\infty(I\times\Omega),\ \varphi(\tau) = 0,\quad \int_0^\tau \int_\Omega \rho_n(-\varphi_t - div(\varphi V_n))\, dx\, dt + \int_\Omega \rho_0 \varphi(0)\, dx = 0$$

□

5.2 The Non-homogeneous Equation

Proposition 52. *Let $V \in L^1(I, L^1(\Omega, \mathbb{R}^N))$ with $divV \in L^1(I \times \Omega)$ and $V.n = 0$ on Γ as an element of $W^{-1,1}(\Gamma)$. Let $\rho_0 \in L^\infty(\Omega)$, $\|\rho_0\| \leq 1$. Assume that $F \in L^1(I, L^\infty(\Omega))$, then there exists a solution $\rho \in L^\infty(I \times \Omega)$ to the transport equation*

$$\rho_t + \nabla\rho.V = F,\ \ \rho(0) = \rho_0 \tag{5.3}$$

Moreover, we have the following estimate

$$\|\rho\|_{L^\infty(I\times D)} \leq \|\rho_0\|_{L^\infty(\Omega)} + \int_0^\tau \|F(s)\|_{L^\infty(\Omega)}\, ds. \tag{5.4}$$

Proof. Let V_n be a smooth field strongly converging to V. Let ρ_n be the solution associated to the smooth vector field V_n, that is:

$$\rho_n(t) = [\rho_0 + \int_0^t F(s) \circ T_s(V_n) ds] \circ T_t(V_n)^{-1} \tag{5.5}$$

then

$$\forall t, \quad \|\rho_n(t)\|_{L^\infty(\Omega)} = \|\rho_0 + \int_0^t F(s) \circ T_s(V_n)\ ds\|_{L^\infty \Omega)}$$

$$\leq ||\rho_0||_{L^\infty(\Omega)} + \int_0^t ||F(s)||_{L^\infty(\Omega)}\ ds. \tag{5.6}$$

Then

$$\|\rho_n(t)\|_{L^\infty(I\times\Omega)} \leq \|\rho_0\|_{L^\infty(\Omega)} + \int_0^\tau \|F(s)\|_{L^\infty(\Omega)}\ ds,$$

The weak formulation gives

$$\varphi(\tau) = 0, \quad \int_0^\tau \int_\Omega \rho_n(-\varphi_t - div(\varphi V_n))\ dxdt + \int_\Omega \rho_0\varphi(0,x)\ dx = \int_0^\tau \int_\Omega F\varphi\ dxdt$$

for all $\varphi \in C^\infty(I \times \Omega)$. Now

$$div(\varphi V_n) = \varphi\ divV_n\ +\ \nabla\varphi.V_n \longrightarrow div(\varphi V) \quad \text{in } L^2,$$

which, together with the weak convergence of ρ_n to some ρ, enables us to pass to the limit and obtain the weak formulation of a solution ρ to equation (5.5). The bound leads to the convergence (up to a subsequence) weakly in $\sigma - *$, and the (weak) limit preserves the estimate. □

6 The Conservation Equation

Proposition 61. *Let $\rho_0 \in L^\infty(\Omega)$ and $V \in L^1(I\times\Omega, \mathbb{R}^N)$, $divV \in L^1(I, L^\infty(\Omega))$, with $V.n = 0$ in $W^{-1,1}(\Gamma)$. Assuming $\|divV\|_{L^1(I,L^\infty(\Omega))} < 1$, there exists a solution $\rho \in L^\infty(I \times \Omega)$ to equation (2.1).*

Proof. Equation (2.1) writes

$$\rho_t + \nabla\rho.V = -\rho\, divV, \quad \rho(0) = \rho_0.$$

Let

$$\rho_t^{n+1} + \nabla\rho^{n+1}.V = -\rho^n\, divV, \quad \rho^{n+1}(0) = \rho_0$$

and

$$\delta_{n,p} = \rho^{n+p} - \rho^n.$$

From **??**, as $\delta_{n,p}(0) = 0$, we get,

$$\begin{aligned}
\|\delta_{n+1,p}\|_{L^\infty(I,L^\infty)} &\leq \|\delta_{n,p}\, divV\|_{L^1(I,L^\infty)} \\
&\leq \|\delta_{n,p}\|_{L^\infty(I,L^\infty)}\, \|divV\|_{L^1(I,L^\infty(\Omega))} \\
&\leq \|\delta_{n-1,p}\|_{L^1(I,L^\infty)}\, \|divV\|^2_{L^1(I,L^\infty(\Omega)} \\
&\ldots \\
&\leq \|\delta_{1,p}\|_{L^\infty(I,L^\infty(\Omega))}\, \|divV\|^n_{L^1(I,L^\infty(\Omega))} \\
&= \|\rho^{p+1} - \rho^1\|_{L^\infty(I,L^\infty(\Omega))}\, \|divV\|^n_{L^1(I,L^\infty(\Omega))}.
\end{aligned}$$

Now

$$
\begin{aligned}
&\|\rho^{p+1}\|_{L^\infty(I,L^\infty(\Omega))}\\
&\leq \|\rho_0\|_{L^\infty(\Omega)} + \|\rho^p\, divV\|_{L^1(I,L^\infty(\Omega))}\\
&\leq \|\rho_0\|_{L^\infty(\Omega)} + \|\rho^p\|_{L^\infty(I,L^\infty(\Omega))}\|divV\|_{L^1(I,L^\infty(\Omega))}\\
&\leq \|\rho_0\|_{L^\infty(\Omega)} + \left(\|\rho_0\|_{L^\infty(\Omega)} + \|\rho^{p-1}\, divV\|_{L^1(I,L^\infty(\Omega))}\right)\|divV\|_{\mathbf{L}^1(I,L^\infty(\Omega)}\\
&\leq \|\rho_0\|_{L^\infty(\Omega)} + \left(\|\rho_0\|_{L^\infty(\Omega)} + \|\rho^{p-1}\|_{L^1(I,L^\infty(\Omega))}\|divV\|_{L^1(I,L^\infty(\Omega))}\right)\|divV\|_{L^1(I,L^\infty(\Omega))}\\
&\ldots\\
&\leq \|\rho_0\|_{L^\infty(\Omega)}\Sigma_{i=0,\ldots,p+1}\|divV\|^i_{L^1(I,L^\infty(\Omega))}
\end{aligned}
$$

We get

$$\forall p,\ \|\rho^p\|_{L^\infty(I,L^\infty(\Omega))} \leq \|\rho_0\|_{L^\infty(\Omega)}\left(1 - \|divV\|_{L^1(I,L^\infty(\Omega))}\right)^{-1}.$$

So $\{\rho^n\}_n$ is a Cauchy sequence in $L^\infty(I, L^\infty(\Omega))$. □

Theorem 61. *Let $\rho_0 \in L^\infty(\Omega)$ and $V \in L^1(I \times \Omega, \mathbb{R}^N)$, $divV \in L^1(I, L^\infty(\Omega))$ with $V.n = 0$ in $W^{-1,1}(\Gamma)$. Then there exists a solution $\rho \in L^\infty(I \times \Omega)$ to equation (2.1).*

Proof. We consider a finite covering of the interval $I = [0, \tau]$ by open intervals $]t_i, t_i + \tau_i[$, $i = 0, \ldots, k$, with $t_0 = 0$, $t_k + \tau_k = \tau$, and such that $\int_{t_i}^{t_i+\tau_i} \|V(t)\|_{L^\infty(\Omega)}\, dt < 1$ for all i. From the next proposition, there exists a solution ρ_1 on the interval $]t_0, t_0 + \tau_0[$ verifying $\rho_1(0) = \rho_0$. This solution is continuous in the following sense:

$$\rho_1 \in C([t_0, t_0 + \tau_0], W^{-1,1}(\Omega)).$$

Then for all $t \in [t_0, t_0+\tau_0]$ the element $\rho_1(t)$ is defined as an element of $W^{-1,1}(\Omega)$, but for a.e. $t \in [t_0, t_0 + \tau_0]$ this element is in $L^\infty(\Omega)$. So we can choose such an element $\tilde{t}_1 \in]t_1, t_0 + \tau_0[$ with $\rho_1(\tilde{t}_1) \in L^\infty(\Omega)$. Then on the interval $I_2 = (\tilde{t}_1, t_1 + \tau_1)$ by the next proposition we built a solution ρ_2, and so on on each interval I_i. We obtain a solution on the whole interval $(0, \tau)$. □

7 Crowd Motion Variational Formulation

Let us denote by I the time interval, $I =]0, \tau[$. To any element $V \in E(\Omega) \subset L^1(I, L^1(\Omega, R^N))$, we associate the set R_V of solutions to the conservation equation (2.1) and we introduce the functionals

(7.1)

$$J(V, \rho) = \|f \circ \rho - |V|\|_{L^1(I,L^1(\Omega))} + \beta\|(1-\rho)\, curlV + 2\nabla\rho \times V\|_{L^1(I,W^{-1,1}(\Omega,\mathbb{R}^N))}$$

(7.2)

$$j(V) = \inf_{\rho \in R_V} J(V, \rho)$$

which can be rewritten as

$$j(V) = \inf_{r \in L^\infty(I, L^\infty \Omega))} \sup_{\theta \in \mathcal{W}} L(V, r, \theta) \tag{7.3}$$

where

$$\mathcal{W} = \left\{ \theta \in L^1(I, W^{1,\infty}(\Omega)) \cap W^{1,1}(I, L^1(\Omega)), : \theta(\tau) = 0 \right\}$$

and

$$L(V, r, \theta) = J(V, r) + \int_I \int_\Omega r \left(\theta_t + \nabla \theta . V\right)) \; dxdt + \int_D \rho_0 \theta(0) \; dx.$$

We have the obvious

Proposition 71. *Let $V \in E(\Omega)$ such that $j(V) = 0$, then it solves the crowd problem in the sense that there exists a solution $\rho \in R_V$ such that (V, ρ) solves the crowd system (2.1), (2.3), (2.4).*

If such a speed vector V exists it minimizes the positive functional j over the space $E(\Omega)$. The variational approach for the crowd problem under consideration is to replace it by the weaker one which is the minimization of the non negative functional j over $E(\Omega)$. Our approach is now to compute the gradient of the functional j to be minimized.

We remark that, if the infimum of the functional j is not zero, then, in some sense, the crowd problem formulated as (2.1), (2.3), (2.4) would have no solution. If the infimum reaches zero it would built a solution.

8 Minimization of the Functional

The main objective is to calculate the gradient of j.

We denote by $j'(V; W) = \liminf_{\varepsilon > 0, \, \varepsilon \to 0} \; j(V + \varepsilon W)$. If the limit exists, it is the classical Gateau semi derivative. For sake of simplicity let us first compute the gradient for a regularized functional j_γ expressed in the following form

$$j_\gamma(V) = j_2(V) + \frac{\gamma}{2} \int_I \int_D \|V(t)\|^2_{\mathcal{H}} dt, \tag{8.1}$$

where $\mathcal{H}$ stands here for a Banach space of function over the domain Ω which will ensure the set R_V of solutions to the conservation equation (2.1) to be a singleton element ρ_V. This will be the case for the following choices :

$\mathcal{H} = \{ V \in H^3(\Omega, R^N), \; \Delta v = 0, \; V.n = v_{max} \; on \; \Sigma \} \subset E(\Omega) \cap C^1(\bar{\Omega})$

or

$\mathcal{H} = \{V \in BV(\Omega, R^N) \; with \; divV \in L^\infty(\Omega)\}$,

and where j_2 is the quadratic version of j, that is $j_2(V) = \inf_{\rho \in R_V} J_2(V, \rho)$ with

(8.2)

$$J_2(V, \rho) = \|f \circ \rho - |V|\|^2_{L^2(I, L^2(\Omega))} + \beta \|(1 - \rho) \, curlV + 2 \nabla \rho \times V\|^2_{L^2(I, L^2(\Omega, \mathbb{R}^N))}$$

We briefly recall now the calculus of the gradient of the functional expressed in Min Max.

8.1 Derivative of a Function in Min Max Form, from [1], [6]

Let E, F be two Banach spaces and $L(s,e,f)$ be a function defined from $[0,1]\times K_E\times K_F$ into $\mathbb{R}$, where K_E, K_F are convex sets, respectively in E and F. Assume that the *Lagrangian* functional L is convex l.s.c. with repect to e , concave u.s.c. with respect to f and continuously differentiable with respect to the parameter s. Assume moreover that there exists a non empty set $S(s)$ of saddle points. Then it always takes the following form:

$$S(s)=A(s)\times B(s),\ A(s)\subset K_E,\ B(s)\subset K_F,\ \textit{such that}:$$

$$\forall a(s)\in A(s),\ \forall b(s)\in B(s),\ \ \forall\gamma\in K_A,\ \ \forall\beta\in K_B,$$

$$L(s,a(s),\beta)\le L(s,a(s),b(s))\le L(t,\gamma,b(s))$$

So that $\forall\gamma'\in K_E,\ \forall\beta'\in K_F$ we have

$$-L(0,\gamma',b(0))\le -L(0,a(0),b(0))\le -L(0,a(0),\beta').$$

By choosing $\gamma=a(0)$, $\beta=b(0)$, $\gamma'=a(s)$, $\beta'=b(s)$, and adding the two previous inequalities we get for any $s>0$:

$$\begin{aligned}\frac{L(s,a(s),b(0))-L(0,a(s),b(0))}{s}&\le\frac{L(s,a(s),b(s))-L(0,a(0),b(0))}{s}\\&\le\frac{L(s,a(0),b(s))-L(0,a(0),b(s))}{s}\end{aligned}$$

Under reasonable smoothness assumptions on L and Kuratowski continuity of the sets $A(s)$ and $B(s)$ we get the semi-derivative of

$$l(s)=\min_{a\in K_E}\max_{b\in K_F}L(s,a,b)$$

i.e.

$$l'(0)=\min_{a\in A(0)}\max_{b\in B(0)}\frac{\partial}{\partial s}L(0,a,b)\tag{8.3}$$

In the following section we shall make use of that semi-derivative in the specific situation in which the set $S(0)$ is reduced to a unique pair, $A(0)=\{y\},\ B(0)=\{p\}$, where y and p will be the "state" and "adjoint-state" solution associated with the wave equation under consideration. In this situation the function l is differentiable at $s=0$ and the derivative (8.3) takes the following form:

$$l'(0)=\frac{\partial}{\partial s}L(0,y,p).$$

8.2 The Optimal System

Just for shortness of the expressions we make $\beta = 0$ and $m = 1$ so that

$$J'_\gamma(V;W) = \int_I \int_\Omega \left(2(|V| - f \circ \rho_V) \frac{V}{|V|}.W + \rho_V \, W.\nabla P \right) dxdt + \gamma \int_I \int_\Omega D\Delta V.D\Delta W \, dxdt,$$

where the adjoint state P is the solution to the following backward adjoint problem:

$$\boxed{P_t + \nabla P.V = 2(f \circ \rho_V - |V|) f' \circ \rho_V, \quad P(\tau) = 0.} \tag{8.4}$$

Obviously the gradient is

$$\boxed{\nabla J_\epsilon(V) = 2(|V| - f \circ \rho_V) \frac{V}{|V|} + \rho_V \nabla P - \gamma \Delta V.}$$

Proposition 81. *If the vector field V minimizes the functional j_γ, then it solves the optimality system*

$$\rho_t + div(\rho \, V) = 0, \qquad \rho(0) = \rho_0,$$

$$P_t + \nabla P.V = -2(1 - \rho + \gamma - |V|), \qquad P(\tau) = 0,$$

$$(2(|V| - f \circ \rho) \frac{V}{|V|} + \rho \nabla P - \gamma \Delta V = 0.$$

$$V(t).n = 0 \text{ on } \Gamma_\omega, \; V(t).n = v_{max} \text{ on } \Gamma_0, \; \Delta V(t) = \Delta^2 V(t) = 0, \; \text{ on } \Sigma$$

References

1. Cuer, M., Zolésio, J.P.: Control of singular problem via differentiation of a min-max. Systems & Control Letters 11(2), 151–158 (1988)
2. Delfour, M.C., Zolésio, J.P.: Shapes and Geometries: Analysis, Differential Calculus and Optimization. SIAM series on Advances in Design and Control. Society for Industrial and Applied Mathematics, Philadelphia (2001) (2nd edn., 2011)
3. Hughes, R.L.: A continuum theory for the flow of pedestrians. Transpn. Res., B 36(6), 507–535 (2002)
4. Sokolowski, J., Zolésio, J.P.: Introduction to shape optimization. Shape sensitivity analysis. Springer Ser. Comput. Math., vol. 16. Springer, Berlin (1992)
5. Zolésio, J.P.: The material derivative (or speed) method for shape optimization. In: Haug, E.J., Céa, J. (eds.) Optimization of Distributed Parameter Structures (Iowa City, Iowa, 1980). NATO Adv. Sci. Inst. Ser. E: Appl. Sci., vol. II, 50, pp. 1089–1151. Sijhofff and Nordhoff, Alphen aan den Rijn, Nijhoff, The Hague (1981)
6. Zolésio, J.P.: Control of moving domains, shape staibilzation and variational tube formulation. Int. series of numerical mathematics, vol. 55, pp. 329–382. Birkhauser Verlag, Basel (2007)

Author Index